nature research

Corresponding author(s): Joshua Z. Levin

☐ Initial submission ☐ Revised version ☒ Final submission

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

Experimental design

1. Sample size

Describe how sample size was determined.

For the 5' end RNA-Seq method comparison, we intentionally used only one RNA sample to control for that as a variable. For the brain-related samples, we chose a range of samples that were relevant with only one sample of each type because of limited availability of samples and the exploratory nature of these experiments. No statistical analysis was done in advance to determine the number of libraries for each method or brain-related samples. Replicate libraries were prepared and sequenced for CAGE, RAMPAGE, and STRT and show that these methods are reproducible.
We sequenced the "main" libraries to a depth of at least 20 million reads or until there was no material left to sequence -- for more details see Supplementary Table 1 and the Methods section.

2. Data exclusions

Describe any data exclusions.

No data was specifically excluded, but filtering was performed as described in the Methods section using Paraclu peak calling and CapFilter to exclude peaks not likely to be derived from true 5' ends.

3. Replication

Describe whether the experimental findings were reliably reproduced.

Replicate libraries were constructed for CAGE, RAMPAGE, and STRT with K-562 for method comparison -- reproducibility was acceptable, see Supplementary Figures 9 and 10. Replicates were not done for the brain-related samples due to limited material, but our key findings were corroborated by analysis of published FANTOM5 data. All attempts at replication were successful.

4. Randomization

Describe how samples/organisms/participants were allocated into experimental groups.

For method comparison, we used the same commercial source for K-562 RNA and document the lot numbers used in Supplementary Table 1. For the brain-related samples, all libraries were prepared and sequenced in the same batch. Samples were not randomized for the experiments.

5. Blinding

Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

Blinding was not relevant to our study. We used computational analysis methods that were intended to be unbiased in their evaluation of methods. We did not use any animals. There were no human participants to randomize -- only commercially available samples.

Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.

6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
☐	☒ A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ A statement indicating how many times each experiment was replicated
☐	☒ The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section)
☐	☒ A description of any assumptions or corrections, such as an adjustment for multiple comparisons
☐	☒ The test results (e.g. *P* values) given as exact values whenever possible and with confidence intervals noted
☐	☒ A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
☐	☒ Clearly defined error bars

See the web collection on statistics for biologists for further resources and guidance.

Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.	Custom code was written in R (version 3.3) and Python (version 2.7) for specific functions as described in the Methods section. This custom code is available on Github at https://github.com/seanken/FivePrime. We also used Capfilter (a modified version is included on our Github), the Rampage peak calling pipeline, the Paraclu peak calling pipeline, and the Strand Invasion filter (see the associated publications and lab websites for code). In addition, we used STAR (version 2.4.2a), RSEM (version 1.2.7), Bedtools (version 2.20.1 for the peak calling pipeline, version 2.26.0 elsewhere), Samtools (version 1.3.1), Tophat (version 1.4.1), Cufflinks (version 2.2.1), and Picard Tools (version 2.16.0).

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* guidance for providing algorithms and software for publication provides further information on this topic.

Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.	Samples used in this study were from commercial sources (Supplementary Table 1) or generated by the authors from human embryonic stem cells that are available from the Harvard Stem Cell Institute and the Wisconsin International Stem Cell (WISC) Bank. Reagents were purchased from commercial sources (Methods). The authors are not aware of any restrictions, though availability of specific samples from commercial sources needs to be checked directly. The Tobacco Acid Pyrophosphatase (TAP) enzyme used in the oligo capping protocol no longer seems to be available from Epicentre (Illumina), as mentioned in the Results section.

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).	No antibodies were used in this study.

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.

K-562 total RNA: Thermo Fisher Scientific.
HUES66: Harvard Stem Cell Institute.
H9: Wisconsin International Stem Cell (WISC) Bank.

b. Describe the method of cell line authentication used.

K-562: Authenticated by Thermo Fisher Scientific.
H9: Cells were authenticated by the provider before shipment including karyotyping, and testing for contamination.

c. Report whether the cell lines were tested for mycoplasma contamination.

K-562: Tested by Thermo Fisher Scientific.
HUES66 & H9: tested with Lonza's MycoAlert™ Mycoplasma Detection kit.

d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.

None of these cell lines are listed in this database.

Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide details on animals and/or animal-derived materials used in the study.

No animals were used in this study.

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.

The cell line K-562 was established from the pleural effusion of a 53-year-old female with chronic myelogenous leukemia in terminal blast crises. (Lozzio CB, Lozzio BB. Blood 45: 321-334, 1975).
HUES66 is a female cell line isolated at the blastocyst stage (Chen A.E., et al. Cell Stem Cell 4:103-106, 2009).
H9 is a female cell line isolated at the blastocyst stage (Thomson, J.A., et al. Science 282:1145-1147, 1998).
No information on donors for the other brain-related samples is available as these were obtained from commercial sources (Sciencell and BioChain).
Information on genotypes, diagnoses, treatments, etc. was not available or relevant to this study.