Published online 1 June 2021

https://doi.org/10.1093/nargab/lqab051

Flexible analysis of TSS mapping data and detection of TSS shifts with TSRexploreR

Robert A. Policastro[1], **Daniel J. McDonald**[2], **Volker P. Brendel**[1,3] **and Gabriel E. Zentner**[1,4,*]

[1]Department of Biology, Indiana University, Bloomington, IN 47405, USA, [2]Department of Statistics, Indiana University, Bloomington, IN 47405, USA, [3]Department of Computer Science, Indiana University, Bloomington, IN 47405, USA and [4]Indiana University Melvin and Bren Simon Comprehensive Cancer Center, Indianapolis, IN 46202, USA

Received February 16, 2021; Revised April 29, 2021; Editorial Decision May 16, 2021; Accepted May 18, 2021

ABSTRACT

Heterogeneity in transcription initiation has important consequences for transcript stability and translation, and shifts in transcription start site (TSS) usage are prevalent in various developmental, metabolic, and disease contexts. Accordingly, numerous methods for global TSS profiling have been developed, including most recently Survey of TRanscription Initiation at Promoter Elements with high-throughput sequencing (STRIPE-seq), a method to profile transcription start sites (TSSs) on a genome-wide scale with significant cost and time savings compared to previous methods. In anticipation of more widespread adoption of STRIPE-seq and related methods for construction of promoter atlases and studies of differential gene expression, we built TSRexploreR, an R package for end-to-end analysis of TSS mapping data. TSRexploreR provides functions for TSS and transcription start region (TSR) detection, normalization, correlation, visualization, and differential TSS/TSR analyses. TSRexploreR is highly interoperable, accepting the data structures of TSS and TSR sets generated by several existing tools for processing and alignment of TSS mapping data, such as CAGEr for Cap Analysis of Gene Expression (CAGE) data. Lastly, TSRexploreR implements a novel approach for the detection of shifts in TSS distribution.

INTRODUCTION

Genome-wide mapping of transcription start sites (TSSs) is crucial to understanding gene regulation. Clusters of TSSs, referred to as transcription start regions (TSRs), are associated with promoter elements and represent genomic positions at which RNA polymerase has initiated synthesis of new RNA molecules. Variation in TSS usage alters the length of 5′ untranslated regions (5′ UTRs), which has been shown to influence transcript stability and translation (1–3) and is recognized as a major contributor to transcript isoform diversity in mammalian cells and tissues (4–6). Alternative initiation has also been described in human cancers (7) and inflammatory bowel diseases (8) as well as during development, particularly in zebrafish (9). Thus, understanding gene regulation under physiologic and pathologic conditions on a global scale requires accurate profiling of TSSs. To this end, numerous techniques have been developed, including Cap Analysis of Gene Expression (CAGE) (10), RNA Annotation and Mapping of Promoters for Analysis of Gene Expression (RAMPAGE) (11), and 5′ global run-on sequencing (GRO-cap) (12).

The recently introduced Survey of TRanscription Initiation at Promoter Elements with high-throughput sequencing (STRIPE-seq) (13) method provides a rapid, efficient, simple, and cost-effective TSS profiling approach, applicable to both genome-wide promoter atlas construction and expression profiling in samples with limited RNA input amounts. Here, we describe the product of synchronous code development to streamline analysis of STRIPE-seq data, as well as data resulting from other popular TSS detection methods. TSRexploreR is distributed as an R package for comprehensive and flexible exploration of TSS mapping data. TSRexploreR accepts pre-processed TSS and TSR data in a variety of common formats, including tab-delimited text files summarizing prior read mapping results or raw mapping data in BAM alignment format. TSRexploreR performs normalization for library size differences and implements a wide array of functions for subsequent derivation of summary and correlation statistics, visualization, and differential TSS/TSR analysis. Furthermore, TSRexploreR implements a novel approach to detect shifts in TSS distributions within TSRs. In sum, TSRexploreR is a feature-rich, interoperable, and easy-to-use software package for comprehensive analysis of TSS mapping data.

[*]To whom correspondence should be addressed. Tel: +1 812 856 7377; Fax: +1 812 855 6082; Email: gzentner@indiana.edu
Present address: Daniel J. McDonald, Department of Statistics, University of British Columbia, Vancouver, BC V6T 1Z4, Canada.

MATERIALS AND METHODS

TSRexploreR implementation

TSRexploreR is fully implemented in R (with the exception of TSS shifting analysis, described below) and makes use of numerous Bioconductor (https://bioconductor.org/) packages and CRAN (https://cran.r-project.org/) libraries such as tidyverse (https://cran.r-project.org/web/packages/tidyverse/index.html) and data.table (https://cran.r-project.org/web/packages/data.table/index.html). Data is stored in a TSRexploreR S4 object in common Bioconductor formats such as GenomicRanges (GRanges) (14) or as a data.table for rapid, memory-efficient manipulation. TSRexploreR accepts bedGraph, bigWig, CTSS, and tab-delimited table files for TSSs and BED and tab-delimited table files for TSRs. Alignment BAM files can also be processed by TSRexploreR, as described below. A full list of TSRexploreR functions, with accompanying documentation, can be found at https://zentnerlab.github.io/TSRexploreR/reference/index.html. TSRexploreR is packaged with STRIPE-seq-detected TSSs along budding yeast chromosome IV alongside the Ensembl release 99 budding yeast V64-1-1 genome sequence and annotation GTF. TSRexploreR is available at https://github.com/zentnerlab/TSRexploreR/releases/tag/v0.2.0 and as a Singularity container from Singularity Library (https://cloud.sylabs.io/library; download command: singularity pull library://zentlab/default/tsrexplorer:main), ensuring prolonged compatibility and reproducibility. We also provide a feature comparison of TSRexploreR with CAGEr (15), the most full-featured TSS analysis software published to this point (Supplementary Table S1).

BAM processing

Alignments in BAM format are loaded into TSRexploreR using the GenomicAlignments package (14) and can be processed as needed during import. GenomicAlignments considers the 5′-most non-soft-clipped base as the start position of the R1 read (and thus the TSS), but 5′ soft-clipped base information is retained in the cigar string (from the BAM file) and thus further exploration is possible. Taking advantage of this fact, an analysis of soft-clipping is performed, where reads having more than a user-specified number of soft-clipped bases are removed. Filtering based on BAM flags is also performed, enabling removal of secondary alignments and, for paired-end reads, removal of unpaired or improperly paired reads and read pairs flagged as duplicates based on identical start and end positions. It has been frequently observed in both CAGE and template-switching reverse transcription (TSRT)-based methods such as nanoCAGE and STRIPE-seq that a non-specific G (corresponding to C on the first-strand cDNA) is often present at the 5′-most position of the R1 read (16,17). This is most likely due to reverse transcription of the cap (18). To correct for this artifact, we determine the frequency of reads with soft-clipped 5′ G bases (that is, cap-templated Gs that do not fortuitously map to the genome). We illustrate the detection of soft-clipped bases by TSRexploreR using nanoCAGE, nAnT-iCAGE, SLIC-CAGE (19), and TSS-seq (1) data from the BY4741 yeast strain. As all of the CAGE-based methods rely on reverse transcription of capped RNA, they are susceptible to the artifact, while TSS-seq, which involves cap removal prior to reverse transcription, is presumably not. We found that ∼41.6–44.4% of reads from each CAGE experiment had soft-clipped bases, the majority of which were single Gs, while <2% of reads from each TSS-seq experiment had soft-clipped bases (Supplementary Figure S1). We assume that the frequency of soft-clipped Gs is similar to that of cap-templated, genome-matching G addition and therefore this frequency to determine if the non-soft-clipped-G should be removed. For each read with a 5′ G following removal of soft-clipped bases, a Bernoulli trial is conducted using the aforementioned soft-clipped G frequency as the 'success' probability to decide if the G should be removed, which is similar in principle to the approach used with CAGE data (16). For methods not subject to this artifact, such as TSS-seq, this step can be skipped. Following the optional G correction, overlapping 5′ read ends are aggregated into TSSs.

TSRexplorer vignettes

Step-by-step vignettes for performing common tasks in TSRexploreR are available at the following URLs:

BAM import and processing:
https://github.com/zentnerlab/TSRexploreR/blob/v0.2.0/documentation/BAM_PROCESSING.pdf

Standard TSS/TSR exploration:
https://github.com/zentnerlab/TSRexploreR/blob/v0.2.0/documentation/STANDARD_ANALYSIS.pdf

Differential feature analysis:
https://github.com/zentnerlab/TSRexploreR/blob/v0.2.0/documentation/DIFF_FEATURES.pdf

TSS shifting analysis:
https://github.com/zentnerlab/TSRexploreR/blob/v0.2.0/documentation/FEATURE_SHIFT.pdf

Data conditioning:
https://github.com/zentnerlab/TSRexploreR/blob/v0.2.0/documentation/DATA_CONDITIONING.pdf

TSS and TSR analysis

Yeast nAnT-iCAGE CTSSs (20) were obtained from www.yeastss.org (21) and imported into TSRexploreR. The nine growth conditions analyzed were: log-phase growth in rich yeast-peptone-dextrose medium (YPD, the control condition), cell cycle arrest with α-factor, DNA damage induced by methyl methanesulfonate (DD), diauxic shift (DSA), YP medium with 16% glucose to induce fermentation (Glc), log-phase growth in yeast-peptone-galactose medium (Gal), oxidative stress induced by H_2O_2, 37°C heat shock (HS), and osmotic stress induced by NaCl. For genome assembly and annotation we used the R packages 'BSgenome.Scerevisiae.UCSC.sacCer3' v1.4.0 (22) and 'TxDb.Scerevisiae.UCSC.sacCer3.sgdGene' v3.2.2 (23), respectively. Code used for yeast CAGE analysis is available at https://github.com/zentnerlab/Policastro_etal_2021/tree/v0.1.1.

Zebrafish TSS shifting analysis

For TSRexploreR analysis, zebrafish developmental CAGE data (24) were obtained as TPM-normalized bigWig files from http://promshift.genereg.net/zebrafish/CAGE/ and imported into TSRexploreR. Scores for negative-strand TSSs were multiplied by -1 to yield positive values. For CAGEr analysis, datasets were imported into CAGEr v1.30.3 from the R package 'ZebrafishDevelopmentalCAGE' v0.99.0 (http://promshift.genereg.net/CAGEr/PackageSource/) and TPM normalized using the power-law approach. For both methods, TSSs supported by ≥3 TPM in one of the two samples were clustered into TSRs with a maximum clustering distance of 25 bp and a maximum TSR width of 250 bp. TSRs supported by at least 10 TPM in both samples were merged if they were within 100 bp of one another. An FDR threshold of 0.05 was used to assess the significance of shifting results from both approaches. Code used for shifting analysis is available at https://github.com/zentnerlab/Policastro_etal_2021/tree/v0.1.1.

RESULTS AND DISCUSSION

Analysis of yeast CAGE data with TSRexploreR

To demonstrate the features of TSRexploreR, we analyzed CAGE CTSSs from yeast cells grown under a variety of conditions (20). In cases where a single plot is shown, this indicates a result from one YPD control replicate.

Genomic annotation and threshold exploration. Using annotations provided in GTF or TxDb format, TSRexploreR links TSSs to known genomic features (25). Assignment of TSSs to such features, particularly promoters, is useful in establishing a read threshold for downstream analyses. TSRexploreR threshold analysis determines the fraction of TSSs that is promoter-proximal and the number of transcripts or genes with at least one unique TSS across a range of raw read count thresholds. This analysis allows selection of a threshold that balances removal of likely artifacts (in particular, weak TSS signals within gene bodies) with detection of authentic lowly abundant promoter-proximal TSSs. Using a promoter definition of -250 to +100 bp relative to annotated gene starts (start codons for mRNA genes and TSSs for ncRNA genes from the TxDb.Scerevisiae.UCSC.sacCer3.sgdGene Bioconductor package (23)), we selected a threshold of 10 counts/TSS, yielding promoter-proximal TSS fractions of 0.603–0.750 (Figure 1B, Supplementary Figure S2). Following annotation and thresholding, the distribution of TSSs relative to known genomic features can be visualized as stacked barplots (Figure 1C). A feature detection plot, wherein the number of genes or transcripts with at least one unique TSS position meeting the specified threshold is displayed, can also be generated (Figure 1D). The functions used to generate the plots described in this section, and many other TSRexploreR plotting functions, return ggplot objects that can be customized according to standard ggplot2 syntax (26). The 'Value' section of each function's documentation indicates what is returned, including whether it returns a ggplot object in the case of plotting functions.

Normalization. TSRexploreR includes three options for normalization. The first, counts per million (CPM), is a simple read number-based normalization approach commonly used for data visualization and is particularly appropriate for replicate comparison. However, CPM normalization is considered simplistic when comparing samples from distinct biological conditions (27) and we thus implemented two additional normalization approaches considered more appropriate for such cases: trimmed mean of M-values (TMM) (27), used in edgeR (28), and median-of-ratios (MOR), used in DESeq2 (29). For this example, data were normalized using the MOR approach. Normalized samples can be compared via a principal component analysis (PCA) plot (30) (Figure 1E) or correlation heatmaps (31) (Supplementary Figure S3).

Visualization of TSS data. TSSs can be exported in bedGraph, bigWig, or tab-delimited table format. TSSs and/or TSRs at a specific gene or its promoter can also be directly visualized using Gviz (32). To demonstrate this feature, we plotted MOR-normalized TSS counts from one replicate each of the YPD and Gal conditions at the promoter of the *GAL2* gene, which encodes a permease required for galactose utilization (Figure 1F). TSS signal around gene starts (start codons for mRNA genes, TSSs for ncRNA genes) can also be displayed as a heatmap (Figure 1G), and the distribution of TSS positions relative to annotated gene starts can be visualized as a density plot (Figure 1H).

TSS sequence analysis. TSRexploreR includes several functions for analyzing the sequence context of TSSs. Furthermore, as it is often desirable to assess specific subsets of TSSs or TSRs, TSRexploreR includes a number of conditioning functions for grouping, ordering, quantiling, and filtering data. Here, we demonstrate these features for the purpose of TSS sequence analysis. We determined the dominant TSS associated with each analyzed transcript, split them into quintiles by score (that is, the total sum of read 5′ end counts at the TSS position), and plotted sequence logos (36) (Figure 2A). This analysis indicates a pyrimidine (Y = C or T) preference at the -1 position and a purine (R = A or G) preference at the +1 position (the TSS itself), as well as an A base in the -8 position (Figure 2A), consistent with previous studies (20,37,38). It is seen that these preferences decrease in terms of information content with decreasing TSS score (that is, the number of 5′ read ends supporting the TSS). Sequences surrounding TSSs can also be visualized as color plots (Figure 2B). Lastly, the frequencies of all observed –1/+1 dinucleotides can be plotted (Figure 2C).

TSR detection and analysis. TSRexploreR uses a simple distance-based clustering approach to aggregate TSSs into TSRs based on a user-specified TSS count threshold that must be met in a specified number of samples and maximum inter-TSS distance. For this analysis, we used a raw count threshold of 10 in at least one sample, a maximum distance of 25 bp, and a maximum TSR width of 250. Many of the analyses described above for TSSs can also be applied to TSRs: correlation, analysis of genomic distribution,

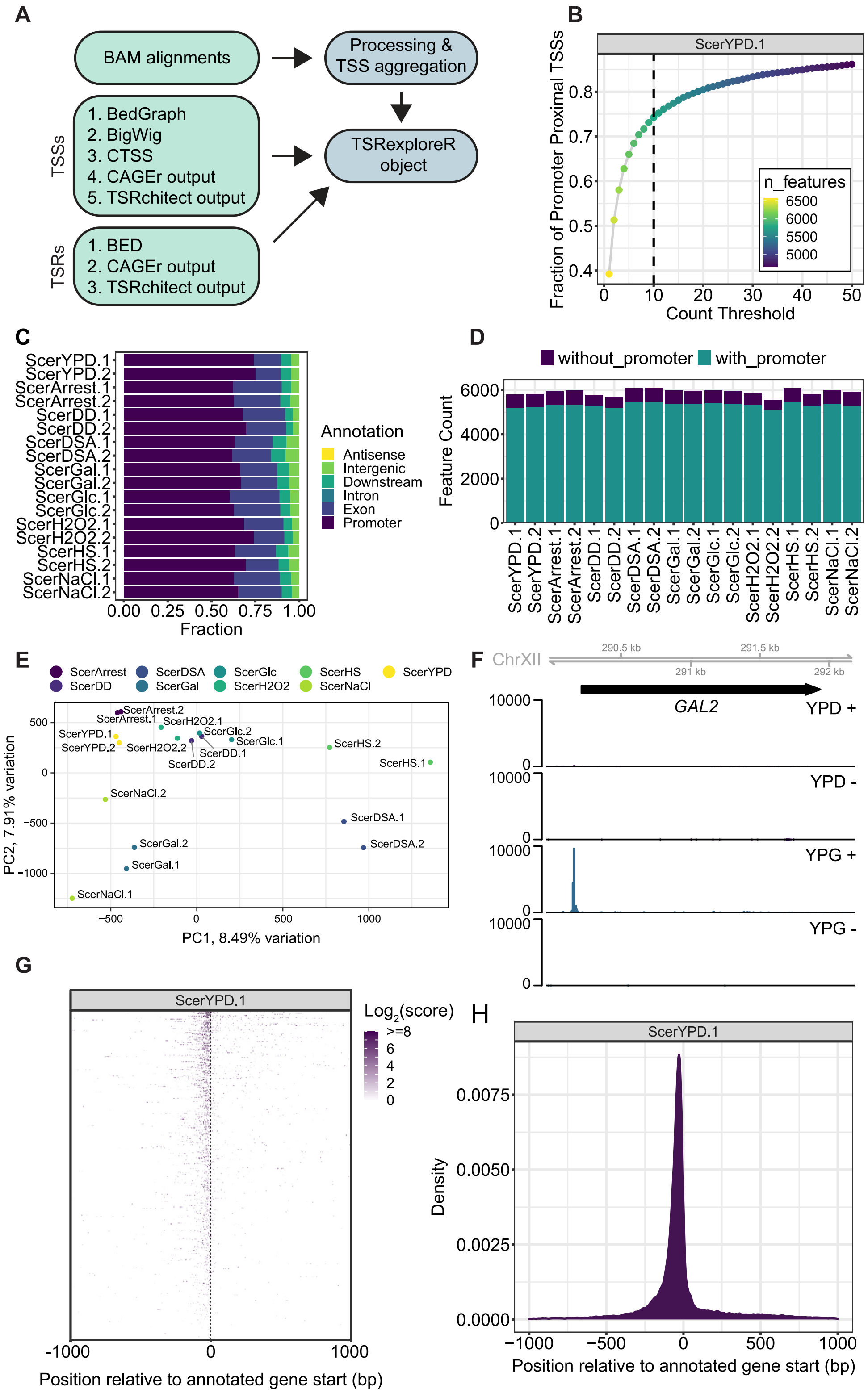


Figure 1. TSS analysis with TSRexploreR. (**A**) Schematic depicting input formats accepted by TSRexploreR and creation of the TSRexploreR object. (**B**) Threshold plot showing the fraction of TSSs that is promoter-proximal (-250 to +100 relative to annotated gene starts (start codons for mRNA genes, TSSs for ncRNA genes)) and the number of features (in this case, transcripts) with at least one unique TSS position at each threshold in YPD replicate 1. (**C**) Barplot of the genomic distribution of TSSs in each sample. (**D**) Barplot of the number of transcripts with a unique TSS position in each sample, and whether that TSS is promoter-proximal or not. (**E**) PCA plot of TSSs detected in each CAGE sample. (**F**) Signal tracks of normalized signal (YPD and Gal replicate 1) at the *GAL2* locus. (**G**) Heatmap of normalized signal from YPD replicate 1 relative to annotated gene starts, sorted descending by total signal. (**H**) Density plot of unique TSS positions relative to annotated gene starts for YPD replicate 1.

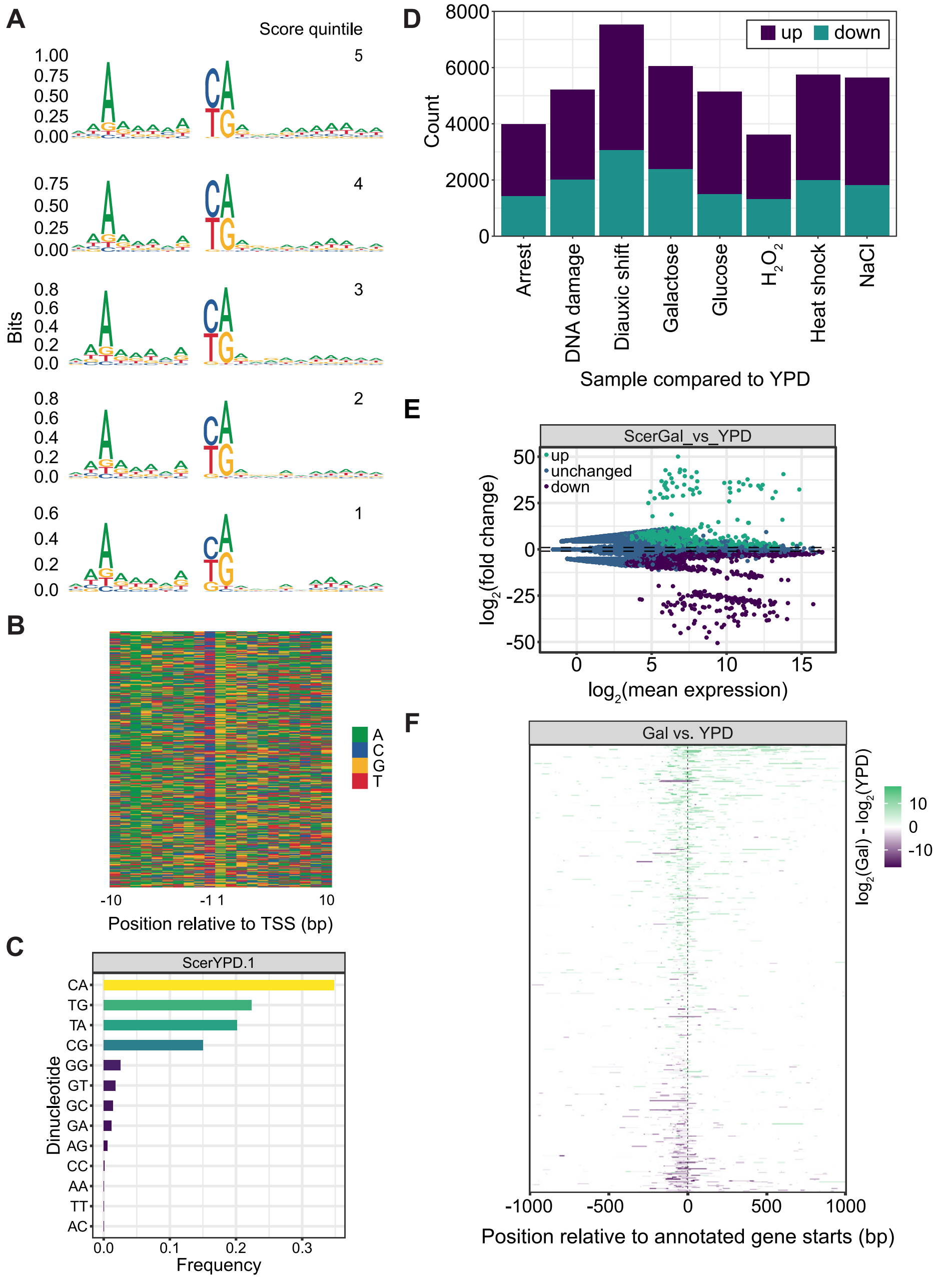


Figure 2. Sequence analysis and differential TSR detection. (**A**) Sequence logos (quintiled in descending order by TSS score (the total sum of read 5′ end counts at the TSS position), (**B**) base color plot (in descending order by TSS score) and (**C**) barplot of dinucleotide frequencies at the dominant TSS of each transcript in YPD replicate 1. (**D**) Barplot of the number of differentially expressed TSRs for comparison of each indicated condition to the YPD control. (**E**) MA plot of differential TSR results for the Gal versus YPD comparison. (F) Difference heatmap of $\log_2$(Gal replicate 1) – $\log_2$(YPD replicate 1) CAGE signal relative to annotated gene starts (start codons for mRNA genes, TSSs for ncRNA genes), sorted descending by magnitude of the difference.

feature detection(iand density/signal relative to annotated gene starts (start codons for mRNA genes, TSSs for ncRNA genes).

Characterization of TSR features. It has been well established that there is a continuum of TSR shapes ranging from sharp or peaked, wherein transcription initiates at one or a few strong TSSs, to broad or dispersed, wherein there are several TSSs of similar strength (33,34). TSRexploreR calculates three metrics relating to TSR shape: (i) shape index (SI), which assesses the shape of TSRs via analysis of the position of each constituent TSS and its strength relative to the overall score of the TSR (33); (ii) inter-quantile range (IQR), which measures the distance between the base positions of the user-specified TSS signal quantiles, providing information on the width of a TSR without being affected by weak TSSs on its edges (15) and (iii) peak balance, which assesses the skew of TSSs around the TSR center, analogous to the torque metric calculated by TSRchitect (35) (Supplementary Figure S4).

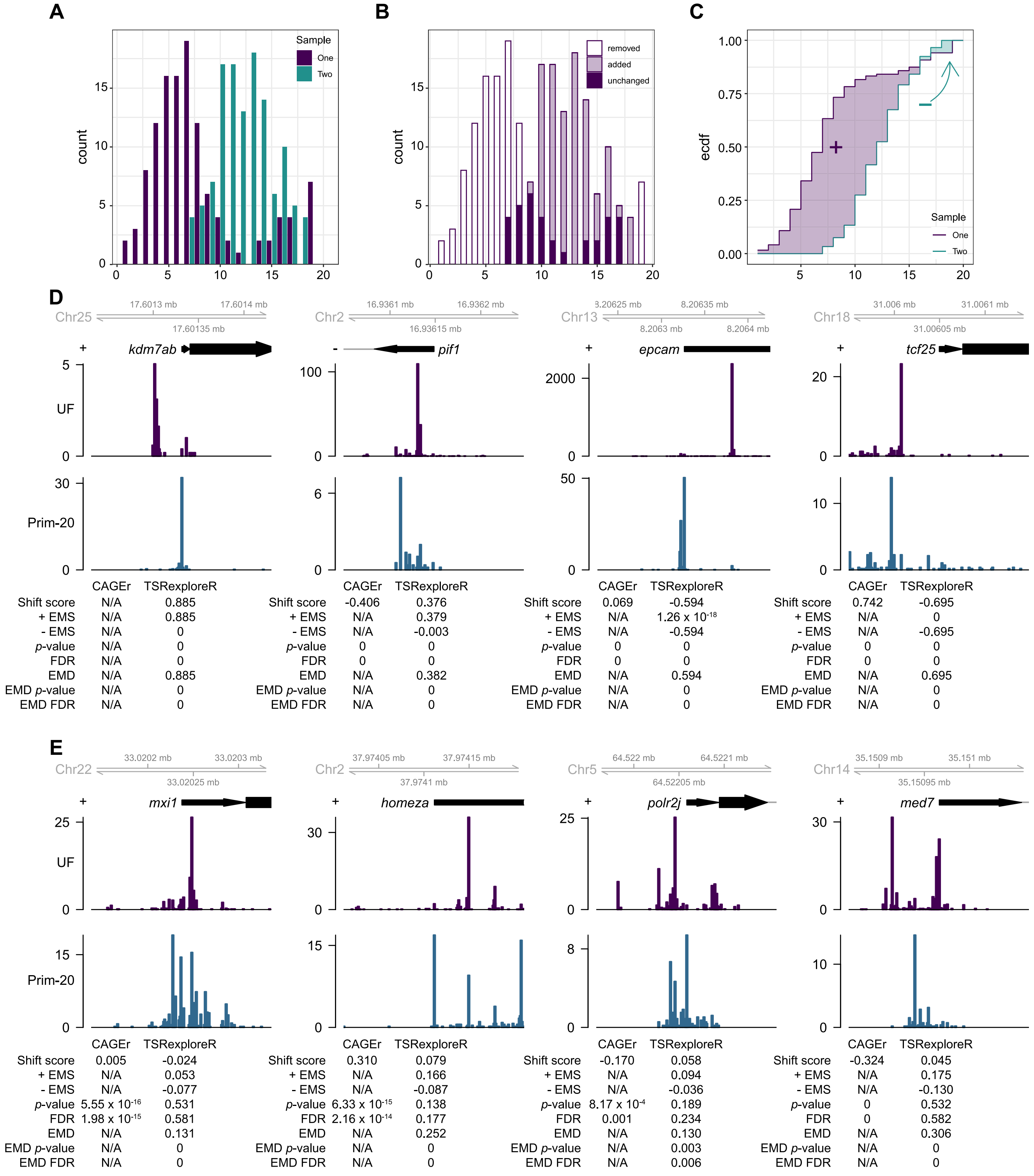
Analysis of differential TSR usage. Transcription is highly dynamic and plastic, able to respond quickly to various stimuli. To enable analysis of differential TSS and TSR usage across distinct conditions, TSRexploreR generates matrices of counts within merged regions that are used as input to edgeR (28) or DESeq2 (29). We used DESeq2 to build a statistical model and then performed contrasts of treated samples versus the control YPD condition (see Supplementary Table S2 for full differential TSR analysis results). As an overview of differential feature analysis, a stacked barplot of the number of changed features in each contrast can be generated (Figure 2D). The results of individual comparisons (for this example, Gal versus YPD) can also be visualized as an MA plot, displaying $\log_2$(fold change) versus mean expression (Figure 2E), and as a volcano plot, displaying $-\log_{10}$(adjusted P-value) versus $\log_2$(fold change) (Supplementary Figure S5). We also demonstrate visualization of differential TSR signal with a differential heatmap. In this example, $\log_2$(YPD signal) is subtracted from $\log_2$(Gal signal) at all annotated gene starts (start codons for mRNA genes, TSSs for ncRNA genes) with an associated TSR in one or both samples (Figure 2F). To facilitate interpretation of differential feature analysis results, TSRexploreR annotates differential features and exports a list of gene names compatible with clusterProfiler, a robust R package for gene ontology (GO) analysis (39). Genes associated with upregulated promoter-proximal TSRs were enriched for GO biological process terms including 'carbohydrate metabolic process' and 'generation of precursor metabolites and energy' (Supplementary Figure S6, Supplementary Table S3). Genes with downregulated promoter-proximal TSRs were enriched for processes related to ribosome biogenesis (Supplementary Figure S6, Supplementary Table S3), consistent with previous work showing reduced levels of ribosomal protein gene transcripts in cells grown continuously in galactose (40).

Detection of TSS shifts with TSRexploreR

Numerous studies indicate that large-scale shifts in TSS distribution are prevalent in various developmental settings (9) and are induced in response to environmental stimuli (41) and mutations in general transcription factors (38). Computational detection of TSS shifts may be approached as testing for differences between two discrete probability distributions. The CAGEr package (15) assesses spatial shifts in TSS usage by generating aggregate TSRs from TSRs identified in all samples and comparing empirical cumulative distribution functions (ECDFs), where the sample with larger total signal has its ECDF rescaled by the ratio of total signal. This results in a score between negative infinity and 1, with larger positive values posited to indicate that a given proportion of signal in the second sample is outside of the TSS-containing region in the first sample. For instance, a CAGEr shift score of 0.4 would indicate that at least 40% of the transcription initiation in the second sample is independent of that in the first sample. This approach only assesses spatial separation between two distributions and thus does not address shifts in signal distribution at largely overlapping positions. Furthermore, it produces a substantial number of negative shift scores, the interpretation of which can be unclear. Lastly, it does not indicate shift direction. In addition to calculating the shifting score, CAGEr also performs a Kolmogorov–Smirnov (K–S) test on the ECDFs, identifying the point of maximal distance between them. The stated purpose of the K–S test is 2-fold: (i) to assess significance of the observed difference in TSS distribution between the two samples and (ii) to capture changes in TSS distributions within mostly overlapping positions that are not captured by the shift score. However, calculation of the K–S statistic is unrelated to the shift score, and its P-value therefore does not indicate the score's significance. Furthermore, the derivation of the P-value formula for the K–S test assumes the data come from a continuous distribution, an assumption not met by TSS distributions, which are composed of observations at discrete locations.

Given these limitations, we implement an alternative approach to detecting TSS shifts using a more intuitive metric. We use a signed version of earth mover's distance (EMD) (42), which we refer to as earth mover's score (EMS), to characterize between-sample differences in TSS distributions within merged TSRs. For this approach, we imagine that the two TSS distributions in question are piles of dirt and ask how much dirt from one pile we would need to move, how far, and in which direction, to recreate the distribution of the other sample. The computed EMS thus represents the minimum 'cost' of converting one distribution into the other. The resulting score is between –1 and 1, with larger magnitudes indicating larger shifts and the sign indicating direction (negative values indicate upstream shifts and positive values indicate downstream shifts). Figure 3A–C illustrates the intuition for calculation of the EMS. We note that the EMS is calculated by summing the overall positive (downstream) and negative (upstream) differences between the two TSS distributions. Thus, EMS indicates the overall direction of the shift, even if a notable amount of shifting occurs in the opposite direction. To allow independent assessment of the degree of upstream and downstream shifting within a given TSR, TSRexploreR also reports the positive and negative components of the EMS alongside the final shift score. We also considered the possibility that relatively balanced upstream and downstream shifting could be obscured by the signed nature of the EMS. Possible examples of balanced shifting include TSR splitting or merging as well as an overall change in the shape of a TSR (e.g., peaked to broad). In order to capture such shifts, we also report the unsigned EMD, which indicates how much total TSS 'mass' has been shifted between two samples without regard to direction. TSRexploreR calculates EMS and EMD as well as a P-value and FDR threshold for both based on permutation tests. TSS shifting analysis is implemented in C++ to enhance execution speed.

To demonstrate the capacity of our EMS/EMD-based approach to detect TSS shifts, we turned to a set of CAGE experiments performed throughout zebrafish embryonic development. Detailed analysis of this dataset revealed distinct distributions of TSSs for the maternally-deposited and zygotic forms of several hundred transcripts (9), and so it provides an appealing test case. We compared the earliest and latest time points assayed (unfertilized egg and Prim-20, respectively) using both the established CAGEr approach and our EMS/EMD-based method. Using CAGEr



D

	CAGEr	TSRexploreR
Shift score	N/A	0.885
+ EMS	N/A	0.885
- EMS	N/A	0
p-value	N/A	0
FDR	N/A	0
EMD	N/A	0.885
EMD *p*-value	N/A	0
EMD FDR	N/A	0

	CAGEr	TSRexploreR
Shift score	-0.406	0.376
+ EMS	N/A	0.379
- EMS	N/A	-0.003
p-value	0	0
FDR	0	0
EMD	N/A	0.382
EMD *p*-value	N/A	0
EMD FDR	N/A	0

	CAGEr	TSRexploreR
Shift score	0.069	-0.594
+ EMS	N/A	1.26 x 10^{-18}
- EMS	N/A	-0.594
p-value	0	0
FDR	0	0
EMD	N/A	0.594
EMD *p*-value	N/A	0
EMD FDR	N/A	0

	CAGEr	TSRexploreR
Shift score	0.742	-0.695
+ EMS	N/A	0
- EMS	N/A	-0.695
p-value	0	0
FDR	0	0
EMD	N/A	0.695
EMD *p*-value	N/A	0
EMD FDR	N/A	0

E

	CAGEr	TSRexploreR
Shift score	0.005	-0.024
+ EMS	N/A	0.053
- EMS	N/A	-0.077
p-value	5.55 x 10^{-16}	0.531
FDR	1.98 x 10^{-15}	0.581
EMD	N/A	0.131
EMD *p*-value	N/A	0
EMD FDR	N/A	0

	CAGEr	TSRexploreR
Shift score	0.310	0.079
+ EMS	N/A	0.166
- EMS	N/A	-0.087
p-value	6.33 x 10^{-15}	0.138
FDR	2.16 x 10^{-14}	0.177
EMD	N/A	0.252
EMD *p*-value	N/A	0
EMD FDR	N/A	0

	CAGEr	TSRexploreR
Shift score	-0.170	0.058
+ EMS	N/A	0.094
- EMS	N/A	-0.036
p-value	8.17 x 10^{-4}	0.189
FDR	0.001	0.234
EMD	N/A	0.130
EMD *p*-value	N/A	0.003
EMD FDR	N/A	0.006

	CAGEr	TSRexploreR
Shift score	-0.324	0.045
+ EMS	N/A	0.175
- EMS	N/A	-0.130
p-value	0	0.532
FDR	0	0.582
EMD	N/A	0.306
EMD *p*-value	N/A	0
EMD FDR	N/A	0

Figure 3. Detection of TSS shifts using earth mover's score. (**A**) Stylized TSS distributions for two samples at a hypothetical region of interest. (B) Illustration of how Sample One would need to be 'moved' in order to match Sample Two. Material (or 'earth') must be moved from the empty bars into the shaded bars while the solid bars remain unchanged. Some material has to be shifted in both directions, but more is moved upstream than downstream. Calculating how much, how far, and which 'piles" to move is a standard constrained optimization problem known by the name 'optimal transport', but here it reduces to a simple integral. (**C**) Calculation of the EMS for the hypothetical example illustrated in (B). The upstream (green, negative) and downstream (purple, positive) areas between the ECDFs are simply integrated and then subtracted from each other. The result is normalized to the number of locations with expression in either sample. This example has an EMS of 0.243 with a *P*-value of 0 based on an approximate permutation test using 1000 resamples. (**D**) Tracks of zebrafish CAGE signal from the unfertilized egg (UF) and Prim-20 stages at four promoters displaying significant shifts in TSS distribution by the EMS-based shift score method implemented in TSRexploreR. Note that only the strand from which the TSS signal originates is shown. (**E**) Same as (D) but at four promoters with significant shifts in TSS distribution by EMD only.

with no shift score threshold, we detected 3,950 significantly shifted TSRs, while applying a shift score threshold of 0.4 yielded 1,314 significant shifts (Supplementary Table S4). EMS/EMD-based analysis yielded 3,782 significantly shifted TSRs (Supplementary Table S5); we note that this number is slightly variable due to the use of a permutation test for determining significance.

To illustrate the relationship between our EMS-based shift score and TSS redistribution, we visualized data at several loci displaying various degrees of shifting (Figure 3D). At *kdm7ab*, CAGE signal was markedly shifted downstream in the Prim-20 sample, yielding an EMS of 0.885 ($P = 0$; all reported P-values reflect a false discovery rate (FDR) correction by the Benjamini–Hochberg procedure) and an equivalent EMD; this shift was not detected by CAGEr. A more modest downstream shift was observed at *pif1* (shift score = 0.376, $P = 0$; EMD = 0.382, $P = 0$). The *pif1* shift was detected as highly significant by CAGEr ($P = 0$), though the shift score was negative (-0.406). At *epcam*, we detected a robust upstream shift (shift score = -0.594, $P = 0$; EMD = 0.594, $P = 0$); this shift was also detected by CAGEr, though with a very small shift score (0.069). Lastly, at *tcf25*, we observed a highly significant upstream shift (EMS = -0.695, $P = 0$; EMD = 0.695, $P = 0$). The *tcf25* shift was also marked as highly significant by CAGEr ($P = 0$), but the sign of the robust shift score (0.742) does not relate to the direction of the shift. Overall, 2913/3782 (77%) of the detected shifts had a significant shift score, while the remaining 869 regions were significant only by EMD (Supplementary Table S5).

The detection of a large number of shifts with a significant EMD but not shift score suggests the detection of balanced shifts, wherein there are similar degrees of positive and negative shifting that would be effectively cancelled out by the signed nature of the shift score. To explore this, we examined a number of these regions. At *mxi1*, we observed expansion of the TSR in both directions in the Prim-20 sample, leading to a small, insignificant shift score (–0.024, $P = 0.581$) (Figure 3E). However, this region was highly significantly shifted by EMD (EMD = 0.131, $P = 0$). The *mxi1* shift was detected by CAGEr, though its reported magnitude was low (shift score = 0.005, $P = 1.98 \times 10^{-15}$). At the *homeza* promoter TSR, we observed markedly increased usage of two TSSs on its edges in the Prim-20 sample, leading to an overall broadening of its TSS signal (shift score = 0.079, $P = 0.177$; EMD = 0.252, $P = 0$; CAGEr shift score = 0.310, $p = 2.16 \times 10^{-14}$). At *polr2j*, we observed the opposite trend, wherein CAGE signal is broad in the unfertilized egg and narrower in Prim-20 (shift score = 0.058, $P = 0.234$; EMD = 0.130, $P = 0.006$; CAGEr shift score = –0.170, $P = 0.001$). At *med7*, there are two strong TSS peaks in the unfertilized egg, while Prim-20 transcription primarily initiates from a position between these peaks (shift score = 0.045, $P = 0.582$; EMD = 0.306, $P = 0$; CAGEr shift score = –0.324, $P = 0$). We conclude that EMD can be used to detect balanced shifts not considered by EMS due to cancellation of upstream and downstream components of the shift.

CONCLUDING REMARKS

TSRexploreR leverages the extensive Bioconductor and tidyverse programming environments to provide a feature-rich and straightforward tool for TSS mapping analysis and also incorporates a novel approach to detecting TSS shifts. While TSRexploreR was originally developed to handle STRIPE-seq data, it has been made highly interoperable and can thus be readily incorporated into workflows using existing TSS analysis software such as CAGEr (15), TSRchitect (35), and CAGEfightR (43). Global TSS profiling methods are currently being used to explore shifts in TSS usage in many biological contexts. In yeast, TSS mapping has been used to explore alternative initiation in many contexts, including the response to external stimuli (20), chromatin remodeler depletion (44,45), transitions between mitotic and meiotic growth (46), promoter evolution (47), and mutation of general transcription factors and RNA polymerase II subunits (38). TSS mapping has also recently been used to assess alternative initiation in mouse germline development (48,49) and sex-specific enhancer and promoter use in *Drosophila* (50). We thus envision that TSRexploreR will be a broadly useful tool for the analysis of data generated by TSS mapping studies in many areas of biological inquiry.

SUPPLEMENTARY DATA

Supplementary Data are available at NARGAB Online.

ACKNOWLEDGEMENTS

We thank Jason Tourigny for assistance in testing TSRexploreR functions and R. Taylor Raborn for helpful comments on the manuscript.

FUNDING

National Science Foundation [IOS-1221984 to V.P.B.]; National Institutes of Health [R35GM128631 to G.E.Z]. Funding for open access charge: National Science Foundation [IOS-1221984 to V.P.B.]; National Institutes of Health [R35GM128631 to G.E.Z].

Conflict of interest statement. None declared.

REFERENCES

1. 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.
2. Rojas-Duran,M.F. and Gilbert,W.V. (2012) Alternative transcription start site selection leads to large differences in translation activity in yeast. *RNA*, **18**, 2299–2305.
3. Arribere,J.A. and Gilbert,W.V. (2013) Roles for transcript leaders in translation and mRNA decay revealed by transcript leader sequencing. *Genome Res.*, **23**, 977–987.
4. Feng,G., Tong,M., Xia,B., Luo,G.-Z., Wang,M., Xie,D., Wan,H., Zhang,Y., Zhou,Q. and Wang,X.-J. (2016) Ubiquitously expressed genes participate in cell-specific functions via alternative promoter usage. *EMBO Rep.*, **17**, 1304–1313.
5. 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.
6. Pal,S., Gupta,R., Kim,H., Wickramasinghe,P., Baubet,V., Showe,L.C., Dahmane,N. and Davuluri,R.V. (2011) Alternative transcription exceeds alternative splicing in generating the transcriptome diversity of cerebellar development. *Genome Res.*, **21**, 1260–1272.
7. Demircioğlu,D., Cukuroglu,E., Kindermans,M., Nandi,T., Calabrese,C., Fonseca,N.A., 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.

8. Boyd,M., Thodberg,M., Vitezic,M., Bornholdt,J., Vitting-Seerup,K., Chen,Y., Coskun,M., Li,Y., Lo,B.Z.S., Klausen,P. *et al.* (2018) Characterization of the enhancer and promoter landscape of inflammatory bowel disease from human colon biopsies. *Nat. Commun.*, **9**, 1661.
9. Haberle,V., Li,N., Hadzhiev,Y., Plessy,C., Previti,C., Nepal,C., Gehrig,J., Dong,X., Akalin,A., Suzuki,A.M. *et al.* (2014) Two independent transcription initiation codes overlap on vertebrate core promoters. *Nature*, **507**, 381–385.
10. Murata,M., Nishiyori-Sueki,H., Kojima-Ishiyama,M., Carninci,P., Hayashizaki,Y. and Itoh,M. (2014) Detecting Expressed Genes Using CAGE. In: Miyamoto-Sato,E., Ohashi,H., Sasaki,H., Nishikawa,J. and Yanagawa,H. (eds). *Transcription Factor Regulatory Networks: Methods and Protocols*. Springer, NY, pp. 67–85.
11. Batut,P., Dobin,A., Plessy,C., Carninci,P. and Gingeras,T.R. (2013) High-fidelity promoter profiling reveals widespread alternative promoter usage and transposon-driven developmental gene expression. *Genome Res.*, **23**, 169–180.
12. Core,L.J., Martins,A.L., Danko,C.G., Waters,C., Siepel,A. and Lis,J.T. (2014) Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. *Nat. Genet.*, **46**, 1311–1320.
13. Policastro,R.A., Raborn,R.T., Brendel,V.P. and Zentner,G.E. (2020) Simple and efficient profiling of transcription initiation and transcript levels with STRIPE-seq. *Genome Res.*, **30**, 910–923.
14. Lawrence,M., Huber,W., Pagès,H., Aboyoun,P., Carlson,M., Gentleman,R., Morgan,M.T. and Carey,V.J. (2013) Software for computing and annotating genomic ranges. *PLoS Comput. Biol.*, **9**, e1003118.
15. Haberle,V., Forrest,A.R.R., 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.
16. Carninci,P., Sandelin,A., Lenhard,B., Katayama,S., Shimokawa,K., Ponjavic,J., Semple,C.A., Taylor,M.S., Engstrom,P.G. and Frith,M.C. (2006) Genome-wide analysis of mammalian promoter architecture and evolution. *Nat. Genet.*, **38**, 626–635.
17. Cumbie,J.S., Ivanchenko,M.G. and Megraw,M. (2015) NanoCAGE-XL and CapFilter: an approach to genome wide identification of high confidence transcription start sites. *BMC Genomics*, **16**, 597.
18. Wulf,M.G., Maguire,S., Humbert,P., Dai,N., Bei,Y., Nichols,N.M., Corrêa,I.R. 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.
19. Cvetesic,N., Leitch,H.G., 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.
20. Lu,Z. and Lin,Z. (2019) Pervasive and dynamic transcription initiation in *Saccharomyces cerevisiae*. *Genome Res.*, **29**, 1198–1210.
21. McMillan,J., Lu,Z., Rodriguez,J.S., Ahn,T.-H. and Lin,Z. (2019) YeasTSS: an integrative web database of yeast transcription start sites. *Database*, **2019**, baz048.
22. BSgenome.Scerevisiae.UCSC.sacCer3. R package version 1.4.0. http://bioconductor.org/packages/BSgenome.Scerevisiae.UCSC.sacCer3/.
23. TxDb.Scerevisiae.UCSC.sacCer3.sgdGene. R package version 3.2.2. http://bioconductor.org/packages/TxDb.Scerevisiae.UCSC.sacCer3.sgdGene/.
24. Nepal,C., Hadzhiev,Y., Previti,C., Haberle,V., Li,N., Takahashi,H., Suzuki,A.M.M., Sheng,Y., Abdelhamid,R.F., Anand,S. *et al.* (2013) Dynamic regulation of the transcription initiation landscape at single nucleotide resolution during vertebrate embryogenesis. *Genome Res.*, **23**, 1938–1950.
25. Yu,G., Wang,L.-G. and He,Q.-Y. (2015) ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinformatics*, **31**, 2382–2383.
26. Wickham,H. (2009) *ggplot2: Elegant Graphics for Data Analysis*. Springer Publishing Company, Incorporated.
27. Robinson,M.D. and Oshlack,A. (2010) A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.*, **11**, R25.
28. Robinson,M.D., McCarthy,D.J. and Smyth,G.K. (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.
29. Love,M.I., Huber,W. and Anders,S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, **15**, 550.
30. Blighe,K. and Lun,A. (2020) *PCAtools: Everything Principal Components Analysis*. R package version 2.2.0, https://github.com/kevinblighe/PCAtools.
31. Gu,Z., Eils,R. and Schlesner,M. (2016) Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics*, **32**, 2847–2849.
32. Hahne,F. and Ivanek,R. (2016) Visualizing Genomic Data Using Gviz and Bioconductor. In: Mathé,E. and Davis,S. (eds). *Statistical Genomics: Methods and Protocols*. Springer, NY, pp. 335–351.
33. Hoskins,R.A., Landolin,J.M., Brown,J.B., Sandler,J.E., Takahashi,H., Lassmann,T., Yu,C., Booth,B.W., Zhang,D., Wan,K.H. *et al.* (2011) Genome-wide analysis of promoter architecture in Drosophila melanogaster. *Genome Res.*, **21**, 182–192.
34. Lenhard,B., Sandelin,A. and Carninci,P. (2012) Metazoan promoters: emerging characteristics and insights into transcriptional regulation. *Nat. Rev. Genet.*, **13**, 233–245.
35. Raborn,R.T., Sridharan,K. and Brendel,V.P. (2017) TSRchitect: promoter identification from large-scale TSS profiling data. https://doi.org/doi:10.18129/B9.bioc.TSRchitect.
36. Wagih,O. (2017) ggseqlogo: a versatile R package for drawing sequence logos. *Bioinformatics*, **33**, 3645–3647.
37. Zhang,Z. and Dietrich,F.S. (2005) Mapping of transcription start sites in Saccharomyces cerevisiae using 5′ SAGE. *Nucleic Acids Res.*, **33**, 2838–2851.
38. Qiu,C., Jin,H., Vvedenskaya,I., Llenas,J.A., Zhao,T., Malik,I., Visbisky,A.M., Schwartz,S.L., Cui,P., Čabart,P. *et al.* (2020) Universal promoter scanning by Pol II during transcription initiation in Saccharomyces cerevisiae. *Genome Biol.*, **21**, 132.
39. Yu,G., Wang,L.-G. and He,Q.-Y. (2012) clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS J. Integr. Biol.*, **16**, 284–287.
40. Gasch,A.P., Spellman,P.T., Kao,C.M., Carmel-Harel,O., Eisen,M.B., Storz,G., Botstein,D. and Brown,P.O. (2000) Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell*, **11**, 4241–4257.
41. Ushijima,T., Hanada,K., Gotoh,E., Yamori,W., Kodama,Y., Tanaka,H., Kusano,M., Fukushima,A., Tokizawa,M., Yamamoto,Y.Y. *et al.* (2017) Light controls protein localization through phytochrome-mediated alternative promoter selection. *Cell*, **171**, 1316–1325.
42. Rubner,Y., Tomasi,C. and Guibas,L.J. (1998) A metric for distributions with applications to image databases. 59–66.
43. 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.
44. Klein-Brill,A., Joseph-Strauss,D., Appleboim,A. and Friedman,N. (2019) Dynamics of chromatin and transcription during transient depletion of the RSC chromatin remodeling complex. *Cell Rep.*, **26**, 279–292.
45. Kubik,S., Bruzzone,M.J., Challal,D., Dreos,R., Mattarocci,S., Bucher,P., Libri,D. and Shore,D. (2019) Opposing chromatin remodelers control transcription initiation frequency and start site selection. *Nat. Struct. Mol. Biol.*, **26**, 744–754.
46. Chia,M., Li,C., Marques,S., Pelechano,V., Luscombe,N.M. and van Werven,F.J. (2021) High-resolution analysis of cell-state transitions in yeast suggests widespread transcriptional tuning by alternative starts. *Genome Biol.*, **22**, 34.
47. Lu,Z. and Lin,Z. (2021) The origin and evolution of a distinct mechanism of transcription initiation in yeasts. *Genome Res.*, **31**, 51–63.
48. Cvetesic,N., Borkowska,M., Hatanaka,Y., Yu,C., Vincent,S.D., Müller,F., Tora,L., Leitch,H.G., Hajkova,P. and Lenhard,B. (2020) Global regulatory transitions at core promoters demarcate the mammalian germline cycle. bioRxiv doi: https://doi.org/10.1101/2020.10.30.361865, 30 October 2020, preprint: not peer reviewed.
49. Yu,C., Cvetesic,N., Hisler,V., Gupta,K., Ye,T., Gazdag,E., Negroni,L., Hajkova,P., Berger,I., Lenhard,B. *et al.* (2020)

TBPL2/TFIIA complex establishes the maternal transcriptome by an oocyte-specific promoter usage. bioRxiv doi: https://doi.org/10.1101/2020.06.08.118984, 09 June 2020, preprint: not peer reviewed.

50. Bhardwaj,V., Semplicio,G., Erdogdu,N.U., Manke,T. and Akhtar,A. (2019) MAPCap allows high-resolution detection and differential expression analysis of transcription start sites. *Nat. Commun.*, **10**, 3219.