Chapter 14

Genome-Wide TSS Identification in Maize

María Katherine Mejia-Guerra, Wei Li, Andrea I. Doseff, and Erich Grotewold

Abstract

Regulation of gene expression is a fundamental biological process that relies on transcription factors (TF) recognizing specific *cis* motifs in the regulatory regions of the genes that they control. In most eukaryotic organisms, *cis*-regulatory elements are significantly enriched around the transcription start site (TSS). However, different from other genic features, TSSs need to be experimentally determined, becoming then important components of genome annotations. One of the methods for experimentally determining TSSs at the genome-wide level is CAGE (cap analysis of gene expression). This chapter describes how to prepare a CAGE library for sequencing, starting with RNA extraction, library construction, and quality controls before proceed to sequencing in the Illumina platform. We then describe how to use a computational pipeline to determine, from the alignment of CAGE tags, the genome-wide location of TSSs, followed with statistical approaches required to cluster TSSs that operate as transcriptional units, and to determine core promoter properties such as shape. The analyses described here focus on maize, since its large and yet deficiently annotated genome creates some unique challenges, but with some modifications can be easily adopted for other organisms as well.

Key words Transcription start site, Transcription factor, Maize, Cap analysis of gene expression, CAGE, Promoter shape

1 Introduction

Transcription is a highly regulated process controlled in large part by transcription factors (TFs) that function in a combinatorial fashion to specify when and how eukaryotic genes are expressed [1]. This is accomplished by TFs interacting with *cis*-regulatory elements (CREs) in the control regions of genes. Regulatory DNA is often enriched in open chromatin regions, which are characterized by DNaseI hypersensitivity or accessibility to transposons. Techniques such as ATAC-Seq, FAIRE-Seq, and DNaseI-Seq permit identifying open chromatin regions and DNaseI hypersensitive sites (DHSs) with high precision [2–6]. Thus, DHSs and open chromatin regions provide a compendium of potentially important functional regulatory elements (*cistrome*). Placing them within the

Nobutoshi Yamaguchi (ed.), *Plant Transcription Factors: Methods and Protocols*, Methods in Molecular Biology, vol. 1830, https://doi.org/10.1007/978-1-4939-8657-6_14, © Springer Science+Business Media, LLC, part of Springer Nature 2018

context of transcriptional units requires identifying and annotating TSSs. Different from other genic features (e.g., open reading frames), TSSs are not characterized by any yet recognizable DNA sequence that permits predicting their position. Moreover, in eukaryotic genes, transcription initiation often occurs at multiple TSSs, resulting in what are commonly known as promoter clusters. In some instances, these clusters are distributed over several hundred or even thousands of base pairs, constituting what are known as broad or dispersed clusters. In other cases, they are concentrated over hundred base pairs or so, resulting in what are known as peaked or focused clusters [7, 8].

There are many different methods for the genome-wide determination of TSSs, including cap analysis of gene expression (CAGE) [9], RNA annotation and mapping of promoters for analysis of gene expression (RAMPAGE) [10], and paired-end analysis of transcription start sites (PEAT) [11]. Here, we describe the application of CAGE (Fig. 1) to the large genome of maize, for which the vast majority of the TSSs are unknown or continue to be inappropriately annotated [12].

2 Materials

All regents must be prepared by using RNA/DNA-free solutions and clean, dedicated equipment.

2.1 Linker Preparation

1. 5′ Bar-coded linkers (HPLC-grade, Invitrogen):

   5′N6-**NNN**:CCACCGACAGGTTCAGAGTTCTACAG**NNN**CAGCAGNNNNNN-P.

   5′GN5-**NNN**:CCACCGACAGGTTCAGAGTTCTACAG**NNN**CAGCAGGNNNNN-P.

   5′ lower-**NNN**: P-CTGCTG**N’N’N’**CTGTAGAACTCTGAACCTGTCGGTGG.

   (**NNN** could be AGA, CTT, GAT, ACA, ACT, ACG, ATC, ATG, AGC, AGT, TAG, TGG, GTA, GAC or GCC. **N’N’N’** correspond to the reverse complement sequences).

2. 3′ linkers:

   Upper: NNTCGTATGCCGTCTTCTGCTTG.

   Lower: CAAGCAGAAGACGGCATACGA.

3. TE buffer: 10 mM Tris–HCl (pH 7.5) and 1 mM EDTA (pH 8.0).

2.2 RNA Extraction

1. Liquid nitrogen.
2. Mortar and pestle.
3. Direct-zol™ RNA MiniPrep kit (Zymo Research).

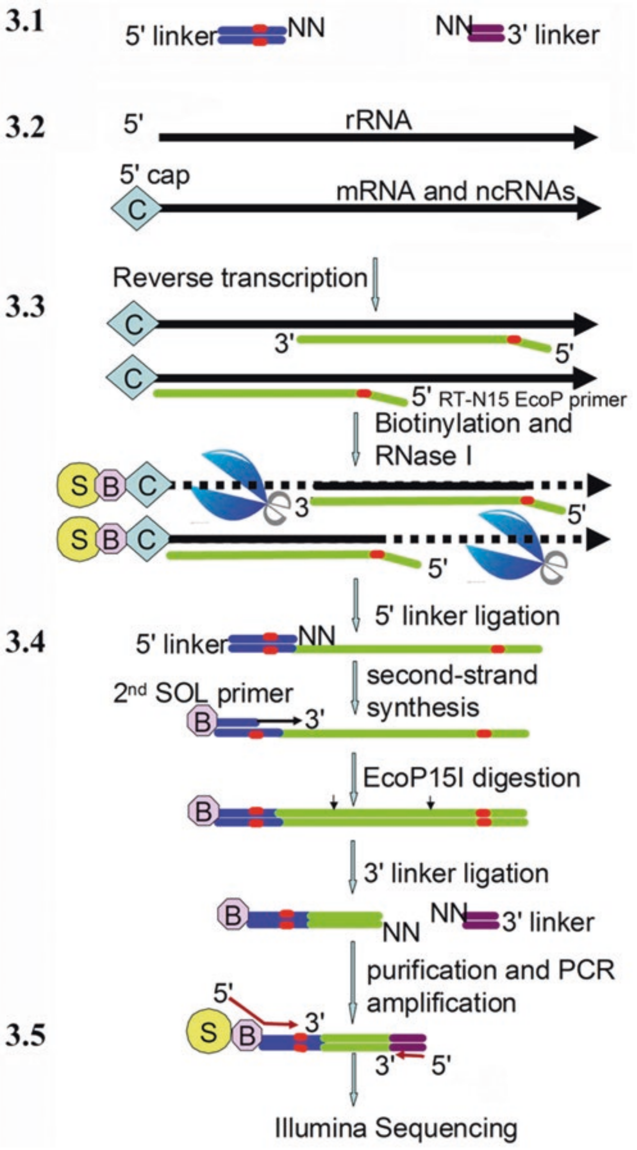


Fig. 1 Workflow of CAGE library preparation. First, prepare 5′ and 3′ linkers. Second, isolate total RNA from tissues. Third, generate single-strand cDNA. Fourth, ligate with 5′ and 3′ linkers. Fifth, amplify and purify CAGE library for sequencing. The numbers indicate the steps in the protocol describe below

4. Agilent RNA 6000 Nano Kit.
5. Agilent 2100 Bioanalyzer.

2.3 Single-Strand cDNA Preparation

1. RT-N15-EcoP primer: AAGGTCTATCAGCAGNNNNNN NNNNNNNNNN.
2. PrimeScript reverse transcriptase (Clontech).

3. 3.3 M Sorbitol/0.66 M trehalose mix, autoclaved at 121 °C for 30 min. Trehalose and sorbitol should be of high quality.
4. Agencourt RNAClean XP kit.
5. 250 mM $NaIO_4$, prepared freshly and kept at room temperature and in the dark until used.
6. 15 mM biotin (long arm) hydrazide, make sure it is completely dissolved.
7. RNase ONE ribonuclease (Promega).
8. 20 μg/μL *E. coli* tRNA: Dissolve 30 mg of *E. coli* tRNA in 400 μL of sterile ddH_2O, pretreat with 30 μL of RQ1 RNase-free DNase at 37 °C for 2 h, and then incubate at 45 °C for 30 min with 10 μL of 10 ng/mL proteinase K, clean up with 500 μL phenol/chloroform, precipitate with 500 μL isopropanol and dissolve in 1.5 mL sterile ddH_2O.
9. MPG® streptavidin (PureBiotech LLC) (MPG = magnetic porous glass particles).
10. Wash buffer 1: Mix well 45 mL of 5 M NaCl and 5 mL of 0.5 M EDTA (pH 8.0).
11. Wash buffer 2: Mix well 3 mL of 5 M NaCl, 100 μL of 0.5 M EDTA (pH 8.0) and 46.9 mL of sterile ddH_2O.
12. Wash buffer 3: Mix well 1 mL of 1 M Tris–HCl (pH 8.5), 100 μL of 0.5 M EDTA (pH 8.0), 25 mL of 1 M sodium acetate (pH 6.1), 2 mL of 10% (wt/vol) SDS and 21.9 mL of sterile ddH_2O. Make sure that there are no crystals before using.
13. Wash buffer 4: Mix well 500 μL of 1 M Tris–HCl (pH 8.5), 100 μL of 0.5 M EDTA (pH 8.0), 25 mL of 1 M sodium acetate (pH 6.1) and 24.4 mL of sterile ddH_2O.
14. 1 M Tris–HCl (pH 7.0).
15. Agencourt AMPure XP kit.
16. Agilent Bioanalyzer RNA Pico Kit.
17. Dynal magnetic stand.
18. Centrifugal concentrator.
19. NanoDrop 1000 spectrophotometer.

2.4 5′ and 3′ Linker Ligation

1. DNA ligation Mighty Mix (Takara Bio USA, Inc.).
2. Second SOL primer: Bio CCACCGACAGGTTCAGAGTT CTACAG.
3. TaKaRa LA Taq.
4. Antarctic phosphatase (New England Biolabs).
5. *Eco*P15I (New England Biolabs).
6. 10 mM Sinefungin (Calbiochem-Novabiochem International).
7. T4 DNA ligase (New England Biolabs).

2.5 Library Preparation and Sequencing

1. PCR primers: Forward: AATGATACGGCGACCACCGACAGGTTCAGAGTTC. Reverse: CAAGCAGAAGACGGCATACGA.
2. Phusion high-fidelity DNA polymerase (New England Biolabs).
3. Bio-Rad T100™ Thermal cycler.
4. Exonuclease I (New England Biolabs).
5. Agilent Bioanalyzer High Sensitivity DNA Assay Kit.
6. MinElute PCR Purification Kit (Qiagen).
7. Illumina Standard Cluster Generation Kit.
8. Sequencing primer: CGGCGACCACCGACAGGTTCAGAGTTCTACAG.
9. Illumina HiSequation 2000 sequencer.

3 Methods

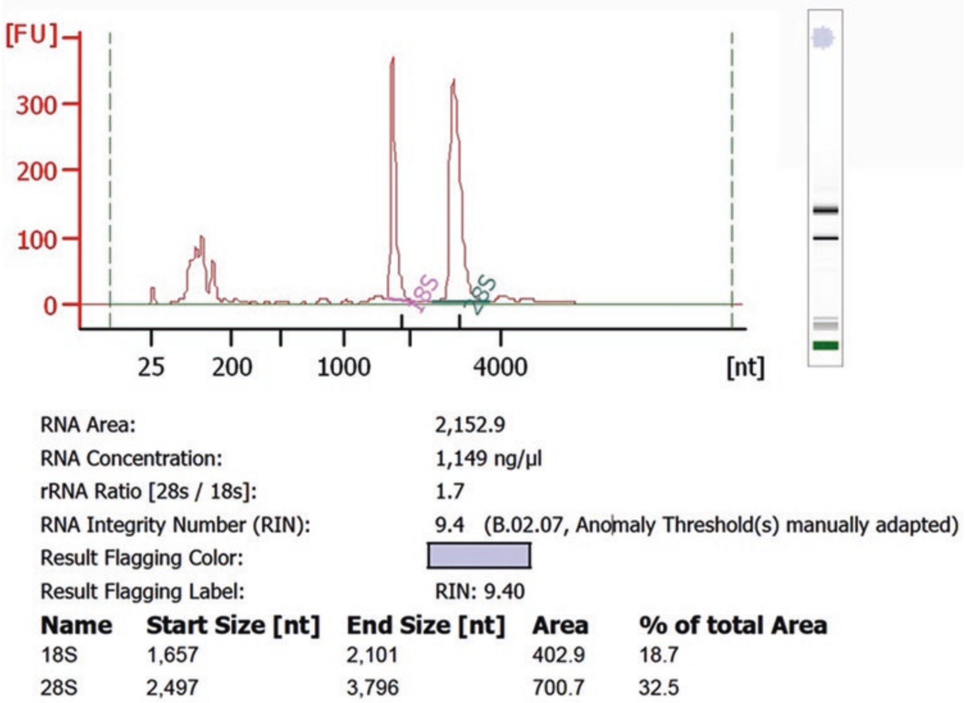
The approach described here is largely based on the CAGE method previously reported [13, 14] with some modifications. We strongly recommend the reader to consult these publications for additional tips.

3.1 Bar-Coded 5′ Linker and 3′ Linker Preparation

1. Prepare each HPLC-grade 5′ and 3′ linker in TE buffer to a final concentration of 2 μg/μL.
2. Set up 7 μL N6-linker reaction with 1.5 μL of each specific 5′ N6 upper linker, 1.5 μL of each specific 5′ lower linker, 0.8 μL of 1 M NaCl and 3.2 μL of sterile ddH_2O.
3. Set up 30 μL GN5-linker reaction with 6 μL of each specific 5′ GN5 upper linker, 6 μL of each specific 5′ lower linker, 3 μL of 1 M NaCl and 15 μL of sterile ddH_2O (*see* **Note 1**).
4. Set up 13.5 μL 3′ linker reaction with 2.5 μL of 3′ upper linker, 2.5 μL of 3′-lower linker, 1.3 μL of 1 M NaCl and 6.2 μL of sterile ddH_2O.
5. Incubate the linker reaction in the thermal cycler as follows: 95 °C, 5 min; decrease by 0.1 °C/s down to 83 °C; 5 min at 83 °C; decrease by 0.1 °C/s down to 71 °C; 5 min at 71 °C; decrease by 0.1 °C/s down to 59 °C; 5 min at 59 °C; decrease by 0.1 °C/s to 47 °C; 5 min at 47 °C; –0.1 °C/s, to 35 °C; 5 min at 35 °C; –0.1 °C/s to 23 °C; 5 min at 23 °C; decrease by 0.1 °C/s to 11 °C, and then hold at 4 °C until ready to process.
6. Mix the N6 and GN5 5′ linker suspensions carrying the same bar code. The final annealed linker concentration should be 800 ng/μL. Keep the linkers at –20 °C until ready to use.

3.2 RNA Isolation and Preparation

1. Grind maize tissues to a very fine powder in liquid nitrogen with mortar and pestle. Do not let the powder thaw at any time.
2. Use the Direct-zol™ RNA MiniPrep kit to isolate RNA from maize seedlings by following the manufacturer's instruction.



RNA Area:	2,152.9
RNA Concentration:	1,149 ng/µl
rRNA Ratio [28s / 18s]:	1.7
RNA Integrity Number (RIN):	9.4 (B.02.07, Anomaly Threshold(s) manually adapted)
Result Flagging Color:	🟪
Result Flagging Label:	RIN: 9.40

Name	Start Size [nt]	End Size [nt]	Area	% of total Area
18S	1,657	2,101	402.9	18.7
28S	2,497	3,796	700.7	32.5

Fig. 2 RNA quality determination with the Agilent RNA 6000 Nano Kit. One µL of RNA was quantified. Obtained RNAs should have a RIN value above 8. RNA concentration should be more than 950 ng/µL

3. Quantify RNA concentrations with the Agilent RNA Nano Kit (*see* **Note 2** and Fig. 2). Five µg of total RNA will be used for the next step.

3.3 Single Strand cDNA Preparation

1. Mix 5 µg of total RNA, 2.2 µL of RT-N15-EcoP primer (210 µM), and add sterile ddH2O to 7.5 µL. Incubate at 65 °C for 5 min and then cool immediately in ice.
2. Set up 30 µL reactions as follows: 7.5 µL 5× PrimeScript buffer, 1.9 µL of 10 mM dNTPs, 7.5 µL of 3.3 M Sorbitol/0.66 M trehalose mix solution, 3.8 µL PrimeScript reverse transcriptase (200 U/µL) and 9.3 µL sterile ddH$_2$O.
3. Add the enzyme mix from **step 2** to the RNA and primer mix obtained in **step 1**, and then mix them by pipetting on ice.
4. Incubate as follows: 25 °C, 30 s; 42 °C, 30 min; 50 °C, 10 min; 56 °C, 10 min; 60 °C, 10 min; finally, keep the mixture on ice until ready to use.
5. Mix 67.5 µL of RNAClean XP and the RT reaction from **step 4** thoroughly by pipetting up and down six times. Incubate at room temperature for 30 min, mixing every 10 min by gently pipetting up and down.
6. Place the reaction on the magnetic stand. After ensuring that the beads are settled on the tube wall, aspirate the cleared supernatant carefully and discard. Be sure not to aspirate beads.

7. Keep the sample on the magnetic stand and wash the beads with 150 μL of 70% (vol/vol) ethanol. After checking that the beads are settled on the tube wall, aspirate the supernatant and discard. Repeat this washing step once.
8. Remove the sample from the magnetic stand, add 40 μL of 37 °C preheated sterile ddH$_2$O. Pipet gently up and down ~25 times to completely elute the nucleic acids from the beads.
9. Incubate the sample at 37 °C for 10 min and then place it on the magnetic stand for 5 min. Transfer the supernatant to a new tube and keep it on ice.
10. Set up the following reaction with NaIO$_4$: 40 μL RNA-cDNA hybrid, 2 μL sodium of 1 M acetate (pH 4.5) and 2 μL of 250 mM NaIO$_4$. Quickly cover the tube(s) with aluminum foil, incubate in ice and in the dark for 45 min.
11. Add 2 μL of 40% glycerol and mix completely to stop the oxidation reaction. Add 14 μL of 1 M Tris–HCl (pH 8.5) to bring the pH above 5.6.
12. Mix 108 μL of RNAClean XP and 60 μL of cDNA from the diol-oxidation reaction solution in **step 11** by pipetting up and down ten times. Incubate the mixture at room temperature for 30 min, mixing every 10 min by pipetting up and down (*see* **Note 3**).
13. Repeat one more time **steps 6–9**. Elute in a final volume of 40 μL sterile ddH$_2$O.
14. Set up the following reaction to biotinylate the RNA diols: 40 μL of purified oxidized cDNA/RNA hybrids from **step 13**, 4 μL of 1 M sodium citrate (pH 6.0) and 13.5 μL of 15 mM biotin hydrazide. Mix by pipetting up and down ten times; incubate at 23 °C for 16 h (overnight) or 37 °C for 3 h.
15. Add 6 μL of 1 M Tris–HCl (pH 8.5), 1 μL EDTA (0.5 M, pH 8.0), and 5 μL RNase ONE ribonuclease (10 U/μL) to the reaction prepared in **step 14**. Mix by pipetting ten times and incubate the mixture at 37 °C for 30 min. Inactivate the enzyme at 65 °C for 5 min and then cool on ice immediately for 2 min.
16. Add 125 μL of RNAClean XP to the cDNA from **step 15**, mix thoroughly by pipetting up and down ten times. Incubate the mixture at room temperature for 30 min, mixing every 10 min by gentle pipetting.
17. Repeat one more time **steps 6–9**. Elute in a final volume of 40 μL sterile ddH$_2$O and keep on ice.
18. Prepare tRNA-coated magnetic beads by adding 1.5 μL of 20 μg/μL *E. coli* tRNA mix to 100 μL of MPG beads and incubate at room temperature for 60 min, mixing every 10 min by moderate vortexing. Separate the beads on a magnetic stand and discard the supernatant. Wash the beads with 50 μL of wash buffer 1. Repeat this washing step once. Resuspend the magnetic beads in 80 μL of wash buffer 1 (*see* **Note 4**).

19. Add 40 μL of the purified cDNA from **step 17** to the 80 μL of washed MPG beads.
20. Incubate at room temperature for 30 min, vortexing moderately every 5 min. Place the mix on the magnetic stand. After checking that the beads are settled to the tube wall, aspirate the supernatant and discard.
21. Wash the beads with 150 μL of the following wash buffers: Wash buffer 1 (once), wash buffer 2 (once), wash buffer 3 (twice), and wash buffer 4 (twice). For each wash, resuspend the beads in the wash buffer and let them separate for 3 min on the magnetic stand before discarding the washing solution (*see* **Note 5**).
22. Add 60 μL of 50 mM NaOH solution to the beads with the cDNA/RNA bound (from **step 21**) and incubate the mixture at room temperature for 10 min, with occasional mixing by vortexing.
23. Place the beads on the magnetic stand and wait for 3 min. Transfer the supernatant to a new tube.
24. Add 12 μL of 1 M Tris–HCl (pH 7.0), keep the cDNA on ice.
25. Add 130 μL of AMPure XP to the cDNA and mix thoroughly by pipetting ten times. Incubate at room temperature for 30 min, mixing every 10 min by pipetting.
26. Repeat **steps 6–9**. Transfer the 35 μL of eluent to a new tube and set 5 μL aside for a quality check.
27. Measure the concentration using 1 μL of purified single-stranded cDNAs using the NanoDrop 1000 spectrophotometer. Measure the size distribution with 1 μL of the cDNA by using the Agilent Bioanalyzer RNA Pico Kit, according to the manufacturer's instructions (*see* **Note 6**).
28. Concentrate the cDNA using a centrifugal concentrator at room temperature in a siliconized tube, and then resuspend in 4 μL of sterile ddH$_2$O (*see* **Note** 7).

3.4 5′ and 3′ Linker Ligation

1. Dilute the 5′ linker from 800 ng/μL to 200 ng/μL.
2. Add 1 μL of the diluted 5′ linker to a labeled empty tube for each cDNA sample and incubate at 37 °C for 5 min. Meanwhile, incubate the 4 μL of resuspended single-stranded cDNA from Subheading 3.3, **step 28** at 65 °C for 5 min. Cool the linker and cDNA on ice for 2 min (*see* **Note 8**).
3. Add the 4 μL of cDNA, 10 μL of DNA ligation Mighty Mix to the 1 μL of 5′ linker tubes. Mix extensively and incubate at 16 °C for 16 h.
4. Add 55 μL of sterile ddH$_2$O to the 15 μL of 5′ linker–ligated cDNAs. For pooling cDNAs with different bar-coded 5′ linkers, pool the ligated cDNAs and adjust the volume to 70 μL with sterile ddH$_2$O. Because of volume constraints, four or less cDNA samples can be pooled.

5. Add 126 μL of Agencourt AMPure XP reagent to the 70 μL of cDNAs, mix extensively. Purify as in Subheading 3.3, **steps 6–9** and repeat the whole purification process one more time. The final elution volume with sterile ddH_2O should be 30.5 μL (*see* **Note 9**).

6. Assemble the second-strand synthesis (50 μL) in ice as follows: 30.5 μL cDNA from **step 5**, 5 μL 10× LA Taq buffer, 5 μL of 25 mM $MgCl_2$, 8 μL dNTPs (2.5 mM each), 1 μL of 24 μM Second SOL primer, 0.5 μL LA Taq (5 U/μL) and gently mix by pipetting up and down six times.

7. Run the thermal cycler as follows: 94 °C for 3 min, 42 °C for 5 min, 68 °C for 20 min, 62 °C for 2 min, hold at 4 °C.

8. Add 4 μL Antarctic phosphatase (5 U/μL) and 6 μL 10× Antarctic phosphatase reaction buffer to the second-strand cDNA reaction solution and mix gently by pipetting up and down ~10 times. Incubate at 37 °C for 1 h, inactivate the enzyme at 65 °C for 5 min, and then cool on ice for 2 min.

9. Add 108 μL of AMPure XP beads to the 60 μL of cDNA, mix by pipetting up and down ~10 times. Incubate at room temperature for 30 min, mixing every 10 min by pipetting.

10. Repeat Subheading 3.3, **steps 6–9**. Elute in a final volume of 30 μL of sterile ddH_2O. This purified cDNA can be frozen at −20 °C. We would probably not recommend storing longer than 1 month.

11. Set up a 40 μL reaction on ice including 30 μL purified cDNA from **step 10**, 4 μL 10× NE buffer, 0.4 μL 100× BSA, 4 μL of 10 mM ATP, 0.4 μL of 10 mM Sinefungin, 0.1 μL *Eco*P15I (10 U/μL) and 1.1 μL sterile ddH_2O. Incubate the mixture at 37 °C for 3 h.

12. Add 1 μL of 0.4 M $MgCl_2$ to the *Eco*P15I-digested cDNA. Incubate the mixture at 65 °C for 20 min. Place in ice until the next step.

13. Set up the ligation solution (80 μL) as follows: 41 μL cDNA from **step 12**, 16 μL 5× 3′ linker ligation buffer, 1 μL 3′ linker (100 ng/μL), 3 μL T4 DNA ligase (400 U/μL) and 19 μL sterile ddH_2O. Mix by pipetting up and down ~10 times in ice. Incubate the reaction solution at 16 °C for 16 h (overnight).

14. Prepare tRNA-coated magnetic beads. Mix 1 μL of 20 μg/μL *E. coli* tRNA mix with 10 μL of MPG beads and incubate at room temperature for 60 min, mixing every 10 min by pipetting up and down. Separate the beads on a magnetic stand and remove the supernatant. Wash the beads with 50 μL of wash buffer 1. Repeat this washing step. Resuspend magnetic beads in 25 μL of wash buffer 1.

15. Mix 80 μL of 3′ linker–ligated cDNA from **step 13** to the 25 μL of washed MPG beads.

16. Incubate the mixture at room temperature for 30 min, vortex moderately every 5 min. Place the reaction solution on the magnetic stand and wait 3 min for the beads to settle. Aspirate and discard the supernatant.
17. Wash the beads with 150 μL of the wash buffers as follows: wash buffer 1 (once), wash buffer 2 (once), wash buffer 3 (twice), and wash buffer 4 (twice). For each wash, resuspend the beads and allow them to separate for 3 min on the magnetic stand before discarding the washing solution (*see* **Note 10**).
18. Keep the sample on the magnetic stand and quickly wash with 50 μL of sterile ddH_2O (*see* **Note 11**).
19. After the sample is removed from the magnetic stand, add 20 μL of sterile ddH_2O to the magnetic beads. This will correspond to the template for subsequent PCR reactions and can be kept frozen for up to 1 month at −20 °C.

3.5 CAGE Library Preparation

1. Set up the PCR reaction as follows: 10 μL 5× High-fidelity buffer, 4 μL dNTPs (2.5 mM each), 0.5 μL each PCR forward and 100 μM of reverse primers, 0.5 μL Phusion polymerase (2 U/μL), 2 μL suspension of cDNA with the magnetic beads from Subheading 3.4, **step 19**, add sterile ddH_2O to 50 μL and mix by pipetting up and down ~10 times in ice.
2. Run the PCR as follows: 98 °C for 30 s; 9,13,15 or 18 cyles of 98 °C for 10 s, 60 °C for 10 s; hold at 4 °C.
3. Use 1 μL of the PCR product to measure the concentration and the product size with the Agilent Bioanalyzer High Sensitivity DNA Assay Kit (*see* **Note 12**).
4. Minimize the number of cycles by using only just enough to detect the desired band. CAGE tag peaks with molarity ~10,000 pmol/L are suitable for bulk PCR (Fig. 4). After determining the optimal PCR cycle number, prepare six tubes of PCR to amplify the 12 μL of remaining cDNAs from Subheading 3.4, **step 19** using the optimal cycle number (*see* **Note 13**).
5. Pool each three of the PCR reactions into one 1.5-mL siliconized tube to yield a total of two tubes for each CAGE library.
6. Add 1 μL of Exonuclease I (20 U/μL) to each of the 150 μL of PCR solutions, and mix by pipetting up and down in ice. Incubate at 37 °C for 1 h to get rid of primers.
7. Use the QIAquick PCR Purification Kit to purify the 151 μL of Exonuclease I-treated CAGE library. Elute in 10 μL EB buffer provided by the kit.
8. Use 1 μL of eluted DNA to check the quantity with the Agilent Bioanalyzer High Sensitivity DNA Assay Kit (*see* **Note 14**).
9. Prepare the CAGE tags from **step 7** for Illumina sequencing following the Illumina cluster generation standard protocol. Set final DNA concentration to 5 pM per lane and sequence using the Illumina HiSeq 2000 Sequencing System.

3.6 CAGE Data Analysis

The computational pipeline for the analysis described here can be divided into two series of steps (Fig. 3). First, raw reads are demultiplexed, adapters are removed to obtain CAGE tags and tags are further aligned to a reference genome. The second set of steps requires parsing the alignment files to determine the position of CAGE TSSs (CTSSs) and clustering of TSSs (TCs) into transcriptional units, according to expression and distance. After obtaining CTSSs and TCs, other downstream analyses are possible, such as determining the shape of transcriptional units (broad and sharp), or to establish the overlap between TCs positions and known gene annotations. All the software here mentioned can be run in Unix-based operating systems and is open-source.

1. Quality checks (QCs) are required to determine that the sequence quality matches expectations. For this task, a commonly used software is FastQC (*see* **Note 15**). The report from FastQC has to be evaluated according to expectations for the experiment. For example, in the case of CAGE libraries, raw reads are clearly biased in nucleotide composition at the end of the reads because of the presence of adapters flanking the CAGE tags. To obtain a FastQC report for your sample run:

```
fastqc raw_reads.fastq
```

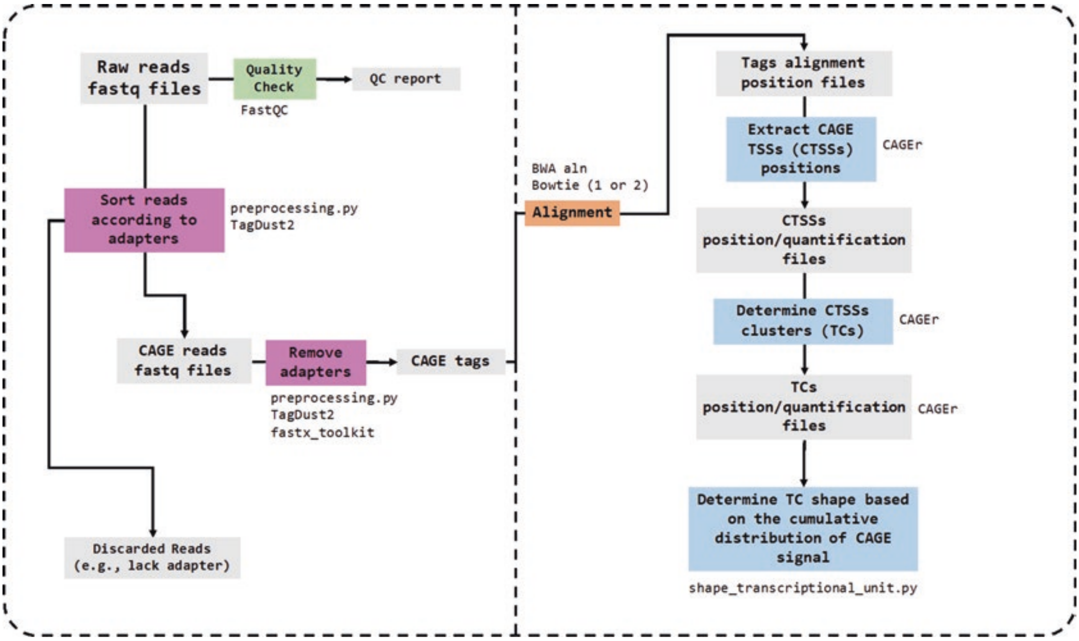

Raw reads fastq files
Quality Check
QC report
FastQC
Sort reads according to adapters
preprocessing.py
TagDust2
CAGE reads fastq files
Remove adapters
CAGE tags
preprocessing.py
TagDust2
fastx_toolkit
Discarded Reads (e.g., lack adapter)
BWA aln
Bowtie (1 or 2)
Alignment
Tags alignment position files
Extract CAGE TSSs (CTSSs) positions
CAGEr
CTSSs position/quantification files
Determine CTSSs clusters (TCs)
CAGEr
TCs position/quantification files
CAGEr
Determine TC shape based on the cumulative distribution of CAGE signal
shape_transcriptional_unit.py

Fig. 3 Computational pipeline for CAGE tag analysis. Flow diagram for a general CAGE analysis. The right half involves the preprocessing of the reads from raw fastq files to CAGE tags of ~27 bp. The left half starts with the alignment of CAGE tags to determine the CTSSs position as the general goal of the pipeline. Once CTSSs positions have been determined, subsequent analyses include CTSSs clustering (calling TCs) and determination of TC shape, which is associated with general promoter properties

2. After QC, CAGE raw reads should be filtered according to quality and sorted based on whether the adaptors are present or not. Next, reads are trimmed to remove the adapters at the 5′ and 3′ ends, and the sequence corresponding to the CAGE tags is extracted. For this steps preprocessing.py, a custom python script (*see* **Note 16**) takes a raw reads file and a CSV file containing the adapter sequences and the desired file prefix to output files with CAGE tags in a fastq format. Instead of a customized script, other programs such as fastx_toolkit and TagDust2 can be equally effective to remove adapters and obtain CAGE tags (*see* **Note 17**). Run:

```
preprocessing.py raw_reads.fastq sample_id_adapter.csv
```

3. Aligner programs such as BWA-aln, Bowtie, and Bowtie2 can be used to align CAGE tags (*see* **Note 18**). Bowtie, Bowtie2, and BWA-aln require building specific genome indexes from a fasta file before alignment (*see* **Note 19**). Run:

```
bowtie --best --stratum –v 2 --sam ZmB73v3_bowtie a.fastq a.sam

bowtie2 –x ZmB73v3_bowtie2 –U a.fastq –S a.sam
```

4. Bowtie alignment files output as SAM and need to be transformed to BAM to be used as input in the next steps using samtools (*see* **Note 20**). Run:

```
samtools view -b -S a.sam > a.bam

samtools sort a.sam –o a_sorted.bam

samtools index a_sorted.bam
```

5. Before parsing the CAGE tags alignment, a forged BSGenome from the reference genome is required (*see* **Note 21**). In the R environment run:

```
library(BSgenome)

forgeBSgenomeDataPkg("path/to/seed_file")

quit()
```

To build the BSGenome package for the maize reference genome run at the command line as follows:

```
R CMD build

R CMD check

R CMD INSTALL
```

6. To use the new BSGenome package in CAGEr [15] run the following commands in the R environment:

```
library(CAGEr)

library(BSgenome.ZmB73.AGPv3)
```

7. Read alignment files with CAGEr and build a CAGEset object as follows:

```
pathsToInputFiles =list.files("path/to/bamfiles/", full.names = TRUE)

samples =c("B73_Shoot_1", "B73_Shoot_2", "B73_Root_1", "B73_Root_2", "Mo17_Shoot_1", "Mo17_Shoot_2", "Mo17_Root_1", "Mo17_Root_2")

CAGEset_ZmB73 =new("CAGEset", genomeName = "BSgenome.ZmB73.AGPv3", inputFiles = pathsToInputFiles, inputFilesType = "bam", sampleLabels=samples)
```

   The biological replicates are merged to get CTSSs information at the sample level, as follows:

```
sample_merged =c("B73_Shoot","B73_Root","Mo17_Shoot","Mo17_Root")

mergeSamples(CAGEset_ZmB73, mergeIndex = c(1,1,2,2,3,3,4,4), mergedSampleLabels = sample_merged)
```

8. CAGEr corrects for known G addition bias at the 5′ end of CAGE tags and extracts the right position of CTSSs, as follows:

```
getCTSS(CAGEset_ZmB73)
```

9. The following step is required to quantify and normalize CTSSs expression before the clustering step, to avoid using the raw reads counts following the method described [16]. This normalization and quantification step are available in the CAGEr package, together with a helper function to fit a power-law distribution and estimate the slope (alpha) of the fitted distribution for each sample, as follows:

```
plotReverseCumulatives(CAGEset_ZmB73, fitInRange = c(3, 1000), onePlot = TRUE)
```

   The plots generated with the command above permit defining the alpha parameter for the normalization function. The total tags for the fitted power-law distribution should be set to one million in order to obtain expression quantified in TPM as unit (TPM: tags per million values).

```
normalizeTagCount(CAGEset_ZmB73, method = "powerLaw", fitInRange = c(3, 1000), alpha = 1.2, T = 1*10^6)
```

   Normalized CTSSs are now ready to be exported as bed, bedGraph, BigWig files all suitable to be observed in a genome browser such as the Integrative Genome Viewer. Run:

```
exportToBed(object = CAGEset_ZmB73, what = "CTSS", qLow = NULL, qUp = NULL, oneFile = TRUE)

exportCTSStoBedGraph(CAGEset_ZmB73, values = "normalized", format = "bedGraph", oneFile = TRUE)
```

10. Clustering TSSs into tag clusters (TCs) using a parametric method on CTSS expression [17] can be accomplished by running:

    ```
    clusterCTSS(object = CAGEset_ZmB73, threshold = 1,
        thresholdIsTpm = TRUE, nrPassThreshold = 1, method =
        "paraclu", maxDist = 150, removeSingletons = TRUE,
        keepSingletonsAbove = 3)
    ```

11. Calculating TCs width based on the cumulative distribution of CAGE signal along the promoter is achieved by running:

    ```
    cumulativeCTSSdistribution(CAGEset_ZmB73, clusters=
        "tagClusters")
    quantilePositions(CAGEset_ZmB73, clusters = "tagClusters",
        qLow = 0.1, qUp = 0.9)
    ```

    The positions of the interquantile range can be exported to bed files to be further visualized in a genome browser as follows:

    ```
    exportToBed(object = CAGEset_ZmB73, what = "tagClusters", qLow = 0.1, qUp = 0.9, oneFile = TRUE)
    ```

12. The interquartile width as calculated above can be used to determine the shape of the transcriptional unit without the need of another program. In the case of zebrafish, a width of <20 bp and <10 bp has been used to classify units as sharp, and the rest as broad [18, 19]. A different approach, similar to the thermodynamic entropy of a given state, was described for Drosophila melanogaster [20], which promised to be more robust to outliers. This approach has been implemented in a custom script that calculates a Shape Index (SI) from TCs positions and quantified TSSs for each sample and classified transcriptional units as sharp when SI > −1 and the rest as broad.

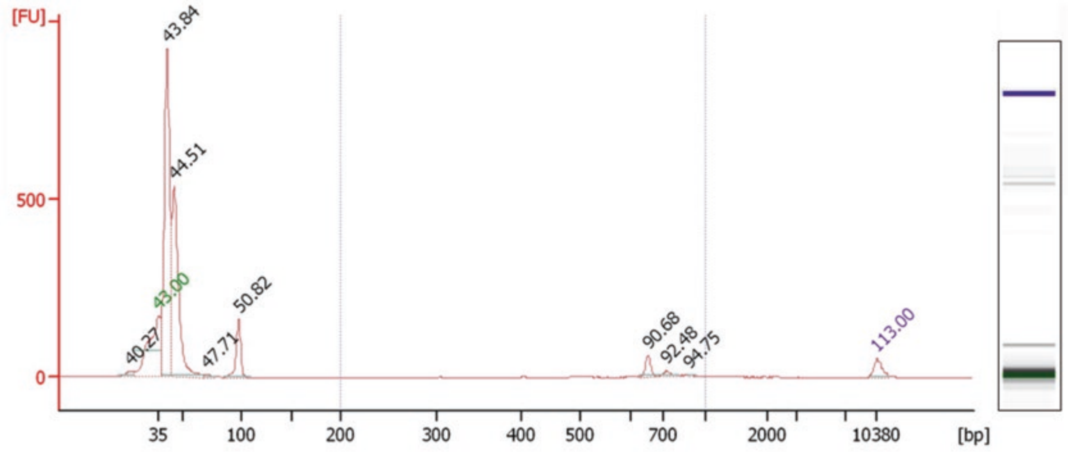
    ```
    shape_transcriptional_unit.py tc.bed tss.bed
    ```

4 Notes

1. The upper and lower linkers with same bar-code sequences are annealed to form a double strand.
2. Obtained RNAs should have a RIN value above 8. Both 260/230 and 260/280 absorbance ratios should be around or higher than 2. High quality RNA is essential for successful CAGE library preparation. Five μg of total RNA are generally used for one CAGE sample. Thus, the concentration of the RNA should be above 950 ng/μL.
3. Mix the Agencourt RNAClean XP reagent and the cDNA at a ratio of 1:1.8. The ratio is important because it is essential to have an efficient capturing of the nucleic acids. Prepare the

15 mM biotin hydrazide (long arm) solution in parallel with this procedure.

4. Coating the beads with tRNA minimizes nonspecific cDNA-bead binding and thus decreases the contamination of cDNAs that did not reach the 5′ end.
5. It is crucial to wash multiple times. This helps prevent contamination of non-capped molecules in the obtained CAGE library. Otherwise, significant amounts of rRNA sequences will be recovered.
6. When starting from 5 μg of total RNA, the concentration of the obtained cDNA will be between 3 and 10 ng/μL. The size range could be wide, from a few hundred base pairs to more than 4 kb.
7. Frequently check the remaining volume of sterile ddH_2O in the tube. Stop the procedure before it is completely dry.
8. It is essential to denature the linker and cDNA secondary structure for efficient ligation.
9. It is important to perform the purification twice to avoid any linker dimers in the final library.
10. Multiple wash iterations are critical to prevent contamination of excess 3′ linkers in the final library.
11. Perform this step as quickly as possible to avoid any DNA loss.
12. The intended product is 96 bp long; a large 40-bp peak constituted by PCR primer dimers will likely also be present. The appearance of other double-strand cDNA contaminants around 70 bp, corresponding to amplified linkers, should be minimal or absent. A ~650 bp peak may also present (Fig. 4), but at a much lower intensity than the CAGE peak, and HiSeq prefers to amplify small rather than long DNAs. Thus, sequencing should not be affected by the ~670 bp peak. The PCR products can be stored at 4 °C for 1 week or −20 °C for at least 1 month.
13. Do not carry out amplification in a single tube with a large amount of beads; the beads may inhibit the PCR (in general, do not amplify more than 2 μL of beads for a 50 μL PCR reaction).
14. After purification, PCR primer dimers should be completely removed. The major peak should be of 96 bp. Extra bands (around 80 bp) comprise linker dimers. If the concentration is much lower compared with the desired CAGE reads, this will not affect the sequencing result. However, if the concentration is high, the library could be run onto an 8% polyacrylamide gel at 120 V for 1 h, and the 96 bp band could be cut and purified with the QIAquick gel extraction kit (Qiagen).
15. Obtain FastqC from (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and follow instructions to install dependencies.



Peak		Size [bp]	Conc. [pg/µl]	Molarity [pmol/l]	Observations
1		25	0.00	0.0	
2	◀	35	125.00	5,411.3	Lower Marker
3		40	5,260.83	197,952.0	PCR primer dimers
4		44	3,507.23	119,467.3	
5		70	43.98	949.4	
6		97	633.78	9,865.1	← CAGE tags
7		653	93.76	217.5	
8		722	46.05	96.6	
9		886	8.75	15.0	
10	▶	10,380	75.00	10.9	Upper Marker

Fig. 4 Analysis of the amplified (14 cycles) PCR products by the Agilent Bioanalyzer High Sensitivity DNA Assay Kit. The measured size may slightly differ from the actual 96 bp because of instrument calibration. CAGE tag peaks with molarity ~10,000 pmol/L are suitable for bulk PCR

16. To use the provided custom scripts, download the readme file and follow instructions to install python dependencies.

17. For instructions to download fastx_toolkit and TagDust2 [21] follow the links http://hannonlab.cshl.edu/fastx_toolkit/download.html, http://sourceforge.net/projects/tagdust/files/tagdust-2.13.tar.gz/download

18. Download Bowtie (http://bowtie-bio.sourceforge.net/index.shtml) or Bowtie 2 (http://bowtie-bio.sourceforge.net/bowtie2/manual.shtml) latest versions [22, 23]. In addition to the reference genome, CAGE tags should be aligned against organelle and rDNA sequences to discard contaminants. Prebuilt Bowtie and Bowtie2 indexes are available to download for the nuclear genome (chromosomes 1–10) of the B73 reference maize genome (RefGen_v3) (http://grassius.org/).

19. Alignments are preferentially restricted to discard CAGE tags that map to multiple positions (multimappers). However, in several scenarios, the multimappers can be recovered to improve coverage or when interested in investigating the roles for the repetitive portion of the genome. Multimappers pro-

grams such as MuMRescue [24] and MuMRescueLite [25] (http://fantom.gsc.riken.jp/software/) can be incorporated into the pipeline, if necessary.

20. Download Samtools from (http://www.htslib.org/download/).
21. This series of steps requires R, and packages BSGenome [26] and CAGEr [15] to be installed. For specific details refer to the programs vignettes.

Acknowledgments

This work was supported by grants from the USA National Science Foundation IOS-1125620 and IOS-1733633 to A.I.D and E.G.

References

1. Brkljacic J, Grotewold E (2017) Combinatorial control of plant gene expression. Biochim Biophys Acta 1860:31–40
2. Meyer CA, Liu XS (2014) Identifying and mitigating bias in next-generation sequencing methods for chromatin biology. Nat Rev Genet 15:709–721
3. Davie K, Jacobs J, Atkins M, Potier D, Christiaens V, Halder G, Aerts S (2015) Discovery of transcription factors and regulatory regions driving in vivo tumor development by ATAC-seq and FAIRE-seq open chromatin profiling. PLoS Genet 11:e1004994
4. Giresi PG, Kim J, McDaniell RM, Iyer VR, Lieb JD (2007) FAIRE (formaldehyde-assisted isolation of regulatory elements) isolates active regulatory elements from human chromatin. Genome Res 17:877–885
5. Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, Sheffield NC, Stergachis AB, Wang H, Vernot B (2012) The accessible chromatin landscape of the human genome. Nature 489:75–82
6. Rodgers-Melnick E, Vera DL, Bass HW, Buckler ES (2016) Open chromatin reveals the functional maize genome. Proc Natl Acad Sci U S A 113:E3177–E3184
7. Juven-Gershon T, Kadonaga JT (2010) Regulation of gene expression via the core promoter and the basal transcriptional machinery. Dev Biol 339:225–229
8. Lenhard B, Sandelin A, Carninci P (2012) Metazoan promoters: emerging characteristics and insights into transcriptional regulation. Nat Rev Genet 13:233–245
9. Shiraki T, Kondo S, Katayama S, Waki K, Kasukawa T, Kawaji H, Kodzius R, Watahiki A, Nakamura M, Arakawa T, Fukuda S, Sasaki D, Podhajska A, Harbers M, Kawai J, Carninci P, Hayashizaki Y (2003) Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. Proc Natl Acad Sci U S A 100:15776–15781
10. Batut P, Gingeras TR (2013) Rampage: Promoter activity profiling by paired-end sequencing of 5′-complete cdnas. Curr Protoc Mol Biol 104:Unit 25B 11
11. Ni T, Corcoran DL, Rach EA, Song S, Spana EP, Gao Y, Ohler U, Zhu J (2010) A paired-end sequencing strategy to map the complex landscape of transcription initiation. Nat Methods 7:521–527
12. Mejia-Guerra MK, Li W, Galeano NF, Vidal M, Gray J, Doseff AI, Grotewold E (2015) Core promoter plasticity between maize tissues and genotypes contrasts with predominance of sharp transcription initiation sites. Plant Cell 27:3309–3320
13. Takahashi H, Kato S, Murata M, Carninci P (2012) CAGE (cap analysis of gene expression): A protocol for the detection of promoter and transcriptional networks. In: Deplancke B, Gheldof N (eds) Gene regulatory networks: methods and protocols, Methods in molecular biology, vol 786. Humana Press Inc., Totowa, NJ, pp 181–200
14. Takahashi H, Lassmann T, Murata M, Carninci P (2012) 5′ end-centered expression profiling using cap-analysis gene expression and next-generation sequencing. Nat Protoc 7:542–561
15. Haberle V, Forrest AR, Hayashizaki Y, Carninci P, Lenhard B (2015) Cager: precise TSS data retrieval and high-resolution promoterome mining for integrative analyses. Nucleic Acids Res 43:e51

16. Balwierz PJ, Carninci P, Daub CO, Kawai J, Hayashizaki Y, Van Belle W, Beisel C, van Nimwegen E (2009) Methods for analyzing deep sequencing expression data: constructing the human and mouse promoterome with deepcage data. Genome Biol 10:R79
17. Frith MC, Valen E, Krogh A, Hayashizaki Y, Carninci P, Sandelin A (2008) A code for transcription initiation in mammalian genomes. Genome Res 18:1–12
18. Nepal C, Hadzhiev Y, Previti C, Haberle V, Li N, Takahashi H, Suzuki AM, Sheng Y, Abdelhamid RF, Anand S, Gehrig J, Akalin A, Kockx CE, van der Sloot AA, van Ijcken WF, Armant O, Rastegar S, Watson C, Strahle U, Stupka E, Carninci P, Lenhard B, Muller F (2013) Dynamic regulation of the transcription initiation landscape at single nucleotide resolution during vertebrate embryogenesis. Genome Res 23:1938–1950
19. Haberle V, Li N, Hadzhiev Y, Plessy C, Previti C, Nepal C, Gehrig J, Dong X, Akalin A, Suzuki AM, van IWF, Armant O, Ferg M, Strahle U, Carninci P, Muller F, Lenhard B (2014) Two independent transcription initiation codes overlap on vertebrate core promoters. Nature 507:381–385
20. Hoskins RA, Landolin JM, Brown JB, Sandler JE, Takahashi H, Lassmann T, Yu C, Booth BW, Zhang D, Wan KH, Yang L, Boley N, Andrews J, Kaufman TC, Graveley BR, Bickel PJ, Carninci P, Carlson JW, Celniker SE (2011) Genome-wide analysis of promoter architecture in *Drosophila melanogaster*. Genome Res 21:182–192
21. Lassmann T, Hayashizaki Y, Daub CO (2009) TAGDUST – a program to eliminate artifacts from next generation sequencing data. Bioinformatics 25:2839–2840
22. Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357–359
23. Langmead B (2010) Aligning short sequencing reads with Bowtie. Curr Protoc Bioinformatics Chapter 11:Unit 11 17
24. Faulkner GJ, Forrest AR, Chalk AM, Schroder K, Hayashizaki Y, Carninci P, Hume DA, Grimmond SM (2008) A rescue strategy for multimapping short sequence tags refines surveys of transcriptional activity by CAGE. Genomics 91:281–288
25. Hashimoto T, de Hoon MJ, Grimmond SM, Daub CO, Hayashizaki Y, Faulkner GJ (2009) Probabilistic resolution of multi-mapping reads in massively parallel sequencing data using mumrescuelite. Bioinformatics 25:2613–2614
26. Pagès H (2018) BSgenome: Software infrastructure for efficient representation of full genomes and their SNPs. R package version 1.48.0