Directionality of eRNA Transcription Burst: Sequence Features and Transcriptional Context

Yifan Dai & Albin Sandelin

Abstract

On bulk level, active enhancers feature with eRNA transcription from both boundaries, but single-cell CAGE sequencing data showed that most eRNAs are unidirectionally transcribed in single-cell level. Utilizing the world's first published single-cell 5' sequencing data, we investigate how is single-cell eRNA directionality determined and constrained. We show that unidirectionality is well-supported by high TPM and multiple TSSs. Single-cell direction preference is prone to keep through time and supported by bias of Pol II preferred sequence features (e.g. PyPu motif, CG-rich region, etc.). Surprisingly, we observe on single-cell level the accordance between eRNA transcription direction and high downstream TSSs number, linking eRNA transcription direction with chromatin accessibility. On the other hand, bidirectional transcription is rare and biased to enhancers with high burst frequency, suggesting superimposing of unidirectional events.

Introduction

Previously defined as intronic or intergenic regulatory elements that boost transcription of target genes, enhancers have been found of ability to initiate transcription by themselves from both boundaries [37]. Like mRNAs, the transcription of enhancer RNAs (eRNAs) is drived by RNA polymerase II (Pol II), based on recruitment of transcription factors (TFs) and other coactivators [48]. Unlike mRNAs, eRNAs are typically short(0.5-2kb), unspliced, exosome-sensitive (7.5 min half life) [31,40], and non-polyadenylated [30]. eRNAs were first discovered by total RNA-seq[40]. Nowadays, Cap Analysis Gene Expression (CAGE)[7], a powerful 5' RNA sequencing technology, is extensively used in permissive RNA (including eRNA) quantification and transcription initiation sites (TSSs) identification[15,30]. Using CAGE, FANTOM consortium accomplished a human eRNA atlas across tissues and cell types[30].

Transcription of eRNA is a nearly universal feature of active enhancers[30]. eRNA abundance is strongly associated with enhancer activity[30]. It is reported as the earliest transcriptional change under environmental stimuli[47]. Their bidirectional TSSs on bulk level are distinct from directional TSSs of mRNAs. For those reasons, balanced, divergent transcription of eRNAs has been used as identification markers for enhancers, replacing previous marker (i.e. H3K4me1-to-H3K4me3 signal ratio [39]) and becoming the state-of-art identification criterion[30,34].

Despite of its strong link with enhancer activity, the biological function of eRNA transcription remains elusive. Many suggest that eRNAs may be transcription noise[44], for its low half-life and low conservation. On the other hand, studies based on eRNA knockdown, ChIP, and FISH has established multiple models for eRNA function mechanisms. Some argues that eRNAs molecules can help initiate or stabilize enhancer-promoter interaction[44,45]. Some propose that eRNAs facilitate recruitment of Pol II by phase separation[46]. Others suggest that enhancer transcription can also function independent of the transcripts, by remodeling histone modification and altering chromatin accessibility [48]. These models are not necessarily mutual exclusive, but they remain to be tested for generalizability. By far, it is hard to use the mentioned wet lab experiments cannot study eRNA on genome-wide level.

A recent break-through in the field is the development of single-cell CAGE data and the discovery it drove. Based on single-cell CAGE data, Tsukasa Kouno et al. reveal that in single cells, most enhancers are transcribed in a exclusive unidirectional burst manner from either side [1], contradictory to the balanced bidirectional transcription model defined on bulk data. In this work, they performed single-cell CAGE analysis on 151 A549 lung cancer cells from three different time course of TGF-beta treatment. The data captures actively transcribed enhancers on single-cell level, thus providing us a new means to study eRNA transcription genome-widely with single-cell resolution.

Tsukasa Kouno et al. focused on performance benchmark of C1 CAGE technology and did not perform a detail analysis on eRNA unidirectional burst. But this data is an important clue that may provide information about cell-specific or chromatin-status-dependent direction decision of eRNA transcription. It is still unknown (i) why bidirectional transcription is under-represented; (ii) whether transcription direction of eRNA is randomly determined, or influenced by other factors, (e.g. sequence features and neighbor transcription). In order to further understand the property of eRNA transcription and mechanism of eRNA function, We utilize their data to conquer these questions.

Material and Methods

Our scripts, research notes, and supplementary data are available at https://github.com/evansdai/eRNA.direction.

1. Raw data

We base our study on the published CAGE data in Tsukasa Kouno et al.'s article about C1 CAGE technology[1]. In this study, A549 lung cancer cells were treated with tumor suppressor Transforming Growth Factor Beta (TGF-β) in 3 time course (0h, 6h, 24h) with 3 technical replicates each. Total 151 cells (40 cells from 0h, 40 cells from 6h, 71 cells from 24h) were loaded to integrated fluidic circuits (IFCs) of C1 Single-Cell Auto Prep System (Fluidigm). Single-cell CAGE data was then collected and processed following C1 CAGE procedure[1]. Bulk CAGE data for each sample was generated following nAnT-iCAGE method[2] using the remaining cells. Linker oligonucleotides (TATAGGG) on 5' end were removed and paired-end reads were mapped to human genome assembly hg19[1]. G addition error[15] was handled by pairedBamToBed12 program with corresponding option. PCR duplicates were collapsed based on unique 5' and 3' ends[1]. OSC files recording CAGE tag counts for each cell/sample were uploaded onto the ZENBU genome browser[3] (https://fantom.gsc.riken.jp/zenbu/gLyphs/#config=NMT9yTLnH59gIVssI9WRfD).

We download the quality-checked data from ZENBU. Bulk data is downloaded from track "[Bulk CAGE] TGF-β Timecourse bulk libraries, TPM-normalized, grouped by timepoint" with options "source type = tagcount" and "Stream processing script = none". Single-cell data is downloaded from track "[C1 CAGE] TGF-β Timecourse libraries normalized, QC filtered and grouped by timepoint" with the same settings. We convert OSC files to strand specific bedgraph formats with a homebrew python script. Then, bedgraph files are converted into BigWig format using UCSC binary utility bedGraphToBigWig[11] for better processing performance.

2. TPM Normalization of CAGE tags abundance

We load the raw counts of CAGE defined TSSs (CTSSs)[7] into R using quantifyCTSSs() function in CAGEfightR[4]. In single-cell data, the raw counts of eRNA TSSs are normalized for each cell by tags per million (TPM)[10]. Similarly, bulk TSSs are normalized in each sample. Function calcTPM() from CAGEfightR[4] is used to apply the normlization.

3. Bidirectional enhancer identification and validation

Instead of using annotated enhancers[1], we identify enhancers de novo, to ensure tissue and time specificity. Bidirectional enhancers in each time course is identified based on bulk CAGE data by CAGEfightR[4]. Firstly, bidirectional CTSSs clusters are identified using quickEnhancer [4]. Then, enhancers are filtered by four criteria: (1) with a balance score Bhattacharyya coefficient (BC) larger than 0.95; (2) with expression in all of 3 replicates (by option minSamples=2); (3) located in intron or intergenic regions (R package TxDb.Hsapiens.UCSC.hg19.knownGene[5] is used for genome region annotation); (4) with single-cell counts in at least on cell.

Identified enhancers are then validated via comparison with FANTOM5[6] enhancer database. One enhancer is validated if its range overlaps with annotated entries in hg19_enhancers_fantom enhancers in R package annotatr version 1.12.1 [8]. Validation rate is used for selecting enhancer identification strategy. We do not filter out enhancers that are not validated.

4. Estimation of burst size and burst frequency for enhancers on single-cell level

We view the expression of eRNA as transcriptional burst events[9]. For each enhancer in each cell, the burst size is defined by the total TPM of CTSSs. For each enhancer, the burst frequency is defined by the fraction of cells that express it.

5. Directionality on bulk and single cell level

Direction of CTSSs is defined based on Waston orientation of chromosome[12]. Plus(+) strand TSSs have 5' end on the side of the chromosome short arm.

On bulk level, direction preference for each enhancer i is calculated as follows:

$$Direction_{P}reference_{i} = log_{2}(Plus_{c}ounts_{i} - Minus_{c}ounts_{i})$$

In the formula, Plus_counts_i means the total plus strand counts of enhancer i of 3 technical replicates, and vice versa for Minus_counts_i. For each enhancer, the strand with larger bulk score is defined as the major strand, and the other strand as the minor strand.

On single-cell level, we calculate direction preference for each enhancer i in each active cell j:

$$Direction_{P}reference_{i}j = log_{2}(Plus_{c}ounts_{i}j - Minus_{c}ounts_{i}j)$$

Based on voting of active cells, enhancers are classified into 4 groups (i.e. Minus, Plus, Bidirectional, Mixed) . One enhancer is classified as Plus if all of its active cells transcribe eRNAs in plus strand, and vice versa for the Minus group. If both strands of eRNA are simultaneously found in (at least) one cell, the enhancer is classified as Bidirectional. At last, if some cells transcribe eRNAs in plus strand and others transcribe them in minus strand, the enhancer is classified as Mixed.

To investigate whether single-cell direction preference is randomly decided or not. We analyze direction preference change of enhancers across time courses. The overlap of enhancer groups across time is visualized by Venn plot (R package eulerr[20]). Significance is tested by Fisher's exact test. To investigate the relation of single-cell eRNA direction and bulk direction preference, we analyze the distribution of bulk preference for each single-cell enhancer classification.

6. Sequence analysis of directional preference

To investigate enrichment of sequence features related with enhancer directional preference, we analyze the CG content and motif enrichment in main strands compared to minor strands. For each enhancer, two divergent TSSs positions, each with the highest bulk score on its strand, are selected. Genomic sequences around are extracted using Biostrings::getSeq() [13] based on reference genome BSgenome.Hsapiens.UCSC.hg19[14].

To test whether direction preference is biased to side with high GC content, we calculate CG content around TSSs (-100 bp ~ 100 bp) for main and minor strand for each enhancer. A paired t test is conducted to assess the significance. Null hypothesis is that main and minor strand TSSs have identical average CG content in -100 ~ 100 bp regions.

In PyPu enrichment analysis, we test if Pyrimidine-Purine dinucleotide (T/C-A/G) are enriched at main strand TSSs[16,23]. We count occurence of PyPu motif at -1,+1 positions separately for main strand TSSs and minor strand TSSs. Odds ratio is calculated and Fisher's exact test is performed.

Motif enrichment analysis is performed with the CentriMo software [17] within MEME suite 5.1.1 [18]. From two motif databases, JASPAR2018 Pol II collection[21] and JASPAR2018_CORE_vertebrates_non-redundant collection[22], we search Pol II binding features and TF binding features around eRNA TSSs (-100 bp ~ 100 bp). Motifs locally enriched are scored by E-value. A secondary criteria, Fisher E-value, is used in those cases of comparing query sequences and background sequences. To improve precision of enrichment, repeat elements in the sequences are masked using RepeatMasder software [19].

7. Single-cell Transcription Context Analysis

Transcription Context is defined for each enhancer for each cell. In each cell, we extract unidirectional CTSSs clusters in the neighbouring genomic region (i.e. -1 Mb to +1 Mb) around an enhancer as its transcriptional context. CTSSs clusters are calculated using quickTSSs function in CAGEfightR[4] with default settings.

To investigate the relation between activity of enhancer and transcription strength of local TSSs, we analyze (1) strand specific cumulative TPM; (2) strand specific cumulative number of CTSS clusters around enhancers for several ranges from -1 kb ~ 1kb to -1Mb ~ 1Mb. Cumulative scores for each enhancer groups (i.e. transcribed/ not transcribed enhancers) are calculated using sum of normalized single-cell scores so that each enhancer-cell pair is equally weighted.

To further test how is single-cell eRNA direction related with its context, we build a regression model to regress single-cell direction score M against context distribution bias scores.

For each enhancer i in cell j:

$$M_{ij}\ Active_{ij} + txType_i + eid_i + A_{ij} + ContextSumBiasDowstream_{ij} + ContextSumBiasPlus_{ij} + ContextNumBiasDownstream_{ij} + ContextNumBiasPlus_{ij}$$

where

$$\begin{gathered}
M_{ij} = log(PlusCounts_{ij} + 1) - log(MinusCounts_{ij} + 1) \\
Active_{ij} = whether\ enhancer\ i\ is\ active\ (TPM > 0)\ in\ cell\ j \\
TxType_i = whether\ enhancer\ i\ is\ located\ at\ intron\ or\ intergenic\ region \\
eid_i = enhancer\ id \\
A_{ij} = log(PlusCounts_{ij} + 1) + log(MinusCounts_{ij} + 1) \\
ContextSumBiasDowstream_{ij} = \frac{TPM.Downstream_{ij}}{TPM.Total_{ij}} \\
ContextSumBiasPlus_{ij} = \frac{TPM.Plus_{ij}}{TPM.Total_{ij}} \\
ContextNumBiasDownstream_{ij} = \frac{NumClusters.Downstream_{ij}}{NumClusters.Total_{ij}} \\
ContextNumBiasPlus_{ij} = \frac{NumClusters.Plus_{ij}}{NumClusters.Total_{ij}}
\end{gathered}$$

We utilize t values and p values of the last four terms to access the relatedness of transcription direction with context. Each term represent a model as presented in Fig.10. Coefficients other than transcription context are regressed out to generalize the model. (e.g. directional preference of enhancer is regressed out in eid_i)

Results

1. Identification of time course specific enhancers

Time	Replicates	Cells	Bulk TCs	Single-cell TCs	Bulk Enhancers	Bulk Enhancers with SC	FANTOM Validation
0h	3	40	473,927	262,227	1330	356	156 (43.8%)
6h	3	41	557,406	277,832	1740	409	170 (41.6%)
24h	3	70	506,550	371,785	1615	596	264 (44.3%)

Table.1 Summary of Enhancer Identification Enhancers are identified on bulk level and filtered for each time course. Time: processing time of TGF-beta. Replicates: number of biological replicates used for bulk data. Cells: number of quality checked cells that were randomly sampled [1]. Bulk/Single-cell TCs: number of tag clusters identified in bulk/single-cell data using CAGEfightR. Bulk enhancers: bidirectional CTSSs clusters on intron or intergenic regions supported by all 3 replicates; Bulk enhancers with SC: Number of enhancers identified on bulk data and that pass our criteria (See Methods). FANTOM Validation: Number of Bulk_Enhancers_with_SC that overlap with FANTOM5 [6] enhancers.

We identify 356, 409, 596 well-supported transcribed enhancers respectively in t0, t6, t24. We select the time-specific strategy to avoid potential artifacts caused by pooling CTSSs under different conditions. Bidirectional tag clusters supported by 3 replicates are identified from bulk data. Only clusters located on intron and intergenic regions are selected as enhancers. They are then filtered by overlapping with single-cell tag clusters. On average 43.2% enhancers are validated by FANTOM5 database[6]. We use enhancers with single cell signals to perform the following analysis. They tend to be more actively transcribed than the other enhancers without single-cell signals, supported by a higher mean bulk TPM (4.38 > 2.34, wilcox test p < 1E-15).

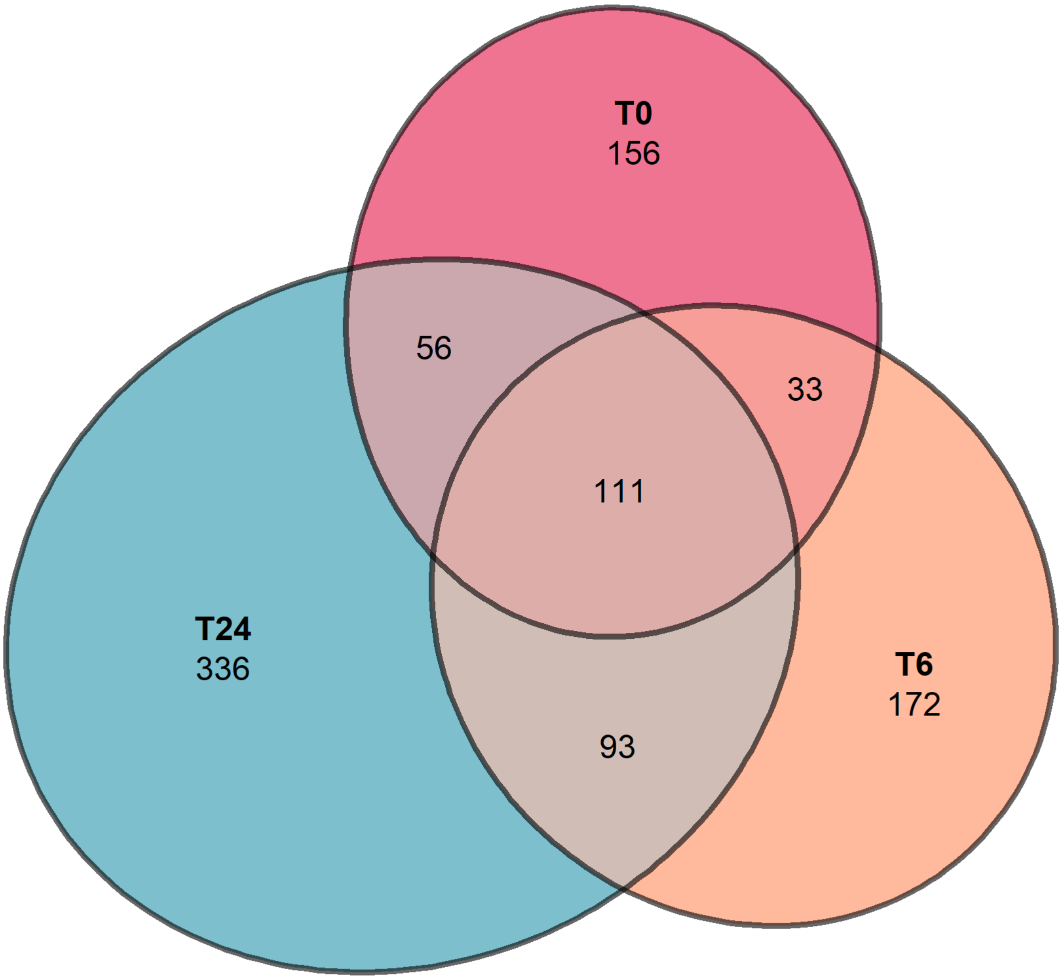


Fig.1 Overlap of transcribed enhancers from 3 time courses Overlap of enhancers from t0, t6, t24 are visualized using venn plot.

Venn diagram (Fig.1) shows that only 111 (11.6%) enhancers are active in all 3 time points, while a large proportion (664/957, 69.4%) of transcribed enhancers are time-course specific (Fig.1), which may reflect the change of cellular regulation under TGF-beta treatment.

2. Transcriptional burst: expression pattern of eRNA on single-cell level.

Transcription initiation from enhancers was believed to be either weak or rare, suggested by its low bulk TPM compared to from promoters. However, Tsukasa Kouno et al. [1] shows that eRNA transcription is rare but not necessarily weak. The distribution of max TPM of eRNA CAGE tags is similar with promoters[1] on single-cell level, supporting that enhancers are transcribed in a transcriptional bursting manner similar as promoters[9].

Utilizing the C1 CAGE data, we confirm that most enhancers rarely start transcription. Less than one-third of the bulk level enhancers have single-cell signals. For each of those enhancer, we calculate the number of cells where eRNAs are detected (Ncell). The mean Ncell is estimated 2.7 in t0, 2.7 in t6 and 3.3 in t24, indicating that eRNA is transcribed discretely and rarely. On the other hand, if transcribed, single-cell eRNAs signals are commonly not weak: the median of max TPM of transcribed enhancers is 34.3. To summarize, our data support that eRNAs are expressed in a transcriptional burst manner. The low bulk TPM of enhancers can be partially resulted from a low frequency of transcription initiation.

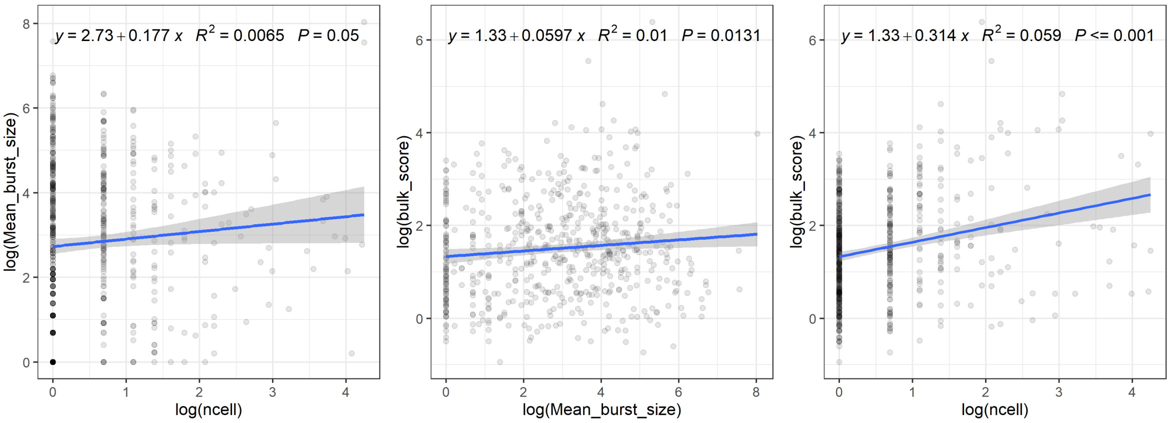


Fig.2 Regression model to explain enhancer bulk score on single-cell level in t24 The regression model log(bulk_score) ~ log(burst_frequency) + log(mean_burst_size) is visualzied using scatter plots. Data of timepoint 24h is used. Burst_frequency is estimated by number of active cells; mean burst size is the mean of total TPM across active cells; bulk score is the total TPM Burst frequency and mean burst size is not significantly correlated. They

To illustrate how single-cell burst frequency and burst size are related with bulk score of enhancers, we utilize linear regression. Fig.2 shows that mean burst size of transcribed enhancers is not significantly correlated with their burst frequency. Comapred to mean burst size, burst frequency can explain more variance of bulk score. The tendency is similar to observations in promoters.[9]

3. Most enhancers burst unidirectionally on single cell level.

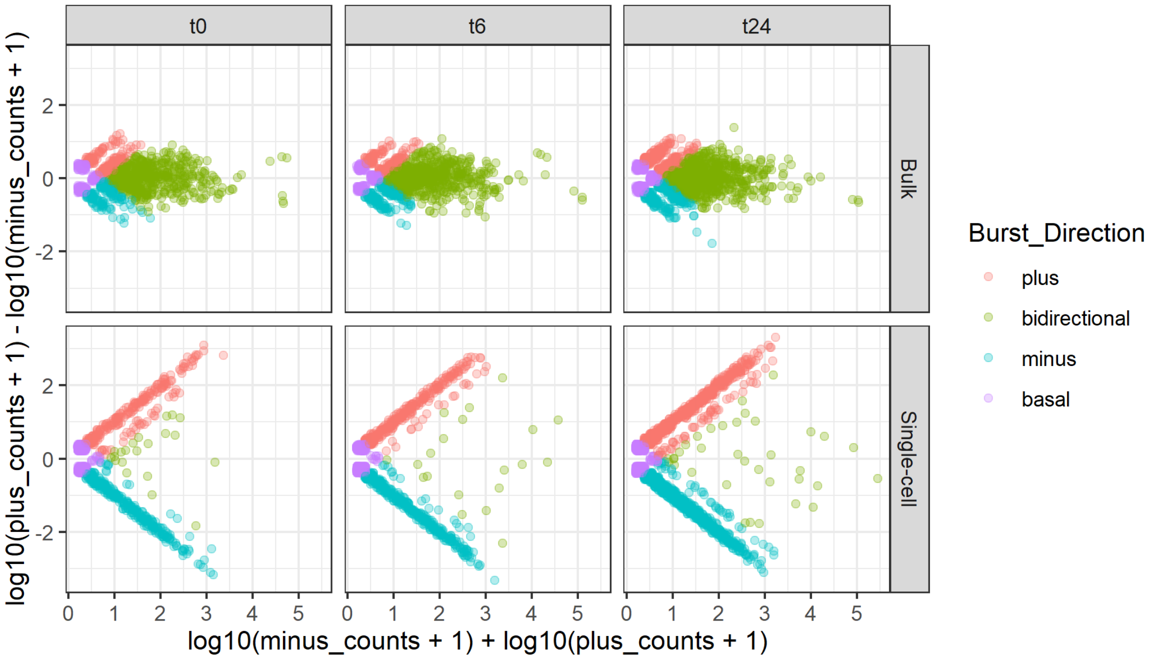


Fig.3 Distribution of plus/minus counts of transcription events on bulk/single-cell level The log counts ratio (y) is plotted against the summed log counts (x) for transcription events of enhancers. On bulk level (the upper row), each point represents an transcribed enhancer in one biological replicates; while on the single-cell level (the lower row), each point stands for an transcribed enhancer in one cell. There are 971 trascribed enhancer-cell pairs in t0, 1091 in t6, 1941 in t24. Points are jittered to enable better visualization. Points are colored based on strand expression values. Events with both plus and minus counts>1 are labeled as "bidirectional"; events with counts>1 on only one strand are labeled respectively "plus" or "minus"; events with counts <= 1 on boths strands are labeled as "basal". Transcription events from two hyper-active outlier enhancers (i.e. chr17:79478552-79478972 and chr21:9827341-9827855) are removed to show the main trend of enhancers.

For each transcribed enhancer in each cell, we calculate its total plus counts and total minus counts. MAplot visualization (Fig.2) shows that if pooled (bulk level), most eRNA transcription events are bidirectional (50.3% 2054/4083), showing that enhancers are capable to start transcription on either strand. However, on single-cell level, most single-cell eRNA transcription events are unidirectional (63.1% 2527/4003). At least 32.2% (1123/3486) of these unidirectional transcription events have multiple 5' TSSs, suggesting that the unidirectional pattern is not an artifact caused by PCR on single eRNA molecule. Instead, the evidence supports that the pattern is formed when multiple Pol II molecules go unidirectionally.

When associating the single-cell burst events with enhancers, we find that enhancers can be classified into 4 groups(Fig.4A) depending on their activity in single cells: some enhancers burst only in one strand, others burst in either one strand, the others are detected to have bidirectional transcription in at least one cell.

Utilizing this classification, we identify two over-transcribed enhancers (chr17:79478552-79478972,median plus TPM = 2744.3, median minus TPM = 67.9; and chr21:9827341-9827855, median plus TPM = 3093.0, median minus TPM = 1.02), whose eRNAs are highly expressed in all the 151 cells. Both enhancers are classified into bidirectional, but their expression are both biased to plus strand.

After removing those two enhancers, we find that the remaining bidirectional transcription events are quite evenly distributed in several enhancers. For instance, the remaining 35 bidirectional events at timepoint 24 are scattered in 22 enhancers.

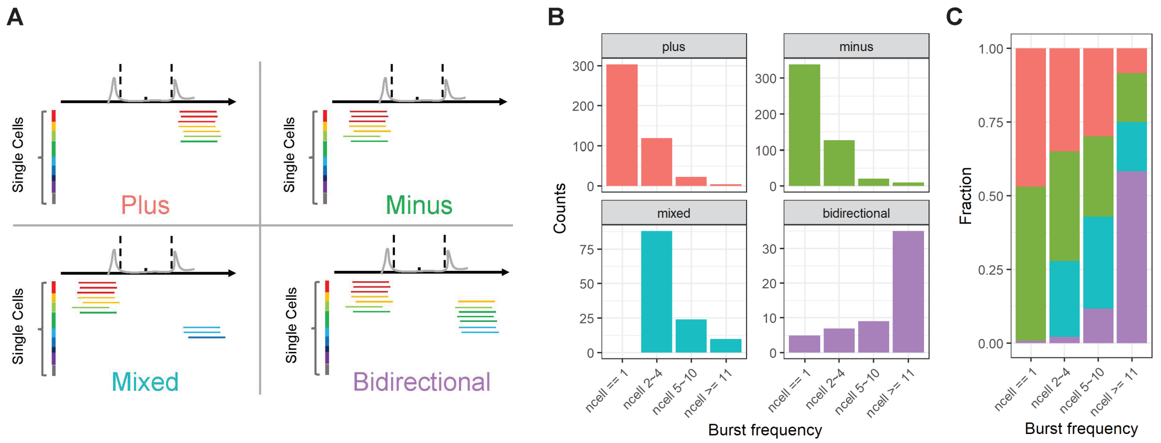


Fig.4 Bidirectional enhancers tend to have higher burst frequency A. Illustration of enhancer classification based on single-cell burst events (in Fig.3). Plus/Minus enhancers: purely transcribed in plus/minus strand on single-cell level; Mixed: transcribed in plus strand in some cells and in minus strand in the other cells; Bidirectional: with single-cell bidirectional transcription in at least one cell. B. Count of enhancers among different burst frequency bins, facet by classification. C. Fraction of enhancer groups among different burst frequency bins. In B and C, enhancers from different timepoints are pooled together.

Furthermore, we observe that the enhancers with bidirectional-events tend to be more frequently bursting (Fig.4B). Vice versa, enhancers with higher burst frequency are more likely to have at least one bidirectional event (Fig.4C). The relation of bidirectionality with burst frequency may infer that bidirectional bursts are related with frequent transcription initiation.

4. Direction preference of unidirectional burst events

Curious about how is unidirectional burst direction determined, we try to test whether direction of unidirectional burst is randomly determined. Firstly, we investigate whether the single-cell direction of enhancers change through time.

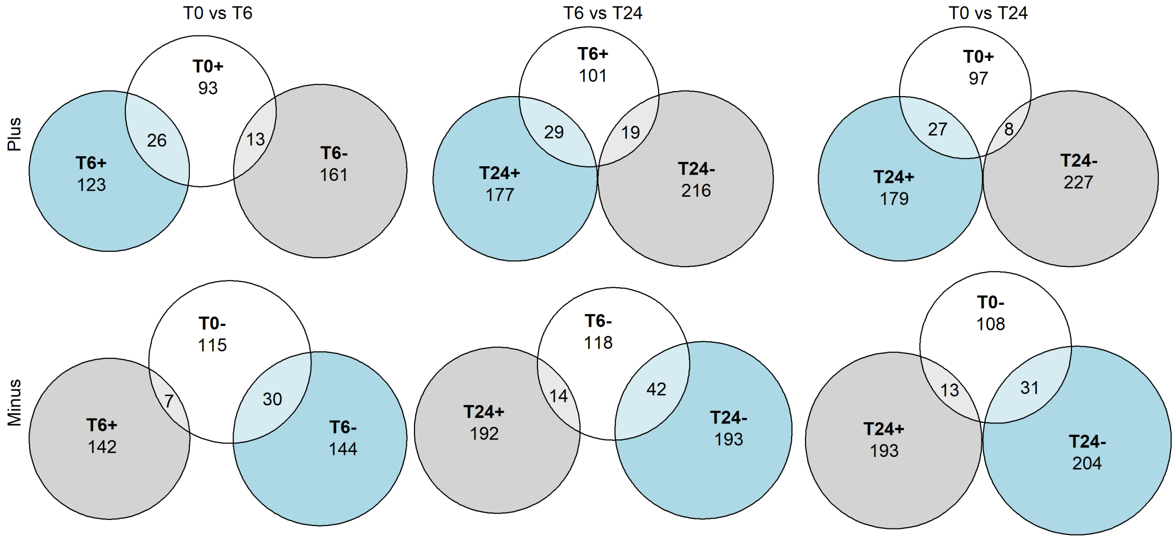


Fig.5 Single-cell unidirectional enhancers are prone to keep their direction through time Number of overlapping unidirectional enhancers between two time points are shown by 6 Venn plots. For example, the top left plot means that there 26 enhancers keep plus in both t0 and t6; while 13 enhancers were original plus, but change to minus in t6. Two-sided Fisher's exact test is done for each Venn plot. (e.g. (26,123),(13,161) is tested in the top left plot) The p-values are (left to right, top to bottom): 0.009472, 0.04738, 0.0001677, 0.0003742, 0.0005108, 0.01711.

Fig.5 shows how many plus enhancer (the same difinition as in Fig.4) in one timepoint keeps to be plus in another timepoint. Although only a small fraction of enhancers are shared among different conditions (Fig.1), we show that plus enhancers are more likely to keep plus than changing to minus (Fig.5 the upper row), and vice versa (Fig.5 the lower row). Significance in Fisher's exact tests support that directional preference is not randomly shuffled after condition changing.

We hypothesize that some enhancers have preferred direction of bursting, so we plot the distribution of plus bulk tags grouped by different single-cell classification.

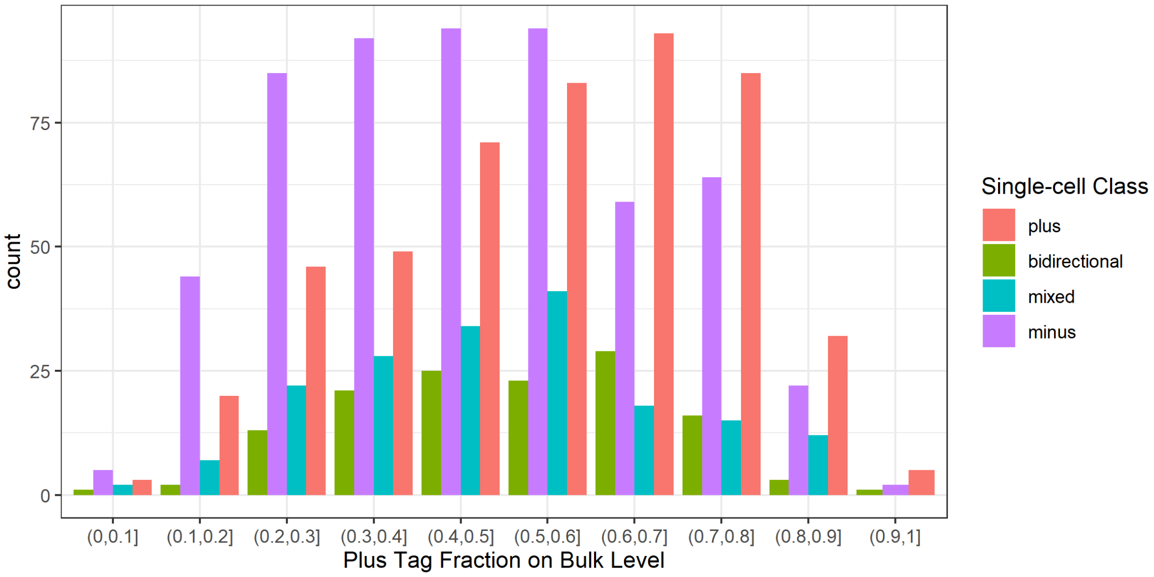
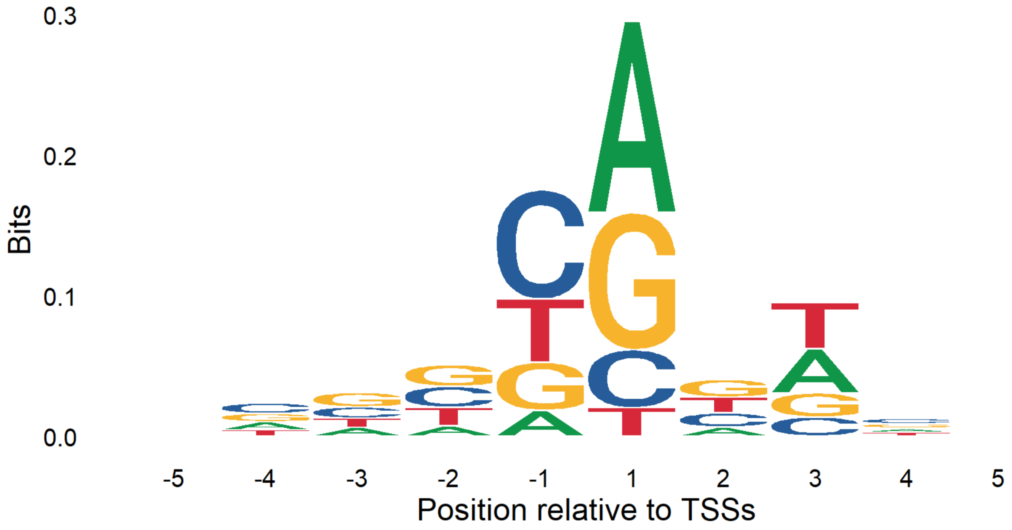


Fig.6 Distribution of bulk direction preference grouped by single-cell classification Bulk direction preference for each enhancer is defined as the fraction of plus TPM on bulk level. Single-cell plus enhancer and minus enhancer are well separated on bulk direction preference. Data from three timepoints is pooled when plotting the histogram.

Fig.5 shows that single-cell plus enhancers tend to have a higher fraction of plus tags, and minus enhancers tend to have lower. On the other hand, mixed and bidirectional enhancers are on average less biased to either side. The histogram indicates that enhancers may have preferred direction of transcription initiation, which can be at least partially described by direction preference on bulk level.

To understand how the directional preference is shaped, we conduct sequence analysis on two sides of enhancers. We define main and minor strand for each enhancer based on its bulk directional preference. Then, we test whether specific motif is enriched on the main strand.

Pyrimidine-purine dinucleotide (PyPu) at the -1,+1 is believed to be a Pol II preferred transcription starting site code for promoters [23]. Using LOGO analysis, we confirm that PyPu is also enrichment on -1,+1 position of enhancer TSSs. (Fig.7)



	T0		T6		T24		Total		
	PyPu	Others	PyPu	Others	PyPu	Others	PyPu	Others	Ratio
Main	207	128	216	159	349	212	772	499	0.607
Minor	193	142	208	167	320	241	721	550	0.567

Fig.7 PyPu dinucleotide enrichment on main and minor strand enhancer TSSs Pyrimidine-purine dinucleotide (C/T-A/G) are enriched on enhancer TSSs. Two-sided Fsisher exact test shows that it is significantly more likely to be TSSs of major strand. (p = 0.0439)

Moreover, we show that PyPu motif is significantly more enriched in main strand TSSs using two-sided Fisher's exact test (p = 0.0439).

High CG content is believed to be associated with pervasive expression of human housekeeping promoters.[24] CpG islands also show association with human enhancers with higher bulk directionality[25]. Here we test, using paired t test, if main strand TSSs tend to have higher CG content.

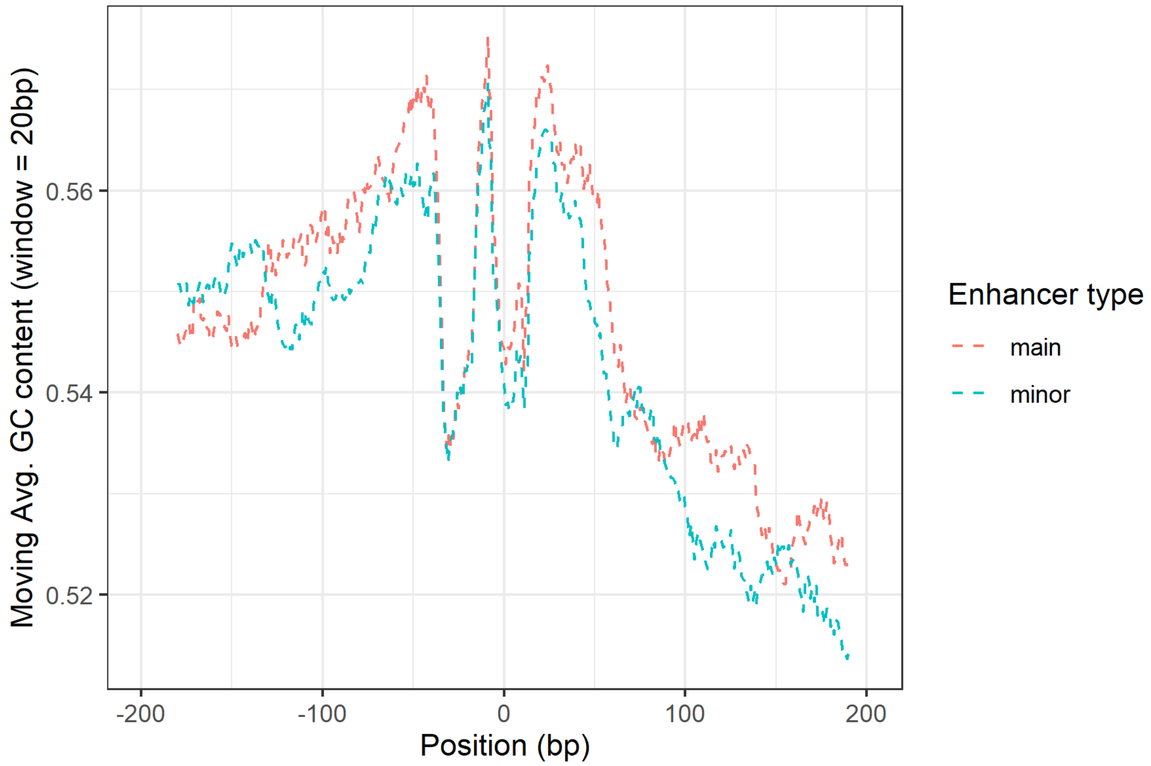
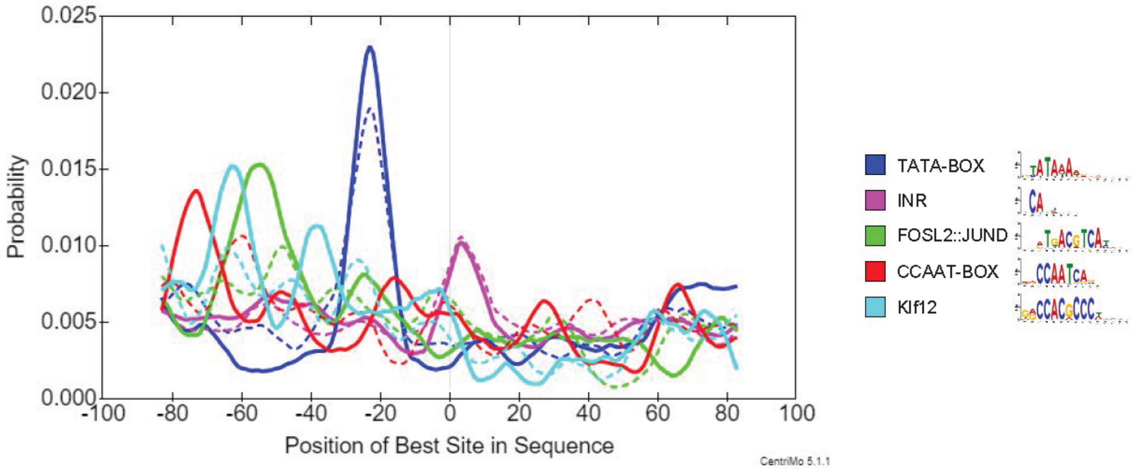


Fig.8 CG content around enhancer TSSs grouped by main and minor strand The -100 ~ +100 bp sequences around enhancer TSSs are extracted. The average CG content is calculated for each position, then plotted using

moving average (window = 20bp). Sequence around main TSSs and minor TSSs are labeled by different colors. Data from t0, t6, t24 are pooled to draw this figure.

Paired t-test proves that main strand TSSs are significantly more likely to have higher CG content around (-100 ~ +100 bp), compared to their corresponding minor strand TSSs. (t = 2.8703, p = 0.004169).

We are also curious about how are other Pol II associated motifs (e.g. TATA box) and TF binding site motifs are differently distributed upstream of main and minor strand TSSs. We conduct a motif enrichment analysis using centrimo software.



Motif	ID	E-value	Bin Position	Bin Width	Input Matches in Bin	Input Matches	Control Matches in Bin	Control Matches	Fisher Adjusted p
TATA-Box	POL012.1	2.70E-37	-23	6	75	298	60	322	1.00E+00
INR	POL002.1	3.50E-37	3	1	60	966	61	996	1.00E+00
FOSL2::JUND(var.2)	MA1145.1	2.50E-05	-56.5	13	42	182	13	164	4.60E-01
CCAAT-box	POL004.1	2.90E-01	-80	5	30	363	11	323	1.00E+00
Klf12	MA0742.1	3.10E-01	-52	34	49	127	34	127	1.00E+00

Fig.9 Motif probability curve around main and minor TSSs Top five non-redundant motifs enriched are shown in the figure. The -100 ~ +100 bp sequences around main TSSs are used as input, and plotted using solid lines. Minor sequences (-100 ~ +100 bp) sequences are used as control, and plotted using dahsed lines. The y-axis means the probability of the given motif at the given position estimated from sequence containing at least one match. Motif databases JASPAR2018 Pol II collection[21] and JASPAR2018_CORE_vertebrates_non-redundant collection[22] are used. Table displays relevant information including overall enrichment significance (E-value), enrichment center (Bin Position) and relative enrichment for main and minor strand. For more motifs, please see Supplementary data.

We identify 126 motifs significantly (E-value < 1) enriched around enhancer TSSs compared to random sequences (1-order Markov model). As expected, TATA-box, INR, CCAAT-box motifs are significantly enriched at -23, +3, and -80 bp, respectively.

We also obeserve motifs of highly-expressed TFs around enhancer TSSs. For instance, we observe a significant enrichment of consensus sequence TGANTCA, which is associated with FOS::JUN family TFs that are proved to be highly expressed in A549 cancer cells and function in proliferation/survival [26]. We also enrich CCACGCCC, a motif associated with KLF family TFs, which can be induced by TGF-beta and arrest cell cycle [27]. It is very likely that some proportion of eRNA transcription in our study is initiated by those TFs.

In the table, we observe a trend, although not significant, that those mentioned motifs are more enriched in main strand than minor strand.

Overall, sequence features, such as Pol II related motifs, TF binding motifs and CG content, biased to main strand may partially explain the directional preference of eRNA expression.

5. Transcriptional context related with single-cell eRNA direction

On single-cell level, eRNA transcription does not strictly follow bulk preference (Fig.6). To further understand direction eRNA transcription on single-cell level, we extract transcription context (i.e. unidirectional TSSs clusters within 1 million bases distance) for each transcribed enhancer in each cell. The 1 million bp boundary is decided by two facts: (i) there exists examples of long-range enhancer[35] interacting with its promoter about 1Mb away; and (ii) the common size TAD is about 0.2-1.0 Mb[36].

Using a regression model (detailly described in Methods Section), we test two hypotheses about relation between direction and context of single-cell eRNA (Fig.10). One hypothesis is that eRNA tends to be transcribed towards direction where there are more/less TSSs clusters. Another hypothesis is that eRNA tends to be transcribed in the similar direction with nearby context (Fig.10).

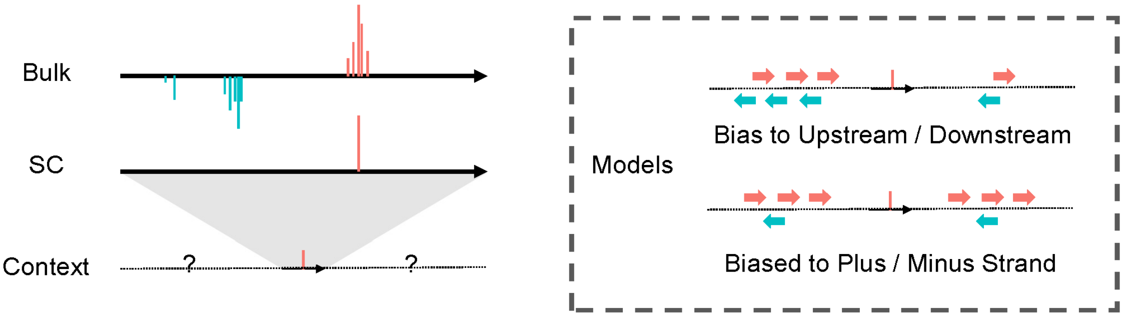


Fig.10 Models of relation between single-cell eRNA direction and transcription context On the left panel, genomic region of a typical bulk-level bidirectional enhancer is illustrated. If we look at one of the cell (SC) where the enhancer is transcribed in plus strand, what does its transcription context tend to be? On the right panel, our hypotheses are illustrated. Colored Arrows stand for unidirectional TSS clusters. We test whether direction of single-cell eRNAs is biased to where there are more/less TSSs clusters. At the same time, we test whether eRNA transcription is biased to the direction of nearby context.

We measure transcription context by two statistics, cumulative TPM and cumulative number of TSSs clusters. The former one measures the pooled transcription activity, while the latter one emphasize on chromosome accessibility by giving more weight on weakly transcribed TSSs. To discriminate their effects, we calculate two types of bias score using those two measurements. For example, ContextSumBiasDowstream (CSB_d) is computed as the fraction of downstream TPM in total TPM, while ContextNumBiasDowstream (CNB_d) is calculated by number of downstream TSS clusters devided by total number of context TSS clusters.

We build the regression model to predict eRNA direction by context and other factors. eRNA direction is calculated by log(Plus counts + 1) - log(Minus counts + 1) As described in Methods Section, context irrelevant factors (e.g. enhancer specific directional preference) are put in the model to make our results generalizable. The result is shown in Table.2.

	Range	t value				p value			
		CSB_d	CSB_p	CNB_d	CNB_p	CSB_d	CSB_p	CNB_d	CNB_p
t0	1.0E+03	0.075	0.748	2.071	-1.815	9.40E-01	4.55E-01	3.85E-02	6.97E-02
	1.0E+04	-2.056	-0.379	4.717	-0.754	3.99E-02	7.04E-01	2.46E-06	4.51E-01
	1.0E+05	-3.203	0.387	4.753	-1.869	1.36E-03	6.99E-01	2.03E-06	6.16E-02
	1.0E+06	-3.885	0.978	3.775	-1.938	1.03E-04	3.28E-01	1.61E-04	5.26E-02
t6	1.0E+03	1.510	-0.435	0.534	0.634	1.31E-01	6.63E-01	5.93E-01	5.26E-01
	1.0E+04	-1.334	-1.580	3.880	0.258	1.82E-01	1.14E-01	1.06E-04	7.96E-01
	1.0E+05	-1.630	-0.030	4.754	-1.628	1.03E-01	9.76E-01	2.02E-06	1.04E-01
	1.0E+06	-1.423	0.643	3.515	-2.662	1.55E-01	5.20E-01	4.41E-04	7.78E-03
t24	1.0E+03	-0.849	0.670	3.835	-0.297	3.96E-01	5.03E-01	1.28E-04	7.67E-01
	1.0E+04	-3.113	-2.822	7.121	2.032	1.85E-03	4.78E-03	1.14E-12	4.22E-02
	1.0E+05	-1.660	1.460	4.913	-1.483	9.69E-02	1.44E-01	9.03E-07	1.38E-01
	1.0E+06	-2.661	0.342	5.396	-0.663	7.80E-03	7.32E-01	6.85E-08	5.07E-01

Table.2 Regression results of eRNA direction prediction by context bias One regression result is shown on each row. Both t-values and p-values of context related terms are shown. T-values are colored base on positive and negative values. P-values smaller than 10E-5 are colored red. Time: the timepoint of enhancers used as input data. Range: the maximum distance of TSSs clusters in calculation. Abbreviation for context bias terms: "C" means "context"; "S"/"N" means "SumTPM"/"Number of TSSs clusters"; "B" means "bias"; "_d" means "downstream"; "_p" means "plus strand". For example, "CSB_d" stands for "ContextSumBiasDowstream". Please refer to the Methods Section for formulas. For the full results of regressions, please see Supplementary data.

Regression results are similar for each time point. The t-value for term ContextNumBiasDowstream is significantly positive, which means that the direction of eRNA tends to accord with the direction where there are more transcribed TSSs clusters (Fig.11A). The effect is stronger when range is set to 10,000 bp or 100,000 bp, suggesting the relation between eRNA transcription and the context bias is local. Regression results do not support other models suggested in Fig.10.

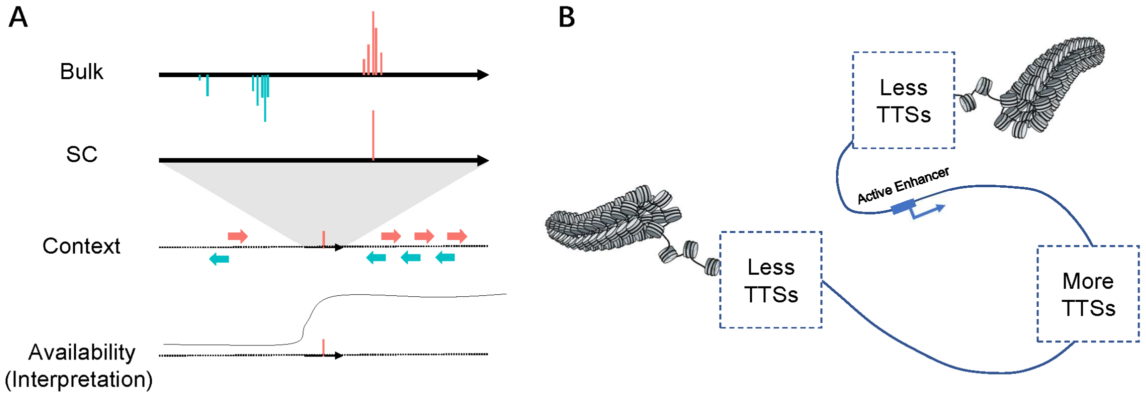


Fig.11 Interpretation of eRNA context bias A. Illustration of the model supported by regression results. There are more TSSs clusters on the downstream of eRNA than on upstream. It is likely that chromatin accessibility are higher in the downstream of eRNA transcription. B. Interpretation of the phenomenon on chromosome.

Our interpretation of the accordance is illustrated in Fig.11B. Our data indicates that the downstream of eRNA is more accessible by pre-initiation complex. One possible explanation is that enhancers are often transcribed when located on the boundary of open chromatin, where transcription towards the open chromatin is favored. Another explanation is that the transcription of eRNA helps open the downstream chromatin. It is hard to tell whether eRNA transcription is the cause, but we propose that single-cell eRNA transcription may relate with chromatin accessibility or interactivity.

Discussion

Sinle-cell CAGE data provides us high enough resolution to see transcription bursts, a discrete, thermodynamic process, while keeping a genome-wide view. It decouples burst size and burst frequency, while in bulk data we have only bulk score. It distinguishes unidirectional burst events, while in bulk data we can only see the mixture. It provides single-cell context around active and inactive enhancer loci, while in bulk we only measures the average. For those reasons, it is very useful in tackling questions about eRNA transcription direction.

In this study, we investigate the eRNA transcription pattern based on C1 CAGE data collected from A549 lung cancer cell line. Based on our observation that most eRNAs are transcribed with TPM high enough, low frequency and multiple TSSs, unidirectional transcription burst pattern is very likely to be realistic but not technical artifact caused by low sequencing size. The unit of eRNA transcription is likely to be unidirectional, where multiple Pol II molecules are recruited to bind chromatin on a single direction. On the other hand, bidirectional bursting events are rare, and prone to happen in enhancer of high burst frequency. The correlation suggests that they could be superimposed discrete burst events.

Direction of single-cell burst is not randomly determined, but are able to keep across time and condition. We show that enhancers are of direction preference based on bulk data. Single-cell unidirectional burst tend to follow the preferece. This is partially because sequence features on the main strand. Main strand of enhancers are significantly more CG-rich, PyPu enriched. We also show tendency that main strand enriches more TATA-box, and specific types of TF binding motifs (e.g. for AP-1 family, KLF family TFs).

Surprisingly, we confirm that single-cell burst direction is significantly correlated with its transcriptional context. eRNAs tend to be transcribed in the direction where there are more downstream TSSs, which suggests relation between eRNA transcription with chromatin accessibility or interactivity. One interpretation is that eRNA transcription increases the probability of downstream transcriptional initiation. Further investigation is needed to attribute causal effects.

Limitations originated from C1 CAGE technology may decrease its statistical power for certain questions. For instance, RNA abundance is globally under-estimated because duplicated alignment with the same 5' and 3' end were removed to alleviate PCR artifacts [1]. It may decrease the accuracy of our estimation of burst size, which may negatively influence our statistical test about relation between bulk score, burst size and burst frequency. Furthermore, the limited coverage (~50%)[1] of single-cell sequencing does not guarantee full sampling of RNAs. It may cause under-estimation of TSSs sites and also lead to neglect of low burst size and low burst frequency enhancers. On the other hand, the short sequencing length prevents C1 CAGE to detect allele-specific transcription, which may lead to over-estimation of bidirectional transcription events. Because of those problems, one should avoid over-interpretation from comparison between bulk and single-cell data. In our study, we bypass those problems by selecting proper control/background model when doing statistical test. For instance, comparing main and minor strand of each enhancer.

Another limitation comes from the short half-life of eRNA. The eRNAs are rapidly degraded by exosomes for lacking pA, so the burst size estimated in our study may not reflect only the transcription rate, but also the degradation rate. We base our burst size estimation on the assumption that eRNAs are equally degraded, but it may not be the reality. To solve this, future researches can be based on exosome depleted cell lines.

By far, there is still a huge knowledge gap between eRNA transcription and mechanism of enhancer function. Our findings, linking eRNA transcription with chromatin status, might be able to provide a novel sight to explain the mechanism of enhancer function. To further investigate the linkage, our future work will focus on integrating chromatin accessibility data (e.g. ATAC data) with eRNA single-cell data. By this means, we plan to test whether eRNA transcription direction preference is related with nearby chromatin accessibility. We are also curious about whether certain types of context TSSs are related with eRNA transcription direction, so we plan to subdivide the downstream TSSs by types for detailed analysis. If available, single-cell data collected from other cell lines, exosome knocked-down or not, shall be examined to test if context bias of eRNA transcription is a general phenomenon.

Reference

1. Kouno, T. et al. C1 CAGE detects transcription start sites and enhancer activity at single-cell resolution. Nat. Commun. 10, (2019).
2. Murata, M. et al. Detecting expressed genes using CAGE. Methods Mol. Biol. 1164,67–85 (2014)
3. Severin, J. et al. Interactive visualization and analysis of large-scale sequencing datasets using ZENBU. Nat. Biotechnol. 32, 217–219 (2014)
4. Thodberg, M., Thieffry, A., Vitting-Seerup, K., Andersson, R. & Sandelin, A. CAGEfightR: Analysis of 5′-end data using R/Bioconductor. BMC Bioinformatics 20, 1–13 (2019)
5. Carlson M, Maintainer BP. TxDb.Hsapiens.UCSC.hg19.knownGene: Annotation package for TxDb object(s). R package version 3.2.2. (2015)
6. Forrest, A. R. R. et al. A promoter-level mammalian expression atlas. Nature 507, 462–470 (2014)
7. Kodzius, R. et al. Cage: Cap analysis of gene expression. Nat. Methods 3, 211 (2006)
8. Cavalcante RG, Sartor MA. “annotatr: genomic regions in context.” Bioinformatics. R package version 1.12.1. (2017)

9. Larsson, A. J. M. et al. Genomic encoding of transcriptional burst kinetics. Nature 565, 251–254 (2019)
10. Carninci, P. et al. Genome-wide analysis of mammalian promoter architecture and evolution. Nat. Genet. 38, 626–635 (2006)
11. Kent, W. J., Zweig, A. S., Barber, G., Hinrichs, A. S. & Karolchik, D. BigWig and BigBed: Enabling browsing of large distributed datasets. Bioinformatics 26, 2204–2207 (2010)
12. Cartwright, R. A. & Graur, D. The multiple personalities of Watson and Crick strands. Biol. Direct 6, 1–5 (2011)
13. Pagès H, Aboyoun P, Gentleman R, DebRoy S . Biostrings: Efficient manipulation of biological strings. R package version 2.56.0. (2020)
14. Team TBD. BSgenome.Hsapiens.UCSC.hg19: Full genome sequences for Homo sapiens (UCSC version hg19, based on GRCh37.p13). R package version 1.4.3. (2020)
15. Carninci, P. et al. Cap-Analysis Gene Expression (CAGE): The Science of Decoding Genes Transcription. CRC Press. 93-100 (2019)
16. Smale, S. T. & Kadonaga, J. T. The RNA Polymerase II Core Promoter. Annu. Rev. Biochem. 72, 449–479 (2003).
17. Bailey, T. L. & MacHanick, P. Inferring direct DNA binding from ChIP-seq. Nucleic Acids Res. 40, 1–10 (2012).
18. Bailey, T. L. et al. MEME Suite: Tools for motif discovery and searching. Nucleic Acids Res. 37, 202–208 (2009).
19. Smit, AFA, Hubley, R & Green, P. RepeatMasker Open-4.0. 2013-2015 http://www.repeatmasker.org.
20. Larsson J. eulerr: Area-Proportional Euler and Venn Diagrams with Ellipses. R package version 6.1.0. (2020)
21. Bryne, J. C. et al. JASPAR, the open access database of transcription factor-binding profiles: New content and tools in the 2008 update. Nucleic Acids Res. 36, 102–106 (2008).
22. Khan, A. et al. JASPAR 2018: Update of the open-access database of transcription factor binding profiles and its web framework. Nucleic Acids Res. 46, D260–D266 (2018).
23. Carninci, P. et al. Genome-wide analysis of mammalian promoter architecture and evolution. Nat. Genet. 38, 626–635 (2006).
24. Landolin, J. M. et al. Sequence features that drive human promoter function and tissue specificity. Genome Res. 20, 890–898 (2010).
25. Steinhaus, R., Gonzalez, T., Seelow, D. & Robinson, P. N. Pervasive and CpG-dependent promoter-like characteristics of transcribed enhancers. Nucleic Acids Res. 48, 5306–5317 (2020).
26. Yadav, S., Kalra, N., Ganju, L. & Singh, M. Activator protein-1 (AP-1): a bridge between life and death in lung epithelial (A549) cells under hypoxia. Mol. Cell. Biochem. 436, 99–110 (2017).
27. Memon, A. & Lee, W. K. KLF10 as a tumor suppressor gene and its TGF-β signaling. Cancers (Basel). 10, (2018).
28. Sherwood, R. I. et al. Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape. Nat. Biotechnol. 32, 171–178 (2014).
29. Shiraki, T. et al. Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. Proc. Natl Acad. Sci. 100, 15776–15781 (2003)
30. Andersson, R. et al. An atlas of active enhancers across human cell types and tissues. Nature 507, 455–461 (2014).
31. De Santa, F. et al. A large fraction of extragenic RNA pol II transcription sites overlap enhancers. PLOS Biol. 8, e1000384 (2010)
32. Kim, T.-K. et al. Widespread transcription at neuronal activity- regulated enhancers. Nature 465, 182–187 (2010).
33. Lambert, S. A. et al. The human transcription factors. Cell 175, 598–599 (2018).
34. Andersson, R. & Sandelin, A. Determinants of enhancer and promoter activities of regulatory elements. Nat. Rev. Genet. (2019). doi:10.1038/s41576-019-0173-8
35. Lettice, L. A. et al. A long- range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. Hum. Mol. Genet. 12, 1725–1735 (2003).
36. Bulger, M. & Groudine, M. Enhancers: the abundance and function of regulatory sequences beyond promoters. Dev. Biol. 339, 250–257 (2010).
37. Andersson, R. et al. Nuclear stability and transcriptional directionality separate functionally distinct RNA species. Nat. Commun. 5, 5336 (2014)

38. Orom, U. A. & Shiekhattar, R. Long noncoding RNAs usher in a new era in the biology of enhancers. Cell 154, 1190 (2013).
39. Sharifi-Zarchi, A. et al. DNA methylation regulates discrimination of enhancers from promoters through a H3K4me1-H3K4me3 seesaw mechanism. BMC Genomics 18, 1–21 (2017).
40. Kim, T. K., Hemberg, M. & Gray, J. M. Enhancer RNAs: A class of long noncoding RNAs synthesized at enhancers. Cold Spring Harb. Perspect. Biol. 7, 3–5 (2015).
41. Andersson, R. et al. Nuclear stability and transcriptional directionality separate functionally distinct RNA species. Nat. Commun. 5, (2014).
42. Cajigas, I. et al. The Evf2 Ultraconserved Enhancer lncRNA Functionally and Spatially Organizes Megabase Distant Genes in the Developing Forebrain. Mol. Cell 71, 956-972.e9 (2018).
43. Tsai, P. F. et al. A Muscle-Specific Enhancer RNA Mediates Cohesin Recruitment and Regulates Transcription In trans. Mol. Cell 71, 129-141.e8 (2018).
44. Arnold, P. R., Wells, A. D. & Li, X. C. Diversity and Emerging Roles of Enhancer RNA in Regulation of Gene Expression and Cell Fate. Front. Cell Dev. Biol. 7, 1–14 (2020).
45. Hsieh, C. L. et al. Enhancer RNAs participate in androgen receptor-driven looping that selectively enhances gene activation. Proc. Natl. Acad. Sci. U. S. A. 111, 7319–7324 (2014).
46. Nair, S. J. et al. Phase [46] of ligand-activated enhancers licenses cooperative chromosomal enhancer assembly. Nat. Struct. Mol. Biol. 26, 193–203 (2019).
47. Weinhold, N. et al. Transcribed enhancers lead waves of coordinated transcription in transitioning mammalian cells. Science (80-.). 347, 1010–1015 (2015).
48. Kaikkonen, M. U. et al. Remodeling of the enhancer landscape during macrophage activation is coupled to enhancer transcription. Mol. Cell 51, 310–325 (2013).