Multiplexed Primer Extension Sequencing Enables High Precision Detection of Rare Splice Isoforms

Hansen Xu[+], Benjamin J. Fair[+], Zach Dwyer, Michael Gildea, and Jeffrey A. Pleiss*

Department of Molecular Biology and Genetics, Cornell University, Ithaca, NY 14853

[+] - equal contribution

[*] - correspondence: jpleiss@cornell.edu

ABSTRACT

Targeted RNA-sequencing aims to focus coverage on areas of interest that are inadequately sampled in standard RNA-sequencing experiments. Here we present a novel approach for targeted RNA-sequencing that uses complex pools of reverse transcription primers to enable sequencing enrichment at user-selected locations across the genome. We demonstrate this approach by targeting pre-mRNA splice junctions in *S. cerevisiae*, revealing high-precision detection of splice isoforms, including rare pre-mRNA splicing intermediates.

Main text

RNA sequencing (RNA-Seq) has greatly expanded our understanding of the variety of splice isoforms that can be generated within a cell. Identification of the small subset of reads that span exon-exon junctions within transcripts has enabled the unambiguous detection of vast numbers of novel splice isoforms in scores of organisms[1,2]. Yet in spite of the power presented by this approach, the sequencing depth necessary to quantitatively detect many splicing events is significantly higher than most experiments generate. While this limitation of whole-transcriptome profiling has been addressed in part by methods that utilize antisense probes[3,4] or PCR enrichment[5] to target sequencing coverage to genomic regions of interest, a deeper understanding of the basic mechanisms by which splicing is regulated, and the pathological consequences of its mis-regulation, will be facilitated by methods that enable higher resolution and precise detection of splicing states within cells. Towards this end, we have designed and implemented a novel targeted sequencing method that enhances splice junction detection and allows for genome-wide resolution of splicing intermediates. Building upon the historically validated use of primer extension as a tool for assessing splicing status, we demonstrate the ability to multiplex primer extension assays and evaluate the products by deep sequencing, an approach we hereafter refer to as Multiplexed Primer Extension sequencing, or MPE-seq (Fig1A).

The method we developed harbors two straightforward yet key features. First, user-selected primers are used to generate complementary DNA (cDNA) during a reverse transcription reaction, enabling targeting of RNA regions of interest. Each primer is appended with a next-generation sequencing adapter, as well as a unique molecular identifier (UMI) to alleviate artifacts associated with PCR amplification during library preparation[6]. The use of

elevated temperatures during the reverse transcription reaction minimizes non-specific primer annealing (FigS1), while the inclusion of derivatized nucleotides allows for the efficient purification of extended products and removal of excess primers. Secondly, a strand-extension step similar to template-switching[7] is used to append the second sequencing adapter onto the 3' terminus of the cDNA molecules. Coupling this approach with paired-end sequencing allows for the simultaneous querying of the 5' and 3' ends of the cDNAs from targeted regions (see Methods for full details).

As an initial demonstration of MPE-seq, we examined pre-mRNA splicing in the budding yeast *Saccharomyces cerevisiae*. For each of the 309 annotated introns in the yeast genome, primers were systematically designed within a 50nt window immediately downstream of the 3' splice site, ensuring that short extensions would be sufficient to cross the splice junctions. Primers were pooled at equimolar concentration and MPE-seq libraries were generated using total cellular RNA from wildtype yeast and sequenced to a depth of only ~5 million reads. As a comparative reference, standard RNA-seq libraries were generated using poly-A selected RNA and sequenced to ~40 million reads. Whereas the standard 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 (Fig1B). Just over 75% of sequenced fragments from MPE-seq mapped to targeted regions (Fig2A, SupplementalInformation_Table1), resulting on average in a greater than 100-fold enrichment in sequencing depth at these regions when compared with RNA-seq (Fig2B, FigS2). The fold-enrichment was similar across transcripts with a wide range of expression levels (FigS2), and from these data we extrapolate that a standard RNA-seq experiment would require

~500 million sequencing reads to achieve a similar level of coverage over the targeted regions as these 5 million MPE-seq reads provided.

Given the increased read depth achieved over targeted regions using MPE-seq, we asked how well rare splice events were sampled. Because each primer extension event corresponded to a single RNA molecule, determining relative isoform expression was simplified because read counts did not need to be adjusted for mapping space as in standard analyses of RNA-seq data (see Methods). Importantly, the levels of intron retention determined from replicate libraries using MPE-seq showed superior internal reproducibility compared with the larger, replicate RNA-seq libraries (Fig2C), likely reflecting the sampling noise associated with RNA-seq data with reduced sequencing depth over the targeted regions. Moreover, while 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 (SupplementalInformation_Table3), consistent with the significantly increased sensitivity of this approach. Nevertheless, while MPE-seq provided increased sensitivity and reproducibility of splicing measurements, the intron retention levels determined from MPE-seq in a wildtype strain only modestly correlated with those determined by RNA-seq (FigS3A, FigS3B). Notably, this correlation improved when comparing how these techniques measured *changes* in splicing between samples assayed by the same methodology (FigS3C), presumably reflecting inherent biases[8] (in fragmentation, ligation, PCR amplification, library size selection, etc.) present in one or both approaches that are internally well controlled.

With the increased resolution provided by this approach, we sought to determine whether we could detect splicing intermediates using MPE-seq. By identifying the locations of reverse transcription stops, primer extension reactions have historically been used to map a variety of

biological features, including among others transcription start sites (TSSs)[9], and the locations of branch sites within the lariat intermediate species of the pre-mRNA splicing reaction[10,11] (Fig3A). Because the approach we developed anticipated the possibility of mapping the 3' ends of the cDNA molecules, we examined the locations of those generated by MPE-seq. As expected, the 3' ends of many cDNAs accumulated at the TSSs as determined by an orthologous method[12] (FigS4), indicating that reverse transcription generally proceeded to the 5' terminus of the RNA. Importantly, we also observed many cDNAs which terminated at or near the annotated branchpoint motifs within introns, with decreased read coverage upstream of the motifs, consistent with the inability of reverse transcriptase to read past the branched adenosine in the lariat intermediate species (Fig3B, Fig3C). This drop in read coverage was not apparent in MPE-seq libraries generated from a strain harboring a conditional mutation in Prp2, an RNA helicase required for catalyzing the 1st step of splicing[13], corroborating that many of these cDNAs originate from lariat intermediates. We noted that these lariat-intermediate derived cDNAs contain a unique signature of mismatches incorporated by reverse transcriptase at the branched adenosine (FigS5), which may serve as a unique tag for *de novo* identification of branch sites in organisms with less well annotated branch sites.

The ability of MPE-seq to differentiate between unspliced isoforms in unfractionated cellular RNA provides a unique opportunity to investigate the relative efficiencies with which transcripts undergo the 1st and 2nd chemical steps in the splicing pathway. Our data revealed that ~10% of unspliced pre-mRNAs at steady state conditions were present in the lariat intermediate form genome-wide (see FigS6, methods), albeit with significant variation between individual pre-mRNAs (Fig3D). Remarkably, a strong correlation was observed between the relative levels of pre-mRNA and lariat intermediate species for a given transcript (FigS7A). Correlations were

also observed between splice intermediate levels and transcript expression level, branch motif strength, and host gene function (FigS7B, FigS7C); however no correlation was seen when considering the ratio of pre-1st step intermediate to lariat intermediate, a metric which we expect to reflect the relative catalysis rates of the 1st and 2nd step of splicing. A complete understanding of the determinants of *in vivo* splicing efficiency will require measurements of the kinetics of these individual steps rather than their steady state levels. The ability of MPE-seq to robustly differentiate pre-mRNA isoforms provides a powerful new opportunity to do just this.

Whereas our initial experiments were performed using individually synthesized oligonucleotides as primers, we sought to increase the utility of this approach by examining methods that would facilitate an increase in the number of targeted regions. Many commercial sources exist that allow for the cost-effective, array-based synthesis of pools of thousands of individual oligonucleotide sequences, so we developed an approach for the pooled synthesis of an equivalent set of the previously described primer sequences in order to test the effectiveness with which they could be used in MPE-seq. By appending a common sequence onto the 3' end of the desired primers and then using a protocol that included PCR amplification, restriction digestion, and targeted strand degradation (FigS8A, FigS8B), we readily prepared a sufficient quantity of single-stranded oligonucleotides with which to generate MPE-seq libraries. Importantly, MPE-seq libraries generated using primers synthesized by this approach also showed strong enrichment for the targeted regions, with levels on par with what we observed using individually synthesized oligonucleotide primers (FigS8C). Moreover, genome-wide splicing efficiencies determined from MPE-seq libraries generated using primers from this pooled synthesis were highly correlated with splice efficiencies derived from individually

synthesized oligos (FigS8D), confirming the capacity of this approach to generate large numbers of unique primers for use in MPE-seq experiments.

Our work here demonstrates the capacity of MPE-seq to facilitate examinations of pre-mRNA splicing status in a targeted, cost-effective way that improves the precision and sensitivity of splice isoform detection. The improved sensitivity of this approach is perhaps best exemplified by our ability to detect the lariat intermediate products of the pre-mRNA splicing pathway. Though other methods have reported large-scale detection of upstream exon splicing intermediates[14–16], MPE-seq uniquely detects the lariat intermediate from unfractionated cellular RNA, and does not require laborious purification of complexes from cellular extracts. Our demonstration that oligonucleotides derived from pooled commercial syntheses can be used in this approach expands the types of applications to which MPE-seq could be applied in a cost-effective manner. While we see no *de facto* limitation to the species or number of unique primer sequences that could be used for MPE-seq, with increasing numbers of primers comes an increasing potential for their cross-reactivity with undesirable RNA targets, highlighting the importance of specificity in primer design. Similarly, the level of enrichment provided by this approach would vary as a function not only of the number of regions being targeted, but also of the distribution of the expression levels of those targets. We expect that the improved sensitivity, precision and flexibility of this approach will enable a higher-resolution understanding of the pre-mRNA splicing pathway. Likewise, primer extension assays have also been used to assay RNA secondary structure after *in vitro*[17] or *in vivo*[18] 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 Methods

Strain Maintenance and Growth Conditions. Unless otherwise indicated, all experiments used the wild type (WT) *S. cerevisiae* strain BY4741 (*MATa*, *his2Δ1*, *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, seeding cultures at OD_{600} ~0.05. Cells were collected by vacuum filtration once cultures reached OD_{600} ~0.7 immediately followed by flash freezing in liquid nitrogen. Cell pellets were then stored at -80°C. For the temperature sensitive mutant *prp2-1*[19] we grew cultures as described above except at 25°C. Once cultures reached OD_{600} ~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 minutes before cell collection as described above.

MPE-seq Primer and Oligo Design.

Gene specific reverse transcription primer design. For each of the 309 annotated spliceosomal introns within the budding yeast genome (annotations obtained from UCSC SacCer3), a reverse transcription primer was designed within the first 50 nucleotides downstream of the intron. Targeting to this region ensured that short-read sequencing of the products generated from reverse transcription with these primers would cross the upstream exon-exon or exon-intron boundaries, enabling determination of the splicing status. Primers were designed using OligoWiz, a program initially developed for microarray probe design, but which enables the selection of primer sequences optimized for target specificity relative to a designated genomic background[20]**.** We used the standalone version of OligoWiz with default parameters for short (24-26bp) oligo design to obtain optimal sequences within each 50bp window. To the 5' end of each of these sequences was appended two additional sequence elements: a random 7-nucleotide

unique molecular index (UMI) which allows for the detection and removal of amplification artifacts arising from library preparation[6]; and the P5XX region of the Illumina sequencing primer to enable the sequencing of the reverse transcription products. Each of these primers was individually synthesized by Integrated DNA Technologies (IDT), the full sequences of which are provided in (Supplemental Information Table 5).

Complex oligo mix amplification method: The above described oligo primers were batch synthesized as a pool on a OligoMix microarray by LC Sciences. These oligos are synthesized at vastly lower quantities than is required for cDNA synthesis in MPE-seq. To generate a quantity of primer pool that is sufficient, PCR amplification, along with several processing steps were used (FigS8A). This was enabled by addition of two key sequence elements appended onto the 3' end of the individually synthesized oligo primers detailed above. From the 5' to 3' direction: 1.) a SapI restriction site, and 2.) a PCR amplification sequence (Supplemental information Table 5). First, the oligos were amplified in a standard PCR using Phusion polymerase with 14 amplification cycles. This 400 µL PCR reaction contained: 1% of the pooled oligonucleotides from LCSciences as a template, a forward primer (oHX093) containing a C3 spacer at its 5' end, and a reverse amplification primer (oHX094) containing a biotin-label at its 5' end (see Supplemental Information Table 6). Cycling conditions were as follows: denaturation at 95°C for 10 sec; annealing at 60°C for 20 sec; and extension at 72°C for 30 sec. Upon completion of this initial reaction, the entire reaction was used as a template to seed a larger (40 mL) PCR reaction. For efficient amplification, this large reaction was performed in four 96-well plates with 100µL in each well. Reaction conditions were identical to those described for the first reaction, and a total of 15 cycles were performed for this second amplification. Reactions were purified and concentrated by isopropanol precipitation. To generate single stranded primers for use in MPE-

seq, the double-stranded amplicons were first digested using SapI (NEB R0569) in a 150 µL reaction containing 30 µL of enzyme. The reaction was incubated at 37°C overnight, after which the reaction products were concentrated by ethanol precipitation. Next, the 5' to 3' lambda exonuclease (NEB M0262) was used to preferentially degrade the two strands containing unmodified 5' ends. This reaction was performed at 37°C for 2 hours according to the manufacturer's protocol. The products of this reaction were then purified using Zymo columns using 7X volume binding buffer (2 M guanidinium-HCl, 75% isopropanol). After this step, the remaining DNA consisted of the desired single stranded RT primer, and an undesired single stranded section containing the SapI site plus the amplification primer. Making use of the 5' biotin tag on the amplification primer, these undesired oligos were removed by affinity capture with streptavidin beads. Specifically, this was accomplished by using 50 µL of Dynabeads MyOne Streptavidin C1 according to the manufacturer's protocol. The unbound supernatant fraction was retained as it contains the desired products. The recovered material was precipitated and verified using 6% native PAGE stained with SyBr Gold.

1st Strand extension template oligo design. The oligos were designed with three key features from the 5' to 3' end of the oligo: 1.) A portion of the Nextera P7XX sequencing adapter. Of the entirety of the P7XX adapter, the region 3' of the I7 barcode was used. This allowed for barcoding and amplification of the sequencing libraries. 2.) a dN9 anchor on the 3' end to randomly anneal to cDNA products. 3.) A 3' carbon block modification. This was done so that the oligo may only be used as a template to append the Nextera sequencing adapter to the end of cDNAs, rather than as a primer. The full sequence of this primer can be found in (Supplemental Information Table 6).

MPE-Seq Library Prep

cDNA synthesis. RNA was isolated following a hot acid phenol extraction protocol[21]. A total of 10ug of total RNA was used to generate each library. cDNA was synthesized by mixing 1 µg of the gene specific primer pool described above with each RNA sample in 50 mM Tris-HCl (pH 8.5), 75 mM KCl in a 20 µL volume. The primers were then annealed in a thermocycler with the following cycle; 70°C for 1 minute, 65°C for 5 minutes, hold at 47°C. An equivalent volume of MMLV reverse transcriptase enzyme mix containing 1 mM dATP, 1 mM dGTP, 1 mM dCTP, 0.4 mM aminoallyl-dUTP, 0.6 mM dTTP, 50 mM Tris-HCl (pH 8.5), 150 mM KCl, 6 mM $MgCl_2$, 10 mM DTT was pre-heated to 47°C and added to the primer-annealed RNA mix resulting in a total reaction volume of 40uL. Maintaining the samples at 47°C was essential for reducing off-target cDNA synthesis. Reactions were incubated at 47°C for 3 hours, followed by heat inactivation at 85°C for 5 minutes. Remaining RNA was hydrolyzed by addition of ½ volume of 0.3 M NaOH, 0.03 M EDTA and incubated at 65°C for 15 minutes. After neutralization with ½ (original) volume of 0.3 M HCl, the cDNA was purified with a Zymo-5 column using 7X volume of binding buffer (2 M guanidinium HCl, 75% isopropanol). Purified cDNA samples were dried in a SpeedVac until all liquid had evaporated.

NHS 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/µL NHS-biotin (ThermoFisher 20217) was added. Reactions were incubated at 65°C for 1 hour followed by purification of biotin coupled cDNA from unreacted NHS-biotin by using Zymo-5 columns using 7X volume of binding buffer (2 M guanidinium HCl, 75% isopropanol).

Streptavidin-biotin purification. 20 µL of Dynabeads MyOne Streptavidin C1 (ThermoFisher 65602) per sample were pre-washed twice in 500 µL of 1X bind and wash buffer (5 mM Tris-HCl (pH 7.5), 0.5 mM EDTA and 1 M NaCl) as per manufacturer's protocol. Washed beads were

resuspended in 50 µL of 2X bind and wash buffer per sample and 50 µL was combined with each 50 µL purified cDNA sample. Biotin-streptavidin binding was allowed to proceed for 30 minutes at room temperature with rotation. Bound material was washed twice with 500 µL of 1X bind and wash buffer, followed by an additional wash 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 minutes and 1 minute, respectively. Finally, the bound material was washed 3 times with 100 µL 1X TE. cDNA was eluted from the beads by heating samples to 90°C for 2 minutes in the presence of 100 µL of 95% formamide, 10 mM EDTA. The eluate was then purified using Zymo-5 columns as described above. cDNA was eluted from columns in 40 µL H_20.

First strand extension. Primers were annealed to purified cDNA by combining: 1uL 1st strand extension oligo (100 µM), 5 µL 10x NEB buffer 2, 40 µL purified cDNA sample, and 1 µL of 10 mM (each) dNTP mix. Samples were then incubated at 65°C for 5 minutes, followed by cooling to room temperature on the bench top. To each sample was added 3 µL of Klenow exo- fragment (NEB M0212) and reactions were incubated for 5 minutes at room temperature, after which they were moved to 37°C for 30 minutes. Samples were subsequently purified via streptavidin beads following the protocol described above. Samples were then purified and concentrated via Zymo-5 columns. Samples were then eluted in 33 µL H_20.

PCR amplification. Amplification of the reaction products was accomplished by using 10 µL of the purified material generated in the 1st strand extension reaction as a template in a PCR reaction. Illumina Nextera (i5) and (i7) indexing primers were used in a standard 50 µL PCR reaction with Phusion polymerase (ThermoFisher F530S). Cycling conditions were as follows: denaturation at 95°C for 10 sec; annealing at 62°C for 20 sec; and extension at 72°C for 30 sec.

Libraries typically required between 14 and 20 cycles of amplification, depending upon the amount of starting material. Libraries were then purified via PAGE. Each 50 μL PCR reaction was run on a 6% native poly-acrylamide gel and DNA was resolved by staining with SyBr gold. Libraries were size selected from 200bp to 800bp 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. Due to the target specific nature of MPE-seq cDNA synthesis, any reverse transcription (RT) events at non-target sites will reduce the fraction of on-target reads. Indeed, these off-target events contribute significantly to the class of nonspecific reads in a typical MPE-seq experiment (Fig2A). One way to reduce off-target RT events is through increasing 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 using the above described protocol with one primary difference: Increased reaction temperatures required the use of a thermostable enzyme. For this reason, Superscript III (ThermoFisher) was used along with the manufacturer supplied buffer (reaction concentration: 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 experiments.

MPE-Seq Data Anlysis

Sequencing and alignment. MPE-seq libraries were sequenced on the NextSeq platform by the BRC Genomics Facility at Cornell university using 60bp (P5) +15bp (P7) paired-end chemistry. PCR duplicates were removed from the dataset by filtering out non-unique reads with respect to all base calls in both reads, including the 7bp 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[22]) using the STAR aligner[23] with the following alignment parameters: {--alignEndsType EndToEnd --alignIntronMin 20 --alignIntronMax 1000 --alignMatesGapMax 400 --alignSplicedMateMapLmin 16 --alignSJDBoverhangMin 1 --outSAMmultNmax 1 --outFilterMismatchNmax 3 --clip3pAdapterSeq CTGTCTCTTATACACATCTCCGAGCCCACGAGAC --clip5pNbases 7 0}. Alignment files were filtered to exclude read mappings deriving from inserts less than 30 bases (including the primer). We believe these short fragments represent unextended reverse transcription 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 they are not derived from cellular RNA.

RNA-seq libraries were sequenced on an Illumina HiSeq2500 by the BRC Genomics Facility at Cornell University using 100bp single end reads. Reads were aligned using the STAR aligner with the following alignment parameters: {--alignEndsType EndToEnd --alignIntronMin 20 --alignIntronMax 1000 --alignSJDBoverhangMin 1 --outSAMmultNmax 1 --outFilterMismatchNmax 3 --clip3pAdapterSeq CTGTCTCTTATACACATCTCCGAGCCCACGAGAC}.

When applicable, replicate libraries were combined prior to alignment. However, to assess technical reproducibility of MPE-seq, replicate libraries were subsampled to varying read depths, aligned separately, and compared to RNA-seq libraries also subsampled to varying read depths.

Estimating Splice isoform abundances from MPE-Seq data.

For each intron, the relative abundance of unspliced and spliced isoforms was determined by counting spliced and unspliced reads. Spliced reads (*S*) were counted using the SJ.out.tab file created by the aligner. Unspliced reads were counted using bedtools[24] to count the number of

reads that cover any part of the intron, considering only the first read of the paired-end reads. Unspliced read counts were further categorized as deriving from a lariat intermediate (*L*) or pre-1st step RNA (*P*) by considering the mapping location of the second read of the paired end reads, which we observed to often terminate near the TSS, or in the case of a lariat-intermediate-derived cDNA, near the branchpoint-A of the intron. Based on paired end mapping locations, each fragment was categorized into one of six categories (See FigS6) and the counts within those six categories were used to calculate *S*, *P*, and *L* as follows:

$$S = C_1 + C_2 \qquad P = C_3\left(1 + \frac{C_5 + C_6}{C_3 + C_4}\right) \qquad L = C_4\left(1 + \frac{C_5 + C_6}{C_3 + C_4}\right)$$

Locations of branchpoints (Supplementary Information Table 7) were determined by consolidating the most used branchpoint from lariat sequencing data[25] and previously described branch locations based on sequence motif searches[26].

Heatmaps and meta-gene plots: To generate metagene plots which illustrate read coverage around features of interest, we used the deepTools ComputeMatrix command[27] in conjunction with a BigWig coverage file of the 3' terminating bases and a bedfile containing TSS-positions as determined by PRO-cap[12] or a bedfile containing the annotated branchpoint regions detailed above. Importantly, this bedfile was filtered to only include branchpoint regions that would produce a lariat intermediate that is within the size range captured by library size-selection of MPE-seq libraries (see column “AttemptedLariatQuantification?” in Supplemental Information Table 2).

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 for Illumina”. Libraries were prepared following the manufacturer’s protocol.

Estimating Splice isoform abundances from RNA-seq data: Similar to MPE-seq data, spliced reads from target introns were counted using the SJ.out.tab file created by the aligner. Unspliced reads were counted using the bedtools software package[24] to count the number of reads which 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 x read length minus the minimum splice junction overhang length. The potential mapping space for an unspliced read is equal to the 2 x read length minus minimum splice junction overhang length plus length of the intron. Reads counts assigned to each feature were then divided by the length. Fraction unspliced was calculated for each intron as the quotient of length normalized unspliced reads and spliced reads. Relative transcript expression was calculated via transcripts per million (TPM) normalization[28], only considering exonic reads and exonic gene-lengths.

Acknowledgements

We thank members of the Pleiss and A. Grimson laboratories for critical feedback on this work. This work was funded by a Research Scholars Grant from the American Cancer Society and NIGMS grant (GM098634) to J.A.P.

Author Contributions

H.X., B.F., Z.D., M.G. and J.A.P. designed research. H.X., Z.D., and M.G. performed research. H.X., B.F., and Z.D. analyzed data. All authors wrote the paper.

Competing interests statement

The authors have no competing interests to declare.

1. Merkin, J., Russell, C., Chen, P. & Burge, C. B. Evolutionary dynamics of gene and isoform regulation in Mammalian tissues. *Science* **338,** 1593–9 (2012).

2. Barbosa-Morais, N. L. *et al.* The evolutionary landscape of alternative splicing in vertebrate species. *Science (80-.).* **338,** 1587–1593 (2012).

3. Mercer, T. R. *et al.* Targeted RNA sequencing reveals the deep complexity of the human transcriptome. *Nat. Biotechnol.* **30,** 99–104 (2011).

4. Mercer, T. R. *et al.* Targeted sequencing for gene discovery and quantification using RNA CaptureSeq. *Nat. Protoc.* **9,** 989–1009 (2014).

5. Blomquist, T. M. *et al.* Targeted RNA-Sequencing with Competitive Multiplex-PCR Amplicon Libraries. *PLoS One* **8,** e79120 (2013).

6. Kivioja, T. *et al.* Counting absolute numbers of molecules using unique molecular identifiers. *Nat. Methods* **9,** 72–74 (2012).

7. Zhu, Y. Y., Machleder, E. M., Chenchik, A., Li, R. & Siebert, P. D. Reverse transcriptase template switching: A SMART^{TM} approach for full-length cDNA library construction. *BioTechniques* **30,** 892–897 (2001).

8. Zheng, W., Chung, L. M. & Zhao, H. Bias detection and correction in RNA-Sequencing data. *BMC Bioinformatics* **12,** (2011).

9. Carey, M. F., Peterson, C. L. & Smale, S. T. The primer extension assay. *Cold Spring Harb. Protoc.* **8,** 164–173 (2013).

10. Coombes, C. E. & Boeke, J. D. An evaluation of detection methods for large lariat RNAs. *RNA* **11,** 323–31 (2005).

11. Padgett, R. A. *et al.* Nonconsensus branch-site sequences in the in vitro splicing of transcripts of mutant rabbit beta-globin genes. *Proc. Natl. Acad. Sci. U. S. A.* **82,** 8349–

8353 (1985).

12. Booth, G. T., Wang, I. X., Cheung, V. G. & Lis, J. T. Divergence of a conserved elongation factor and transcription regulation in budding and fission yeast. *Genome Res.* **26,** 799–811 (2016).

13. Kim, S. H. & Lin, R. J. Spliceosome activation by PRP2 ATPase prior to the first transesterification reaction of pre-mRNA splicing. *Mol. Cell. Biol.* **16,** 6810–6819 (1996).

14. Chen, W. *et al.* Transcriptome-wide interrogation of the functional intronome by spliceosome profiling. *Cell* **173,** 1031–1044 (2018).

15. Burke, J. *et al.* Spliceosome profiling visualizes the operations of a dynamic RNP in vivo at nucleotide resolution. *Cell* **173,** 1014–1030 (2018).

16. Nojima, T. *et al.* Mammalian NET-Seq Reveals Genome-wide Nascent Transcription Coupled to RNA Processing. *Cell* **161,** 526–540 (2015).

17. Lucks, J. B. *et al.* Multiplexed RNA structure characterization with selective 2'-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq). *Proc. Natl. Acad. Sci.* **108,** 11063–11068 (2011).

18. Rouskin, S., Zubradt, M., Washietl, S., Kellis, M. & Weissman, J. S. Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo. *Nature* **505,** 701–5 (2014).

19. Hartwell, L. H., McLaughlin, C. S. & Warner, J. R. Identification of ten genes that control ribosome formation in yeast. *MGG Mol. Gen. Genet.* **109,** 42–56 (1970).

20. Wernersson, R. & Nielsen, H. B. OligoWiz 2.0 - Integrating sequence feature annotation into the design of microarray probes. *Nucleic Acids Res.* **33,** (2005).

21. Collart, M. A. & Oliviero, S. in *Current Protocols in Molecular Biology* (2001).

doi:10.1002/0471142727.mb1312s23

22. Engel, S. R. *et al.* The reference genome sequence of Saccharomyces cerevisiae: then and now. *G3 (Bethesda).* **4,** 389–98 (2014).

23. Dobin, A. *et al.* STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29,** 15–21 (2013).

24. Quinlan, A. R. & Hall, I. M. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26,** 841–842 (2010).

25. Mayerle, M. *et al.* Structural toggle in the RNaseH domain of Prp8 helps balance splicing fidelity and catalytic efficiency. *Proc. Natl. Acad. Sci.* **114,** 4739–4744 (2017).

26. Grate, L. & Ares, M. Searching yeast intron data at Ares lab web site. *Methods in Enzymology* **350,** 380–392 (2002).

27. Ramírez, F. *et al.* deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.* **44,** W160–W165 (2016).

28. Conesa, A. *et al.* A survey of best practices for RNA-seq data analysis. *Genome Biology* **17,** (2016).

Figure legends:

Figure1: MPE-seq uses complex pools of reverse transcription primers to target sequencing to regions of interest

(A) MPE-seq comprises the following steps: (1) Primers containing sequencing adapter overhangs and a target-specific region are designed for genomic regions of interest. (2) Primers are pooled and used to reverse transcribe total cellular RNA in the presence of amino-allyl dUTP (aa-dUTP), which allows (3) biotin coupling and purification of cDNAs from other reaction components. (4) A second sequencing adapter is appended at the 3' terminus of the cDNA through first strand extension using Klenow. Libraries are further purified (5), PCR amplified (6) and sequenced (7).

(B) Genome browser screenshot of a targeted region in MPE-seq (pink) and conventional RNA-seq (purple)

Figure2: MPE-seq enrichment enables high-precision measurements of splicing

(A) The percentage of reads mapped to target and off-target regions is depicted for MPE-seq and conventional RNA-seq. In MPE-seq a small fraction of reads were categorized as "Unextended primer" which corresponds to short primer extension products (0-5 bases extended past the primer) and thus they were not categorized as cDNAs derived from RNA targets.

(B) Each point represents the fold enrichment of a target region in MPE-seq over conventional RNA-seq.

(C) Scatter plots depict intron-retention measurements in replicate libraries in MPE-seq and conventional RNA-seq at matched or greater read depth.

Figure3: MPE-seq enables genome-wide profiling of lariat intermediates

(A) Schematic depicting cDNA products derived from pre 1st step (P), lariat intermediate (L) and spliced mRNA (S) isoforms.

(B) Meta-intron coverage plot surrounding predicted branchpoints in a wild-type (Prp2) and step1 splicing mutant strain (*prp2-1*). The region between the +10 position downstream of the annotated branchpoint and the 3' splice site (3'ss) was re-scaled for each intron.

(C) Heatmap plots showing the relative coverage at each intron for which reads were detected.

(D) Estimates of the relative abundance of each isoform for each targeted intron for which reads were detected.

FigureS1: Elevated temperatures in reverse transcription reactions increase specificity

The fraction of on-target and off-target reads from replicate MPE-seq libraries generated from reverse transcription reactions performed at various temperatures. A small fraction of reads were categorized as "Unextended primer" which corresponds to short primer extension products (0-5 bases extended past the primer) and thus they were neither categorized as cDNAs derived from RNA targets or unamappable.

FigS2: Expression measurements as determined by MPE-seq and RNA-seq

(A) A scatter plot depicts gene expression measurements (RNA-seq) in replicate datasets. Genes containing splice-events that were among those chosen for targeted sequencing are depicted in red. These targeted genes range in expression levels by orders of magnitude.

(B) A scatter plot depicts gene expression measurements in replicate MPE-seq datasets (red). Similar to conventional RNA-seq, expression measurements in MPE-seq are highly reproducible between replicates, even for the small proportion of mis-priming events that map to off-target locations (grey).

(C) A scatter plot depicts gene expression measurements in RNA-seq and MPE-seq. The right shift of targeted genes reflects successful enrichment of targets by orders of magnitude. The observation that even highly expressed genes as measured by RNA-seq are proportionally highly expressed in MPE-seq suggests that primers are not limiting during reverse transcription.

FigS3: Splicing measurements as determined by MPE-seq and RNA-seq

(A) A scatter plot depicts intron-retention measurements in MPE-seq and conventional RNA-seq using wildtype (Prp2) RNA.

(B) A scatter plot depicts intron-retention measurements in MPE-seq and conventional RNA-seq using RNA from a splicing mutant strain (*prp2-1*).

(C) A scatter plot depicts the fold-change (*prp2-1/*Prp2) in intron-retention as measured by MPE-seq and conventional RNA-seq

FigS4: Transcription start site profiling by MPE-seq

Metagene profile of 3' ends mapped by MPE-seq, centered around transcription start sites (TSS) as determined by PRO-cap, an orthologous method for mapping transcription start sites. The high abudance of read ends that pile up at TSSs indicates that MPE-seq can be used to profile cDNA termini.

FigS5: Lariat-intermediate derived cDNAs contain a unique signature of mismatches incorporated by reverse transcriptase at the branched adenosine

(A) Genome browser screenshot of 3' reads from paired-end sequenced fragments illustrates the unique signature of non-templated base incorporation by reverse transcriptase at a branched-adenosine vs the 5' RNA terminus

(B) Genome-wide quantification of the mismatch frequencies at 3' cDNA termini near the TSS (left) versus at the annotated branchpoint (right).

FigS6: Schematic for assigning reads to splice intermediate isoforms

To quantify the abundance of pre 1st step, lariat intermediate, and spliced isoforms for each targeted splice event, we categorized fragments into six classes based on paired-end read alignments. Fragments containing a splice junction (C_1 and C_2) are indicative of a spliced RNA (S). Fragments that are unspliced and traverse the branchpoint region (C_3) are classified as pre-1st step RNA (P). Fragments that are unspliced but terminate within a -3 to +5bp window from the previously determined branchpoint (C_4) are classified as lariat intermediate (L). Fragments that are unspliced and either terminate downstream of the branchpoint (C_5) or the terminus could not be mapped (C_6) are ambiguous between P and L. Therefore, for accounting purposes, the counts for these fragments were coerced into P and L classifications based on the ratio of P and L determined by unambiguous mappings (C_3 and C_4). See methods for more details.

FigS7: Transcript features that correlate with the abundance of lariat intermediate

(A) Scatter plot depicting the correlation between the abundance of pre-1st step RNA and the abundance of lariat intermediate

(B) The abundance of pre-1st step RNA and lariat intermediate RNA is significantly correlated with classification of introns into those that are in ribosomal protein genes (RPG) and non-RPG. However, the abundance of lariat intermediate relative to pre-1st step RNA, a metric of the efficiency of the second step of splicing, does not correlate.

(C) Spearman correlations of various features to the abundance of pre-1st step RNA, lariat intermediate, the abundance of lariat intermediate relative to pre-1st step RNA. Errorbars indicate 95% confidence intervals as estimated by Fisher transformation of Spearman's correlation coefficient.

FigS8: Array-based oligonucleotide synthesis can be used to generate primer pools for use in MPE-seq

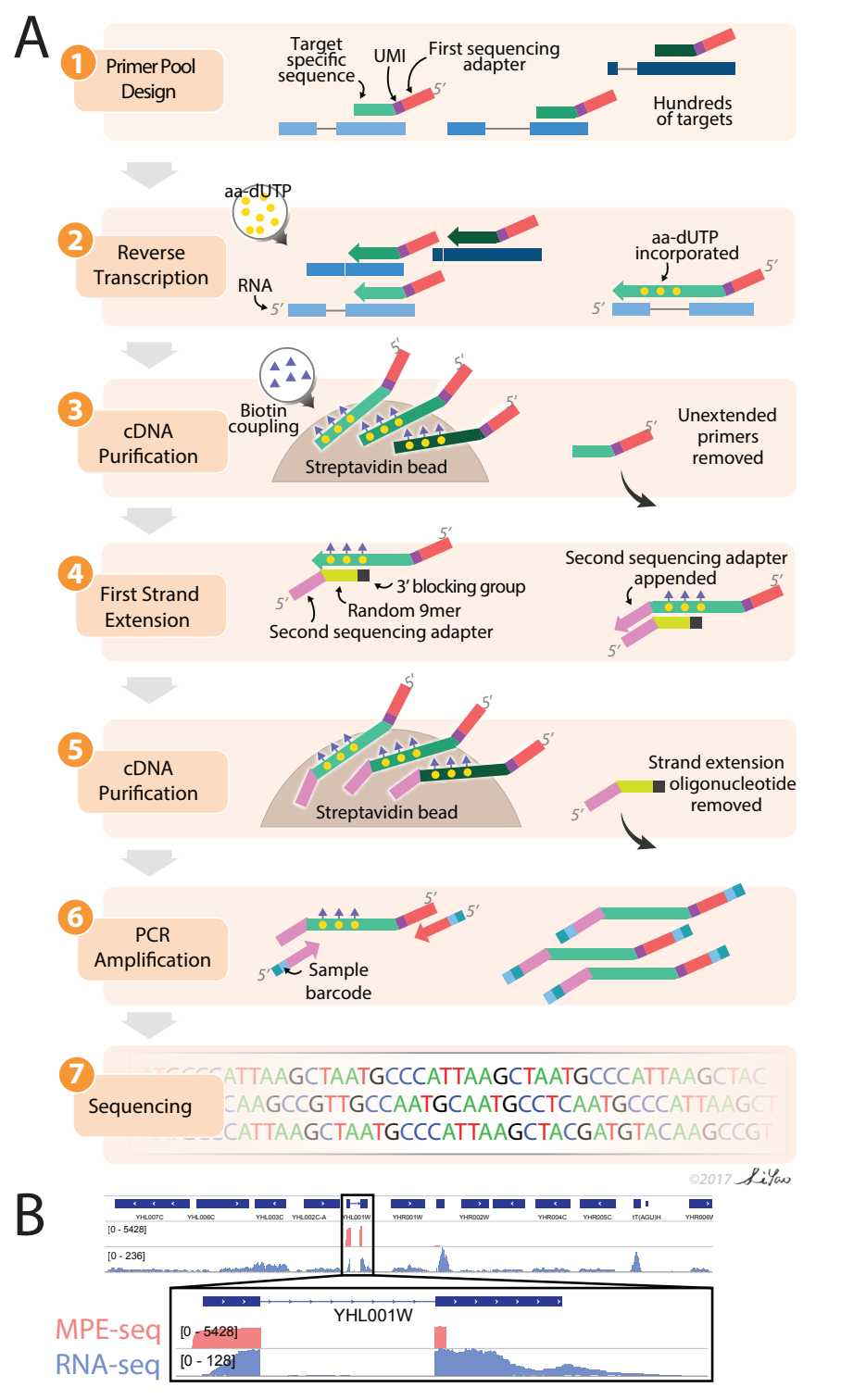
(A) Obtaining adequate amounts of primer pools for MPE-seq from cost-effective array-based oligonucleotide synthesis can be achieved in four steps: (1) PCR amplification of the oligonucleotide synthesis pool using a 5' blocked sense primer and a biotinylated antisense primer. (2) Restriction digestion to cleave off the PCR primer handle. (3) Lambda exonuclease digestion of free 5' ends. (4) Streptavidin purification of biotinylated PCR handle. The unbound fraction is the desired primer pool product.

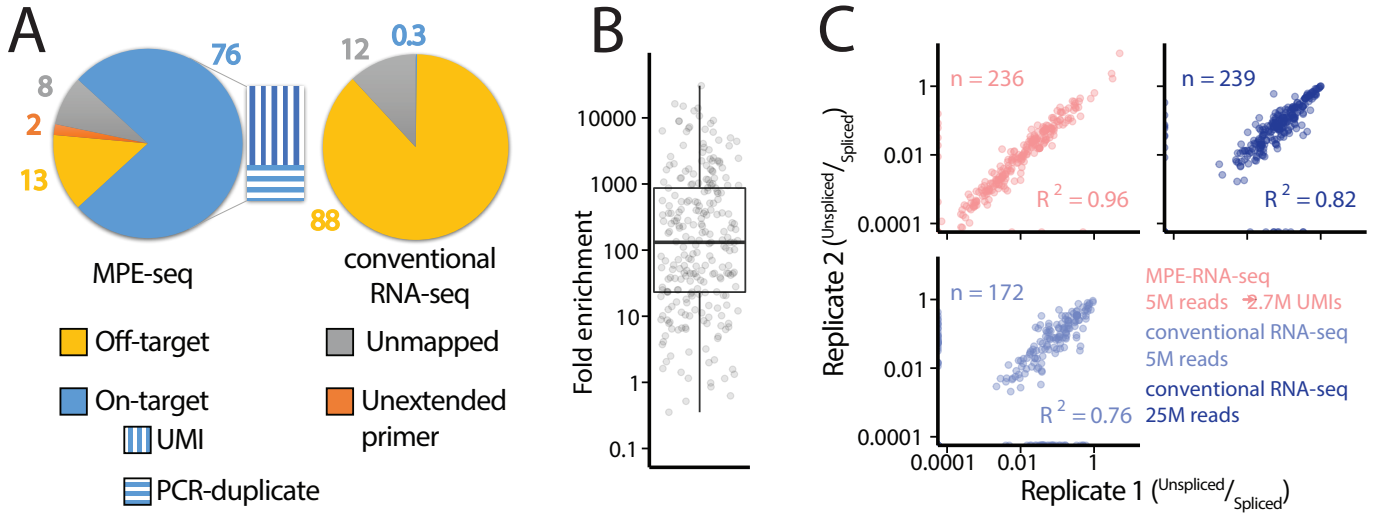
(B) Steps during the amplification and purification of array-synthesized primer pools are monitored via native gel electrophoresis. The control lane represents a pool of individually synthesized MPE-seq primers which did not require amplification and purification.

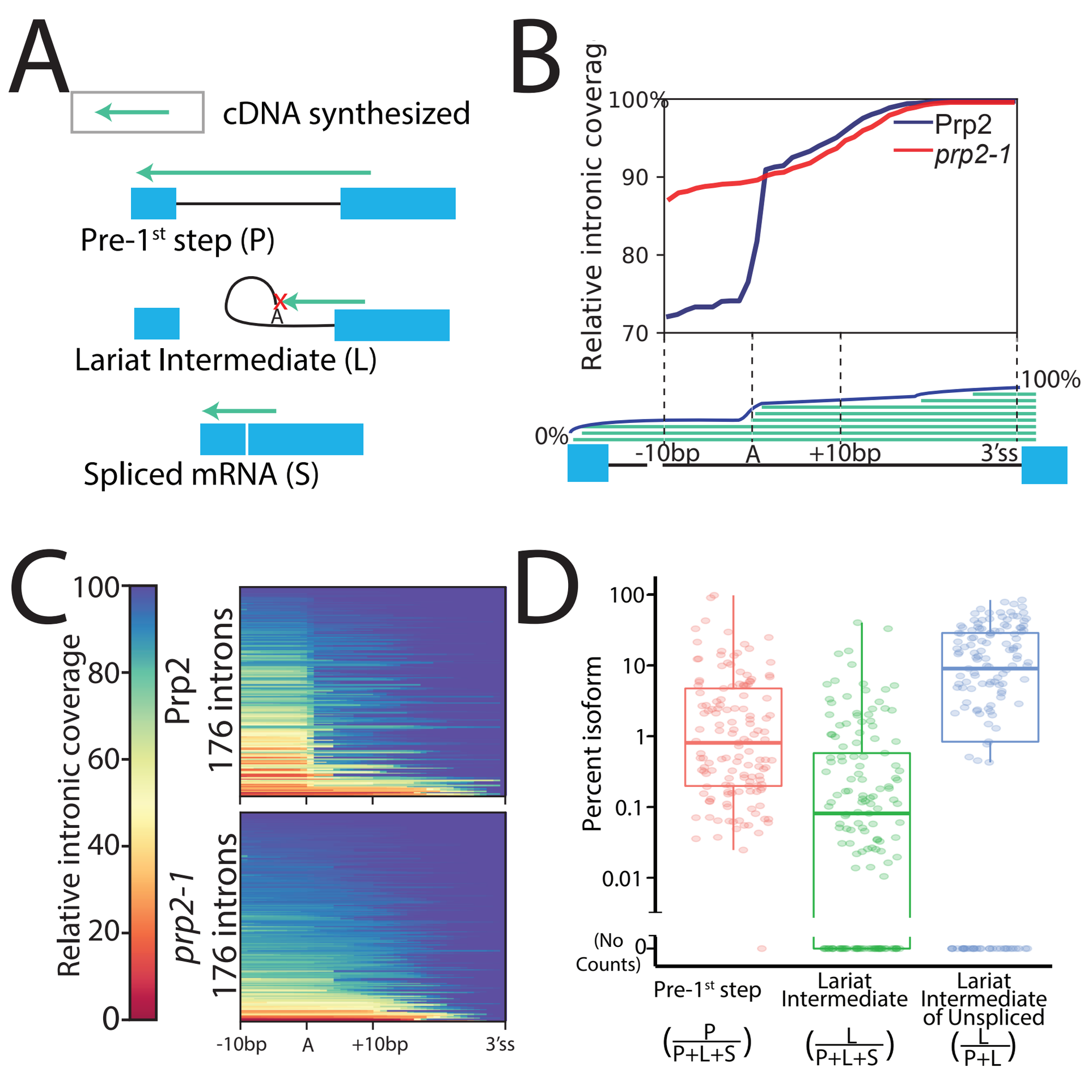
(C) The percentage of reads mapped to target and off-target regions is depicted for MPE-seq using array-synthesized primers. In MPE-seq a small fraction of reads were categorized as "Unextended primer" which corresponds to short primer extension products (0-5 bases

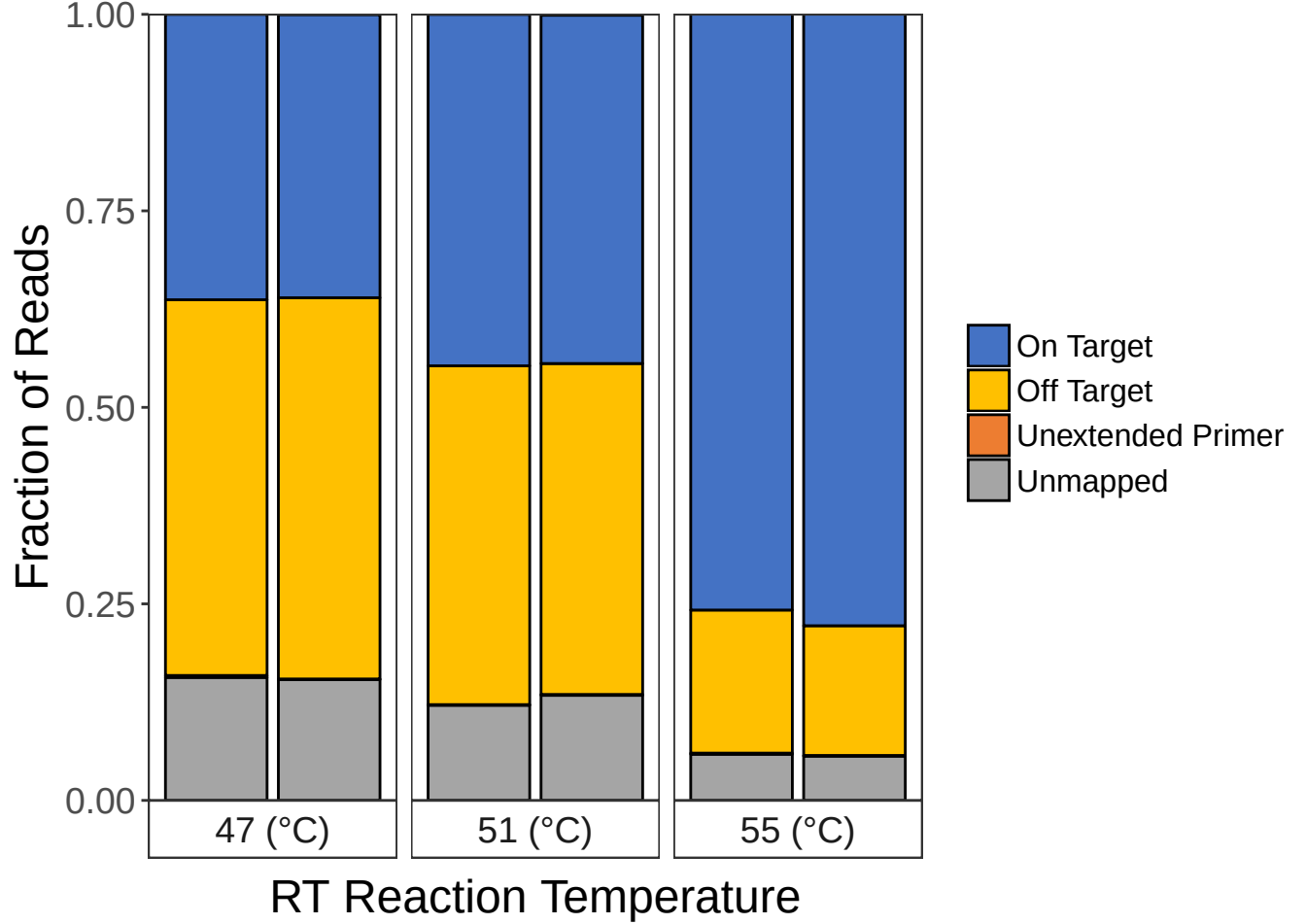
extended past the primer) and thus they were not categorized as cDNAs derived from RNA targets.

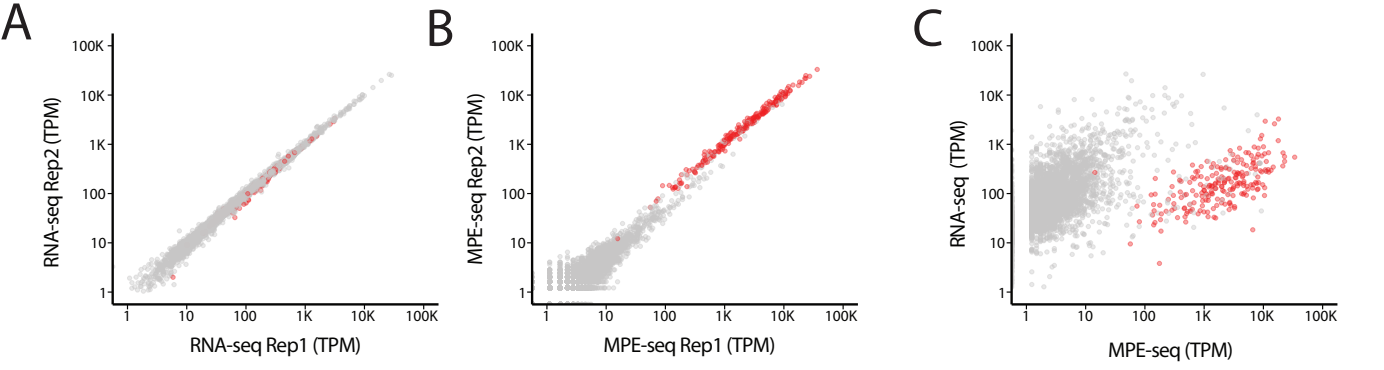
(D) A scatter plot depicts intron-retention measurements in MPE-seq libraries which used individually synthesized primer pools and array-based synthesis of primer pools

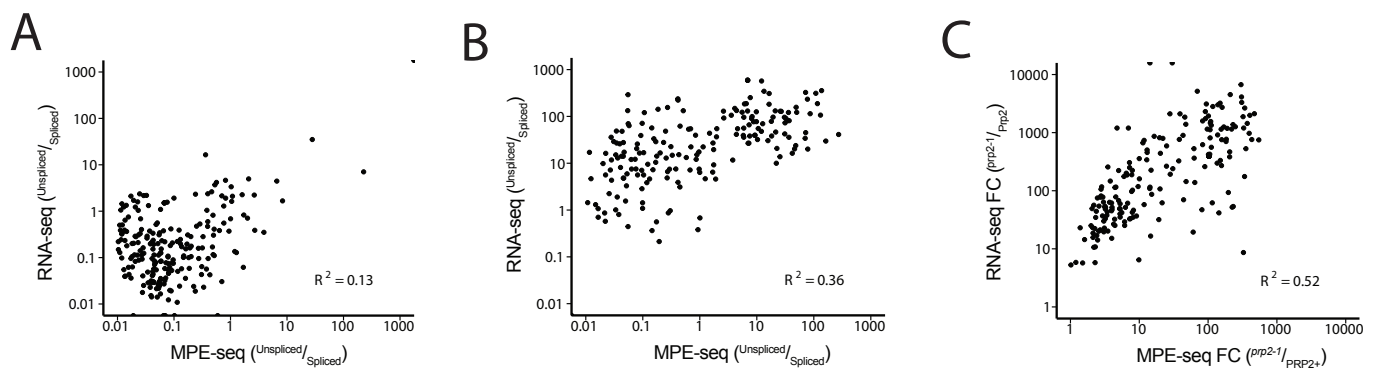

A
1 Primer Pool Design
Target specific sequence
UMI
First sequencing adapter
5′
Hundreds of targets
2 Reverse Transcription
aa-dUTP
RNA
5′
aa-dUTP incorporated
3 cDNA Purification
Biotin coupling
Streptavidin bead
Unextended primers removed
4 First Strand Extension
3′ blocking group
Random 9mer
Second sequencing adapter
Second sequencing adapter appended
5 cDNA Purification
Streptavidin bead
Strand extension oligonucleotide removed
6 PCR Amplification
Sample barcode
7 Sequencing
ATTAAGCTAATGCCCATTAAGCTAATGCCCATTAAGCTAC
CAAGCCGTTGCCAATGCAATGCCTCAATGCCCATTAAGCT
CATTAAGCTAATGCCCATTAAGCTACGATGTACAAGCCGT
©2017 Liew
B
YHL007C
YHL006C
YHL003C
YHL002C-A
YHL001W
YHR001W
YHR002W
YHR004C
YHR005C
YHR006W
[0 - 5428]
[0 - 236]
YHL001W
MPE-seq
[0 - 5428]
RNA-seq
[0 - 128]

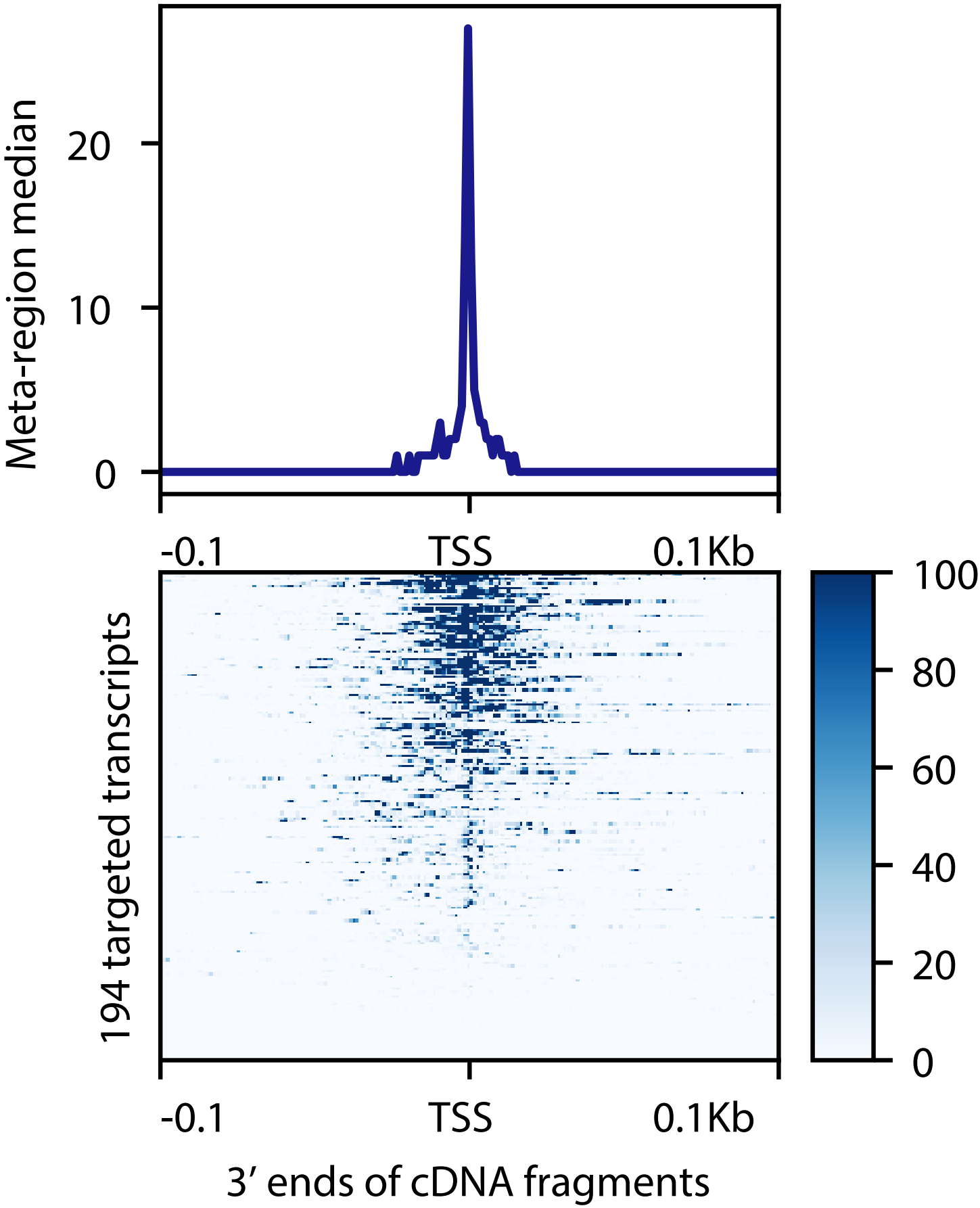

A
76
8
2
13
12
0.3
88
MPE-seq
conventional RNA-seq
Off-target
Unmapped
On-target
Unextended primer
UMI
PCR-duplicate
B
Fold enrichment
10000
1000
100
10
1
0.1
C
n = 236
R 2 = 0.96
n = 239
R 2 = 0.82
n = 172
R 2 = 0.76
Replicate 2 (Unspliced/Spliced)
Replicate 1 (Unspliced/Spliced)
1
0.01
0.0001
MPE-RNA-seq
5M reads ~2.7M UMIs
conventional RNA-seq
5M reads
conventional RNA-seq
25M reads


A
cDNA synthesized
Pre-1st step (P)
A
Lariat Intermediate (L)
Spliced mRNA (S)
B
Relative intronic coverag
100%
90
80
70
Prp2
prp2-1
100%
0%
-10bp
A
+10bp
3'ss
C
Relative intronic coverage
100
80
60
40
20
0
Prp2
176 introns
prp2-1
176 introns
-10bp
A
+10bp
3'ss
D
Percent isoform
100
10
1
0.1
0.01
0
(No Counts)
Pre-1st step
Lariat Intermediate
Lariat Intermediate of Unspliced
(P/(P+L+S))
(L/(P+L+S))
(L/(P+L))

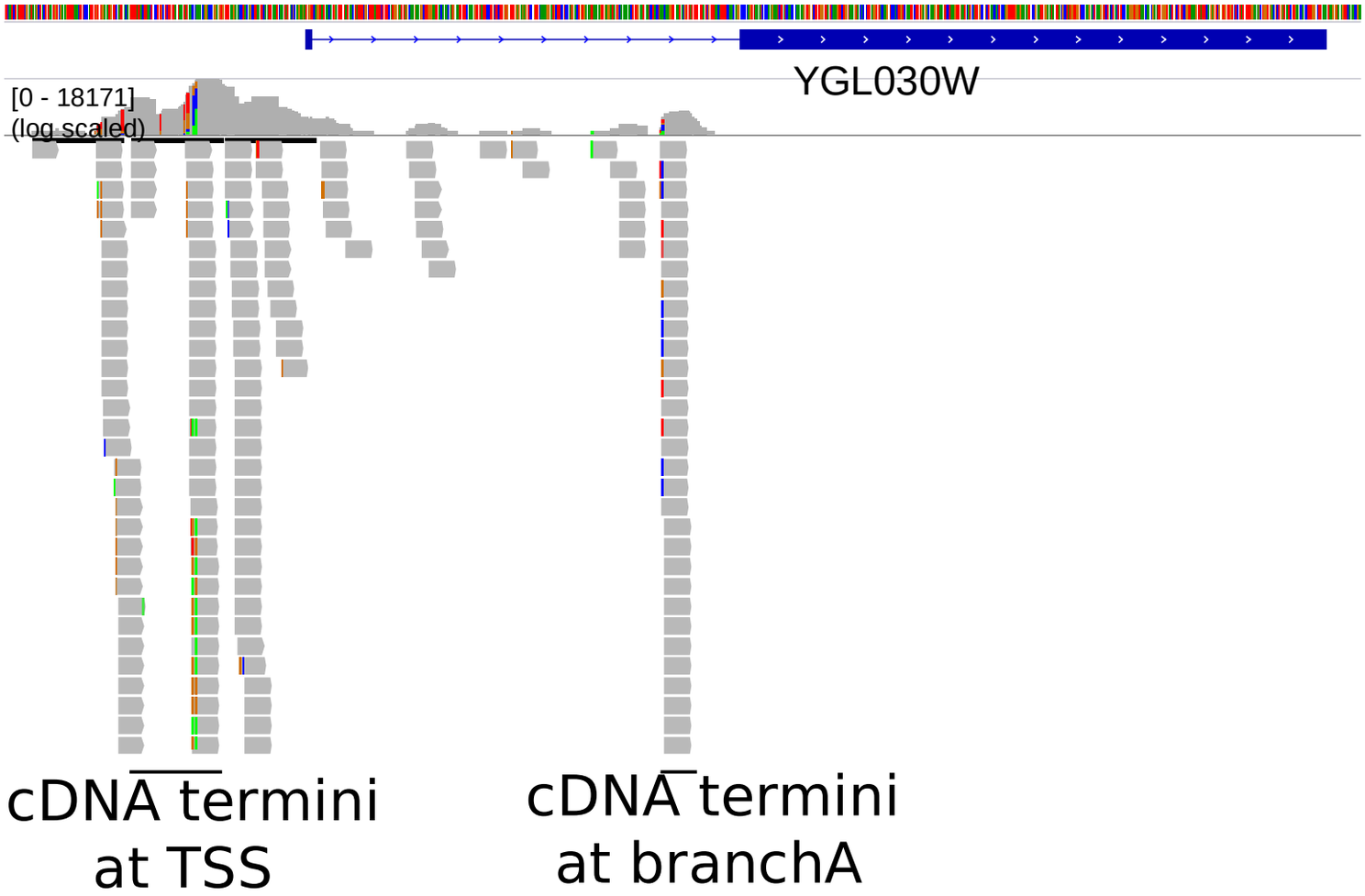

1.00
0.75
0.50
0.25
0.00
Fraction of Reads
47 (°C)
51 (°C)
55 (°C)
RT Reaction Temperature
On Target
Off Target
Unextended Primer
Unmapped


A
RNA-seq Rep2 (TPM)
RNA-seq Rep1 (TPM)
B
MPE-seq Rep2 (TPM)
MPE-seq Rep1 (TPM)
C
RNA-seq (TPM)
MPE-seq (TPM)
100K
10K
1K
100
10
1

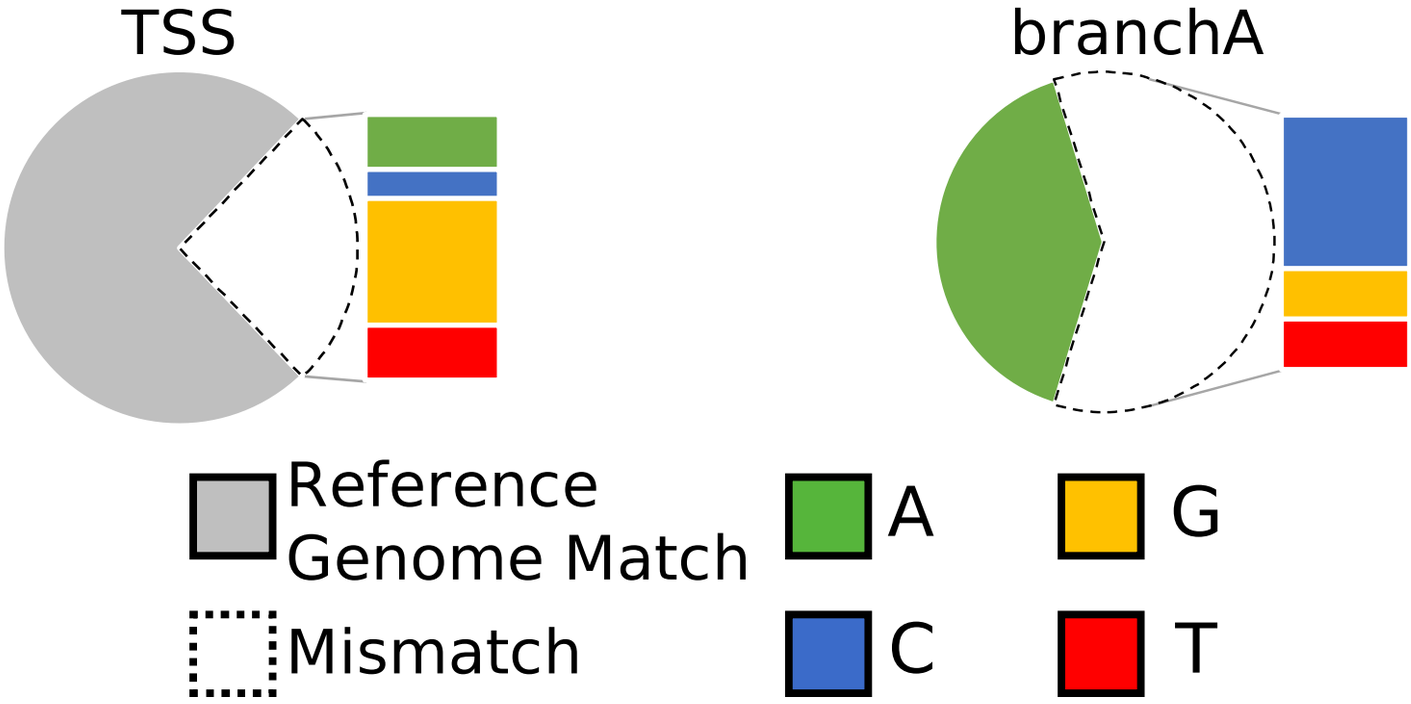

A
RNA-seq (Unspliced/Spliced)
MPE-seq (Unspliced/Spliced)
R 2 = 0.13
B
RNA-seq (Unspliced/Spliced)
MPE-seq (Unspliced/Spliced)
R 2 = 0.36
C
RNA-seq FC (prp2-1/Prp2)
MPE-seq FC (prp2-1/PRP2+)
R 2 = 0.52

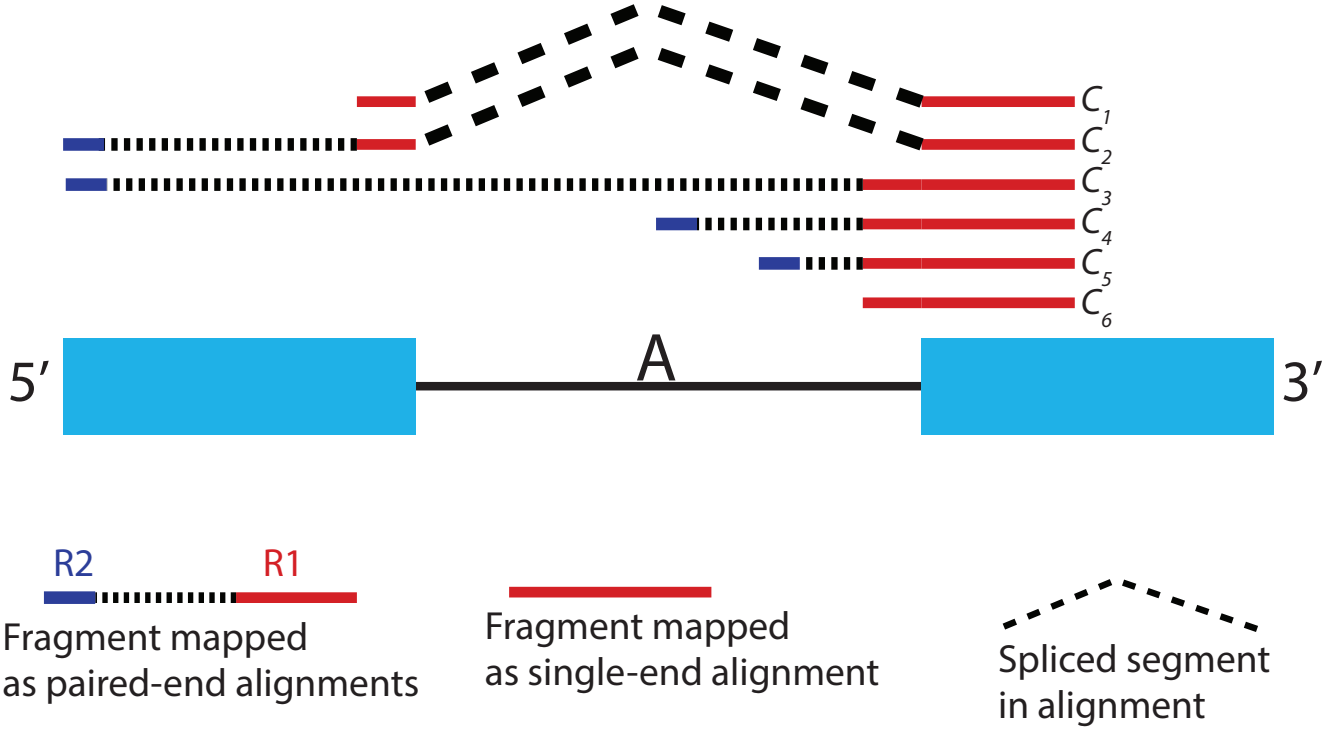

Meta-region median
20
10
0
-0.1
TSS
0.1Kb
194 targeted transcripts
100
80
60
40
20
0
-0.1
TSS
0.1Kb
3' ends of cDNA fragments

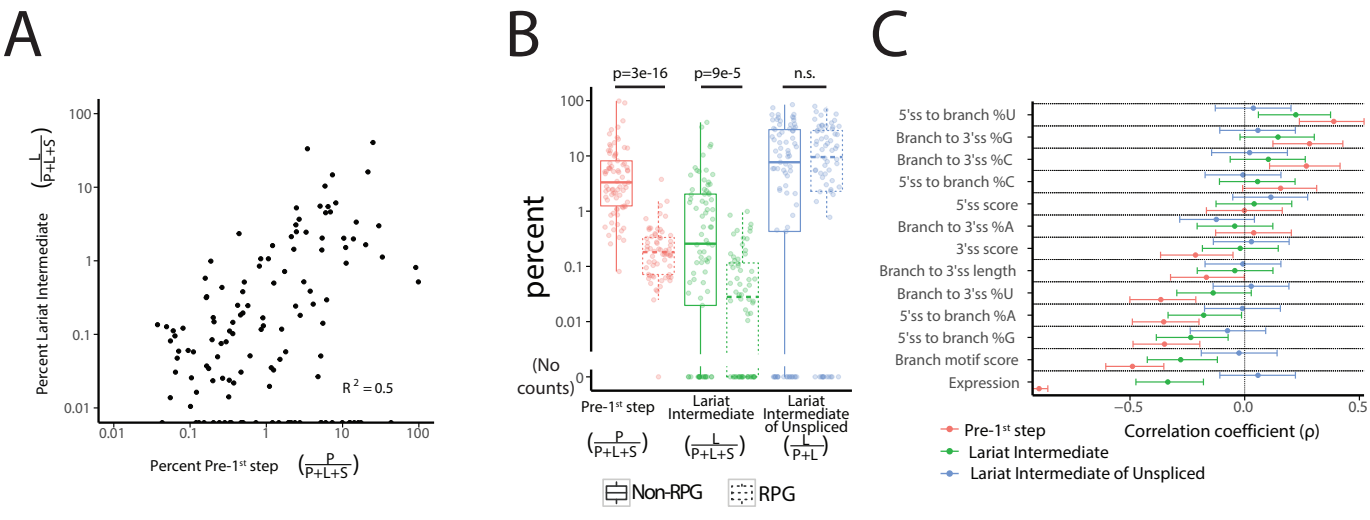
A


YGL030W
[0 - 18171]
(log scaled)
cDNA termini
at TSS
cDNA termini
at branchA

B


TSS
branchA
Reference
Genome Match
Mismatch
A
C
G
T


C1
C2
C3
C4
C5
C6
5′
A
3′
R2
R1
Fragment mapped
as paired-end alignments
Fragment mapped
as single-end alignment
Spliced segment
in alignment


A
Percent Lariat Intermediate (L/(P+L+S))
100
10
1
0.1
0.01
R 2 = 0.5
0.01
0.1
1
10
100
Percent Pre-1st step (P/(P+L+S))
B
p=3e-16
p=9e-5
n.s.
percent
100
10
1
0.1
0.01
(No counts)
0
Pre-1st step
Lariat Intermediate
Lariat Intermediate of Unspliced
(P/(P+L+S))
(L/(P+L+S))
(L/(P+L))
Non-RPG
RPG
C
5'ss to branch %U
Branch to 3'ss %G
Branch to 3'ss %C
5'ss to branch %C
5'ss score
Branch to 3'ss %A
3'ss score
Branch to 3'ss length
Branch to 3'ss %U
5'ss to branch %A
5'ss to branch %G
Branch motif score
Expression
−0.5
0.0
0.5
Correlation coefficient (ρ)
Pre-1st step
Lariat Intermediate
Lariat Intermediate of Unspliced

A

Array synthesis of oligos

5′ UMI SapI digest site 3′

5′ First sequencing adapter Gene specific variable region Reverse amplification primer region 3′

1 PCR Amplification

Blocked 5′ end

Forward amplification primer

biotin

Reverse amplification primer

2 SapI Digestion

SapI digestion

3 Lambda Exonuclease Digestion

Lambda exonuclease

4 Streptavidin Purification

Streptavidin bead

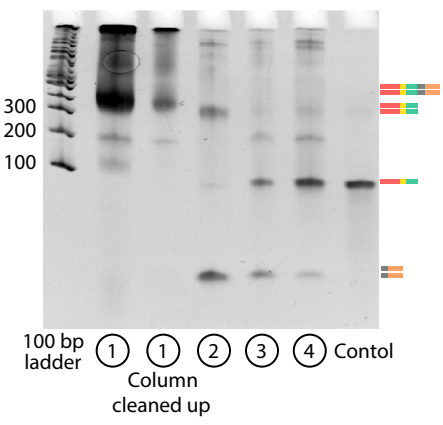
Removed

Complex Primer Pool

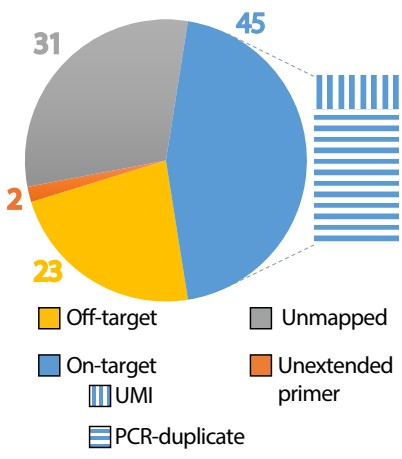
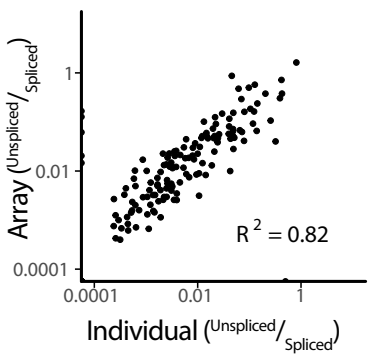
~300 different gene specific variable regions

©2017 Lilao

B



C



D