Title
Computational Methods for Analyzing RNA Sequencing to Study Post-Transcriptional Gene Regulation

Author
Cass, Ashley Anne

Publication Date
2018

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Computational Methods for Analyzing RNA Sequencing to Study Post-Transcriptional Gene Regulation

A dissertation submitted in partial satisfaction of the requirements for the degree Doctor of Philosophy in Bioinformatics

by

Ashley Anne Cass

2018

ABSTRACT OF THE DISSERTATION

Computational Methods for Analyzing
RNA Sequencing to Study
Post-Transcriptional Gene Regulation

by

Ashley Anne Cass
Doctor of Philosophy in Bioinformatics
University of California, Los Angeles, 2018
Professor Xinshu Xiao, Chair

Since the completion of the Human Genome Project in 2003, massive DNA sequencing efforts enabled gene mapping and enhanced our understanding of genetic variation. However, exactly how the same DNA sequence in every cell of one individual leads to vast biological variation is still not fully understood. In particular, the DNA sequence does not directly contain information regarding which genes are expressed in different cell types, tissues, and disease states. With the advent of high-throughput RNA sequencing (RNA-Seq), gene expression and RNA isoform variation can be assayed cost- and time-efficiently in different conditions. In this work, we aimed to develop computational methods to analyze RNA-Seq for the purpose of elucidating mechanisms of post-transcriptional gene regulation. The first chapter briefly introduces RNA biology, including co- and post-transcriptional gene regulation concepts. The second chapter describes the identification of small cleavage-inducing RNAs and their RNA targets for degradation through bioinformatic integration of small RNA sequencing and

Degradome Sequencing, the latter capturing RNA degradation products. This work revealed an expanded repertoire of small cleavage-inducing RNAs (sciRNAs) and their targets, suggesting that small RNA-mediated cleavage is more widespread than previously appreciated. Post-transcriptional regulation is often mediated by cis-regulatory elements in 5' and 3' untranslated regions (UTRs), including sciRNA target motifs. Thus, alternative transcription start sites (ATSS) and alternative polyadenylation (APA) often impact post-transcriptional gene regulation through the inclusion or exclusion of cis-regulatory elements in UTRs. In chapter three, we describe mountainClimber, a novel method that overcomes several limitations of existing approaches to identify ATSS and APA from RNA-Seq. In chapter four, we applied mountainClimber to thousands of RNA-Seq datasets derived from many human tissues in the largest study of ATSS and APA to date. In chapter five, we applied mountainClimber to chromatin-associated and poly(A)-selected RNA-Seq in murine macrophages with or without previous exposure to an endotoxin. This analysis revealed ATSS, APA, and alternative transcription end sites associated with tolerization of macrophages to endotoxins. Finally, we summarize our conclusions in chapter six.

The dissertation of Ashley Anne Cass is approved.

Amander Therese Clark

Matteo Pellegrini

Xia Yang

Xinshu Xiao, Committee Chair

University of California, Los Angeles

2018

To my family

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

Funding

This work was partially supported by the National Institute of Health (NIH) predoctoral training grant T90DE022734 (2014-2017) and the UCLA Dissertation Year Fellowship (2017-2018).

Chapter contributions

Chapter 2 is a version of the published article: **Cass AA**, Bahn JH, Lee JH, Greer C, Lin X, Kim Y, Hsiao YH, Xiao X. "Global analyses of endonucleolytic cleavage in mammals reveal expanded repertoires of cleavage-inducing small RNAs and their targets." Nucleic Acids Research, 2016. 44(7):3253-63. X.X, A.A.C., C.G. designed the study. A.A.C., J.H.L, C.G. and Y.H.H. analyzed the data. J.H.B. made all sequencing libraries. J.H.B. and X.L. conducted in vitro cleavage assays. Y.K. conducted H1 cell culture. A.A.C. and X.X. wrote the paper with input from other authors.

Chapter 3 and 4 are currently prepared as a single manuscript for submission: **Cass, AA**, Xiao X. "The landscape of alternative transcription start and polyadenylation sites in human tissues." A.A.C and X.X. designed the study and wrote the paper. A.A.C. implemented the software and performed all analysis. I would also like to thank Yi-Wen Yang for sharing the LOESS regression approach used in mountainClimberTest and Giovanni Quinones Valdez for sharing code related to bam file processing.

Chapter 5 describes work as part of the Ribonomics of Gene Regulation project in collaboration with Dr. Alexander Hoffmann's lab at UCLA. A.A.C. carried out all the described data analyses. I would also like to acknowledge Jae Hoon Bahn, Emily Chen, and Eddie Park who generated the RNA-Seq.

Very important people

First and foremost, I must thank Grace, not only as a research advisor, but also as a great mentor and example of a strong leader. I always felt reenergized and ready for the next steps after our meetings and look back fondly on the meetings where we ended up on topics other than science. She always puts her students first, even as she got busier with your own career advancements. Her passion for science is obvious, infectious, and inspiring to all of us.

I would also like to thank my committee members for their advice at each meeting. Matteo Pellegrini has always offered invaluable suggestions for my projects, starting from my undergraduate coursework, through a lab rotation, and ultimately as part of my doctoral committee. Xia Yang provided not only scientific but also very useful career advice, for which I'm very grateful. Thanks also to Amander Clark for inviting me to contribute to one of her lab's projects, and for always providing her expertise and a fresh perspective.

Thomas Graeber, with whom I worked with as an undergraduate researcher, deserves a special thanks. My experience there inspired me to pursue a research career in the first place.

Thanks to all of my coauthors, especially those mentioned above who contributed to the work presented in this dissertation. I would also like to thank Yibin Wang for having me on several of his group's projects. Thanks to Haiyin Chen and Matthew Albert at Genentech for an outstanding internship experience and including me as co-second author for that work.

Thanks to all of my fellow Xiao lab members, who consistently offered help and suggestions and made this experience so memorable. Our lab is as quiet as a library most of the time, but I love that people in neighboring offices can hear our laughs in the breakroom. Thanks especially to Esther Hsiao who I leaned on throughout my whole PhD, and Kiku Koyano who is (almost) always willing to go to Trader Joe's with me for lunch. Thanks to all current and former members Jae Hoon Bahn, Hyun-Ik Jun, Yi-Wen Yang, Kofi Amoah, Christina Burghard, Tracey Chan, Mudra Choudhury, Ting Fu, Giovanni Quinones Valdez, Stephen Tran, Yi-Wei Sun, Boon

Xin Tan, Jae-Hyung Lee, Xianzhi Lin, Yun Yang, Anneke Bruemmer, Jianlin Shao, Adel Azghadi, Elizabeth Chin, and Ru Mi Moon.

I would also like to thank the student affairs officers who helped me throughout this program, especially Pamela Hurley, Allison Taka, and Mandy McWeeney.

Thanks to my friends, without whom I would not have been nearly as happy throughout this program. I've been at UCLA for around 1/3 of my life, and Artur Jaroszewicz has been one of my closest friends during that time. Robert Brown has been a great friend, retreat co-coordinator, and advice-giving peer to me, and my experience here would not have been the same without him. Thanks to Adriana Sperlea, who was a joy to serve with as our program student representatives. Finally, Chelsea Fenerin and Alex Ow deserve a huge thanks for being there for me for about 25 years.

Thank you to all of my family, who I would not be here without. Thanks to Tim for being my wise little brother with a calm confidence that I look up to. A special thanks to Nonnie and Papa, Theresa and Frank Cardinale, who gave me a giant Italian family and were not just grandparents, but my second set of parents growing up. I wish you could be here, and I think about you every day.

Very special thanks to Kyle Bayles, the person who I met and created a home with during my PhD. Thank you for all of your support. Coming home to you made the weight of a PhD much lighter. I don't think I will ever be able to fully repay you, but I will certainly always try.

Finally, a million thanks to my parents, Marlowe and Mary. Your support and encouragement throughout my life has gotten me to this point. It's hard to find the words to express my gratitude. Dad – thank you for always being there, especially with your sense of humor. Mom – I am becoming more like you every day, and I am so proud of that, even if it means becoming a worry wart. Knowing you both were always there to support me gave me the courage and drive to pursue this path.

VITA

2007-2011 B.S., Computational and Systems Biology
University of California, Los Angeles

2009-2012 Undergraduate Research Assistant, Laboratory of Thomas Graeber
University of California, Los Angeles

2015 Teaching Assistant, PhySci125: Molecular Systems Biology
University of California, Los Angeles

2017 Intern, Cancer Immunology
Genentech, South San Francisco

2012-2018 Ph.D. Candidate, Bioinformatics IDP
University of California, Los Angeles

AWARDS

2017-2018 UCLA Dissertation Year Fellowship
University of California, Los Angeles

2014-2017 NIH/NIDCR T90 Oral Health-Scientist Training Program
University of California, Los Angeles

2016 RNA Conference 2016 Travel Fellowship
Kyoto, Japan

2015 Poster Award, UCLA Quantitative and Computational Biosciences Institute
Lake Arrowhead, California

PUBLICATIONS

Cass AA, Xiao X. The landscape of alternative transcription start and polyadenylation sites in human tissues. In preparation.

Kong Y, Rose C*, **Cass A***, Darwish M, Lianoglou S, Haverty P, Tong AJ, Blanchette C, Mellman I, Bourgon R, Greally J, Jhunjhunwala S, Albert M, Chen-Harris H. Transposable element expression in tumors is associated with immune infiltration and increased antigenicity. bioRxiv, 2018. 10.1101/388215 (*Contributed equally)

Touma M, Kang X, Gao F, Zhao Y, **Cass AA**, Biniwale R, Xiao X, Eghbali M, Coppola G, Reemsten B, Wang Y. Wnt11 regulates cardiac chamber development and disease during perinatal maturation. JCI Insight, 2017. 2(17). 10.1172/jci.insight.94904

Graham NA*, Minasyan A*, Lomova A, **Cass A**, Balanis NG, Friedman M, Chan S, Zhao S, Delgado A, Go J, Beck L, Hurtz C, Ng C, Qiao R, Ten Hoeve J, Palaskas N, Wu H, Müschen M,

Multani AS, Port E, Larson S, Schultz N, Braas D, Christofk H[#], Mellinghoff I[#], Graeber TG. Recurrent patterns of DNA copy number alterations in tumors reflect metabolic selection pressures. Molecular Systems Biology, 2017. 13(2):914. 10.15252/msb.20167159 (*Contributed equally, [#]Contributed equally)

Wang Z, Zhang XJ, Ji XY, Zhang P, Deng KQ, Gong J, Ren S, Wang X, Chen I, Wang H, Gao C, Yokota T, Ang YS, Li S, **Cass A**, Vondriska TM, Li G, Deb A, Srivastava D, Yang HT, Xiao X, Li H, Wang Y. The long noncoding RNA Chaer defines an epigenetic checkpoint in cardiac hypertrophy. Nature Medicine, 2016. 22(10):1131-1139. 10.1038/nm.4179

Touma M, Kang X*, Zhao Y*, **Cass AA**, Gao F, Biniwale R, Coppola G, Xiao X, Reemsten B, Wang Y. Decoding the long noncoding RNA during cardiac maturation: a roadmap for functional discovery. Circulation Cardiovascular Genetics, 2016. 9(5):395-407.10.1161/CIRCGENETICS. 115.001363 (*Contributed equally)

Cass AA, Bahn JH, Lee JH, Greer C, Lin X, Kim Y, Hsiao YHE, Xiao X. Global analyses of endonucleolytic cleavage in mammals reveal expanded repertoires of cleavage-inducing small RNAs and their targets. Nucleic Acids Research, 2016. 44(7):3253-63. 10.1093/nar/gkw164

Hsiao YHE, **Cass AA**, Bahn JH, Lin X, Xiao X. Global approaches to alternative splicing and its regulation – recent advances and open questions. In Transcriptomics and Gene Regulation (ed. Jiaqian Wu) (Springer Netherlands, 2016)

Li Z, Yu J, Hosohama L, Nee K, Gkountela S, Chaudhari S, **Cass AA**, Xiao X, Clark AT. The Sm protein methyltransferase PRMT5 is not required for primordial germ cell specification in mice. EMBO Journal, 2014. 34(6):748-58. 10.15252/embj.201489319

Escuin-Ordinas H, Atefi M, Fu Y, **Cass A**, Ng C, Huang RR, Yashar S, Comin-Anduix B, Avramis E, Cochran AJ, Marais R, Lo RS, Graeber TG, Herschman HR, Ribas A. COX-2 inhibition prevents the appearance of cutaneous squamous cell carcinomas induced by BRAF inhibitors. Molecular Oncology, 2014. 8(2):250-60. 10.1016/j.molonc.2013.11.005

Ahler E*, Sullivan WJ*, **Cass A**, Braas D, York AG, Bensinger SJ, Graeber TG, Christofk HR. Doxycycline alters metabolism and proliferation of human cell lines. PLoS One, 2013. 8(5):e64561. 10.1371/journal.pone.0064561 (*Contributed equally)

Knight DA, Ngiow SF, Li M, Parmenter T, Mok S, **Cass A**, Haynes NM, Kinross K, Yagita H, Koya RC, Graeber TG, Ribas A, McArthur GA, Smyth MJ. Host immunity contributes to the anti-melanoma activity of BRAF inhibitors. Journal of Clinical Investigation, 2013. 123(3):1371-81. 10.1172/JCI66236

Chapter 1: Introduction

The central dogma of molecular biology describes the process of expressing genetic information from DNA to RNA to protein. In humans, the DNA sequence is three billion base pairs long, of which only 1-2% is comprised of 20,000 functional units called genes that code for proteins. Genes are transcribed into RNA messages that function as templates for protein production. RNA is further processed after transcription and before translation, including 5' cap addition, intron removal, and finally 3' cleavage and polyadenylation. After cleavage and polyadenylation, the RNA is localized, e.g. to the cytoplasm in the case of coding RNA. Although the DNA sequence is the same in every cell of an individual, differential gene expression and RNA processing lead to different RNA and protein contents in different biological contexts.

Since the human genome project was completed [1,2], many efforts have focused on how genetic information is regulated to produce the observed variation in RNA and protein content. With the advent of high throughput sequencing (HTS) and mass spectrometry technologies, large-scale studies of the genome, transcriptome, and proteome have begun to characterize this variation. While it was previously accepted that 1-2% of the genome is functional and the rest of the DNA is "junk", The Encyclopedia of DNA Elements (ENCODE) project applied various HTS technologies to assay RNA, transcription factor binding on chromatin, chromatin structure, and histone modifications and concluded that 80% of the genome is functional and regulates the 1-2% of coding genes [3]. Much of these transcripts function in roles other than coding for a protein. Several long non-protein-coding RNAs (lncRNAs) function in heterochromatin formation and gene expression regulation (reviewed in [4]). More recently, thousands of circular RNAs were discovered from RNA-Seq analyses and regulate gene expression, transcription, and splicing (reviewed in [5]). Small RNAs, such as microRNAs which are typically 21-22 nucleotides long, specifically regulate mRNAs and lncRNAs through reverse complementary sequence matching. Nascent RNA,

including upstream antisense RNAs and enhancer RNAs, functions in promoter definition, gene expression enhancement, and recruitment of transcription factor and regulatory proteins (reviewed in [6]). These discoveries, greatly facilitated by RNA-Seq, shifted the paradigm of RNAs from simple messengers between DNA and protein to complex molecules with a diverse array of functions.

The protein-coding potential and stability of RNA are often determined by alternative RNA processing. Alternative promoters, or alternative transcription start sites (ATSSs), often affect translation through addition or removal of cis-regulatory sequence and structural motifs (reviewed in [7]). Alternative transcription termination also functions in gene regulation; early termination can induce RNA degradation, while late termination may inhibit expression of a nearby downstream gene (reviewed in [8]). Alternative 3' cleavage and polyadenylation results in the alternative inclusion of 3'UTR sequences, which contain many cis-regulatory elements, leading to changes in RNA stability, localization, nuclear export, and translation, as well as protein localization and stability (reviewed in [9,10]). These cis-acting elements, including microRNA-binding sites, are enriched between proximal and distal polyadenylation sites [11]. Additionally, the 3'UTR is under higher selective pressure than other gene regions, supporting its functionality [12]. Thus, ATSS and APA are functionally important in gene regulation. In this work, we aim to better understand some of these mechanisms of post-transcriptional gene regulation.

Chapter two investigates the repertoires of small cleavage-inducing RNAs and their cognate targets in different murine tissues. Although 60% of protein-coding genes are microRNA targets [13], few are known to induce cleavage of target RNAs. However, the catalytic cleavage activity of Ago2, the small RNA-binding component of the RNA induced silencing complex, is conserved in mammals suggesting that small RNA-mediated cleavage may be more widespread than previously appreciated [14]. Through integrative analysis of small RNA-Seq and Degradome-Seq, we identified hundreds of small cleavage-inducing RNAs and their cognate targets, not

limited to microRNAs, demonstrating that small RNA-guided cleavage is more widespread than previously appreciated.

ATSS and APA regulate the majority of human genes, often in a cell type-specific manner. Motivated by the plethora of publicly available RNA-Seq datasets, we developed a novel algorithm, mountainClimber, for ATSS and APA identification from RNA-Seq. Our approach, described in chapter 3, overcomes several limitations of similar existing methods; it is fully de novo, identifies multiple change points per gene, and detects change point downstream of the first exon and upstream of the last exon.

Chapter 4 describes the application of mountainClimber to thousands of samples from human tissues generated by the GTEx consortium. Although previous studies identified tissue-specific APA in RNA-Seq, they were limited to annotated poly(A) sites. In contrast to APA, tissue-specific ATSS is less well studied. In the largest study of simultaneous ATSS and APA prediction to date, we describe the landscape of ATSS and APA in humans. Tissue specificity was the main driver of observed 5' and 3' end variation, and there were thousands of significantly differential ATSS and APA events.

In chapter 5, we describe ATSS and APA identification in murine macrophages with or without re-exposure to an endotoxin. While exposure to a new toxin induces widespread changes in macrophage gene expression, re-exposure to the same toxin is characterized by poor induction of those same genes. We identified alternative 5' and 3' ends in both chromatin-associated and poly(A)-selected RNA-Seq in tolerized (re-exposed to an endotoxin) and naïve macrophages. Because mountainClimber is robust to high levels of RNA-Seq non-uniformity, we were able to predict change points in chromatin-associated RNA-Seq for the first time to our knowledge. This enabled identification of alternative transcription termination events in the chromatin-associated fraction in addition to ATSS and APA in poly(A)-selected RNA. We identified many alternative 5' and 3' events, suggesting that alternative RNA processing contributes to the tolerized phenotype.

Chapter 2: Global analyses of endonucleolytic cleavage in mammals reveal expanded repertoires of cleavage-inducing small RNAs and their targets

2.1. Abstract

In mammals, small RNAs are important players in post-transcriptional gene regulation. While their roles in mRNA destabilization and translational repression are well appreciated, their involvement in endonucleolytic cleavage of target RNAs is poorly understood. Very few microRNAs are known to guide RNA cleavage. Endogenous small interfering RNAs are expected to induce target cleavage, but their target genes remain largely unknown. We report a systematic study of small RNA-mediated endonucleolytic cleavage in mouse through integrative analysis of small RNA and degradome sequencing data without imposing any bias toward known small RNAs. Hundreds of small cleavage-inducing RNAs and their cognate target genes were identified, significantly expanding the repertoire of known small RNA-guided cleavage events. Strikingly, both small RNAs and their target sites demonstrated significant overlap with retrotransposons, providing evidence for the long-standing speculation that retrotransposable elements in mRNAs are leveraged as signals for gene targeting. Furthermore, our analysis showed that the RNA cleavage pathway is also present in human cells but affecting a different repertoire of retrotransposons. These results show that small RNA-guided cleavage is more widespread than previously appreciated. Their impact on retrotransposons in non-coding regions shed light on important aspects of mammalian gene regulation. [*]

---

[*] The work appearing in this chapter is published: Cass AA, Bahn JH, Lee JH, Greer C, Lin X, Kim Y, Hsiao YH, Xiao X. Global analyses of endonucleolytic cleavage in mammals reveal expanded repertoires of cleavage-inducing small RNAs and their targets. Nucleic Acids Research, 2016. 44(7):3253-63. 10.1093/nar/gkw164

2.2. Introduction

In mammals, the best known small RNA targeting pathways include destabilization or translational repression of target mRNAs [15,16]. A third mechanism, small RNA-guided endonucleolytic cleavage of target RNAs, is assumed to be very rare in animals, although it is prevalent in plants [17]. Thus far, only a small number of microRNAs (miRNAs) were predicted to have this function in mammals [18–22], affecting a very small number of target genes. Endogenous small interfering RNAs (endo-siRNAs) are expected to induce target cleavage (reviewed in [23]). However, their targetome is not yet well characterized.

The catalytic function of Ago2, which carries out the slicing reaction on mRNA targets, is highly conserved throughout mammals [14]. This observation suggests that small RNA-directed cleavage may be an essential aspect of mammalian gene regulation and more widespread than currently appreciated. Three factors may have hindered progress in this research area. One is the possibility that small RNA-directed cleavage is highly cell type-specific. The specific cell types examined by previous studies may have failed to reveal the bulk of such events. Second, a diverse panel of small RNAs, not limited to miRNAs or siRNAs, may mediate mRNA cleavage, an aspect that has not been explored. Third, technical challenges, such as the enrichment of repetitive elements in the target sites or small RNAs, may have prevented discovery of the full spectrum of small RNA-mediated cleavage events.

Our study aimed to address the above challenges and better characterize small RNA-mediated cleavage in mammals. We analyzed a large amount of small RNA and Degradome Sequencing data (Deg-Seq, also known as PARE), with the latter capturing the 5' ends of RNA degradation products [24,25], in mouse embryonic stem cells (mESCs), testis and cerebellum. This analysis allowed a systematic characterization of small cleavage-inducing RNAs (sciRNAs) and their targets simultaneously. Our bioinformatic method captures any type of sciRNAs, unlimited to known RNA classes, and accommodates existence of repetitive sequences in the RNA. As a

result, we identified 398 sciRNAs and 810 cognate cleavage target genes, much more than previously known in the literature. Interestingly, about 40% of sciRNAs overlap known miRNAs, endo-siRNAs or piwi-interacting RNAs (piRNAs), revealing novel targets of these RNA regulators. This observation also indicates that sciRNAs, defined to conveniently refer to their function, may have diverse biogenesis pathways. sciRNAs demonstrated a high degree of cell type-specificity, developmental stage-specificity, and diversity in possible functional pathways. A striking feature of both sciRNAs and their target cleavage sites is their significant overlap with retrotransposable elements, providing evidence for the first time that retrotransposons in transcripts are leveraged as signals for gene targeting. Additionally, our analysis showed that the sciRNA pathway is also present in human cells but affecting a different repertoire of retrotransposons. Thus, sciRNA targeting is a conserved mechanism between human and mouse but involves different sciRNA molecules and targets, possibly reflecting the divergence of retrotransposons between the two genomes.

2.3. Methods

Bioinformatic prediction of sciRNAs and their targets

Preprocessing: Deg-Seq and small RNA-Seq reads were trimmed with cutadapt [26] to remove adapters and PCR primers. For mESCs, Deg-Seq and small RNA-Seq datasets were acquired from GSE21975 [22] and GSE35368 (SRR402760, SRR402761, SRR402762, SRR402766) [27], respectively, while other datasets were generated in-house. 3' end regions with quality less than 20 were also trimmed from Deg-Seq reads. A minimum length of 19 nt was required for small RNA-Seq since typical known small RNAs are longer than 19nt. A minimum length of 25 nt was required for Deg-Seq reads to ensure specific mapping to the genome while retaining as many reads as possible. The first step of the pipeline was the exclusion of small RNA-Seq and Deg-Seq reads with low complexity since such reads tend to base-pair with each other by random

chance. Low complexity reads were defined as those with tandem repeats of mono-, di-, tri-, or quad-nucleotides of 5, 3, 2, 2 respectively. The length cutoffs were determined by examining repeat patterns of known functional small RNAs. Small RNA sequences were required to have length 19-24nt and read count ≥20.

Gene Annotation: To define a comprehensive set of annotated mRNAs, we merged the following gene annotation databases: Ensembl, UCSC knowngene, RefSeq, VegaGene, GENCODE, Pseudogene.org, and NONCODEv4 [28].

Define significant peaks: Deg-Seq reads were aligned only to annotated regions (listed above) of genome mm10 or hg19 using Bowtie v.0.12.7 [29] requiring no mismatches and reporting up to 100 valid alignments. Reads that were mapped non-uniquely to the genome were counted as 1/n in calculating Deg-Seq coverage, where n is the total number of mapped loci. To identify Deg-Seq peaks (i.e., high coverage sites), we applied a binomial test to each continuous stretch of ≤4 nucleotides with ≥3 reads in each transcript. The expected probability of observing a Deg-Seq peak is 1/l where l is the total number of nucleotides in the transcript of interest with read coverage ≥1. A p-value cutoff was determined as the smaller of the Bonferroni-corrected p-value or 0.05. These significant peaks were considered candidate cleavage sites.

Small RNA-target alignment & parsing: The candidate cleavage sites with their upstream and downstream 25nt were aligned to unique small RNA-Seq reads that passed the length and coverage filters. This alignment was conducted using miRanda [30] requiring a score of at least 60. miRanda was chosen as a convenient local alignment tool that aligns sequences by complementary (as opposed to matching) nucleotides and allows GU wobbles. However, the scoring option for miRNA seed match was not used because we require complementarity beyond

the seed region for candidate sciRNAs. Additionally, the thermodynamic energy calculation was not used in order to minimize the number of assumptions we make and obtain a large initial list that can be later filtered using customized criteria. Nucleotides 9-11 relative to the 5' end of the small RNA were required to match perfectly and overlap the Deg-Seq peak since this is required for cleavage-competent pairing [31]. Gaps and G=U wobble base pairing were allowed, counting G=U base pairing as mismatch 0.5. Unique alignments with at most 4 mismatches were retained for further analyses, which we call "candidate sciRNAs" and their targets.

100x shuffled sciRNAs: Given the large number of small RNAs and Deg-Seq peaks, control analyses were carried out to ensure that the base-pairing relationship was more significant than expected by chance. 100 shuffled controls were generated for each candidate sciRNA, maintaining di-nucleotide frequencies in the sciRNA and masking simple repeats. Simple repeats were defined as tandem repeats of mono-, di-, tri- or quad-nucleotides (number of repeats > 3, 2, 2, 2, respectively). Unique controls were then aligned to the significant Deg-Seq peaks and their flanking regions followed by parsing as described above for the true small RNA-Seq data. Although it is desirable to use a larger number of shuffled controls, we found that the majority (mESCs, 73%; testis, 74%; cerebellum, 70%) of small RNAs had fewer than 100 unique shuffles due to low complexity and the constraints we imposed in shuffling (maintaining di-nt frequencies and simple repeats). Approximately half had less than 90 unique shuffled controls (mESCs, 50%; testis, 52%; cerebellum, 43%). These data suggest that the usage of 100 shuffled controls was a reasonable choice.

Calculate signal-to-noise ratio (SNR): To identify sciRNAs with more targets than expected by chance, a signal-to-noise ratio was calculated using the true and control sciRNA-target alignments. First, an individual SNR (iSNR) was calculated for each candidate sciRNA at

mismatch cutoffs ranging from 0 to 4 at 0.5 intervals. iSNR is defined as the ratio of total targets of the candidate sciRNA to the total targets of all shuffled small RNAs (plus a pseudocount) normalized by the total number of unique shuffled small RNAs (required to be >10). To avoid over-counting targets due to sequence similarity among small RNAs, those small RNAs sharing at least one common 17-mer were grouped together. In other words, for a given group, at least 2 small RNAs share a 17-mer. The results were not very sensitive to this parameter within the range of 15-18. For a range of iSNR cutoffs, a group SNR was calculated for each group of small RNAs as the ratio of total targets of candidate sciRNAs in the group to the total targets of all shuffled small RNAs in the group normalized by the total number of unique shuffled small RNAs. A minimum iSNR cutoff of 10 was chosen, although the resulting sciRNA-target predictions with less than 3 mismatches were insensitive to iSNR cutoffs. Finally, an average SNR was calculated for a given dataset as the average of all group SNRs. The output of this pipeline is the small RNAs that have significantly more targets compared to their controls, which we call “predicted sciRNAs,” and their targets. A signal-to-noise ratio was chosen as an alternative to, for example, an empirical p-value using the 100 shuffled controls as a null distribution. The SNR method affords higher resolution to detect highly confident sciRNA-target pairs, as most empirical p-values were very small.

Total RNA samples

Total RNA samples for whole brain embryo E10, cerebellum embryo E14, cerebellum embryo E18, cerebellum post-natal (PN) 3 weeks, cerebellum PN 6 months, testis embryo E14, testis embryo E18, testis PN 3 weeks and testis PN 6 month were purchased from Zyagen. All RNAs were obtained from the same BALB/C mouse strain. Total RNA of H1 cells was isolated using Trizol (Life Technologies). Additional column purification and DNaseI treatment were applied using Direct-zol RNA kit (Zymo Research).

Construction of small RNA sequencing libraries

Spike-in RNAs (Exiqon) were added into 1μg total RNA before library construction for the normalization control between tissue samples. Small RNA sequencing libraries were generated using NEBNext Small RNA library Prep kit and NEBNext multiplex oligos for Illumina according to the manufacturer's instructions (NEB). The final libraries were purified from 6% PAGE gel, and their concentrations were measured using Qubit fluorometric assay (Life Technologies).

Construction of RNA sequencing libraries

rRNA was depleted using RiboMinus Transcriptome isolation kit (Life Technologies) from 10μg total RNA. ERCC Spike-in RNA (Life Technologies) was added to 500 ng of rRNA depleted RNA. mRNA was isolated using the NEBNext Poly(A) mRNA magnetic isolation module. mRNA sequencing libraries were generated using the NEBNext Ultra Directional RNA library Prep kit and NEBNext multiplex oligos for Illumina according to the manufacturer's instruction (NEB). Final libraries were examined using the Qubit fluorometric assay (Life Technologies) and Bioanalyzer (Agilent) for quality confirmation.

Construction of Degradome sequencing libraries

To generate the degradome sequencing libraries, we used the global 5' RACE library preparation method [22] with some modifications. Briefly, poly(A)+ mRNA was isolated from 400-500 μg of total RNA using Dynabead Oligo(dT) (Life Technologies) according to the manufacturer's instructions. This procedure was repeated to increase the effectiveness of poly(A) selection. The NEBNext Small RNA library Prep kit was used for 5' adapter ligation, followed by reverse transcription using random hexamer primers containing the 3' SR adapter sequence 5'-AGACGTGTGCTCTTCCGATCTNNNNNN. PCR was conducted in 25 cycles at 94°C 15s, 60°C

30s, 70°C 1min. The final libraries were purified from 6% PAGE gel followed by AMPure XP Beads size selection (Beckman Coulter).

2.4. Results

2.4.1. Prediction of sciRNA-mediated RNA cleavage events

We first examined whether the Deg-Seq peaks identified in our study could be artifacts due to PCR amplification bias. Since PCR amplification bias is known to be associated with or reflected in biases in GC content, nucleotide composition and read length [32], we examined these features for reads overlapping Deg-Seq peaks and reads outside of Deg-Seq peaks. We further separated each group of reads into those that have unique sequences (compared to other reads in the same peak) and those that are duplicated. Overall, there appears to be very little PCR amplification bias (Fig. 2.S1).

Next, to evaluate whether it is potentially feasible to identify RNA cleavage events using Deg-Seq and small RNA sequences, we aligned a set of predicted endo-siRNAs [33] to the +/- 25nt flanking sequence of significant Deg-Seq peaks in mESCs [22]. At nucleotides 9-11 from the 5' end of the endo-siRNA, there was an enrichment of Deg-Seq reads in wild type (WT) mESCs (Fig. 2.S2, red) and a depletion of reads in Ago2-/- mESCs (Fig. 2.S2, grey). This result is consistent with the known biochemical properties of Ago2 which cleaves the phosphodiester bond corresponding to bases 10-11 of the small RNA [31,34], suggesting that combined usage of Deg-Seq and small RNA-Seq with appropriate controls may enable identification of functional sciRNAs and their targets.

To achieve the above goal, we analyzed Deg-Seq and small RNA-Seq data as illustrated in Fig. 2.1a (see Section 2.3 Methods). This analysis was carried out for three cell types: mESCs, adult mouse cerebellum post-natal 6 months (6M PN), and adult mouse testis (6M PN) (see Table

2.S1 for all datasets in this study). Cerebellum and testis were chosen in order to compare and contrast sciRNA-mediated RNA cleavage in a tissue containing mature non-dividing cells with a tissue containing frequently dividing germ cells, respectively. mESCs were included because previously published small RNA-Seq and Deg-Seq data were available. Since complementary base pairing along the small RNA is likely required to induce cleavage [18], mismatches were counted in the entire small RNA-target alignment rather than the seed region alone. Across all datasets, the optimal mismatch cutoff corresponding to the highest average SNR was at most 1.5 (Fig. 2.1b, Fig. 2.S3). Thus, we allowed up to 1.5 mismatches for all downstream analyses.

Table 2.1 summarizes the total small RNAs predicted to induce cleavage and the set of Deg-Seq peaks that they target. Notably, many sciRNAs had more than one target, resulting in more than 1,000 total sciRNA-target pairs. Table 2.S2 describes these sciRNAs and targets in detail. The relative scarcity of sciRNAs and targets in cerebellum is unlikely due to low sequencing depth since testis has the lowest number of Deg-Seq peaks and unique small RNA species in the initial sequencing data (Table 2.S1). The vast majority of sciRNAs were identified in only one cell type (Fig. 2.1c), suggesting either a high degree of cell type-specificity or that there are more sciRNAs to be discovered. In addition, about 40% of sciRNAs were known miRNAs, endo-siRNAs, or piRNAs (Fig. 2.1d), with additional sciRNAs being novel small RNA species. Fig. 2.1e shows an example of a novel small RNA inducing Ago2-dependent cleavage of Mtrr. These results suggest that small RNA-mediated target cleavage in mouse may be much more widespread than previously appreciated.

2.4.2. Experimental and genomic validations of sciRNA-target predictions

To provide experimental support, we carried out in vitro cleavage assays for four predicted cleavage events using HeLa S100 extracts and synthetic sciRNAs (Fig. 2.1f, Table 2.S3). These events were picked to represent different types of target genes: a protein-coding gene (Kpna4)

and non-coding genes including a lncRNA (NONMMUG002900), a pseudogene (Zfp389), and an antisense transcript of Traf3ip2. We observed an increasing amount of cleavage products with increasing S100, confirming the validity of the predicted targets.

To further validate our predictions, we applied the pipeline to Deg-Seq of Ago2-/- mESCs [22]. This analysis yielded only 58 sciRNA-target pairs, about 5% of those predicted using WT Deg-Seq (Table 2.1). This result is consistent with the expectation that Ago2 is the main executor of target RNA cleavage and serves as validation of our method. The false discovery rate of our method is at most 5%, which could be an over-estimate since the above 58 sciRNA-target pairs likely include true cleavage events mediated by proteins other than Ago2.

To complement this analysis, we next examined whether sciRNAs were frequently bound by Ago2 in mESCs using Ago2 CLIP-Seq data [35]. Compared to control small RNAs (see Section 2.8 Supplementary Methods), sciRNAs were bound by Ago2 more often in wild type mESCs (Fig. 2.2a). To rule out the possibility that canonical miRNAs were driving the observed sciRNA association with Ago2, we excluded miRNAs from the pool of sciRNAs. The remaining sciRNAs were still enriched in Ago2 CLIP (Fig. 2.2b). Again, these data confirm that sciRNA function is dependent on Ago2.

We next asked whether the cleavage sites in predicted target genes were associated with Ago2. Compared to controls with similar read coverage (see Section 2.8 Supplementary Methods), we observed a highly significant enrichment of Ago2 CLIP-Seq reads for the target sites (Fig. 2.2c). In addition to Ago2-association, we examined whether the Deg-Seq abundance of the target sites was dependent on Ago2 using Deg-Seq of Ago2-/- mESCs [22]. We observed that sciRNA targets sites had significantly reduced Deg-Seq abundance in Ago2-/- mESCs compared to wild type cells (Fig. 2.2d). Together, these results strongly support the validity of the predicted sciRNAs and their cleaved targets.

2.4.3. Small RNAs from diverse classes function as sciRNAs

Since sciRNAs are defined based on a common function (i.e., target cleavage), we hypothesized that a universal pathway may not explain their biogenesis. Rather, sciRNAs may include multiple types of annotated or novel small RNAs. To better understand sciRNA biogenesis, we examined their i) annotation, ii) dependence on the microprocessor in mESCs [33], and iii) long hairpin RNA (hpRNA) structure (see Section 2.8 Supplementary Methods, Fig. 3a).

In mESCs and testis, miRNAs only explained 18% and 9.7% of sciRNAs, respectively, whereas 76.2% of sciRNAs were miRNAs in cerebellum (Fig. 2.3a). In mESCs and cerebellum, many miRNAs had canonical microprocessor dependence (Dicer- and Dgcr8-dependent) based on data derived from mESCs [33]. In contrast, no miRNAs in testis had the canonical microprocessor signature. These may be incorrectly annotated miRNAs, miRNAs generated by non-canonical pathways, or canonical miRNAs in testis but with no microprocessor dependence in mESCs (since microprocessor dependence was evaluated using data from mESCs). The 3 cell types also differed dramatically in the number of predicted endo-siRNAs (Dicer-dependent and Dgcr8-independent), with mESCs having the most endo-siRNAs. Many (64%) of these endo-siRNAs in mESCs had long hairpin structure, consistent with their biogenesis model. Notably, an additional 9.4% and 12.6% of sciRNAs in mESCs and testis, respectively, also had predicted long hairpin structure (Fig. 2.3a), thus are likely endo-siRNAs. Fig. 2.3b illustrates an example of sciRNA-hosting long hairpin structure generated by inverted B1 sequences in mESCs.

Another category of sciRNAs consists of those that appear to be shorter forms of full-length non-coding RNAs from Rfam and piRNABank databases. For example, in testis, a large fraction (27.2%) of sciRNAs overlapped piRNA sequences, consistent with the high abundance of piRNAs in this tissue. piRNAs appeared to be trimmed from the 5' end, 3' end, or both, to generate sciRNAs (Fig. 2.S4), indicating existence of additional processing mechanisms. Similarly, tRNAs, snRNAs and rRNAs were also identified as possible sciRNA-generating RNAs,

all of which were reported previously to produce small RNAs [36,37]. The last category of sciRNAs aligned to annotated genes that are not miRNA/endo-siRNA/Rfam/piRNA genes ("other genes" in Fig 2.3a). Their biogenesis mechanisms remain unknown.

2.4.4. sciRNA expression varies during testis and cerebellum development

Since sciRNA populations in mESCs, adult testis and adult cerebellum were largely distinct, we examined the divergence process of sciRNA profiles from mESCs to the adult cells during development. We obtained small RNA sequencing data to examine sciRNA expression in several developmental stages of testis and cerebellum. We then compared expression profiles of sciRNAs between mESCs and testis (Fig. 2.3c), or between mESCs and cerebellum (Fig. 2.3d). Specifically, sciRNAs identified in mESCs or the adult tissue (testis 6M PN or cerebellum 6M PN) were labeled as mESC-specific (if predicted in mESC data only), adult tissue-specific (if predicted in adult tissue only), or common to both. Interestingly, we observed reciprocal changes in the relative enrichment of expressed mESC-specific and adult tissue-specific sciRNAs during the development of both testis and cerebellum. Thus, mESC-specific sciRNAs were gradually replaced by tissue- and adult-specific sciRNAs as the cells mature.

Notably, cerebellum and testis demonstrated different patterns of sciRNA expression during development. A considerable portion (25-30%) of sciRNAs in cerebellum was also present in mESCs, which was a general observation for all developmental stages ("Both," Fig. 2.3d). In contrast, sciRNAs common to both mESCs and testis were rare in all developmental stages ("Both," Fig. 2.3c). In testis stages embryonic day 18 (E18) and later, the majority of sciRNAs were testis-specific. On the other hand, there were few testis-specific sciRNAs at E14. This time point approximately precedes the development and proliferation of prospermatogonia [38]. Thus, it is possible that sciRNAs in testis are primarily generated during spermatogenesis and largely distinct from those in mESCs or other tissues (e.g., brain).

In striking contrast to sciRNAs, miRNA profiles (excluding sciRNAs) were much more stable across all developmental stages included in this study (Fig. 2.S5). A much larger fraction of miRNAs was common to mESCs and different stages of testis or cerebellum. Additionally, the difference between testis and cerebellum was not as pronounced as that observed for sciRNAs. The considerable distinction in the developmental- and tissue-specific profiles of sciRNAs and non-sciRNA miRNAs indicates that these two classes of small RNAs may have distinct cellular functions.

2.4.5. sciRNAs target non-coding regions of genes spanning diverse functional categories

Target genes in the three cell types demonstrated little overlap, with only 40 genes in common between any two samples (Fig. 2.4a). This apparent tissue specificity is mainly due to the tissue-specific expression of sciRNAs. The number of sciRNAs expressed in a particular tissue but not predicted to induce cleavage was a small minority (Table 2.S4).

Strikingly, the majority of cleavage sites within coding genes was located in 3' UTRs in all cell types, much more than expected by chance (Fig. 2.4b). Since our search of cleavage sites was across the entire mRNAs, this 3' UTR enrichment strongly testifies to the validity of our results. It should be noted that miRNAs are primarily known to target 3' UTRs [39], which could arguably be partially due to the intense focus on 3' UTRs in prediction algorithms and the usage of evolutionary conservation as a requirement of target sites. Thus, our study supports 3' UTR targeting by small RNAs in an unbiased manner.

Besides the non-coding 3' UTRs, many of the sciRNA targets are non-coding transcripts, derived from lncRNAs, pseudogenes or other non-coding RNAs in GENCODE and NONCODE annotations (Fig. 2.4b,c). In all cell types, lncRNAs account for the majority of non-coding targets. A relatively large fraction of targets in testis and cerebellum was regulated by miRNAs (Fig. 2.S6),

whereas novel small RNAs derived from other genes account for the majority of targeting in mESCs.

Among the predicted sciRNA target genes, many are associated with important functional relevance. Fig 2.4d shows a subset of such genes grouped into transcription factors [40], ubiquitin related genes, splicing related genes, and cancer-testis antigens [41]. Importantly, most of these target genes demonstrated negative correlation in gene expression levels (measured by RNA-Seq of testis samples at different developmental stages) relative to their corresponding sciRNA expression (Section 2.8 Supplementary Methods), further confirming the predicted functional relationship of sciRNAs and targets.

We also carried out pathway, ontology, and Ingenuity network analyses for protein-coding and non-coding target genes to obtain a comprehensive view of functional relevance (Table 2.S5). Overall, sciRNA targets are involved in a diverse spectrum of functional categories, enriched with developmental-related processes and basic cellular function (cellular assembly and organization, cell morphology, and cell cycle).

2.4.6. sciRNAs and target genes are enriched with repetitive elements

Although the biogenesis pathways of sciRNAs appear diverse, a unifying feature of the sciRNAs and their targets is their substantial overlap with repetitive elements. The majority of sciRNAs in mESCs and testis are repetitive, with most aligned to SINE elements, especially the B1 subclass (comparable to human Alus) (Fig. 2.5a). Repetitive sciRNAs often target more RNAs than non-repetitive sciRNAs (Fig. 2.S7a). Furthermore, we observed that B1-derived sciRNAs mapped to specific sub-regions of the consensus B1 sequence (Repbase [42]) in both sense and antisense orientations (Fig. 2.S7b). Thus, many sciRNAs may be derived from pairs of inverted B1 repeats, as shown for Ccdc30 (Fig. 2.3b). Since the above observation applies to both mESCs and testis,

sciRNA biogenesis likely shares similar pathways and genomic features in the two cell types despite the involvement of different sciRNA species.

Similar to sciRNAs, the majority of target cleavage sites were in B1 elements (Fig. 2.5b), and their +/- 5nt sequences mapped to similar regions of the consensus B1 sequence as sciRNAs (Fig. 2.S7b vs. S7c). Because the majority of sciRNA cleavage sites are located in SINEs, we next tested whether this is a unique feature of sciRNA-directed degradation or common in the global Degradome. In contrast to the significant enrichment of SINEs in sciRNA-targeted cleavage sites, the remaining cleavage sites in the rest of the Degradome were rarely in SINE regions (Fig. 2.5c). The fraction in repetitive regions only slightly increased when all types of repeats were considered, suggesting that SINE elements are driving this phenomenon (Fig. 2.S7d).

To ensure that the relative enrichment of SINEs in target sites was not artificially inflated as a result of non-unique mapping of the Deg-Seq reads, we examined the sequence uniqueness of the flanking regions of predicted cleavage sites. The majority of target sequences were unique among all predicted targets of a specific cell type regardless of the length of flanking regions, although targets in testis had the smallest level of uniqueness (Fig. 2.S7e). We then reexamined the overlap between target cleavage sites and repetitive elements after removing redundant target sequences. B1 elements were still enriched, confirming that SINE elements are enriched in the target pool (Fig. 2.S7f). Thus, SINE-targeting, especially B1-targeting, is a unique feature of sciRNA-mediated cleavage in mouse.

2.4.7. Repetitive elements as signals for sciRNA targeting

It was previously speculated that SINEs are used as signals for miRNA targeting [43–45]. However, other studies presented evidence against this postulation, showing that canonical miRNA targeting avoided Alu elements [46]. Here, we suggest that B1 elements in mice serve as signals for small RNA targeting through endonucleolytic cleavage instead of the canonical miRNA

pathway. If this speculation holds, then Ago2 should bind to sciRNA targets in B1 regions more often than to predicted canonical miRNA targets in B1 regions. To test this hypothesis, for miRNAs expressed in mESCs, we focused on their predicted canonical targets (as defined in microrna.org [47]) where the target sites are located in B1 elements. These targets were separated into two groups: those with target sites overlapping our predicted sciRNA target, and those that neither overlapped a sciRNA target nor contained a Deg-Seq peak (Section 2.8 Supplementary Methods). It should be noted that only 2% (68,005 / 3,316,252) of the predicted canonical miRNA targets were in B1 regions. We observed that Ago2 binds to the first group more often than the second (Fig. 2.S7g). The above results support our hypothesis that B1 elements are likely signals for sciRNA-mediated cleavage.

Next, we asked whether sciRNA-mediated targeting of B1 elements is under evolutionary selection. Since repetitive regions are poorly conserved across species, a conventional multi-species sequence conservation analysis was not feasible. Instead, we conducted an analysis of SNP enrichment in sciRNA target sites using known mouse SNPs (Section 2.8 Supplementary Methods). Strikingly, we observed that SNPs were significantly depleted in sciRNA-targeted B1 sequences compared to the flanking B1 regions (Fig. 2.5d), suggesting that sciRNA targets are under selection for sequence conservation. This finding also indicates that sciRNA-mediated regulation has potential functional significance.

2.4.8. Small RNA-guided endonucleolytic cleavage in human ESCs also targets retrotransposons

To investigate sciRNA-guided cleavage in human cells, we obtained small RNA-Seq and Deg-Seq data from human H1 ESCs (Table 2.S1) and conducted the same analysis as for the mouse datasets. A total of 34 sciRNAs and 23 target genes were identified (allowing up to two mismatches in the alignment), with about 50% sciRNAs being annotated miRNAs (Fig. 2.S8a).

The lower numbers of sciRNAs and targets compared to mESCs could be explained by lower depth of small RNA sequencing in human (Table 2.S1). Alternatively, differences in the repetitive sequences and their distribution in human and mouse genomes may also account for this difference. Nevertheless, these results allow an examination of the global properties of human sciRNAs and targets. Similar to their mouse counterparts, they were enriched with sequences overlapping retrotransposons (Fig. 2.S8b,c). However, in addition to SINE (Alu) elements, LINE (L2) elements were considerably enriched among human sciRNAs and their target sites. Similar to mouse sciRNA targets, human target sites were often located in non-coding genes or 3' UTRs of coding genes (Fig. 2.S8d). Furthermore, functional analysis of human targets revealed similar categories as for mouse targets (Fig. 2.S8e, Table 2.S5).

Despite the above high similarities in general properties of sciRNA targeting between human and mouse, the specific types of retrotransposons enriched in the human data are different from those in mouse. This is likely explained by the apparent difference in abundance, sequence composition, and activity of retrotransposons across the two species [48,49]. Thus, the sciRNA pathway is a conserved mechanism between human and mouse but leverages different sciRNA molecules, possibly to adapt to the divergence of retrotransposons between the two genomes.

2.5. Discussion

We report a global analysis of endonucleolytic RNA cleavage events in mouse ESCs, testis, and cerebellum. In mammals, mRNA cleavage was not previously considered a major pathway for small RNA-guided mRNA degradation, with a small number of genes predicted as targets of this mechanism [18–22]. Our analysis revealed an expanded repertoire of hundreds of sciRNAs and their corresponding target genes in mouse and human, suggesting that this regulatory pathway is conserved and relatively prevalent in a cell-type specific manner. Given the potential functional significance of the target genes in development and essential cellular processes, sciRNA-

mediated cleavage may have a much more profound impact on gene regulation and cellular function than previously appreciated.

We defined sciRNAs based on a unifying function, that is, those that are predicted to cleave target RNAs via near perfect sequence complementarity. Thus, it is not surprising to find sciRNAs potentially reflecting diverse biogenesis mechanisms and overlapping known small RNAs of different categories (miRNAs, siRNAs, piRNAs). Despite this diversity, sciRNA expression appears to be under close regulation, as manifested by their striking expression specificity to developmental stages and cell types in contrast to all miRNAs (Fig. 2.3c,d, Fig. 2.S5). In addition to known categories of small RNAs, many sciRNAs were novel, with unknown biogenesis mechanisms and derived from genomic regions of known genes. These data suggest that the biogenesis and regulated expression of sciRNAs need further investigation.

Despite their heterogeneity in biogenesis, a salient feature of sciRNAs and their target regions is the enrichment of repetitive sequences, especially of the B1 class in mouse. Retrotransposons are very prevalent in mammals, accounting for more than 40% of the human and mouse genomes. However, little is known regarding the functional implication of their presence within genes. It was speculated that miRNAs or other small RNAs may target SINE elements embedded in mRNAs, and therefore the SINEs are used as signals for gene targeting [43–45]. Yet, supporting data for this speculation was lacking. Studies that imposed the canonical miRNA targeting rules (requiring seed matching) predicted that Alu elements avoid targeting by miRNAs, thus providing data against the above speculation [46]. Our results reconcile the seemingly conflicting hypotheses and data by supporting that B1 elements within murine RNA transcripts serve as signals for small RNA targeting, but through the endonucleolytic cleavage pathway instead of the canonical miRNA targeting based on seed matches alone. As retrotransposable elements spread across the genome and into non-coding regions of genes, sciRNA-mediated

regulatory mechanisms may have evolved to leverage the abundant repetitive elements as signals for gene targeting, although this hypothesis remains to be tested.

Capturing such targeting events may have been difficult due to the repetitive nature of the small RNAs and their target sites. Non-uniquely mapped reads in sequencing data analysis are often excluded because they are difficult to interpret. In this study, non-unique alignments of Deg-Seq reads were retained, with their abundance normalized by total number of non-unique matches to the genome (Section 2.3 Methods). Nevertheless, we observed that the majority of cleavage site-flanking sequences were unique, suggesting that enrichment of repetitive targets was not overestimated (Fig. 2.S7e,f). The recognition of retrotransposons in RNA transcripts allows for targeting of multiple repeat-containing transcripts by a single sciRNA (Fig. 2.S7a). However, it should be noted that the number of targets of a typical sciRNA is much smaller than that of canonical miRNAs. sciRNAs are still highly specific to their respective targets given their extended sequence complementarity and the high degree of divergence and uniqueness among retrotransposable elements [50].

It should be noted that our method imposed stringent criteria in predicting sciRNA-target relationships. In using the SNR approach, we assumed that sciRNAs should have more targets than expected by chance. Due to the requirement of extended sequence complementarity, many true sciRNAs may only target a small number of genes. As a result, a true sciRNA may not have a high SNR. Thus, it is possible that many more sciRNAs exist than presented in our study.

In summary, we report the discovery of a large number of sciRNAs and their cognate targets in mouse and human cells. This mode of gene regulation was previously poorly characterized in mammals. We demonstrate that this pathway mainly targets retrotransposons in mammalian genomes, and likely plays essential roles in gene regulation in a developmental stage- and cell type-specific manner.

2.6. Figures

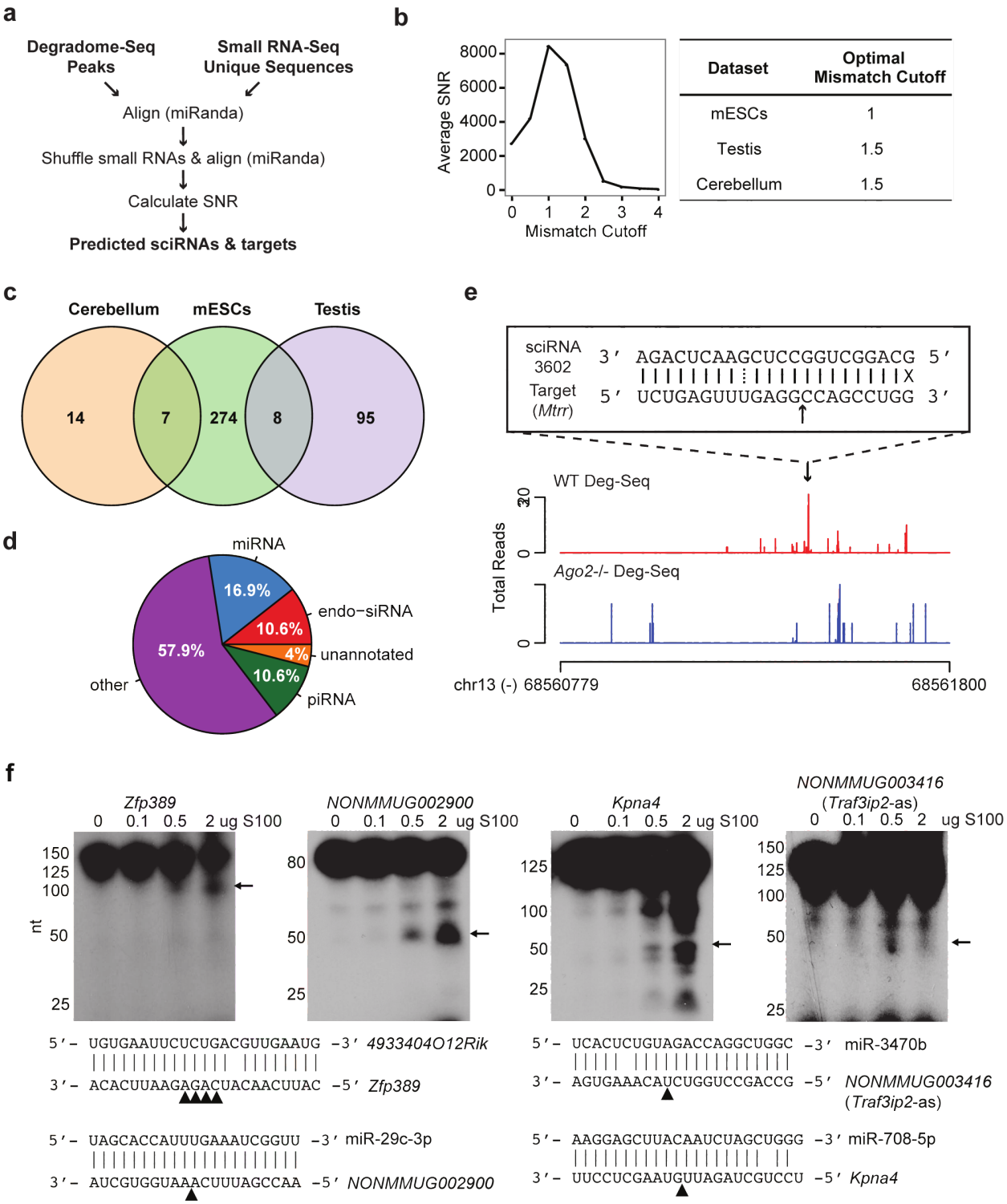


Figure 2.1 | Prediction of sciRNA-mediated mRNA cleavage events.

(**a**) Bioinformatic pipeline schematic. (**b**) Left panel: average SNR for each mismatch cutoff in mESCs; right panel: optimal mismatch cutoff corresponding to the maximum average SNR for each dataset. (**c**) Venn diagram of sciRNAs identified in mESCs, testis 6M PN, and cerebellum 6M PN. (**d**) Pie chart of sciRNA annotations (total = 398, combining sciRNAs from 3 cell types). (**e**) An example of predicted sciRNA-mediated cleavage. Read distributions are shown for the 3'UTR of the Mtrr gene in Deg-Seq data of wild type (WT, red) and Ago2 knockout (blue) mESCs. Alignment of the sciRNA to the Deg-Seq peak is shown in a box, where a solid line indicates a base pair match, dotted line indicates a G=U wobble, X indicates a mismatch, and black arrow indicates location of the Deg-Seq peak. (**f**) Experimental validation of target RNA cleavage mediated by small RNAs. 200ng of Kpna4, NONMMUG002900, Zfp389, or NONMMUG003416 (Trafip2-as) RNA were incubated with different amount of HeLa S100 loaded with 50 nM synthetic sciRNA at 37°C for 30 min. Arrows indicate cleaved 5' RNA fragments (whose sizes are consistent with predicted cleavage products). Small RNA/target RNA sequences are shown with arrowheads indicating the predicted cleavage sites (4 sites identified as Deg-Seq peaks in Zfp389).

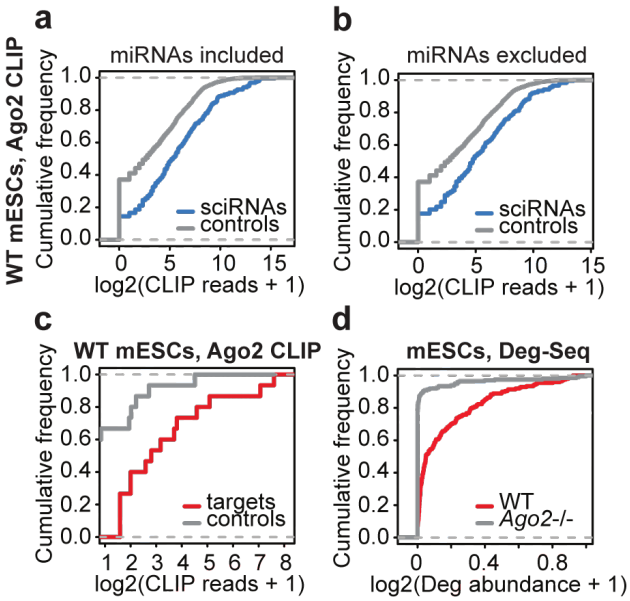


Figure 2.2 | Genomic data supporting the validity of sciRNA-target predictions.
(**a**) Empirical cumulative frequency of abundance of Ago2 CLIP-Seq reads containing sciRNA sequences (blue) or control sequences randomly picked from Dicer-independent Dgcr8-independent small RNAs (grey) in WT mESCs (P = 2.2e-16, two-sided Kolmogorov-Smirnov (KS) test, same below). (**b**) Similar to (**a**), excluding miRNAs (P = 2.1e-10). (**c**) Similar to (**a**), comparing abundance of CLIP reads covering Deg-Seq peaks in target genes (red) or controls (grey, see Section 2.8 Supplementary Methods) (P = 0.003). (**d**) Deg-Seq peak abundance in WT and Ago2 knockout mESCs (P < 2.2e-16).

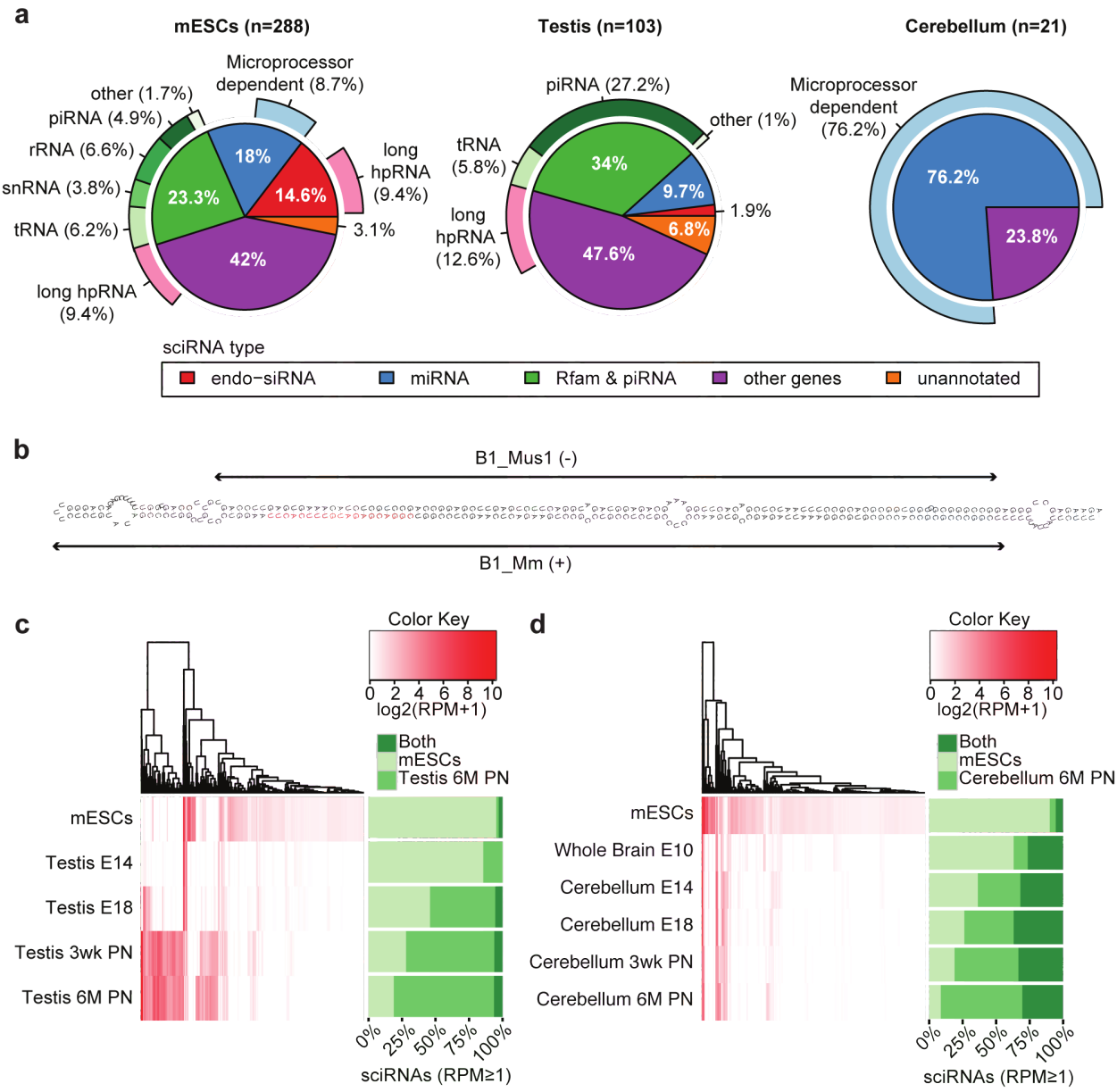


Figure 2.3 | Characterization of sciRNAs.
(**a**) Categorization of sciRNAs in mESCs, testis and cerebellum according to: (i) annotation (inner circle), (ii) dependence on the microprocessor (outer circle, light blue) and (iii) long hpRNA structure (outer circle, pink). In mESCs, 1 unmapped sciRNA was excluded. (**b**) RNAfold structure of sciRNA 24792 (red) and flanking regions (within the Ccdc30 gene). This sciRNA was predicted in mESCs. Two inverted B1 repeats (Repeatmasker) are labeled. (**c**) Hierarchical clustering of sciRNA expression levels (reads per million, RPM) in mESCs and different stages of testis development. E14: embryonic day 14; E18: embryonomic day 18; 3wk PN: 3-week postnatal; 6M PN: 6-month postnatal. In the heatmap, RPM values of all sciRNAs that were identified originally in mESCs or 6M PN testis were visualized for each sample. Stacked bars on the right show the percentage of sciRNAs (among those with RPM ≥ 1) specific to mESCs (defined as those that were only identified in mESCs by the pipeline in Fig. 2.1A, but not in the testis 6M PN data), testis 6M PN (similarly as defined above) or common to both. Note that some sciRNAs predicted

originally in mESCs or testis may be excluded in the stacked bars due to low RPM. (**d**) Similar to (**c**), for cerebellum development. E10: embryonic day 10.

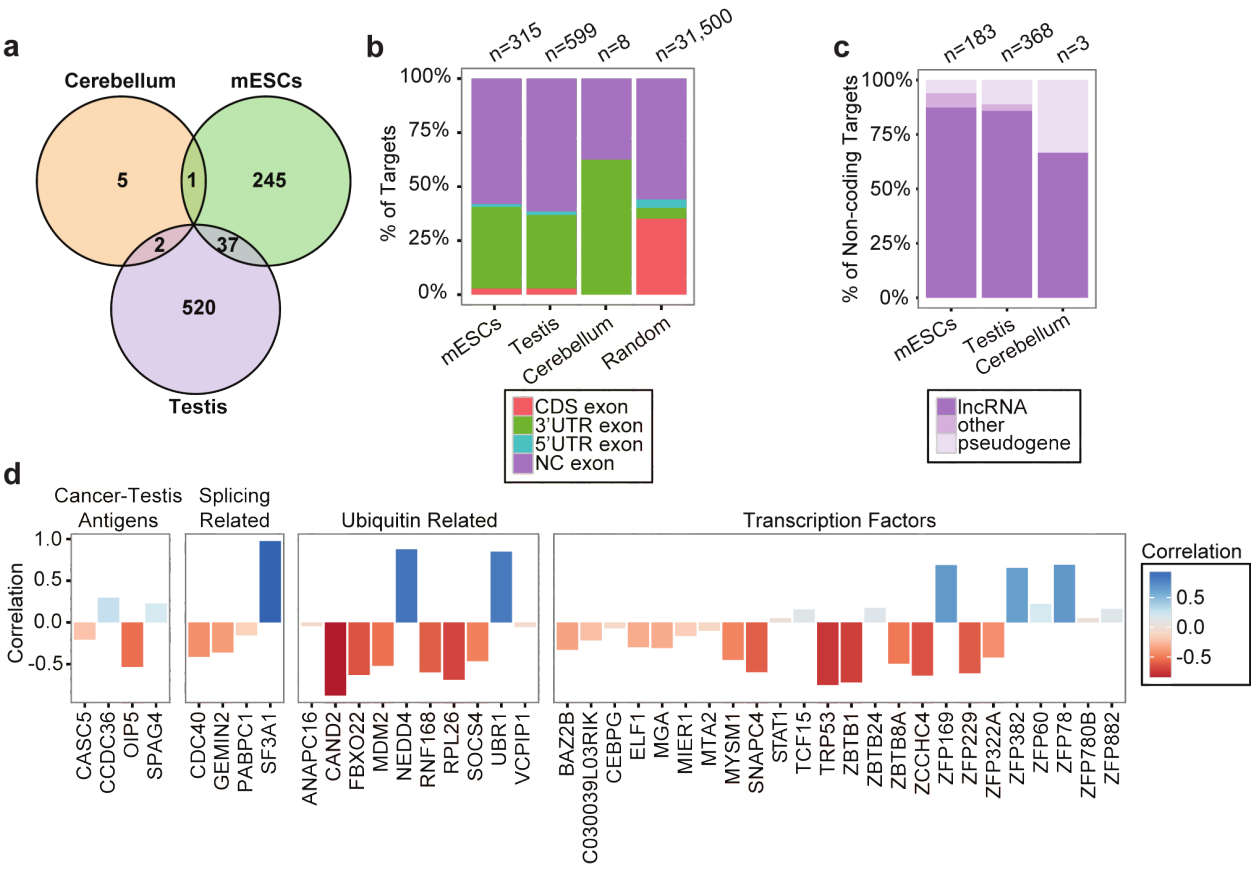


Figure 2.4 | Characterization of sciRNA targets.
(**a**) Venn diagram of target genes predicted in mESCs, testis and cerebellum. (**b**) Distribution of target cleavage sites (Deg-Seq peaks) in different types of regions of the transcriptome. CDS: coding sequence; NC exon: exon of non-coding transcript. Random: random positions from random transcripts (see Section 2.8 Supplementary Methods). (**c**) Types of non-coding transcripts among sciRNA targets, prioritized as pseudogene > lncRNA > other. (**d**) Pearson correlation of target mRNA expression and sciRNA expression for four example categories of target genes (see Section 2.8 Supplementary Methods).

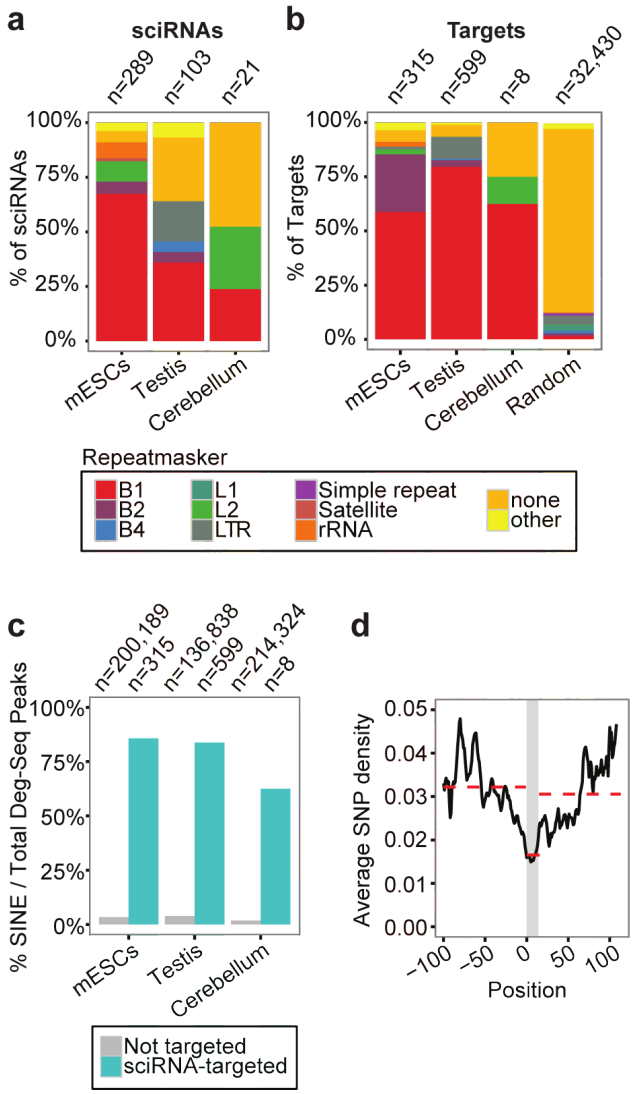


Figure 2.5 | Small RNA guided endonucleolytic cleavage targets retrotransposons.
(**a**) Distribution of sciRNAs in different types of repeats (Repeatmasker). If more than one Repeatmasker annotation was identified, the following prioritization was used: B1 > B2 > B4 > others. Random: similar to Fig. 2.4B. (**b**) Similar to (**a**), for target cleavage sites. (**c**) Percentage of Deg-Seq peaks overlapping SINE regions among all Deg-Seq peaks. Two groups of Deg-Seq peaks are shown: those targeted by sciRNAs (sciRNA-targeted) or otherwise (Not targeted). (**d**) Average SNP density per position in all B1 sequences bound by sciRNAs and their ±100 nt flanking region (chi-square P-value P < 2.2e-16). The y-axis is the average SNP density per nucleotide (see Section 2.8 Supplementary Methods). A smoothing window of 10nt was applied to all data points. The grey region indicates the sciRNA-binding region. It ranges from 0–14 because the maximum length of targeted region was 24 and the smoothing window spanned 10 nt. Red dashed lines indicate the average SNP density of the three corresponding regions.

2.7. Tables

Table 2.1 | Summary of the final sets of predicted sciRNAs, targeted cleavage sites, and their combinations.

	Mismatches allowed	Predicted sciRNAs	Predicted sciRNA cleavage sites	Total sciRNA-target pairs
Cerebellum	1.5	21	8	30
Testis	1.5	103	599	1,772
mESC wild-type	1.5	289	315	1,108
mESC Ago2-/-	1.5	53	23	58

2.8. Supplementary Methods

Bioinformatic analysis of Ago2 dependence

Ago2 CLIP-Seq in wild type mESCs (3 samples) were used (accession GSE25310: SRR072951, SRR072952, SRR072953, SRR072954, SRR072955) [35]. To evaluate the presence of sciRNAs in Ago2 CLIP-Seq reads, the total number of CLIP reads containing each sciRNA was summed across the 3 samples. As a control, the presence of Dicer-independent, Dgcr8-independent small RNAs [33] in Ago2 CLIP-Seq reads was calculated similarly.

To calculate the enrichment of Deg-Seq peak regions in Ago2 CLIP-Seq data, the total number of CLIP reads at each nucleotide of the Deg-Seq peak (up to 4 nucleotides) was summed across each of the 3 wild type samples and normalized by the length of the Deg-Seq peak. For each peak with non-zero CLIP coverage, 100 random peaks with the same abundance were chosen by Fisher-Yates shuffle, and normalized CLIP coverage was calculated similarly as described above. Then, the average CLIP-Seq coverage of the 100 random peaks was calculated.

To calculate Deg-Seq abundance of sciRNA targets in wild type mESCs, the total number of Deg-Seq reads at the peak was normalized by the total number of reads in the gene. For each target gene, there is often no Deg-Seq peak present in Ago2-/- data, as expected. Thus, to calculate Deg-Seq abundance of targeted genes in Ago2-/- Deg-Seq (accession GSE21975 [22]), the maximum number of Deg-Seq reads at any position in the gene was normalized by the total Deg-Seq reads in the gene.

RNA-Seq analysis

The following public mESC RNA-Seq datasets were used: SRR921480, SRR921481, SRR921482, SRR921483 (GSE48252) [51]. For both public and in-house RNA-Seq data, sequencing reads were aligned to mm10 using tophat2 [52] with parameters -a 9 -g 1 for all

datasets. RPKM was calculated using in-house scripts for genes annotated in the databases described in Materials and Methods.

Pearson correlation analysis: The Pearson correlation was calculated using RNA-Seq and small RNA-Seq data of the following samples: mESCs, testis E14, testis E18, testis 3 weeks post-natal, testis 6 months post-natal. For mESCs, the average RPKM of the above four datasets from GSE48252 was used in the Pearson correlation analysis of target and sciRNA expression. If more than one sciRNA targeted a gene, the minimum Pearson correlation was kept.

sciRNA characterization and expression

sciRNAs were aligned to genome mm10 or hg19 using Bowtie 0.12.7 [29] with parameters -v 1 -a --best --strata. Because many sciRNAs may align to thousands of genomic locations, a stepwise prioritization process was used to annotate them (listed in descending priority): [1] the sciRNA was found to be Dicer-dependent and Dgcr8-independent [33], [2] the sciRNA overlaps a miRBase [53] miRNA or Rfam [54] pre-miRNA +/- 3nt, [3] the sciRNA overlaps a non-miRNA Rfam annotation +/- 3nt, [4] the sciRNA is a sub-sequence of a piRNABank [55] annotated piRNA, [5] the sciRNA overlaps a region from our custom merged annotation described earlier, [6] no annotation, [7] unmapped. If a category had 3 or fewer members, it was labeled as “other” in Fig. 2.3a.

We next predicted whether a sciRNA was derived from a long hairpin RNA (hpRNA) structure. For each sciRNA, we applied RNAfold [56] to the region flanking its genomic alignment (+/-500nt). If the sciRNA was aligned to multiple genomic locations, the region (+/-500nt) with the highest read coverage was used. Then, RNAfold’s dot bracket notation was used to examine whether the sciRNA aligned to the stem of a long secondary structure, i.e. “long hpRNA.” Namely, two criteria were used: stem length ≥70, and ≥70% of the sciRNA nucleotides (length 19-24) were structured (brackets). Stem length was calculated as the distance to the next opposite-facing

bracket (equal to zero if the sciRNA contained two opposite facing brackets). These thresholds were chosen by checking that previously identified endo-siRNAs (e.g. miR-1195 and miR-1965) in mESCs were included and manually checking the RNAfold structure prediction for some novel examples (e.g. Ccdc30 shown in Fig 2.3b). In addition, several thresholds were tested, and although stem length ≥70 and ≥70% structured were the optimal thresholds, the results were largely robust to choice of stem length and percent nucleotides structured.

Small RNA expression was calculated using reads per million (RPM), where sequencing depth for each sample only included small RNAs that passed the preprocessing steps in the pipeline (i.e. masked small repeats and length [19,24]). sciRNAs identified in mESCs and/or testis 6M PN were clustered across all testis developmental stages, and similarly for mESCs and/or cerebellum 6M PN. sciRNAs with RPM ≥ 1 were grouped based on whether they targeted a transcript in mESCs, adult tissue, or both.

miRNAs with minimum read count 20 in at least one sample were clustered since this minimum read count was used in the preprocessing pipeline to identify sciRNAs. miRNAs with RPM ≥ 1 were labeled as adult tissue, mESCs, or both based on read count ≥ 20 in these samples.

R function heatmap.2 was used for hierarchical clustering with Euclidean distance and complete linkage.

Target characterization

Cleavage sites (Deg-Seq peaks) were characterized by genic location and by type of transcript. If the Deg-Seq peak overlapped multiple transcripts, the genic regions were prioritized as follows: coding exon (CDS exon) > 5'UTR exon > 3'UTR exon > exon in non-coding transcripts (NC exon) > intron in coding genes > 5'UTR intron > 3'UTR intron > intron in non-coding transcripts. Transcripts harboring the cleavage sites were also examined for their types: coding, pseudogene, lncRNA, or other non-coding RNAs. Finally, Repeatmasker was used to decide whether a

cleavage site overlaps a repetitive sequence. To test if the observed gene region and Repeatmasker enrichment at cleavage sites significantly differed from the transcriptome overall, 100 random positions per Deg-Seq peak were chosen from any annotated transcript. If a random position overlapped multiple transcripts, the genic regions were prioritized using the above prioritization schemes (Fig. 2.4b, 2.5b).

Functional analysis of target genes

Gene set enrichment: Ingenuity pathway analysis was conducted for sciRNA targets in the mouse and human datasets (IPA®, QIAGEN Redwood City, www.qiagen.com/ingenuity) using the default parameters. Gene Ontology (GO) analysis was carried out with our previous approach [57] for non-coding genes according to the functional annotations provided by NONCODEv4 [28].

SNP enrichment: The union of all sciRNA binding sites located within B1 regions (i.e. the Alu Repeatmasker family) in predicted target genes of mESC, testis, and cerebellum were used (n = 1,510). Each position in the targeted region and 100 nt flanking region were interrogated for SNPs (dbSNP 138 [58]). To calculate a SNP density per nucleotide, the sum of SNPs at each position was normalized by the total sequences interrogated at that position. These values were smoothed using a sliding window of size 10nt and step size 1nt (Fig. 2.5b). The smoothed values were anchored on the rightmost nucleotide (e.g. the smoothed SNP density at position -100 is the average SNP density of the window -100 to -90). A chi-square p-value was calculated using a contingency table of total SNPs vs. the sum of length of B1 annotation (up to 100 nt) within vs. outside the target region.

Ago2 binding in canonical miRNA targets: Predicted canonical miRNA target sites were downloaded from microrna.org [47]. Targets of expressed miRNAs in mESCs (miRNA read count ≥

20) located within B1 (Alu family) regions were separated into two groups based on whether or not they overlapped a sciRNA target sequence. A total of 102 unique miRNA targets overlapped a sciRNA target, and 54,766 did not. Among the latter, 53,333 (97%) did not contain a Deg-Seq peak, as expected. We refer to these 53,333 targets as “putative canonical miRNA B1 targets” and the 102 unique miRNA targets overlapping a sciRNA target as “putative sciRNA targets”. To compare Ago2 CLIP-Seq overlap between the two groups, one “putative canonical miRNA B1 target” was randomly chosen for each “putative sciRNA target” (i.e. 102 vs. 102), and this process was repeated 1,000 times. Ago2 CLIP-Seq overlap was calculated as the sum of reads in the target site, divided by the length of target sequence. For each of the 1,001 total target and control sets, targets with zero Ago2 CLIP-Seq overlap were excluded and the average CLIP-Seq overlap of the remaining targets was calculated. An empirical p-value was calculated by counting the total sets of random “putative canonical miRNA B1 targets” with higher average Ago2 CLIP-Seq density than the “putative sciRNA targets.”

In Vitro Transcription

sciRNA targets were amplified using OneTaq DNA polymerase (NEB) from mouse genomic DNA followed by TOPO TA cloning (Life Technologies). The target-specific primers are listed in Table 2.S3. The TOPO cloning products were then transformed into DH5α competent cells and were later plated for overnight incubation at 37°C. PCR and Sanger sequencing were used to verify the constructs.

Target RNAs were in vitro transcribed using 1 µg of template DNA and the HiScribe™ T7 High Yield RNA Synthesis Kit (NEB). To remove template DNA, 20U RNase-free DNase I (Roche Diagnostics) was applied for 15 min at 37°C followed by phenol extraction. In vitro transcribed RNA was purified from 10% PAGE gel. RNA was dephosphorylated by 10U calf intestinal alkaline phosphatase (NEB) at 37°C for 60 min and then purified by phenol-chloroform extraction. Two µg

dephosphorylated RNA was labeled with γ-^{32}P ATP 150Ci (MPbio) by T4 Polynucleotide Kinase (NEB) at 37°C for 60 min followed by 12% PAGE gel isolation.

In Vitro Cleavage Assays

In vitro cleavage assays were performed as described previously [21]. HeLa cytoplasmic S100 extract was obtained from Speed BioSystems. For endogenous sciRNA cleavage (miR-708-5p and miR-29c-3p), HeLa cytoplasmic S100 extract (0, 0.1, 0.5, and 2 μg respectively) was incubated with 200 ng of ^{32}P-labeled target RNA at 37°C for 30 min in the cleavage buffer (20mM HEPES KOH pH7.9, 100mM KCl, 1.5mM $MgCl_2$, 0.5mM DTT, 0.5mM PMSF, 1mM ATP, 0.2mM GTP). The cleavage reaction was terminated by adding 2X RNA gel loading buffer and incubated at 60°C for 5 min. Cleaved RNA was loaded onto 10-12% PAGE gel and exposed to X-ray film at –80°C.

For cleavage of Traf3ip2-as and Zfp389 target genes, endogenous sciRNA levels were relatively low. Thus, sciRNA +/- strands were annealed to form sciRNA duplexes as follows: [1] prepare the +/- strand sciRNAs (Table 2.S3) at a final concentration of 100 μM; [2] mix 2 μL of the two sciRNA strands with 5μL 10X Annealing Buffer (100 mM Tris-HCl, pH 7.5, 1 M NaCl, 10 mM EDTA); [3] add nuclease-free H_2O to reach a total volume of 50 μL; [4] heat at 94°C in water bath for 4 min, 70°C for 10 min, and then allow cooling to room temperature; [5] annealed sciRNA duplex was further purified by 10% PAGE gel and precipitated with 2.5 volumes of absolute ethanol. HeLa cytoplasmic S100 extract was then preincubated with 50 nM purified sciRNA duplex at 37°C for 30 min before adding the ^{32}P-labeled target RNA. The cleavage reaction was otherwise carried out in the same way as described above.

2.9. Supplementary Figures

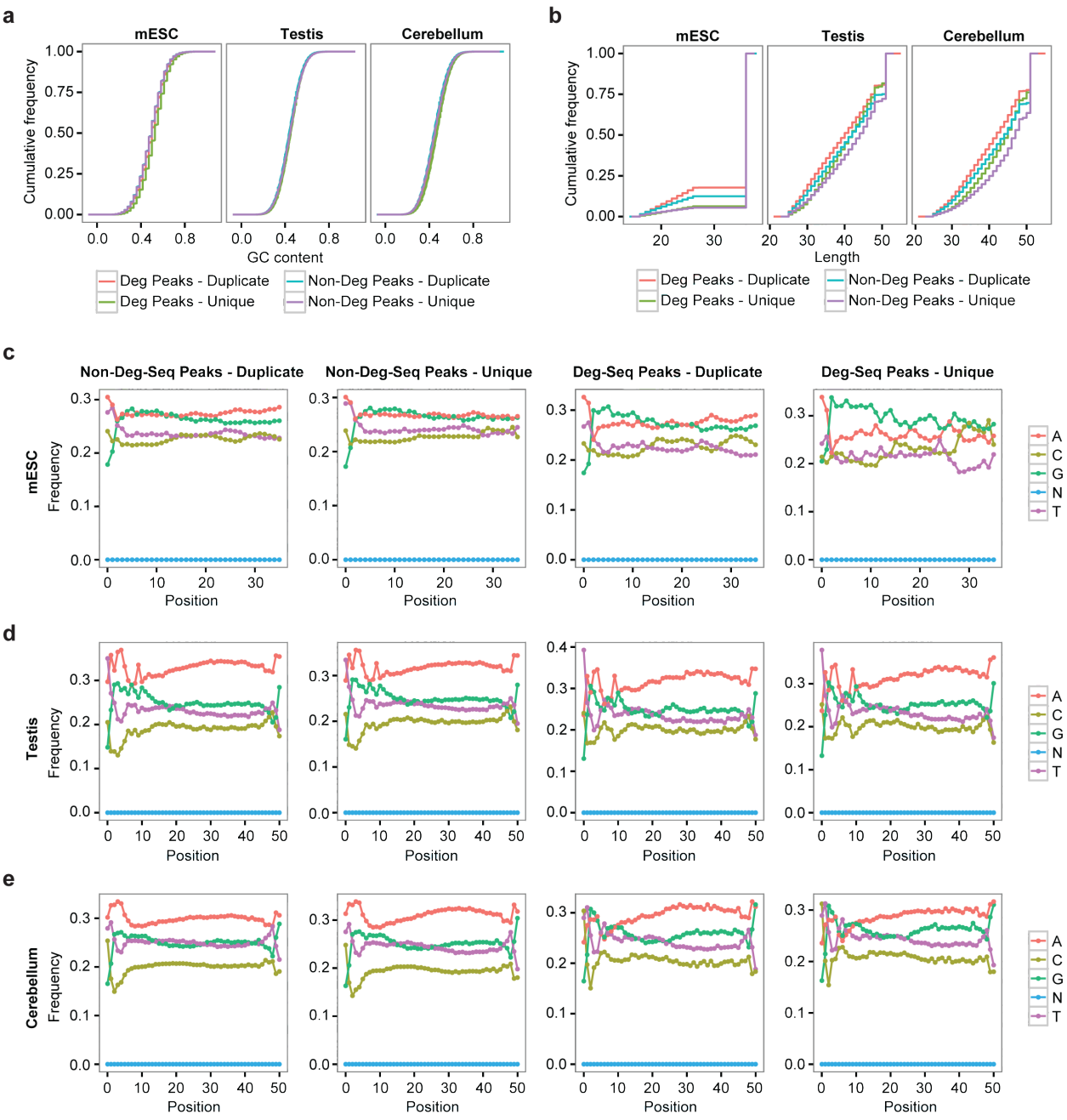


Figure 2.S1 | PCR amplification bias detection.
All reads in each Deg-Seq library were separated into four groups: unique reads within Deg-Seq peaks, duplicate reads within Deg-Seq peaks, unique reads outside of Deg-Seq peaks, and duplicate reads outside of Deg-Seq peaks. Three criteria were used for comparison: (**a**) GC content, (**b**) read length after trimming adapters, and (**c-e**) nucleotide composition.

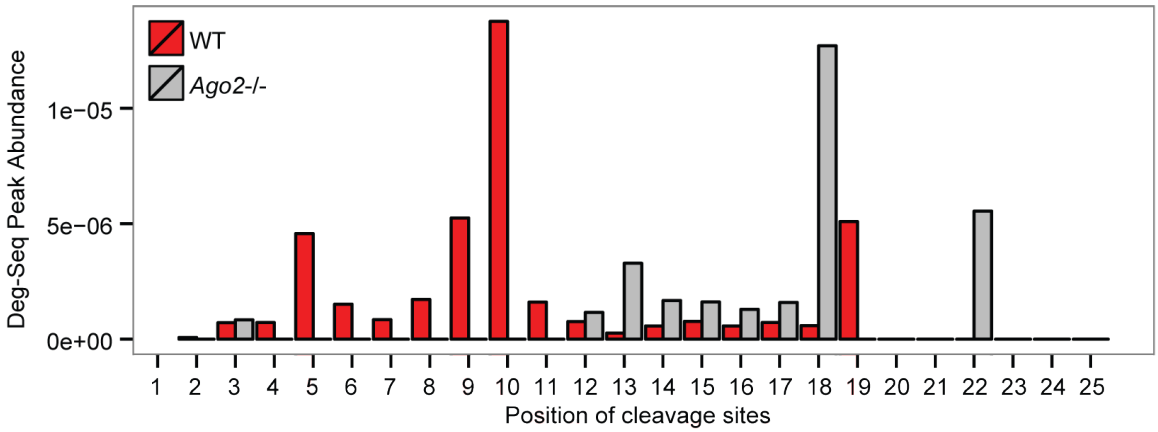


Figure 2.S2 | Alignment of predicted endo-siRNAs to Degradome-Seq supports existence of small RNA-guided cleavage.
Deg-Seq peak abundance (total reads in the Deg-Seq peak / total mapped reads) per nucleotide along predicted predicted endo-siRNAs [33] from 5' to 3' (left to right on the x-axis) in wild type (WT, red) and Ago2 knockout mESCs (Ago2-/-, grey). Results are shown for Deg-Seq peaks that aligned to endo-siRNAs with up to 1 mismatch. The enrichment of reads at nt 9-11 in WT is eliminated in Ago2-/-. Moreover, the Deg-Seq abundance is within the level of background noise for most positions in AGO2-/- with the exception of nt 18 and 22. These could be due to unknown artifacts or mechanisms.

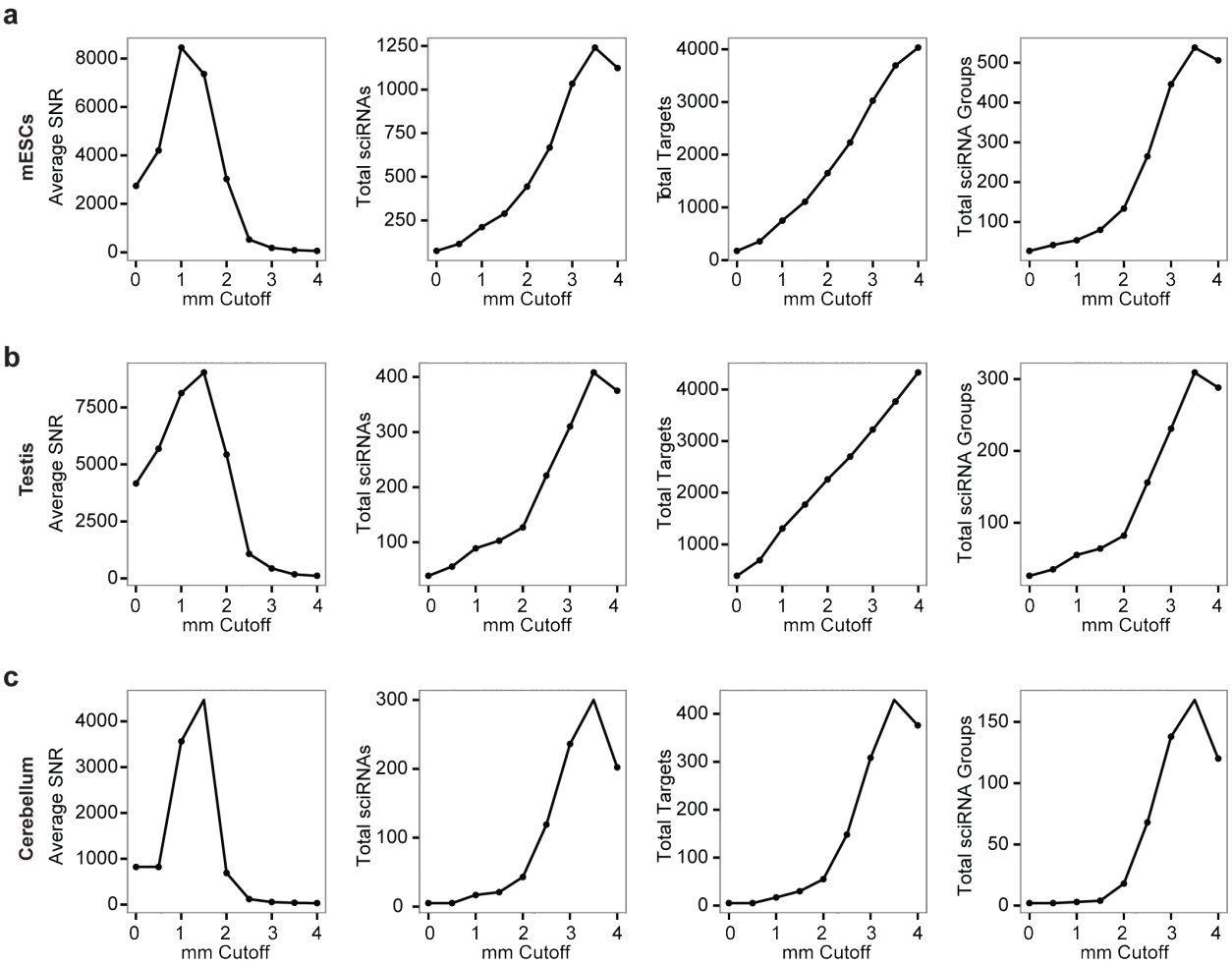


Figure 2.S3 | sciRNA-target prediction at varying mismatch cutoffs.
The average SNR (Methods), total sciRNAs, total targets, and total sciRNA groups for each mismatch (mm) cutoff varying from 0 to 4 in 0.5 intervals in (**a**) mESCs, (**b**) testis, and (**c**) cerebellum.

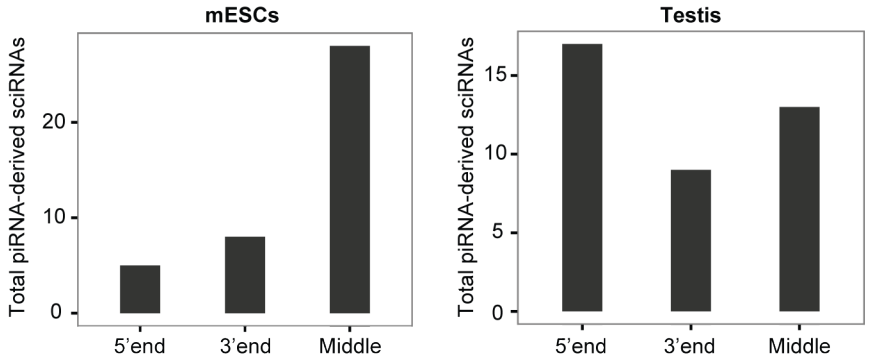


Figure 2.S4 | Characterization of piRNA-derived sciRNAs.
piRNA-derived sciRNAs grouped by the sciRNA's relative location in the piRNA in mESCs and testis. 5' end: sciRNA starts at the first nt of the piRNA; 3' end: sciRNA ends at the last nt of the piRNA; middle: sciRNA starts and ends at internal nt of the piRNA.

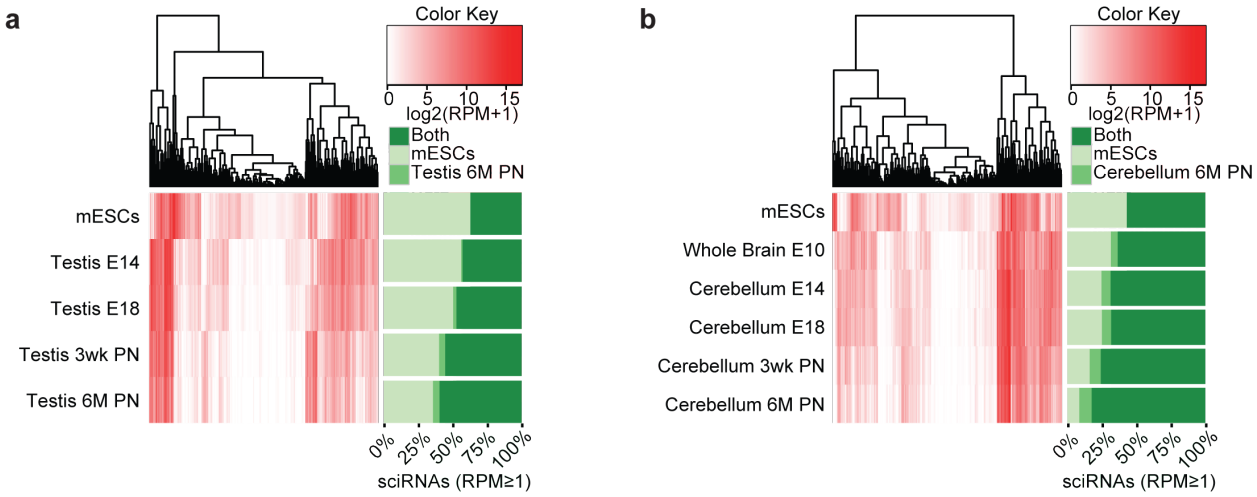


Figure 2.S5 | miRNA expression in testis and cerebellum development.
Similar to Fig. 2.3c and 2.3d, but instead clustering all non-sciRNA miRNAs and labeling them based on having read count ≥ 20 in adult tissue (Testis 6M PN or Cerebellum 6M PN), mESCs, or both.

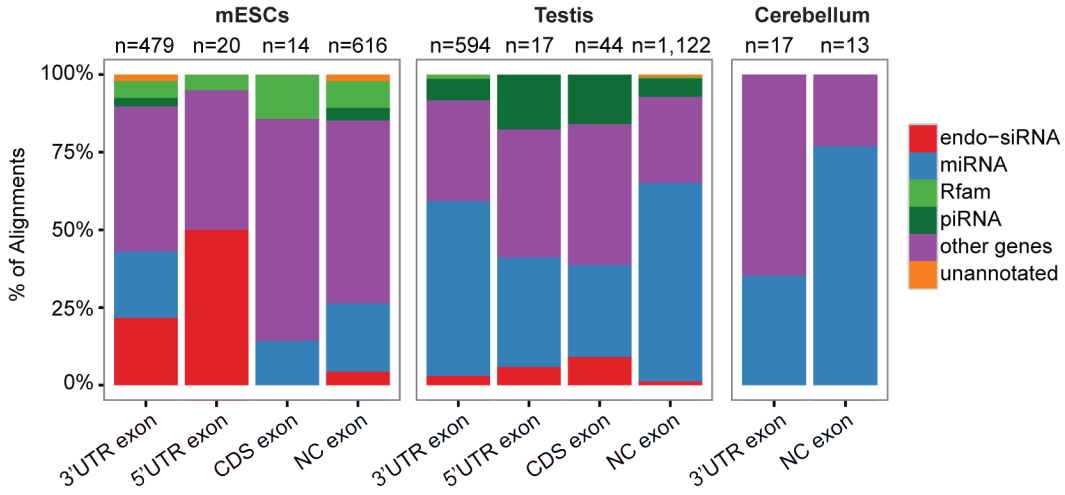


Figure 2.S6 | Characterization of sciRNA targets.
The type of sciRNAs targeting each type of gene region is shown for each sciRNA-target alignment. CDS: coding sequence; NC exon: exon in non-coding transcript. The total number of sciRNA-target alignments is shown above each bar.

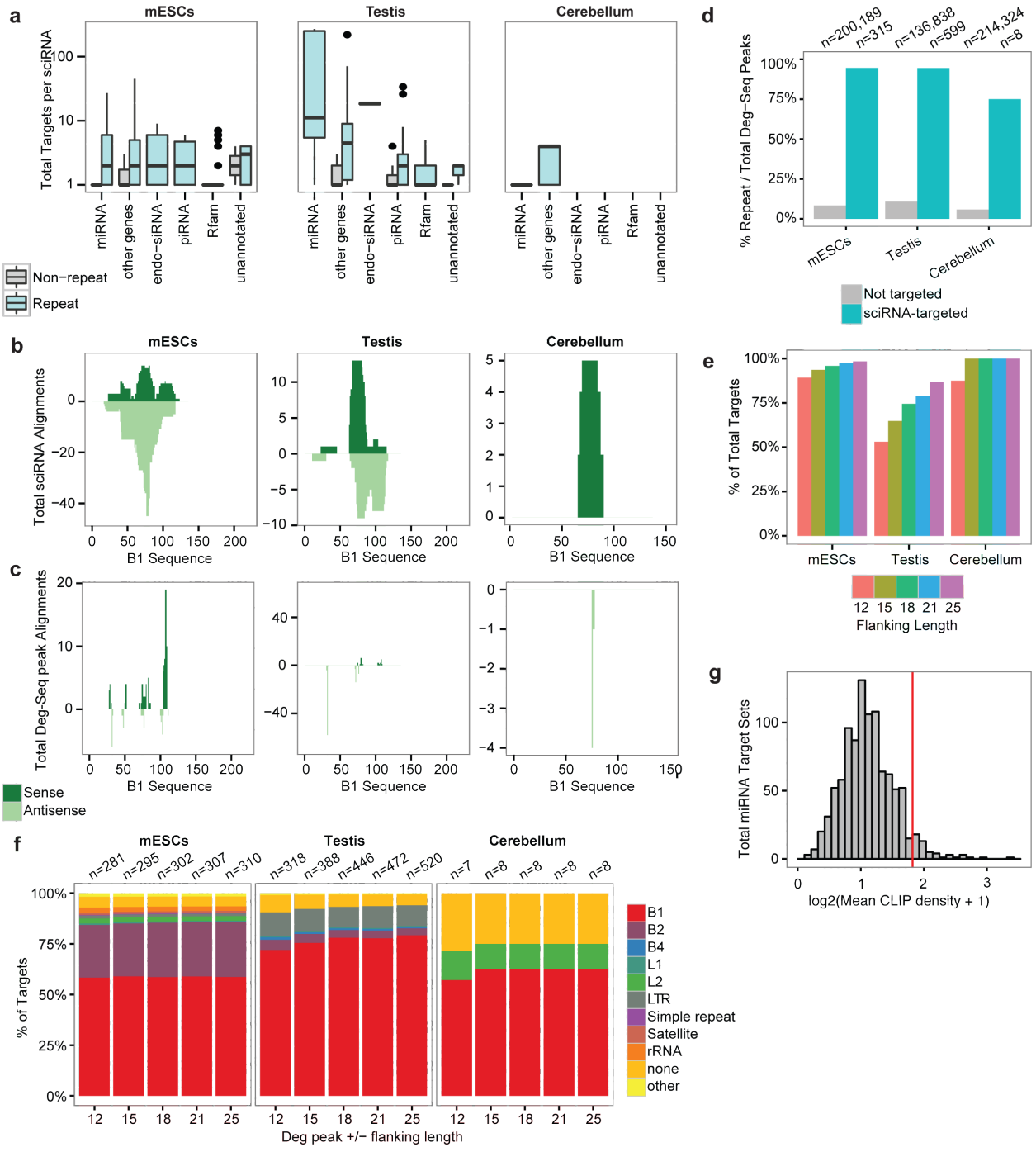


Figure 2.S7 | Repetitive nature of sciRNAs and targets.
(**a**) Total number of targets per sciRNA separated by the subtype and repetitive vs. non-repetitive nature (Repeatmasker) of sciRNAs. Alignment of sciRNA (**b**) or Deg-Seq peak +/-5nt (**c**) to the B1 consensus sequence using BLASTN. Dark (light) green and positive (negative) values indicate sense (antisense) alignment. (**d**) Percentage of Deg-Seq peaks overlapping repetitive regions (any type in Repeatmasker) among all Deg-Seq peaks. Two groups of Deg-Seq peaks are shown: those targeted by sciRNAs (sciRNA-targeted) or otherwise (Not targeted). (**e**) Percent of total targeted cleavage sites with unique flanking sequence. Results shown for varying +/-12 to 25 nt. (**f**) Distribution of Repeatmasker annotations for target cleavage sites with unique flanking

sequence for varying flanking sequence length (x-axis). For targets with the same flanking sequence, the following prioritization was used: B1 > B2 > B4 > others. (**g**) Histogram of 1,001 sets of 102 microrna.org miRNA targets based on their average Ago2 CLIP-Seq density (empirical P = 0.049). 1,000 controls in grey; sciRNA target-overlapping miRNA targets in red.

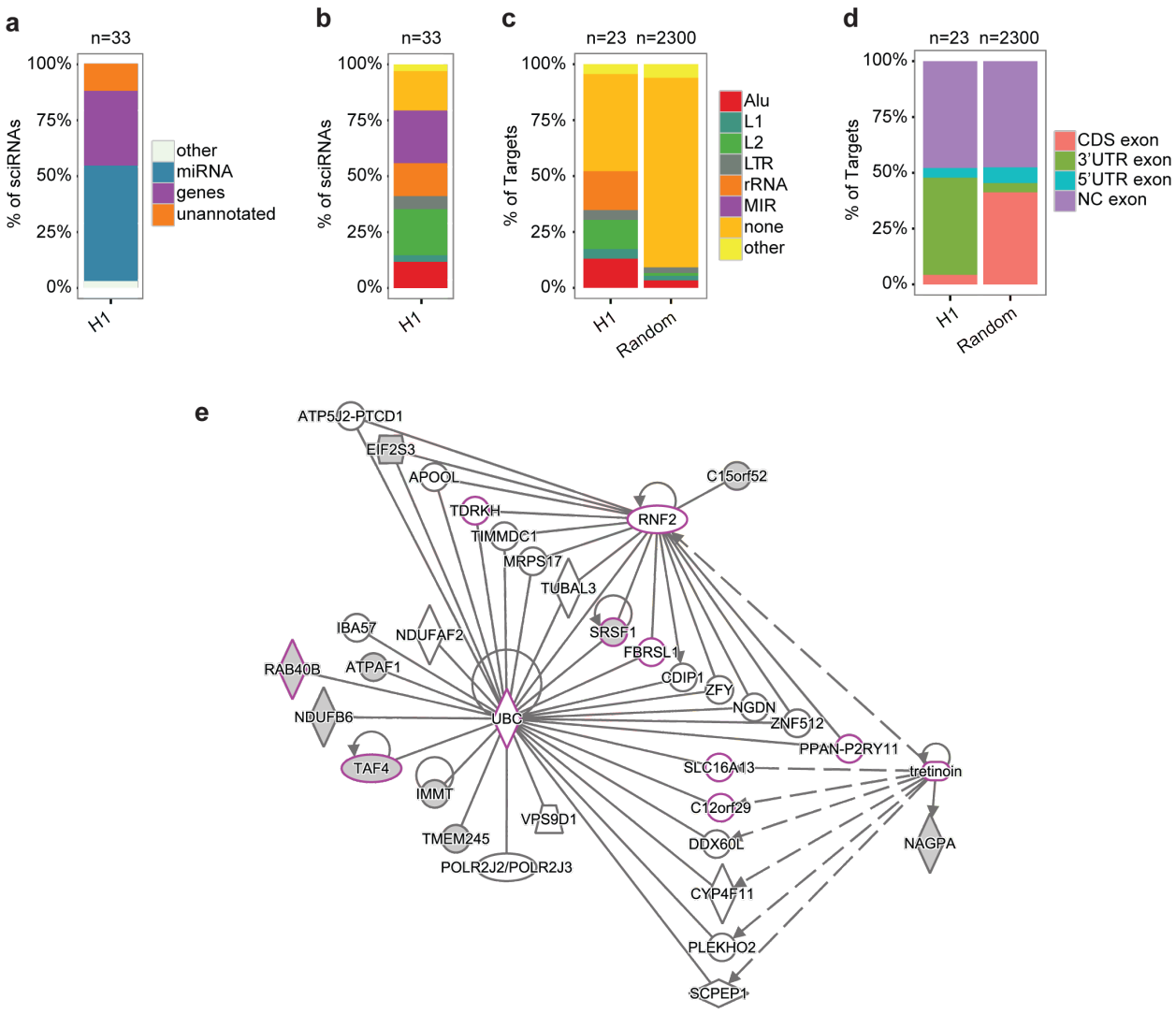


Figure 2.S8 | Small RNA guided endonucleolytic cleavage in H1 ESCs.
(**a**) sciRNA annotation. One unmapped sciRNA was excluded. (**b**) Repeatmasker family annotation of sciRNAs (**c**) Repeatmasker family annotation of target cleavage sites. Random: random positions from any transcript were chosen as a control cleavage site (see Section 2.3 Methods). (**d**) Distribution of target cleavage sites (Deg-Seq peaks) in different regions of the transcriptome. CDS: coding sequence; NC exon: exon in non-coding transcript. Random: similarly defined as in (**c**). (**e**) One network was identified by Ingenuity Pathway Analysis of sciRNA targets. Grey-shaded nodes: sciRNA targets; magenta-outlined nodes: sciRNA targets associated with the top three diseases/functions shown below the network.

2.10. Supplementary Tables

Table 2.S1 | Datasets used in this study.

Mouse embryo (E10) whole brain	H1	Mouse adult (PN 6 months)	Mouse adult (PN 6 months)	Mouse ESC WT				
								Karginov et al., 2010 (GSE21975)
	146,290,212	81,487,812	107,172,831					124,939,742
	129,112	137,437	214,332					200,504
-	-	-	-					Toedling et al., 2012 (GSE35368)
11,962,085	12,187,485	12,519,963	13,213,728					72,165,395
	454,507	536,942	274,022					1,273,306
	7,146,855	2,748,030	12,026,992					55,796,555
		-		Hou et al,. 2013 (GSE48243) SRR921483	Hou et al,. 2013 (GSE48243) SRR921482	Hou et al,. 2013 (GSE48243) SRR921481	Hou et al,. 2013 (GSE48243) SRR921480	
		50,246,960		16,162,090	18,990,358	24,068,096	10,951,783	
								Leung et al., 2011 (GSE25310)
								24,505,777

		Mouse ESC AGO2 KO	Mouse adult (PN 3 weeks) testis	Mouse adult (PN 3 weeks) cerebellum	Mouse embryo (E18) testis	Mouse embryo (E18)	Mouse embryo (E14) testis	Mouse embryo (E14)
Degradome-Seq	**Reference**	Karginov et al., 2010 (GSE21975)						
	Total Reads	29,889,865						
	Total Deg-Seq Peaks	70,900						
sRNA-Seq	**Reference**		-	-	-	-	-	-
	Total Reads		16,407,546	13,004,933	15,761,194	15,159,178	13,346,437	10,313,260
	Total small RNAs with repeats							
	Total copies of small RNAs with							
mRNA-Seq	**Reference**		-		-		-	
	Total Reads		42,883,607		48,175,122		56,866,472	
AGO2 CLIP-Seq	**Reference**							
	Total Reads							

Table 2.S2 | List of sciRNAs and target genes identified in mouse ESCs, adult testis and cerebellum.

Please refer to the published article for Table 2.S2.

Table 2.S3 | Primers used in this study.

PCR primers for in vitro Cleavage Assay	
Primer name	**Sequence (5'to 3')**
KPNA4-T7-Fw	GAAATAATACGACTCACTATAGGGTCTATAGTATCTGTTCACTCATTG
KPNA4-Rv	GAACTCATGGTACCTTCACG
NONMMUG002900-T7-Fw	GCGTAATACGACTCACTATAGGGCACAAGGTTCTGAGGCTGGTG
NONMMUG002900-Rv	GTTCAGAGTCTGTTTTTGTCTAGC
TRAF3IP2-as-T7-Fw	GCGTAATACGACTCACTATAGGGACCTTTAATCCCAGCACTCGG
TRAF3IP2-as-Rv	TCATGCAGTTGAGGTTTTGTGA
Zfp389-T7-Fw	GCGTAATACGACTCACTATAGGGAGAAGACTATAATAAAGGCAGGCC
Zfp389-Rv	TCATGCACAGTGAGGGTTGA
Annealing primers for sciRNAs	
Primer name	**Sequence (5'to 3')**
miR-3470b-	UCACUCUGUAGACCAGGCUGGCUU
miR-3470b+	GCCAGCCUGGUCUACAGAGUGAUU
4933404O12RIK-	UGUGAAUUCUCUGACGUUGAAUGUU
4933404O12RIK+	CAUUCAACGUCAGAGAAUUCACAUU

Table 2.S4 | sciRNA expression vs. targeting.

cerebellum	expressed	not expressed
targeting	21	0
not targeting	2	375
mESCs	expressed	not expressed
targeting	289	0
not targeting	9	100
testis	expressed	not expressed
targeting	103	0
not targeting	3	292

Table 2.S5 | Functional analysis of target genes of sciRNAs.

Please refer to the published article for Table 2.S5.

Chapter 3: De novo identification of alternative transcription start and polyadenylation sites in RNA-Seq

3.1. Abstract

Alternative transcription start sites (ATSSs) and alternative polyadenylation sites (APAs) modulate transcriptional and post-transcriptional gene regulation, often in a tissue-specific manner. Several methods are available for identifying APA from RNA-Seq through change point detection, but most are limited by requiring a 3'UTR annotation and/or only identifying one APA site among other limitations. Additionally, most change point detection methods ignore the 5' end. Here, we developed mountainClimber, a de novo approach for the simultaneous identification of ATSS and APA that overcomes several limitations of existing approaches and outperforms a similar method. In the subsequent chapters, we apply this approach to poly(A)-selected RNA from a variety of human tissues (Chapter 4) and both chromatin-associated RNA and poly(A)-selected RNA in murine macrophages (Chapter 5). Upon publication, the software will become publicly available.

3.2. Introduction

Alternative polyadenylation and alternative promoter usage are well-appreciated mechanisms of gene regulation. More than 70% of mammalian genes utilize APA sites, which can affect mRNA stability, mRNA translation, mRNA nuclear export, mRNA localization, and protein localization through addition or removal of RNA binding protein sites, miRNA target sites, and other regulatory motifs in the 3' untranslated region (UTR) (reviewed in [9,10]). Additionally, poly(A) site usage can influence transcription, as it occurs co-transcriptionally [59]. The choice of polyadenylation (poly(A)) site has been associated with polyadenylation factors, nucleosome density, splicing activity

(especially U1 snRNP), and other factors including TSS (reviewed in [9]). ATSS occurs in 40-50% of mammalian genes, which can affect translation through addition or removal of upstream open reading frames, secondary structure motifs, and RNA binding protein recognition sites in the 5'UTR [60–62]. Thus, identification and characterization of APA and ATSS in different biological conditions will enhance our understanding of gene regulation.

Recent advances in sequencing technologies improved detection of ATSS and APA. The FANTOM consortium, among others, identified TSSs with CAGE-Seq [63] while others developed approaches to sequence poly(A) sites including PolyA-Seq [64], 3'READS [65], and 3'-Seq [66]. These assays revealed ATSS and APA are more widespread than previously appreciated. In particular, APA is not only tandem (i.e. in the last exon), but also frequently in exons and introns upstream of the 3'UTR [65]. Still, RNA-Seq is significantly more widely used than these approaches, motivating the development of methods for ATSS and APA identification from RNA-Seq. While several methods exist for identifying APA from RNA-Seq, they have one or more of the following limitations: reliance on gene annotation, searching for tandem APA within last exons only, identifying no more than one change point (i.e. two poly(A) sites), requiring two biological conditions, or handling only one sample [67–73].

To identify both ATSS and APA while overcoming the aforementioned limitations, we developed mountainClimber, a novel de novo approach for change point identification. In contrast to existing methods, our approach is de novo, scans the entire transcription unit (TU), can identify any number of change points, works on single samples, and tests for differential ATSS and APA across biological conditions. Thus, mountainClimber identifies not only APA within 3'UTRs, but also ATSS in 5'UTRs, intronic APA, and coding sequence APA. We demonstrate it outperforms an existing approach, IsoSCM[70].

3.3. Methods

de novo change point identification

There are three major steps of the mountainClimber approach: defining de novo TUs in each sample with mountainClimberTU, calling change points in TUs of each sample with mountainClimberCP, and detecting differential ATSS and APA usage with mountainClimberTest. The pipeline is written in Python 2.7.2 and R 3.4.3, and relies on python modules pybedtools[74,75], scipy, numpy, peakutils, bisect, itertools, sklearn[76], and pysam, and R packages ggplot2 [77], reshape2 [78], and dplyr. Module and package versions used were: pybedtools 0.6.2, scipy 0.15.1, numpy 1.10.4, peakutils 1.0.3, sklearn 0.18.1, pysam 0.9.0, ggplot 2.2.1, reshape2 1.4.3, and dplyr 0.7.4. Each step can optionally be used in isolation.

mountainClimberTU(c, n, p, w): Define TUs de novo based on RNA sequencing. 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, ends are extended or trimmed until there are no zero-coverage bases. If a bed or bedgraph file of split reads is input (see get_junction_counts below), then introns spanned by at least c exon-exon junction read counts (default = 2) will be included in order to join exons from the same transcript in the same TU.

mountainClimberCP(a, d, w, t, l, e, f, s, u, n, z): Given one or two bedgraph files for non-strand-specific and strand-specific RNA-Seq libraries respectively, a bed file of split reads (see get_junction_counts below), and the TUs from mountainClimberTU, change points are called in each TU with length ≥ l (default = 1000) and average reads/bp in exons ≥ e (default = 10). The Cumulative Read Sum (CRS) is defined as:

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

$$\mathrm{CRS}_i = \frac{1}{\max_i \mathrm{CRS}} \sum_{j=1}^{i} v_j$$

Where x_i is the total reads at position i in a TU. The Kolmogorov-Smirnov (KS) test is used to test whether the CRS significantly deviates from the uniform distribution (i.e. the diagonal line in Fig. 3.1b "1. Calculate CRS"). If the KS test p-value < t (default = 0.001), then proceed with calling change points. To identify the elbows in this distribution, i.e. candidate change points, the distance between the CRS and the diagonal line is calculated and denoised using a median filter with window size w. Then, the 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. 3.1b "2. Identify local extrema"). This motivates the weighted CRS (wCRS), which effectively amplifies this signal at these positions:

$$w_i = \frac{v_i^2}{\max_i v}$$

$$\mathrm{wCVS}_i = \frac{1}{\max_i \mathrm{wCVS}} \sum_{j=1}^{i} |w_i - w_{i-1}|$$

On the other hand, change points in last exons are more gradual due to the end-effect of the reads. Thus, we call change points in two steps: (1) wCRS to call junction change points, and (2) CRS in UTRs to call UTR change points. Finally, the change points are called again in 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 bp, x, 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). If non-strand-specific RNA-Seq is used, then the strand is inferred by choosing the strand with the maximum number of supporting GT-AG splice site signals at exon-intron junctions. 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, TandemTSS, 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 if RNA-Seq was strand-specific or strand was able to be inferred from non-strand-specific RNA-Seq; DistalLeft, DistalRight, TandemLeft, TandemRight, Exon, and Intron if strand could not be inferred.

mountainClimberTest: Cluster change points across replicates and conditions and test for differential change point usage. This is comprised of three major steps: (1) mountainClimberTest_cluster, (2) mountainClimberTest_ru, (3) mountainClimberTest_diff. While any number of conditions can be clustered in the first step, the last step tests for differential end usage between only two conditions.

mountainClimberTest_cluster(n, e, f, d, m): Cluster change points first across replicates, then across conditions using sklearn.cluster DBSCAN and report the median position of each cluster.

DBSCAN was chosen because it is not parameterized by the total number of clusters, but rather by the minimum number of points n and neighborhood size e. If e = -1.0 (default), then use the optimal window size, w, from mountainClimberCP. Users can also filter change points using minimum fold change f and average reads/bp in exons m, though these are not considered by default. If clustering across multiple conditions, then require at least d conditions to be clustered together (default = 1) and use the minimum n across all conditions. For example, let's say we have five replicates in two conditions. If d = 2 and n = 4, and a DBSCAN cluster contained 4 replicates from condition A but only 2 replicates from condition B, then this cluster is ignored because it was not reproducible in n replicates in condition B. However, if the DBSCAN cluster contained 4 replicates from condition A and none from condition B, then this cluster is retained because it represents a condition-specific change point. Change point labels are prioritized as follows: Junction > DistalTSS > DistalPolyA > TandemTSS > TandemAPA > DistalLeft > DistalRight > Exon > Intron > TandemRight > TandemLeft.

mountainClimberTest_readCounts: Calculate the average reads/bp for each segment after clustering.

mountainClimberTest_ru: Given the output segments and change points from mountainClimberTest_cluster and bed file of average read counts per segment from mountainClimberTest_readCounts, define transcript ends and calculate their relative usage (RU) for each condition. Genes with no clustered change points and/or less than 4 segments are excluded. End segments are defined as (1) all distal ends regardless of their change point label, (2) all DistalLeft, DistalRight, DistalTSS, TandemTSS, DistalPolyA, TandemAPA change points, (3) all change points before the first and after the last Junction change points, and (4) any condition-specific non-Junction change points between the first and last Junction change points.

Ambiguous segments that could not be assigned to either 5' or 3' end are ignored. If a proximal segment has lower coverage than the distal segment (e.g. in the case of intron retention) then the RU is set to 0 for the proximal (e.g. intronic) segment. Otherwise,

$$\mathrm{RU_k} = \frac{1}{\mathrm{n}} \sum_{\mathrm{s}=0}^{\mathrm{n}} \frac{\mu_{\mathrm{k,s}}}{\mu_{\mathrm{k-1,s}}}$$

Where $\mu_{k,s}$ is the average reads/bp in each segment k in proximal to distal order in sample s, and n is the total number of samples.

mountainClimberTest_diff (d, p, t, m): Test for differential usage of 5' and 3' ends in two conditions. Only consider TUs for which change points were called in both conditions (otherwise, the gene was either not expressed or was too noisy to cluster change points). TUs are ignored if: the distal segment is the same in both conditions and RU = 1 (no difference between conditions), mean distal segment coverage < d (default = 5) in all conditions, proximal coverage is < p in at least one condition, or coverage is increasing from proximal to distal. If two segments are consecutive in at least one condition, then that change point is tested for differential usage; otherwise, the RU is reported without testing. See Table 3.S1 for all scenarios.

Given two conditions, a change point is considered differential if there is a significant difference in mean read counts of the distal segment. In order for the distal segments to be comparable, the distal segment coverage d_s in sample s is first scaled to the maximum proximal coverage p_s across all samples. Formally, each d_s is multiplied by λ_s:

$$\lambda_{\mathrm{s}} = \frac{\max_{\mathrm{s}} \mathrm{p_s} + 1}{\mathrm{p_s} + 1}$$

Because the number of replicates per condition is typically small in a given experiment, a standard t-test is underpowered to detect a difference between two means. Instead, we use a data-driven estimation of expected variance across all tested distal segments. LOESS regression was used to predict CV_c^2:

$$\mathrm{CV}_c^2 = \left(\frac{\sigma_c}{\mu_c}\right)^2$$

given the $\log_2\mu_c$ where μ_c and σ_c are the mean and standard deviation of λ_s d_s in condition c respectively. As described in voom[79], the normal distribution is appropriate to model read count data. Assuming a normal distribution and two conditions A and B, a p-value is then calculated by testing the probability of observing $\min(\mu_A, \mu_B)$ from distribution:

$$\mathrm{N}\left(\max(\mu_A, \mu_B), \underset{\sigma_c}{\mathrm{argmax}}(\mu_A, \mu_B)\right)$$

These p-values are corrected with the Benjamini-Hochberg (BH) procedure. Change points with BH-corrected p-value < t (default = 0.05) and $|\mathrm{RU}_{k,A} - \mathrm{RU}_{k,B}|$ > m (default = 0.05) are considered significantly differential.

If only a single sample is input, then mountainClimberTest_diff will label each 5' and 3' end as follows. The distal ends are labeled Distal or DistalOnly if there were or were not alternative ends, respectively. If strand could be inferred, then proximal ends are labeled alternative first exon (AFE) or alternative last exon (ALE) when segments are non-consecutive, and TandemTSS or TandemAPA when segments are consecutive. If strand could not be inferred, then proximal ends are either alternative terminal exon (ATE) or Tandem.

Methods supporting the change point identification pipeline

annotate_change_points(j): Annotate mountainClimberCP output with gene regions given an annotation file. If a change point is within j bp (default = 10) of an annotated TSS, poly(A) site, or exon-intron junction, then it is labeled TSS, POLYA, and JXN respectively; otherwise, it is given one of the other labels. Prioritization of annotation is as follows: JXN/TSS (overlapping both JXN and TSS) > JXN/POLYA > JXN > TSS > POLYA > CDS (coding exon) > 3UTR > 5UTR > NC (exon in non-coding gene) > INTRON > INTERGENIC (outside of annotated TUs).

get_junction_counts(h, m, a): Given an input bam file, identify exon-exon junction reads that cover ≥ h bp in each exon (default = 8), and span introns ≥ m (default = 30) and ≤ a (default = 500,000) bps long.

merge_tus: Merge TUs that overlap across samples and merge those within the same gene annotation.

Mapping pipeline

There are two mapping pipeline options: (1) aligning to the genome and calling all de novo TUs followed by change point identification, or (2) skip genome alignment and align directly to the transcriptome. In both pipelines, RSEM [80] was used to assign the most likely location for multi-mapped reads. While pipeline #1 will yield more (unannotated) TUs, pipeline #2 is significantly faster because there's only one alignment instead of two, and the RSEM reference is only built once instead of once per biological condition. See Fig. 3.S1 for a stepwise summary of the two pipelines. MAQC was run with both mapping pipelines for comparison: pipeline #1 with STAR v2.5.2a [81] and GENCODE v25, and pipeline #2 with hisat2 v2.0.5 [82] and Ensembl release 75, both aligned to hg19. Simulated RNA-Seq reads were aligned with pipeline #1 using hisat2 and mm10 with Ensembl release 84.

Alignment to genome: For MAQC, 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.

After alignment to the genome, bam files were converted to bedgraphs using bedtools genomecov, get_junction_counts, mountainClimberTU, and merge_tus (see above).

For simulations, hisat2 was run with the following parameters: --dta-cufflinks --mp 6,4 --no-softclip --no-mixed --no-discordant --add-chrname -k 100.

Prepare RSEM reference: The RSEM reference is prepared with rsem-prepare-reference from the gtf file output from merge_tus in pipeline #1 or from a gtf of annotated transcripts +/-10kb in pipeline #2. It is recommended to merge de novo TUs with an existing annotation so that the aligner can better identify splice sites, as the de novo TUs do not contain splice site information.

Alignment to transcriptome: For MAQC alignment with pipeline #1, STAR was used with the following parameters to allow 200 multi-mapped reads and be compatible with RSEM: --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.

For MAQC alignment with pipeline #2 and simulation alignment with pipeline #1, hisat2 was used with the following parameters to allow 100 multi-mapped reads and be compatible with RSEM (i.e. force hisat2 not to report indels): --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.

RSEM: rsem-calculate-expression was run with arguments --paired-end --append-names --seed 0 --estimate-rspd --sampling-for-bam --output-genome-bam, and the alignment with maximum posterior probability was kept for each multi-mapped read.

Simulations

We used Flux Simulator [83] with default parameters to simulate 100bp paired-end RNA-Seq of annotated TandemUTRs downloaded from the MISO website and derived from PolyA DB [84,85]. TandemUTR transcript isoforms were simulated with 1, 2, or 3 change points with varying distal / proximal expression (Fig. 3.S2a,b). Fasta files were generated and reads were aligned using the mapping pipeline #1 with hisat2 as described above. Transcripts with at least 10 average reads per bp in exons were considered for precision and recall calculations. Predicted DistalPolyA and TandemAPA change points were considered for precision and recall calculations. To compare varying degrees of difficulty in identifying 3' ends, we created equal sized bins by fold change of (proximal + 1) / (distal + 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.

IsoSCM was run with default parameters. To calculate the fold change of (proximal + 1) / (distal + 1), we used the coverage reported in the coverage.gtf file, which represents the median read coverage of each segment. Predicted '3p_exon' change points were considered for precision and recall calculations ('5p_exon' and 'internal_exons' were ignored). To assign true fold changes of annotated genes to IsoSCM's predictions, we overlapped the predictions with the true TandemUTR 3' ends +/-51bp.

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 [86]. PolyA-Seq sites were downloaded from the UCSC Genome Browser: MAQC UHR 1, UHR 2, Brain 1, Brain 2 [64]. Polyadenylation sites with at least 1 RPM were considered.

RNA-Seq reads were mapped as described below, except junction reads were not used in mountainClimberTU due to many split reads spanning multiple genes for unknown reasons. Adapters were trimmed with cutadapt -a AGATCGGAAGAG -n 5 -m 35 -O 5 -e 0.2 -q 20 [26]. After calling TUs with mountainClimberTU and merging overlapping TUs across samples, change points were called with mountainClimberCP only in TUs with at most one annotated gene and at least 10 average reads per bp. Predicted 5' and 3' ends of multi-exon genes for which the strand could be inferred were interrogated for FANTOM CAT TSSs [63] and PolyA-Seq poly(A) sites within 200bp on the same strand at the 5' and 3' ends respectively. If a prediction was close to multiple FANTOM CAT or PolyA-Seq sites, only the closest one was counted to avoid counting any predictions, FANTOM CAT, or PolyA-Seq sites more than once. Similar to the simulation analysis, we reported the precision binned by fold change at the predicted change points, with PolyA-Seq poly(A) sites or FANTOM CAT sites considered as true positives.

3.4. Results

3.4.1 A de novo approach for change point detection in RNA-Seq

Our approach is based on a notion of measuring non-uniformity in RNA-Seq and finding positions where the degree of non-uniformity changes significantly. By measuring the non-uniformity, we are inherently robust to it when calling change points (Fig. 3.1 and Fig. 3.S1). This process is

comprised of three major steps: (1) identifying de novo TUs by finding all continuous regions with RNA-Seq reads in each sample with mountainClimberTU (Fig. 3.1a), (2) identifying change points in the entire TU in each sample without restricting to either 5' or 3' end with mountainClimberCP (Fig. 3.1b), and (3) combining replicates and testing for differential 5' and 3' end usage with mountainClimberTest (Fig. 3.1c) (Section 3.3 Methods).

mountainClimberTU identifies de novo TUs by merging consecutive 1kb bins that meet specified RNA-Seq depth and breadth criteria (Section 3.3 Methods). By identifying TUs de novo, we are not limited by gene annotation, which may exclude TSSs or poly(A) sites outside of the gene annotation. Additionally, transcriptional read-through may be detected, depending on the RNA-Seq library type (for example in chromatin-associated RNA-Seq).

To call change points while maintaining robustness to RNA-Seq non-uniformity, mountainClimberCP leverages the non-uniformity by calculating the Cumulative Read Sum (CRS) and finding positions where the CRS significantly deviates from the uniform distribution. The premise of this approach is similar to that of the classical CUSUM approach [87] in that it measures the cumulative sum of differences in read counts of neighboring base pairs, but differs in the following capacities: (1) it does not require any weights to be assigned to each base pair, (2) RNA-Seq features including exon-exon junction reads are leveraged to optimize parameters, and (3) the change point detection approach differs. While change points are presented as elbows in the CRS as a function of position (Fig. 3.1b "1. Calculate CRS"), highly non-uniform RNA-Seq will not significantly deviate from the uniform distribution (diagonal line in Fig. 3.1b "1. Calculate CRS"). Since we expect to identify significant change points at exon-exon junctions, the parameters are optimized by maximizing the number of exon-exon junctions with predicted change points. Additionally, because this approach is inherently robust to RNA-Seq non-uniformity, it can be used in chromatin-associated RNA-Seq which is more non-uniform or "spiky" than poly(A)-selected RNA-Seq [88] (see Chapter 5).

mountainClimberTest is comprised of three major steps: (1) cluster change points within and across any number of experimental conditions using DBSCAN, (2) calculate the relative usage (RU), a value between 0 and 1 indicating the usage of each change point at the 5' and 3' ends relative to the proximal-most segment, and (3) test for differential change point usage in two conditions. Because change points are identified in the entire TU, mountainClimber identifies differential 5' and 3' ends that other methods cannot capture, e.g. intronic APA and alternative first or last exons (Fig. 3.2g).

3.4.2 mountainClimber performance evaluation

To evaluate the performance of mountainClimber, we simulated known alternative 3' ends (TandemUTRs downloaded from MISO and derived from PolyA DB) with Flux Simulator [83] with up to three change points per gene and with varying levels of distal vs. proximal 3' end usage (Section 3.3 Methods; Fig. 3.S2a,b). As expected, recall and precision both increased with higher fold change at the 3' end (Fig. 3.2a,b and Fig. 3.S2c,d). We chose to compare mountainClimber with IsoSCM [70] because it identifies both 5' and 3' ends while many other existing methods identify alternative 3' ends only. While mountainClimber and IsoSCM obtained similar recall, mountainClimber outperformed IsoSCM in terms of precision. Upon manual investigation of some examples, this appears to be caused by IsoSCM predicting multiple 3' ends very close together in the same TU.

To further evaluate the performance of mountainClimber on real data, we analyzed RNA-Seq from MAQC Universal Human Reference RNA and Ambion Human Brain Reference RNA with paired poly(A) sites experimentally identified with PolyA-Seq [64] (Section 3.3 Methods). 18,964 total TUs were called, 13,977 (74%) of which were annotated and 10,513 (55%) of which had at most one gene annotation. From these 10,513, an average 49,610 change points per sample

were identified, 75% of which were exon-exon junctions (Table 3.S1). On average, we found 2,565 TandemAPA and 776 TandemTSS cases per sample.

mountainClimberCP achieved higher precision than IsoSCM (Fig. 3.2c,d). In the highest fold change bin, mountainClimberCP reached a maximum precision of 75% while IsoSCM reached 56%. This was especially apparent for 3' ends with lower fold change, where mountainClimberCP and IsoSCM reached a maximum precision of 42% and 23% respectively. This maximum achieved precision is lower than what Shenker et al. reported (approximately 80%), which may be due to the difference in bin definitions or the use of different MAQC datasets. Moreover, mountainClimberCP identified an average of 2.34 fold (2,259 vs. 964) more 3' ends overlapping PolyA-Seq compared to IsoSCM (Table 3.S2). The precision of TandemAPA and DistalPolyA sites was highest in annotated poly(A) sites and 3'UTRs (Fig. 3.S2e). Similar to PolyA-Seq, the poly(A) signal motif A[A/T]TAAA was enriched at predicted 3' ends with higher fold change (Fig. 3.S2f). Cases that were missed by mountainClimberCP may be due to differences in limitations of both RNA-Seq and PolyA-Seq. PolyA-Seq may detect smaller magnitude APA events that are undetectable in RNA-Seq. On the other hand, mountainClimberCP may have higher sensitivity in other cases such as regions with multi-mapped reads, as PolyA-Seq only considered uniquely mapped reads.

Analogous to PolyA-Seq at the 3' end, we used FANTOM CAT [63] to evaluate mountainClimberCP predictions at the 5' end (Section 3.3 Methods) and again found that it outperforms IsoSCM (Fig. 3.2e,f). Both methods had lower precision at the 5' end compared to the 3' end, which may be due to biases in RNA-Seq and CAGE. Considering that TSS regions are hard to predict given RNA-Seq alone, mountainClimberCP performs reasonably well.

To test whether accuracy suffers by skipping the genome alignment step and aligning directly to the transcriptome (Fig. 3.S1), and to additionally compare aligners, we compared STAR with genome alignment to hisat2 with transcriptome alignment only. Evaluation of MAQC PolyA-

Seq and FANTOM CAT TSS enrichment in hisat2 with transcriptome alignment showed no distinguishable difference compared to STAR with genome alignment, suggesting that mountainClimberCP does not suffer by excluding the genome alignment step for poly(A)-selected RNA-Seq (Fig. 3.S2g,h).

Finally, mountainClimberTest was used to test for significant differential end usage across the two conditions (Section 3.3 Methods). 942 / 6,854 total ends met the criteria for testing: at least 5 average reads/bp in the distal segment in at least 1 condition, non-zero proximal coverage in all samples, and strictly decreasing average reads/bp from proximal to distal. 867 of these were tandem in at least one condition and tested for significant alternative end usage. 348 had BH-corrected p-value < 0.05, and 246 of these additionally had an absolute relative usage (RU) difference > 0.05 (Fig. 3.2g). The top ATSS and APA examples were supported by the annotation, supporting the validity of our approach (Fig. 3.2h,i).

3.5. Discussion

Here, we presented mountainClimber, a novel de novo approach for identifying alternative 5' and 3' ends from RNA-Seq. Unlike several existing approaches, it identifies change points anywhere in a transcription unit and identifies more than two change points[67–69,71–73]. Compared to another approach that identifies change points in the entire TU, IsoSCM[70], mountainClimber achieved higher precision. Additionally, mountainClimberTest tests for significant differential change points across two groups of samples, while IsoSCM handles only one sample per group and recommends pooling bam files from replicate samples, thereby potentially diminishing useful signal.

One limitation of our approach as well as all other existing alternative 5' and 3' end identification approaches is that they focus on ATSS and APA but ignore alternative splicing. Many tools are built for alternative splicing analysis, but these do not identify ATSS or APA [84,89,90].

Recently, full-length isoform sequencing revealed that >60% of genes with multiple transcripts showed coupling between TSS, poly(A) sites, and splicing [91]. In the future, it will be useful to build software that simultaneously detects alternative splicing, ATSS, and APA. While transcriptome assembly-based approaches such as Cufflinks and Scripture [92,93] can theoretically achieve this, methods specifically build for ATSS and APA identification typically outperform them in terms of ATSS and APA accuracy [70], suggesting that there is much room for improvement.

3.6. Figures

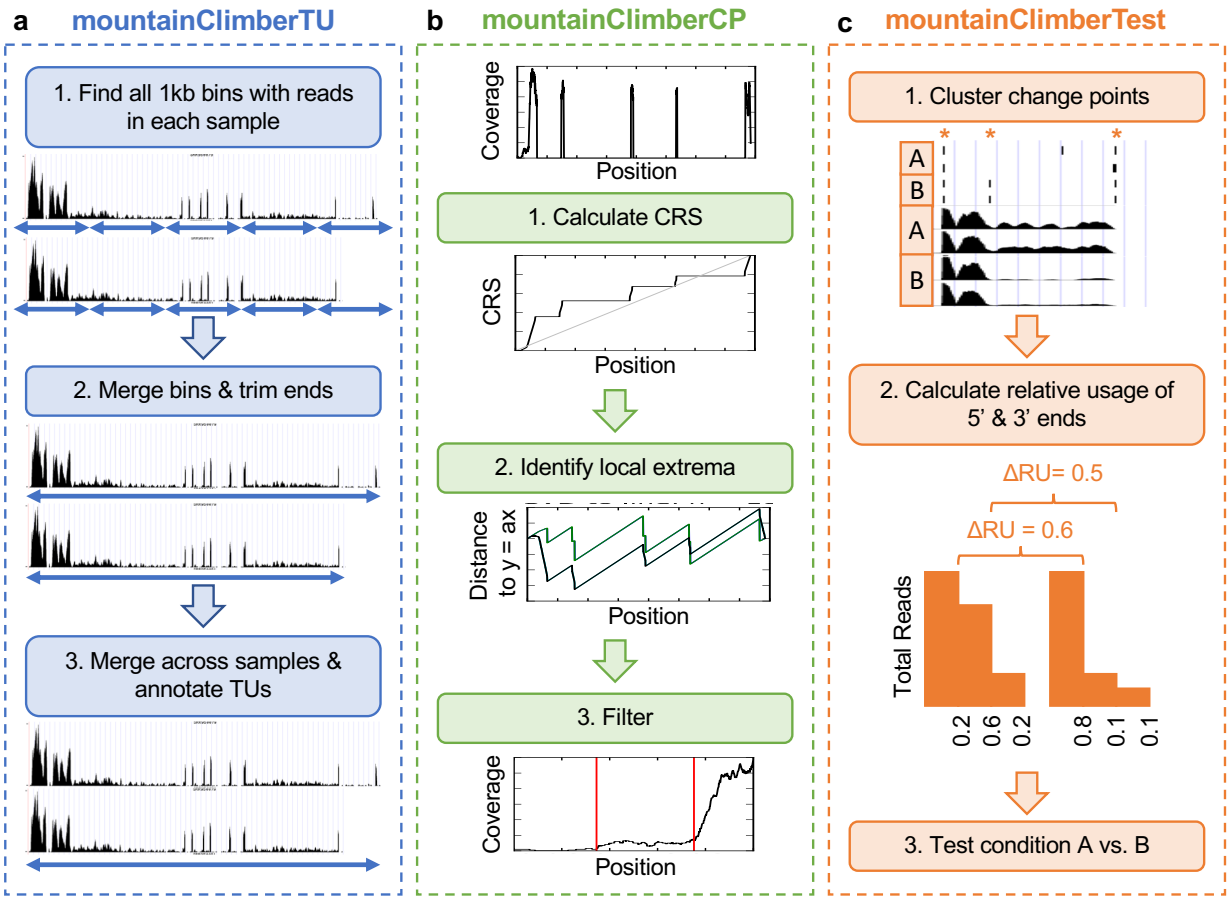


Figure 3.1 | Schematic of the complete de novo change point identification pipeline. (**a**) The mountainClimberTU approach for calling de novo transcription units (TUs). Blue double-headed arrows indicate 1kb bins in step #1, and ultimately TUs in step #3. (**b**) The mountainClimberCP approach for identifying change points in each TU in each sample. The first panel shows poly(A)-selected RNA-Seq reads per base pair in an example gene. Step #1 shows the cumulative read sum (CRS) distribution in black compared to the uniform distribution (diagonal line y = x) in grey. Step #2 shows the identification of elbows in the CRS distribution by calculating the distance from the CRS and weighted CRS (wCRS) to the diagonal line y = x in black and green respectively. Note that wCRS is needed to observe elbows corresponding to both exon-intron junctions for each exon. Step #3 shows the final results after filtering, where red lines indicate change points identified from the RNA-Seq read distribution in black. (**c**) The mountainClimberTest approach for identifying differential 5' and 3' end usage for a toy example in conditions A & B with two replicates each. Step #1 shows the change points (orange asterisks) identified in each sample above the corresponding read distributions (one outlier in sample A was not clustered). Step #2 shows a cartoon example of relative usage (RU) calculation based on average reads/bp in each segment. Finally, step #3 is the test for significant differential 5' and 3' end usage. For more details, see Section 3.3 Methods and Fig. 3.S1.

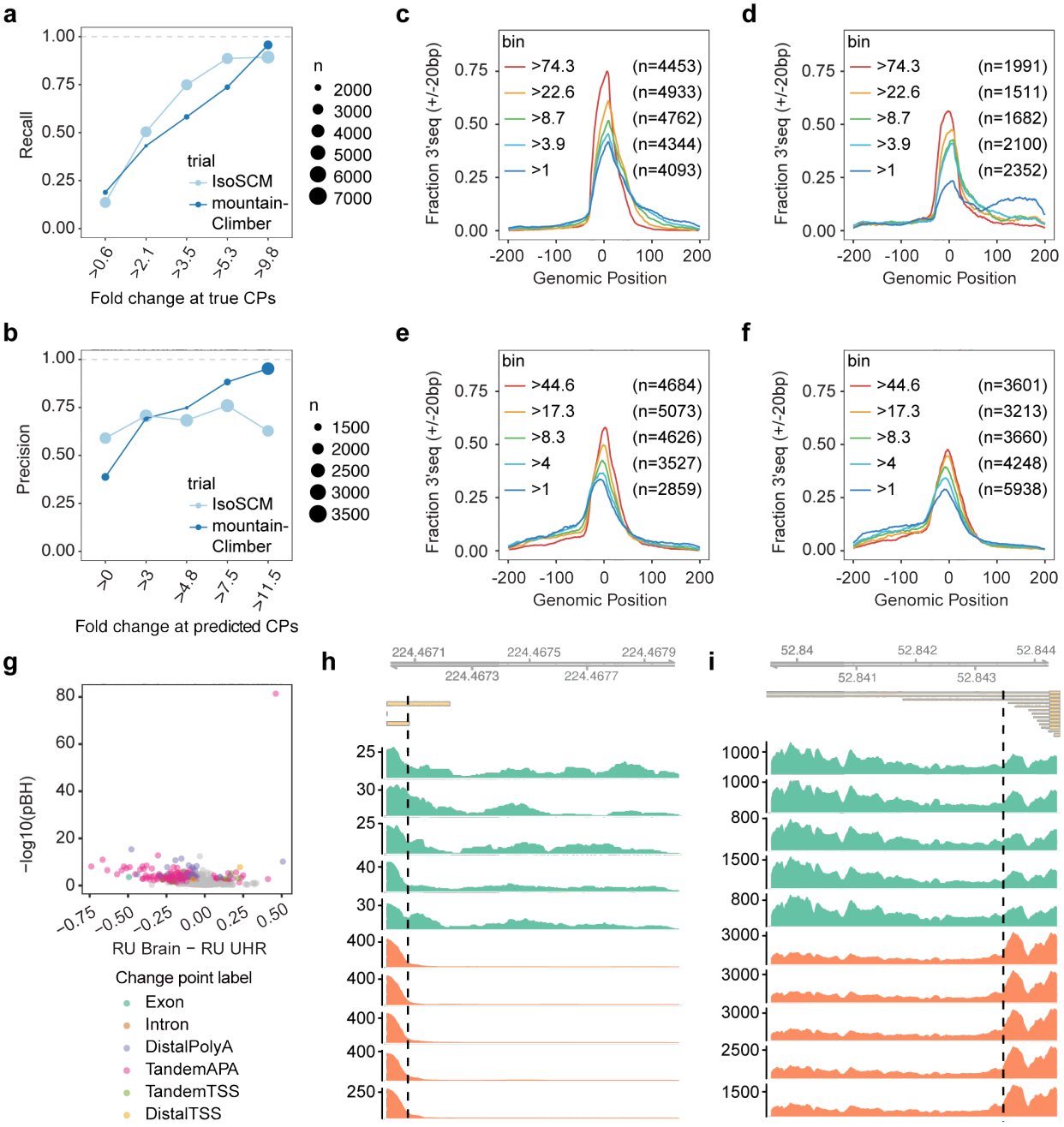


Figure 3.2 | mountainClimber performance evaluation.
(**a**) Recall of simulated 3' ends stratified by the fold change of average reads/bp of proximal vs. distal segments at the true simulated change points (CPs). Point size indicates the number of predicted 3' ends. (**b**) Precision of simulated 3' ends stratified by the fold change of average reads/bp of proximal vs. distal segments at the predicted 3' ends. Point size indicates the number of predicted 3' ends. (**c, d**) Overlap of 3' end predictions from RNA-Seq with PolyA-Seq for mountainClimberCP (**c**) and IsoSCM (**d**). The x-axis indicates the positional difference between predicted and PolyA-Seq sites, so positive (negative) values mean the prediction is downstream (upstream) of the PolyA-Seq site. The y-axis indicates the total predictions with PolyA-Seq within +/-20bp as a fraction of the total predictions within 200bp for each bin. The inset indicates the minimum fold change for each bin and the number of change points in each bin. (**e, f**) Overlap of

5’ end predictions from RNA-Seq with FANTOM CAT TSSs for mountainClimberCP (**e**) and IsoSCM (**f**), similar to (**c,d**)**.** (**g**) Volcano plot of mountainClimberTest BH-corrected p-value vs. relative usage (RU) difference in brain vs. UHR for 867 tandem ends. (**h,i**) Top examples of ATSS and APA by BH-corrected p-value. Genomic coordinates in the top panel are reported on megabase scale. The top five read distributions are UHR and the bottom five are Brain. Black dashed vertical lines indicate change points identified.(**h**) ATSS in Scg2 (BH-corrected p = 2.61e-10, RU difference = 0.23), and (**i**) APA in Arpp19 (BH-corrected p-value = 4.57e-85; RU difference = 0.46).

3.7. Supplementary Figures

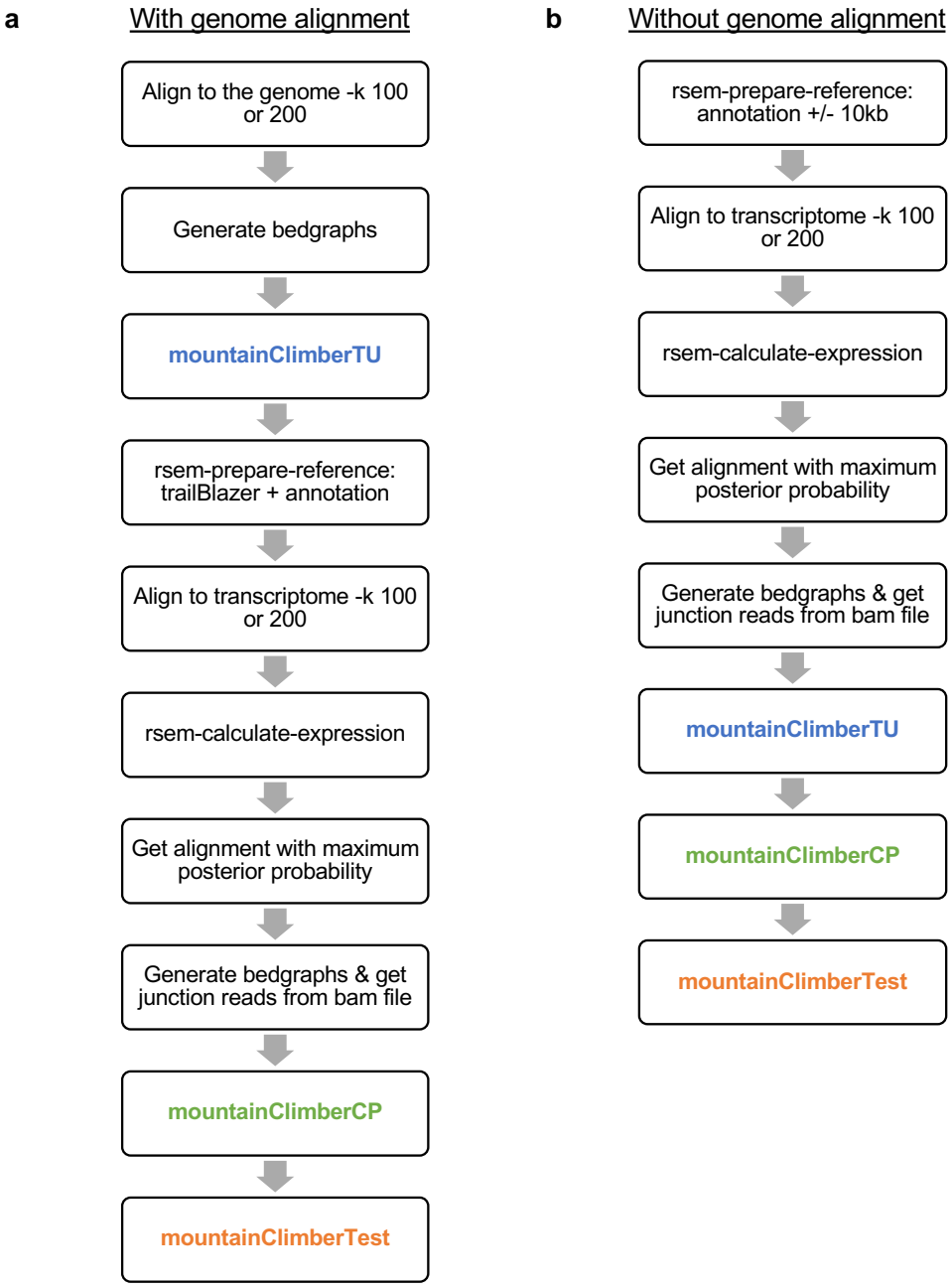


Figure 3.S1 | Mapping and change point identification pipeline.
(**a**) Approach with genome alignment (used for MAQC in Chapter 3). This approach is slower but identifies more novel transcription units. (**b**) Approach without genome alignment (used for GTEx in Chapter 4). This approach is faster and identifies change points in annotated genes.

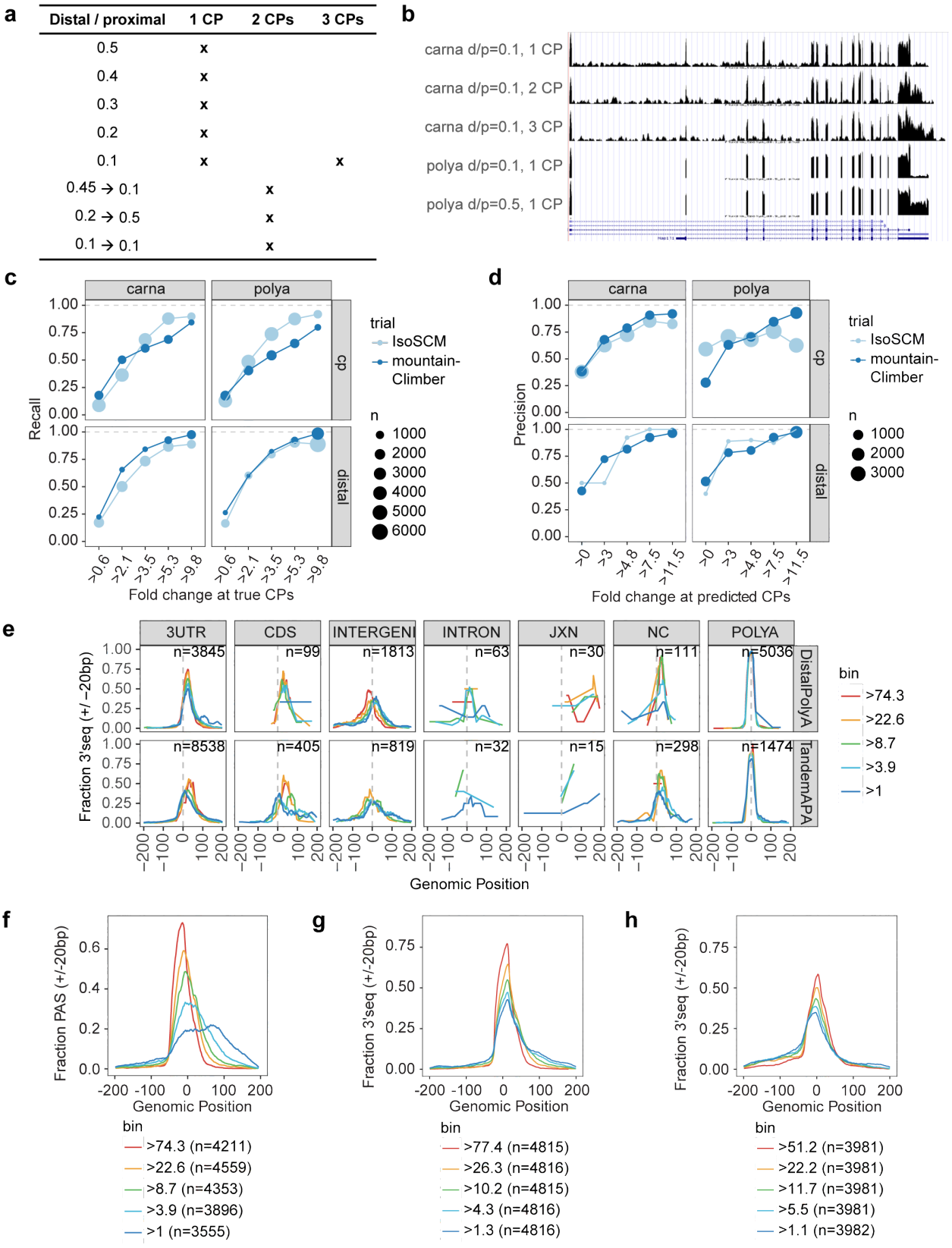


Figure 3.S2 | mountainClimberCP performance evaluation.

(**a**) Schematic of the simulation approach. “Distal / proximal” indicates the expression of the distal segment relative to the proximal. (**b**) Examples of simulated change points (CPs) for chromatin-associated RNA-Seq (carna) and poly(A)-selected RNA-Seq (polya). (**c, d**) Recall (**c**) and precision (**d**) of carna and polya simulations stratified by whether the true 3’ end was a tandem proximal change point (cp) or distal end (distal). (**e**) Overlap of 3’ end predictions from RNA-Seq with PolyA-Seq as in Fig. 3.2c, stratified by change point label (rows) and overlap with GENCODE v25 (columns). Inset text indicates the total number of predicted change points. GENCODE regions with less than 5 DistalPolyA or TandemAPA were excluded (these included JXN/POLYA and TSS). (**f**) Similar to Fig 3.2c, but plotting poly(A) signal (PAS) motif overlap (A[A/T]TAAA). (**g,h**) PolyA-Seq (**g**) and FANTOM CAT TSS (**h**) overlap using hisat2 and Ensembl 75 alignment directly to the transcriptome (Fig. 3.S1b).

3.8. Supplementary Tables

Table 3.S1 | mountainClimberTest_diff cases given two conditions, A and B.

Distal A	Proximal A	Distal B	Proximal B	Segments Consecutive	Description
X	X	X	X	X	Test if min distal ≥ d in A or B & proximal ≥ p in both A & B.
X	X	X	X		Output without testing.
X	X	X		X	Test if min distal ≥ d in A or B & proximal ≥ p in both A & B.
X	X	X			Output without testing.
X		X	X	X	Test if min distal ≥ d in A or B & proximal ≥ p in both A & B.
X		X	X		Output without testing.
X	X			X	Test if min distal ≥ d in A or B & proximal ≥ p in both A & B.
X	X				Output A only if any CPs were called in this gene in B (otherwise, gene not expressed in B à do not output).
X					Output A only if any CPs were called in this gene in B (otherwise, gene not expressed in B à do not output).
		X	X	X	Test if min distal ≥ d in A or B & proximal ≥ p in both A & B.
		X	X		Output B only if any CPs were called in this gene in A (otherwise, gene not expressed in A à do not output).
		X			Output B only if any CPs were called in this gene in A (otherwise, gene not expressed in A à do not output).
X		X			No difference. Do no output.

Table 3.S2 | MAQC mountainClimberCP total change points.

Sample	Type	Total genes	Total change points	Tan-dem-TSS	Distal-TSS	Tan-dem-APA	Distal-PolyA	Junc-tion	Intron	Tan-dem-Right	Tan-dem-Left	Exon	Distal-Right	Distal-Left
SRR950078	UHR	3,093	49,912	758	2,580	2,617	2,580	36,486	1,612	-	1	2,252	513	513
SRR950079	Brain	3,068	51,110	793	2,567	2,732	2,567	37,846	1,389	1	1	2,212	501	501
SRR950080	UHR	2,953	49,538	727	2,541	2,499	2,541	37,427	1,272	-	1	1,706	412	412
SRR950081	Brain	3,092	51,266	795	2,586	2,786	2,586	38,066	1,300	1	2	2,132	506	506
SRR950082	UHR	2,586	45,213	619	2,325	2,336	2,325	34,469	1,337	-	-	1,280	261	261
SRR950083	Brain	3,139	51,816	821	2,577	2,792	2,577	37,814	1,602	1	1	2,507	562	562
SRR950084	UHR	3,338	52,255	747	2,670	2,617	2,670	38,042	1,493	2	2	2,676	668	668
SRR950085	Brain	2,880	49,129	718	2,481	2,649	2,481	37,018	1,282	-	-	1,702	399	399
SRR950086	UHR	2,789	49,533	893	2,485	2,297	2,485	38,842	860	1	1	1,061	304	304
SRR950087	Brain	2,561	46,323	888	2,311	2,322	2,311	36,137	872	-	-	982	250	250
	mean	**2,950**	**49,610**	**776**	**2,512**	**2,565**	**2,512**	**37,215**	**1,302**	**1**	**1**	**1,851**	**438**	**438**

Table 3.S3 | MAQC mountainClimberCP vs. IsoSCM.

	Total change points		Total with ≥ 50 reads/bp in exons		Total 3' (3p_exon for IsoSCM; TandemAPA & DistalPolyA for mountainClimber)		Total with PolyA-Seq within 200bp		Total after filtering non-unique PolyA-Seq sites	
	mountain-Climber	IsoSCM	mountain-Climber	IsoSCM	mountain-Climber	IsoSCM	mountain-Climber	IsoSCM	mountain-Climber	IsoSCM
SRR950078	47,025	44,592	47,025	38,405	5,197	19,014	2,872	800	2,208	769
SRR950079	48,215	42,709	48,215	37,786	5,299	18,433	2,990	1,128	2,325	1,100
SRR950080	47,315	42,867	47,315	36,866	5,040	18,329	2,889	807	2,227	772
SRR950081	48,446	42,847	48,446	37,996	5,372	18,565	3,033	1,118	2,370	1,094
SRR950082	43,724	41,989	43,724	35,216	4,661	17,474	2,789	829	2,113	804
SRR950083	48,535	43,323	48,535	38,267	5,369	18,727	3,039	1,172	2,348	1,145
SRR950084	48,621	44,367	48,621	38,780	5,287	19,314	2,918	872	2,249	828
SRR950085	46,905	41,884	46,905	36,581	5,130	17,856	2,996	1,154	2,309	1,130
SRR950086	48,054	40,091	48,054	34,012	4,782	16,868	2,924	851	2,209	823
SRR950087	45,017	39,370	45,017	33,370	4,633	16,229	2,955	1,193	2,227	1,171
mean	**47,186**	**42,404**	**47,186**	**36,728**	**5,077**	**18,081**	**2,941**	**992**	**2,259**	**964**

Chapter 4: The landscape of alternative transcription start and polyadenylation sites in human tissues

4.1. Abstract

As described in Section 3.1, alternative transcription start sites (ATSS) and alternative polyadenylation (APA) affect the majority of mammalian genes, often in a tissue-specific manner. However, most prior studies of tissue-specific APA were limited to annotated poly(A) sites. Additionally, tissue-specific ATSS is less well studied compared to APA. We discovered ATSS and APA in 2,342 GTEx samples (36 tissues, 215 individuals) with mountainClimber, constituting the largest study of ATSS and APA to our knowledge. 70% and 56% of tested transcription units (TUs) exhibited differential APA and ATSS respectively across tissues. Globally, 3'UTRs were longest in the brain and shortest in blood and testis, consistent with previous studies. On the other hand, 5'UTRs were longest in testis and shortest in skeletal muscle, which was not previously reported. Interestingly, we found that the cerebellum was distinct from other brain regions in both 5' and 3' ends. Overall, this study reports the most comprehensive characterization of 5' and 3' ends across human tissues to date.

4.2. Introduction

APA and ATSS affect the majority of genes in mammals. APA often results in changes in mRNA stability, translation, nuclear export, and localization (reviewed in [9,10]). For example, shorter 3'UTRs were observed in tumors compared to normal cells and were associated with more stable mRNA [69]. APA can also affect membrane protein localization; for example, the long 3'UTR isoform of CD47 localizes to the cell surface while the shorter 3'UTR isoform localizes to the endoplasmic reticulum [94]. In terms of nuclear export, shorter 3'UTRs were observed in the cytoplasmic

compared to the nucleoplasmic cell fraction[95]. While APA often occurs in the last exon, termed tandem APA, it was recently shown that there are also many intronic APA events, especially in blood-derived immune cells [96]. This was reported to often remove significant portions of the coding region leading to either truncated proteins or non-coding transcripts. While tissue-specific APA isoforms in human were previously observed in RNA-Seq, these studies were limited by restricting to annotated poly(A) sites [97,98]. Other smaller scale studies of tissue-specific APA employed RNA sequencing protocols to specifically identify poly(A) sites [64,66].

While APA often affects mRNA levels, localization, and translation, ATSS most often results in translational changes. Various sequence and structural motifs in the 5'UTR can affect translation including upstream open reading frames, internal ribosomal entry sites, as well as binding sites for long non-coding RNAs and RNA binding proteins [7,60]. Additionally, upstream TSS usage often inhibits translation of the canonical translation start site. For example, differential ATSS isoform expression drives protein level changes at different stages of meiosis [99]. In this way, different transcription factors can orchestrate usage of different TSSs to control protein levels without necessarily changing mRNA levels.

Here, we conducted the largest study of tissue-specific ATSS and APA to our knowledge using GTEx RNA-Seq [12] and the mountainClimber pipeline described in Section 3.3. Our findings expand the repertoire of tissue-specific APA and ATSS sites in humans.

4.3. Methods

GTEx sample selection

GTEx [100] RNA-Seq was downloaded through dbGaP under accession phs000424.v6.p1, except for metadata which was downloaded under accessions phs000424.v7.pht002742.v7.p2.c1 (subject phenotypes) and phs000424.v7.pht002743.v7.p2.c1 (sample attributes). A subset of GTEx samples were chosen as follows to maximize the number of tissues available per donor:

individuals with at least 20 tissues (excluding Cells), tissues with less than 20 samples, excluding 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 read1 and read2 or too few reads, and those tissues from the same individual that were released multiple times (the most recent release was kept). Those with read length other than 76bp, median transcript integrity (TIN) score < 70 (calculated using tin.py from RSeQC[101]), and mapping rate < 50% were retained, resulting in 2,342 samples from 36 tissues and 215 individuals (Table 4.S1).

Mapping pipeline and change point identification

Adapter trimming and quality check: Adapters were trimmed with cutadapt [26] -m 35 -n 3 -a AGATCGGAAGAGC -a CAAGCAGAAGACGGCATACGAG -g AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC -A AGATCGGAAGAGC -A CAAGCAGAAGACGGCATACGAG -G AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC, adding the following parameters if FastQC returned "WARN" or "FAIL" for respective categories: adapter content, per-base sequence quality, -q 20,20; sequence content, --trim-n (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/).

Change point identification: The mountainClimber mapping pipeline #2 was used (Fig. 3.S1) with hisat2 v2.0.5[82] and Ensembl release 75 aligned to hg19 standard chromosomes.

Overlap between Ensembl and mountainClimberTU TUs

Ensembl TUs were compared with mountainClimberTU TUs to assess whether the total number of mountainClimberTU TUs containing a single annotated gene was reasonable (Fig. 4.S1a). Because the GTEx RNA-Seq is non-strand-specific, we first merged all overlapping genes on either strand, resulting in 34,008 total Ensembl TUs. Next, mountainClimberTU TUs were

stratified based on whether they overlapped one or multiple Ensembl TUs. Those overlapping one Ensembl TU were further stratified by whether the Ensembl TU contained one or multiple genes. The distance between Ensembl TUs within the same mountainClimberTU TU are shown in Fig. 4.S1b. Manual inspection of some examples of non-overlapping Ensembl TUs in the same mountainClimberTU TU revealed high levels of RNA-Seq reads between Ensembl TUs, suggesting that the genes could not be confidently separated into separate TUs based on RNA-Seq.

GTEx weighted mean extension length (WMEL) calculation & downstream analysis

Relative usage calculation for each sample: mountainClimberTest_ru was modified to calculate relative usage (RU) of individual change points rather than change points clustered across replicates. Because only one sample was considered, there were by definition no condition-specific internal change points, so this step was ignored (see Section 3.3 for more details). Finally, mountainClimberTest_diff was used on each sample to label ends as TandemTSS, TandemAPA, Distal, or ATE (alternative terminal exon).

5' and 3' end grouping and weighted mean length: After labeling the 5' and 3' end segments as Tandem, Distal, or ATE (alternative terminal exon), TUs were separated into four categories for 5' and 3' ends separately: [1] alternative terminal exon (ATE) within a single sample, [2] ATE across different samples, [3] the last segment is disrupted by an annotated intron, [4] tandem and distal UTRs (Fig. 4.S3a). To check for ATE 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 last segment was entirely 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, weighted by each segment's RU. Finally, the weighted mean extension length (WMEL) was calculated by subtracting the minimum WML across all 2,342 samples from each WML (Fig. 4.S2b).

Covariate correction: 14 covariates selected in a previous study to correct gene expression were used to correct WMEL [102]: AGE, HGHT, WGHT, BMI, ETHNCTY, SEX, MHCANCERNM, SMRIN, SMATSSCR, SMCAT, SMCENTER, TRTPTREF, SMNABTCH, COHORT. We used the lm function in R (wmel_lm = lm(wmel ~ covariates)) and residuals(wmel_lm) as the corrected values.

Analysis of the contribution of tissue and individual to WMEL variation: To assess the contribution of either tissue or individual to the observed variation in WMEL, we adopted an approach similar to [103]. 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 [104] as follows: lmer(log2(WMEL) ~ (1|tissue) + (1|subject) + AGE + ETHNCTY + SEX). To calculate the contribution of tissue and subject, we divided their REML-estimated variance by the sum of the estimated variance of tissue + individual + residual.

GTEx tissue specificity analysis

WMEL variation per TU across tissues: To identify the subset of TUs with varying WMEL across tissues, we tested for differential log2(WMEL) using the R pairwise.t.test() function across all pairs

of tissues for each 5' and 3' end. A maximum Bonferroni-adjusted p-value of 0.01 was required, resulting in 301 and 321 total TUs for 5' and 3' ends respectively. Tissue-wise correlation of the 301 5' and 321 3' end lengths was conducted using the mean WMEL across individuals in each tissue. R^2 was calculated with R's cor(method = "spearman", use = "complete.obs") ** 2. To visualize global tissue-wise WMEL mean and variance, we took the mean and variance of WMEL across individuals for all TUs in each tissue. TU complexity was defined as the mean number of change points with RU ≥ 10% across individuals in each tissue.

mountainClimberTest: To identify significantly differential change points, mountainClimberTest was run for all pairwise tissue comparisons. For chrY, only males were clustered. Default parameters were used for mountainClimberTest_cluster, except the minimum number of points required for DBSCAN to call a cluster, -n, was set to 50% of the number of individuals in that tissue.

Gene ontology analysis

Gene ontology (GO) analysis was performed controlling for gene length and GC content as described previously [57]. The union of all genes that were differential in each of the five tissues with the most differential genes (Whole-Blood, Testis, Muscle-Skeletal, Brain-Cerebellum, and Brain-Cerebellar-Hemisphere) compared to all 35 other tissues were combined for GO analyses at the 5' and 3' ends. Background gene lists were chosen from the 516 and 420 genes with ubiquitous expression and tandem UTRs respectively. Empirical p-values were corrected with the Benjamini-Hochberg procedure.

4.4. Results

4.4.1. Identification of alternative 5' and 3' ends in human tissues

To investigate ATSS and APA in humans, we analyzed 2,342 samples from 36 tissues and 215 individuals from GTEx [100] after excluding samples with low mapping rate and median TIN scores (Section 4.3 Methods and Table 4.S1). To align thousands of samples in a reasonable amount of time and focus on analyzing alternative 5' and 3' ends of annotated genes, we used mapping pipeline #2 described in Section 3.3.

30,442 total TUs were identified over all GTEx samples. 21,231 overlapped at most one gene annotation. The remaining 9,211 TUs contained more than one gene annotation, most often because they either overlap in the Ensembl annotation (Fig. 4.S1a) or were in close proximity and joined due to high RNA-Seq levels (Fig. 4.S1b). Of the 21,231 with at most one gene annotation, 12,720 (7,367 annotated) were at least 1kb long and eligible for calling change points with mountainClimber. Although we were only able to analyze ATSS and APA in 7,367 annotated genes which may be lower than expected, this is mainly due to closely spaced or overlapping genes and is not a reflection of our method's sensitivity.

Out of 12,720 total TUs, 11,993 were expressed highly enough to call change points and 9,283 had at least three segments in order to calculate RU of 5' and 3' ends (Section 3.3 Methods). Overall, DistalPolyA and TandemAPA predictions most often overlapped annotated 3'UTRs and DistalPolyA sites, while DistalTSS and TandemTSS predictions most often overlapped annotated DistalTSS and 5'UTRs, as expected (Fig. 4.S2a,b). DistalTSS and DistalPolyA were often predicted within 10bp of an annotated TSS and poly(A) sites respectively, but many were within annotated UTRs or intergenic (Fig. 4.S2c) and may indicate ATSS within closely spaced promoters[105] or 3' transcriptional readthrough due to incomplete selection of polyadenylated RNA. Many TUs utilized TandemAPA and TandemTSS as well, primarily in

3'UTRs and 5'UTRs respectively. Additionally, some poly(A) sites and TSSs were in coding exons and introns, which has been reported but is rarely observed [9,10,96]. Overall, most TUs utilize one TSS, but many use multiple APA sites (Fig. 4.1a).

4.4.2. Tissue type drives the observed variation in 5' and 3' end length

To systematically compare 5' and 3' ends across all samples, we summarized the change point information in each tandem 5' and 3' end in each sample by calculating the weighted mean extension length (WMEL), weighting by the relative usage (RU) in each sample (Section 4.3 Methods and Fig. 4.S3a,b). It should be noted that other methods that require two conditions defined a priori to identify APA, such as DaPars [69], would not support this type of single-sample analysis. Because the WMEL is not interpretable in ATE ends, 5' and 3' ends were separated into 4 groups based on their ATE status across all 2,342 samples: [1] ATE within a single sample, [2] ATE in different samples, [3] the last segment overlaps an annotated intron, [4] tandem and distal ends (Section 4.3 Methods and Fig. 4.S3a,b,c). Most TUs were in group #4 in the vast majority of samples (Fig. 4.1b, Fig. 4.S3d), consistent with previous observations (reviewed in [9,10]). In total, 4,655 and 4,731 TUs were in group #4 for the 5' and 3' UTRs respectively. Interestingly, there are more ATE across samples at the 5' end than the 3' end, suggesting that ATSS influences exon inclusion more often than APA (Fig. 4.1b and Fig. 4.S3e). Because group #4 was most prevalent and WMEL is easily interpretable in this case, we focused on tandem and distal ends for the remaining analysis.

To first check whether variation in WMEL is driven by tissue or individual, we used a linear mixed model similar to the approach used in [103]. We focused this analysis on ubiquitously expressed TUs in group #4, as it was previously shown that APA occurs more often in ubiquitously expressed genes than in tissue-specifically expressed genes [66]. Ubiquitously expressed TUs (group #4 in ≥ 90% of samples) with strand successfully inferred were considered, resulting in

420 TUs at the 5' end and 516 at the 3' end. Overall, tissue was the driver of WMEL variation for both 5' and 3' ends for the vast majority of TUs (Fig. 4.1c,d), affirming previous observations that APA and ATSS are tissue-specific [9,106]. Interestingly, subject was the driver for a small subset of genes including HLA-C, which is known to contain many polymorphisms.

Consistent with the linear modeling results, clustering of WMEL in ubiquitously expressed TUs showed primary clustering by tissue for both 5' and 3' ends (Fig. 4.S4a,b). Brain, Heart, Muscle-Skeletal, Whole-Blood, and Testis have especially tissue-specific 5' and 3' end lengths, while other tissues have more similar 5' and 3' WMELs. To exclude the possibility that covariates such as age, sex, ethnicity, batch, and sample quality contribute to these patterns, we regressed out 14 covariates and found that samples still largely clustered the same way (Fig. 4.S4c,d and Section 4.3 Methods). Because covariate regression did not appreciably change the clustering pattern, and 5' and 3' end lengths are not interpretable after covariate correction, all downstream analyses were conducted without regressing out covariates.

4.4.3. Global patterns of 5' and 3' end lengths across tissues

Given that tissue specificity is driving the observed variation, we sought to identify tissue-specific patterns of 5' and 3' WMEL. On a global scale, it was previously shown that 3'UTRs are shorter in cancer, proliferating cells, testis, blood, ovaries, placenta, and early developmental stages, while longer 3'UTRs are utilized in neuronal cells and in later developmental stages [10,69]. Consistent with these observations, Brain samples were longest, and Testis and Whole-Blood 3'UTRs were among the shortest (Fig. 4.2a). Additionally, we make several novel observations. First, Brain-Cerebellum and Brain-Cerebellar-Hemisphere 3' ends were the longest among all tissues. Second, although Testis had among the shortest 3' ends, it had the longest 5' ends (Fig. 4.2b). Third, Brain samples had the most variable 5' and 3' lengths across individuals (Fig. 4.S5a,b). Neither median TIN nor RIN scores correlated with the global trends in 5' and 3' lengths

across tissues, suggesting that the observed differences across tissues are biological rather than technical artifact (Fig. 4.S5c,d).

Because mountainClimberCP identifies change points in single samples, the degree of 5' and 3' complexity in an individual sample can be interrogated. Complexity was defined as the average number of 5' or 3' ends per TU with at least 10% relative usage across individuals in each tissue. Across all tissues, the 3' end was more complex than the 5' end (Fig. 4.S5e,f). Testis and cerebellum had the most complex 5' ends. Interestingly, genes with an average of 1.5+ 5' ends in Testis and Brain-Cerebellum were enriched for similar GO terms, including ubiquitin protein ligase, intracellular protein transport, and GTPase-related gene sets (Table S4.2). On the other hand, protein dephosphorylation was specific to Brain-Cerebellum, while transcription- and splicing-related gene sets were specific to Testis. Whole-Blood had the most complex 3' ends, and genes with 2+ change points were enriched for proliferation and viral-related GO terms (Table S4.3). Because the data is bulk RNA-Seq, it should be noted that 5' and 3' end complexity likely reflects the cell type heterogeneity in each tissue.

After identifying global patterns of 5' and 3' end length differences, we next aimed to compare TU lengths across tissues. We restricted our analysis to 301 and 321 ubiquitously expressed TUs with highly variable lengths across tissues in the 5' and 3' ends respectively (Section 4.3 Methods). Strikingly, the majority of TU's with variable 5' ends had significantly different lengths in testis compared to other tissues (Fig. 4.S6a), while five tissues stood out at the 3' end: Whole-Blood, Muscle-Skeletal, Testis, Brain-Cerebellum, and Brain-Cerebellar-Hemisphere (Fig. 4.S6b). Still, there were many TUs that had different lengths in only a small number of pairwise tissue comparisons (Fig. 4.S6c,d).

WMEL correlation of highly variable 5' and 3' ends across tissues recapitulated the findings that testis stands out at the 5' end, and the five tissues mentioned above stood out at the 3' end (Section 5.3 Methods and Fig. 4.2c,d). Interestingly, brain regions clustered together, with

the exception of cerebellar regions, in both 5' and 3' ends. The cerebellum was also distinct from other brain regions in previous GTEx studies of RNA editing [107] and gene expression [103]. On the other hand, the patterns of tissue similarity in the 5' and 3' ends differ in some tissues such as Testis, which stands alone at the 5' end but clusters with Muscle-Skeletal and Whole-Blood at the 3' end. Additionally, Prostate and Thyroid cluster together at the 5' end but not at the 3' end. These results suggest that some tissues have similar transcriptional mechanisms (reflected by their 5' end similarity) and simultaneously different post-transcriptional control mechanisms (reflected by their 3' end dissimilarity).

4.4.4. Alternative transcription start sites and polyadenylation sites across tissues

Since there are tissue-specific patterns of 5' and 3' end length, we next used mountainClimberTest to identify significantly differential ATSS and APA sites in all pairwise tissues (Section 4.3 Methods). To focus on events driven by tissue specificity, change points were required to be reproducible (i.e. clustered by mountainClimberTest) in at least 50% of individuals. On average, change points were reproducible in at least 50% of individuals in 1,420 / 2,597 (55%) total TUs in each pairwise comparison. This suggests that in addition to tissue-specific ATSS and APA, there is a substantial amount of individual variability. For the purpose of this study, we focused on tissue-specific events.

In a given pairwise tissue comparison, up to 266 differential change points were identified (Fig. 4.S7a,b and Fig. 4.3a,b). Of these TUs, 1,707 contained at least 3 segments and were tested for differential change points, corresponding to 1,069 and 618 tested APA and ATSS events respectively. 745 / 1,069 (70%) and 347 / 618 (56%) of TUs had significantly differential APA and ATSS respectively (BH-corrected p-value ≤ 0.05 and absolute RU difference ≥ 0.05), consistent

with the previous observations that 70% of mammalian genes utilize APA and 40-50% utilize ATSS [9,10,60,62,108].

To characterize ATSS and APA sites, we focused on the five main tissues of interest identified previously: Brain-Cerebellar-Hemisphere, Brain-Cerebellum, Muscle-Skeletal, Testis, and Whole-Blood. Most of the change points identified in these tissues vs. all 35 other tissues were tandem change points (Fig. 4.S7c). Significantly differential tandem change points were most often APA events in 3'UTRs, with the exception of Testis which additionally had many differential ATSS change points (Fig. 4.3c and Fig. 4.S7d). APA in Samd4a illustrates a typical example of APA in a 3'UTR, where an annotated proximal poly(A) site is utilized in Testis, while the distal poly(A) site is utilized in Heart-Atrial-Appendage (Fig. 4.3d). A typical ATSS example is illustrated by Cpne5 in Brain-Cortex vs. Heart-Atrial-Appendage (Fig. 4.3e). Both of these examples are known events annotated in Ensembl, supporting the validity of our approach.

Recently, intronic polyadenylation was reported to be more widespread than previously appreciated [96], so the ability to identify it from RNA-seq is an important advantage our approach has compared to other approaches. A minority of significantly differential tandem change points overlapped annotated introns (Fig. 4.S7e). Because this RNA-Seq is derived from bulk tissue, intronic APA in more distinct cell types within a tissue may not be detectable. Still, some examples of significant tissue-specific APA were detectable, such as Spdl1 in Brain-Frontal-Cortex-BA9 vs. Brain-Spinal-cord-cervical-c-1 (Fig. 4.3g). While both tissues express the intronic APA isoform, it is the primary expressed isoform in Brain-Spinal-cord-cervical-c-1, while Brain-Frontal-Cortex-BA9 additionally expresses the isoform skipping this exon.

Another advantage of our approach is the ability to identify more than two APA sites. Out of 67,789 total 5' and 3' ends with significantly differential change points in union across all pairwise tissue comparisons, 4,435 (7%) had two significantly differential change points and 88 (0.1%) had 3 significantly differential change points (illustrated for the 5 tissues of interest in Fig.

4.S7f). Ube2j1 is one example with three differential change points, where the distal poly(A) sites are preferentially used in Breast-Mammary-Tissue while the proximal-most poly(A) site is preferred in Testis (Fig. 4.3f).

To test whether APA or ATSS affects biological functions, we performed Gene Ontology (GO) analysis on genes with APA or ATSS. APA- and ATSS-containing genes were defined as those with significantly differential WMEL by t-test in a pairwise tissue comparison. With this definition, as opposed to a mountainClimberTest-based definition, gene sets could be grouped by whether their WMEL was longer or shorter. The five tissues with the most differential WMELs described above were considered (Fig. 4.S6a,b, Table S4.5-7, and Sections 4.3 Methods). Membrane-related genes had shorter 3' ends in Muscle-Skeletal compared to other tissues (Table S4.6). Recently, it was shown that the 3'UTR can affect the localization of membrane proteins, e.g. CD47 from the endoplasmic reticulum membrane to the cell membrane [94]. Thus, it is possible that genes with shorter 3' ends in skeletal muscle have different cellular localization compared to other tissues. Nucleotide-excision repair was also enriched in genes with longer 5' ends in testis, which may be related to DNA damage repair in spermatogenesis, where reactive oxygen species, sperm chromatin packaging, and apoptosis were proposed to contribute to infertility (reviewed in [109]). Transcription-related terms were enriched for genes that had shorter 3' ends and longer 5' ends in Testis compared to other tissues (Tables 4.S6,S7). This suggests that 5' and 3' end usage may be correlated for subsets of genes.

4.5. Discussion

Here, we applied mountainClimber, a novel de novo approach for identifying alternative 5' and 3' ends from RNA-Seq, to 2,342 human tissue samples from the GTEx consortium. 56% and 70% of tested TUs were regulated by ATSS and APA respectively, consistent with previous reports that 50% and 70% of mammalian genes are regulated by ATSS and APA [9,10,62]. Consistent with

our findings, ATSS and APA were recently shown to drive isoform differences across human tissues [98]. The study presented here enhances these findings through application of a novel approach, mountainClimber, and novel observations summarized below.

Most strikingly, Testis, Brain-Cerebellum, Brain-Cerebellar-Hemisphere, Muscle-Skeletal, and Whole-Blood had significantly different 3' ends compared to all other tissues. This suggests that alternative 3' ends are acting on different sets of genes to contribute to tissue specificity. This notion was supported by GO analysis, which revealed different gene sets enriched in different tissues.

On the other hand, Testis had the most distinct 5' ends compared to all other tissues. In contrast to APA, which has been extensively studied, ATSS identification in RNA-Seq is less well studied. This may be due to the difficulty in change point detection at 5' ends in RNA-Seq, which mountainClimberCP improves (Fig. 3.2e,f). In particular, Testis' distinction from all other tissues is a phenomenon that was not previously appreciated. Interestingly, GO analysis of genes with different lengths in testis compared to other tissues revealed similar gene sets enriched at both 5' and 3' ends, suggesting some level of coordination between 5' and 3' ends in testis. While coordination among 5' and 3' ends was reported in the Drosophila nervous system [110] and MCF-7 breast cancer cells [91], it was not previously reported in human testis. As APA often affects stability and localization of RNA while ATSS often affects translation, this coordination may have a significant impact on gene regulation and will be interesting to pursue in the future.

Additionally, some tissues exhibited similar 5' end lengths and simultaneously dissimilar 3' end lengths, suggesting that some tissues utilize similar transcriptional mechanisms and simultaneously different post-transcriptional control mechanisms. Similar transcriptional mechanisms may possibly be driven by similar transcription factor expression while dissimilar 3' processing may be driven by differential expression of cleavage and polyadenylation factors or differential stability of alternative 3' isoforms.

In this study, tissue type drove the observed variation in 5' and 3' end length in ubiquitously expressed genes. Still, individual variation is a significant factor. This was apparent in mountainClimberTest analysis, where requiring that the change point was observed in 50% of individuals led to exclusion of 55% of TUs. While the scope of this study includes tissue specificity, it will be interesting to investigate individual variability in APA and ATSS in the future.

4.6. Figures

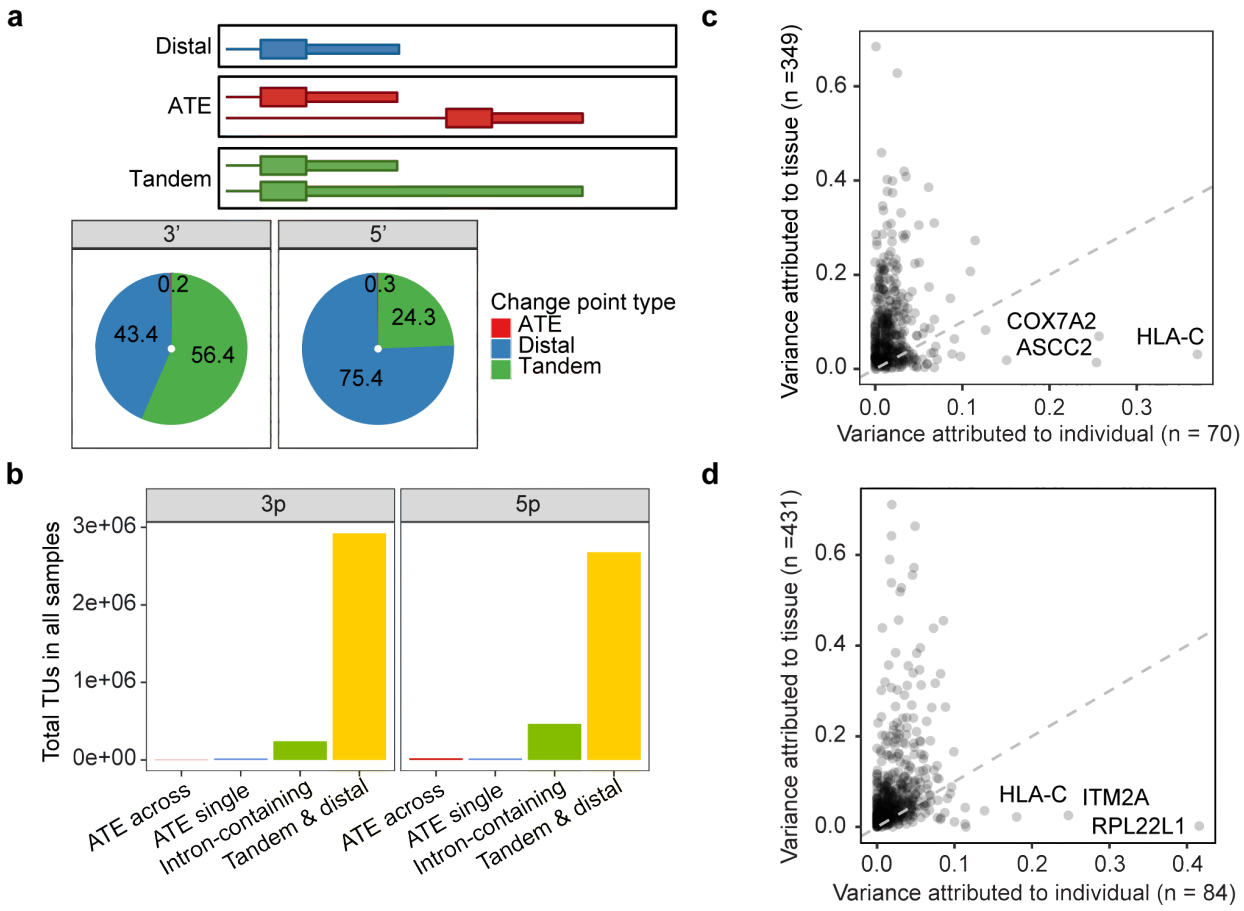


Figure 4.1 | Identification of alternative 5' and 3' ends in human tissues.
(**a**) Proportion of each type of change point with maximum relative usage (distal, alternative terminal exon (ATE), or tandem) at each end. (**b**) Total genes summed over all samples in each of the 4 categories (Section 4.3 Methods and Fig. 4.S3). (**c, d**) Contribution of tissue and individual to length variation for 5' ends (**c**) and 3' ends (**d**) in ubiquitously expressed TUs with tandem ends. The numbers in the axes labels indicate the number of genes for which the tissue explained more variation than the individual (y-axis) or vice versa (x-axis).

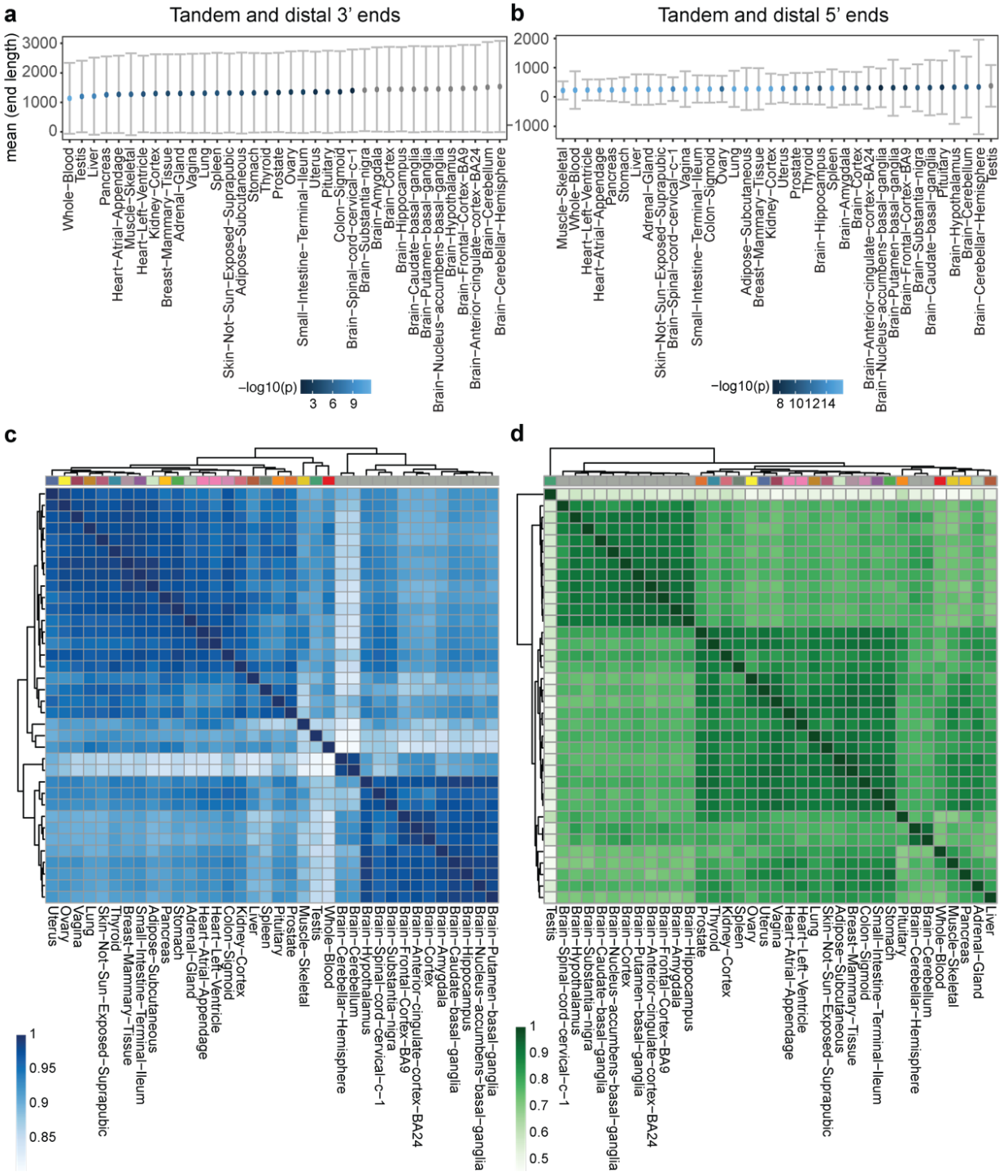


Figure 4.2 | Global trends of tandem 5' and 3' end lengths across human tissues.
(**a**) Mean weighted mean extension length (WMEL) across all subjects per TU and all TUs per tissue for 2,274 3' ends that were tandem in ≥ 12 tissues and ≥ 71 subjects. Color indicates -log10(KS test p-value) (grey indicates p >= 0.05) of mean WMELs of each tissue vs. Brain-Cerebellar-Hemisphere. (**b**) Similar to (**a**), but for 2,204 5' ends. Top panel, KS test p-value of each tissue vs. Testis. (**c, d**) Hierarchical clustering of Spearman correlation r^2 of WMEL for 321 genes with highly variable 3' WMELs (**c**) and 301 highly variable 5' WMELs (**d**).

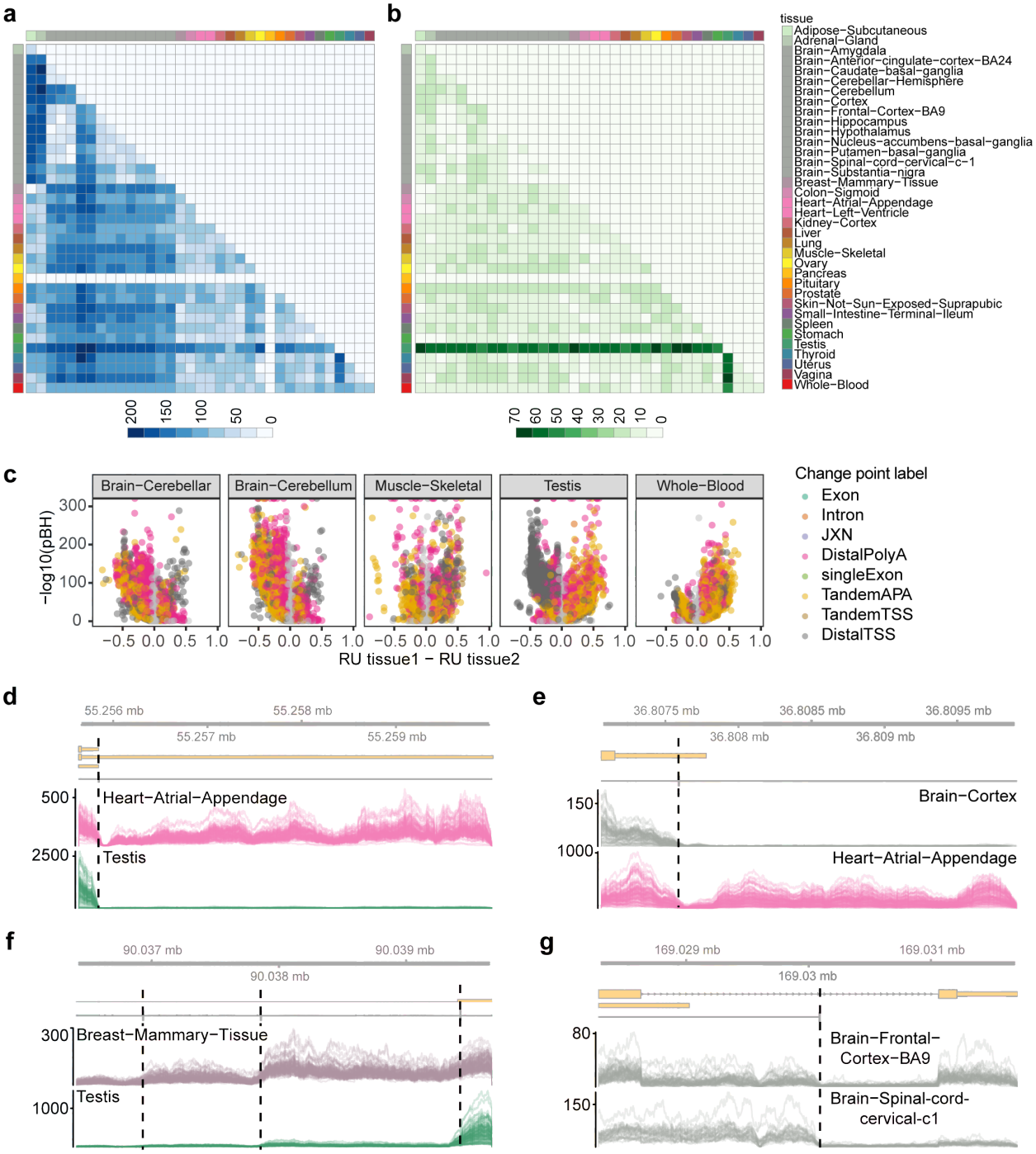


Figure 4.3 | Differential ATSS and APA sites across human tissues.
(**a,b**) Total significantly differential change points identified by mountainClimberTest in each pairwise comparison in the 3' (**a**) and 5' (**b**) ends. (**c**) Volcano plot of mountainClimberTest_diff results where the x-axis indicates RU difference in tissue1 (the tissue of interest indicated in each panel) vs. tissue2 (all 35 other tissues). (**d-g**) Examples of alternative 5' and 3' ends. Each read coverage line indicates one individual, and change points are indicated by black dashed lines. (**d**) APA in Samd4a in Testis vs. Heart-Atrial-Appendage (p = 2e-314, RU difference = -0.51). (**e**) ATSS in Cpne5 in Brain-Cortex vs. Heart-Atrial-Appendage (p = 7.49e-236, RU difference = 0.59).

(**f**) Three APA change points in Ube2j1 in Breast-Mammary-Tissue vs. Testis. (**g**) Intronic APA in Spdl1 in Brain-Frontal-Cortex-BA9 vs. Brain-Spinal-cord-cervical-c-1 (p = 4.72e-43, RU difference = -0.39).

4.7. Supplementary Figures

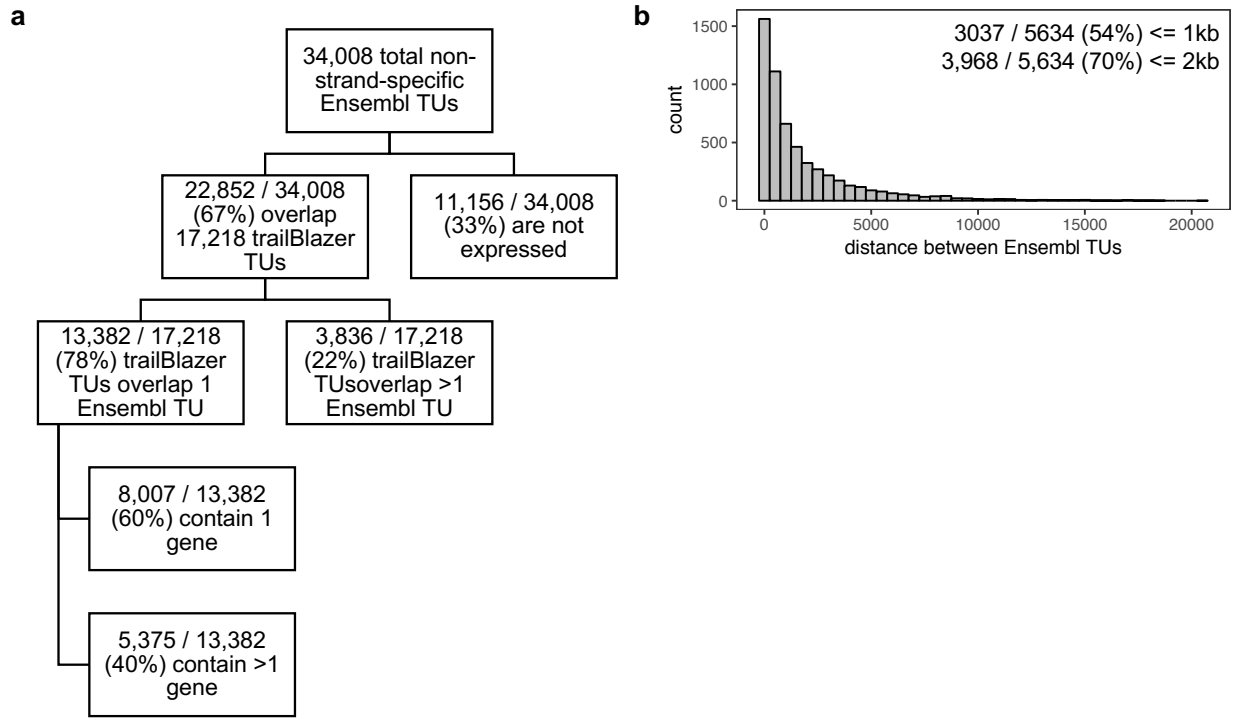


Figure 4.S1 | Overlap between mountainClimberTU and Ensembl transcription units.
(**a**) Flow chart indicating the overlap of Ensembl TUs with mountainClimberTU TUs (Section 4.3 Methods). (**b**) Histogram of 5,634 pairwise distances among the 3,836 mountainClimberTU TUs that overlapped more than one Ensembl TU. The inset text describes the total within 1 or 2kb.

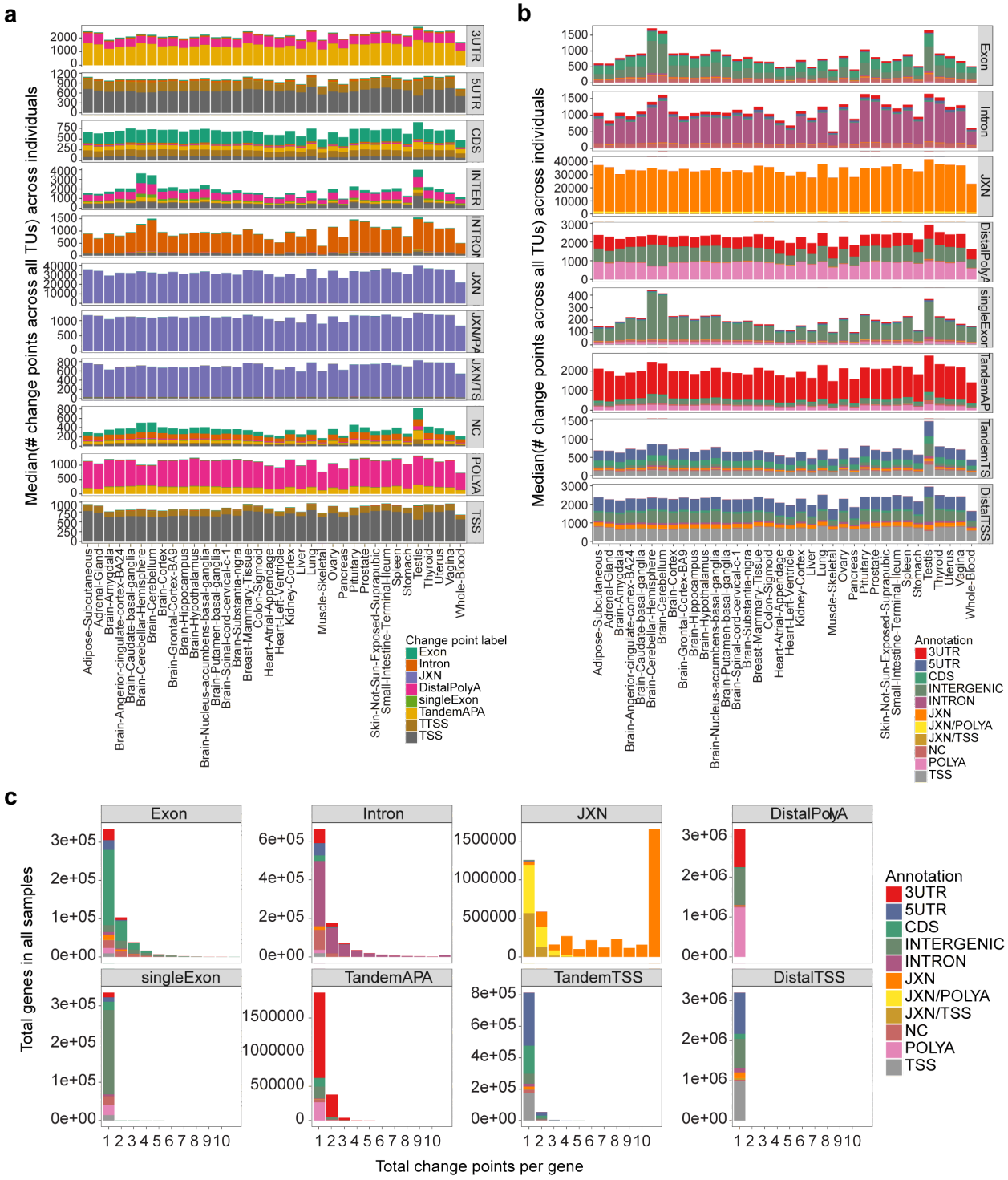


Figure 4.S2 | Identification of alternative 5' and 3' ends in human tissues.
(**a, b**) Median total number of change points per TU in individuals with each change point label (colored in (**a**), rows in (**b**) and overlapping Ensembl 75 gene regions (rows in (**a**), colored in (**b**)). (**c**) The total change points per TU in all individuals stratified by each change point label (columns) and colored by Ensembl 75 gene region.

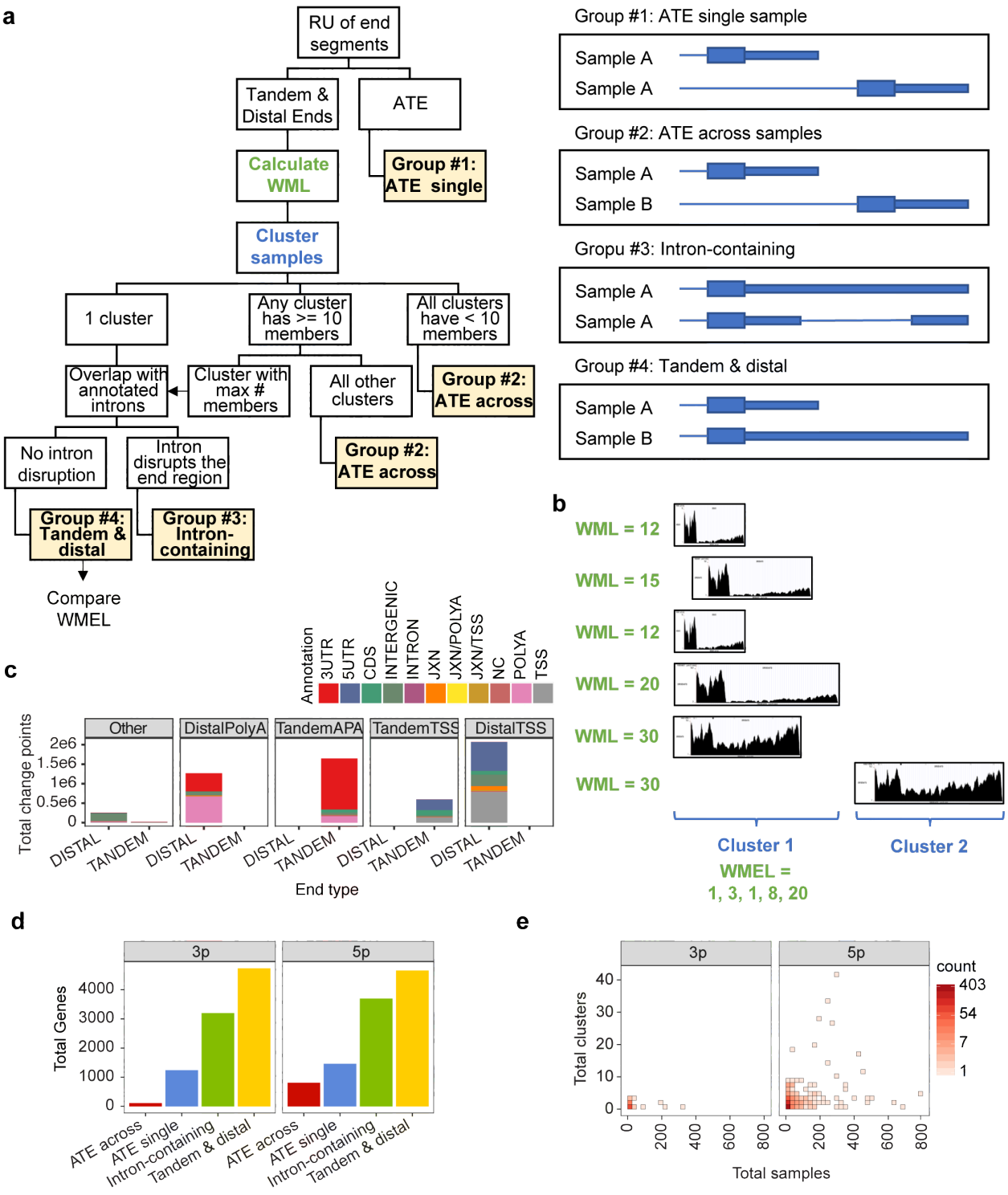


Figure 4.S3 | Grouping of 5' and 3' ends into four categories.
(**a**) Schematic outlining the grouping into 4 categories (left) with cartoon examples (right). ATE, alternative terminal exon (Section 4.3 Methods). (**b**) Toy example illustrating clustering, weighted mean length (WML), and weighted mean extension length (WMEL) calculations. In this case, there is ATE because there is >1 cluster, but since Cluster 2 contains < 10 members, we ignore it and only consider Cluster 1 (Section 4.3 Methods). (**c**) End type and change point type identified

by mountainClimberCP (DistalPolyA, TandemAPA, TandemTSS, DistalTSS, or otherwise Other) of Group #4: tandem genes for the end with maximum relative usage (RU), colored by overlap with Ensembl 75. (**d**) Total genes in each group. Note that groups are not mutually exclusive - a gene is often in a different group in a different sample. (**e**) Two-dimensional histogram of total individuals vs. total clusters in Group #2: ATE across samples, excluding change points in non-strand-specific TUs.

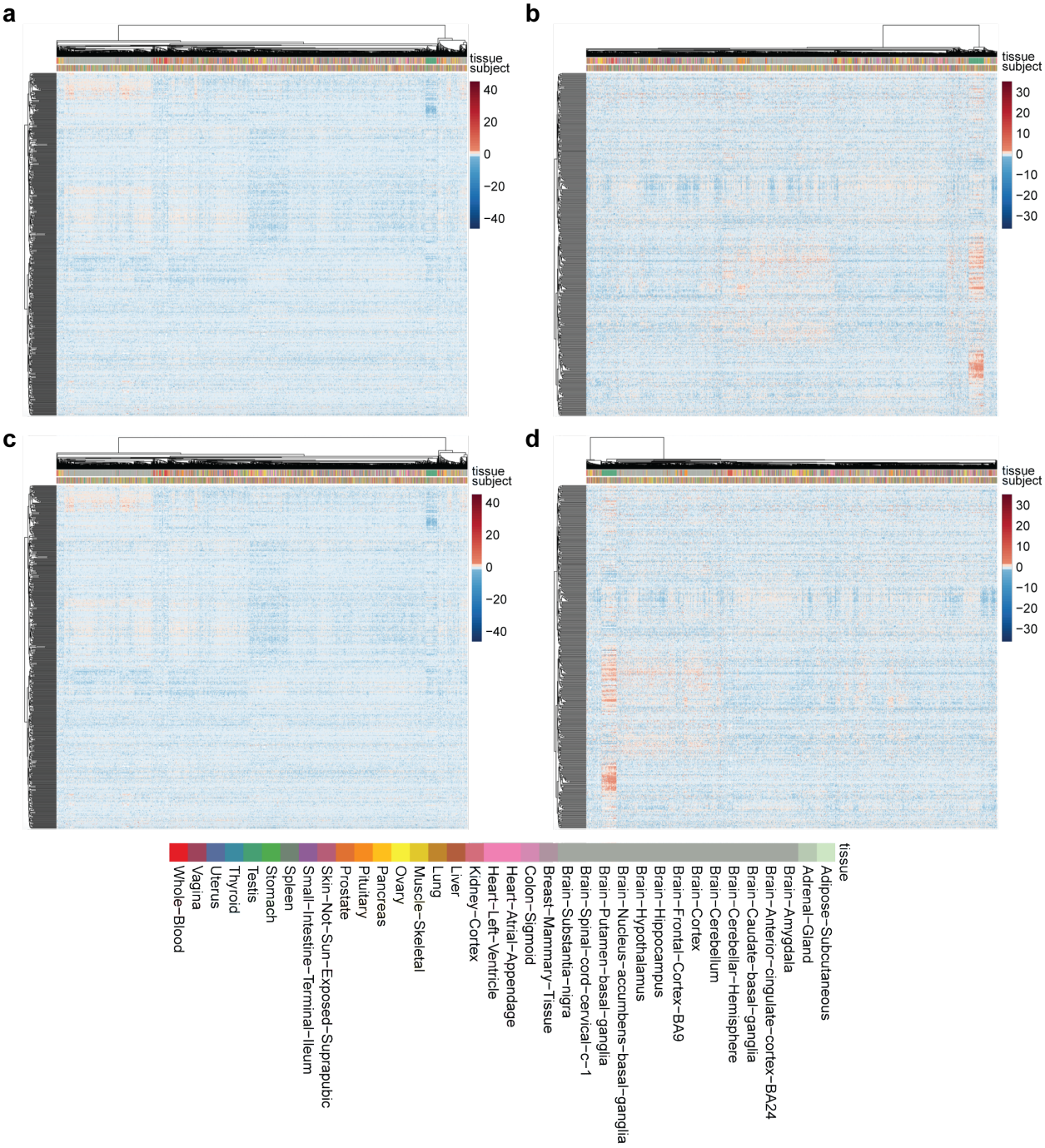


Figure 4.S4 | Tissue specificity of 5’ and 3’ ends across human tissues.
(**a, b**) Heatmap of row-scaled log2(relative weighted mean extension length (WMEL)) for all 516 ubiquitously expressed 3’ ends (**a**) and 420 5’ ends (**b**). Column colors indicate tissue and individual. (**c, d**) Similar to (**a**) and (**b**), but with covariates regressed out.

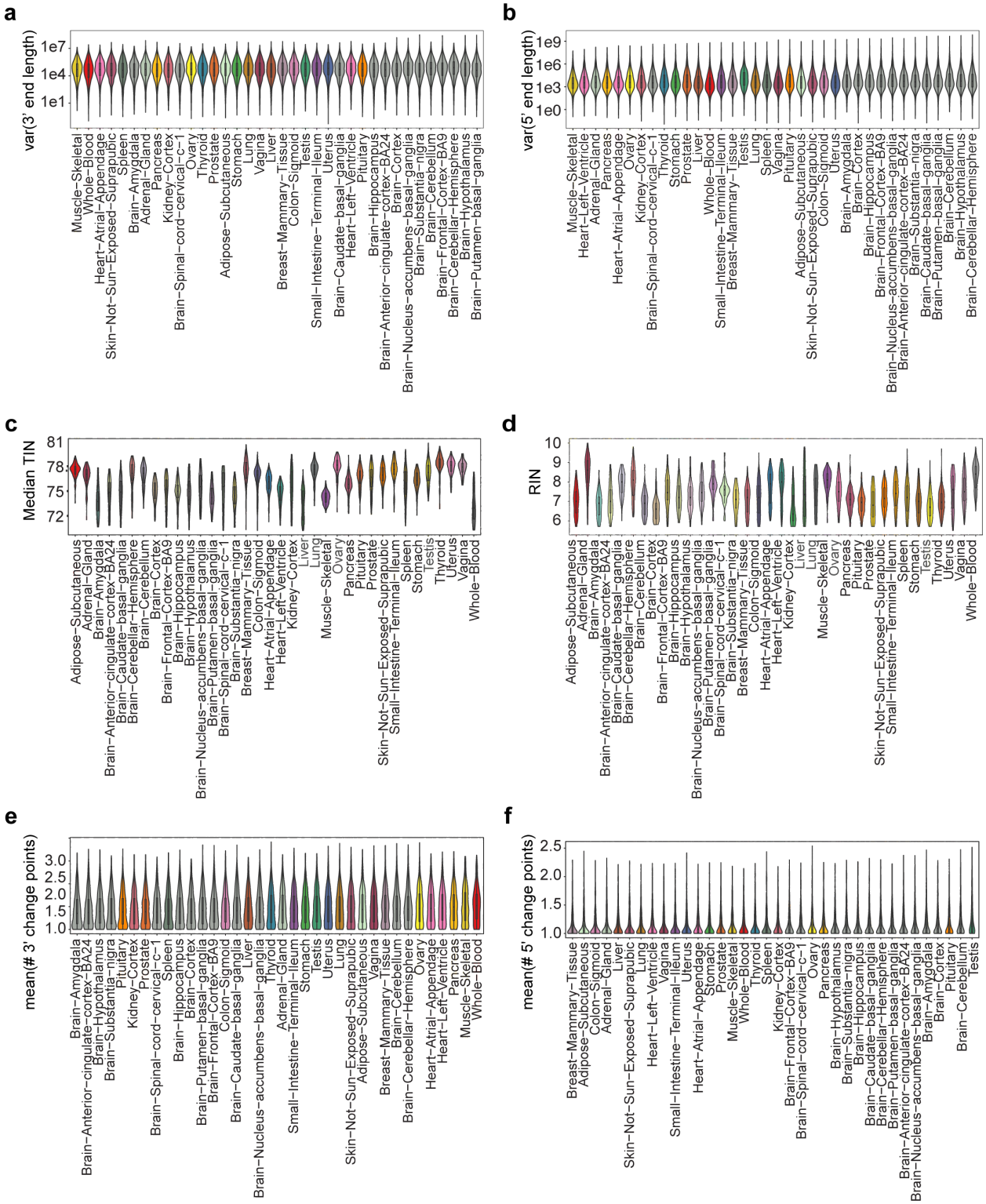


Figure 4.S5 | Variation in 5' and 3' ends across human tissues.
(**a, b**) Variance of relative weighted mean length of all 2,274 3' (**a**) and 2,204 5' (**b**) ends for each TU across individuals within each tissue, sorted from low to high variance. (**c, d**) median TIN score across all transcripts (**c**) and RIN score (**d**) for each sample in each tissue. Only the 2,342

samples analyzed are shown, so all have median TIN ≥ 70 in (**c**). (**e, f**) Mean number of change points with at least 10% relative usage in 3' (**e**) and 5' (**f**) ends for each TU across individuals within each tissue, sorted from low to high mean number of change points.

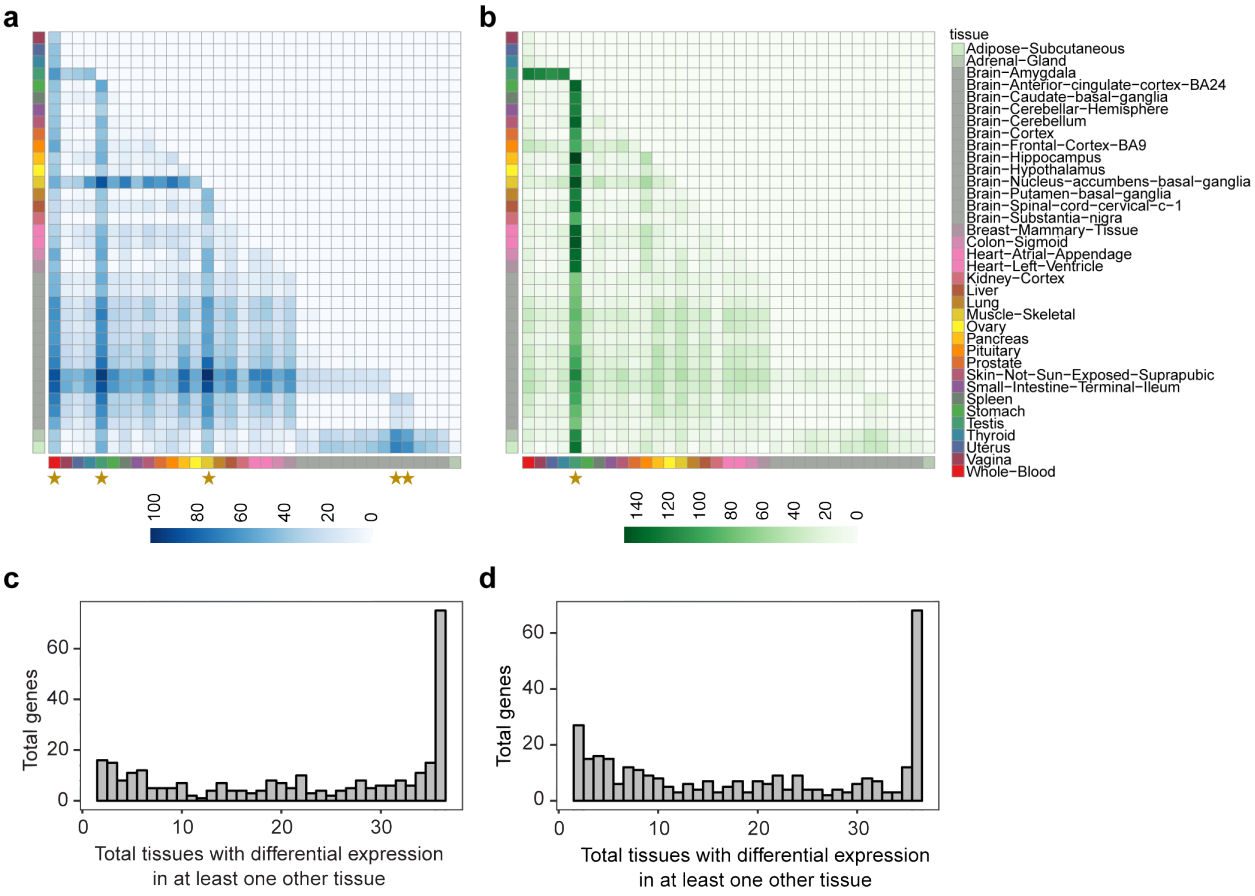


Figure 4.S6 | Identification of TUs with highly variable 5' or 3' end lengths.
(**a, b**) Total genes with differential 3' (**a**) and 5' (**b**) WMEL by t-test. Yellow stars indicate the tissues that were most different compared to all other tissues. (**c, d**) Total tissues with differential WMEL in at least one pairwise comparison for 3' ends (**c**) and 5' ends (**d**).

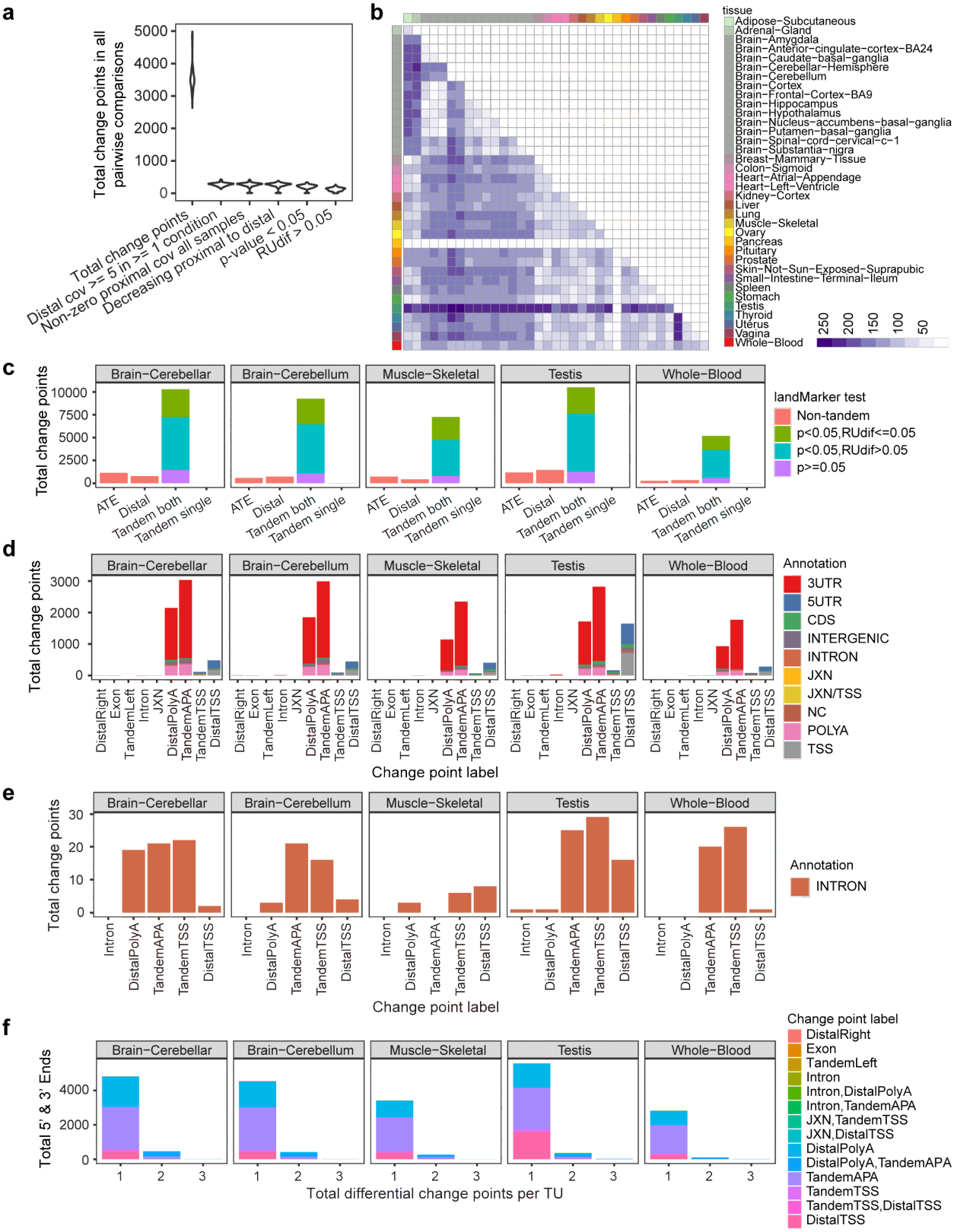


Figure 4.S7 | Significantly differential change points identified by mountainClimberTest.

(**a**) Total change points identified at each step of mountainClimberTest_diff over all pairwise tissue comparisons (see Section 4.3 Methods for details on these steps). (**b**) Total change points identified in either 5' or 3' ends, i.e. Fig. 4.3a and 4.3b combined. (**c**) Types of change points identified by mountainClimberTest_diff across all 35 comparisons for each tissue colored by whether they were significant (see mountainClimberTest methods in Section 3.3 for more details on the x-axis labels). (**d, e**) Annotation of all significantly differential change points across all 35 comparisons that were tandem in at least one condition (**d**) and those in introns only (**e**). (**f**) Total significantly differential change points in each TU vs. all 35 other tissues.

4.8. Supplementary Tables

Table 4.S1 | GTEx Sample Totals.

Tissue	Total samples downloaded	Filtered samples			Total samples analyzed
		Low median TIN	Read length != 76	Mapping rate < 50%	
Adipose - Subcutaneous	99	3	6	5	85
Adrenal gland	78	1			77
Brain - Amygdala	44	6		6	32
Brain - Anterior cingulate cortex (BA24)	54	9		4	41
Brain - Caudate (basal ganglia)	67	3		5	59
Brain - Cerebellar Hemisphere	67	1	3	2	61
Brain - Cerebellum	81		9	2	70
Brain - Cortex	70	4	4	1	61
Brain - Frontal Cortex (BA9)	64	7	3	2	52
Brain - Hippocampus	59	13		4	42
Brain - Hypothalamus	56	8		5	43
Brain - Nucleus accumbens (basal ganglia)	65	8		3	54
Brain - Putamen (basal ganglia)	60	14		5	41
Brain - Spinal cord (cervical c-1)	37	5		3	29
Brain - Substantia nigra	41	7		3	31
Breast - Mammary Tissue	104	1			103
Colon - Sigmoid	105	2		1	102
Heart - Atrial Appendage	99	1		2	96
Heart - Left Ventricle	94	22	3	5	64
Kidney - Cortex	36	2			34
Liver	64	14			50
Lung	70	2	5	3	60
Muscle - Skeletal	97	2		3	92
Ovary	63				63
Pancreas	107	2	4		101
Pituitary	71	1		2	68
Prostate	59	1		1	57
Skin - Not Sun Exposed (Suprapubic)	99	3			96
Small Intestine - Terminal Ileum	69	1			68
Spleen	67	3			64
Stomach	108				108
Testis	95		4		91
Thyroid	99		6	7	86
Uterus	55				55
Vagina	54				54
Whole Blood	99	39	6	2	52

Table 4.S2 | Gene Onotology analysis of complex 5' ends.

GO term	pBF	Total genes	Gene list	Com-parison
positive regulation of NF-kappaB transcription factor activity	<1e-04	4	CAT,MAP3K7,NFKBIA,PRDX3	Testis longer
negative regulation of transcription, DNA-dependent	<1e-04	4	CITED2,COMMD7,NR2C1,ZNF282	Testis longer
positive regulation of transcription from RNA polymerase II promoter	<1e-04	11	CALCOCO1,CDK7,CITED2,ERCC3,FOS,HSF2,MED10,NFKBIA,PARP1,STAT3,TFE3	Testis longer
endoplasmic reticulum	<1e-04	16	APOD,ATL2,CAMLG,CAT,CYB5A,EMC2,FOS,PDIA4,RCN2,RRBP1,SRP72,STT3B,TRAPPC3,UBXN4,VIMP,YIPF6	Testis longer
protein C-terminus binding	<1e-04	6	CALCOCO1,CDK7,ERCC3,IFT52,KPNA3,PRDX3	Testis longer
transcription initiation from RNA polymerase II promoter	<1e-04	6	CDK7,ERCC3,GTF2B,MED10,NR2C1,PARP1	Testis longer
DNA replication	<1e-04	4	MCM3,ORC5,RPA2,WRNIP1	Testis longer
molecular function	<1e-04	8	DHRS7B,MKRN1,MRPS10,ORC5,PTTG1IP,TMEM128,TMEM57,VPS45	Testis longer
nucleotide-excision repair	<1e-04	4	CDK7,ERCC3,RAD23B,RPA2	Testis longer
biological process	<1e-04	11	BCL7B,DHRS7B,FAM3C,MKRN1,MRPS10,PROSC,SDC4,SGTA,TMEM128,TMEM57,ZNF330	Testis longer
transforming growth factor beta receptor signaling pathway	<1e-04	4	CITED2,FOS,MAP3K7,PARP1	Testis longer
cellular component	<1e-04	7	BCL7B,COMMD7,DHRS7B,MKRN1,PIP4K2A,TMEM128,VPS45	Testis longer
transcription regulatory region DNA binding	<1e-04	4	CALCOCO1,FOS,STAT3,TFE3	Testis longer
aging	<1e-04	4	APOD,CAT,FOS,IGFBP2	Testis longer
nucleobase-containing small molecule metabolic process	<1e-04	4	ATIC,CAT,CMPK1,NUDT9	Testis longer
endocytosis	<1e-04	4	LMBR1L,NECAP1,RAC1,USP33	Testis longer
protein serine/threonine kinase activity	<1e-04	5	CDK7,CSNK2A1,MAP3K7,OXSR1,STK40	Testis longer
transcription elongation from RNA polymerase II promoter	<1e-04	4	CDK7,ELP3,ERCC3,GTF2B	Testis longer
protein phosphorylation	<1e-04	6	CDK7,CSNK2A1,ERCC3,MAP3K7,OXSR1,STK40	Testis longer
positive regulation of type I interferon production	<1e-04	4	DHX36,DHX9,NFKBIA,POLR3E	Testis longer
modulation by virus of host morphology or physiology	<1e-04	9	CAMLG,ERCC3,GTF2B,KPNA3,NFKBIA,SGTA,SLC25A4,STAT3,YWHAE	Testis longer
protein kinase activity	<1e-04	6	CDK7,CSNK2A1,ERCC3,MAP3K7,OXSR1,STK40	Testis longer

Fc-epsilon receptor signaling pathway	<1e-04	4	FOS,MAP3K7,NFKBIA,RAC1	Testis longer
nucleotide-excision repair, DNA damage removal	<1e-04	4	CDK7,ERCC3,RAD23B,RPA2	Testis longer
cytoplasmic vesicle	<1e-04	4	BICD2,IGFBP2,RAB13,RAC1	Testis longer
intracellular	0.0418	14	APOD,BIRC2,CAPN2,IFT27,NUDT9,PICALM,PROSC,RAB13,RAC1,RPL19,RPL22L1,RPS19,UBL3,ZNF282	Testis longer
protein transport	<1e-04	5	AP2S1,RAB9A,SELENBP1,SNX12,TIMM50	Testis shorter
microtubule	<1e-04	4	CCT2,DYNC1LI1,MAP2K2,SPRY2	Testis shorter
plasma membrane	0.0187	10	ADIPOR1,ANXA1,AP2S1,DYNC1LI1,HLA-DRA,MGLL,PKM,RAB9A,SLC2A3,SPRY2	Testis shorter

GO Term, Gene Ontology Term; pBF, Bonferroni-corrected p-value; Total genes, total genes from the test set in the enriched term; Gene list, genes from the test set in the enriched term; Comparison, indicates the tissue of interest and whether the genes in the test set were significantly longer or shorter in the tissue of interest compared to at least one other tissue.

Table 4.S3 | Gene Onotology analysis of complex 3' ends.

GO term	pBF	Total genes	Gene list	Comparison
ribosome	<1e-04	7	DNAJC21,MRPL22,MRPS10,MRPS21,RPL29,RPS14,RPS19	Brain-Cereb longer
ribosome	<1e-04	4	DNAJC21,MRPL22,RPL29,RPS14	Muscle-Skeletal shorter
ribosome	<1e-04	6	DNAJC21,MRPL22,MRPS10,MRPS21,RPS14,RPS19	Testis shorter
ribosome	<1e-04	4	DNAJC21,MRPS21,RPL29,RPS14	Whole-Blood shorter
condensed chromosome kinetochore	<1e-04	4	BOD1,BUB3,DYNC1LI1,PPP1CC	Brain-Cereb longer
condensed chromosome kinetochore	<1e-04	4	BOD1,BUB3,DYNC1LI1,PPP1CC	Testis shorter
mitosis	<1e-04	4	ARPP19,BOD1,DYNC1LI1,RAN	Brain-Cereb longer
mitosis	<1e-04	4	ARPP19,BOD1,DYNC1LI1,RAN	Testis shorter
protein kinase binding	<1e-04	5	CDK5RAP1,MAP2K3,PPP1CC,RAC1,RPS19	Testis shorter
protein kinase binding	<1e-04	4	MAP2K3,PPP1CC,RAC1,RPS19	Whole-Blood longer
RNA metabolic process	<1e-04	4	EIF4A3,PSMD12,RPL29,RPS14	Brain-Cereb vs Brain longer
RNA metabolic process	<1e-04	5	EIF4A3,LSM5,PSMD12,RPL29,RPS14	Whole-Blood shorter
small molecule metabolic process	<1e-04	25	ACSS1,AKR1B1,ALAD,AMD1,B4GALT5,CERK,CMPK1,COX5A,COX7B,CYB5A,FH,GLUL,HADH,LPCAT4,MARCKS,MGLL,NDUFB4,NUDT9,PPP1CC,PSMB2,PSMD12,RAN,SDC2,SLC25A4,SLC2A3	Brain-Cereb longer
small molecule metabolic process	<1e-04	16	AKR1B1,B4GALT5,COX7B,FH,GLUL,HADH,MGLL,PHYH,PPP1CC,PSMB2,PSMD12,RAN,SDC2,SDHB,SLC25A4,SUCLG1	Muscle-Skeletal longer
transcription, DNA-templated	<1e-04	15	ARGLU1,CALCOCO1,DNTTIP2,ETS2,HINFP,KLF9,MAP3K7,PARP1,PHF2,POLR3E,PWP1,SFPQ,SUV420H1,TXNIP,YBX3	Testis shorter
transcription, DNA-templated	<1e-04	7	ARGLU1,HTATSF1,MLXIP,PARP1,PHF2,SP1,TXNIP	Whole-Blood longer
viral process	<1e-04	8	GTF2B,PSMB2,PSMD12,RAC1,RAN,RPS14,RPS19,SLC25A4	Testis shorter
viral process	0.036	6	GTF2B,PSMD12,RAN,RPL29,RPS14,SLC25A4	Whole-Blood shorter
cell differentiation	<1e-04	4	FHL1,GADD45B,HTATIP2,NUS1	Brain-Cereb longer
centrosome	<1e-04	4	FBXW11,NFU1,PCGF5,PCM1	Muscle-Skeletal longer

cytosol	<1e-04	11	CTSH,FHL1,GLUL,LSM4,MAP2K3,PPP1CC,RAC1,RPL19,RPS19,TRAPPC3,TXNIP	Whole-Blood longer
endoplasmic reticulum membrane	<1e-04	11	CYB5A,LPCAT4,MTDH,NUS1,PJA2,SSR2,TMED9,TMEM106C,UBE2J1,VAPB,VMA21	Muscle-Skeletal shorter
enzyme binding	<1e-04	5	CTNNBL1,CYB5A,HINFP,PARP1,VAPB	Brain-Cereb vs Brain longer
histone deacetylase binding	<1e-04	4	GMNN,HES1,NR2C1,YWHAE	Testis longer
hydrolase activity	<1e-04	4	MGLL,NUDT9,PPP2CB,PSMC6	Testis longer
intracellular protein transport	0.0318	6	RAB13,RAN,RHOU,TIMM17A,VPS18,VPS45	Testis longer
mRNA metabolic process	<1e-04	4	EIF4A3,PSMD12,RPL29,RPS14	Brain-Cereb vs Brain longer
negative regulation of transcription from RNA polymerase II promoter	<1e-04	7	MTDH,NR2C1,PARP1,RPS14,SFPQ,TXNIP,YBX3	Testis shorter
nuclear membrane	<1e-04	4	AKIRIN1,MTDH,RANGAP1,SEPHS1	Muscle-Skeletal shorter
nucleolus	<1e-04	5	MTDH,PARP1,PHF2,PPP1CC,RPS19	Whole-Blood longer
oxidoreductase activity	<1e-04	5	AKR1B1,CREG1,CYB5R1,HADH,SDHB	Muscle-Skeletal longer
peptidase activity	<1e-04	4	CTSH,LGMN,PEPD,THOP1	Muscle-Skeletal shorter
protein polyubiquitination	<1e-04	4	FBXW11,PSMB2,PSMD12,UBE2V2	Muscle-Skeletal longer
protein transport	<1e-04	7	GOSR1,IFT27,NECAP1,RAB18,RAB9A,RAC1,RAN	Testis shorter
proteolysis	<1e-04	7	ABHD10,ADAMTS1,CTSH,CTSL,LGMN,PEPD,THOP1	Muscle-Skeletal shorter
regulation of transcription, DNA-dependent	<1e-04	11	ARGLU1,CLPX,DNTTIP2,ETS2,GTF2B,HINFP,HSF2,KLF9,NR2C1,SUV420H1,YBX3	Testis shorter
sequence-specific DNA binding	0.0382	4	CALCOCO1,ETS2,HSF2,NR2C1	Testis shorter
SRP-dependent cotranslational protein targeting to membrane	<1e-04	4	RPL29,RPS14,SRP9,SSR2	Muscle-Skeletal shorter
structural constituent of ribosome	<1e-04	4	MRPL15,MRPS21,RPL29,RPS14	Whole-Blood shorter
transcription coactivator activity	<1e-04	7	CALCOCO1,HINFP,HSF2,HTATIP2,MTDH,PHF2,RAN	Testis shorter
transcription factor complex	<1e-04	4	CREG1,LMO4,NFIC,SKI	Testis longer
zinc ion binding	<1e-04	11	ALAD,CRIP2,GTF2B,LMO4,MT1E,NR2C1,PJA2,SKI,ZMIZ2,ZMYM3,ZNF330	Testis longer

GO Term, Gene Ontology Term; pBF, Bonferroni-corrected p-value; Total genes, total genes from the test set in the enriched term; Gene list, genes from the test set in the enriched term; Comparison, indicates the tissue of interest and whether the genes in the test set were significantly longer or shorter in the tissue of interest compared to at least one other tissue.

Table 4.S4 | mountainClimberTest_cluster totals.

Tissue1	Tissue2	Total Samples	--minpts	Total Genes	Total genes clustered per condition	Total genes clustered across conditions	% Clustered / Total Genes
Adipose-Subcutaneous	All but Ovary, Uterus, Vagina	2,008	992	2,585	1,517	1,517	59%
Adipose-Subcutaneous	Ovary, Uterus, Vagina	2,040	1,008	2,585	1,517	1,517	59%
Adrenal-Gland	All but Ovary, Uterus, Vagina	1,806	891	2,260	1,421	1,421	63%
Adrenal-Gland	Ovary, Uterus, Vagina	1,848	912	2,260	1,419	1,419	63%
Brain-Amygdala	All but Ovary, Uterus, Vagina	756	378	2,199	1,358	1,358	62%
Brain-Amygdala	Ovary, Uterus, Vagina	768	384	2,199	1,358	1,358	62%
Brain-Anterior-cingulate-cortex-BA24	All but Ovary, Uterus, Vagina	972	474	2,406	1,432	1,432	60%
Brain-Anterior-cingulate-cortex-BA24	Ovary, Uterus, Vagina	984	480	2,406	1,432	1,432	60%
Brain-Caudate-basal-ganglia	All but Ovary, Uterus, Vagina	1,397	687	2,566	1,450	1,450	57%
Brain-Caudate-basal-ganglia	Ovary, Uterus, Vagina	1,416	696	2,566	1,450	1,450	57%
Brain-Cerebellar-Hemisphere	All but Ovary, Uterus, Vagina	1,445	711	2,986	1,442	1,442	48%
Brain-Cerebellar-Hemisphere	Ovary, Uterus, Vagina	1,464	720	2,986	1,442	1,442	48%
Brain-Cerebellum	All but Ovary, Uterus, Vagina	1,657	828	2,641	1,401	1,401	53%
Brain-Cerebellum	Ovary, Uterus, Vagina	1,680	840	2,641	1,401	1,401	53%
Brain-Cortex	All but Ovary, Uterus, Vagina	1,444	710	2,523	1,413	1,413	56%
Brain-Cortex	Ovary, Uterus, Vagina	1,464	720	2,523	1,413	1,413	56%
Brain-Frontal-Cortex-BA9	All but Ovary, Uterus, Vagina	1,230	615	2,443	1,449	1,449	59%
Brain-Frontal-Cortex-BA9	Ovary, Uterus, Vagina	1,248	624	2,443	1,449	1,449	59%
Brain-Hippocampus	All but Ovary, Uterus, Vagina	995	497	2,520	1,395	1,395	55%
Brain-Hippocampus	Ovary, Uterus, Vagina	1,008	504	2,520	1,395	1,395	55%
Brain-Hypothalamus	All but Ovary, Uterus, Vagina	1,020	498	2,573	1,501	1,501	58%
Brain-Hypothalamus	Ovary, Uterus, Vagina	1,032	504	2,573	1,500	1,500	58%

Brain-Nucleus-accumbens-basal-ganglia	All but Ovary, Uterus, Vagina	1,277	638	2,752	1,498	1,498	54%
Brain-Nucleus-accumbens-basal-ganglia	Ovary, Uterus, Vagina	1,296	648	2,752	1,498	1,498	54%
Brain-Putamen-basal-ganglia	All but Ovary, Uterus, Vagina	969	473	2,427	1,404	1,404	58%
Brain-Putamen-basal-ganglia	Ovary, Uterus, Vagina	984	480	2,427	1,404	1,404	58%
Brain-Spinal-cord-cervical-c-1	All but Ovary, Uterus, Vagina	686	331	2,326	1,389	1,389	60%
Brain-Spinal-cord-cervical-c-1	Ovary, Uterus, Vagina	696	336	2,326	1,388	1,388	60%
Brain-Substantia-nigra	All but Ovary, Uterus, Vagina	731	354	2,505	1,404	1,404	56%
Brain-Substantia-nigra	Ovary, Uterus, Vagina	744	360	2,505	1,404	1,404	56%
Breast-Mammary-Tissue	All but Ovary, Uterus, Vagina	2,423	1,200	2,908	1,524	1,524	52%
Breast-Mammary-Tissue	Ovary, Uterus, Vagina	2,472	1,224	2,908	1,524	1,524	52%
Colon-Sigmoid	All but Ovary, Uterus, Vagina	2,405	1,202	2,784	1,432	1,432	51%
Colon-Sigmoid	Ovary, Uterus, Vagina	2,448	1,224	2,784	1,432	1,432	51%
Heart-Atrial-Appendage	All but Ovary, Uterus, Vagina	2,266	1,133	2,249	1,289	1,289	57%
Heart-Atrial-Appendage	Ovary, Uterus, Vagina	2,304	1,152	2,249	1,289	1,289	57%
Heart-Left-Ventricle	All but Ovary, Uterus, Vagina	1,505	752	2,105	1,212	1,212	58%
Heart-Left-Ventricle	Ovary, Uterus, Vagina	1,536	768	2,105	1,212	1,212	58%
Kidney-Cortex	All but Ovary, Uterus, Vagina	809	404	2,627	1,407	1,407	54%
Kidney-Cortex	Ovary, Uterus, Vagina	816	408	2,627	1,407	1,407	54%
Liver	All but Ovary, Uterus, Vagina	1,181	590	2,241	1,174	1,174	52%
Liver	Ovary, Uterus, Vagina	1,200	600	2,241	1,174	1,174	52%
Lung	All but Ovary, Uterus, Vagina	1,416	708	2,773	1,585	1,585	57%
Lung	Ovary, Uterus, Vagina	1,440	720	2,773	1,585	1,585	57%
Muscle-Skeletal	All but Ovary, Uterus, Vagina	2,169	1,084	1,953	1,019	1,019	52%
Muscle-Skeletal	Ovary, Uterus, Vagina	2,208	1,104	1,953	1,019	1,019	52%
Ovary		1,512	744	2,438	1,430	1,430	59%
Pancreas	All but Ovary, Uterus, Vagina	2,375	1,176	2,231	1,088	1,088	49%

Pancreas	Ovary, Uterus, Vagina	2,424	1,200	2,231	1,086	1,086	49%
Pituitary	All but Ovary, Uterus, Vagina	1,613	806	2,864	1,441	1,441	50%
Pituitary	Ovary, Uterus, Vagina	1,632	816	2,864	1,441	1,441	50%
Prostate	All but Ovary, Uterus, Vagina	1,368	672	2,678	1,504	1,504	56%
Skin-Not-Sun-Exposed-Suprapubic	Ovary, Uterus, Vagina	2,271	1,135	2,709	1,516	1,516	56%
Skin-Not-Sun-Exposed-Suprapubic	All but Ovary, Uterus, Vagina	2,304	1,152	2,709	1,516	1,516	56%
Small-Intestine-Terminal-Ileum	Ovary, Uterus, Vagina	1,605	802	2,888	1,581	1,581	55%
Small-Intestine-Terminal-Ileum	All but Ovary, Uterus, Vagina	1,605	802	2,888	1,581	1,580	55%
Small-Intestine-Terminal-Ileum	Ovary, Uterus, Vagina	1,632	816	2,888	1,581	1,581	55%
Small-Intestine-Terminal-Ileum	All but Ovary, Uterus, Vagina	1,632	816	2,888	1,581	1,580	55%
Spleen	Ovary, Uterus, Vagina	1,504	752	2,604	1,528	1,528	59%
Spleen	All but Ovary, Uterus, Vagina	1,536	768	2,604	1,528	1,528	59%
Stomach	Ovary, Uterus, Vagina	2,541	1,270	2,634	1,262	1,262	48%
Stomach	All but Ovary, Uterus, Vagina	2,592	1,296	2,634	1,261	1,261	48%
Testis		2,184	1,080	4,784	2,020	2,020	42%
Thyroid	Ovary, Uterus, Vagina	2,030	1,015	2,856	1,577	1,577	55%
Thyroid	All but Ovary, Uterus, Vagina	2,064	1,032	2,856	1,577	1,577	55%
Uterus		1,320	648	2,468	1,508	1,508	61%
Vagina		1,296	648	2,695	1,573	1,573	58%
Whole-Blood	Ovary, Uterus, Vagina	1,224	612	2,297	983	983	43%
Whole-Blood	All but Ovary, Uterus, Vagina	1,248	624	2,297	983	983	43%

Table 4.S5 | Gene ontology analysis of differential WMEL.

Comparison	Testis vs. all others	Whole-Blood vs. all other tissues	Testis vs. all other tissues	Muscle-Skeletal vs. all other tissues	Brain-Cereb* vs. all other Brain	Brain-Cereb* vs. all other tissues
End	**5'**	**3'**	**3'**	**3'**	**3'**	**3'**
Background genes to choose from: ubiquitously expressed genes	420	516	516	516	516	516
Total background genes with GO terms	410	503	503	503	503	503
Total background genes with similar GC content & length: longer	160	30	77	87	38	152
Total background genes with similar GC content & length: shorter	41	96	109	120	12	30
Total genes longer in tissue of interest (pBF <= 0.01)	210	37	98	109	45	193
Total genes shorter in tissue of interest (pBF <= 0.01)	50	120	129	141	16	42

Table 4.S6 | Gene ontology analysis of differential WMEL: 3' end.

GO term	pBF	Total genes	Gene list	Comparison
ribosome	<1e-04	7	DNAJC21,MRPL22,MRPS10,MRPS21,RPL29,RPS14,RPS19	Brain-Cereb longer
ribosome	<1e-04	4	DNAJC21,MRPL22,RPL29,RPS14	Muscle-Skeletal shorter
ribosome	<1e-04	6	DNAJC21,MRPL22,MRPS10,MRPS21,RPS14,RPS19	Testis shorter
ribosome	<1e-04	4	DNAJC21,MRPS21,RPL29,RPS14	Whole-Blood shorter
condensed chromosome kinetochore	<1e-04	4	BOD1,BUB3,DYNC1LI1,PPP1CC	Brain-Cereb longer
condensed chromosome kinetochore	<1e-04	4	BOD1,BUB3,DYNC1LI1,PPP1CC	Testis shorter
cytosol	0.0339	22	AKR1B1,BUB3,FBXW11,FHL1,GLUL,LSM4,MAP2K3,NFU1,PCM1,POLR3E,POMP,PPP1CC,PPP2CB,PSMB2,PSMD12,RAN,RARS,RPL19,RPS19,SMAD7,SNRPE,TXNIP	Muscle-Skeletal longer
cytosol	<1e-04	11	CTSH,FHL1,GLUL,LSM4,MAP2K3,PPP1CC,RAC1,RPL19,RPS19,TRAPPC3,TXNIP	Whole-Blood longer
mitosis	<1e-04	4	ARPP19,BOD1,DYNC1LI1,RAN	Brain-Cereb longer
mitosis	<1e-04	4	ARPP19,BOD1,DYNC1LI1,RAN	Testis shorter
protein kinase binding	0.0384	5	CDK5RAP1,MAP2K3,PPP1CC,RAC1,RPS19	Testis shorter
protein kinase binding	<1e-04	4	MAP2K3,PPP1CC,RAC1,RPS19	Whole-Blood longer
RNA metabolic process	<1e-04	4	EIF4A3,PSMD12,RPL29,RPS14	Brain-Cereb vs Brain longer
RNA metabolic process	0.0361	5	EIF4A3,LSM5,PSMD12,RPL29,RPS14	Whole-Blood shorter
transcription, DNA-templated	<1e-04	15	ARGLU1,CALCOCO1,DNTTIP2,ETS2,HINFP,KLF9,MAP3K7,PARP1,PHF2,POLR3E,PWP1,SFPQ,SUV420H1,TXNIP,YBX3	Testis shorter
transcription, DNA-templated	0.0204	7	ARGLU1,HTATSF1,MLXIP,PARP1,PHF2,SP1,TXNIP	Whole-Blood longer
cell differentiation	<1e-04	4	FHL1,GADD45B,HTATIP2,NUS1	Brain-Cereb longer
centrosome	<1e-04	4	FBXW11,NFU1,PCGF5,PCM1	Muscle-Skeletal longer
endoplasmic reticulum membrane	<1e-04	11	CYB5A,LPCAT4,MTDH,NUS1,PJA2,SSR2,TMED9,TMEM106C,UBE2J1,VAPB,VMA21	Muscle-Skeletal shorter
enzyme binding	<1e-04	5	CTNNBL1,CYB5A,HINFP,PARP1,VAPB	Brain-Cereb vs Brain longer
histone deacetylase binding	<1e-04	4	GMNN,HES1,NR2C1,YWHAE	Testis longer

hydrolase activity	<1e-04	4	MGLL,NUDT9,PPP2CB,PSMC6	Testis longer
mRNA metabolic process	<1e-04	4	EIF4A3,PSMD12,RPL29,RPS14	Brain-Cereb vs Brain longer
negative regulation of transcription from RNA polymerase II promoter	<1e-04	7	MTDH,NR2C1,PARP1,RPS14,SFPQ,TXNIP,YBX3	Testis shorter
nuclear membrane	<1e-04	4	AKIRIN1,MTDH,RANGAP1,SEPHS1	Muscle-Skeletal shorter
nucleolus	<1e-04	5	MTDH,PARP1,PHF2,PPP1CC,RPS19	Whole-Blood longer
oxidation-reduction process	<1e-04	5	CREG1,CYB5R1,FADS3,HADH,SDHB	Muscle-Skeletal longer
oxidoreductase activity	<1e-04	5	AKR1B1,CREG1,CYB5R1,HADH,SDHB	Muscle-Skeletal longer
peptidase activity	<1e-04	4	CTSH,LGMN,PEPD,THOP1	Muscle-Skeletal shorter
protein polyubiquitination	<1e-04	4	FBXW11,PSMB2,PSMD12,UBE2V2	Muscle-Skeletal longer
protein transport	<1e-04	7	GOSR1,IFT27,NECAP1,RAB18,RAB9A,RAC1,RAN	Testis shorter
proteolysis	<1e-04	7	ABHD10,ADAMTS1,CTSH,CTSL,LGMN,PEPD,THOP1	Muscle-Skeletal shorter
regulation of transcription, DNA-dependent	0.0384	11	ARGLU1,CLPX,DNTTIP2,ETS2,GTF2B,HINFP,HSF2,KLF9,NR2C1,SUV420H1,YBX3	Testis shorter
sequence-specific DNA binding	0.0384	4	CALCOCO1,ETS2,HSF2,NR2C1	Testis shorter
small molecule metabolic process	<1e-04	16	AKR1B1,B4GALT5,COX7B,FH,GLUL,HADH,MGLL,PHYH,PPP1CC,PSMB2,PSMD12,RAN,SDC2,SDHB,SLC25A4,SUCLG1	Muscle-Skeletal longer
SRP-dependent cotranslational protein targeting to membrane	<1e-04	4	RPL29,RPS14,SRP9,SSR2	Muscle-Skeletal shorter
structural constituent of ribosome	<1e-04	4	MRPL15,MRPS21,RPL29,RPS14	Whole-Blood shorter
transcription coactivator activity	<1e-04	7	CALCOCO1,HINFP,HSF2,HTATIP2,MTDH,PHF2,RAN	Testis shorter
transcription factor complex	<1e-04	4	CREG1,LMO4,NFIC,SKI	Testis longer
viral process	<1e-04	8	GTF2B,PSMB2,PSMD12,RAC1,RAN,RPS14,RPS19,SLC25A4	Testis shorter
zinc ion binding	<1e-04	11	ALAD,CRIP2,GTF2B,LMO4,MT1E,NR2C1,PJA2,SKI,ZMIZ2,ZMYM3,ZNF330	Testis longer

GO Term, Gene Ontology Term; pBF, Bonferroni-corrected p-value; Total genes, total genes from the test set in the enriched term; Gene list, genes from the test set in the enriched term; Comparison, indicates the tissue of interest and whether the genes in the test set were significantly longer or shorter in the tissue of interest compared to at least one other tissue.

Table 4.S7 | Gene ontology analysis of differential WMEL: 5' end.

GO term	pBF	Total genes	Gene list	Comparison
endoplasmic reticulum	<1e-04	16	APOD,ATL2,CAMLG,CAT,CYB5A,EMC2,FOS,PDIA4,RCN2,RRBP1,SRP72,STT3B,TRAPPC3,UBXN4,VIMP,YIPF6	Testis longer
endoplasmic reticulum	<1e-04	16	APOD,ATL2,CAMLG,CAT,CYB5A,EMC2,FOS,PDIA4,RCN2,RRBP1,SRP72,STT3B,TRAPPC3,UBXN4,VIMP,YIPF6	Testis longer
intracellular	0.0418	14	APOD,BIRC2,CAPN2,IFT27,NUDT9,PICALM,PROSC,RAB13,RAC1,RPL19,RPL22L1,RPS19,UBL3,ZNF282	Testis longer
positive regulation of transcription from RNA polymerase II promoter	<1e-04	11	CALCOCO1,CDK7,CITED2,ERCC3,FOS,HSF2,MED10,NFKBIA,PARP1,STAT3,TFE3	Testis longer
biological process	<1e-04	11	BCL7B,DHRS7B,FAM3C,MKRN1,MRPS10,PROSC,SDC4,SGTA,TMEM128,TMEM57,ZNF330	Testis longer
plasma membrane	0.0187	10	ADIPOR1,ANXA1,AP2S1,DYNC1LI1,HLA-DRA,MGLL,PKM,RAB9A,SLC2A3,SPRY2	Testis shorter
modulation by virus of host morphology or physiology	<1e-04	9	CAMLG,ERCC3,GTF2B,KPNA3,NFKBIA,SGTA,SLC25A4,STAT3,YWHAE	Testis longer
molecular function	<1e-04	8	DHRS7B,MKRN1,MRPS10,ORC5,PTTG1IP,TMEM128,TMEM57,VPS45	Testis longer
cellular component	<1e-04	7	BCL7B,COMMD7,DHRS7B,MKRN1,PIP4K2A,TMEM128,VPS45	Testis longer
protein C-terminus binding	<1e-04	6	CALCOCO1,CDK7,ERCC3,IFT52,KPNA3,PRDX3	Testis longer
transcription initiation from RNA polymerase II promoter	<1e-04	6	CDK7,ERCC3,GTF2B,MED10,NR2C1,PARP1	Testis longer
protein phosphorylation	<1e-04	6	CDK7,CSNK2A1,ERCC3,MAP3K7,OXSR1,STK40	Testis longer
protein kinase activity	<1e-04	6	CDK7,CSNK2A1,ERCC3,MAP3K7,OXSR1,STK40	Testis longer
protein serine/threonine kinase activity	<1e-04	5	CDK7,CSNK2A1,MAP3K7,OXSR1,STK40	Testis longer
protein transport	<1e-04	5	AP2S1,RAB9A,SELENBP1,SNX12,TIMM50	Testis shorter
positive regulation of NF-kappaB transcription factor activity	<1e-04	4	CAT,MAP3K7,NFKBIA,PRDX3	Testis longer
negative regulation of transcription, DNA-dependent	<1e-04	4	CITED2,COMMD7,NR2C1,ZNF282	Testis longer
DNA replication	<1e-04	4	MCM3,ORC5,RPA2,WRNIP1	Testis longer
nucleotide-excision repair	<1e-04	4	CDK7,ERCC3,RAD23B,RPA2	Testis longer
transforming growth factor beta receptor signaling pathway	<1e-04	4	CITED2,FOS,MAP3K7,PARP1	Testis longer
transcription regulatory region DNA binding	<1e-04	4	CALCOCO1,FOS,STAT3,TFE3	Testis longer

aging	<1e-04	4	APOD,CAT,FOS,IGFBP2	Testis longer
nucleobase-containing small molecule metabolic process	<1e-04	4	ATIC,CAT,CMPK1,NUDT9	Testis longer
endocytosis	<1e-04	4	LMBR1L,NECAP1,RAC1,USP33	Testis longer
transcription elongation from RNA polymerase II promoter	<1e-04	4	CDK7,ELP3,ERCC3,GTF2B	Testis longer
positive regulation of type I interferon production	<1e-04	4	DHX36,DHX9,NFKBIA,POLR3E	Testis longer
Fc-epsilon receptor signaling pathway	<1e-04	4	FOS,MAP3K7,NFKBIA,RAC1	Testis longer
nucleotide-excision repair, DNA damage removal	<1e-04	4	CDK7,ERCC3,RAD23B,RPA2	Testis longer
cytoplasmic vesicle	<1e-04	4	BICD2,IGFBP2,RAB13,RAC1	Testis longer
microtubule	<1e-04	4	CCT2,DYNC1LI1,MAP2K2,SPRY2	Testis shorter

GO Term, Gene Ontology Term; pBF, Bonferroni-corrected p-value; Total genes, total genes from the test set in the enriched term; Gene list, genes from the test set in the enriched term; Comparison, indicates the tissue of interest and whether the genes in the test set were significantly longer or shorter in the tissue of interest compared to at least one other tissue.

Chapter 5: Alternative RNA processing in macrophages upon endotoxin re-exposure

5.1. Abstract

Upon exposure to a new pathogen, widespread epigenetic transcriptional changes are induced in macrophages. However, upon re-exposure to the same pathogen, the same genes are poorly induced, constituting a form of innate immune memory referred to as tolerization. Although the lack of transcriptional response is well appreciated in tolerized macrophages, its mechanism is poorly understood. Here, we identified alternative transcription start sites (ATSS) and alternative polyadenylation sites (APA) in poly(A)+ and chromatin-associated RNA of macrophages naïve (previously unexposed) and tolerized to Lipid A (LPA), the active component of lipopolysaccharide (LPS). In the chromatin-associated cell fraction, APA is interpreted as alternative transcription termination (ATT), as this fraction contains both new transcripts and transcripts that remain associated with the chromatin after transcription terminates. While APA has been studied in the macrophage response to LPS, bacteria, and virus, this is to our knowledge the first study of ATSS and APA in tolerized cells, and the first time predicting ATSS and ATT in chromatin-associated RNA. We identified many APA and some TSS events in tolerized vs. naïve macrophages. In the subset of APA events, transcription typically terminated after the distal poly(A) site, suggesting that transcription termination has little influence on APA definition. Overall, this study suggests that APA and ATSS contribute to the tolerized phenotype.

5.2. Introduction

Macrophages, a type of white blood cell in the immune system, undergo extensive transcriptional changes upon exposure to a new pathogen or toxin, including expression induction of many

inflammatory genes. Upon repeated or prolonged exposure to the same pathogen, macrophages develop a tolerance to the pathogen to avoid overstimulation. Premature or prolonged tolerization may lead to immune system suppression, e.g. in some sepsis patients, which can lead to higher incidence of infection and is sometimes fatal [111]. LPS, a component of the outer membrane in Gram-negative bacteria, is an example toxin that can induce macrophage tolerization. Compared to naïve macrophages, i.e. those that were not previously exposed to LPS, tolerized macrophages are associated with poor induction of genes induced in the naïve state, including genes involved in the inflammatory NFkB and MAPK pathways [112]. Tolerization was also previously associated with epigenetic changes due to microRNA-mediated regulation of chromatin remodeling factors [113,114]. In addition to epigenetic and transcriptional changes, inefficient splicing and alternative 3' end regulation are other possible mechanisms underlying poor inflammatory gene induction in the tolerized state.

Changes in 3' end regulation were recently associated with different immune response phenotypes. Shorter 3'UTRs due to APA were observed transcriptome-wide in macrophages upon bacterial or viral infection [115,116] and in LPS-stimulated human monocyte-derived macrophages [117]. In addition to APA, transcription termination efficiency can change in response to stress. While transcription termination typically occurs soon after cleavage and polyadenylation (reviewed in [8]), various cell stresses including oxidative stress specifically reduced transcription termination efficiency in mouse fibroblasts [118].

In addition to alternative 3' end regulation, ATSS may also contribute to the immune response. Generally, promoters in close proximity can give rise to new alternative TSSs [105], often leading to translational changes [7,60,99]. Many alternative promoter events were observed in LPA-stimulated human monocyte-derived macrophages, often changing the coding sequence by >100bp [117].

Based on this evidence, we hypothesize that ATSS and APA may contribute to the tolerized macrophage phenotype. Previously, cellular fractionation into cytoplasmic, nucleoplasmic, and chromatin-associated RNA-Seq in time series following LPA treatment revealed novel insights into gene regulation in the LPS response [88,119]. In particular, transcripts remain on the chromatin until splicing has completed, even after transcription completion. Here, we used a similar time series experiment to identify ATSS and alternative 3' end regulation in chromatin-associated and whole cell poly(A)+ cellular fractions of naïve and tolerized macrophages. Because the algorithm described in Section 3.3 is inherently robust to RNA-Seq non-uniformity and identifies change points in de novo TUs, our algorithm is well-poised to robustly predict ATSS and ATT events in chromatin-associated RNA-Seq, which is highly non-uniform compared to poly(A)+ RNA-Seq and may exhibit ATSS or 3' readthrough far beyond the annotated gene. We identified several ATSS and APA events that may contribute to the tolerized phenotype.

5.3. Methods

Cell fractionation and RNA sequencing

Bone marrow-derived macrophages (BMDMs) were established and RNA was isolated from cellular fractions as previously described [88,119]. Briefly, red-blood-cell-depleted murine bone marrow cells were cultured with M-CSF-containing medium (L929 cell conditioned medium) for 7 days, followed by replating for an additional 2 days. BMDMs were stimulated with 100 ng/ml Lipid A (Invivogen) and harvested at 0, 10, 15, 20, 25, 30, 40, 50, 60, 75, 90, and 120 minutes after stimulation for chromatin-associated RNA, and at 0, 30, 60, 120, 180, 300, and 480 minutes after stimulation for whole cell poly(A)-selected RNA. Lipid A-tolerized BMDMs were generated by stimulation with Lipid A for 24 hours starting 8 hours after replating, followed by 16 hours of rest.

Then, cells were additionally stimulated with Lipid A and harvested at the time points described above (Fig. 5.1).

RNA-Seq pre-processing and alignment

The mountainClimber tool suite and mapping pipeline described in Section 3.3 was used.

Adapter trimming: The Illumina Universal Human Adapter was trimmed from both poly(A)+ and chromatin cell fractionation RNA-Seq with cutadapt [26] as follows: cutadapt -a AGATCGGAAGAG -n 5 -m 35 -O 5 -e 0.2. For poly(A)+, poly(A) and poly(T) sequences were additionally trimmed.

Genome alignment: Reads were aligned to the mm10 genome with hisat2 [82] and the following parameters: --dta-cufflinks --mp 6,4 --no-softclip --no-mixed --no-discordant --add-chrname -k 100. After alignment, exon-exon junction reads were retrieved from the bam files.

de novo change point identification

mountainClimberTU: TUs were called using mountainClimberTU. For poly(A)+ RNA, the default parameters were used. Due to the high degree of non-uniformity in chromatin-associated RNA, depth and breadth parameters were relaxed to -n 3 (3 average reads per bp in 1kb windows) and -p 0.5 (50% of the 1kb window must contain reads). TUs identified in all samples from both fractions were merged to create the RSEM reference.

Transcriptome alignment and RSEM: hisat2 and RSEM were used with the GENCODE vM10 annotation as described in Section 3.3.

mountainClimberCP: 5' and 3' ends may be vastly different in the chromatin fraction in different conditions, for example due to 3' readthrough into downstream genes. For simplicity, we only analyze change points in TUs with a single annotated gene. Therefore, change points were called in TUs merged across replicates rather than merged across all samples in order to maximize the number of TUs with a single annotated gene. mountainClimberCP was run with default parameters for poly(A)+ RNA-Seq and with relaxed expression minimum -e 3 and stringent fold change minimum -f 3 for chromatin-associated RNA-Seq due to the high degree of non-uniformity.

mountainClimberTest: Differential change points were tested between all pairwise conditions with two samples after removing outliers and only testing between the same batch (see "Batch effect identification and outlier removal" below). First, TUs were merged among all samples in the two sample groups of interest (e.g. tolerized vs. naïve or time point vs. time point) and filtered to include only TUs with one annotated gene. Because there were some sample quality issues (described below), mountainClimberTest_cluster was run with parameters -n 2 and -d 2 to stringently require that change points were observed in both replicates and both conditions, respectively. Because the chromatin fraction distal ends are noisier across replicates (data not shown), we set the DBSCAN neighborhood size to the maximum optimal mountainClimberCP window size across replicates, while the minimum across replicates was used for poly(A)+ RNA-Seq (see Section 3.3 for further mountainClimberTest details). mountainClimberTest_readCounts and mountainClimberTest_ru were run with default parameters as described in Section 3.3. Finally, mountainClimberTest_diff was run with maximum p-value -t 0.01 and minimum relative usage difference -m 0.1 for added stringency in both cell fractions. For poly(A)+ RNA, we additionally set the mountainClimberTest_diff minimum mean reads/bp in the proximal segment -p to 50 for added stringency. -p 50 was not required in the chromatin-associated RNA given the high degree of non-uniformity.

Change point annotation: Change point annotation was done as described in Section 3.3 with GENCODE vM10.

Relative usage calculation for each sample: To calculate the relative usage for each sample, we used a similar approach as described in Section 3.3.

Weighted 3' length analyses

Weighted average end calculation: The weighted average end is defined as the weighted average of the 5' or 3' end in each sample, using the relative usage of each end as the weights. To calculate the relative weighted 3' length in poly(A)+ RNA, the weighted 3' end is compared to a reference sample.

Batch effect identification and outlier removal: Poly(A)+ samples were clustered by weighted 3' length relative to PolyA.Naive.000.b2 (i.e. poly(A)+ RNA naïve macrophage at time 0, replicate #2) to systematically identify sample outliers. This revealed five candidate outlier samples. Upon manual inspection of the read distribution in the housekeeping gene ACTB across all samples, we observed that the RNA-Seq was apparently artificially non-uniform in these five samples, and additionally in the reference sample PolyA.Naive.000.b2. Therefore, the following six samples were removed from all downstream analysis: PolyA.LPA.000.b2, PolyA.LPA.000.b3, PolyA.LPA.120.b3, PolyA.LPA.480.b2, PolyA.Naive.000.b2, and PolyA.Naive.300.b3. Because there were batch effects, we kept all batches separate (replicates b0 and b1 in batch "b0b1", and replicates b2 and b3 in batch "b2b3") for all mountainClimberTest analysis.

Gene Ontology

Gene Ontology (GO) analysis was performed as described previously [57]. Briefly, background gene sets with similar gene length and GC content were chosen from all genes with change points called in the comparison.

Comparison of chromatin-associated vs. poly(A)+ 5' and 3' ends

To compare the 5' and 3' ends across fractions, we calculated the median distal 5' and 3' ends across replicates in each fraction and the distance between them for each time both with both fractions available (times 0, 30, 60, and 120). Only annotated TUs were considered. To ensure reproducibility across replicates, TUs with distal ends having at least 200bp standard deviation were ignored. In total, 209,022 / 291,670 (91%) and 204,790 / 281,670 (89%) total TUs across all samples were analyzed at the 3' and 5' ends respectively (i.e. ~10% of TUs were ignored)

To compare significantly differential APA in poly(A)+ with the chromatin-associated 3' end, we separated genes in to two categories: (1) genes with change points detected in the chromatin, and (2) genes with no change points detected in the chromatin. The two fractions were compared by subtracting the proximal and distal poly(A) 3' end from either the proximal 3' end (category 1) or the distal 3' end (category 2). Plotting the difference between chromatin 3' and proximal poly(A) vs. the difference between chromatin 3' and distal poly(A) allows us to interpret each of the four quadrants as follows: quadrant 1, the chromatin 3' end is past both APA sites; quadrant 3, the chromatin 3' end is before both APA sites; quadrant 4, the chromatin 3' end is between the APA sites (quadrant 2 is always empty).

5.4. Results

5.4.1 Identification of alternative 5' and 3' ends

Prior evidence of alternative 5' and 3' ends in naïve macrophages and transcriptional changes in tolerized immune cells suggest that alternative 5' and 3' ends may contribute to the tolerized phenotype. Analogous to previous splicing studies in cellular fractions of macrophages after LPA treatment, here we identified alternative 5' and 3' ends in cellular fractions of tolerized and naïve macrophages (Fig. 5.1). Overall, there were approximately 20 to 70 million reads per poly(A)+ sample after merging sequencing lanes and 50 to 125 million reads per chromatin-associated RNA sample (data not shown). We observed an average of 90% and 74% mapping rate in poly(A)+ and chromatin-associated RNA respectively.

We first applied mountainClimberTU to identify TUs in each sample (Fig. 5.S1a) and identified 80,130 TUs after merging across all samples. Of these, 65,705 (82%) were at least 1kb long and eligible for calling change points with mountainClimber. Of those, 51,789 (79%) were unannotated, and 47,739 (92%) of these were specific to the chromatin-associated fraction, suggesting that we captured many TUs that may be enhancer RNAs, upstream antisense transcription in the promoter region, or other unannotated transcriptional products. Of the 13,916 annotated TUs, 8,713 (63%) overlapped one gene while the rest contained more than one gene. To gain more TUs overlapping a single gene, we only merged TUs across replicates instead of across all samples (see Section 5.3 Methods).

After calling TUs, we identified change points in each sample with mountainClimberCP. Similar to our observations in poly(A)+ RNA from human tissues in Section 4.4, most predicted distal and tandem poly(A) sites were near annotated poly(A) sites or within annotated 3'UTRs (Fig. 5.2a and Fig. 5.S1b). Briefly, tandem APA sites were defined as change points predicted after the last observed exon-exon junction. Analogously, predicted distal and tandem TSS were

most often overlapping annotated distal TSS or within annotated 5'UTRs. As expected, most predicted exon-exon junctions overlapped annotated junctions within 10bp. Additionally, similar to the results across human tissues, APA was observed more often than ATSS (Fig. 5.2b) and most change points were tandem events (Fig. 5.S1c).

On the other hand, distal 3' and 5' ends were often intergenic in the chromatin-associated fraction (Fig. 5.2c and Fig. 5.S1d). Due to known 3' readthrough beyond the cleavage and polyadenylation site, this level of intergenic 3' ends is not surprising. Additionally, there were relatively fewer exon-exon junctions identified compared to the poly(A)+ fraction (Fig. 5.2c) and relatively more change points in segments containing introns in the chromatin-associated fraction than the poly(A)+ fraction (Fig. 5.S1e). These observations are reasonable, as the chromatin-associated fraction contains many short unannotated TUs (described above), intermediate splicing products, and may capture RNA prior to transcription completion. In contrast to APA in the poly(A)+ fraction, there were fewer alternative transcription end sites (Fig. 5.2d). Still, some tandem ATT events (labeled TandemAPA) are predicted in the chromatin-associated fraction, suggesting observable cleavage and polyadenylation followed by 3' readthrough. This is investigated further below.

5.4.2 Batch effect identification and outlier removal

To compare the combined effects of ATSS, ATT, and APA across all samples from both fractions, we clustered samples by TU length, defined as the distance from the weighted average 5' end to weighted average 3' end (see Section 5.3 Methods). As expected, the chromatin-associated and poly(A)+ fractions are distinct (Fig. 5.2e). Interestingly, the chromatin-associated samples cluster primarily by time, and secondarily by condition, indicating alternative transcription over time after LPA treatment. Chromatin-associated samples did not strongly cluster by sequencing library date, library preparation maker, or RNA isolation, indicating there were no observable batch effects in

the chromatin-associated RNA-Seq. On the other hand, the poly(A)+ samples primarily cluster by library preparation, date, and RNA isolation, indicating presence of batch effects.

Because sample clustering by weighted TU length revealed apparent batch effects in the poly(A)+ RNA-Seq, and 3' end bias is often observed in poor quality RNA-Seq data, we clustered poly(A)+ samples by weighted 3' length relative to PolyA.Naive.000.b2 (i.e. poly(A)+ RNA naïve macrophage at time 0, replicate #2) to systematically identify sample outliers (Fig. 5.S2). This analysis combined with manual inspection of the RNA-Seq in a housekeeping gene (data not shown) revealed six sample outliers which were excluded from all downstream analysis. Intriguingly, there were no strong observable batch effects based on gene expression (data not shown). This suggests that while gene expression may not necessarily be affected by poor quality RNA-Seq, high quality RNA-Seq is essential for change point identification. For the remaining analysis, we will refer to two batches based on the library preparation / date / RNA isolation batch: batch b0b1 (containing replicates b0 and b1) and batch b2b3 (containing replicates b2 and b3).

5.4.3 Alternative transcription start and polyadenylation sites in tolerized vs. naïve macrophages

After excluding the six problematic samples, we tested for differential change points in tolerized vs. naïve macrophages at each time point. While there were few significantly differential ATSSs, there were several tandem APA events observed in poly(A)+ RNA between tolerized and naïve macrophages at each time point (Fig. 5.3a). 3'UTR shortening was more frequent than 3'UTR lengthening in tolerized cells compared to naïve, though both were observed. One of the most significant differential ATSS was in Ubl7, ubiquitin-like 7 (bone marrow stromal cell-derived) (Fig. 5.3b). One of the most significant differential APA sites in poly(A)+ samples was in Papd4, a cytoplasmic poly(A) RNA polymerase in (Fig. 5.3c). As post-transcriptional polyadenylation by

Papd4 can influence stability and translation of target mRNAs, its APA may have widespread downstream effects.

GO analysis revealed that APA and ATSS events were enriched in membrane- and vesicle-related genes e.g. APA in endoplasmic reticulum membrane, COPI vesicle coat, exocyst, and intracellular membrane-bounded organelle, and ATSS in phagocytic membrane vesicle (Fig. 5.3d). Alternative polyadenylation can affect membrane protein localization [94], suggesting the possibility that membrane proteins may be differentially localized in tolerized cells. Additionally, both APA and ATSS were enriched with immune pathways (e.g. APA in genes related to positive regulation of NF-kappaB import into nucleus and positive regulation of interferon-beta biosynthetic process, and ATSS in genes related to defense response and cytokine activity) (Fig. 5.3d,e). Two of the 4 genes enriched in the GO term positive regulation of NF-kappaB import into nucleus, Ptgs2 and Tnf, were also induced upon LPA treatment in naive cells. APA in genes related to endoplasmic reticulum, vesicle, and Golgi may affect cytokine release, as cytokines with signal peptides are trafficked from the ER through the Golgi to the cell surface (reviewed in [120]). Intriguingly, siRNA binding was also enriched with alternative polyadenylation, including the genes Ago2, Tarbp2, Tlr7, and Tlr9. Ago2, the catalytic component of the RNA induced silencing complex, is essential for microRNA-mediated mRNA silencing. Thus, APA of Ago2 may contribute to differential gene silencing in tolerized vs. naïve cells. Together, these results suggest that APA and ATSS may be additional previously unappreciated mechanisms of differential response to toxins in naïve and tolerized cells.

Analogous to the poly(A)+ RNA, change points were also identified in the chromatin-associated RNA. 287 total differential events were identified in tolerized vs. naïve cells across all time points (Fig. 5.3f). For example, Gramd1b exhibited ATSS in tolerized time 30 vs. 120, illustrating that mountainClimber is robust to the RNA-Seq non-uniformity in the chromatin-associated fraction (Fig. 5.3g). Note that many other change point detection approaches that only

identify tandem events would miss the ATSS shown in Fig. 5.3g as the proximal TSS is several exons away from the distal TSS.

5.4.4 Comparison of cellular fractions reveals kinetics of alternative polyadenylation regulation

Previous studies across cellular fractions revealed that transcription terminates close to the cleavage and polyadenylation site for most genes, with some exceptional genes having extensive 3' readthrough [88]. Here, we first compared chromatin-associated and poly(A)+ 3' ends as before, but with more precise de novo TU definitions. Next, we tested whether ATT co-occurs with APA, which has not been previously addressed with cell fractionation data to our knowledge.

To compare the chromatin-associated vs. poly(A)+ 5' and 3' ends, we calculated the median distal 5' and 3' ends across fractions (see Section 5.3 Methods). As expected, the TSS was highly consistent across both fractions; 25,464 / 31,308 (81%) total TSSs across all samples were within +/-200bp (Fig. 5.4a). At the 3' end, most genes terminated transcription near the cleavage and polyadenylation site, consistent with previous observations [88]. Still, there was observable 3' readthrough for many genes; 13,756 / 32,238 (43%) total 3' ends across all samples extended > 1kb beyond the 3' end observed in the poly(A)+ fraction (Fig. 5.4b). There was apparent negative readthrough for some genes; upon manual inspection of some examples, we found that longer TUs may be identified in poly(A)+ than chromatin-associated RNA if exon-exon junction reads were present poly(A)+ but not chromatin-associated RNA. However, these were a minority of cases. Overall, the 5' and 3' consistency across fractions for most genes as well as 3'readthrough for a significant portion of genes was expected.

Since we were able to predict change points in both chromatin-associated and poly(A)+ fractions, we next tested for concordance between ATT and APA (see Section 5.3 Methods). If transcription extends beyond the distal poly(A) site, then trans factor activity drives the observed

APA. On the other hand, if ATT is detected between the APA sites in the mature mRNA, then transcription rate influences APA. Out of 2,212 total significantly differential change points across all comparisons, a corresponding change point was identified in the same TU in the same comparison in the chromatin fraction for 23 cases. For 78% (18 / 23) cases, the chromatin 3' end surpassed both APA sites (quadrant 1 Fig 5.3c). Of the remaining 2,189 change points without a corresponding change point identified in the chromatin fraction, 1,293 (59%) had a corresponding distal 3' end in the chromatin fraction. This set excludes, for example, those TUs in the chromatin containing multiple annotated genes. Comparing the distal chromatin 3' end to each APA site identified in the poly(A)+ fraction, including non-differential APA sites, revealed a similar pattern; 1,064 / 1,293 (82%) were in quadrant 1, often with transcription terminating soon after the distal poly(A) site (Fig. 5.4d). This suggests that transcription terminates near or beyond the distal APA site, and trans factors are responsible for the observed APA. Wdr18 is one example from quadrant 1 of APA in tolerized vs. naïve at time 30, and the chromatin 3' end is near the distal end identified in poly(A) (Fig. 5.4e). Interestingly, there was 3' readthrough several kb downstream of Papd4 into the next gene (Fig. 5.4f). Because the chromatin TU contained multiple annotated genes, Papd4 was not considered in the analysis across cell fractions. Thus, the results reported here are likely a lower limit of the total TUs with extended 3' readthrough relative to APA.

5.5. Discussion

In summary, we report de novo TU definition and alternative 5' and 3' end prediction in tolerized and naïve macrophages. To our knowledge, this is the first report of ATSS and APA observed in tolerized macrophages. While epigenetic and transcriptional changes in tolerized macrophages are well established, we provide evidence that alternative polyadenylation and alternative transcription start site usage may additionally contribute to the tolerized phenotype.

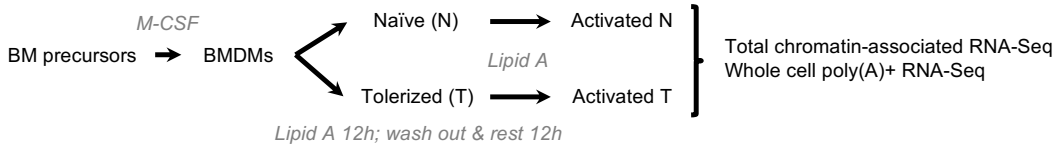
Alternative transcription events in tolerized vs. naïve macrophages were dominated by tandem APA events. While the magnitude of total APA events identified was less than, for example, tissue-specific APA described in Section 4, we identified APA in several genes that may lead to widespread downstream effects. In particular, we highlight APA in Papd4 and Ago2. APA of these genes may remove cis-regulatory elements and lead to changes in translation efficiency, localization, or stability (reviewed in Section 1). Papd4 was recently shown to regulate cell cycle and senescence (mediated by CPEB or QKI) by influencing stability and translation of target mRNAs [121,122]. Although the APA events we identified are both known poly(A) sites, the functional consequence of APA in Papd4 is still poorly understood. Recently, two microRNAs were reported to regulate murine macrophage tolerization by targeting chromatin remodeling factors, thereby leading to transcriptional silencing of inflammatory genes [114]. Because this functionality is dependent on Ago2, APA-induced changes of Ago2 may widely disrupt microRNA-mediated gene silencing. In the future, further investigation of the impact of APA in Papd4 and Ago2 may add further insight into the mechanism of macrophage tolerization.

For the first time, to our knowledge, we applied change point prediction to chromatin-associated RNA. We suspect this was not done previously due to technical difficulties in identifying change points in highly non-uniform data. With this capability, we were able to analyze co-occurrence of ATT and APA events and determine whether transcription termination or trans factors dominate the definition of APA sites. Although the role of trans factors in regulating APA is well appreciated, the role of transcription rate was not previously well understood. We concluded that transcription termination likely plays little role in APA definition compared to trans factors. This may be related to the fact that proximal poly(A) sites typically lack the canonical polyadenylation signal motif (reviewed in [10]), which has been implicated in transcription termination (reviewed in [8]). Still, 3' readthrough has been associated with other functional roles. Recently, unspliced nascent transcripts were reported to have extended 3' readthrough and

increased degradation in the nucleus in S. pombe, suggesting that splicing efficiency and 3' readthrough can control the transcriptional response [123]. In the future, de novo TU definition and change point identification in chromatin-associated RNA-Seq will help elucidate the functional roles of 3' readthrough.

While this dataset in time series after LPA treatment lends itself to kinetic studies and lengthening or shortening of TUs over time, we focused on comparing tolerized vs. naïve macrophages in this study. Due to batch effects and sample outliers, batch b0b1 only contained two time points, while batch b2b3 did not contain the baseline timepoint 0. Thus, interpretation of kinetic studies proved difficult. In the future, kinetic studies of ATSS and APA, analogous to splicing kinetic studies done previously, will provide further insight into how these mechanisms impact gene regulation.

5.6. Figures



		0	10	15	20	25	30	40	50	60	75	90	120	180	300	480
caRNA	b1		T N	T N	T N	T N	T N	T N	T N	T N		T N	T N			
	b2	T N	T N	T N	T N	T N	T N	T N	T N	T N	T N	T N	T N			
	b3	T N	T N	T N	T N	T N	T N	T N	T N	T N	T N	T N	T N			
poly(A)+ RNA	b0	T N								T N						
	b1	T N								T N				T N		T N
	b2	~~T N~~					T N			T N			T N	T N	T N	~~T~~ N
	b3	~~T~~					T N			T N			~~T~~ N	T N	T ~~N~~	T N

Figure 5.1 | Experimental overview.
Top: overview of the experimental protocol for macrophage tolerization. Bottom: table of chromatin-associated RNA-Seq (caRNA) and poly(A)+ RNA-Seq across different replicates (b0, b1, b2, b3) and time points measured in minutes (columns). Yellow T and blue N indicate tolerized and naïve samples respectively. Grey crossed out samples indicate those samples were removed due to poor quality.

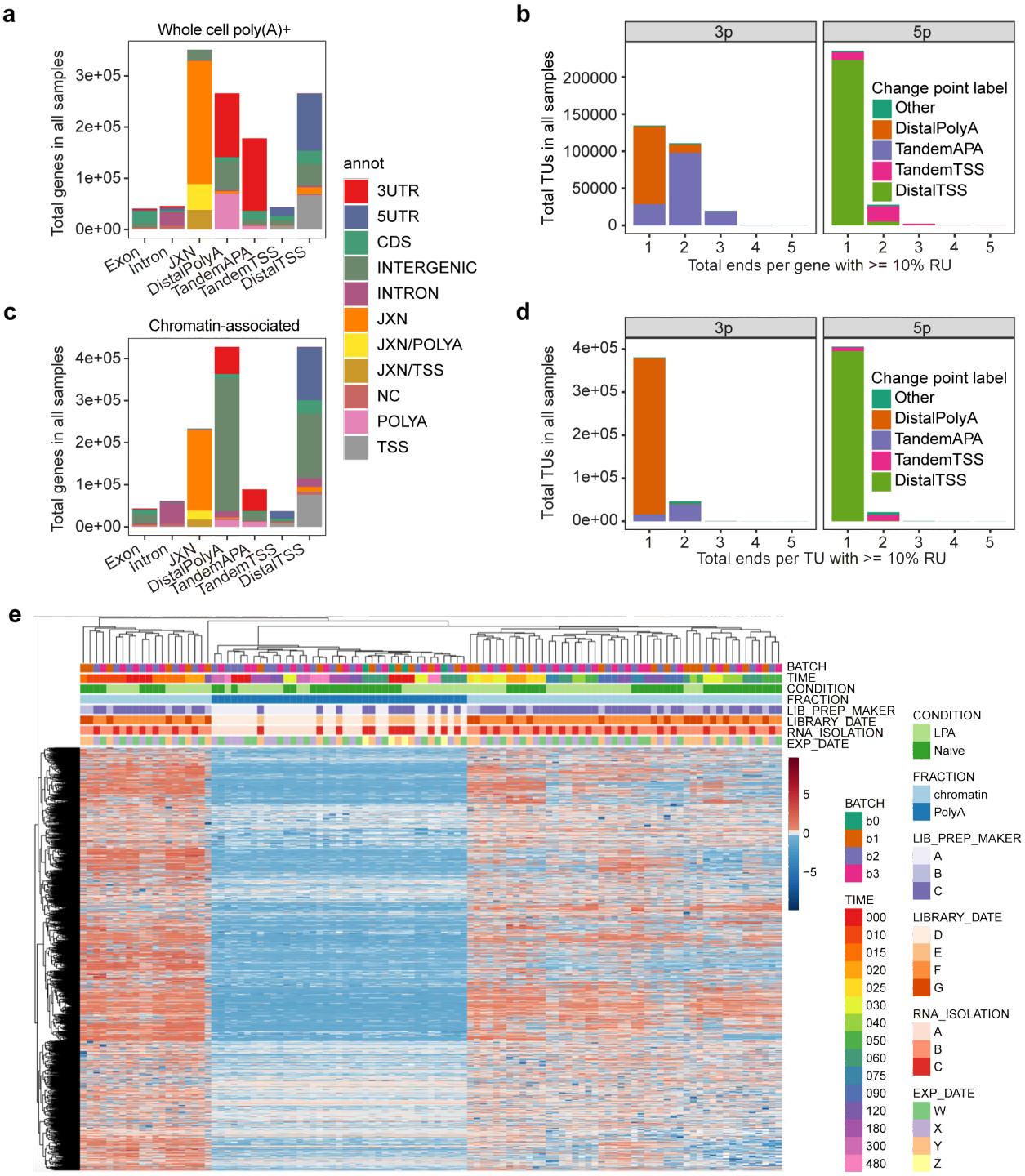


Figure 5.2 | Identification of alternative 5' and 3' ends in macrophage cellular fractions. (**a, c**) The total change points per TU in all samples stratified by each change point type (x-axis) and colored by GENCODE vM10 region in poly(A)+ (**a**) and chromatin-associated RNA (**c**). (**b, d**) Total 5' and 3' ends per TU with at least 10% RU colored by the change point type of the end with maximum RU in poly(A)+ (**b**) and chromatin-associated RNA (**d**), prioritized as follows: Tandem > Distal > Other. (**e**) Clustering of all samples by TU length from weighted 5' end to weighted 3' end, including only group #4 for poly(A)+ and group #3+4 genes for chromatin (see Fig. 5.S1c,e).

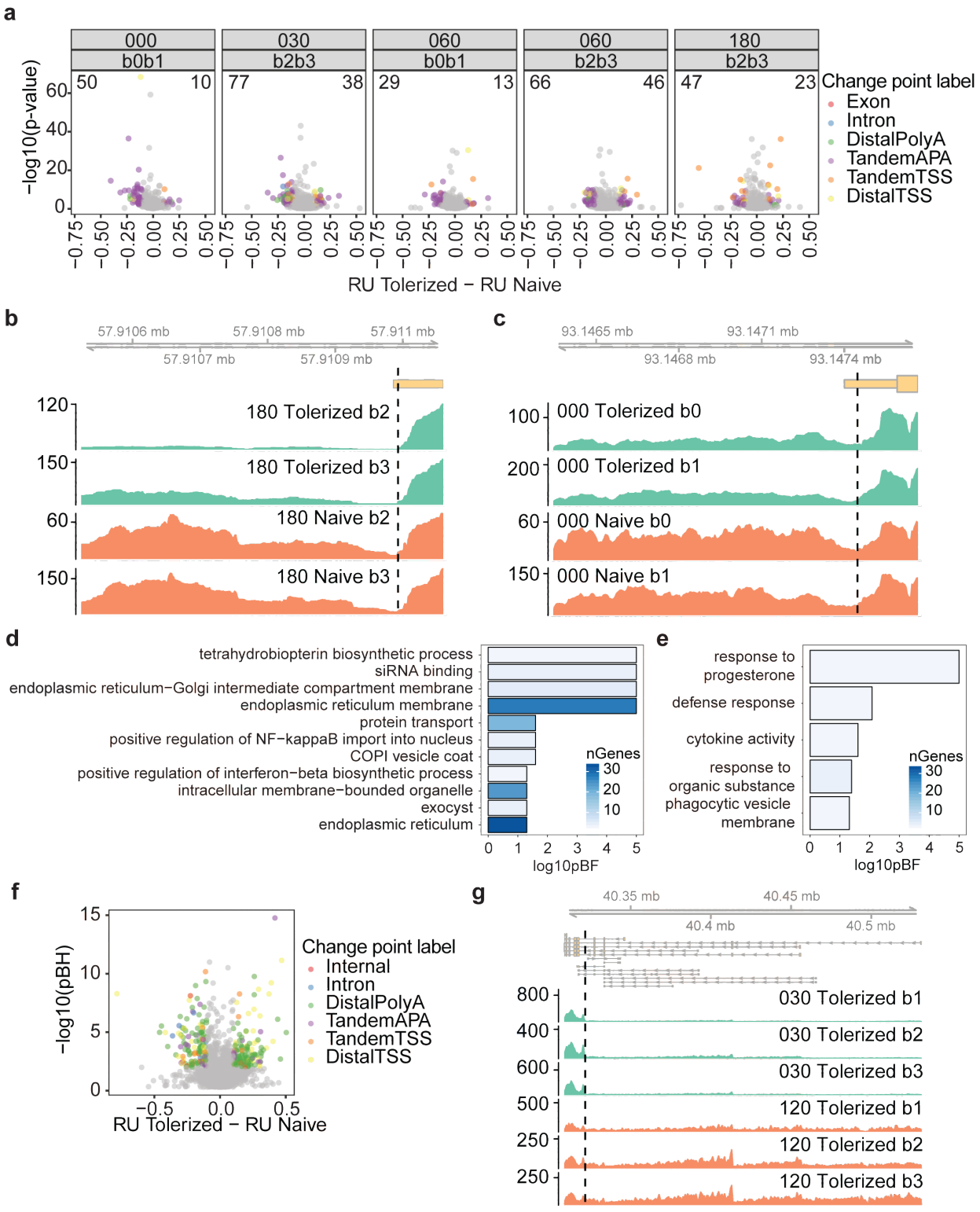


Figure 5.3 | Alternative transcription start and polyadenylation site usage in tolerized vs. naïve macrophages.
(**a**) Volcano plot of change points in tolerized vs. naïve poly(A)+ samples at each time point. For each panel, the top label indicates the time point and the bottom label indicates the batch. The x-

axis is the RU difference (negative indicates shorter UTR in tolerized cells) and the y-axis is the -log10(BH-corrected p-value) from mountainClimberTest. Change points are colored by change point type (grey indicates non-significant). (**b**) Alternative TSS in Ubl7 in poly(A)+ tolerized vs. naïve at time 180 (BH-corrected p = 6.25e-22, RU difference = 0.56). The black dotted line indicates the predicted change point. (**c**) Alternative polyadenylation in Papd4 in poly(A)+ tolerized vs. naïve at time 000 (BH-corrected p = 2.64-15, RU difference = -0.41).The black dotted line indicates the predicted change point. (**d, e**) Top enriched Gene Ontology (GO) terms with ≥ 3 genes and -log10(Bonferroni-corrected p-value) ≤ 0.05 in tolerized vs. naïve in poly(A)+ RNA at any time point the 3' end (**d**) and 5' end (**e**). Bars are colored by the total genes enriched per GO term. (**f**) Volcano plot of change points in tolerized vs. naïve chromatin-associated samples, combining all time points. (**g**) Alternative TSS in chromatin-associated tolerized time 30 vs. 120 in Gramd1b (BH-corrected p = 2.45e-28, RU difference = -0.6). The black dotted line indicates the predicted change point.

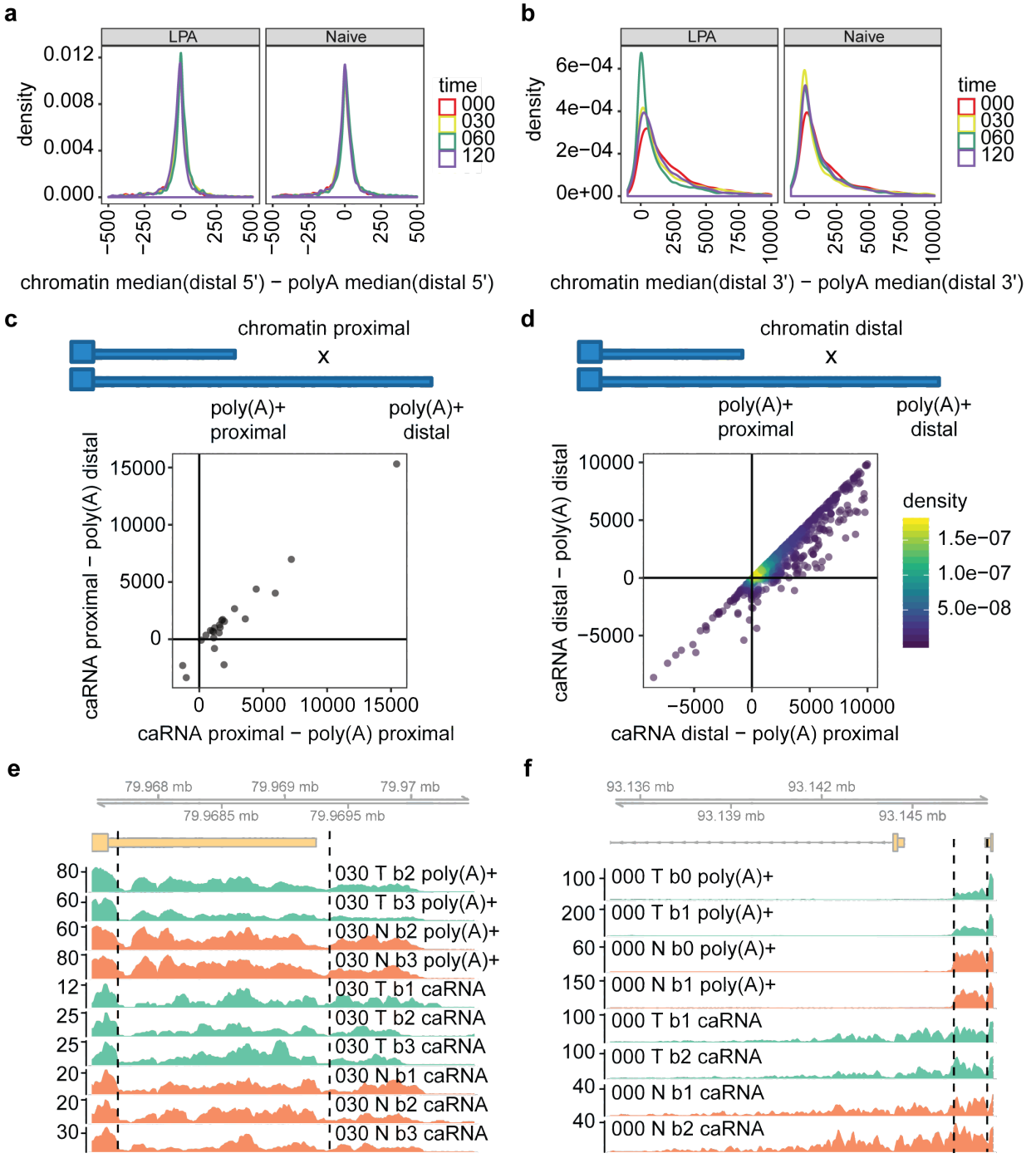


Figure 5.4 | Chromatin-associated vs. poly(A)+ 5' and 3' ends.
(**a, b**) Median distal 5' (**a**) and 3' (**b**) end across replicates in chromatin-associated vs. poly(A)+ RNA, zoomed in to +/- 500bp (**a**) and -1kb to +10kb (**b**). (**c**) chromatin-associated change point 3' end (proximal, by definition) relative to significantly differential 3' end in poly(A)+ in the same pairwise comparison. (**d**) distal 3' end in chromatin-associated RNA relative to all change points in poly(A)+, excluding those genes with change points in the chromatin-associated RNA (i.e. excluding the genes shown in (**c**)). (**e**) Wdr18, an example of a significantly differential 3' end (pBH = 2.65e-06, RU difference = -0.29) in quadrant 1 in panel (**d**). (**f**) Papd4, as in Fig. 5.3d

shown with corresponding chromatin-associated RNA-Seq. T, tolerized; N, naïve. Both proximal and distal change points are shown in dashed black lines.

5.7. Supplementary Figures

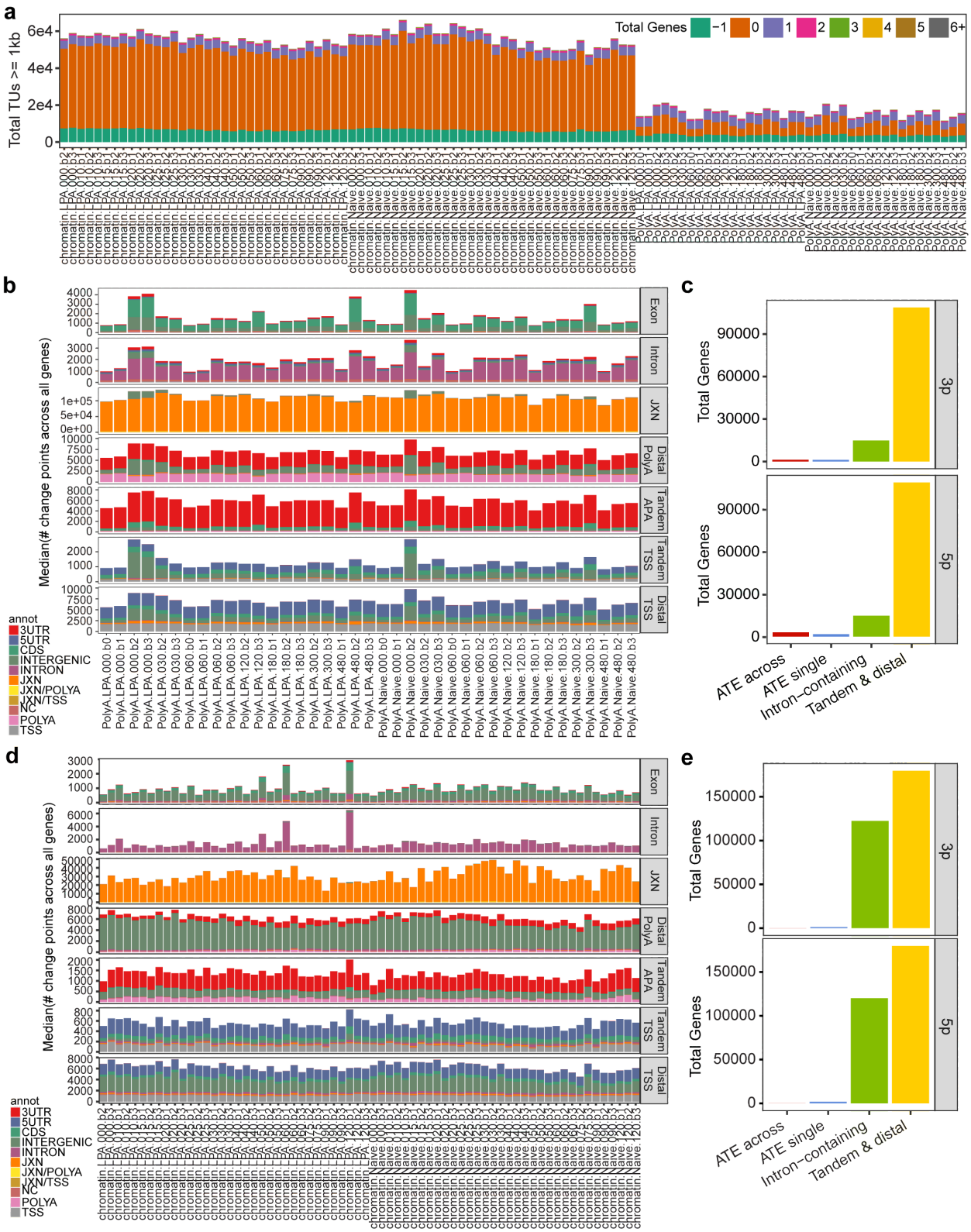


Figure 5.S1 | Overview of mountainClimberTU and mountainClimberCP results.
(**a**) Total transcription units at least 1kb long in each sample, colored by the number of overlapping genes. -1 indicates TUs with no annotation on the sense strand, but with overlap on the antisense strand. (**b, d**) Median total change points per TU in each sample with each change point label (rows) and overlapping GENCODE vM10 gene regions (colored) in poly(A)+ (**b**) and chromatin-associated (**d**) fractions. (**c, e**) Total TUs across all samples in poly(A)+ (**c**) or chromatin-associated RNA (**e**) in each of the four categories described in Section 3.3 (see Fig. 4.S3a for further change point category descriptions). ATE across, alternative terminal (first- or last-) exon across different samples; ATE single, ATE within the same sample; intron-containing, the last segment contains an annotated intron; tandem, the last segment does not overlap any annotated intron.

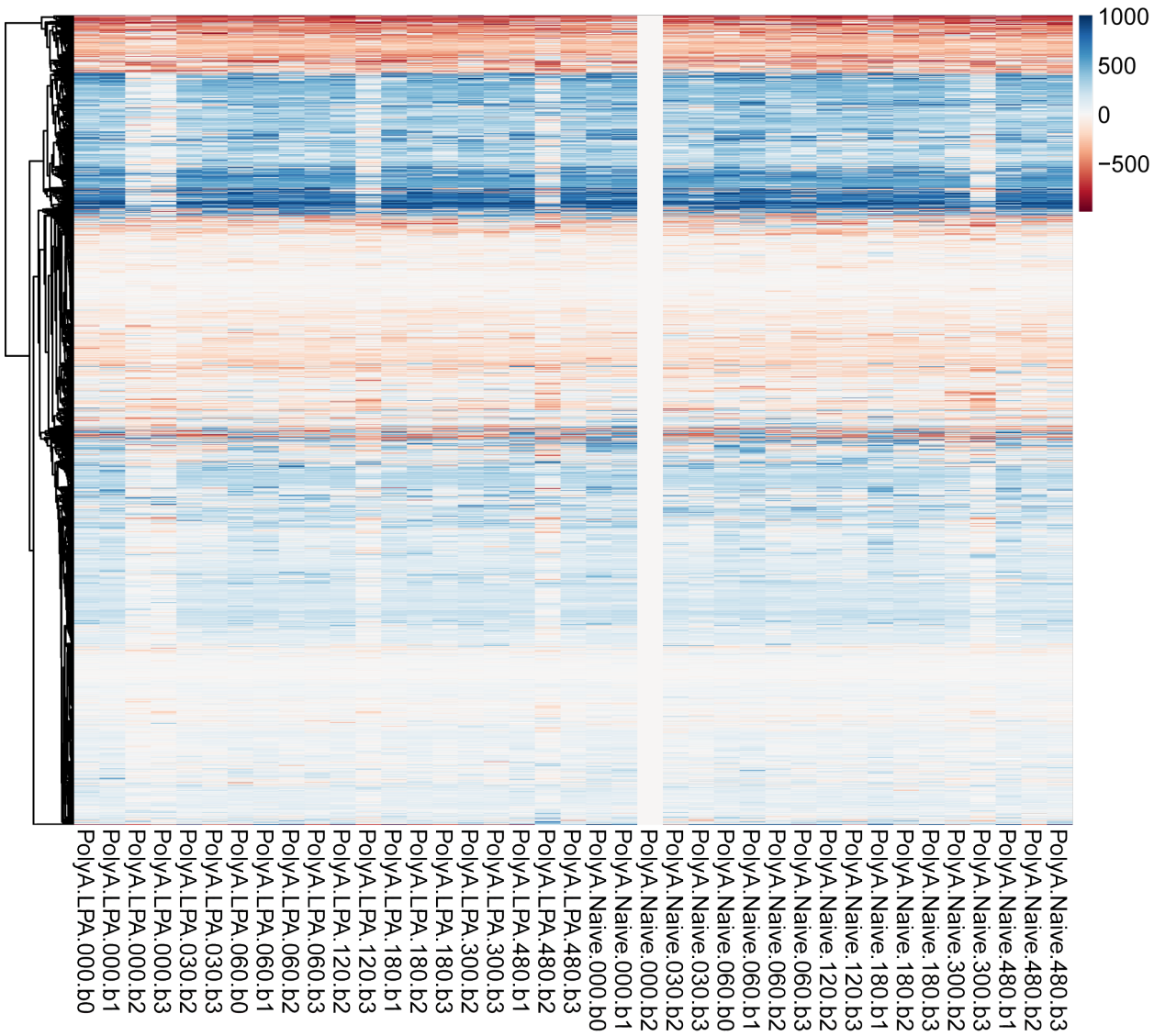


Figure 5.S2 | Poly(A)+ sample clustering by relative weighted 3' length.
Clustering of the weighted 3' length relative to PolyA.Naïve.000.b2 for 2,673 TUs with change points identified, at most 5 missing samples per gene, and abs(rowMax) ≤ 1000 for visualization purposes. Rows are scaled.

Chapter 6: Conclusions

As the cost of high throughput sequencing decreases, more datasets and new high throughput experimental protocols continue to be developed. The development of bioinformatic tools to not only analyze new types of data but also integrate different types of data together is imperative to understand biology. In this dissertation, this we described computational pipelines developed for analyzing different types of RNA-Seq to understand post-transcriptional mechanisms of gene regulation.

In chapter two, we reported hundreds of sciRNAs and their cognate targets through integrative analysis of small RNA-Seq and Degradome-Seq. By using Degradome-Seq as a functional readout of RNA cleavage and developing an analysis pipeline to consider unannotated small RNAs, we were able to identify many novel sciRNAs. Additionally, we found that sciRNAs primarily target retrotransposons, suggesting that retrotransposons may be leveraged as signals for small-RNA mediated cleavage. This work illustrates how novel biological insights can be drawn through integrative analysis of different types of high throughput sequencing.

In chapter three, we developed a novel algorithm, mountainClimber, for change point identification in RNA-Seq based on a notion of measuring signal non-uniformity. It outperformed an existing method and overcame several limitations of other existing approaches. Importantly, it simultaneously identifies alternative transcription start sites (ATSS) and alternative polyadenylation (APA) while other methods ignore ATSS prediction.

The mountainClimber pipeline was applied to GTEx RNA-Seq to identify ATSS and APA in human tissues in chapter four. Because mountainClimber was applied to single samples prior to testing across two biological conditions, we were able to estimate the contribution of individual and tissue to the observed ATSS and APA variability. Additionally, we estimated the length of each 5' and 3' end in each individual sample and demonstrated tissue-specific differences in

global 5’ and 3’ lengths over all genes. Genes with shorter or longer 5’ and 3’ ends in different tissues were enriched in different functional categories, suggesting that ATSS and APA functionally regulate different gene sets in different biological contexts. These analyses of single samples are not possible with other existing approaches that require two biological conditions to be defined a priori in order to identify ATSS and APA. From our analysis of differential change points across pairwise tissues, we estimated 70% and 56% of genes utilize ATSS and APA in different tissues, consistent with previous reports. In summary, we reported the largest characterization of 5’ and 3’ ends in human tissues.

In chapter five, we demonstrated the applicability of the mountainClimber pipeline to different types of RNA-Seq due to its inherent robustness to RNA-Seq non-uniformity. By predicting ATSS and alternative transcription termination in chromatin-associated RNA as well as ATSS and APA in poly(A)-selected RNA, we provided a more complete picture of RNA processing from the chromatin to the mature RNA product. For the first time, we identified several alternative 5’ and 3’ ends in macrophages tolerized to toxin exposure compared to naïve cells. While APA is known to be regulated by trans factors, the role of transcription termination in APA was not previously well understood. We showed that transcription typically terminates beyond the distal poly(A) site, suggesting that transcription is not a primary factor in proximal poly(A) site definition.

In summary, our computational pipelines were used to expand the known repertoire of small cleavage-inducing RNAs, alternative transcription start sites, and alternative polyadenylation sites, thereby providing a more complete picture of post-transcriptional gene regulation. Although previous methods existed for these types of analyses, we overcame several limitations and were able to draw novel insights. While we generated new sequencing datasets for parts of these studies, it should also be noted that our methods were able to draw novel insights from existing publicly available datasets as well. As more datasets become available, their re-analysis by novel bioinformatic algorithms will enrich our understanding of biology.

References

1. Consortium, H. G. S. Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).

2. Venter, J. C. *et al.* The sequence of the human genome. *Science* **291,** 1304–51 (2001).

3. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74 (2012).

4. Guttman, M. & Rinn, J. L. Modular regulatory principles of large non-coding RNAs. *Nature* **482,** 339–46 (2012).

5. Chen, L. The biogenesis and emerging roles of circular RNAs. *Nat. Rev. Mol. Cell Biol.* 15–17 (2016). doi:10.1038/nrm.2015.32

6. Skalska, L., Beltran-Nebot, M., Ule, J. & Jenner, R. G. Regulatory feedback from nascent RNA to chromatin and transcription. *Nat. Rev. Mol. Cell Biol.* (2017). doi:10.1038/nrm.2017.12

7. Hinnebusch, A. G., Ivanov, I. P. & Sonenberg, N. Translational control by 5'-untranslated regions of eukaryotic mRNAs. *Science* **352,** 1413–6 (2016).

8. Porrua, O. & Libri, D. Transcription termination and the control of the transcriptome: why, where and how to stop. *Nat. Rev. Mol. Cell Biol.* **16,** 190–202 (2015).

9. Tian, B. & Manley, J. L. Alternative polyadenylation of mRNA precursors. *Nat. Rev. Mol. Cell Biol.* **18,** 18–30 (2017).

10. Elkon, R., Ugalde, A. P. & Agami, R. Alternative cleavage and polyadenylation: extent, regulation and function. *Nat. Rev. Genet.* **14,** 496–506 (2013).

11. Yoon, O. K., Hsu, T. Y., Im, J. H. & Brem, R. B. Genetics and regulatory impact of alternative polyadenylation in human B-lymphoblastoid cells. *PLoS Genet.* **8,** e1002882 (2012).

12. Mu, X. J., Lu, Z. J., Kong, Y., Lam, H. Y. K. & Gerstein, M. B. Analysis of genomic variation in non-coding elements using population-scale sequencing data from the 1000 Genomes Project. *Nucleic Acids Res.* **39,** 7058–76 (2011).

13. Friedman, R. C., Farh, K. K.-H., Burge, C. B. & Bartel, D. P. Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res.* **19,** 92–105 (2009).

14. Liu, J. *et al.* Argonaute2 is the catalytic engine of mammalian RNAi. *Science (80-.).* **305,** 1437–41 (2004).

15. Bartel, D. P. MicroRNAs: target recognition and regulatory functions. *Cell* **136,** 215–33 (2009).

16. Lewis, B. P., Shih, I., Jones-Rhoades, M. W., Bartel, D. P. & Burge, C. B. Prediction of mammalian microRNA targets. *Cell* **115,** 787–98 (2003).

17. Rhoades, M. W. *et al.* Prediction of plant microRNA targets. *Cell* **110,** 513–20 (2002).

18. Yekta, S., Shih, I.-H. & Bartel, D. P. MicroRNA-directed cleavage of HOXB8 mRNA. *Science (80-.).* **304,** 594–6 (2004).

19. Davis, E. *et al.* RNAi-mediated allelic trans-interaction at the imprinted Rtl1/Peg11 locus. *Curr. Biol.* **15,** 743–9 (2005).

20. Bracken, C. P. *et al.* Global analysis of the mammalian RNA degradome reveals widespread miRNA-dependent and miRNA-independent endonucleolytic cleavage. *Nucleic Acids Res.* **39,** 5658–68 (2011).

21. Shin, C. *et al.* Expanding the microRNA targeting code: functional sites with centered pairing. *Mol. Cell* **38,** 789–802 (2010).

22. Karginov, F. V *et al.* Diverse endonucleolytic cleavage sites in the mammalian transcriptome depend upon microRNAs, Drosha, and additional nucleases. *Mol. Cell* **38,** 781–8 (2010).

23. Okamura, K. & Lai, E. C. Endogenous small interfering RNAs in animals. *Nat. Rev. Mol. Cell Biol.* **9,** 673–8 (2008).

24. German, M. A. *et al.* Global identification of microRNA-target RNA pairs by parallel analysis of RNA ends. *Nat. Biotechnol.* **26,** 941–6 (2008).

25. Addo-Quaye, C., Eshoo, T. W., Bartel, D. P. & Axtell, M. J. Endogenous siRNA and miRNA Targets Identified by Sequencing of the Arabidopsis Degradome. *Curr. Biol.* **18,** 758–762 (2008).

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

27. Toedling, J. *et al.* Deep-sequencing protocols influence the results obtained in small-RNA sequencing. *PLoS One* **7,** e32724 (2012).

28. Xie, C. *et al.* NONCODEv4: exploring the world of long non-coding RNA genes. *Nucleic Acids Res.* **42,** D98-103 (2014).

29. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10,** R25 (2009).

30. Enright, A. J. *et al.* MicroRNA targets in Drosophila. *Genome Biol.* **5,** R1 (2003).

31. Elbashir, S. M., Martinez, J., Patkaniowska, A., Lendeckel, W. & Tuschl, T. Functional anatomy of siRNAs for mediating efficient RNAi in Drosophila melanogaster embryo lysate. *EMBO J.* **20,** 6877–88 (2001).

32. Schwartz, S., Oren, R. & Ast, G. Detection and removal of biases in the analysis of next-

generation sequencing reads. *PLoS One* **6,** e16685 (2011).

33. Babiarz, J. E., Ruby, J. G., Wang, Y., Bartel, D. P. & Blelloch, R. Mouse ES cells express endogenous shRNAs, siRNAs, and other Microprocessor-independent, Dicer-dependent small RNAs. *Genes Dev.* **22,** 2773–85 (2008).

34. Elbashir, S. M., Lendeckel, W. & Tuschl, T. RNA interference is mediated by 21- and 22-nucleotide RNAs. *Genes Dev.* **15,** 188–200 (2001).

35. Leung, A. K. L. *et al.* Genome-wide identification of Ago2 binding sites from mouse embryonic stem cells with and without mature microRNAs. *Nat. Struct. Mol. Biol.* **18,** 237–44 (2011).

36. Lee, Y. S., Shibata, Y., Malhotra, A. & Dutta, A. A novel class of small RNAs: tRNA-derived RNA fragments (tRFs). *Genes Dev.* **23,** 2639–49 (2009).

37. Li, Z. *et al.* Extensive terminal and asymmetric processing of small RNAs from rRNAs, snoRNAs, snRNAs, and tRNAs. *Nucleic Acids Res.* **40,** 6787–99 (2012).

38. Phillips, B. T., Gassei, K. & Orwig, K. E. Spermatogonial stem cell regulation and spermatogenesis. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **365,** 1663–78 (2010).

39. Bartel, D. MicroRNAs: Genomics, Biogenesis, Mechanism, and Function Genomics. *Cell* **116,** 281–297 (2004).

40. Zhang, H.-M. *et al.* AnimalTFDB: a comprehensive animal transcription factor database. *Nucleic Acids Res.* **40,** D144-9 (2012).

41. Almeida, L. G. *et al.* CTdatabase: A knowledge-base of high-throughput and curated data on cancer-testis antigens. *Nucleic Acids Res.* **37,** D816-9 (2009).

42. Jurka, J. *et al.* Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* **110,** 462–7 (2005).

43. Schmitz, J. SINEs as driving forces in genome evolution. *Genome Dyn.* **7,** 92–107 (2012).

44. Lehnert, S. *et al.* Evidence for co-evolution between human microRNAs and Alu-repeats. *PLoS One* **4,** e4456 (2009).

45. Smalheiser, N. R. & Torvik, V. I. Alu elements within human mRNAs are probable microRNA targets. *Trends Genet.* **22,** 532–6 (2006).

46. Hoffman, Y., Dahary, D., Bublik, D. R., Oren, M. & Pilpel, Y. The majority of endogenous microRNA targets within Alu elements avoid the microRNA machinery. *Bioinformatics* **29,** 894–902 (2013).

47. Betel, D., Wilson, M., Gabow, A., Marks, D. S. & Sander, C. The microRNA.org resource: targets and expression. *Nucleic Acids Res.* **36,** D149-53 (2008).

48. Chalopin, D., Naville, M., Plard, F., Galiana, D. & Volff, J.-N. Comparative analysis of transposable elements highlights mobilome diversity and evolution in vertebrates. *Genome*

Biol. Evol. **7,** 567–80 (2015).

49. Sela, N. *et al.* Comparative analysis of transposed element insertion within human and mouse genomes reveals Alu's unique role in shaping the human transcriptome. *Genome Biol.* **8,** R127 (2007).

50. Umylny, B., Presting, G., Efird, J. T., Klimovitsky, B. I. & Ward, W. S. Most human Alu and murine B1 repeats are unique. *J. Cell. Biochem.* **102,** 110–121 (2007).

51. Hou, P. *et al.* Pluripotent Stem Cells Induced from Mouse Somatic Cells by Small-Molecule Compounds. *Sci. (New York, NY)* **341,** 651–654 (2013).

52. Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14,** R36 (2013).

53. Kozomara, A. & Griffiths-Jones, S. MiRBase: Annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res.* **42,** 68–73 (2014).

54. Burge, S. W. *et al.* Rfam 11.0: 10 years of RNA families. *Nucleic Acids Res.* **41,** 1–7 (2013).

55. Sai lakshmi, S. & Agrawal, S. piRNABank: A web resource on classified and clustered Piwi-interacting RNAs. *Nucleic Acids Res.* **36,** 173–177 (2008).

56. Hofacker, I. L. *et al.* Fast folding and comparison of RNA secondary structures. *Monatshefte fur Chemie Chem. Mon.* **125,** 167–188 (1994).

57. Lee, J.-H. *et al.* Analysis of transcriptome complexity through RNA sequencing in normal and failing murine hearts. *Circ. Res.* **109,** 1332–41 (2011).

58. Sherry, S. T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29,** 308–311 (2001).

59. Mapendano, C. K., Lykke-Andersen, S., Kjems, J., Bertrand, E. & Jensen, T. H. Crosstalk between mRNA 3' end processing and transcription initiation. *Mol. Cell* **40,** 410–22 (2010).

60. Leppek, K., Das, R. & Barna, M. Functional 5′ UTR mRNA structures in eukaryotic translation regulation and how to find them. *Nat. Rev. Mol. Cell Biol.* **19,** 158–174 (2017).

61. Baek, D., Davis, C., Ewing, B., Gordon, D. & Green, P. Characterization and predictive discovery of evolutionarily conserved mammalian alternative promoters. *Genome Res.* **17,** 145–55 (2007).

62. Carninci, P. *et al.* The transcriptional landscape of the mammalian genome. *Science* **309,** 1559–63 (2005).

63. Hon, C.-C. *et al.* An atlas of human long non-coding RNAs with accurate 5' ends. *Nature* **543,** 199–204 (2017).

64. Derti, A. *et al.* A quantitative atlas of polyadenylation in five mammals. *Genome Res.* **22,** 1173–83 (2012).

65. Hoque, M. *et al.* Analysis of alternative cleavage and polyadenylation by 3' region extraction and deep sequencing. *Nat. Methods* **10,** 133–9 (2013).

66. 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.* **27,** 2380–96 (2013).

67. Lu, J. & Bushel, P. R. Dynamic expression of 3' UTRs revealed by Poisson hidden Markov modeling of RNA-Seq: Implications in gene expression profiling. *Gene* **527,** 616–623 (2013).

68. Wang, W., Wei, Z. & Li, H. A change-point model for identifying 3'UTR switching by next-generation RNA sequencing. *Bioinformatics* **30,** 2162–70 (2014).

69. Xia, Z. *et al.* Dynamic analyses of alternative polyadenylation from RNA-seq reveal a 3'-UTR landscape across seven tumour types. *Nat. Commun.* **5,** 5274 (2014).

70. Shenker, S., Miura, P., Sanfilippo, P. & Lai, E. C. IsoSCM: improved and alternative 3' UTR annotation using multiple change-point inference. *RNA* **21,** 14–27 (2015).

71. Ye, C., Long, Y., Ji, G., Li, Q. Q. & Wu, X. APAtrap: identification and quantification of alternative polyadenylation sites from RNA-seq data. *Bioinformatics* **33,** 2699–2705 (2018).

72. Kim, M., You, B. H. & Nam, J. W. Global estimation of the 3' untranslated region landscape using RNA sequencing. *Methods* (2015). doi:10.1016/j.ymeth.2015.04.011

73. Arefeen, A., Liu, J., Xiao, X. & Jiang, T. TAPAS: tool for alternative polyadenylation site analysis. *Bioinformatics* **34,** 2521–2529 (2018).

74. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26,** 841–2 (2010).

75. Dale, R. K., Pedersen, B. S. & Quinlan, A. R. Pybedtools: a flexible Python library for manipulating genomic datasets and annotations. *Bioinformatics* **27,** 3423–4 (2011).

76. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12,** 2825–2830 (2011).

77. Wickham, H. ggplot2: Elegant Graphic for Data Analysis. *Springer* 1–210 (2009). doi:10.1007/978-0-387-98141-3

78. Wickham, H. Reshaping Data with the **reshape** Package. *J. Stat. Softw.* **21,** (2007).

79. Law, C. W., Chen, Y., Shi, W. & Smyth, G. K. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.* **15,** R29 (2014).

80. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12,** 323 (2011).

81. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29,** 15–21 (2013).

82. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12,** 357–60 (2015).

83. Griebel, T. *et al.* Modelling and simulating generic RNA-Seq experiments with the flux simulator. *Nucleic Acids Res.* **40,** 10073–83 (2012).

84. Katz, Y., Wang, E., Airoldi, E. & Burge, C. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat. Methods* **7,** (2010).

85. Zhang, H., Hu, J., Recce, M. & Tian, B. PolyA_DB: A database for mammalian mRNA polyadenylation. *Nucleic Acids Res.* **33,** 116–120 (2005).

86. Rapaport, F. *et al.* Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome Biol.* **14,** R95 (2013).

87. PAGE, E. S. Continuous Inspection Schemes. *Biometrika* **41,** 100–115 (1954).

88. Bhatt, D. M. *et al.* Transcript dynamics of proinflammatory genes revealed by sequence analysis of subcellular RNA fractions. *Cell* **150,** 279–90 (2012).

89. Shen, S. *et al.* MATS: a Bayesian framework for flexible detection of differential alternative splicing from RNA-Seq data. *Nucleic Acids Res.* **40,** e61 (2012).

90. Li, Y. I. *et al.* Annotation-free quantification of RNA splicing using LeafCutter. *Nat. Genet.* **50,** 151–158 (2018).

91. Anvar, S. Y. *et al.* Full-length mRNA sequencing uncovers a widespread coupling between transcription initiation and mRNA processing. *Genome Biol.* **19,** 46 (2018).

92. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28,** 511–515 (2010).

93. Guttman, M. *et al.* Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat. Biotechnol.* **28,** 503–510 (2010).

94. Berkovits, B. D. & Mayr, C. Alternative 3' UTRs act as scaffolds to regulate membrane protein localization. *Nature* **522,** 363–7 (2015).

95. Neve, J. *et al.* Subcellular RNA profiling links splicing and nuclear DICER1 to alternative cleavage and polyadenylation. *Genome Res.* **26,** 24–35 (2016).

96. Singh, I. *et al.* Widespread intronic polyadenylation diversifies immune cell transcriptomes. *Nat. Commun.* **9,** 1716 (2018).

97. Ji, Z. *et al.* Transcriptional activity regulates alternative cleavage and polyadenylation. *Mol. Syst. Biol.* **7,** 534 (2011).

98. Reyes, A. & Huber, W. Alternative start and termination sites of transcription drive most transcript isoform differences across human tissues. *Nucleic Acids Res.* 1–11 (2017).

doi:10.1093/nar/gkx1165

99. Cheng, Z. *et al.* Pervasive, Coordinated Protein-Level Changes Driven by Transcript Isoform Switching during Meiosis. *Cell* **172,** 910–923.e16 (2018).

100. GTEx Consortium. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348,** 648–60 (2015).

101. Wang, L. *et al.* Measure transcript integrity using RNA-seq data. *BMC Bioinformatics* **17,** 58 (2016).

102. Ferreira, P. G. *et al.* The effects of death and post-mortem cold ischemia on human tissue transcriptomes. *Nat. Commun.* **9,** 490 (2018).

103. Melé, M. *et al.* The human transcriptome across tissues and individuals. *Science* **348,** 660–5 (2015).

104. Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting Linear Mixed-Effects Models Using lme4. *J. Stat. Softw.* **67,** (2015).

105. Chen, Y. *et al.* Principles for RNA metabolism and alternative transcription initiation within closely spaced promoters. *Nat. Genet.* **48,** 984–94 (2016).

106. FANTOM Consortium and the RIKEN PMI and CLST (DGT) *et al.* A promoter-level mammalian expression atlas. *Nature* **507,** 462–70 (2014).

107. Tan, M. H. *et al.* Dynamic landscape and regulation of RNA editing in mammals. *Nature* **550,** 249–254 (2017).

108. Baek, D., Davis, C., Ewing, B., Gordon, D. & Green, P. Characterization and predictive discovery of evolutionarily conserved mammalian alternative promoters. *Genome Res.* **17,** 145–155 (2007).

109. Gunes, S., Al-Sadaan, M. & Agarwal, A. Spermatogenesis, DNA damage and DNA repair mechanisms in male infertility. *Reprod. Biomed. Online* **31,** 309–19 (2015).

110. Oktaba, K. *et al.* ELAV links paused Pol II to alternative polyadenylation in the Drosophila nervous system. *Mol. Cell* **57,** 341–8 (2015).

111. Hotchkiss, R. S., Monneret, G. & Payen, D. Sepsis-induced immunosuppression: From cellular dysfunctions to immunotherapy. *Nature Reviews Immunology* **13,** 862–874 (2013).

112. Mages, J., Dietrich, H. & Lang, R. A genome-wide analysis of LPS tolerance in macrophages. *Immunobiology* **212,** 723–737 (2008).

113. Saeed, S. *et al.* Epigenetic programming of monocyte-to-macrophage differentiation and trained innate immunity. *Science* **345,** 1251086 (2014).

114. Seeley, J. J. *et al.* Induction of innate immune memory via microRNA targeting of chromatin remodelling factors. *Nature* **559,** 114–119 (2018).

115. Pai, A. A. *et al.* Widespread Shortening of 3' Untranslated Regions and Increased Exon Inclusion Are Evolutionarily Conserved Features of Innate Immune Responses to Infection. *PLoS Genet.* **12,** e1006338 (2016).

116. Jia, X. *et al.* The role of alternative polyadenylation in the antiviral innate immune response. *Nat. Commun.* **8,** 14605 (2017).

117. Alasoo, K. *et al.* Transcriptional profiling of macrophages derived from monocytes and iPS cells identifies a conserved response to LPS and novel alternative transcription. *Sci. Rep.* **5,** 12524 (2015).

118. Vilborg, A. *et al.* Comparative analysis reveals genomic features of stress-induced transcriptional readthrough. *Proc. Natl. Acad. Sci. U. S. A.* **114,** E8362–E8371 (2017).

119. Pandya-Jones, A. *et al.* Splicing kinetics and transcript release from the chromatin compartment limit the rate of Lipid A-induced gene expression. *RNA* **19,** 811–27 (2013).

120. Lacy, P. & Stow, J. L. Cytokine release from innate immune cells: association with diverse membrane trafficking pathways. *Blood* **118,** 9–18 (2011).

121. Novoa, I., Gallego, J., Ferreira, P. G. & Mendez, R. Mitotic cell-cycle progression is regulated by CPEB1 and CPEB4-dependent translational control. *Nat. Cell Biol.* **12,** 447–56 (2010).

122. Yamagishi, R., Tsusaka, T., Mitsunaga, H., Maehata, T. & Hoshino, S. The STAR protein QKI-7 recruits PAPD4 to regulate post-transcriptional polyadenylation of target mRNAs. *Nucleic Acids Res.* gkw118 (2016). doi:10.1093/nar/gkw118

123. Herzel, L., Straube, K. & Neugebauer, K. M. Long-read sequencing of nascent RNA reveals coupling among RNA processing events. *Genome Res.* 1–43 (2018). doi:10.1101/gr.232025.117