Quantitative analysis of dosage compensation in flies using promoter profiling

Vivek Bhardwaj[1,2]*, Giuseppe Semplicio[1,2]*, Thomas Manke[1], Asifa Akhtar[1#]

[1]Max Planck Institute for Immunobiology and Epigenetics, 79108, Freiburg, Germany

[2]Faculty of Biology, University of Freiburg, 79104, Freiburg, Germany

*equal contribution

#corresponding author

akhtar@ie-freiburg.mpg.de

Phone: +49 (0)7615108565

Fax: +49 (0)7615108566

Abstract

Promoter architecture, shape and position of transcription start sites (TSS) play an important role in the regulation of eukaryotic gene expression. Promoter profiling methods like CAGE (Cap Analysis of Gene Expression) are widely used to detect transcription start sites and alternative promoter usage between tissues or during development. However, these methods are rarely used for differential expression analysis. In this study, we describe an approach to combine promoter profiling and differential expression analysis in a single setup, using a fast and simple protocol, MAPCap (Multiplexed Affinity Purification of Capped RNA) along with a new tool "icetea" (https://bioconductor.org/packages/icetea). icetea enables detection of TSS at high-resolution after UMI-based removal of PCR duplicates and detects differential gene expression and promoter usage using both internal and external normalization controls. Using MAPCap and icetea, we analyzed TSS expression in the brains of *Drosophila melanogaster* larvae, and observed the effects of knockout of MLE (male-less) helicase on X chromosome dosage compensation at promoters. Our results expand the scope of TSS profiling methods to differential expression analysis.

Introduction

Most genes in eukaryotes express multiple isoforms and transcript isoform expression is a major mechanism behind tissue-specific regulation of gene expression. Isoform diversity could be achieved by the usage of alternative exons, UTRs, transcript start and end sites. Recent analysis of Flies and Human genome has suggested that transcript start and end site selection is a major driver of alternative isoform usage across tissues [1]. Promoter profiling methods, such as CAGE [2], RAMPAGE [3], NanoCAGE [4] and GRO-cap [5] are widely to identify transcription start sites (TSSs), therefore making them useful for the detection of alternative isoform usage across tissues or developmental stages. A recent comparative analysis of six

such methods identifies CAGE as the overall best method [6] while RAMPAGE [3] comes close second. The amount of time and number of steps required per library was found to be highest for CAGE. RAMPAGE reduces the processing time (to 2 days) and the required input material (up to 5uG) therefore providing a suitable alternative. Despite this improvement, the CAGE methods have not gained wide usage besides the detection of new TSS and analysis of promoter architecture.

In this study, we developed a short and easy to perform 5' profiling method, termed as MAPCap (Multiplexed Affinity Purification of Capped RNA), which allows multiplexing of samples and produces paired-end reads. Synthetically designed random barcodes allow removal of PCR duplicates, and external spike-in controls allow accurate relative quantification of TSS expression changes. Further we developed an R/Bioconductor package **icetea** (Integrating Cap Enrichment and Transcript Expression Analysis) (https://bioconductor.org/packages/icetea), which allows easy processing and analysis of data obtained from protocols such as MAPCap and RAMPAGE. We performed MAPCap on brains isolated from 3rd instar Fly larvae and quantified the defect of dosage compensation upon knockout of male-less (MLE) gene on X chromosome at transcription start site resolution.

Results

MAPCap enriches Capped RNA from multiplexed samples

A new type of adapter oligo developed previously by our group (called the s-oligo) has dramatically increased the speed to perform transcriptome-wide RNA-protein interaction assays [7] . One main reason is the suppression of so-called "adapter-dimers" which can frequently occur in ligation-mediated clonings of RNA, a widely used approach in RNA library preparation, including promoter-profiling methods. The successful application of the s-oligo in CLIP experiments prompted us to develop a similar approach for promoter-profiling. MAPCap is a method that combines the power of the s-oligo with the easy handling of bead-based affinity purifications (Fig. 1A, see methods). This allows a for a fast and reproducible processing of even low RNA input amounts. Abundant RNA species such as sn- and snoRNAs are selectively

degraded by targeted antisense oligos and RNase H increasing the recovery of other capped RNA species. The s-oligo incorporates the sequences of both standard sequencing adapters which omits the usage of an RT-primer and allows for a highly efficient intramolecular ligation. The linear PCR amplification product creates a uniform library with ideal insert sizes of around 150 nt (Fig. S1A).

We performed MAPCap on stage 15 *Drosophila* embryos, in four replicates, and obtained ~10 Million reads per sample after de-multiplexing (Fig. S1B, see methods). For comparison, we analysed the CAGE data downloaded from modENCODE [8] and the RAMPAGE data [3] from embryos corresponding to the same stage. Reads obtained from both CAGE and RAMPAGE protocols show a high 'G' nucleotide bias due to the template-free activity of the reverse-transcriptase during cDNA preparation in RAMPAGE [9], and the design of the attached linkers in CAGE [10] . This demands post-mapping correction and sometimes affects the accuracy of TSS detection [11] . MAPCap, on the other hand, shows no such bias due to the use of s-oligo (Fig. 1B).

In order to evaluate the performance of MAPCap for TSS detection, we first performed TSS calling from MAPCap data using the paraclu method, and compared them to the TSS detected from CAGE and RAMPAGE data using the same method and filtering parameters (see methods). We then evaluated the TSS detection sensitivity, specificity, precision and F1-score between the methods, by comparing them to all annotated TSS present in the *Drosophila* ensembl annotation (release 76) as well as the RNA-Seq data obtained from modENCODE (see methods). CAGE performed the best amongst the three methods, as observed before [6], followed by RAMPAGE and MAPCap (Fig S1C).

We then correlated the depth-normalized counts on 5'-UTRs of known genes between MAPCap, CAGE, RAMPAGE and total RNA-Seq data from the same stage obtained from modENCODE project (see methods). We find that although MAPCap signal shows good correlation with other protocols, CAGE and RAMPAGE show better correlation with each other while MAPCap showed better correlation with RNA-Seq (Fig 1B-C, S1D). An independently performed MAPCap experiment with S2 cells shows the same relationship between the protocols, suggesting that MAPCap produces gene expression estimates closer to RNA-Seq compared to these protocols.

Figure 1

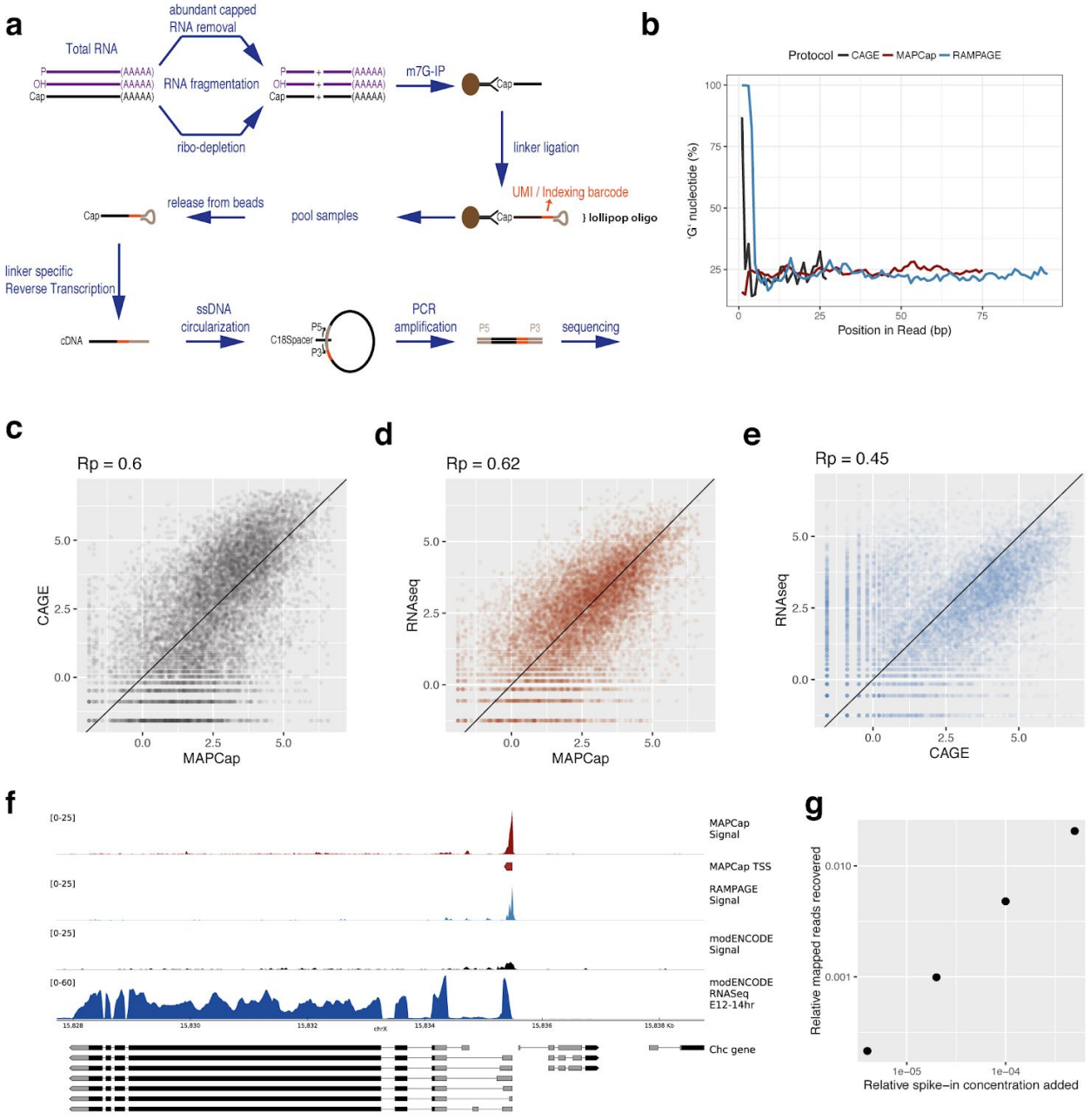

Fig 1. MAPCap protocol and its data quality. A. *Overview of the MAPCap protocol. After fragmentation and ribo-depletion, the Capped RNA is immunoprecipitated using an antibody, and the s-oligos are attached, afterwards the samples are pooled for PCR and library preparation steps.* ***B.*** *Nucleotide content in read positions. CAGE and RAMPAGE show a high*

artificial G-bias, due to their cloning steps, while MAPCap shows low bias for any specific nucleotide. ***C-E.*** *Correlation of MAPCap counts with CAGE, RNA-seq and the correlation of counts between CAGE and RNA-seq, on 3'UTR of genes. MAPCap shows better correlation with RNA-seq than CAGE.* ***F.*** *Genome track of the de-duplicated counts from MAPCap, RAMPAGE and CAGE on TSS. For MAPCap and RAMPAGE, the de-duplication was performed using 5'-position of the reads and the UMIs, while for CAGE it was performed using only 5'-position. RNA-seq track is shown for comparison.* ***G.*** *Added relative concentration (w.r.t. total RNA concentration) vs recovered relative counts (w.r.t total read counts) for the embryos. The samples were added with increasing relative concentration of ERCCs.*

The random barcodes present in lollipop oligos allow us to remove PCR duplicates, while preserving the transcript expression signal. Similar de-duplication can be performed for data obtained from the RAMPAGE protocol, where the oligos used as RT-PCR primers also serve as 'pseudo-random' barcodes [3]. A comparison of PCR duplicate removed signal from the three protocols show that both MAPCap and RAMPAGE protocols preserve the signal on the TSS, while de-duplication in absence of UMIs lead to near-complete loss of signal from the CAGE protocol (Fig. 1E).

Finally, we tested the sensitivity and relative quantification accuracy of MAPCap protocol for RNAs at different concentration by using external ERCC controls. We prepared spike-in mix containing 10 in-vitro capped ERCC spikes (see methods) in a 2-fold relative concentration ranging from 15.6 pM to 8 nM. We then mixed each replicate of the embryo sample with different concentrations of this spike-in mix (from 0.0004% to 0.05% of isolated RNA), before the beginning of the protocol. Processing of data shows that the relative concentration of spike-ins betweens sample can be faithfully recovered after sequencing (Fig. 1G). Relative ratio between individual spike-ins within each mix could also be accurately recovered (Fig. S1F), suggesting that MAPCap provides good sensitivity and accuracy to detect original transcript concentrations.

High resolution TSS detection and differential expression analysis using biological replicates

The ease of use and multiplexing ability of MAPCap protocol allows performing biological replicates without adding additional time and effort. We therefore sought to develop analysis methods which could benefit from biological replicates. Popularly used methods for TSS detection (parclu and distclu) are performed on within sample clustering of tags, and don't incorporate biological replicates to improve the performance of TSS detection. We developed a window-based TSS detection method that borrows ideas from window based peak calling methods developed for ChIP-Seq [12,13]. Reads counts are modelled using negative binomial distribution and the TSSs are detected as windows of enrichment in the genome, with respect to a local background (Fig. S2A, see methods). Consecutively enriched windows are then merged to detect both short and long TSSs. We compared peaks obtained from our method with those from paraclu method on the embryos to evaluate sensitivity and specificity. TSSs detected from the new method show higher sensitivity as well as specificity (Fig. 2a).

Apart from the accuracy of TSS detection, the shape of TSSs have also been shown to reflect biologically meaningful signal. Genes with sharp, focussed TSSs are mostly tissue specific and developmentally regulated, while the genes with long TSSs are shown to have housekeeping functions [14]. GO term enrichment analysis of the sharp and broad TSSs obtained from our method confirms previous results (see methods). Genes with sharp TSSs are enriched for processes like morphogenesis and development (Fig. 2B) while genes with broad TSSs are enriched for processes such as protein localization, metabolism and membrane organization (Fig. S2B), suggesting that the new method successfully detects biologically meaningful TSS shapes. Motif enrichment analysis of sharp and broad TSSs further confirm these results, with sharp TSSs being enriched for the Inr element, while broad TSS being enriched for core promoter motif M1BP and others (Fig. 2C, see methods).

Figure 2

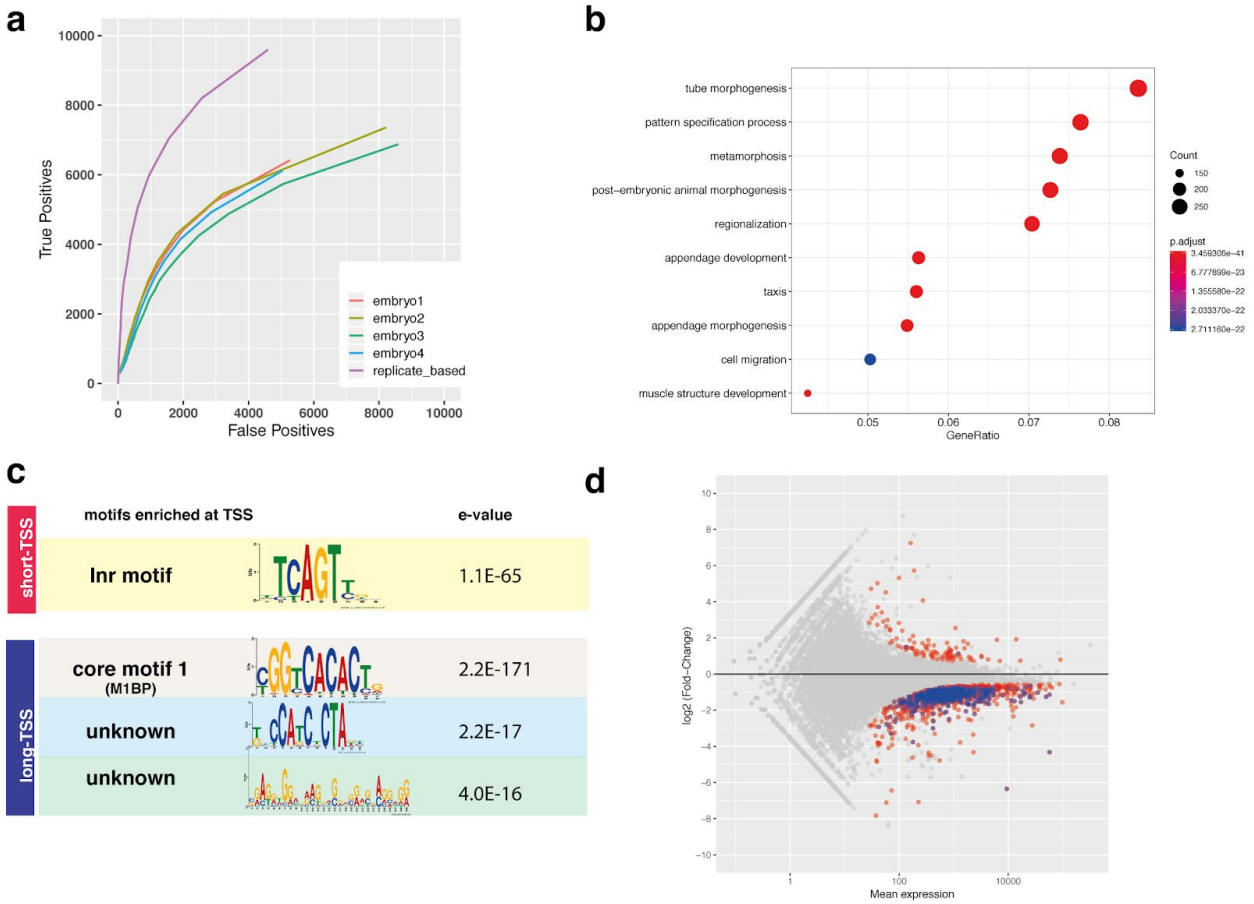

Fig 2. A replicate-based method of TSS detection. A. *ROC curve of MAPCap data on embryos, samples labelled embryo1 to 4 were processed using paraclu method and compared with the new method (replicate based) that uses all 4 samples as replicated for TSS detection. The detection accuracy improves using the new method.* ***B.*** *GO enrichment of "sharp" TSSs (<20bp) detected by the new method.* ***C.*** *Motif enrichment of the "sharp" (<20bp) and "broad" (>100bp) TSS detected by the method.* ***D.*** *Gene-level differential expression estimates obtained from MAPCap data in brains of MLE KO males. Differentially expressed TSSs on X-chromosome are highlighted in blue.*

Finally, we also sought to utilize the capped ERCC controls added to the protocol for the analysis of differential TSS usage between groups of samples. To this end, we performed MAPCap on RNA isolated from flies which are knock-outs of *maleless* (MLE) RNA helicase, and compared them with RNA from wild-type flies. The MLE helicase is an important component of the MSL complex which recognizes roX RNAs on the X-chromosome and helps guide the MSL complex to the X, leading to 2-fold upregulation of the male X-chromosome. In absence of MLE,

we therefore expected X chromosome to be downregulated due to failure of dosage compensation. We adopted popularly used method of gene-level differential expression analysis to the MAPCap data and utilized the spike-ins for normalization (see methods). Results show that most genes in MLE KO were downregulated, and most of the downregulated genes were on X chromosome (Fig. 2D). In absence of spike-in normalization, however, we saw more balanced number of up and down-regulated genes, where a bias towards the X-chromosome was not clearly visible (Fig. S2C), suggesting that spike-in normalization provides more useful biological insights.

Effect of MLE KO on dosage compensation of male promoters

After confirming the validity of our experimental and analysis method, we applied MAPCap on total RNA isolated from brains of male and female *Drosophila melanogaster* larvae of both wild-type and MLE KO background (see methods). We then deployed a pipeline that performs de-multiplexing, mapping, de-duplication, TSS detection and annotation of the TSS (Fig. S3A-B). We detected TSSs using our new method, using a fold-change threshold of 4x over the background, followed by comprehensive functional annotation of the detected TSS (see methods). While most of the detected TSS originated from previously annotated TSSs, X% of detected TSS in all samples came from promoter-proximal or intergenic enhancers (Fig 3A).

We then compared the MLE KO and wild-type genotypes in both male and female brains in order to detect differential promoter usage after spike-in normalization. Comparison of wild-type male and female brains showed that most promoters are equally utilized between sexes (Fig. S3C). Similar to our gene-level differential expression analysis, our differential promoter usage analysis revealed a significant downregulation of TSS from X-chromosome in KO males (Fig. 3B), while the females showed almost no effect in promoter usage (Fig. S3D).

Figure 3

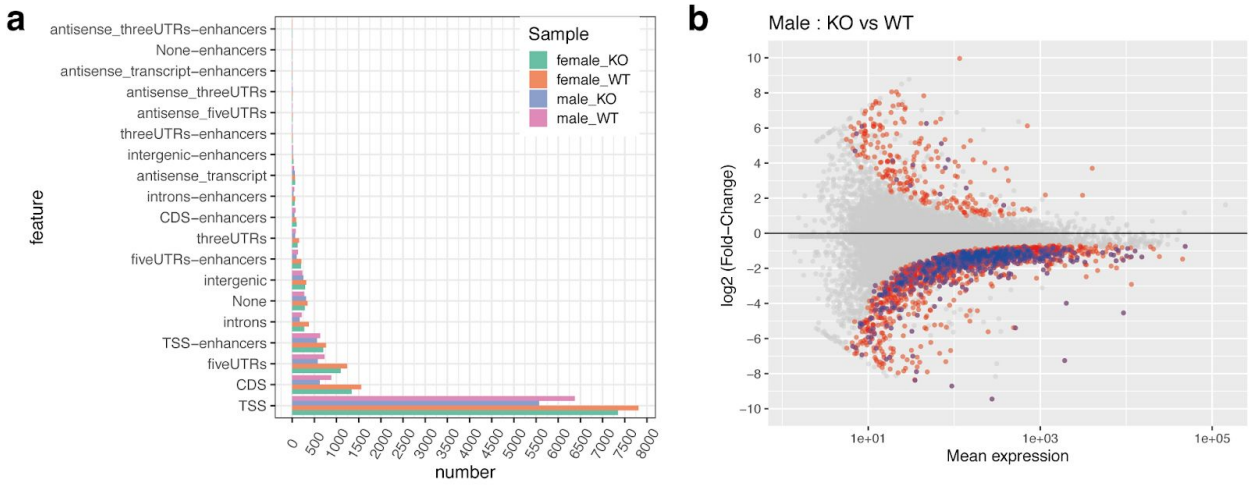

Fig 3. Assessment of promoter usage in the male and female fly brains. A. *Annotation of detected TSS into different functional categories. Both wild-type and KO samples show similar enrichment of TSS in each category.* ***B.*** *Differential TSS (promoter usage) in male KO, compared to wild-type. Differentially expressed TSSs from X-chromosome are highlighted in blue.*

icetea simplifies TSS detection and expression analysis from promoter profiling data

We implemented the processing and analysis methods described in this manuscript in an easy to use R package **icetea** (Integrating Cap Enrichment with Transcript Expression Analysis). **Icetea** performs sample de-multiplexing, PCR de-duplication as well as employs the new TSS detection approach described previously that takes advantage of biological replicates (Fig. 4A) . Further functions for quality control (Fig. 4B-C) and quick annotation of detected TSS (Fig. 4D) are also implemented. Differential TSS expression analysis can be performed between group of samples, using either internal or external (spike-in) normalization, allowing accurate quantification of relative gene and isoform expression changes (Fig. 4E). **Icetea** is especially suitable for end-to-end analysis of paired-end 5' profiling techniques such as MAPCap and RAMPAGE, however it can easily be used for analysis of CAGE, GRO-Cap and other promoter profiling protocols. **icetea** is open source and available for use via Bioconductor (https://bioconductor.org/packages/icetea) and the source code is available on GitHub (https://github.com/vivekbhr/icetea).

Figure 4

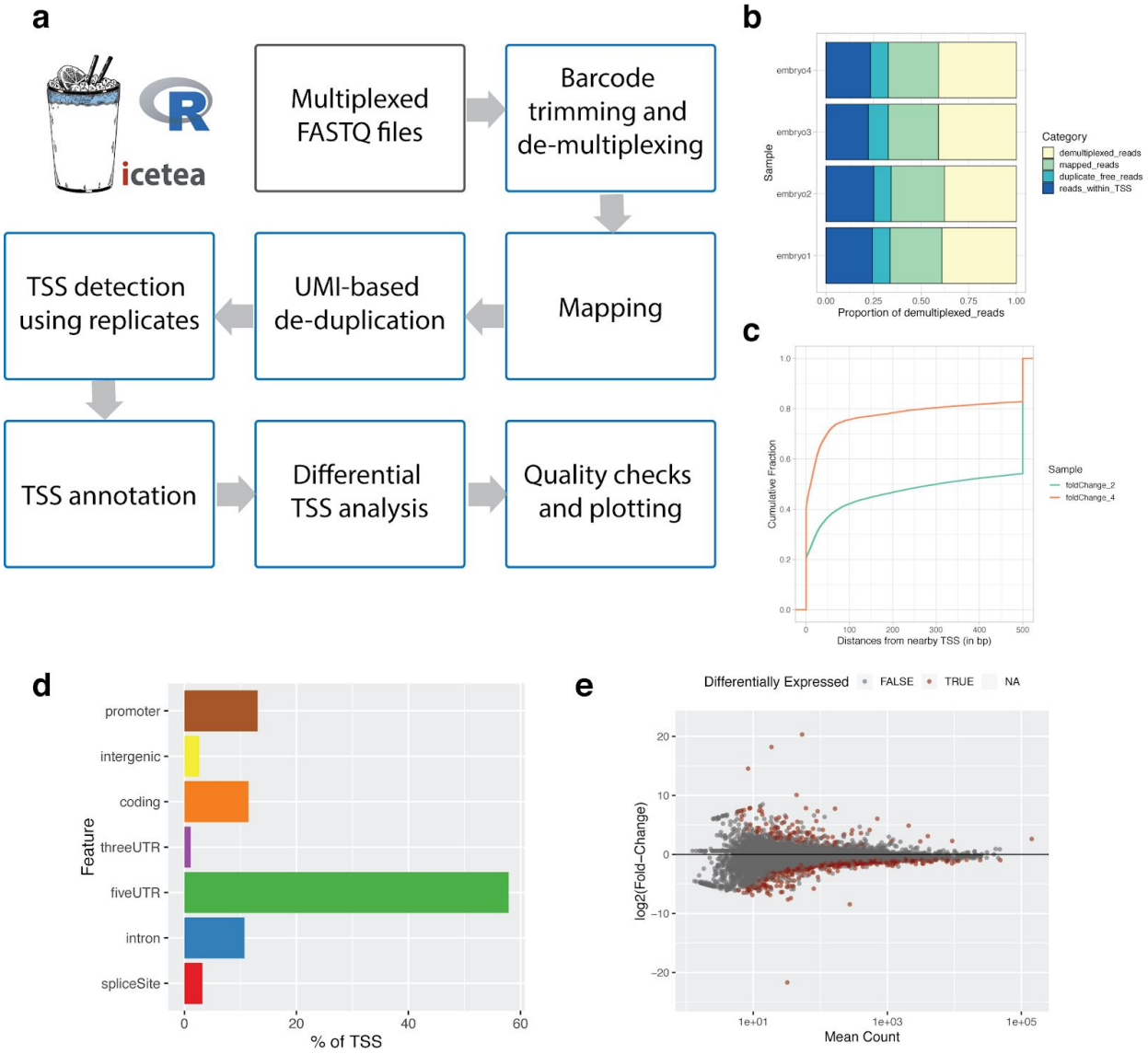

Fig 4. ***Description of the icetea bioconductor package for analysis of promoter-profiling data. A.*** *Steps of data analysis implemented in icetea. Read de-multiplexing and de-duplication based on UMI is supported for MAPCap and RAMPAGE data.* ***B-E.*** *Some examples of results obtained from icetea. B) Read mapping statstistics for embryo data (plotReadStats) C) TSS distance precision (distance of detected TSS to nearby annotated TSS) at different fold-change cutoffs on embryo data (plotTSSprecision) D) Annotation of detected TSS for male WT brains (annotateTSS). E) Differentially expressed TSS between male and female KO samples (detectDiffTSS).*

Discussion

In this study we introduce an easy to perform promoter profiling technique, along with a new analysis approach which simplifies the integration of TSS discovery and transcript expression analysis. The use of MAPCap protocol along with icetea analysis provides most benefits of CAGE and RNA-seq, at a fraction of total cost and time of performing both the protocols. Unlike CAGE or RAMPAGE which utilize protocol-specific optimization, MAPCap utilizes protocol-agnostic s-oligos, which, apart from promoter profiling, could also be used for iCLIP experiments [7], as well as for RNA-Seq, allowing for wider scope of integrative analysis. We propose that this approach would prove optimal for pilot projects of transcript discovery and gene expression analysis of newly assembled as well as annotated genomes.

Methods

Cells

S2 cells (gift from the butros lab, Heidelberg) were cultured in Express Five SFM media (Thermo Fisher) supplemented with 10% (v/v) Glutamax (Thermo Fisher). Cultures were maintained adherent or in shaking incubators at 27ºC at a speed of 80 rpm. Cells were kept at a density of 1-16 million/ml.

Generation of capped ERCC spikes

Ten spikes sequences were chosen from the ERCC spike mix, trimmed to ~500 bp length and ordered as gBlocks from IDT. At the 5' end we inserted a T7 class II promoter ϕ2.5, which has been shown to create more homogenous 5' ends transcription promoter sequence [15]. Spikes were in vitro transcribed using T7-FLASHScribe Transcription kit (CellScript) according to manufacturer's instructions and purified using MegaClear kit. In vitro capping was performed with Vaccinia capping system. Potentially uncapped RNAs were degraded by treatment of spikes with polyphosphatase and terminator. Samples were cleaned using OCC, concentrations

measured on Qubit. A master mix was created where each subsequent spike was added at half the concentration of the previous spike, starting from 8 fmol/ul.

MAPCap library preparation

RNA from S2 cells was extracted using the Quick RNA Kit (Zymo Research). RNA from dissected brains of embryos was isolated using the DirectZol Kit (Zymo Research). RNA was eluted in 25 ul of RNase-free water. The concentrations are adjusted and capped ERCC spikes and HEK polyA RNA were added at 0.05% of input amount. To remove abundant capped RNAs (snRNA, snoRNAs) as well as rRNA contamination, we added antisense DNA oligos (see table) targeting the RNA species detected from a preliminary MAPCap run. 8ul of oligo mix were added together with 4ul of 10x terminator buffer A (Epicentre). The RNA was heated to 70 ºC for 2min followed by an active cooling in the Thermomixer (Eppendorf) to 37 ºC. Upon reaching this temperature, 1ul of RNaseH (Life/Invitrogen) was added and incubated at 37 ºC for 30 min. The samples were then heated to 70 ºC for 2 min, put immediately on ice for 1 min and 1 ul of Terminator exonuclease was added for 1 hr at 30 ºC. RNA was purified using RNA clean and concentrator (Zymo Research) and eluted with 100ul TE buffer. The samples were fragmented using a Covaris E220 Ultrasonicator (200 cycles/burst, Duty cycle 5, 175 W, 10%) for 180 s per sample. Fragmented RNA was incubated with 2.5-5 ug of anti-m7G antibody (SYSY, cat no. 201001) pre-coupled to Protein G magnetic beads for 1 hour in IPP buffer (50 mM Tris-HCl pH 7.4, 150 mM NaCl, 0.1% NP-40) rotating at 4 ºC. Beads were washed three times with IPP and RNA 3' ends were dephosphorylated using PNK for 30 min at 37 ºC. Beads were washed and the s-oligo was ligated using T4 RNA Ligase 1 for 1 hr at 25 ºC. S-oligos contain barcode and random nucleotides in the following pattern NNNNNTTTTTTNN (N=random nucleotide; T=barcode nucleotide). Excess s-oligos were washed away with IPP buffer and samples were pooled together. After 30 min of treatment with rSAP at 30 ºC to dephosphorylate the s-oligo, the RNA was released from the beads using Proteinase K treatment and column purification (Oligo Clean and Concentrator (OCC), Zymo Research). Isolated RNA was reverse-transcribed using SuperScript III (Invitrogen) for 10 consecutive minutes at 42, 50, 55 and 65 ºC. After 30 min treatment with RNaseH the cDNA was column-purified using OCC and circularized with CircLigase2 for 2-16 hr. 1 ul of circularized cDNA was taken to determine the amplifications

cycles using qPCR. After PCR amplification the libraries were cleaned up twice using 1x Ampure beads (Beckman Coulter), quantified with Qubit (Thermo Fisher Scientific) and the quality was assessed on Bioanalyzer (Agilent). MAPCap libraries were sequenced on Illumina NextSeq 500, 3000 or HiSeq 2000.

Processing of MAPCap data

Paired-end FASTQ files were trimmed for adaptors using Trimmomatic [16] (v 0.3.7). Samples were de-multiplexed by icetea (v0.99, demultiplexFastq) using provided barcode information, and mapped to the dm6 genome using Rsubread [17] (v 1.22.3, mapping wrapper provided in icetea). For de-duplication, we consider all reads mapping to the same 5'-position and having the same random barcode as duplicates and only keep the first instance of each such alignment (using icetea - filterDuplicates). BigWigs were created using deepTools [18] (v3.0.2) bamCoverage and bamCompare, with the option `*--offSet 1 --binSize 1*`. Quality control was performed using deepTools and multiQC [19] (v1.3). Genomic regions were plotted using pyGenomeTracks [20] (v2.0). The full MAPCap data processing workflow (described in Fig. S3A) is available at : https://github.com/vivekbhr/cage_pipeline

Comparison with external methods

For evaluation of TSS detection accuracy, we used the *paraclu* method [21] to cluster CAGE tags from CAGE, RAMPAGE and MAPCap data. We used the 12-14hr Sample from modENCODE, and merged the 12, 13 and 14hr samples from RAMPAGE to compare with merged (embryo 1-4) samples from MAPCap data. All samples were then downsampled to 10 million reads and paraclu was run with the parameters : *min_value = 1, min_density_rise = 1, min_pos_with_data = 1, min_sum = 1, min_width = 3, max_width = 300* (i.e. all criteria such as minimum reads used for clustering and minimum density of reads per cluster etc. were kept to the lowest, and tag clusters of length 3 to 300 bp were considered for analysis).

The "density score" provided by paraclu for each tag cluster was used to calculate precision and sensitivity. Scores between 10 and 500 were plotted for ROC curve. For the evaluation of true and false positives, we used the RNA-seq data of 12-14 hr embryo from modENCODE and

calculated transcript-level TPMs using Salmon [22]. Transcripts with TPM > 1.0 were considered "expressed" and the TSSs of expressed transcripts which were not detected by the promoter-profiling methods considered "false negatives". TSSs detected by the methods which did not overlap with a known TSS in dm6 (ensembl-79) annotation were considered "false positives".

For comparison between replicate-based and paraclu method, we ran paraclu on samples with same criteria, while for our new method we obtained TSSs using a 2-fold local background cutoff. The score per TSS (mean fold-change across enriched windows) was used to evaluate the true and false positives for the analysis. Since the range of scores obtained per TSS is very different between paraclu and our method, there is no comparable cutoff for comparison of precision and sensitivity.

TSS detection and differential TSS usage analysis using replicates

For the detection of transcription start sites using replicates, we first count the 5'-end of reads in 10 bp sliding windows (w) across the genome for all samples (with a slide of 5 bp). For each window, we also calculate all 5'-ends of the reads falling into the corresponding 2 kb background region (b) centered at the window. Counting is done in a strand-specific way, using the *intersectionStrict* mode. We then calculate the fold change (delta) of each window with respect to the background as :

$$\delta = Avg(\hat{\omega}) / Avg(\hat{b})$$

Where $Avg(\hat{\omega})$ and $Avg(\hat{b})$ are average logCPM values across replicates, obtained by a fitting single group negative binomial glm :

$$\hat{Y}wi \sim NB(Mi\rho wj, \phi w)$$

implemented in *mglmOnegroup* function of the edgeR package [23].

For differential TSS usage analysis, strand-specific counting is performed in the same way, on the union of TSSs detected across samples. Library sizes were normalized using the size factors obtained from ERCC counts using median of ratios method from DESeq2. The differential expression analysis was then performed in DESeq2 using *nbinomWaldTest* function.

TSSs with and adjusted p < 0.05 were considered significantly different between tissues and sexes.

To perform differential gene expression analysis from MAPCap data, we summed the counts obtained from all 3'UTRs of a gene into one, and performed the normalization and differential expression using DESeq2. Spike-in normalization was performed the same way as above.

TSS annotation

For a comprehensive annotation of our detected TSSs, we first created a mutually exclusive set of annotations from dm6 (ensembl-79) GTF file, by first separating genic from intergenic regions , followed by ranking them in this order (5'UTR > CDS > 3'UTR > Introns; and sense > antisense). Further the features were re-annotated by overlapping them with enhancers [24] and repeats (RepBase release 20140131). The annotation pipeline is available as part of the full MAPCap data processing pipeline at : https://github.com/maxplanck-ie/cage_pipeline

Evaluation of promoter width

To evaluate the promoter width distribution obtained from icetea analysis, we divided our 12921 detected TSS into "broad" and "sharp" categories, by taking arbitrary cutoffs : >50 bp (8.1%) and <20 bp (36%), respectively. We performed GO enrichment analysis of the two categories for biological processes (BP) terms and plotted them using the *clusterProfiler* bioconductor package [25] ($p < 0.01$, $q < 0.05$). Further we, extracted the FASTA sequences associated with the two categories from the dm6 genome using the *BSgenome* package and performed de-novo motif enrichment via *meme* [26]. We sampled 10000 3'UTR sequences (100 bp regions) and used them as control for the motif enrichment. *Meme* was then run with the parameters: *-mod zoops -nmotifs 3 -minw 6 -maxw 30 -dna -revcomp*

Acknowledgements

The authors acknowledge the deep-sequencing unit at MPI-IE for data production. AA and TM acknowledge funding from the German Science Foundation (CRC992 "Medical Epigenetics").

Author Contributions

VB performed the analysis of data with input from GS, developed the icetea bioconductor package, and wrote the manuscript with input from all authors. GS developed the MAPCap protocol with analysis input from VB and performed all the experiments. GS and VB conceived the project with input from AA. TM and AA supervised VB and GS during the project.

Code availability

Icetea is available open source at https://github.com/vivekbhr/icetea. All the data presented in the manuscript has been processed via the cage analysis pipeline available at https://github.com/vivekbhr/cage_pipeline .

Conflict of interest

The authors declare no conflict of interest.

References

1. Reyes A, Huber W. Alternative start and termination sites of transcription drive most transcript isoform differences across human tissues. Nucleic Acids Res. 2018;46:582–92.

2. Kodzius R, Kojima M, Nishiyori H, Nakamura M, Fukuda S, Tagami M, et al. CAGE: cap analysis of gene expression. Nat Methods. 2006;3:211–22.

3. Batut P, Dobin A, Plessy C, Carninci P, Gingeras TR. High-fidelity promoter profiling reveals widespread alternative promoter usage and transposon-driven developmental gene expression. Genome Res. 2013;23:169–80.

4. Salimullah, Sakai M, Plessy C, Carninci P. NanoCAGE: A High-Resolution Technique to Discover and Interrogate Cell Transcriptomes. Cold Spring Harb Protoc. 2011;2011:db.erratum2011_01.

5. Core LJ, Martins AL, Danko CG, Waters CT, Siepel A, Lis JT. Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. Nat Genet. 2014;46:1311–20.

6. Adiconis X, Haber AL, Simmons SK, Levy Moonshine A, Ji Z, Busby MA, et al. Comprehensive comparative analysis of 5'-end RNA-sequencing methods. Nat Methods [Internet]. 2018; Available from: http://dx.doi.org/10.1038/s41592-018-0014-2

7. Aktaş T, Avşar Ilık İ, Maticzka D, Bhardwaj V, Pessoa Rodrigues C, Mittler G, et al. DHX9 suppresses RNA processing defects originating from the Alu invasion of the human genome. Nature. 2017;544:115–9.

8. The modENCODE Consortium, Roy S, Ernst J, Kharchenko PV, Kheradpour P, Negre N, et al. Identification of Functional Elements and Regulatory Circuits by Drosophila modENCODE. Science. American Association for the Advancement of Science; 2010;330:1787–97.

9. Batut P, Gingeras TR. RAMPAGE: Promoter Activity Profiling by Paired-End Sequencing of 5′-Complete cDNAs. Current Protocols in Molecular Biology. John Wiley & Sons, Inc.; 2001.

10. Carninci P, Sandelin A, Lenhard B, Katayama S, Shimokawa K, Ponjavic J, et al. Genome-wide analysis of mammalian promoter architecture and evolution. Nat Genet. 2006;38:626–35.

11. Kawaji H, Lizio M, Itoh M, Kanamori-Katayama M, Kaiho A, Nishiyori-Sueki H, et al. Comparison of CAGE and RNA-seq transcriptome profiling using clonally amplified and single-molecule next-generation sequencing. Genome Res. 2014;24:708–17.

12. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based

analysis of ChIP-Seq (MACS). Genome Biol. 2008;9:R137.

13. Lun ATL, Smyth GK. csaw: a Bioconductor package for differential binding analysis of ChIP-seq data using sliding windows. Nucleic Acids Res. 2016;44:e45.

14. Lenhard B, Sandelin A, Carninci P. Metazoan promoters: emerging characteristics and insights into transcriptional regulation. Nat Rev Genet. 2012;13:233–45.

15. Huang F, He J, Zhang Y, Guo Y. Synthesis of biotin-AMP conjugate for 5' biotin labeling of RNA through one-step in vitro transcription. Nat Protoc. 2008;3:1848–61.

16. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30:2114–20.

17. Liao Y, Smyth GK, Shi W. The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. Nucleic Acids Res. 2013;41:e108.

18. Ramírez F, Ryan DP, Grüning B, Bhardwaj V, Kilpert F, Richter AS, et al. deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic Acids Res. 2016;44:W160–5.

19. Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics. 2016;32:3047–8.

20. Ramírez F, Bhardwaj V, Arrigoni L, Lam KC, Grüning BA, Villaveces J, et al. High-resolution TADs reveal DNA sequences underlying genome organization in flies. Nat Commun. 2018;9:189.

21. Frith MC, Valen E, Krogh A, Hayashizaki Y, Carninci P, Sandelin A. A code for transcription initiation in mammalian genomes. Genome Res. 2008;18:1–12.

22. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods. 2017;14:417–9.

23. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26:139–40.

24. Kvon EZ, Kazmar T, Stampfel G, Yáñez-Cuna JO, Pagani M, Schernhuber K, et al. Genome-scale functional characterization of Drosophila developmental enhancers in vivo. Nature. 2014;512:91–5.

25. Yu G, Wang L-G, Han Y, He Q-Y. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS. 2012;16:284–7.

26. Bailey TL, Elkan C, Others. Fitting a mixture model by expectation maximization to discover motifs in bipolymers. Department of Computer Science and Engineering, University of California, San Diego; 1994; Available from: http://www.cs.toronto.edu/~brudno/csc2417_15/10.1.1.121.7056.pdf

Supplementary Figures

Figure S1

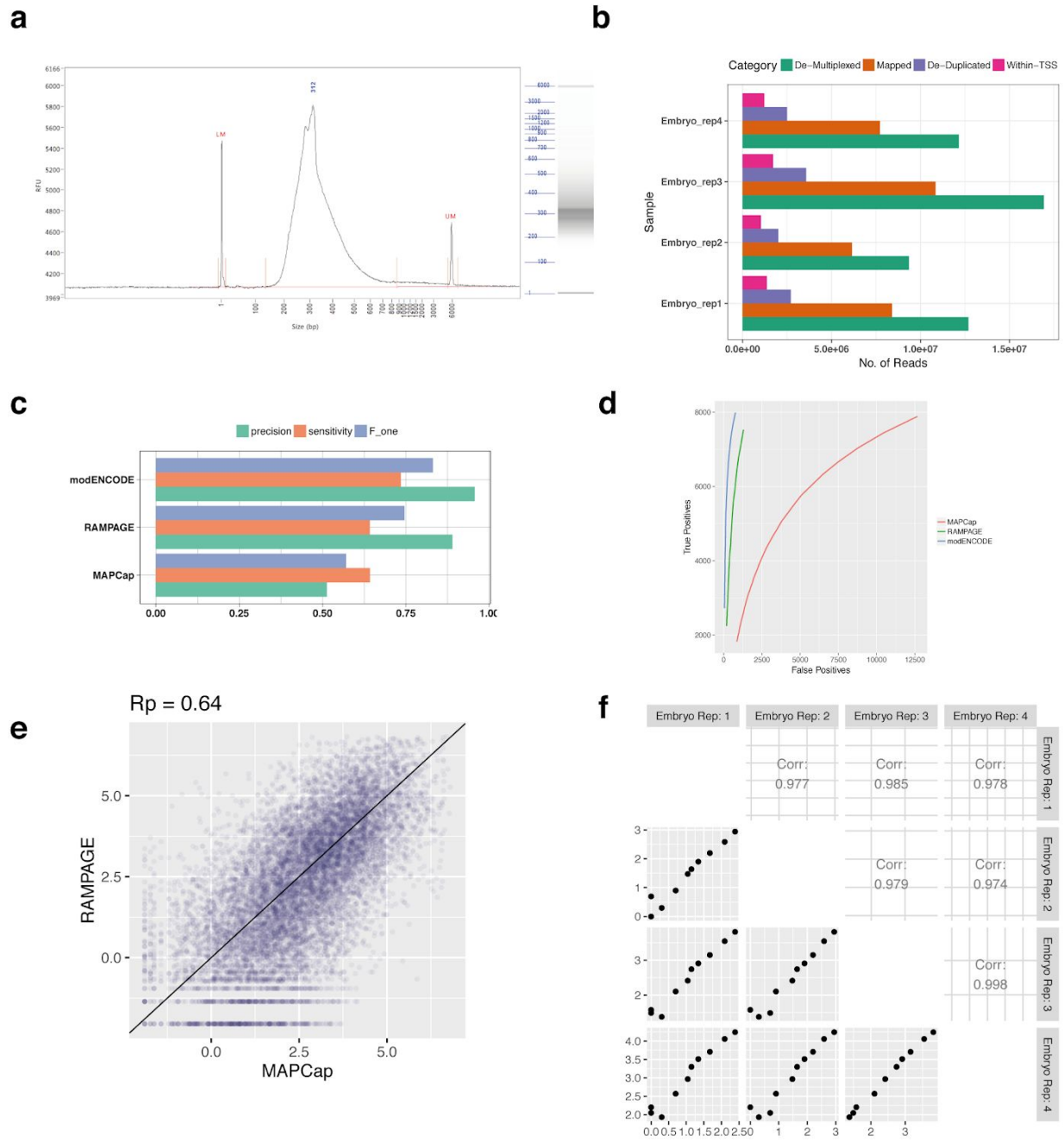

Fig. S1. Assessment of the MAPCap data quality. A. *BioAnalyzer profile of the MAPCap library.* ***B.*** *Number of reads kept at each step of MAPCap analysis, for the embryo samples.* ***C.***

Precision, Sensitivity and F1-score analysis of MAPCap and other protocols. ***D.*** *ROC curve of the MAPCap protocol compared to other protocols. TSS overlapping with those present in dm6 annotation and also expressed in modENCODE RNAseq data were considered true positives for these analysis.* ***E.*** *Same as Fig. 1C; correlation of depth-normalized read counts for MAPCap and RAMPAGE.* ***F.*** *Correlation of recovered counts of individual ERCC oligos between samples. Oligo mix was created after 2-fold serial dilution of individual oligos, which is also reflected in the data.*

Figure S2

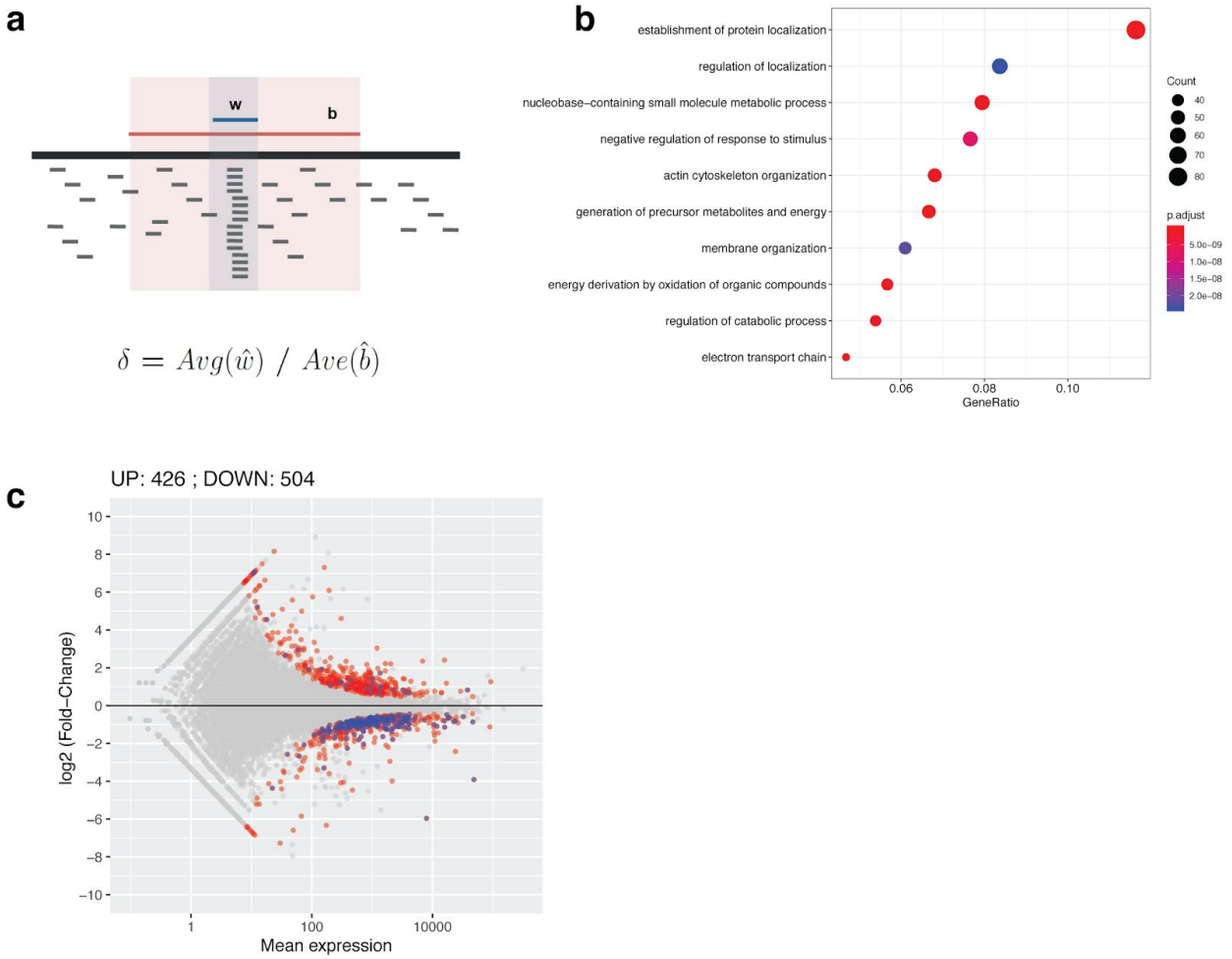

Fig S2. Evaluation of the new TSS detection and differential expression approach. A. *Schematic diagram of "local enrichment" method, the fold-change (**δ**) for windows (w) over background (b) is calculated as average fold change of replicates after depth-normalization.* ***B.*** *GO enrichment of "broad" (>100bp) TSS detected using our method shows enrichment of basic cellular functions.* ***C.*** *MA plot of gene-level differential expression estimates from MAPCap data using internal normalization.*

Figure S3

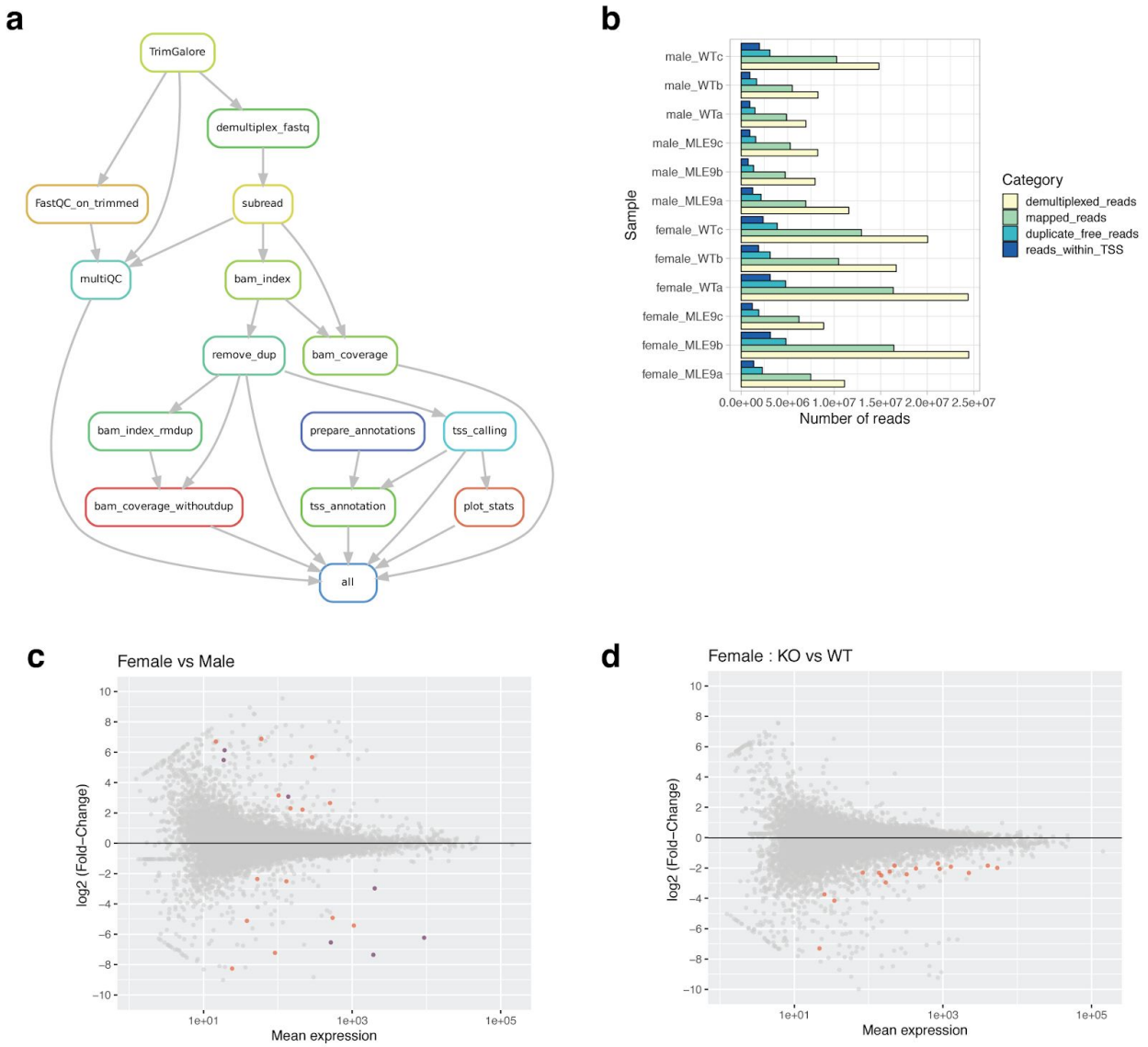

Fig S3. Analysis of dosage compensation defects in MLE KO flies. A. *Workflow used for the analysis of data. It involves read trimming, demultiplexing using sample barcodes, mapping, duplicate removal (using random barcodes), TSS detection, TSS annotation and generation of coverage files and plots (see methods).* ***B.*** *Number of reads kept at each step of the analysis.* ***C-D.*** *MA plot of differentially used promoters between wild-type male and female brains, and between MLE KO female brain over wild-type.*

Supplementary table 1. List of oligos used for abundant RNA depletion :

CCATAAGGCCGAGAAGCGAT
CCTCTACGCCAGGTAAGTAT
TATCGCCTCTGCGCAAAGAT
TATTGCCACTGCGCAAAGAT
TGATGATCCCCGACACTCGA
CAGTCTACCTCTACTAATGA
TGAAGCGGCGATCGAGACAT
ATCCTGTGAAGTATAGTCTT
AATTGAAGAGAAACCAGAGT
AGAGAATAAAAATTTTCAAT
ACCCAATCGTCACCTCTCGCA
CCAGGACGAGCACCCTTTTT
CTCCCCAAGACAAGGAAGGT
TACTCATTAGTTTGAGGCAC
TTCATCATATCATCTAGAGA
TGTTCTGCCGAAGCAAGAAC
GCTCTCCTTCCAAACAACAC
ATGTAATGTTCATCATGTCG