HHS Public Access

Author manuscript

Cell Rep Methods. Author manuscript; available in PMC 2021 October 07.

Published in final edited form as:

Cell Rep Methods. 2021 September 27; 1(5): . doi:10.1016/j.crmeth.2021.100081.

Global approaches for profiling transcription initiation

Robert A. Policastro[1,3], **Gabriel E. Zentner**[1,2,3,*]

[1]Department of Biology, Indiana University, Bloomington, IN 47401, USA

[2]Indiana University Melvin and Bren Simon Comprehensive Cancer Center, Indianapolis, IN 46202, USA

[3]Present address: eGenesis Inc, Cambridge, MA 02139, USA

SUMMARY

Transcription start site (TSS) selection influences transcript stability and translation as well as protein sequence. Alternative TSS usage is pervasive in organismal development, is a major contributor to transcript isoform diversity in humans, and is frequently observed in human diseases including cancer. In this review, we discuss the breadth of techniques that have been used to globally profile TSSs and the resulting insights into gene regulation, as well as future prospects in this area of inquiry.

INTRODUCTION

The first base of a gene to be transcribed by an RNA polymerase, corresponding to the 5′-most base of the resulting transcript, is referred to as the transcription start site (TSS). Within a given gene promoter, there is generally not a single TSS, but rather a cluster of TSSs, referred to as a transcription start region (TSR). Furthermore, a gene might have multiple TSRs interspersed throughout the locus, indicating the presence of alternative promoters. The phenomenon of alternative transcription initiation is widespread in biology. For instance, several studies have described large-scale shifts in patterns of initiation during development (Batut et al., 2013; Danks et al., 2018; Zhang et al., 2017). This is perhaps most strikingly apparent in zebra-fish, wherein the maternal and zygotic forms of more than 900 transcripts display differential TSS usage (Haberle et al., 2014), a phenomenon that appears to be conserved in mice (Cvetesic et al., 2020). Alternative promoter usage has also been reported to be widespread in human cancers, and use of alternative promoters is predictive of patient survival in some cases (Demircioğlu et al., 2019).

Broadly speaking, TSS shifting impacts gene regulation by altering the length of 5′ transcript leaders (5′ TLs). The 5′ TLs have been shown to play a large role in modulating both the stability and translation of mRNAs (Figure 1A) (Arribere and Gilbert, 2013;

*Correspondence: gabe.zentner@egenesisbio.com.

DECLARATION OF INTERESTS

R.A.P. and G.E.Z. are employees of eGenesis, Inc. The topics covered in this review are not related to any current work at the company.

Dieudonné et al., 2015; Malabat et al., 2015; Rojas-Duran and Gilbert, 2012; Wang et al., 2016). In many cases, 5′ TLs encode upstream open reading frames (uORFs), which are short peptide-coding regions upstream of the primary ORF of a given transcript. uORFs might repress translation by preventing ribosomes from reaching the start codon of the primary ORF and might also lead to nonsense-mediated decay if the uORF stop codon is recognized as premature (Barbosa et al., 2013). A striking example of counteraction of uORF-mediated repression is observed in *Arabidopsis thaliana*, wherein exposure of seedlings to blue light causes downstream shifts in the TSS usage of 220 uORF-containing genes (Kurihara et al., 2018). Secondary structures within the 5′ TL, such as G-quadruplexes and pseudo-knots, can also influence transcript stability and translation, and internal ribosome entry sites (IRESs) can promote cap-independent translation (Leppek et al., 2018).

In addition to altering transcript regulation via 5′ TL length modulation, alternative TSS selection can give rise to transcript isoforms with distinct protein-coding potential. Indeed, it is estimated that the contributions of alternative promoters and transcription termination sites to transcript isoform diversity exceeds that of alternative splicing in multiple contexts, including normal human tissues (Reyes and Huber, 2018; Shabalina et al., 2014). Studies in plants have provided notable examples of alternative TSS selection giving rise to distinct transcript and protein isoforms. In maize, transcription of the *mybr35* gene, encoding a MYB-family transcription factor, initiates from downstream TSSs in shoots, which gives rise to a protein lacking an N-terminal zinc finger domain. In contrast, upstream TSSs are used in roots, generating a full-length protein (Mejía-Guerra et al., 2015). In *Arabidopsis*, exposure of seedlings to red light globally alters TSS usage, leading to the production of N-terminally truncated protein isoforms with distinct subcellular localizations (Ushijima et al., 2017).

Although much work has been done on the impact of TSS selection on the regulation of the resulting transcript, it is becoming increasingly clear that transcription itself can inhibit TSS usage (Gowthaman et al., 2020). This phenomenon, termed transcriptional interference, refers to suppression in *cis* of transcription from one transcript unit by the act of RNAPII transcription of a second overlapping transcription unit. For instance, initiation from upstream TSSs might inhibit the use of downstream TSSs, a process termed tandem transcriptional interference (tTI). tTI might be induced by transcription of an mRNA isoform with an extended 5′ TL (Chen et al., 2017; Chia et al., 2017; Hollerer et al., 2019; Jorgensen et al., 2020; Nielsen et al., 2019) or an upstream noncoding RNA (Lin et al., 2018). In some cases, the long mRNA isoform might contain a uORF in its 5′ TL, downregulating translation (Chen et al., 2017; Chia et al., 2017; Hollerer et al., 2019), leading to a model of integrated transcriptional and translational repression.

Because of the importance of TSS selection in numerous biological contexts, a large number of methods for global TSS profiling have been developed. In this review, we describe the general enzymatic approaches used for this purpose, and the specific techniques that have used them. We also lay out computational strategies for TSS mapping data analysis. Last, we discuss current challenges and future prospects for the field, with a particular focus on single-cell mapping of heterogeneity in TSS usage.

MOLECULAR APPROACHES TO GLOBAL TSS MAPPING

Oligo-capping

Oligo-capping was originally developed to facilitate recovery of 5′-complete cDNAs (Kazuo and Sumio, 1994; Suzuki and Sugano, 2003). Oligo-capping involves enzymatic removal of the 7-methylguanosine (m7G) cap of mRNAs (and other RNA species such as long non-coding RNAs, pre-micro RNAs, and enhancer RNAs) followed by ligation of a synthetic oligonucleotide (Figure 2A). In practice, the original oligo-capping protocol first uses Bacterial Alkaline Phosphatase (BAP) to hydrolyze the 5′ phosphates of uncapped RNA molecules, preventing subsequent adapter ligation. Tobacco Acid Pyrophosphatase (TAP) is then used to remove m7G caps, leaving a 5′ mono-phosphate suitable for adapter ligation. Prior to the widespread adoption of high-throughput sequencing, oligo-capping was used for an iteration of 5′ Serial Analysis of Gene Expression (5′ SAGE) (Hashimoto et al., 2004), wherein transcript 5′ sequences (5′ SAGE tags) are concatemerized, cloned, and sequenced. Oligo-capping was then adapted to high-throughput sequencing by ligation of 5′ tags to Solexa sequencing adapters and sequencing on the Illumina GA platform (Tsuchihara et al., 2009; Wakaguri et al., 2008), an approach later termed TSS-seq (Yamashita et al., 2011). Since the original implementation of TSS-seq, numerous TSS mapping methods have employed oligo-capping. Firstly, the Paired-End Analysis of TSSs (PEAT) approach (Ni et al., 2010) adapted oligo-capping for paired-end sequencing. Other oligo-capping approaches varied the enzymes used for RNA 5′ end processing. CapSeq (Gu et al., 2012), Transcript Leader sequencing (TL-seq) (Arribere and Gilbert, 2013), and Transcript IsoForm sequencing (TIF-seq) (Pelechano et al., 2013) used Calf Intestinal alkaline Phosphatase (CIP) instead of BAP for dephosphorylation of uncapped RNA, given that it can be heat inactivated, whereas Start-seq employed RNA 5′ Polyphosphatase (Nechaev et al., 2010). Start-seq and CapSeq also added a Terminator 5′ phosphate-dependent exonuclease (TEX) treatment to further reduce uncapped RNA (predominantly rRNA) levels, whereas TIF-seq provides both 5′and 3′ transcript end sequences thanks to a circular ligation step in the protocol. Simultaneous Mapping of RNA Ends by sequencing (SMORE-seq) enables 5′and 3′ transcript end mapping through sequential adapter ligation and also omits the phosphatase treatment prior to decapping, enabling capture of RNA degradation intermediates (Park et al., 2014).

One note for oligo-capping methods is that TAP is no longer commercially available, so alternative enzymes are required. The *E.coli* RNA 5′ pyrophosphohydrolase RppH has been successfully used for decapping in Start-seq (Scheidegger et al., 2019), and a recombinant fusion of the *Schizosaccharomyces pombe* Dcp1-Dcp2 to its activator Edc1 has also shown promise in this regard (Paquette et al., 2018). Oligo-capping methods also tend to have high input requirements. For instance, it was reported that 30 μg of total RNA was necessary to construct an *Arabidopsis* PEAT library (Morton et al., 2014). Lastly, oligo-capping methods might suffer from sequence and/or structure biases in the RNA ligases used to add adapters (Baldrich et al., 2020; Fuchs et al., 2015; Hafner et al., 2011; Jayaprakash et al., 2011).

Cap-trapping

In addition to oligo-capping, early efforts to generate libraries of 5′-complete cDNAs led to the development of the cap-trapping approach (Carninci et al., 1996), in which the m7G cap is oxidized and biotinylated to allow streptavidin purification of 5′-complete cDNAs after reverse transcription (Figure 2B). In addition to its use in the generation of large cDNA libraries (Kawai et al., 2001; Okazaki et al., 2002), cap-trapping serves as the basis of Cap Analysis of Gene Expression (CAGE), perhaps the most widely known TSS mapping method. In the initial iteration of CAGE (Kodzius et al., 2006; Shiraki et al., 2003), reverse transcription and cap-trapping are performed, followed by ligation of a 5′ linker containing XmaJI and MmeI restriction sites. After second-strand synthesis, the double-stranded cDNA is digested with MmeI. MmeI is a Type IIs restriction enzyme that cuts 20 and 18 nucleotides downstream of the 3′ end of its recognition site, generating an asymmetric overhang to which a second adapter containing an XmaJI site can be ligated. XmaJI cleavage then releases the ligated cDNA fragments (referred to as CAGE tags), which are concatemerized, cloned, and sequenced by using the RIKEN Integrated Sequence Analysis (RISA) system (Shibata et al., 2000). A further iteration of CAGE, DeepCAGE, adapted the method to the 454 Life Sciences GS20 pyrosequencer (Balwierz et al., 2009; Valen et al., 2009). The first no-amplification version of CAGE, HeliScopeCAGE, greatly reduced input requirements and removed PCR biases by directly sequencing first-strand cDNA with the HeliScope Genetic Analysis System (Kanamori-Katayama et al., 2011). Switching to the EcoP15I restriction enzyme allowed for longer CAGE tags, facilitating more confident mapping to the genome (Takahashi et al., 2012), and eventually the need for restriction enzymes was removed in no-amplification non-tagging CAGE for Illumina sequencers (nAnT-iCAGE) (Murata et al., 2014). The major limitation of standard CAGE methods has been their high input requirements, though the nAnT-iCAGE protocol has been adapted to nanogram levels of RNA input via the use of capped, selectively degradable carrier RNA as Super-Low-Input Carrier CAGE (SLIC-CAGE) (Cvetesic et al., 2018).

Whereas CAGE relies on chemical modification of the cap structure to facilitate isolation of 5′-complete cDNAs, the recently published Multiplexed Affinity Purification of Capped RNA (MAPCap) method instead uses an antibody against m7G to isolate capped RNA prior to reverse transcription (Bhardwaj et al., 2019). Notably, MAPCap was reported to work well with as little as 100 ng RNA, suggesting that it is suitable for TSS mapping in precious samples. Natural proteins could also be used to isolate capped RNA for TSS mapping: a high-affinity mutant of the eukaryotic translation initiation factor 4E (eIF4E) (Choi and Hagedorn, 2003) has successfully been used to isolate mRNA (Blower et al., 2013) and nascent RNA (Matveeva et al., 2019) for sequencing.

Template-switching reverse transcription

Much like oligo-capping and cap-trapping, template-switching reverse transcription (TSRT) was initially leveraged as a means to generate 5′-complete cDNA molecules (Schmidt and Mueller, 1999; Zhu et al., 2001). TSRT leverages the propensity of Moloney murine leukemia virus (MMLV)-based reverse transcriptases to add 1–3 nontemplated bases to the 5′ end of a cDNA molecule. The prevailing view is that TSRT adds 1–3 predominantly cytosine residues, which can then serve as a handle for annealing of a template-switching

oligo (TSO) with a three-riboguanosine overhang, allowing direct incorporation of a sequencing adapter at the 5′ end of the first-strand cDNA (Figure 2C). Interestingly, recent work indicates that the CCC overhang is a relatively rare result of TSRT (Wulf et al., 2019). Given that TSOs with riboguanosine overhangs have long been effective in TSRT despite the apparent paucity of CCC addition, it stands to reason that the mechanism of template switching does not necessarily rely on base pairing between templates.

Prior to the widespread adoption of high-throughput sequencing, TSRT was used to profile budding yeast TSSs in a modification of 5′ SAGE (Zhang and Dietrich, 2005). TSRT-based TSS mapping was brought into the high-throughput sequencing era via the original iteration of nanoCAGE (hereafter referred to as nanoCAGE [2010]) and CAGEscan (Plessy et al., 2010). After TSRT, nanoCAGE (2010) uses semi-suppressive PCR to reduce the prevalence of small artifactual fragments, followed by EcoP15I cleavage, adapter ligation, and library PCR, similar to first-generation CAGE approaches. CAGEscan simplified the nanoCAGE (2010) protocol by only requiring library PCR after semi-suppressive PCR, and also enabled paired-end sequencing of CAGE tags. CAGEscan, originally designed for the Illumina Genome Analyzer$_{IIX}$ system, was subsequently adapted to the more modern Illumina HiSeq 2000 platform as NanoCAGE-XL (Cumbie et al., 2015). A further iteration of nanoCAGE (hereafter referred to as nanoCAGE [2017]), took advantage of Tn5 tagmentation to reduce input requirements and improve library size distribution, and added an optional TEX treatment step to reduce the prevalence of uncapped RNAs (such as rRNA) in the final libraries (Poulain et al., 2017). RNA Annotation and Mapping of Promoters for the Analysis of Gene Expression (RAMPAGE) combined TSRT with cap-trapping to provide additional specificity for capped transcripts (Batut et al., 2013). TSRT was later adapted to mapping of TSSs at single-cell resolution in methods such as C1-CAGE (Kouno et al., 2019), Tn5Prime (Cole et al., 2018), and low-input Parallel Analysis of RNA Ends (nanoPARE) (Schon et al., 2018). Finally, Survey of TRanscription Initiation and Promoter Elements with high-throughput sequencing (STRIPE-seq) (Policastro et al., 2020) removed the need for semi-suppressive PCR or tagmentation for modest RNA input amounts, and further reduced the prevalence of artifactual reads through improved oligo design and methodological optimizations.

TSRT-based methods are vulnerable to artifactual TSSs arising from a process termed strand invasion, wherein the TSO hybridizes to the first-strand cDNA before polymerization is complete, resulting in an artificially truncated cDNA molecule (Tang et al., 2013). Such artifacts can be reduced by introducing a spacer sequence into the TSO between the ribo-G overhang and the remaining sequence (e.g., TATAGGG in STRIPE-seq). This enables more confident filtering of such artifacts, as there are fewer matches to such a sequence within transcripts versus GGG. TSRT can also result in TSO chaining, wherein once RT reaches the 5′ end of the TSO, it performs another round of non-templated nucleotide addition, allowing another TSO to bind. The prevalence of the resulting TSO concatemers can be reduced by 3′ modification of the TSO with non-natural nucleotides (Kapteyn et al., 2010) or biotin (Turchinovich et al., 2014).

MAPPING TSSs FROM NASCENT RNA

The methods discussed to this point profile the TSSs of stable, mature transcripts. However, in some cases, capture of TSSs from nascent transcripts might be desired. For instance, this is useful for unstable transcripts, which would be underrepresented in the steady-state RNA pool of the cell because of their rapid turnover after synthesis. The first iterations of nascent RNA 5′ end capture involved modifications of the Global Run-On sequencing (GRO-seq) and Precision Run-On sequencing (PRO-seq) protocols. The GRO-seq protocol involves isolation of nuclei, extension of nascent RNA, and simultaneous incorporation of the ribonucleotide analog 5-bromouridine 5′-triphosphate (BrUTP), and pulldown of nascent RNA with an anti-BrUTP antibody. PRO-seq, on the other hand, employs four parallel run-on reactions, each of which contains one of four biotinylated nucleotides, resulting in the incorporation of the biotinylated base and subsequent stalling of transcription, and streptavidin pulldown to capture nascent RNA (Kwak et al., 2013). GRO/PRO-cap (Core et al., 2014; Kwak et al., 2013) are modifications to the GRO/PRO-seq protocols designed to capture TSSs of nascent transcripts wherein, after the run-on reactions, nascent RNA is treated with TEX in the case of GRO-cap, and for both methods the subsequent addition of an oligo-capping strategy to the protocol. CAGE has also been extended to mapping TSSs from nascent RNA as Native Elongating Transcript sequencing and CAGE (NET-CAGE) (Hirabayashi et al., 2019). NET-CAGE combines CAGE with an approach to nascent RNA isolation used in 3′NT (Weber et al., 2014) and one iteration of mammalian NET-seq (Mayer et al., 2015). Both methods take advantage of the extraordinary resistance of the RNAPII:DNA:RNA ternary complex to harsh conditions such as high salt and urea (Cai and Luse, 1987; Wuarin and Schibler, 1994) to isolate chromatin-associated nascent RNAs.

COMPUTATIONAL PROCESSING CONSIDERATIONS

There are a few important points of consideration for preprocessing of TSS mapping data. TSS mapping data require precise alignment of reads to the genome, and for this task the STAR aligner (Dobin et al., 2013) is most often used. Another preprocessing consideration is the presence of PCR duplicates. Although more common in low-input methods such as nanoCAGE and STRIPE-seq, PCR duplicates can lead to inaccurate TSS and TSR quantification in all methods. In some methods (e.g., STRIPE-seq), a random sequence called a unique molecular identifier (UMI) is included in the R1 read, which for single-end sequencing data allows for the computational removal of PCR duplicates, as reads with the same genomic position and UMI are not expected to occur frequently by chance. For this task, UMI-tools (Smith et al., 2017) is recommended because of its ability to correct for PCR and sequencing errors that could lead to incorrect bases called in UMIs. For paired-end data, the read not anchored to the TSS tends to be somewhat randomly positioned because of random priming of RT or tag-mentation. This can be used to remove PCR duplicates by using Samtools (Li et al., 2009), as unique reads would not be expected to have the same R1 and R2 positions by chance.

After PCR duplicate removal, there are two main steps shared by most methods. First, the 5′-most aligned bases should be processed for various artifacts. In TSRT-based methods, it is expected that 1–3 nontemplated bases will be added to the 3′ end of the cDNA. During

alignment, these are generally marked as soft-clipped (that is, they are present in the read but are not part of the read's alignment to the genome). After removal of soft-clipped bases, additional processing is necessary. Reverse transcription of capped RNA in cap-trapping and TSRT-based methods often leads to the addition of a C to the cDNA, likely templated by the m7G cap itself, which results in a G at this position in the R1 sequencing read. If this base does not match the genome during alignment, it will be soft-clipped; however, it is possible (particularly in mammalian genomes, where promoters tend to be GC-rich [Fenouil et al., 2012]) that this cap-templated base will match the genome and it is thus impossible to determine if it represents the true TSS. Both TSRexploreR (Policastro et al., 2021) and CAGEr (Haberle et al., 2015) implement probability-based stochastic removal of mapped 5′ Gs to mitigate this artifact. It is worthwhile to note that analysis of data generated by oligo-capping methods such as TSS-seq do not require G correction, as the native cap is removed prior to reverse transcription.

Data generated by various TSS mapping techniques can also be used to quantify transcript abundance. Thus, an RNA-seq-like analysis of such data can be performed, using software such as HTSeq (Anders et al., 2014) or featureCounts (Liao et al., 2019) to quantify counts at the gene level, or software such as RSEM (Li and Dewey, 2011), Salmon (Patro et al., 2017), or Kallisto (Bray et al., 2016) to estimate counts at the transcript level. For transcript quantification, it is recommended to extend the transcript sequences upstream roughly 100 to 200 base pairs (bp), given that the TSS-containing read will often be fully or partially upstream of the annotated 5′ TL, complicating analysis.

SOFTWARE AND ANALYTICAL CONSIDERATIONS

After preprocessing of TSS mapping data, there exists a wide breadth of tools available to explore and analyze the data (Figure 3). Because of the large number of possible analyses, this section is not intended to be comprehensive, but rather a show-case of important and interesting computational considerations. For more information, refer to the vignettes and papers for CAGEr (Haberle et al., 2015), TSRexploreR (Policastro et al., 2021), CAGEfighteR (Thodberg et al., 2019), and other tools mentioned to learn the full breadth of available analyses.

Calling TSSs from raw data

After processing and alignment of sequencing reads, the first step in analyzing TSS mapping data is aggregating read 5′ ends into TSS positions. Although conceptually simple, all global TSS mapping technologies suffer from spurious background reads over gene bodies, and so some form of thresholding is necessary to remove background TSS signal while retaining true TSSs. Thus, a number of thresholding approaches have been put forth. In CAGEr (Haberle et al., 2015) and CAGEfightR (Thodberg et al., 2019), read 5′ ends are first aggregated into CAGE-detected TSSs without thresholding and normalized (discussed below). In CAGEr, TSS filtering is performed during clustering, wherein the user specifies the number of samples that must have at least n normalized counts at a given TSS position for it to be considered for TSS clustering. Similarly, CAGEfightR allows users to discard TSSs not meeting the above-described sample number and count thresholds prior to

clustering. In TSRexploreR, a genome annotation is used to determine the fraction of TSSs within a specified distance from an annotated TSS, as well as the number of features (genes or transcripts) with at least one unique TSS position (Policastro et al., 2021). We found that the promoter proximal TSS fraction increases markedly as the threshold is increased from a single read, likely indicating loss of weak TSSs within gene bodies. This gain in promoter proximal TSS fraction levels off as the threshold increases, and the number of features with a unique TSS decreases as weaker promoter proximal TSSs are progressively eliminated. We therefore suggest selecting a threshold within the inflection point of the promoter proximal fraction curve to balance removal of likely artifacts with retention of weak true TSSs. We found that, in STRIPE-seq data from yeast and human cells, a threshold of three raw counts per TSS is a suitable threshold for TSS retention (Policastro et al., 2020).

Normalization

An important step in the analysis of TSS mapping data is normalization. CAGEr and CAGEfighteR utilize transcripts per million normalization (Haberle et al., 2015; Thodberg et al., 2019), and CAGEr additionally allows for power-law normalization due to CAGE datasets being previously shown to follow a power-law distribution (Balwierz et al., 2009). TSRexploreR, on the other hand, utilizes either counts per million or DESeq2/edgeR normalization (Policastro et al., 2021). There are two important considerations for TSS normalization: the number of total sequenced reads, and compositional bias. The total number of sequenced reads (or library size) affects read quantification because of variations in counts per feature and the chance for lowly expressed features to drop out at lower sequencing depths. Compositional bias is a consequence of read counts being relative and not absolute values. At a fixed library size, as the number of reads captured for a set of genes increases, fewer reads are available to capture other sets of genes, which can give a false impression of altered expression between conditions. These considerations are similar to those experienced in bulk RNA-seq analysis, so software such as DESeq2 (Love et al., 2014) or edgeR (Robinson et al., 2010) can be used to correct for them. DESeq2 uses a geometric mean approach and edgeR uses the trimmed mean of M-values approach (Robinson and Oshlack, 2010), both of which will correct for library sequencing depth and compositional bias.

Clustering TSSs into TSRs

A number of different methods for clustering TSSs into TSRs have been developed. The simplest approach, here referred to as naive distance clustering, is available in tools such as CAGEr, TSRexploreR, TSRchitect (Raborn et al., 2017), and CAGEfightR (Thodberg et al., 2019). In this approach, TSSs meeting a specific score threshold and within a specified distance of one another (commonly 20–25 bp) are clustered into TSRs. CAGEr additionally offers a parametric clustering (PARACLU) algorithm that aims to find regions within chromosomes of maximal local TSS density (Frith et al., 2008). Reproducible clustering (RECLU) is a further iteration of PARACLU introduced by the functional annotation of the mammalian genome (FANTOM) consortium that incorporated a number of improvements, including irreproducible discovery rate analysis to enable assessment of TSR reproducibility (Ohmiya et al., 2014). Hypergeometric Optimization of Motif EnRichment (HOMER) employs a chromatin immunoprecipitation sequencing (ChIP-seq)-like clustering algorithm

to cluster densities of TSSs, and can optionally perform additional quality control and filtering steps by using an available genome annotation (Duttke et al., 2019). Integrating Cap Enrichment with Transcript Expression Analysis (icetea), the software accompanying MAPCap, calls peaks with a ChIP-seq-like window-based method utilizing a negative binomial distribution (Bhardwaj et al., 2019). ADAPT-CAGE adopted a machine learning approach by using support vector machines and stochastic gradient boosting, whereby DNA shape features, motifs, and CAGE signal are used to discern strong putative TSSs from background (Georgakilas et al., 2020).

Differential TSSs and TSRs

Differential TSS and TSR state between conditions has received much interest because of the implications for transcript expression and isoform behavior. CAGEexploreR utilizes a user-supplied set of promoters to find differential promoter usage (Dimont et al., 2014), and a classification approach has been proposed to find differential TSS distributions (shape changes) between conditions (Liang et al., 2014). CAGEr and TSRexploreR use a more traditional approach, employing either edgeR or DESeq2 for differential TSS or TSR analysis (Haberle et al., 2015; Policastro et al., 2021).

TSS cluster shifting

It is increasingly clear that large-scale shifts in TSS usage are commonplace during development, disease, and in response to environmental perturbations. Shifting of TSS clusters is defined as TSS density, shifting a relatively short but meaningful distance (often ~100 bases) upstream or downstream between conditions. For instance, a comprehensive CAGE study of zebrafish development uncovered over 900 transcripts with shifted TSRs corresponding to the transition between maternal and zygotic gene expression (Haberle et al., 2014). Shifting of TSSs at thousands of genes is also seen in budding yeast strains bearing various mutations in RNAPII and general transcription factors (Qiu et al., 2020), and also during certain meiotic and mitotic cell cycle stages (Chia et al., 2021).

Because of the emerging importance of TSS clustering shifting, various computational methods have been developed to detect this phenomenon. Generally, such approaches merge TSRs within a given distance to generate a set of consensus regions in which shifting will be assessed. As the distribution of TSSs within a TSR is essentially a discrete probability distribution, computational detection of TSS shifts might be approached as testing for differences between two such distributions. CAGEr introduced the first such method, which performs two distinct calculations. First, a shift score is calculated based on the maximal difference between the empirical cumulative distribution functions of the two TSS distributions. The magnitude of the shift score is reported to reflect the degree to which the TSS signal in the compared distributions is nonoverlapping (e.g., a CAGEr shift score of 0.4 indicates that 40% of the initiation between the two samples does not overlap). Second, a two-sample Kolmogorov-Smirnov (KS) test is performed to test for a significant difference between distributions. The KS test allows detection of shifts in TSS “mass” within a given TSR when the positions of initiation are largely or completely overlapping. Although this approach has been used to great effect in detecting TSS shifts in various contexts, there are limitations. The shift score does not capture shifts in TSS distribution that take place in

largely overlapping positions, and its scale is somewhat unintuitive, ranging from negative infinity to 1 with a sign that does not reflect the directionality of the shift. Calculation of the KS test is also independent of the shift score, and the discrete nature of TSS distributions violates the KS test assumption of continuous distributions.

To address these limitations, we implemented an alternative approach based on the earth mover's distance (EMD) that we termed earth mover's score (EMS) Policastro et al., 2021. EMS has an intuitive scale that spans from −1 to 1: in accordance with conventions for denoting sequence positions in relation to a reference point, a negative EMS indicates an upstream shift and a positive EMS indicates a downstream shift. Furthermore, the p value is computed directly from the test statistic by using a permutation test, which facilitates agreement between EMS and p value. One limitation of the EMS is that a very small score, with attendant lack of significance, could mask "balanced" shifts, meaning expansion or contraction of a TSS cluster in which the movement of TSS density is symmetrical in both directions. To capture these cases, we also report standard unsigned EMD, which reports the overall difference between two distributions without regard to direction, and a corresponding permuted p value and false discovery rate threshold. EMD spans from 0 to 1, with 0 indicating identical TSS positions and 1 indicating no overlap of TSS positions. Balanced shifts are marked by an EMS score near 0, often without a significant p value, and a significant EMD with high magnitude. It is possible that these balanced shifts could be related to changes in peak shape (broad versus peaked), so more exploration of this phenomenon is required.

TSR shape

Early global studies found that TSRs can often be functionally classified by their overall shape. The first detailed global study of TSR shape generated human and mouse CAGE data for the FANTOM consortium (Carninci et al., 2006). In this paper, TSRs were classified into one of four shape categories (single dominant peak [SP], broad [BP], multimodal [MU], or broad with dominant peak [PB]) on the basis of the interquartile range (the distance between the TSS positions encompassing specified upper and lower quantiles of a TSR's signal) and distance between TSSs. They found that TATA box-containing promoters tended to have more peaked TSS distributions, whereas TATA-less, CpG-rich promoters had broader TSS distributions on average. A study of *Drosophila melanogaster* TSRs derived from EST data used a simpler classification scheme wherein TSRs with a single TSS were considered peaked and those with more than one were annotated as broad (Rach et al., 2009). In addition to the TATA box, peaked promoters were also associated with the Initiator element (Inr), Downstream Promoter Element (DPE), and the Motif Ten Element (MTE). In contrast, broad promoters were enriched for the DNA Replication Element (DRE), and the Ohler 1, 6, and 7 elements. The original PEAT paper introduced another classification strategy wherein a smoothed density estimate was fit to each TSR followed by condensation of each TSR to the shortest width containing 95% of its constituent reads and classification into three patterns (narrow with peak, broad with peak [BP], or weak peak) on the basis of the width of the smoothed TSS density (Ni et al., 2010). This classification scheme yielded results similar to those of Rach et al. (2009) in terms of motif enrichment in peaked versus broad promoters.

The most popular contemporary TSR shape measurement is perhaps the shape index (SI). SI quantifies the number of TSSs at each position within a TSR. The scale of SI ranges from negative infinity to one, with scores > −1 classified as peaked and ≤ −1 as broad (Hoskins et al., 2011). SI is a more robust measure of TSR shape than TSR width, as it is not sensitive to low-scoring outlier TSSs. Furthermore, the continuous nature of SI allowed description of a continuum of TSR shapes. Application of SI to *Drosophila* CAGE data revealed many of the same motifs as Rach et al., 2009 and Ni et al. (2010), but additionally found the Pause Button and GAGA motifs enriched in peaked promoters and the NDM1 and DMv1 motifs in broad promoters. Furthermore, it was found that genes with broad promoters tended to be constitutively expressed through development, whereas peaked promoters tended to be activated at certain times during development Hoskins et al., 2011. TSRchitect (Raborn et al., 2021) introduced the Modified Shape Index, which ranges from −1 to 1 rather than negative infinity to 1. Another approach devised a two-step clustering method that utilized a dissimilarity metric called generalized minimum distance of distributions (GM-distance) which is a modified form of minimum distance of pair assignments, and a peakedness score (Zhao et al., 2011). Using this strategy, three TSR cluster shapes emerged: scattered, dense, and ultradense, with most dense and ultradense TSRs corresponding to the SP class in Carninci et al., 2006, and the scattered TSR corresponding to BP, MU, and PB. In addition to motif findings similar to Carninci et al., 2006, exploration of various ChIP-seq datasets showed interesting correlations such as H3K4 methylation, H2A.Z, and H3K79me3 being more associated with scattered promoters, and H3K27me3, H3K9me3, and DNA methylation being more associated with dense and ultradense promoters.

FUTURE PROSPECTS

Single-cell TSS profiling

The past several years have seen a rapid proliferation of techniques for measuring transcript levels (Ramsköld et al., 2012; Tang et al., 2009), DNA methylation (Guo et al., 2013), chromosome conformation (Nagano et al., 2013), chromatin accessibility (Buenrostro et al., 2015; Cusanovich et al., 2015), and protein-DNA interactions (Bartosovic et al., 2021; Grosselin et al., 2019; Rotem et al., 2015; Wu et al., 2021) in single cells. However, little work has been done on cell-to-cell variation in TSS usage: to our knowledge, only two such studies have been performed. Tn5Prime (Cole et al., 2018) uses TSRT of RNA from a single lysed cell followed by tagmentation with Tn5 for library construction. In the second study, nanoCAGE (2017) was combined with the C1 microfluidic platform to yield C1 CAGE (Kouno et al., 2019). The resulting analysis of TSS usage in 136 single cells revealed heterogeneity in the transcriptional response of cells to transforming growth factor (TGF)-β as well as unidirectional enhancer transcription in each cell, suggesting that the bidirectional transcription often observed at enhancers is a result of sampling a population of cells.

Moving forward, how else might single-cell TSS profiling be performed? A number of scRNA-seq approaches use TSRT, already in wide use as a means for TSS mapping. Smart-seq3 (and its predecessors) (Hagemann-Jensen et al., 2020) and Single-cell Tagged Reverse Transcription (STRT) (Islam et al., 2011) use TSRT on single isolated cells, and STRT has been tested for TSS profiling from bulk RNA (Adiconis et al., 2018). The

10X Genomics Chromium microfluidic platform uses TSRT for 5′-centric gene expression profiling, suggesting that a combination of cell isolation and TSRT on the Chromium instrument with a TSS-focused library preparation protocol (e.g., STRIPE-seq) could yield single-cell TSS maps. Further development of such approaches will undoubtedly yield further insight into cell-to-cell variability in gene regulation.

Long-read sequencing

Although short-read sequencing (e.g., on Illumina platforms) is the standard readout for functional genomics methods, long-read sequencing technologies (Oxford NanoPore Technologies (ONT) and Pacific Biosciences Single-Molecule Real-Time (SMRT) sequencing) have gained popularity for applications such as improving existing genome assemblies and assessing structural variation in chromosomes (Logsdon et al., 2020). ONT and SMRT have also been used to determine full-length transcript sequences via cDNA sequencing (Byrne et al., 2017; Sharon et al., 2013), and ONT is capable of direct RNA sequencing for both transcriptome profiling (Garalde et al., 2018; Workman et al., 2019) and detection of modified bases (Leger et al., 2019; Liu et al., 2019). For TSS analysis, long-read sequencing would be advantageous in that a single read would contain both a transcript's TSS and complete coding sequencing, enabling one-to-one assignment of a TSS to its corresponding transcript. However, it has been observed that ONT direct RNA-seq reads, originating from a transcript's 3′ end, are often truncated before the transcript's true TSS (Workman et al., 2019). This might arise from electrical abnormalities because of enzyme stalling during RNA translocation or voltage spikes of unknown origin. Extremely rapid translocation of the 5′-most 10–15 nucleotides of a transcript though a pore also prevents reading of these terminal nucleotides. Thus, dedicated methods are still required to confidently detect TSSs when ONT direct RNA-seq is performed.

CONCLUDING REMARKS

Despite the development of a wide variety of techniques for global TSS mapping over the past few decades, such methods are integrated into global studies of gene expression far less frequently than approaches for quantifying transcript levels (e.g., RNA-seq) or mapping protein-DNA interactions (e.g., ChIP-seq, CUT&RUN). Given that heterogeneity in TSS usage is a major driver of transcript isoform diversity (Reyes and Huber, 2018; Shabalina et al., 2014), likely plays important roles in development (Cvetesic et al., 2020; Haberle et al., 2014), is involved in the response to environmental stimuli (Kurihara et al., 2018; Lu and Lin, 2019; Ushijima et al., 2017), and is altered in cancer (Demircioğlu et al., 2019), we argue that TSS mapping techniques can provide insights into gene regulation complementary or inaccessible to those obtained with other more commonly used techniques. Indeed, CAGE has been extensively used alongside methods such as RNA-seq and ChIP-seq in the context of large consortia such as FANTOM (Forrest et al., 2014) and encyclopedia of DNA elements (ENCODE) (ENCODE Project Consortium, 2012) groups, where it has provided great insight into promoter-level gene regulation. Furthermore, many TSS mapping techniques can also provide information on transcript levels comparable to those obtained with various RNA-seq approaches, increasing the cost efficiency of each experiment. In some cases, widespread adoption of TSS mapping techniques might have been hampered

by barriers of cost and/or technical difficulty. For instance, we calculated the per-sample cost of nAnT-iCAGE and SLIC-CAGE to be > $100 USD, with protocols spanning multiple days (Policastro et al., 2020), whereas a commercial nAnT-iCAGE kit (https://cage-seq.com/cage_kit/index.html) has a cost of 25,000 JPY (~$230 USD) per sample. However, these methods are currently regarded as the gold standards for TSS mapping thanks to their sensitivity, resolution, and low bias (Cvetesic et al., 2018) and so are preferred for applications in which high sensitivity is essential. More routine profiling of TSS usage can easily be performed with TSRT-based methods, which trade a moderate degree of sensitivity for reduced cost and simpler, faster protocols (Policastro et al., 2020). Given the range of methods available, we surmise that any researchers interested in profiling transcription initiation will be able to find a method that suits their needs. To facilitate methodological comparisons, we provide Table 1, which outlines the general advantages and disadvantages associated with each of the three enzymatic approaches discussed here as well as salient features of specific techniques, and Table 2, which lists reported RNA input amounts for each protocol.

ACKNOWLEDGMENTS

Work in the Zentner Lab was supported by NIH grant R35GM128631 to G.E.Z.

REFERENCES

Adiconis X, Haber AL, Simmons SK, Levy Moonshine A, Ji Z, Busby MA, Shi X, Jacques J, Lancaster MA, Pan JQ, et al. (2018). Comprehensive comparative analysis of 5′-end RNA-sequencing methods. Nat. Methods 15, 505–511. [PubMed: 29867192]

Anders S, Pyl PT, and Huber W (2014). HTSeq—a Python framework to work with high-throughput sequencing data. Bioinformatics 31, 166–169. [PubMed: 25260700]

Arribere JA, and Gilbert WV (2013). Roles for transcript leaders in translation and mRNA decay revealed by transcript leader sequencing. Genome Res. 23, 977–987. [PubMed: 23580730]

Baldrich P, Tamim S, Mathioni S, and Meyers B (2020). Ligation Bias Is a Major Contributor to Nonstoichiometric Abundances of Secondary siRNAs and Impacts Analyses of microRNAs. BioRxiv, 2020.09.14.296616. 10.1101/2020.09.14.296616.

Balwierz PJ, Carninci P, Daub CO, Kawai J, Hayashizaki Y, Van Belle W, Beisel C, and van Nimwegen E (2009). Methods for analyzing deep sequencing expression data: constructing the human and mouse promoterome with deepCAGE data. Genome Biol. 10, R79. [PubMed: 19624849]

Barbosa C, Peixeiro I, and Romão L (2013). Gene expression regulation by upstream open reading frames and human disease. PLoS Genet. 9, e1003529. [PubMed: 23950723]

Bartosovic M, Kabbe M, and Castelo-Branco G (2021). Single-cell CUT&Tag profiles histone modifications and transcription factors in complex tissues. Nat. Biotechnol 39, 825–835. [PubMed: 33846645]

Batut P, Dobin A, Plessy C, Carninci P, and Gingeras TR (2013). High-fidelity promoter profiling reveals widespread alternative promoter usage and transposon-driven developmental gene expression. Genome Res. 23, 169–180. [PubMed: 22936248]

Bhardwaj V, Semplicio G, Erdogdu NU, Manke T, and Akhtar A (2019). MAPCap allows high-resolution detection and differential expression analysis of transcription start sites. Nat. Commun 10, 3219. [PubMed: 31363093]

Blower MD, Jambhekar A, Schwarz DS, and Toombs JA (2013). Combining different mRNA capture methods to analyze the transcriptome: analysis of the Xenopus laevis transcriptome. PLoS One 8, e77700. [PubMed: 24143257]

Bray NL, Pimentel H, Melsted P, and Pachter L (2016). Near-optimal probabilistic RNA-seq quantification. Nat. Biotechnol 34, 525–527. [PubMed: 27043002]

Buenrostro JD, Wu B, Litzenburger UM, Ruff D, Gonzales ML, Snyder MP, Chang HY, and Greenleaf WJ (2015). Single-cell chromatin accessibility reveals principles of regulatory variation. Nature 523, 486–490. [PubMed: 26083756]

Byrne A, Beaudin AE, Olsen HE, Jain M, Cole C, Palmer T, DuBois RM, Forsberg EC, Akeson M, and Vollmers C (2017). Nanopore long-read RNAseq reveals widespread transcriptional variation among the surface receptors of individual B cells. Nat. Commun 8, 16027. [PubMed: 28722025]

Cai H, and Luse DS (1987). Transcription initiation by RNA polymerase II in vitro. Properties of preinitiation, initiation, and elongation complexes. J. Biol. Chem 262, 298–304. [PubMed: 2432061]

Carninci P, Kvam C, Kitamura A, Ohsumi T, Okazaki Y, Itoh M, Kamiya M, Shibata K, Sasaki N, Izawa M, et al. (1996). High-efficiency full-length cDNA cloning by biotinylated CAP trapper. Genomics 37, 327–336. [PubMed: 8938445]

Carninci P, Sandelin A, Lenhard B, Katayama S, Shimokawa K, Ponjavic J, Semple CAM, Taylor MS, Engström PG, Frith MC, et al. (2006). Genome-wide analysis of mammalian promoter architecture and evolution. Nat. Genet 38, 626–635. [PubMed: 16645617]

Chen J, Tresenrider A, Chia M, McSwiggen DT, Spedale G, Jorgensen V, Liao H, van Werven FJ, and Ünal E (2017). Kinetochore inactivation by expression of a repressive mRNA. Elife 6, e27417. [PubMed: 28906249]

Chia M, Tresenrider A, Chen J, Spedale G, Jorgensen V,Ünal E, and van Werven FJ (2017). Transcription of a 5' extended mRNA isoform directs dynamic chromatin changes and interference of a downstream promoter. Elife 6, e27420. [PubMed: 28906248]

Chia M, Li C, Marques S, Pelechano V, Luscombe NM, and van Werven FJ (2021). High-resolution analysis of cell-state transitions in yeast suggests widespread transcriptional tuning by alternative starts. Genome Biol. 22, 34. [PubMed: 33446241]

Choi YH, and Hagedorn CH (2003). Purifying mRNAs with a high-affinity eIF4E mutant identifies the short 3′ poly(A) end phenotype. Proc. Natl. Acad. Sci 100, 7033–7038. [PubMed: 12777618]

Cole C, Byrne A, Beaudin AE, Forsberg EC, and Vollmers C (2018). Tn5Prime, a Tn5 based 5′ capture method for single cell RNA-seq. Nucleic Acids Res. 46, e62. [PubMed: 29548006]

Core LJ, Martins AL, Danko CG, Waters CT, Siepel A, and Lis JT (2014). Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. Nat. Genet 46, 1311–1320. [PubMed: 25383968]

Cumbie JS, Ivanchenko MG, and Megraw M (2015). NanoCAGE-XL and CapFilter: an approach to genome wide identification of high confidence transcription start sites. BMC Genomics 16, 597. [PubMed: 26268438]

Cusanovich DA, Daza R, Adey A, Pliner H, Christiansen L, Gunderson KL, Steemers FJ, Trapnell C, and Shendure J (2015). Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing. Science 348, 910–914. [PubMed: 25953818]

Cvetesic N, Borkowska M, Hatanaka Y, Yu C, Vincent SD, Müller F, Tora L, Leitch HG, Hajkova P, and Lenhard B (2020). Global regulatory transitions at core promoters demarcate the mammalian germline cycle. BioRxiv. 10.1101/2020.10.30.361865.

Cvetesic N, Leitch HG, Borkowska M, Müller F, Carninci P, Hajkova P, and Lenhard B (2018). SLIC-CAGE: high-resolution transcription start site mapping using nanogram-levels of total RNA. Genome Res. 28, 1943–1956. [PubMed: 30404778]

Danks GB, Navratilova P, Lenhard B, and Thompson EM (2018). Distinct core promoter codes drive transcription initiation at key developmental transitions in a marine chordate. BMC Genomics 19, 164. [PubMed: 29482522]

Demircioğlu D, Cukuroglu E, Kindermans M, Nandi T, Calabrese C, Fonseca NA, Kahles A, Lehmann K-V, Stegle O, Brazma A, et al. (2019). A pan-cancer transcriptome analysis reveals pervasive regulation through alternative promoters. Cell 178, 1465–1477.e17. [PubMed: 31491388]

Dieudonné F-X, O'Connor PBF, Gubler-Jaquier P, Yasrebi H, Conne B, Nikolaev S, Antonarakis S, Baranov PV, and Curran J (2015). The effect of heterogeneous Transcription Start Sites (TSS)

on the translatome: implications for the mammalian cellular phenotype. BMC Genomics 16, 986. [PubMed: 26589636]

Dimont E, Hofmann O, Ho Sui SJ, Forrest ARR, Kawaji H, and Hide W; the FANTOM Consortium (2014). CAGExploreR: an R package for the analysis and visualization of promoter dynamics across multiple experiments. Bioinformatics 30, 1183–1184. [PubMed: 24675730]

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

Duttke SH, Chang MW, Heinz S, and Benner C (2019). Identification and dynamic quantification of regulatory elements using total RNA. Genome Res. 29, 1836–1846. [PubMed: 31649059]

ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57–74. [PubMed: 22955616]

Fenouil R, Cauchy P, Koch F, Descostes N, Cabeza JZ, Innocenti C, Ferrier P, Spicuglia S, Gut M, Gut I, et al. (2012). CpG islands and GC content dictate nucleosome depletion in a transcription-independent manner at mammalian promoters. Genome Res. 22, 2399–2408. [PubMed: 23100115]

Forrest ARR, Kawaji H, Rehli M, Kenneth Baillie J, de Hoon MJL, Haberle V, Lassmann T, Kulakovskiy IV, Lizio M, Itoh M, et al. (2014). A promoter-level mammalian expression atlas. Nature 507, 462–470. [PubMed: 24670764]

Frith MC, Valen E, Krogh A, Hayashizaki Y, Carninci P, and Sandelin A (2008). A code for transcription initiation in mammalian genomes. Genome Res. 18, 1–12. [PubMed: 18032727]

Fuchs RT, Sun Z, Zhuang F, and Robb GB (2015). Bias in ligation-based small RNA sequencing library construction is determined by adaptor and RNA structure. PLoS One 10, e0126049. [PubMed: 25942392]

Garalde DR, Snell EA, Jachimowicz D, Sipos B, Lloyd JH, Bruce M, Pantic N, Admassu T, James P, Warland A, et al. (2018). Highly parallel direct RNA sequencing on an array of nanopores. Nat. Methods 15, 201–206. [PubMed: 29334379]

Georgakilas GK, Perdikopanis N, and Hatzigeorgiou A (2020). Solving the transcription start site identification problem with ADAPT-CAGE: a Machine Learning algorithm for the analysis of CAGE data. Sci. Rep 10, 877. [PubMed: 31965016]

Gowthaman U, García-Pichardo D, Jin Y, Schwarz I, and Marquardt S (2020). DNA processing in the context of noncoding transcription. Trends Bio-chem. Sci 45, 1009–1021.

Grosselin K, Durand A, Marsolier J, Poitou A, Marangoni E, Nemati F, Dahmani A, Lameiras S, Reyal F, Frenoy O, et al. (2019). High-throughput single-cell ChIP-seq identifies heterogeneity of chromatin states in breast cancer. Nat. Genet 51, 1060–1066. [PubMed: 31152164]

Gu W, Lee H-C, Chaves D, Youngman EM, Pazour GJ, Conte D, and Mello CC (2012). CapSeq and CIP-TAP identify pol II start sites and reveal capped small RNAs as C. elegans piRNA precursors. Cell 151, 1488–1500. [PubMed: 23260138]

Guo H, Zhu P, Wu X, Li X, Wen L, and Tang F (2013). Single-cell methylome landscapes of mouse embryonic stem cells and early embryos analyzed using reduced representation bisulfite sequencing. Genome Res. 23, 2126–2135. [PubMed: 24179143]

Haberle V, Li N, Hadzhiev Y, Plessy C, Previti C, Nepal C, Gehrig J, Dong X, Akalin A, Suzuki AM, et al. (2014). Two independent transcription initiation codes overlap on vertebrate core promoters. Nature 507, 381–385. [PubMed: 24531765]

Haberle V, Forrest ARR, Hayashizaki Y, Carninci P, and Lenhard B (2015). CAGEr: precise TSS data retrieval and high-resolution promoterome mining for integrative analyses. Nucleic Acids Res. 43, e51. [PubMed: 25653163]

Hafner M, Renwick N, Brown M, Mihailović A, Holoch D, Lin C, Pena JTG, Nusbaum JD, Morozov P, Ludwig J, et al. (2011). RNA-ligase-dependent biases in miRNA representation in deep-sequenced small RNA cDNA libraries. RNA 17, 1697–1712. [PubMed: 21775473]

Hagemann-Jensen M, Ziegenhain C, Chen P, Ramsköld D, Hendriks G-J, Larsson AJM, Faridani OR, and Sandberg R (2020). Single-cell RNA counting at allele and isoform resolution using Smart-seq3. Nat. Biotechnol 38, 708–714. [PubMed: 32518404]

Hashimoto S, Suzuki Y, Kasai Y, Morohoshi K, Yamada T, Sese J, Morishita S, Sugano S, and Matsushima K (2004). 5′-end SAGE for the analysis of transcriptional start sites. Nat. Biotechnol 22, 1146–1149. [PubMed: 15300261]

Hirabayashi S, Bhagat S, Matsuki Y, Takegami Y, Uehata T, Kanemaru A, Itoh M, Shirakawa K, Takaori-Kondo A, Takeuchi O, et al. (2019). NET-CAGE characterizes the dynamics and topology of human transcribed cis-regulatory elements. Nat. Genet 51, 1369–1379. [PubMed: 31477927]

Hollerer I, Barker JC, Jorgensen V, Tresenrider A, Dugast-Darzacq C, Chan LY, Darzacq X, Tjian R,Ünal E, and Brar GA (2019). Evidence for an integrated gene repression mechanism based on mRNA isoform toggling in human cells. G3 GenesGenomesGenetics 9, 1045–1053.

Hoskins RA, Landolin JM, Brown JB, Sandler JE, Takahashi H, Lassmann T, Yu C, Booth BW, Zhang D, Wan KH, et al. (2011). Genome-wide analysis of promoter architecture in Drosophila melanogaster. Genome Res. 21, 182–192. [PubMed: 21177961]

Islam S, Kjällquist U, Moliner A, Zajac P, Fan J-B, Lönnerberg P, and Linnarsson S (2011). Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. Genome Res. 21, 1160–1167. [PubMed: 21543516]

Jayaprakash AD, Jabado O, Brown BD, and Sachidanandam R (2011). Identification and remediation of biases in the activity of RNA ligases in small-RNA deep sequencing. Nucleic Acids Res. 39, e141. [PubMed: 21890899]

Jorgensen V, Chen J, Vander Wende H, Harris DE, McCarthy A, Breznak S, Wong-Deyrup SW, Chen Y, Rangan P, Brar GA, et al. (2020). Tunable transcriptional interference at the endogenous alcohol dehydrogenase gene locus in Drosophila melanogaster. G3 GenesGenomesGenetics 10, 1575–1583.

Kanamori-Katayama M, Itoh M, Kawaji H, Lassmann T, Katayama S, Kojima M, Bertin N, Kaiho A, Ninomiya N, Daub CO, et al. (2011). Unamplified cap analysis of gene expression on a single-molecule sequencer. Genome Res. 21, 1150–1159. [PubMed: 21596820]

Kapteyn J, He R, McDowell ET, and Gang DR (2010). Incorporation of non-natural nucleotides into template-switching oligonucleotides reduces background and improves cDNA synthesis from very small RNA samples. BMC Genomics 11, 413. [PubMed: 20598146]

Kawai J, Shinagawa A, Shibata K, Yoshino M, Itoh M, Ishii Y, Arakawa T, Hara A, Fukunishi Y, Konno H, et al. (2001). Functional annotation of a full-length mouse cDNA collection. Nature 409, 685–690. [PubMed: 11217851]

Kazuo M, and Sumio S (1994). Oligo-capping: a simple method to replace the cap structure of eukaryotic mRNAs with oligoribonucleotides. Gene 138, 171–174. [PubMed: 8125298]

Kodzius R, Kojima M, Nishiyori H, Nakamura M, Fukuda S, Tagami M, Sasaki D, Imamura K, Kai C, Harbers M, et al. (2006). CAGE: cap analysis of gene expression. Nat. Methods 3, 211–222. [PubMed: 16489339]

Kouno T, Moody J, Kwon AT-J, Shibayama Y, Kato S, Huang Y, Böttcher M, Motakis E, Mendez M, Severin J, et al. (2019). C1 CAGE detects transcription start sites and enhancer activity at single-cell resolution. Nat. Commun 10, 360. [PubMed: 30664627]

Kurihara Y, Makita Y, Kawashima M, Fujita T, Iwasaki S, and Matsui M (2018). Transcripts from downstream alternative transcription start sites evade uORF-mediated inhibition of gene expression in Arabidopsis. Proc. Natl. Acad. Sci 115, 7831–7836. [PubMed: 29915080]

Kwak H, Fuda NJ, Core LJ, and Lis JT (2013). Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. Science 339, 950–953. [PubMed: 23430654]

Leger A, Amaral PP, Pandolfini L, Capitanchik C, Capraro F, Barbieri I, Migliori V, Luscombe NM, Enright AJ, Tzelepis K, et al. (2019). RNA Modifications Detection by Comparative Nanopore Direct RNA Sequencing. BioRxiv, 843136.

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

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

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, and Durbin R; 1000 Genome Project Data Processing Sub-group (2009). The sequence alignment/map format and SAMtools. Bioinformatics 25, 2078–2079. [PubMed: 19505943]

Liang K, Suzuki Y, Kumagai Y, and Nakai K (2014). Analysis of changes in transcription start site distribution by a classification approach. Gene 537, 29–40. [PubMed: 24389500]

Liao Y, Smyth GK, and Shi W (2019). The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads. Nucleic Acids Res. 47, e47. [PubMed: 30783653]

Lin D, Hiron TK, and O'Callaghan CA (2018). Intragenic transcriptional interference regulates the human immune ligand MICA. EMBO J. 37, e97138. [PubMed: 29643123]

Liu H, Begik O, Lucas MC, Ramirez JM, Mason CE, Wiener D, Schwartz S, Mattick JS, Smith MA, and Novoa EM (2019). Accurate detection of m6A RNA modifications in native RNA sequences. Nat. Commun 10, 4079. [PubMed: 31501426]

Logsdon GA, Vollger MR, and Eichler EE (2020). Long-read human genome sequencing and its applications. Nat. Rev. Genet 21, 597–614. [PubMed: 32504078]

Love MI, Huber W, and Anders S (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 15, 550. [PubMed: 25516281]

Lu Z, and Lin Z (2019). Pervasive and dynamic transcription initiation in Saccharomyces cerevisiae. Genome Res. 29, 1198–1210. [PubMed: 31076411]

Malabat C, Feuerbach F, Ma L, Saveanu C, and Jacquier A (2015). Quality control of transcription start site selection by nonsense-mediated-mRNA decay. Elife 4, e06722.

Matveeva EA, Al-Tinawi QMH, Rouchka EC, and Fondufe-Mittendorf YN (2019). Coupling of PARP1-mediated chromatin structural changes to transcriptional RNA polymerase II elongation and cotranscriptional splicing. Epigenetics Chromatin 12, 1–18. [PubMed: 30602389]

Mayer A, di Iulio J, Maleri S, Eser U, Vierstra J, Reynolds A, Sandstrom R, Stamatoyannopoulos JA, and Churchman LS (2015). Native elongating transcript sequencing reveals human transcriptional activity at nucleotide resolution. Cell 161, 541–554. [PubMed: 25910208]

Mejía-Guerra MK, Li W, Galeano NF, Vidal M, Gray J, Doseff AI, and Grotewold E (2015). Core promoter plasticity between maize tissues and genotypes contrasts with predominance of sharp transcription initiation sites. Plant Cell 27, 3309–3320. [PubMed: 26628745]

Morton T, Petricka J, Corcoran DL, Li S, Winter CM, Carda A, Benfey PN, Ohler U, and Megraw M (2014). Paired-end analysis of transcription start sites in Arabidopsis reveals plant-specific promoter signatures. Plant Cell 26, 2746–2760. [PubMed: 25035402]

Murata M, Nishiyori-Sueki H, Kojima-Ishiyama M, Carninci P, Hayashizaki Y, and Itoh M (2014). Detecting expressed genes using CAGE. In Transcription Factor Regulatory Networks: Methods and Protocols, Miyamoto-Sato E, Ohashi H, Sasaki H, Nishikawa J, and Yanagawa H, eds. (Springer), pp. 67–85.

Nagano T, Lubling Y, Stevens TJ, Schoenfelder S, Yaffe E, Dean W, Laue ED, Tanay A, and Fraser P (2013). Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. Nature 502, 59–64. [PubMed: 24067610]

Nechaev S, Fargo DC, dos Santos G, Liu L, Gao Y, and Adelman K (2010). Global analysis of short RNAs reveals widespread promoter-proximal stalling and arrest of pol II in Drosophila. Science 327, 335–338. [PubMed: 20007866]

Ni T, Corcoran DL, Rach EA, Song S, Spana EP, Gao Y, Ohler U, and Zhu J (2010). A paired-end sequencing strategy to map the complex landscape of transcription initiation. Nat. Methods 7, 521–527. [PubMed: 20495556]

Nielsen M, Ard R, Leng X, Ivanov M, Kindgren P, Pelechano V, and Marquardt S (2019). Transcription-driven chromatin repression of Intragenic transcription start sites. PLoS Genet. 15, e1007969. [PubMed: 30707695]

Ohmiya H, Vitezic M, Frith MC, Itoh M, Carninci P, Forrest AR, Hayashizaki Y, and Lassmann T; The FANTOM Consortium (2014). RECLU: a pipeline to discover reproducible transcriptional start sites and their alternative regulation using capped analysis of gene expression (CAGE). BMC Genomics 15, 269. [PubMed: 24779366]

Okazaki Y, Furuno M, Kasukawa T, Adachi J, Bono H, Kondo S, Nikaido I, Osato N, Saito R, Suzuki H, et al. (2002). Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. Nature 420, 563–573. [PubMed: 12466851]

Paquette DR, Mugridge JS, Weinberg DE, and Gross JD (2018). Application of a Schizosaccharomyces pombe Edc1-fused Dcp1–Dcp2 decapping enzyme for transcription start site mapping. RNA 24, 251–257. [PubMed: 29101277]

Park D, Morris AR, Battenhouse A, and Iyer VR (2014). Simultaneous mapping of transcript ends at single-nucleotide resolution and identification of widespread promoter-associated non-coding RNA governed by TATA elements. Nucleic Acids Res. 42, 3736–3749. [PubMed: 24413663]

Patro R, Duggal G, Love MI, Irizarry RA, and Kingsford C (2017). Salmon provides fast and bias-aware quantification of transcript expression. Nat. Methods 14, 417–419. [PubMed: 28263959]

Pelechano V, Wei W, and Steinmetz LM (2013). Extensive transcriptional heterogeneity revealed by isoform profiling. Nature 497, 127–131. [PubMed: 23615609]

Plessy C, Bertin N, Takahashi H, Simone R, Salimullah M, Lassmann T, Vitezic M, Severin J, Olivarius S, Lazarevic D, et al. (2010). Linking promoters to functional transcripts in small samples with nanoCAGE and CAGEscan. Nat. Methods 7, 528–534. [PubMed: 20543846]

Policastro RA, Raborn RT, Brendel VP, and Zentner GE (2020). Simple and efficient profiling of transcription initiation and transcript levels with STRIPE-seq. Genome Res. 30, 910–923. [PubMed: 32660958]

Policastro RA, McDonald DJ, Brendel VP, and Zentner GE (2021). Flexible analysis of TSS mapping data and detection of TSS shifts with TSRexploreR. NAR Genom. Bioinform 3, lqab051.

Poulain S, Kato S, Arnaud O, Morlighem J-É, Suzuki M, Plessy C, and Harbers M (2017). NanoCAGE: a method for the analysis of coding and non-coding 5′-capped transcriptomes. In Promoter Associated RNA: Methods and Protocols, Napoli S, ed. (Springer), pp. 57–109.

Qiu C, Jin H, Vvedenskaya I, Llenas JA, Zhao T, Malik I, Visbisky AM, Schwartz SL, Cui P, Čabart P, et al. (2020). Universal promoter scanning by Pol II during transcription initiation in Saccharomyces cerevisiae. Genome Biol. 21, 132. [PubMed: 32487207]

Raborn RT, Sridharan K, and Brendel VP (2017). TSRchitect: promoter identification from large-scale TSS profiling data. https://dx.org/doi:10.18129/b9.bioc.TSRchitect.

Raborn RT, Brendel VP, and Sridharan K (2021). TSRchitect: Promoter Identification from Large-Scale TSS Profiling Data (Bioconductor version: Release), p. 3.12.

Rach EA, Yuan H-Y, Majoros WH, Tomancak P, and Ohler U (2009). Motif composition, conservation and condition-specificity of single and alternative transcription start sites in the Drosophila genome. Genome Biol. 10, R73. [PubMed: 19589141]

Ramsköld D, Luo S, Wang Y-C, Li R, Deng Q, Faridani OR, Daniels GA, Khrebtukova I, Loring JF, Laurent LC, et al. (2012). Full-Length mRNA-Seq from single cell levels of RNA and individual circulating tumor cells. Nat. Biotechnol 30, 777–782. [PubMed: 22820318]

Reyes A, and Huber W (2018). Alternative start and termination sites of transcription drive most transcript isoform differences across human tissues. Nucleic Acids Res. 46, 582–592. [PubMed: 29202200]

Robinson MD, and Oshlack A (2010). A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. 11, R25. [PubMed: 20196867]

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

Rojas-Duran MF, and Gilbert WV (2012). Alternative transcription start site selection leads to large differences in translation activity in yeast. RNA 18, 2299–2305. [PubMed: 23105001]

Rotem A, Ram O, Shoresh N, Sperling RA, Goren A, Weitz DA, and Bernstein BE (2015). Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state. Nat. Biotechnol 33, 1165–1172. [PubMed: 26458175]

Scheidegger A, Dunn CJ, Samarakkody A, Koney NK-K, Perley D, Saha RN, and Nechaev S (2019). Genome-wide RNA pol II initiation and pausing in neural progenitors of the rat. BMC Genomics 20, 477. [PubMed: 31185909]

Schmidt WM, and Mueller MW (1999). CapSelect: a highly sensitive method for 5′ CAP-dependent enrichment of full-length cDNA in PCR-mediated analysis of mRNAs. Nucleic Acids Res. 27, e31–i. [PubMed: 10518626]

Schon MA, Kellner MJ, Plotnikova A, Hofmann F, and Nodine MD (2018). NanoPARE: parallel analysis of RNA 5′ ends from low-input RNA. Genome Res. 28, 1931–1942. [PubMed: 30355603]

Shabalina SA, Ogurtsov AY, Spiridonov NA, and Koonin EV (2014). Evolution at protein ends: major contribution of alternative transcription initiation and termination to the transcriptome and proteome diversity in mammals. Nucleic Acids Res. 42, 7132–7144. [PubMed: 24792168]

Sharon D, Tilgner H, Grubert F, and Snyder M (2013). A single-molecule long-read survey of the human transcriptome. Nat. Biotechnol 31, 1009–1014. [PubMed: 24108091]

Shibata K, Itoh M, Aizawa K, Nagaoka S, Sasaki N, Carninci P, Konno H, Akiyama J, Nishi K, Kitsunai T, et al. (2000). RIKEN integrated sequence analysis (RISA) system—384-format sequencing pipeline with 384 multicapillary sequencer. Genome Res. 10, 1757–1771. [PubMed: 11076861]

Shiraki T, Kondo S, Katayama S, Waki K, Kasukawa T, Kawaji H, Kodzius R, Watahiki A, Nakamura M, Arakawa T, et al. (2003). Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. Proc. Natl. Acad. Sci 100, 15776–15781. [PubMed: 14663149]

Smith TS, Heger A, and Sudbery I (2017). UMI-tools: modelling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. Genome Res. 27, 491–499. [PubMed: 28100584]

Suzuki Y, and Sugano S (2003). Construction of a full-length enriched and a 5′-end enriched cDNA library using the oligo-capping method. In Generation of CDNA Libraries: Methods and Protocols, Ying S-Y, ed. (Humana Press)), pp. 73–91.

Takahashi H, Lassmann T, Murata M, and Carninci P (2012). 5′ end–centered expression profiling using cap-analysis gene expression and next-generation sequencing. Nat. Protoc 7, 542–561. [PubMed: 22362160]

Tang DTP, Plessy C, Salimullah M, Suzuki AM, Calligaris R, Gustincich S, and Carninci P (2013). Suppression of artifacts and barcode bias in high-throughput transcriptome analyses utilizing template switching. Nucleic Acids Res. 41, e44. [PubMed: 23180801]

Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, Wang X, Bodeau J, Tuch BB, Siddiqui A, et al. (2009). mRNA-Seq whole-transcriptome analysis of a single cell. Nat. Methods 6, 377–382. [PubMed: 19349980]

Thodberg M, Thieffry A, Vitting-Seerup K, Andersson R, and Sandelin A (2019). CAGEfightR: analysis of 5′-end data using R/Bioconductor. BMC Bioinformatics 20, 487. [PubMed: 31585526]

Tsuchihara K, Suzuki Y, Wakaguri H, Irie T, Tanimoto K, Hashimoto S, Matsushima K, Mizushima-Sugano J, Yamashita R, Nakai K, et al. (2009). Massive transcriptional start site analysis of human genes in hypoxia cells. Nucleic Acids Res. 37, 2249–2263. [PubMed: 19237398]

Turchinovich A, Surowy H, Serva A, Zapatka M, Lichter P, and Burwinkel B (2014). Capture and amplification by tailing and switching (CATS). RNA Biol. 11, 817–828. [PubMed: 24922482]

Ushijima T, Hanada K, Gotoh E, Yamori W, Kodama Y, Tanaka H, Kusano M, Fukushima A, Tokizawa M, Yamamoto YY, et al. (2017). Light controls protein localization through phytochrome-mediated alternative promoter selection. Cell 171, 1316–1325.e12. [PubMed: 29129375]

Valen E, Pascarella G, Chalk A, Maeda N, Kojima M, Kawazu C, Murata M, Nishiyori H, Lazarevic D, Motti D, et al. (2009). Genome-wide detection and analysis of hippocampus core promoters using DeepCAGE. Genome Res. 19, 255–265. [PubMed: 19074369]

Wakaguri H, Yamashita R, Suzuki Y, Sugano S, and Nakai K (2008). DBTSS: database of transcription start sites, progress report 2008. Nucleic Acids Res. 36, D97–D101. [PubMed: 17942421]

Wang X, Hou J, Quedenau C, and Chen W (2016). Pervasive isoform-specific translational regulation via alternative transcription start sites in mammals. Mol. Syst. Biol 12, 875. [PubMed: 27430939]

Weber CM, Ramachandran S, and Henikoff S (2014). Nucleosomes are context-specific, H2A.Z-modulated barriers to RNA polymerase. Mol. Cell 53, 819–830. [PubMed: 24606920]

Workman RE, Tang AD, Tang PS, Jain M, Tyson JR, Razaghi R, Zuzarte PC, Gilpatrick T, Payne A, Quick J, et al. (2019). Nanopore native RNA sequencing of a human poly(A) transcriptome. Nat. Methods 16, 1297–1305. [PubMed: 31740818]

Wu SJ, Furlan SN, Mihalas AB, Kaya-Okur HS, Feroze AH, Emerson SN, Zheng Y, Carson K, Cimino PJ, Keene CD, et al. (2021). Single-cell CUT&Tag analysis of chromatin modifications in differentiation and tumor progression. Nat. Biotechnol 39, 819–824. [PubMed: 33846646]

Wuarin J, and Schibler U (1994). Physical isolation of nascent RNA chains transcribed by RNA polymerase II: evidence for cotranscriptional splicing. Mol. Cell. Biol 14, 7219–7225. [PubMed: 7523861]

Wulf MG, Maguire S, Humbert P, Dai N, Bei Y, Nichols NM, Corrêa IR, and Guan S (2019). Non-templated addition and template switching by Moloney murine leukemia virus (MMLV)-based reverse transcriptases co-occur and compete with each other. J. Biol. Chem 294, 18220–18231. [PubMed: 31640989]

Yamashita R, Sathira NP, Kanai A, Tanimoto K, Arauchi T, Tanaka Y, Hashimoto S, Sugano S, Nakai K, and Suzuki Y (2011). Genome-wide characterization of transcriptional start sites in humans by integrative transcriptome analysis. Genome Res. 21, 775–789. [PubMed: 21372179]

Zhang Z, and Dietrich FS (2005). Mapping of transcription start sites in Saccharomyces cerevisiae using 5′ SAGE. Nucleic Acids Res. 33, 2838–2851. [PubMed: 15905473]

Zhang P, Dimont E, Ha T, Swanson DJ, Hide W, and Goldowitz D; the FANTOM Consortium (2017). Relatively frequent switching of transcription start sites during cerebellar development. BMC Genomics 18, 461. [PubMed: 28610618]

Zhao X, Valen E, Parker BJ, and Sandelin A (2011). Systematic clustering of transcription start site landscapes. PLoS One 6, e23409. [PubMed: 21887249]

Zhu Y.y., Machleder E.m., Chenchik A, Li R, and Siebert P.d. (2001). Reverse transcriptase template switching: a SMART™ approach for full-length cDNA library construction. BioTechniques 30, 892–897. [PubMed: 11314272]

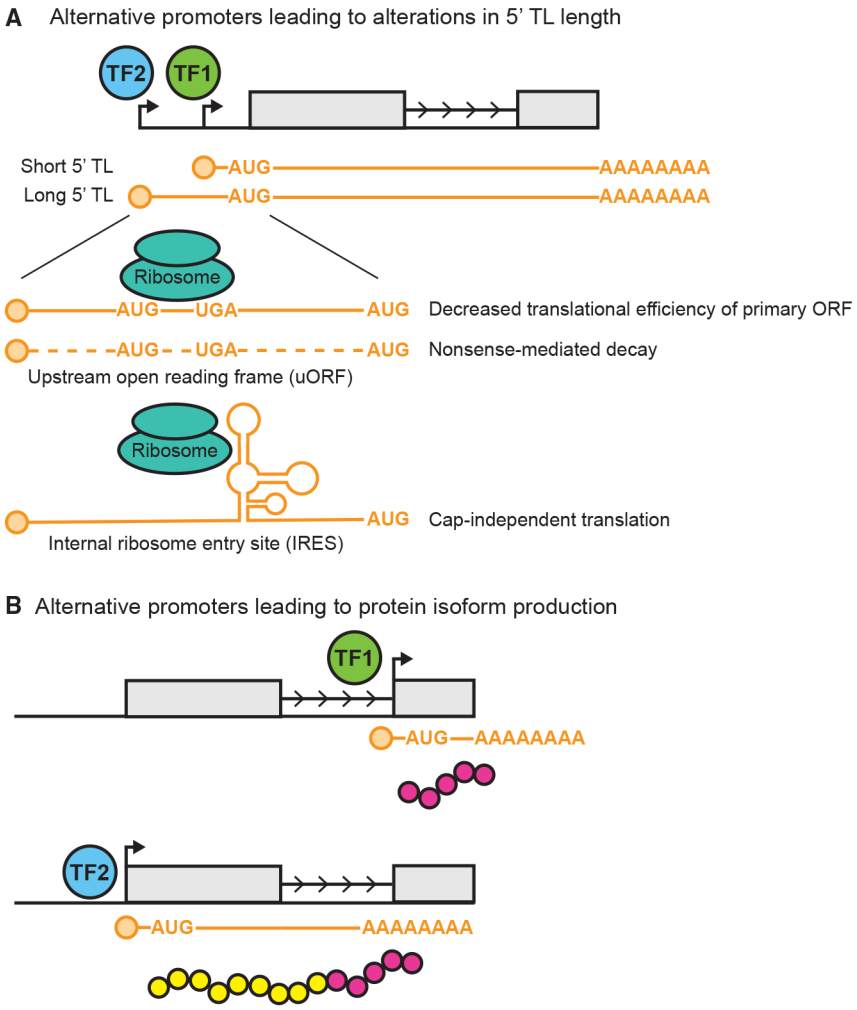

Figure 1. Effects of TSS selection on gene expression

(A) Possible effects of 5′ TL lengthening on transcript stability and translation. Transcription factor (TF) 1 specifies use of a proximal promoter, leading to a transcript with a short 5′ TL, while TF2 activates an upstream promoter that produces a transcript with a long 5′ TL. The extended 5′ TL may contain a uORF, which can act as a "sponge" for ribosomes by preventing them from reaching the transcript's primary ORF and may also lead to destruction of the transcript via NMD if the uORF stop codon is recognized as premature. The 5′ TL may also contain an IRES, enabling cap-independent translation. We note that these 5′ TL features are not mutually exclusive and direct interested readers to a recent comprehensive review on the roles of 5′ TLs in gene regulation (Leppek et al., 2018). (B) Production of transcripts encoding distinct protein isoforms by TF-mediated activation of alternative promoters.

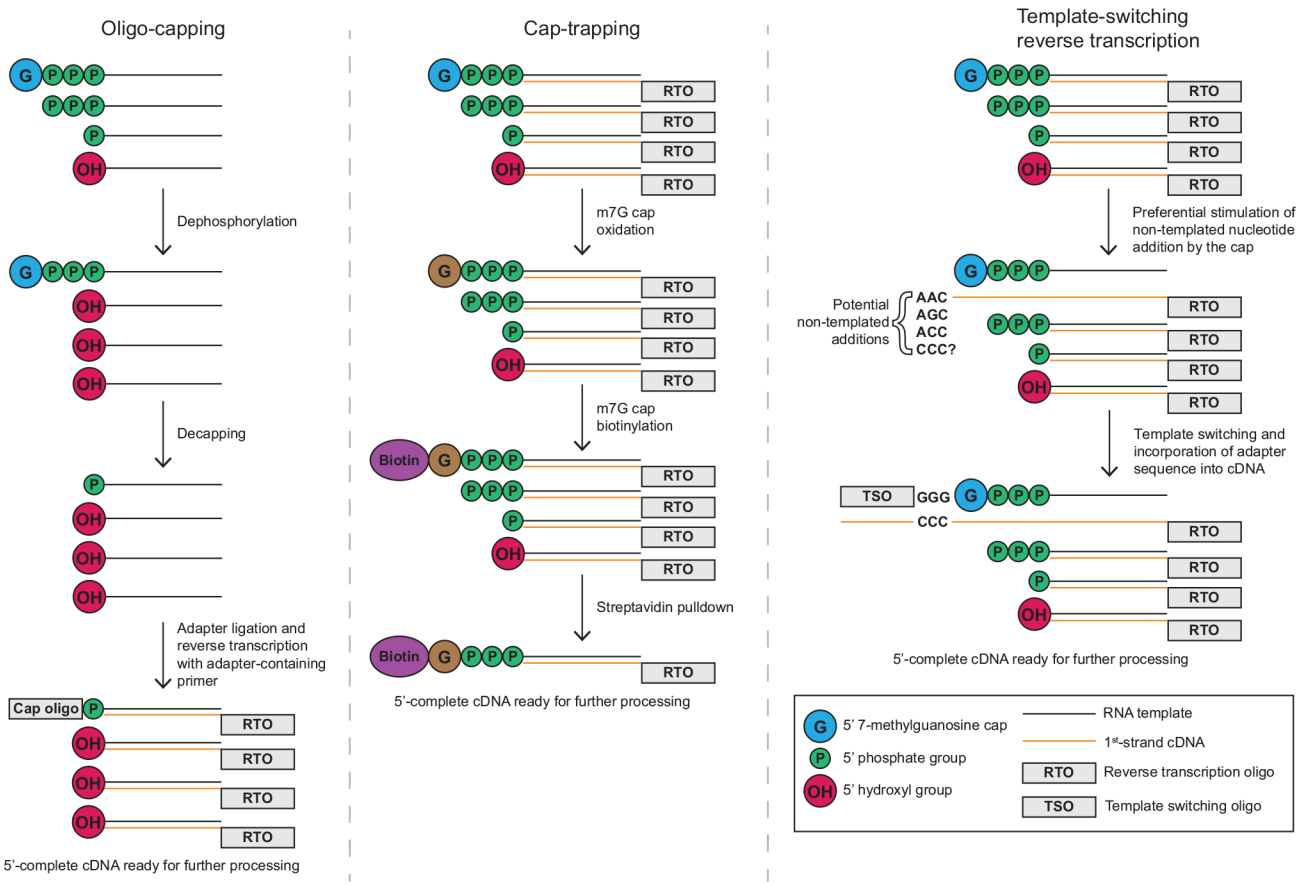

Figure 2. General approaches for TSS mapping

In oligo-capping, total RNA is first treated enzymatically to dephosphorylate uncapped RNAs. Caps are then removed, leaving 5′ monophosphates compatible with ligation. The cap oligo is ligated to the decapped RNAs and reverse transcription is performed, yielding 5′-complete cDNA ready for further processing. In cap-trapping, RNA:cDNA hybrids are chemically treated to oxidize RNA caps, which are then biotinylated. Streptavidin purification is then used to selectively enrich capped hybrids for further processing. In TSRT, total RNA is reverse transcribed, and the cap stimulates the addition of nontemplated nucleotides to the 3′ end of the first-strand cDNA. A TSO then interacts with the additional nucleotides and reverse transcriptase incorporates the complement of the TSO sequence into the first-strand cDNA, resulting in 5′-complete cDNA ready for further processing. See Table 1 for advantages and disadvantages of each approach and Table 2 for RNA input requirements.

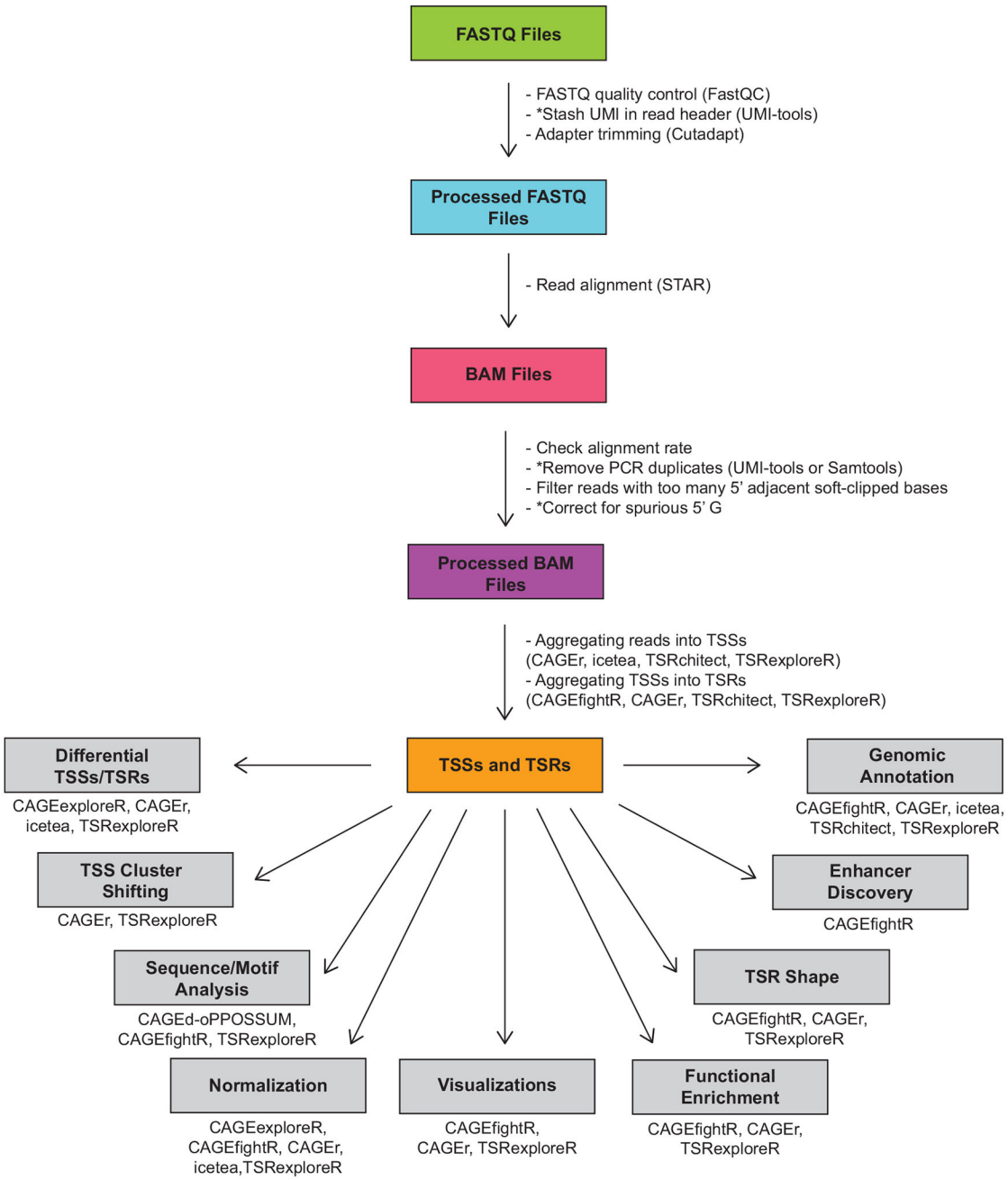

Figure 3. Computational processing of TSS mapping data

A general workflow for processing and analysis of TSS mapping data is shown, with software that can be used for each step indicated. Asterisks indicate optional steps. More information on each piece of software listed here can be found at the following URLs: FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc); UMI-tools (https://github.com/CGATOxford/UMI-tools); Cutadapt (https://cutadapt.readthedocs.io/en/stable); STAR (https://github.com/alexdobin/STAR); Samtools (http://www.htslib.org); CAGEr (https://www.bioconductor.org/packages/release/bioc/html/CAGEr.html); icetea (https://www.bioconductor.org/packages/release/bioc/html/icetea.html); TSRchitect (https://

www.bioconductor.org/packages/release/bioc/html/TSRchitect.html); TSRexploreR (https://zentnerlab.github.io/TSRexploreR/index.html); CAGEexploreR (https://github.com/edimont/CAGExploreR); CAGEd-oPPOSSUM (http://cagedop.cmmt.ubc.ca/CAGEd_oPOSSUM/); CAGEfightR (https://www.bioconductor.org/packages/release/bioc/html/CAGEfightR.html).

Table 1.

Advantages and disadvantages of TSS mapping approaches

Enzymatic approach	Methods	General comments	Method-specific features
Oligo-capping	TSS-seq, PEAT, CapSeq, TL-seq, TIF-seq, Start-seq, SMORE-seq	Removal of m7G cap prior to reverse transcription reduces prevalence of the 5′ G artifact, thus providing high TSS specificity. However, oligo-capping methods generally have high total RNA input requirements (see Table 2), complex protocols, and may suffer from the sequence biases of ligases used to attach oligo caps.	• TIF-seq/SMORE-seq: simultaneous mapping of 5′ and 3′ ends of transcripts. • Start-seq: enhanced TSS specificity due to isolation of short transcripts from nuclear RNA.
Cap-trapping	nAnT-iCAGE, SLIC-CAGE, MAPCap	Oligo-capping methods generally have lower input requirements than cap-trapping methods (see Table 2) and provide high spatial resolution and sensitivity but suffer from the 5′ G artifact due to reverse transcription of capped RNA. Cap-trapping-based protocols are relatively complex and can be expensive.	• SLIC-CAGE: uses selectively degradable carriers to facilitate processing of very small amounts of input RNA. • MAPCap: isolation of capped RNA with m7G immunoprecipitation versus the cap oxidation, biotinylation, and streptavidin pulldown used in CAGE methods simplifies this portion of the protocol; reduced prevalence of 5′ G artifact due to RT reaction conditions.
Template-switching reverse transcription	nanoCAGE-XL[a], nanoCAGE 2017[a], RAMPAGE[a] Tn5Prime, nanoPARE[a], STRIPE-seq	TSRT-based approaches generally have the lowest input requirements of all TSS mapping methods (SLIC-CAGE excepted, see Table 2). Their protocols tend to be simpler than those of oligo-capping and cap-trapping methods. NanoCAGE 2017, Tn5Prime, and nanoPARE use Tn5 tagmentation for library preparation, while STRIPE-seq uses stringent bead purifications to optimize library size distribution. These methods may suffer from reduced sensitivity in complex transcriptomes and are susceptible to the 5’ G artifact. In addition, several TSRT-based methods use custom sequencing primers, complicating pooling with other types of libraries.	• nanoCAGE-XL: the companion software, CapFilter, uses the 5′ G artifact as a “cap signature” to enhance TSS detection. • RAMPAGE: combines cap-trapping and TSRT for enhanced TSS specificity. • nanoPARE: enables parallel profiling of gene body RNA signal from a single sample; companion software provided (EndGraph). • STRIPE-seq: very simple and rapid protocol; low cost; companion software provided (GoSTRIPEs/TSRexploreR).

[a]Indicates that a custom sequencing primer is required for libraries of this type.

Table 2.

Reported input requirements of TSS mapping methods

Method	Enzymatic approach	Reported inputs
TSS-seq	Oligo-capping	200 μg total RNA (Yamashita et al., 2011); 500 ng poly(A)+ RNA (Malabat et al., 2015)
PEAT	Oligo-capping	1–2 μg poly(A)+ RNA (Ni et al., 2010); 30 μg total RNA (Morton et al., 2014)
CapSeq	Oligo-capping	500 ng–2 μg total RNA (Gu et al., 2012)
TL-seq	Oligo-capping	1 μg poly(A)+ RNA (Arribere and Gilbert, 2013)
TIF-seq	Oligo-capping	60 μg total RNA (Pelechano et al., 2013)
SMORE-seq	Oligo-capping	500 ng poly(A)+ RNA (Park et al., 2014)
nAnT-iCAGE	Cap-trapping	5 μg total RNA (Murata et al., 2014)
SLIC-CAGE	Cap-trapping	1–100 ng total RNA brought up to 5 μg with carrier (Cvetesic et al., 2018)
MAPCap	Cap-trapping	100 ng–5 μg total RNA (Bhardwaj et al., 2019)
nanoCAGE-XL	Template-switching reverse transcription	200 ng rRNA-depleted RNA (Cumbie et al., 2015); 7.5 μg total RNA (Adiconis et al., 2018)
nanoCAGE 2017	Template-switching reverse transcription	50–500 ng total RNA (Poulain et al., 2017); Single cell (C1 CAGE [Kouno et al., 2019])
RAMPAGE	Template-switching reverse transcription/cap-trapping	5 μg total RNA (Batut et al., 2013)
Tn5Prime	Template-switching reverse transcription	Single cell – 5 ng total RNA (Cole et al., 2018)
nanoPARE	Template-switching reverse transcription	10 pg (single-cell equivalent) – 5 ng total RNA (Schon etal., 2018)
STRIPE-seq	Template-switching reverse transcription	50–250 ng total RNA (Policastro et al., 2020)