Detection of splice isoforms and rare intermediates using multiplexed primer extension sequencing

Hansen Xu1,2, Benjamin J. Fair1,2, Zachary W. Dwyer1, Michael Gildea1 and Jeffrey A. Pleiss1*®

Targeted RNA sequencing (RNA-seq) aims to focus coverage on areas of interest that are inadequately sampled in standard RNA-seq experiments. Here we present multiplexed primer extension sequencing (MPE-seq), an approach for targeted RNA-seq that uses complex pools of reverse-transcription primers to enable sequencing enrichment at user-selected locations across the genome. We targeted hundreds to thousands of pre-mRNA splice junctions and obtained high-precision detection of splice isoforms, including rare pre-mRNA splicing intermediates.

Identification of the small subset of RNA-seq reads that span exon–exon junctions in transcripts has allowed the unambiguous detection of vast numbers of novel splice isoforms in scores of organisms1–3. Yet in spite of the power of this approach, the sequencing depth necessary to quantitatively detect many splicing events is substantially higher than what most experiments generate. Although this limitation of whole-transcriptome profiling has been addressed in part by methods that use antisense probes4 or PCR enrichment5 to target sequencing coverage to genomic regions of interest, a deeper understanding of the basic mechanisms by which splicing is regulated, and of the pathological consequences of its misregulation, will be facilitated by methods that enable the detection of splicing states within cells with higher resolution and greater precision.

Toward this end, we have designed MPE-seq, a targeted sequencing method based on primer extension that improves splice-junction detection and allows for resolution of splicing intermediates. We demonstrate the ability to multiplex hundreds to thousands of primer extension assays and evaluate the products by deep sequencing (Fig. 1a). User-selected primers are extended to generate complementary DNA (cDNA) during a reverse-transcription reaction, thus enabling the user to target RNA regions of interest. The use of elevated temperatures during reverse transcription minimizes nonspecific primer annealing (Supplementary Fig. 1), and each primer is appended with a next-generation sequencing adaptor and a unique molecular identifier6. A strand-extension step similar to template switching7 appends the second sequencing adaptor onto the 3′ terminus of each cDNA molecule. Coupling this approach with paired-end sequencing allows for simultaneous querying of the 5′ and 3′ ends of the cDNAs from targeted regions (full details are presented in the Methods).

As an initial demonstration of MPE-seq, we examined pre-mRNA splicing in the budding yeast Saccharomyces cerevisae. For each of the 309 annotated introns in the yeast genome, primers were systematically designed within a 50-n window immediately downstream of the 3′ splice site, ensuring that short extensions would cross splice junctions. Primers were pooled at equimolar concentration, and MPE-seq libraries were generated with total cellular RNA from wild-type yeast and sequenced to a depth of ~5 million reads. As a comparative reference, we generated conventional RNA-seq libraries using poly(A)-selected RNA and sequenced them to ~40 million reads. Whereas the conventional RNA-seq libraries yielded read coverage that comprised full gene bodies across the transcriptome, MPE-seq coverage was focused on the selected genes, precisely targeted to the regions upstream of the designed primers (Fig. 1b). Just over 75% of sequenced fragments from MPE-seq mapped to targeted regions (Supplementary Fig. 2, Supplementary Table 1), resulting on average in >100-fold enrichment in sequencing depth at these regions compared with that obtained with RNA-seq (Fig. 2a, Supplementary Fig. 3). Although the fold enrichment varied on a target-by-target basis, it was similar across transcripts with a wide range of expression levels (Supplementary Fig. 3). From these data, we extrapolate that a standard RNA-seq experiment would require ~500 million sequencing reads to achieve a level of coverage over the targeted regions similar to what these 5 million MPE-seq reads provided. Given the increased read depth achieved over targeted regions with MPE-seq, we asked how well unspliced isoforms were sampled. Measurements of the fraction of unspliced messages from replicate libraries obtained with MPE-seq showed superior internal reproducibility compared with that in the larger, replicate RNA-seq libraries (Fig. 2b), probably reflecting the sampling noise associated with RNA-seq data with reduced sequencing depth over the targeted regions. Moreover, although MPE-seq is not amenable to de novo discovery of novel splicing events across the entire genome, it did allow for the identification of scores of rare, previously unannotated splicing events at the targeted regions (Supplementary Table 2). Nevertheless, although MPE-seq provided increased sensitivity and reproducibility of splicing measurements, estimates of the unspliced fraction determined from MPE-seq in a wild-type strain only modestly correlated with those determined by RNA-seq (Supplementary Fig. 4a,b). Notably, this correlation improved when when we compared these techniques’ measurement of changes in splicing between samples assayed by the same methodology (Supplementary Fig. 4c), presumably reflecting inherent technical biases8 present in one or both approaches that are internally well controlled.

We next sought to determine whether we could detect splicing intermediates with MPE-seq. Primer extension reactions, which can reveal the locations of reverse-transcription stops, have historically been used to map a variety of biological features such as transcription start sites9 and the locations of branch sites within the lariat intermediate (LI) species of the pre-mRNA splicing reaction10.11.
Our approach anticipated the possibility of mapping the 3′ ends of the cDNA molecules, and indeed we found in our MPE-seq libraries that the 3′ ends of many cDNAs accumulated at the transcription start sites, as determined by an orthologous method12 (Supplementary Fig. 6), indicating that reverse transcription generally proceeded to the 5′ terminus of the RNA. We also observed many cDNAs that terminated at or near the annotated branch-point motifs in introns, with decreased read coverage upstream of the motifs, consistent with the inability of reverse transcriptase to read past the branched adenosine in the LI (Fig. 3a,b). This drop in read coverage was not apparent in MPE-seq libraries generated from a strain that harbored a conditional mutation in Prp2, an RNA helicase required for catalysis of the first step of splicing13, thus corroborating that these cDNAs originate from LIs. We note that these LI-derived cDNAs often contained at the 3′ terminus a unique signature of mismatches incorporated by reverse transcriptase at the branched adenosine (Supplementary Fig. 7), which may serve as a tag for de novo identification of branch sites14. The ability of MPE-seq to differentiate between unspliced isoforms allowed us to estimate that ~10% of unspliced pre-mRNAs are of the LI form genome-wide under steady-state conditions (Supplementary Fig. 5b, Supplementary Table 3, Methods), albeit with considerable variation between individual pre-mRNAs (Fig. 3b,c). Although we identified correlations between transcript- and intron-level features and the abundances of these species (Supplementary Fig. 8), none of these correlations held when we considered the abundance of

Fig. 1 | MPE-seq uses complex pools of reverse-transcription primers to target sequencing to regions of interest. a. Outline of the MPE-seq protocol. UMI, unique molecular identifier. b. Genome browser screenshot of a targeted region in MPE-seq (pink) and conventional RNA-seq (purple). The location of a targeting primer is indicated by a green arrow.

Fig. 2 | MPE-seq enrichment allows high-precision measurements of splicing. a. Each point represents the fold enrichment of a target region in MPE-seq compared with values for conventional RNA-seq. In the box plot, the center line represents the 50th percentile, and lower and upper hinges represent the 25th and 75th percentiles, respectively. Whiskers end at the 0th and 100th percentiles. n is the number of quantified intron-retention events, with at least one spliced read and one unspliced read required in both experiments.

Fig. 3 | Multiplexed primer extension (MPE) sequencing of alternative splicing (LI) RNAs. a, Outline of the MPE-seq protocol. UMI, unique molecular identifier. b, Genome browser screenshot of a targeted region in MPE-seq (pink) and conventional RNA-seq (purple). The location of a targeting primer is indicated by a green arrow.
pre-first-step RNA relative to LIs, a metric that we expect would reflect variation between the relative catalytic rates of the first and second steps of splicing. A more complete understanding of the determinants of in vivo splicing efficiency will require kinetic measurements of the individual steps of splicing, rather than the steady-state levels measured here. The ability of MPE-seq to robustly distinguish these splice isoforms provides an opportunity to do just this.

In our initial experiments we used individually synthesized oligonucleotides as primers; we next sought to increase the utility of this approach by examining methods that would facilitate an increase in the number of targeted regions. We developed an approach that used pools of primers derived from array-based syntheses of thousands of oligonucleotides (Supplementary Fig. 9a,b). Using this approach, we re-created the 309 previously described S. cerevisiae primers, and generated an additional 3,918 primers that targeted splice junctions in the relatively intron-rich fission yeast Schizosaccharomyces pombe. Genome-wide splicing efficiencies determined from MPE-seq libraries generated with primers from pooled syntheses correlated highly with those in libraries derived from individually synthesized oligos (Supplementary Fig. 9), thus validating the utility of this approach. Moreover, MPE-seq libraries generated with primers derived from pooled synthesis also showed strong enrichment for the targeted regions, with levels on par with what we observed with individually synthesized oligonucleotide primers (Supplementary Figs. 2 and 9c). We observed a modest increase in off-target reads when we used primers from the pooled synthesis, consistent with the decreased sequence fidelity of array-based oligo synthesis15 and the increased capacity of these aberrant oligos to prime reverse transcription at undesirable locations. Additionally, as the fraction of the transcriptome that is targeted becomes larger, the fold enrichment over RNA-seq is naturally expected to decrease. Accordingly, when we used the ~4,000 targeting primers in fission yeast, we achieved a median enrichment of sixfold at targeted regions (Supplementary Table 4). Nevertheless, this enrichment enabled us to detect rare but natural alternative splicing events16 that are poorly sampled with standard RNA-seq library-preparation methods (Supplementary Fig. 9e). Although we see no de facto limitation to the number of unique primer sequences or species that could be used for MPE-seq, with increasing numbers of primers comes increasing potential for their cross-reactivity with undesirable RNA targets, which highlights the importance of specificity and fidelity in primer design and synthesis.

The improved sensitivity of MPE-seq is perhaps best exemplified by our ability to detect the LI products of the pre-mRNA splicing pathway. In contrast to studies using other recently described methods14-18 that have reported large-scale detection of upstream-exon splice intermediates and excised lariats, MPE-seq uniquely detects LIs, not excised lariats, from unfractionated cellular RNA. Moreover, these profiling methods that detect RNAs physically associated with the spliceosome require protein tagging and/or purification steps that necessitate large amounts of starting material, which limits their application. Conversely, MPE-seq can be implemented in virtually any system of interest with a need for only microgram quantities of RNA. Additionally, the ability of MPE-seq to query RNA from a wide variety of sources (e.g., cytoplasmic/nuclear fractionated RNA, polysome-fractionated RNA, poly(A)-selected RNA, metabolically labeled RNA) allows for analysis of the cellular location, translational or polyadenylation status, and turnover rates of splice isoforms and intermediates. Overall, we expect that the sensitivity, precision, and flexibility of this approach will lead to a higher-resolution understanding of the splicing pathway. Likewise, primer extension assays have been used to assay RNA secondary-structure after in vitro19 or in vivo20 chemical probing, and we expect that MPE-seq could be readily adapted to RNA-structure interrogation and other approaches where primer extension assays or targeted RNA sequencing is applicable.

Online content
Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41592-018-0258-x.

Received: 14 July 2018; Accepted: 8 November 2018; Published online: 20 December 2018

References
1. Merkin, J., Russell, C., Chen, P. & Burge, C. B. Science 338, 1593–1599 (2012).
2. Barbosa-Morais, N. L. et al. Science 338, 1587–1593 (2012).
3. Mercer, T. R. et al. Nat. Biotechnol. 30, 99–104 (2011).
4. Mercer, T. R. et al. Nat. Protoc. 9, 989–1009 (2014).
5. Blosquart, T. M. et al. PLoS One 8, e79120 (2013).

Fig. 3 | MPE-seq allows genome-wide profiling of lariat intermediates. a, Meta-intron coverage plot surrounding predicted branch points in a wild-type (Prp2) and step-1 splicing mutant strain (prp2-1). The region between the +10 position downstream of the annotated branch point and the 3’ splice site (3’ss) was rescaled for each intron. b, Heat maps showing the relative coverage at each intron for which lariat intermediate reads were detected. c, Estimates of the relative abundance of each isoform for each targeted intron for which reads were detected (P, pre-first-step RNA; L, lariat intermediate; S, spliced mRNA). In box plots, the center line represents the 50th percentile, and lower and upper hinges represent the 25th and 75th percentiles, respectively. Whiskers end at the 0th and 100th percentiles. n = 141 introns for which we attempted lariat quantification and found at least one spliced read.
6. Kivioja, T. et al. Nat. Methods 9, 72–74 (2011).
7. Zhu, Y. Y., Machleder, E. M., Chenchik, A., Li, R. & Siebert, P. D. Biotecniques 30, 892–897 (2001).
8. Zheng, W., Chung, L. M. & Zhao, H. BMC Bioinformatics 12, 290 (2011).
9. Carey, M. F., Peterson, C. L. & Smale, S. T. Cold Spring Harb. Protoc. 2013, 164–173 (2013).
10. Coombes, C. E. & Boeke, J. D. RNA 11, 323–331 (2005).
11. Padgett, R. A. et al. Proc. Natl Acad. Sci. USA 82, 8349–8353 (1985).
12. Booth, G. T., Wang, I. X., Cheung, V. G. & Lis, J. T. Genome Res. 26, 799–811 (2016).
13. Kim, S. H. & Lin, R. J. Mol. Cell. Biol. 16, 6810–6819 (1996).
14. Chen, W. et al. Cell 173, 1031–1044 (2018).
15. Wan, W., Lu, M., Wang, D., Gao, X. & Hong, J. Sci. Rep. 7, 6119 (2017).
16. Stepanskiw, N., Raghavan, M., Fogarty, E. A., Grimson, A. & Pleiss, J. A. Nucleic Acids Res. 43, 8488–8501 (2015).
17. Nojima, T. et al. Cell 161, 526–540 (2015).
18. Burke, J. E. et al. Cell 173, 1014–1030 (2018).
19. Lucks, J. B. et al. Proc. Natl Acad. Sci. USA 108, 11063–11068 (2011).
20. Rouskin, S., Zubradt, M., Washietl, S., Kellis, M. & Weissman, J. S. Nature 505, 701–705 (2014).

Acknowledgements
We thank members of the J.A.P., H. Kwak, and A. Grimson laboratories, as well as the anonymous reviewers for critical feedback on this work. We thank L. Yao for initial drafts of Fig. 1a and Supplementary Fig. 9a. We thank P. Schweitzer, J. Grenier, and the BRC Genomics Facility at Cornell for outstanding technical support with Illumina sequencing. This work was funded by the American Cancer Society (Research Scholars Grant to J.A.P.) and the NIH (grant R01GM098634 to J.A.P.).

Author contributions
H.X., B.J.F., Z.W.D., M.G., and J.A.P. contributed to research design. H.X., B.J.F., Z.W.D., and M.G. performed research and analyzed data. All authors wrote the paper.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41592-018-0258-x.
Reprints and permissions information is available at www.nature.com/reprints.
Correspondence and requests for materials should be addressed to J.A.P.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.
© The Author(s), under exclusive licence to Springer Nature America, Inc. 2018
Methods

**Strain maintenance and growth conditions.** Unless otherwise indicated, all S. cerevisiae experiments used the wild-type strain BY4741 (MATa, his3Δ1, leu2Δ0, met15Δ0, ura3Δ0). Single colonies were inoculated into liquid YPD media and grown overnight at 30 °C. Overnight cultures were then inoculated into fresh liquid YPD media, with cultures seeded at OD₆₀₀ ~ 0.05. Cells were collected by vacuum filtration once cultures reached OD₆₀₀ ~ 0.7 and were then immediately flash-frozen in liquid nitrogen. Cell pellets were stored at -80 °C. For the temperature-sensitive strain harboring the prp2-1 mutation, we grew cultures as described above, but at 25 °C. Once cultures reached OD₆₀₀ ~ 0.7, an equal volume of fresh 50 °C YPD media was added to shift cells to the non-permissive temperature of 37 °C.

The cultures were then maintained at 37 °C for 15 min before cell collection as described above. All S. pombe experiments used the wild-type strain JP002 (h-; ade6-M210; leu1-32, ura4-D18). Single colonies were inoculated into liquid YES media and grown overnight at 30 °C. Overnight cultures were then inoculated into fresh liquid YES media and extension, samples were held at 55 °C for 1 h and then subjected to heat inactivation at 85 °C for 5 min. The remaining RNA was then reverse transcribed with the addition of 0.3 M NaCl and incubation at 65 °C for 15 min. After neutralization with half the original volume of 0.3 M HCl, the cDNA was purified on a Zymo-5 column with a 7x volume of binding buffer (2 M guanidinium HCl, 75% isopropanol). Purified cDNA samples were dried to completion in a SpeedVac. For S. pombe libraries, RNA was isolated as described above. Polyadenylated RNA was then isolated from 60 μg of total RNA with the NEBNext Poly(A) Oligo (dT) Magnetic Isolation Module. RNA was then fragmented to an average size of 200 nt by incubation in 10 mM ZnCl₂, 10 mM Tris-HCl (pH 7.0) for 10 min at 65 °C. The reaction was then quenched by the addition of EDTA (pH 8.0) to a final concentration of 50 mM. The cDNA synthesis reactions were performed as above, with some modifications. For reasons described below, 4 μl of SuperScript III (Thermo Fisher) was used along with the manufacturer-supplied 5x buffer. For primer annealing and extension, samples were held at 55 °C for 1 h and then subjected to heat inactivation at 85 °C for 5 min.

**NSH-ester biotin coupling.** Dried cDNA samples were resuspended in 18 μl of fresh 0.1 M sodium bicarbonate (pH 9.0) to which 2 μl of 0.1 mg/ml NSH-biotin (Thermo Fisher; 20217) was added. Reactions were incubated at 65 °C for 1 h, after which biotin-coupled cDNA was purified from unreacted NSH-biotin on Zymo-5 columns with a 7x volume of binding buffer (2 M guanidinium HCl, 75% isopropanol).

**Streptavidin–biotin purification.** 20 μl of Dynabeads MyOne streptavidin C1 (Thermo Fisher; 65602) per sample was prewashed twice in 500 μl of 1x binding and wash buffer (5 mM Tris-HCl (pH 7.5), 0.5 mM EDTA, and 1 μM NaCl) per the manufacturer's protocol. Washed beads were resuspended in 50 μl of 2x binding and wash buffer per sample, and 50 μl of 2x binding and wash buffer containing 50 μl of purified cDNA sample. Biotin–streptavidin binding was allowed to proceed for 30 min at room temperature with rotation. Bound material was washed twice with 500 μl of 1x binding and wash buffer, and then once with 100 μl of 1x SSC. To ensure purification of only single-stranded cDNAs, beads were then incubated with 0.1 M NaOH for two consecutive room-temperature washes for 10 min and 1 min, respectively. Finally, the bound material was washed three times with 100 μl of 1x TE. The cDNA was eluted from the beads by heating to 90 °C for 2 min in the presence of 100 μl of 95% formamide, 10 mM EDTA. The eluate was then purified on Zymo-5 columns as described above, and the cDNA was eluted from columns in 40 μl of water.

**First-strand extension.** We amplified the reactions to purified cDNA by combining 1 μl of first-strand extension oligo (100 μM of qP789 for S. cerevisiae and qP789 for S. pombe), 5 μl of 10X NEB buffer 2, 40 μl of purified cDNA sample, and 1 μl of 10 mM (each) dNTP mix. Samples were incubated at 65 °C for 5 min and then cooled to room temperature on the benchtop. To each sample we added 3 μl of Klenow exon–fragment (NEB M0212), and then we incubated the reactions for 5 min at room temperature and subsequently 37 °C for 30 min. Samples were then purified with streptavidin beads according to the protocol described above. Samples were concentrated on Zymo-5 columns as described above, and the cDNA was eluted from columns in 40 μl of water.
PCR reaction was then run on a 6% native polyacrylamide gel, and the DNA was resolved by staining with Sybr gold. Libraries were size-selected from 200 bp to 800 bp, and DNA was extracted from gel fragments via passive diffusion overnight in 0.3 M sodium acetate (pH 5.3). Libraries were then ethanol-precipitated and quantified.

cDNA synthesis temperature experiment. Because of the target-specific nature of MPE-seq cDNA synthesis, any reverse-transcription (RT) events at nontarget sites will reduce the fraction of on-target reads. Indeed, these off-target events contribute substantially to the nonspecific class reads in a typical MPE-seq experiment (Fig. 2a). One way to reduce off-target RT events is to increase the specificity of the RT primers. We assessed this by testing the effect of increased temperature during the RT reaction on off-target sequencing reads. MPE-seq libraries were generated via the above-described protocol, with one primary difference: increased reaction temperature required the use of a thermophilic enzyme. For this reason, we used Superscript III (Thermo Fisher) along with the manufacturer-supplied buffer (reaction concentrations: 50 mM Tris–HCl (pH 8.3), 75 mM KCl, 3 mM MgCl2). Primer annealing and reactions were carried out at 47 °C, 51 °C, and 55 °C in replicate.

MPE-seq data analysis. Sequencing and alignment. S. cerevisiae MPE-seq libraries were sequenced on the NextSeq platform by the BRC Genomics Facility at Cornell University with 60-bp (P5) + 15-bp (P7) paired-end chemistry. We removed PCR duplicates from the dataset by filtering out non-unique reads with respect to all base calls in both reads, including the 7-bp UMI. In other words, for each set of identical paired-end reads, a single read pair was retained for analysis. MPE-seq reads were aligned to the yeast genome (reference genome assembly R64-1-1 3-17) with the STAR aligner 28 with the following alignment parameters: --alignEndsType EndToEnd --alignIntronMin 20 --alignIntronMax 1000 --alignMateGapMax 400 --alignSplicedMateMapLmin 56 --alignSplicedMateMapLmax 34 --clip3pAdapterSeq CTGTCTCTTATACACATCTCCGAGCCCACGAGAC --clip5pNbases 70. Alignment files were filtered to exclude read mappings deriving from inserts of less than 30 bases. We believe that these small fragments represent unextended RT primers that were retained in the sequencing libraries. These small fragments can sometimes erroneously map to splice junctions or target introns, even though we believe that they are not derived from cellular RNA.

S. pombe MPE-seq libraries were sequenced on the MiSeq platform by the BRC Genomics Facility at Cornell University using 100-bp (P5) + 50-bp (P7) paired-end chemistry. Reads were trimmed to 60 and 15 bp and processed as described above for Supplementary Fig. 9c, whereas full-length reads were processed as described above for Supplementary Fig. 9e.

S. cerevisiae RNA-seq libraries were sequenced on an Illumina HiSeq 2500 by the BRC Genomics Facility at Cornell University using 100-bp single-end reads. S. pombe RNA-seq data were downloaded from the NCBI BioSample database (accession SRS167019), and read 2 of read pairs was discarded to make read lengths comparable to those in our other libraries. Reads were aligned with the STAR aligner with the following alignment parameters: --alignEndsType EndToEnd --alignIntronMin 20 --alignIntronMax 1000 --alignSJDoverglapMin 1 –outSAM multilineNmax 1 --outFilterMismatchNmax 3 --clip3pAdapterSeq CTGTCTCTTATACACATCTCCGAGCCCACGAGAC --clip5pNbases 70. Alignments were filtered to exclude read mappings deriving from insertions of less than 30 bases. We believe that these small fragments represent unextended RT primers that were retained in the sequencing libraries. These small fragments can sometimes erroneously map to splice junctions or target introns, even though we believe that they are not derived from cellular RNA.

Because of the target-specific nature of MPE-seq cDNA synthesis, any reverse-transcription (RT) events at nontarget sites will reduce the fraction of on-target reads. Indeed, these off-target events contribute substantially to the nonspecific class reads in a typical MPE-seq experiment (Fig. 2a). One way to reduce off-target RT events is to increase the specificity of the RT primers. We assessed this by testing the effect of increased temperature during the RT reaction on off-target sequencing reads. MPE-seq libraries were generated via the above-described protocol, with one primary difference: increased reaction temperature required the use of a thermophilic enzyme. For this reason, we used Superscript III (Thermo Fisher) along with the manufacturer-supplied buffer (reaction concentrations: 50 mM Tris–HCl (pH 8.3), 75 mM KCl, 3 mM MgCl2). Primer annealing and reactions were carried out at 47 °C, 51 °C, and 55 °C in replicate.

We determined the locations of branch points (Supplementary Table 8) by consolidating the most used branch point from lariat sequencing data 28 and previously described branch locations based on sequence motif searches 29.

Heat maps and meta-gene plots. To generate meta-gene plots, which illustrate read coverage around features of interest, we used the deepTools ComputeMatrix command 30 in conjunction with a BigWig coverage file of the 3′ terminating bases and a bedfile containing transcription start site positions as determined by PRO-capt 24 or a bedfile containing the annotated branch-point regions detailed above. Importantly, this bedfile was filtered to include only branch-point regions that would produce an L1 within the size range captured by library size-selection of MPE-seq libraries (see column "Attempted.ariat.Quantity?" in Supplementary Table 3).

RNA-seq experiments. Library prep. For each RNA-seq library, 1 μg of total RNA was input into the NEBNext Ultra Directional RNA library prep kit (Illumina). Libraries were prepared according to the manufacturer’s protocol.

Estimating splice isoform abundances from RNA-seq data. Similarly to MPE-seq data, spliced reads from target introns were counted with the $\text{S}\text{.}\text{out}\text{-}\text{tab}$ file created by the aligner. Unspliced reads were counted with the bedtools software package 28, which counted the number of reads that overlapped an intron. Spliced and unspliced read counts for each intron were then length-normalized for the feature’s potential mapping space. The potential mapping space for a spliced read is equal to 2 × the read length minus the minimum splice-junction overhang length. The potential mapping space for an unspliced read is equal to 2 × the read length minus the minimum splice-junction overhang length plus the length of the intron. Read counts assigned to each feature were then divided by the length. The unspliced fraction was calculated for each intron as the quotient of length-normalized unspliced reads and spliced reads.

Gene expression normalization. Relative transcript expression was calculated from RNA-seq data via transcripts per million (TPM) normalization 30, considering only exonic reads and exonic gene lengths. For S. cerevisiae MPE-seq data, we calculated a similar TPM metric by summing the reads per gene and dividing by the number of library mapped reads. Given that a single RNA corresponds to a single primer extension event, and because nearly all targeted transcripts have only a single targeting primer, normalization by gene length was not done in this calculation of TPM.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Code availability. Code for basic analysis steps is available at https://github.com/bfairkun/SceSpliceSeq.

Data availability. All sequencing data are available through NCBI’s Sequence Read Archive (SRA) under accession number SRP148810.
Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

- The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- The statistical test(s) used AND whether they are one- or two-sided
- Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- A description of all covariates tested
- A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- For null hypothesis testing, the test statistic (e.g. F, t, r) with confidence intervals, effect sizes, degrees of freedom and P value noted
  Give P values as exact values whenever suitable.
- For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- Clearly defined error bars
  State explicitly what error bars represent (e.g. SD, SE, CI)

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

| Data collection | no software was used |
|-----------------|----------------------|
| Data analysis   | STAR_2.5.2b, bedtools v2.26.0, deepTools 2.5.4. Use of that software is in shell scripts available at https://github.com/bfairkun/ScerSpliceSeq |

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:
- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All sequencing data are available through NCBI’s Sequence Read Archive (SRA) at accession number SRP148810.
Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

- Life sciences
- Behavioural & social sciences
- Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

| Sample size       | No sample size calculation was performed a priori. Independently grown yeast cultures were used as replicates (n=2) and compared to RNA-seq (n=2) to demonstrate reproducibility of the novel method. |
|-------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Data exclusions   | Some introns were removed from branchpoint or lariat intermediate analysis due to the predicted size of those fragments being smaller than the range for which we size selected sequencing libraries. The introns included in analysis are listed in Supplemental Data. |
| Replication       | We have repeatedly used (>10 times) this technique with similar results using budding yeast total RNA with individually synthesized oligos in our lab. We have only once used this technique with array-synthesized oligos as described and presented in the manuscript using fission yeast poly-A selected RNA, or budding yeast total RNA. |
| Randomization     | Not relevant, we were not testing any variables that had to be controlled for by randomization.                                                                                                          |
| Blinding          | Investigators were not blinded to any sample identities during data acquisition or processing.                                                                                                           |

Reporting for specific materials, systems and methods

| Materials & experimental systems | Methods |
|----------------------------------|---------|
| n/a                              | n/a     |
| involved in the study            | involved in the study |
| xxx unique biological materials  | ⊗        |
| antibodies                       | ⊗        |
| eukaryotic cell lines            | ⊗        |
| palaeontology                    | ⊗        |
| animals and other organisms      | ⊗        |
| human research participants      | ⊗        |
| chip-seq                         | ⊗        |
| flow cytometry                   | ⊗        |
| mri-based neuroimaging           | ⊗        |

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)  prp2-1 strain, Pleiss lab

Authentication none of the cell lines were authenticated. though results were as expected for this unique line

Mycoplasma contamination not tested for mycoplasma contamination

Commonly misidentified lines (See ICCLAC register) No commonly misidentified lines were used