Detailed explanation of the Smart-3SEQ method

See Figure S1 for a detailed visualization of the reactions and Figure S2 for the oligonucleotide sequences compatible with the Illumina platform.

First, the input RNA is fragmented to a desirable size by a divalent cation in high heat. Whole cells or tissue can be lysed with the addition of a detergent in the same step; lysing fixed tissue may require adding proteinase K at this step and an inhibitor afterward. There is no need to purify the RNA after fragmentation as it performed in the presence of some components of the next reaction mix, nor is mRNA enrichment or rRNA depletion required. Instead, immediately after RNA fragmentation, reverse transcription is primed with an oligonucleotide comprising a 3′ anchored oligo(dT) primer, which hybridizes to the beginning of the RNA template's poly(A) tail (so only the single fragment that contains this site is reverse-transcribed), and a non-complementary 5′ sequence matching the innermost portion of the downstream sequencing adapter (P7 on the Illumina platform). This incorporates the partial adapter into the first cDNA strand and eliminates the need to add that adapter by ligation later.

After extending the first cDNA strand, MMLV-derived reverse transcriptase tends to extend several non-template bases at the 3′ end, which are primarily dC. This provides a target for hybridization with a second oligonucleotide, which comprises a short 3′ oligo(G) primer and the innermost portion of the upstream sequencing adapter (Illumina P5). Reverse transcriptase then performs a "template switch", extending a second cDNA strand from this new primer. Thus the reverse transcription produces both the first and second cDNA strands in a single incubation, and this ds-cDNA has partial sequencing adapters at both ends. Note that the template-switch oligonucleotide consists mainly of DNA but the 3′ guanine residues are RNA to reduce off-target strand invasion. Each oligonucleotide in the template-switching reverse transcription also includes a blocking group (biotin) at its 5′ end to discourage concatenation of additional adapters.

All that remains to produce a sequencing-ready library is to extend the adapters to full length, including the multiplexing barcodes, and to amplify the library to sufficient concentration for quality control and pooling. Both purposes are served by PCR with primers matching the sequences of the entire adapters, which anneal to the partial adapters on the cDNA and extend them to full length. Finally, the only cleanup step in the protocol is to purify the amplified dsDNA library by a single SPRI procedure, using stringent conditions to avoid retaining molecules that are too short to be useful. Optionally, because they are now labeled with separate barcodes, the amplified libraries can be pooled before cleanup to combine them into a single small volume, reducing the amount of downstream handling and yielding acceptable concentrations from lower numbers of PCR cycles.

When the library is sequenced, each read contains up to five sections (Figure S3A): 1) the UMI, a set of random bases included in the second-strand primer to discriminate PCR duplicates from fragmentation duplicates; 2) a short stretch of Gs; 3) cDNA sequence matching the source transcript; 4) a long stretch of As, if the read length is longer than the cDNA insert; and 5) potentially the downstream adapter sequence, if the read length is much longer than the cDNA insert, though in practice bases downstream of a homopolymer tend to be poorly read. When the cDNA sequences are aligned to the reference transcriptome, they align in the sense orientation slightly upstream of the polyadenylation site (Figure S3B) and the read count is directly proportional to the abundance of the source transcript, regardless of the transcript's length.

Validation of the laser ablation method

Dissecting single cells with LCM is difficult and sometimes more than one cell is recovered on the cap. We ensured that our libraries came from true single cells by destroying the extraneous cells with the UV laser (Figures S18A, S18B). To verify that this eliminates the signal from the ablated cells, we performed an experiment on a larger scale. On each cap we collected a roughly equal number of macrophages and DCIS cells, then on some caps we ablated all cells of one type, while on other caps we performed no ablation as a control (3 bulk ablations for each tissue type plus 2 no-ablation controls). Compared with the bulk samples (6 per tissue type), the gene-expression data from the no-ablation controls resembled a mix of both cell types, while the profiles from the ablated samples resembled the bulk data from their unablated cell types (Figures S18C, S18D).

Table S1: Comparison of Smart-3SEQ with selected RNA-seq methods. Cost per library includes all reagents (kits, SPRI beads, enzymes) but not consumables (tubes, pipet tips).

	TruSeq Stranded Total RNA (Illumina)	SMART-Seq v4 Ultra Low (Takara)	SMARTer Pico (Takara)	3SEQ (Beck et al. 2010)	Smart-3SEQ
Cost per library (USD)	$160	$110	$45	$120	$5
Required total RNA	100 ng	10 pg	250 pg	10 µg	10 pg
Protocol time	2 working days	2.5 working days	5 hours	2 working days	3 hours
Strand-specific?	Yes	No	Yes	Yes	Yes
Supports damaged RNA?	Yes	No	Yes	Yes	Yes
Works directly on cells?	No	Yes	No	No	Yes
Works directly on FFPE tissue?	No	No	No	No	Yes

Table S2: Software and configurations used

Program (suite)	Version	Important arguments
bcl2fastq	2.17.1.14	--minimum-trimmed-read-length 0 --mask-short-adapter-reads 0
NovoAlign	3.08.00	-H --trim3HP -a AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
STAR	2.5.3a	--outFilterMultimapNmax 1 --outFilterMismatchNmax 999 --clip3pAdapterMMp 0.2 --clip3pAdapterSeq AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
featureCounts (subread)	1.5.0-p2	-s 1 --read2pos 5

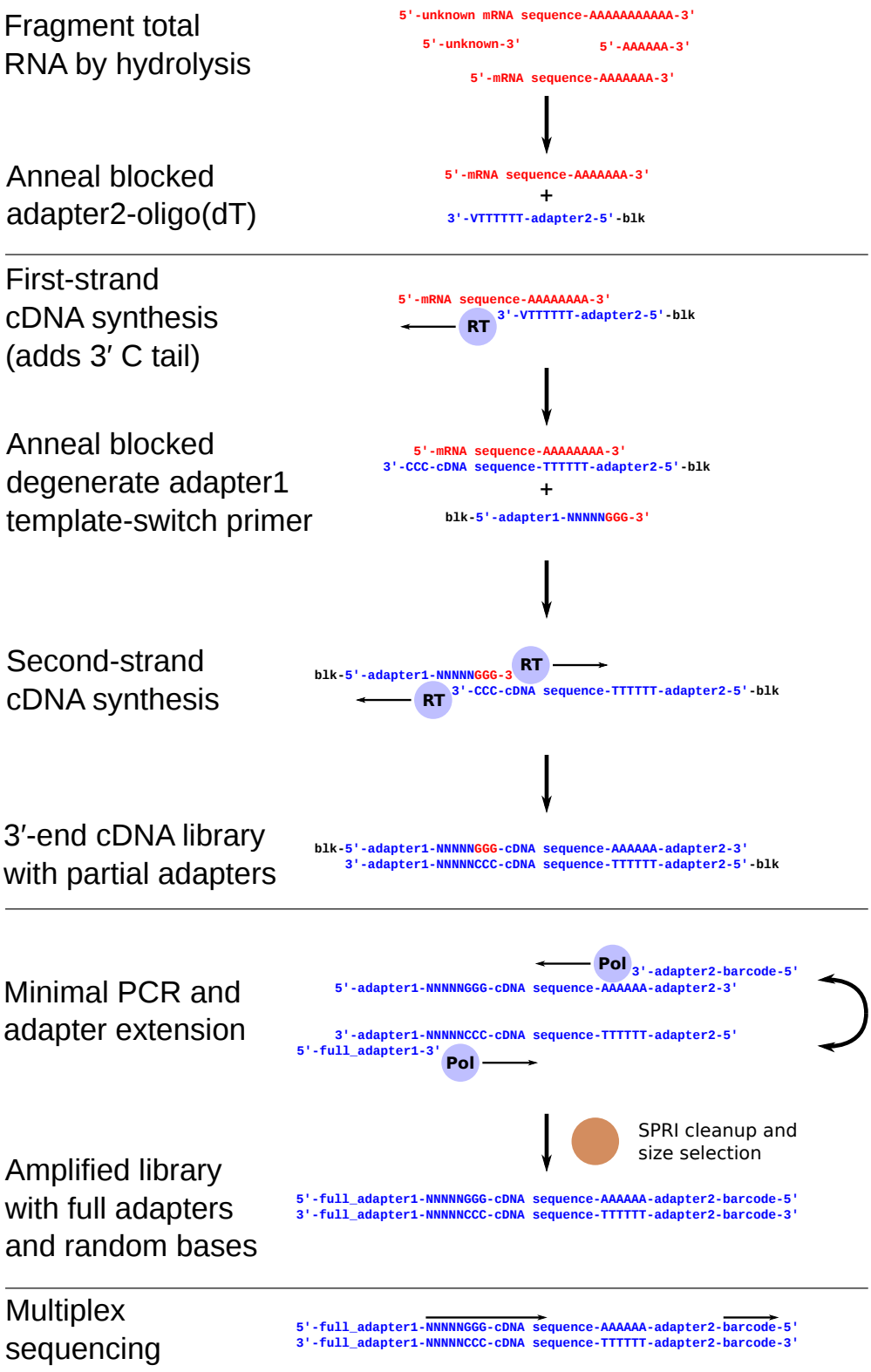


Figure S1: Schematic of the Smart-3SEQ library preparation method.

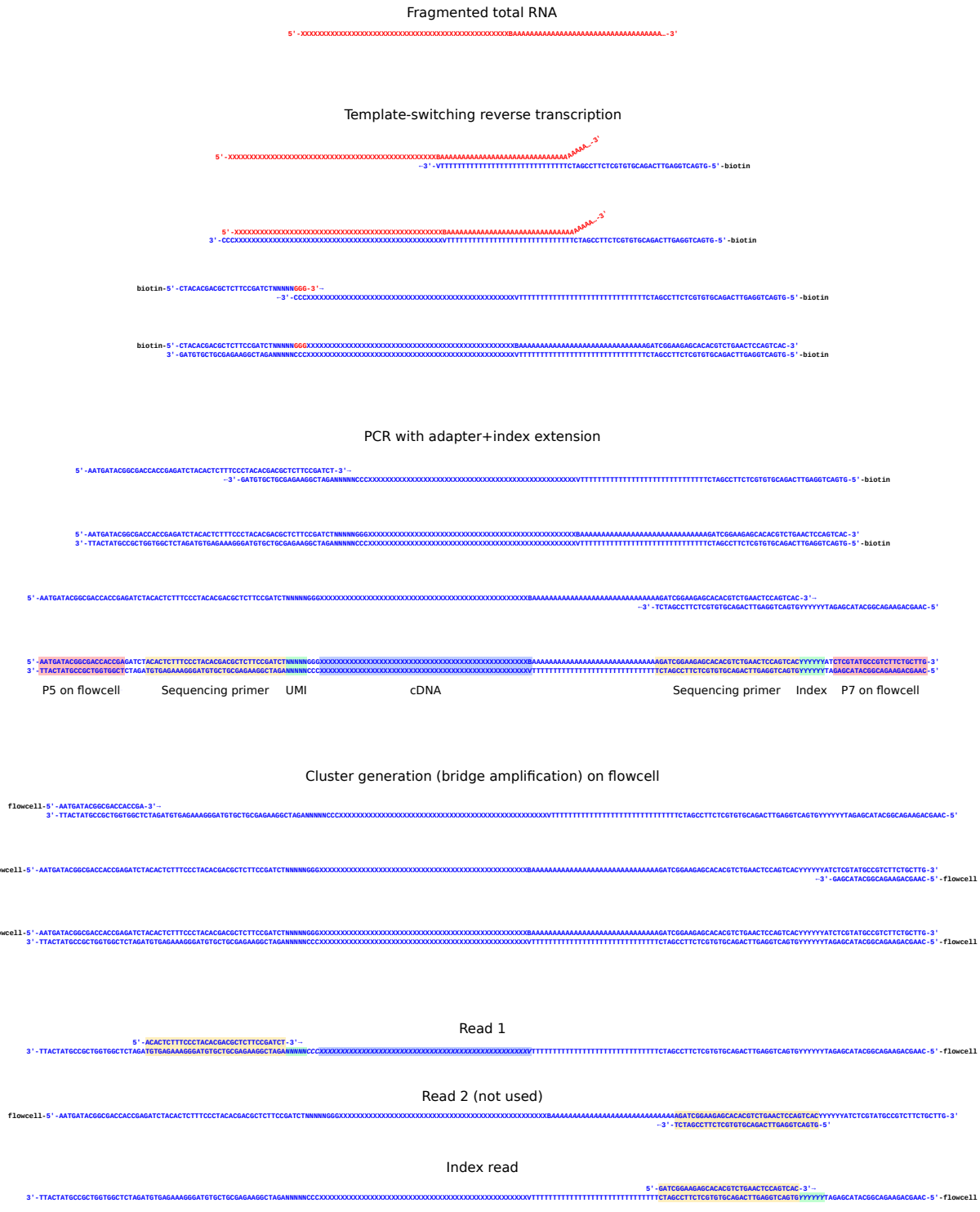
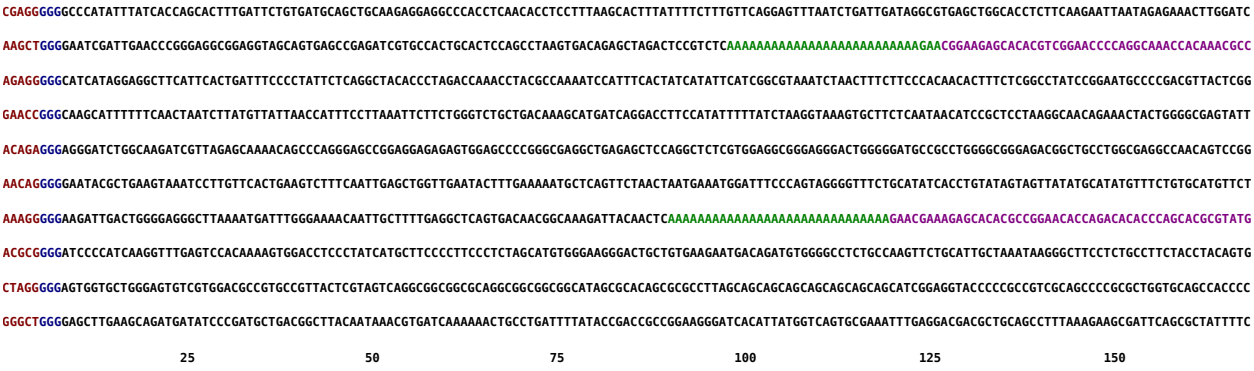
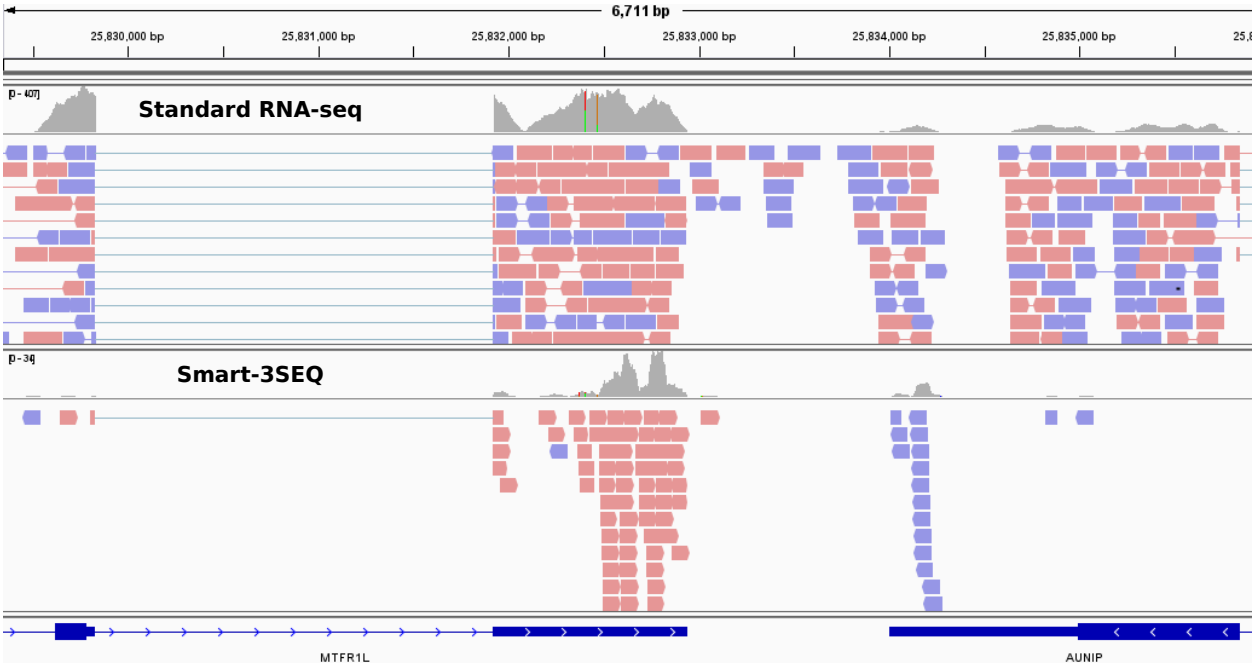


Figure S2: Schematic of oligonucleotide sequences used in Smart-3SEQ library preparation and sequencing on the Illumina platform with single indexing. Oligonucleotide sequences © 2017 Illumina, Inc. All rights reserved. Derivative works created by Illumina customers are authorized for use with Illumina instruments and products only. All other uses are strictly prohibited.



(A)



(B)

Figure S3: Example data from Smart-3SEQ. A: Read sequences. Before the cDNA sequence, each read begins with a 5 nt UMI (red), which is tracked for deduplication, then a G_3 stretch (blue) derived from the overhang required for template-switching, which is discarded. The length of the remaining cDNA insert (black) depends on the integrity of the input RNA: damaged RNA will yield shorter fragments. If the reads are longer than the inserts, they may continue into the poly(A) sequence (green), whose length is expected to match that of the oligo(dT) reverse-transcription primer, 30 nt. Any further base calls (purple) are unreliable because of the difficulty of sequencing through the homopolymer. B: IGV screenshot of example alignments, alongside RNA-seq data (top) from the same sample. RNA-seq reads span all exons, while Smart-3SEQ yields sense-oriented reads that only align directly upstream of the transcription termination site. They may still straddle an exon–exon junction if the final exon is short.

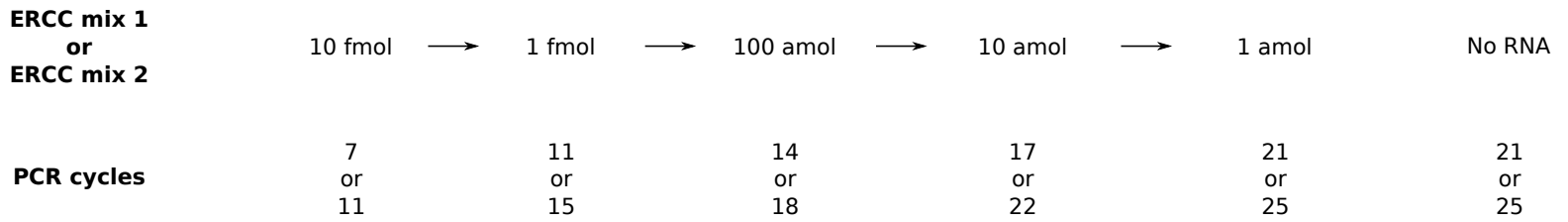


Figure S4: Experimental design of validation with ERCC standards.

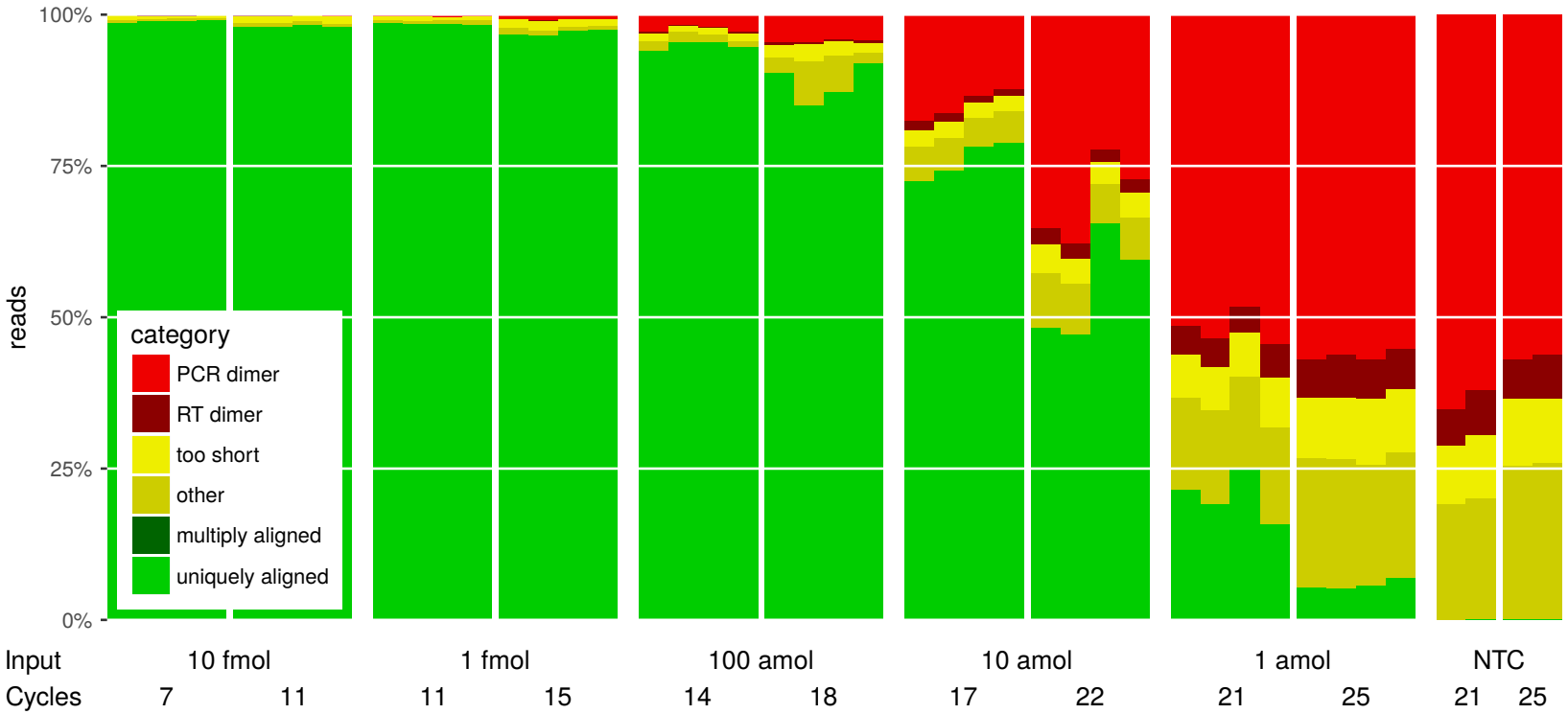


Figure S5: Alignability of Smart-3SEQ reads from ERCC dilutions and no-template controls.

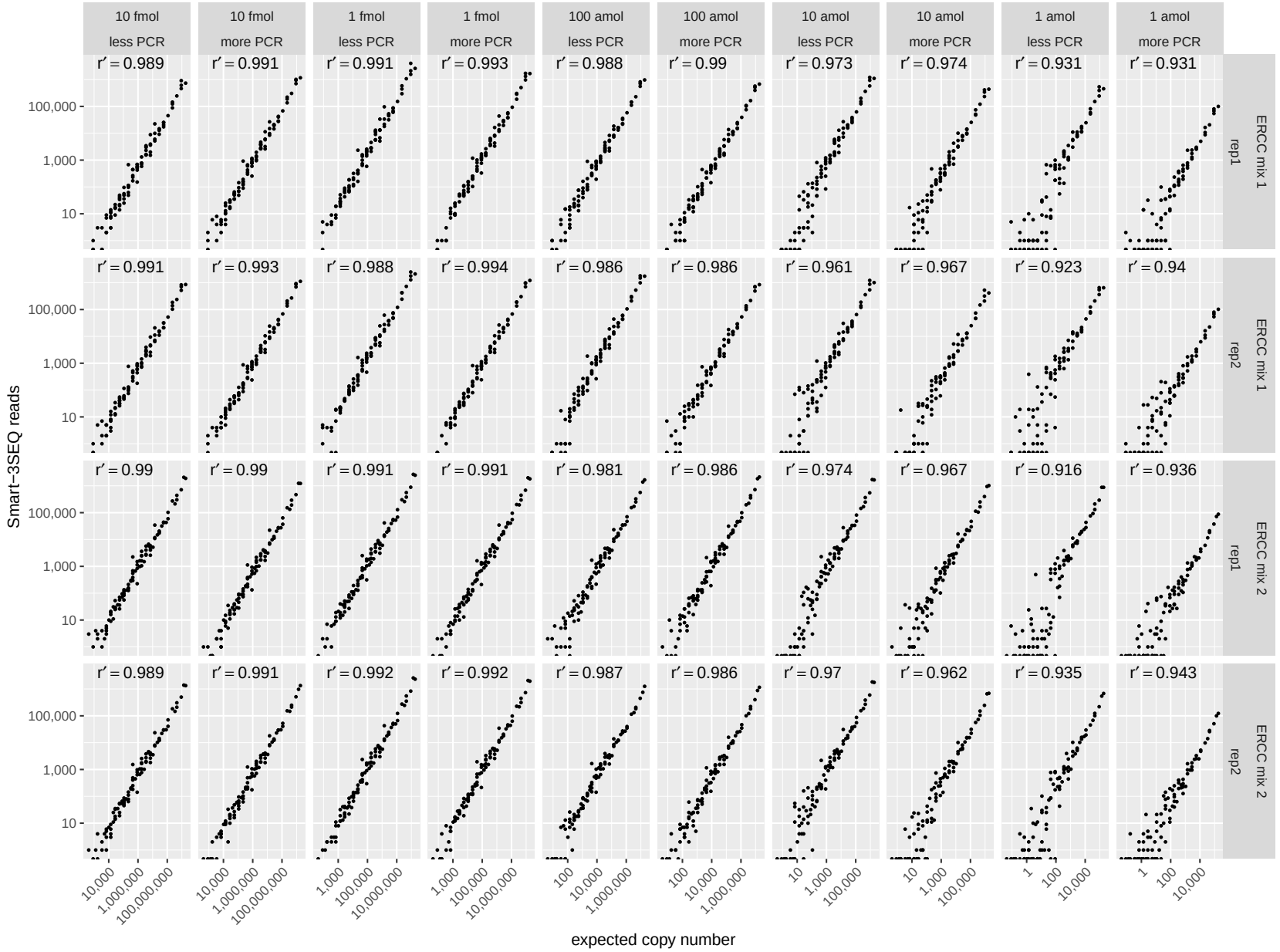


Figure S6: Accuracy of Smart-3SEQ. Smart-3SEQ standard curves from ERCC experiment. Each point is a single transcript.

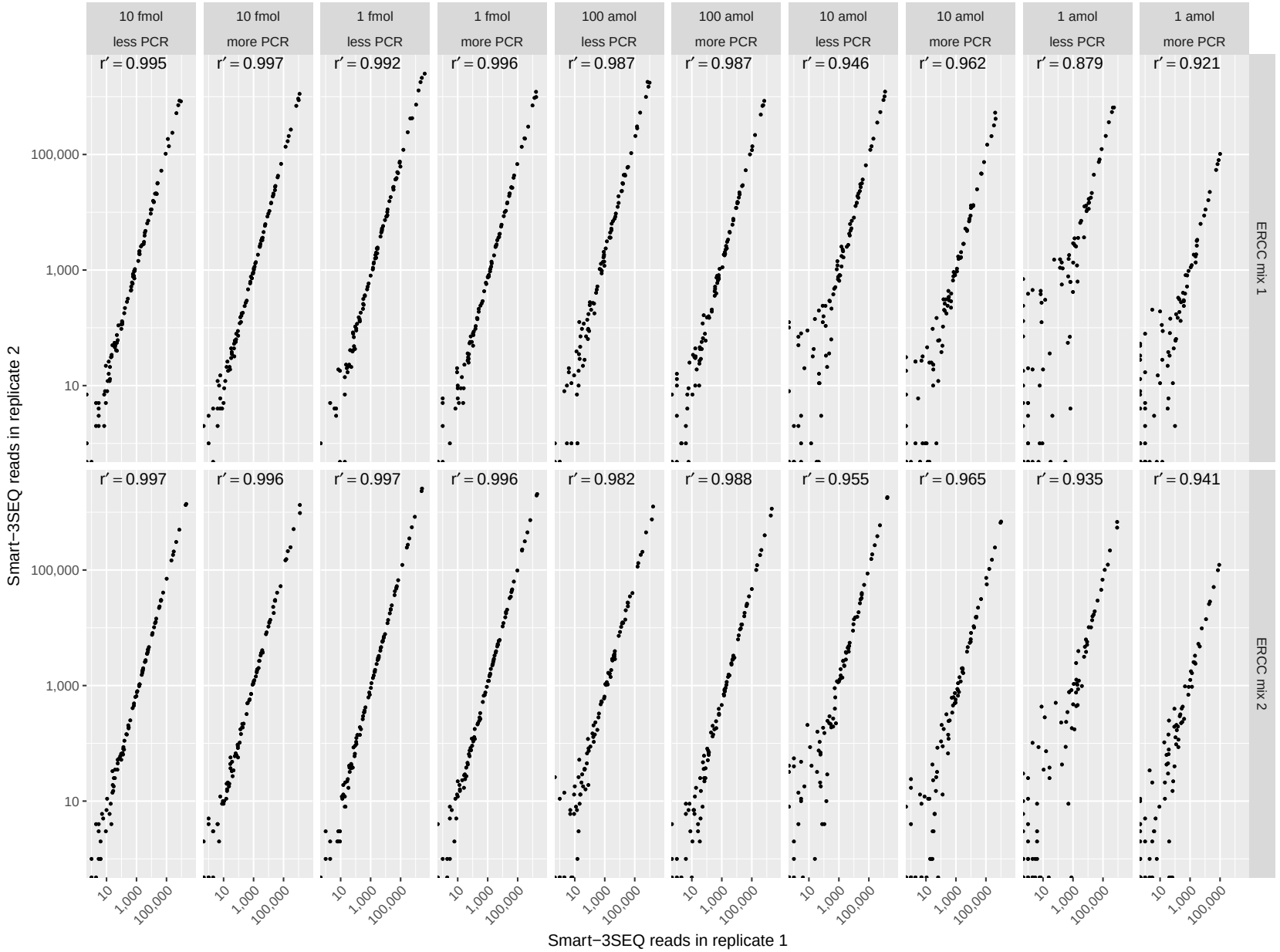


Figure S7: Precision of Smart-3SEQ. Correlation of technical replicates in ERCC experiment.

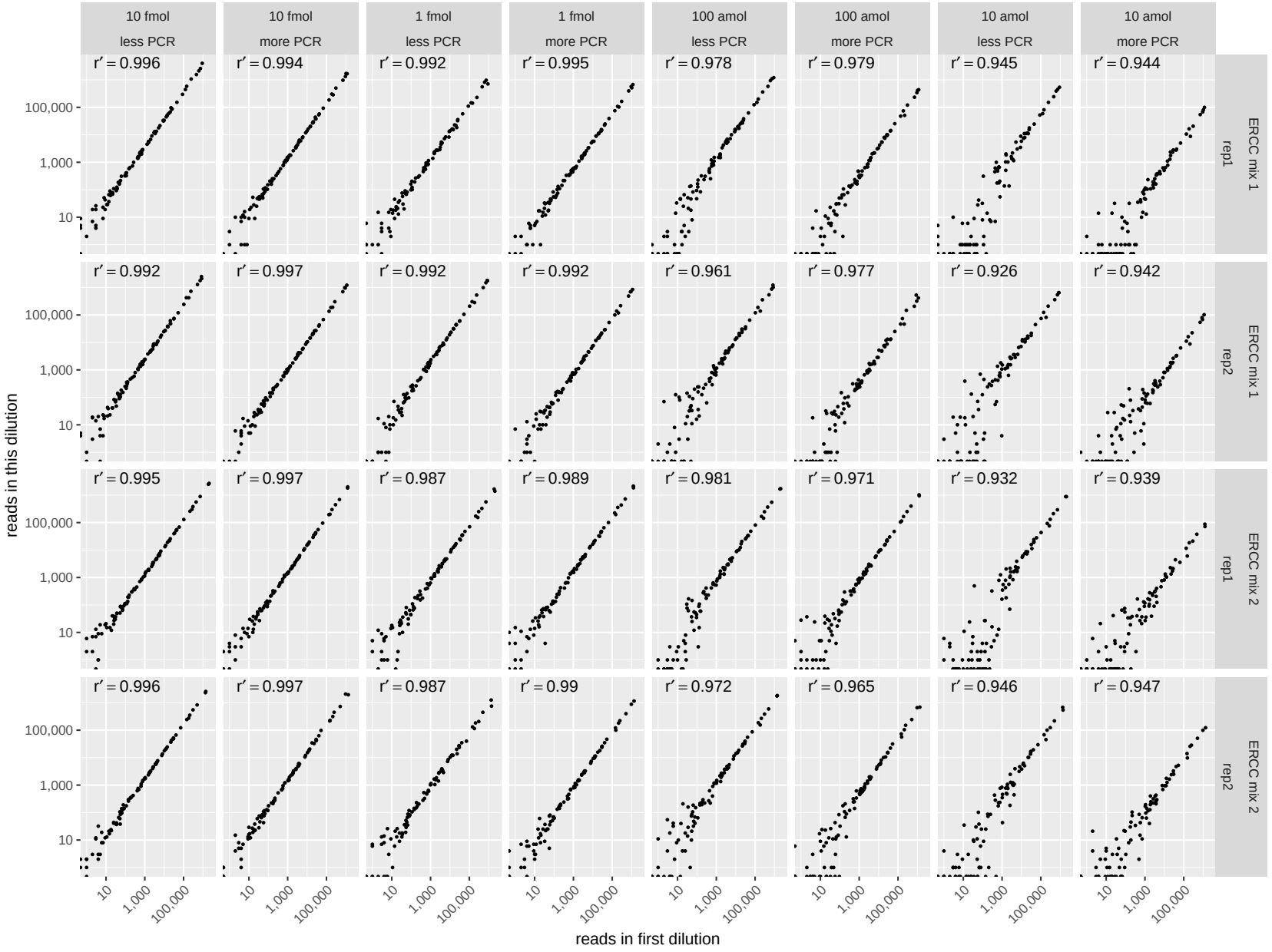


Figure S8: Sensitivity of Smart-3SEQ. Correlation of read counts in each subsequent dilution with those in the first dilution.

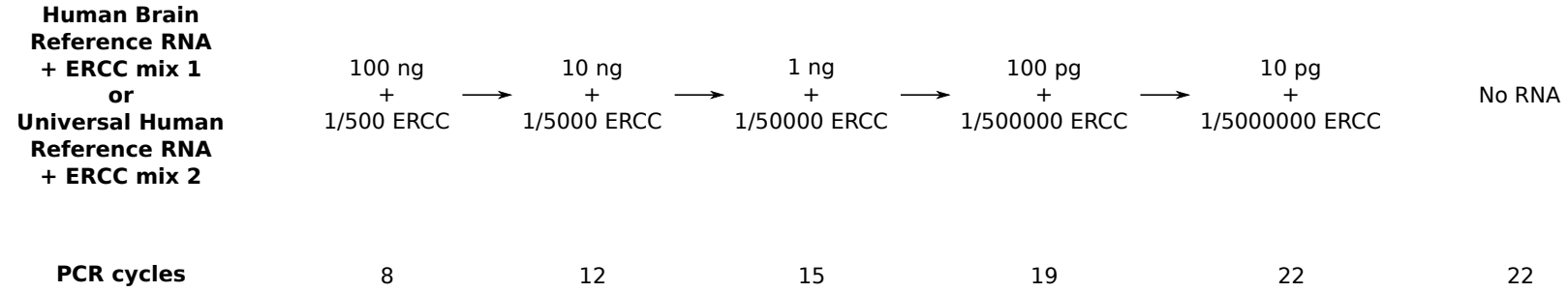


Figure S9: Experimental design of validation with human reference RNAs.

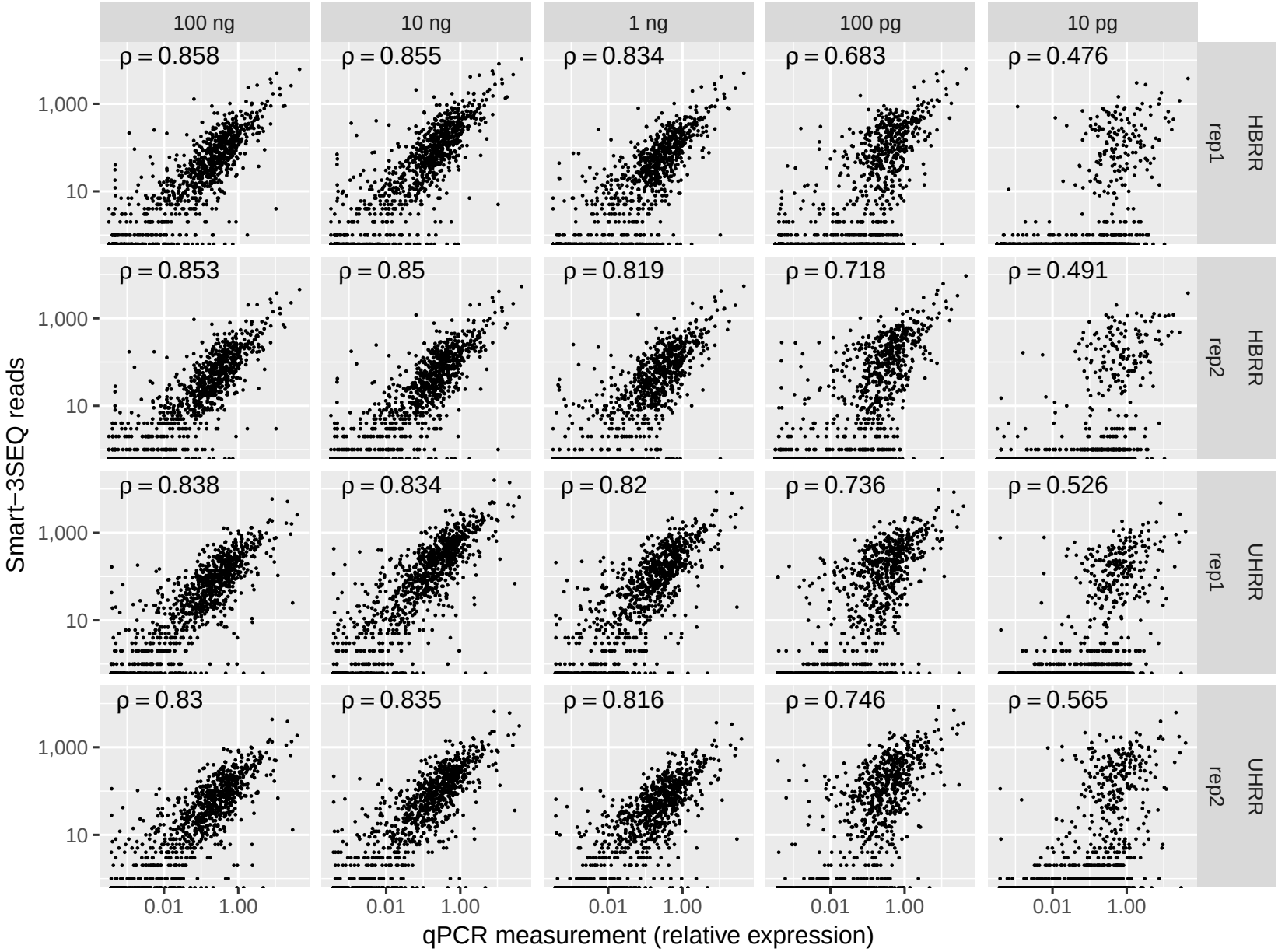


Figure S10: Correspondence of Smart-3SEQ results with TaqMan qPCR measurements. Each point is a single gene with available TaqMan qPCR data. ρ: Spearman's rank correlation.

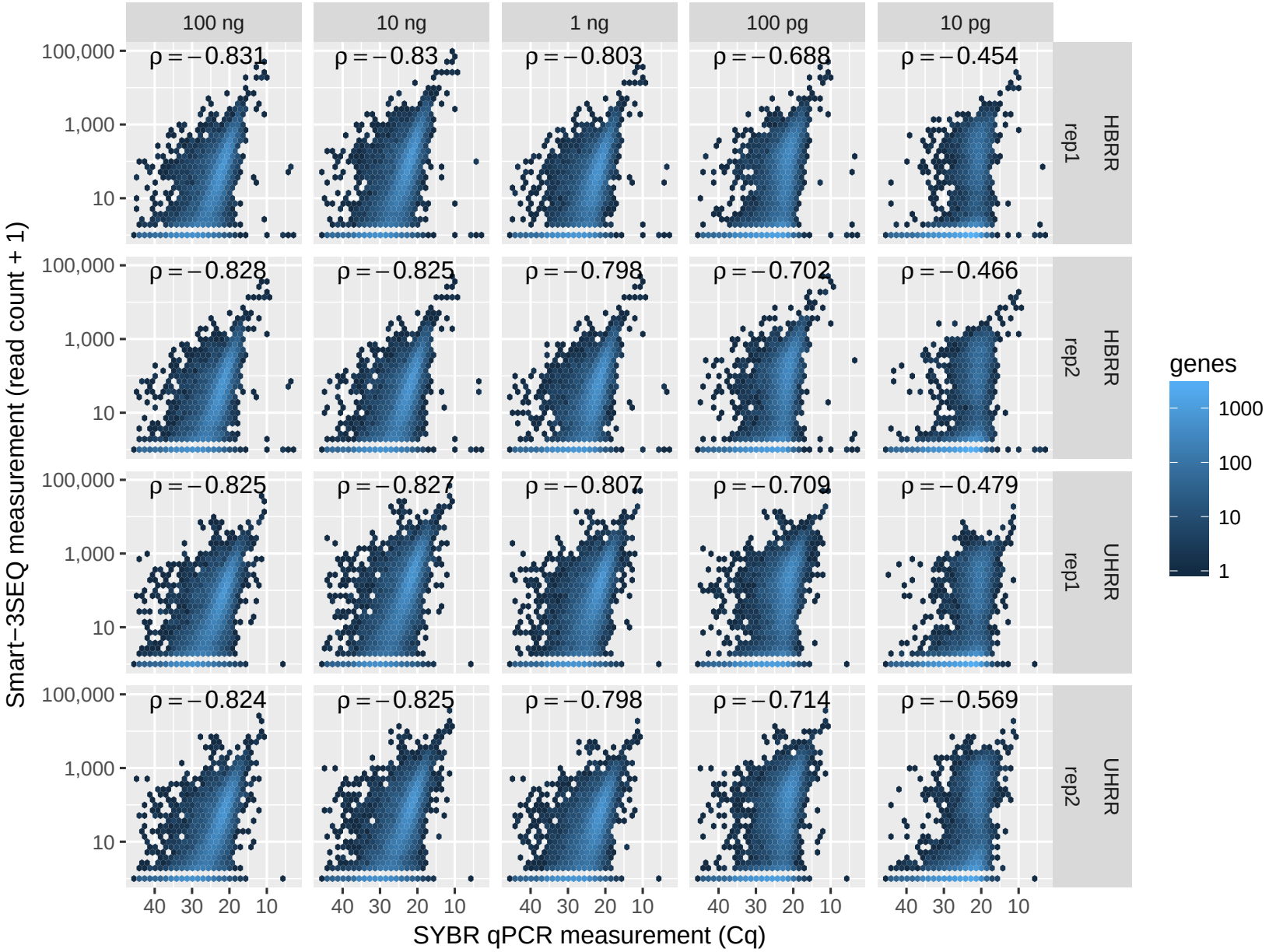


Figure S11: Correspondence of Smart-3SEQ results with SYBR qPCR measurements. Hexagonal bins are colored, on a logarithmic scale, by the number of genes among those with available SYBR qPCR data.

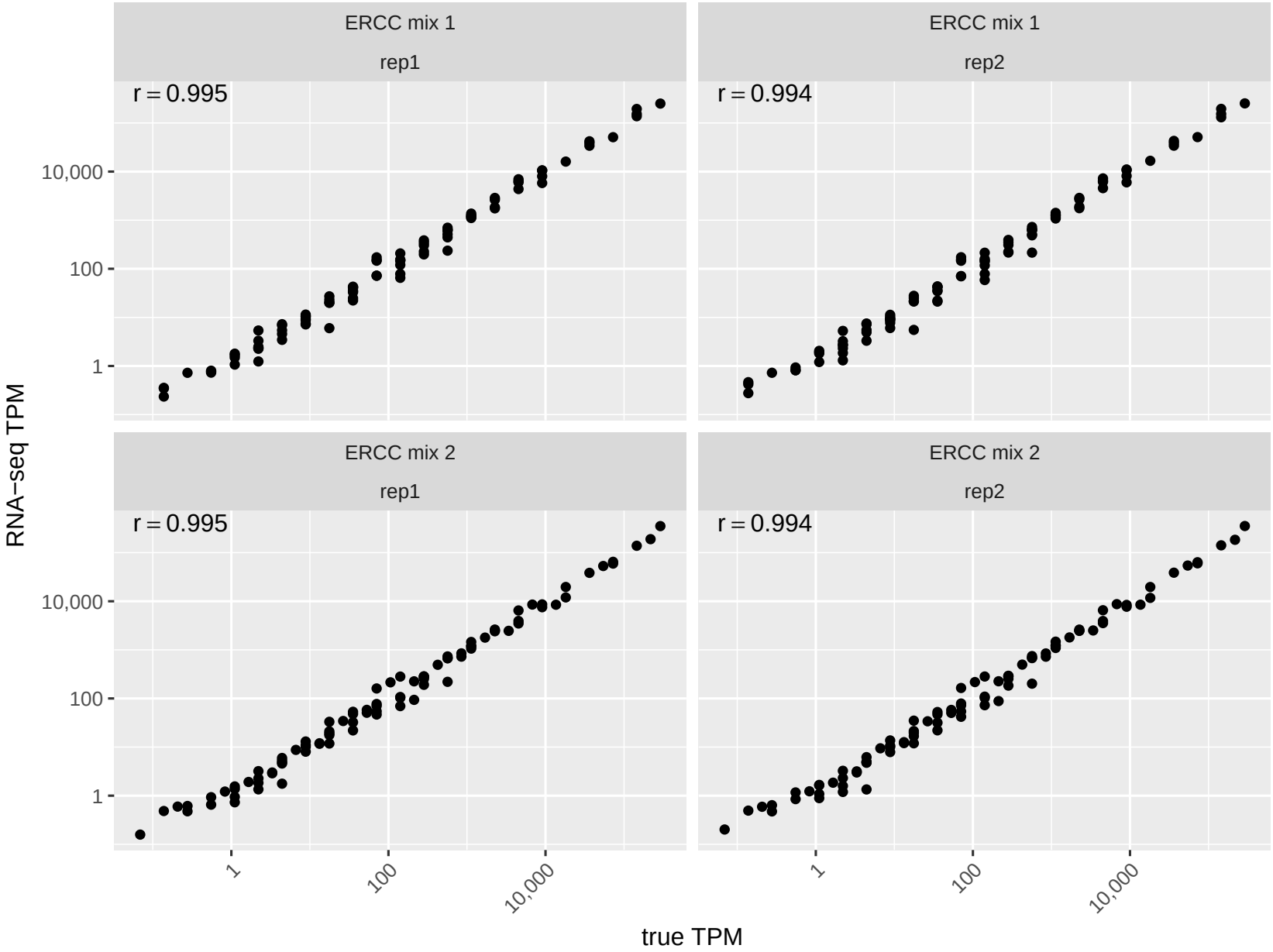


Figure S12: Accuracy of RNA-seq on ERCC standards.

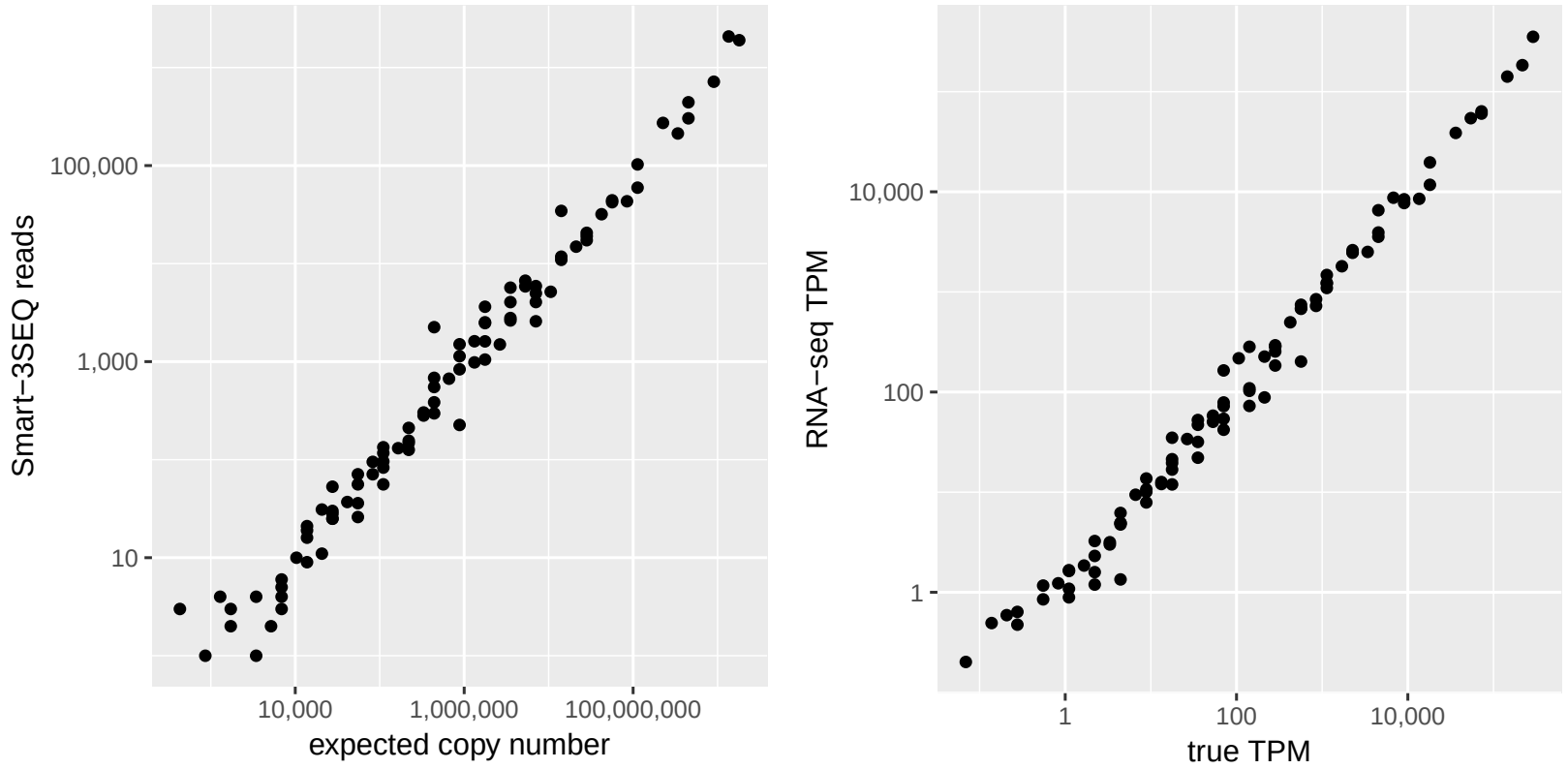


Figure S13: Comparative accuracy of Smart-3SEQ and RNA-seq on ERCC standards using the libraries with the greatest sequencing depth. Smart-3SEQ $r' = 0.990$, RNA-seq $r = 0.994$.

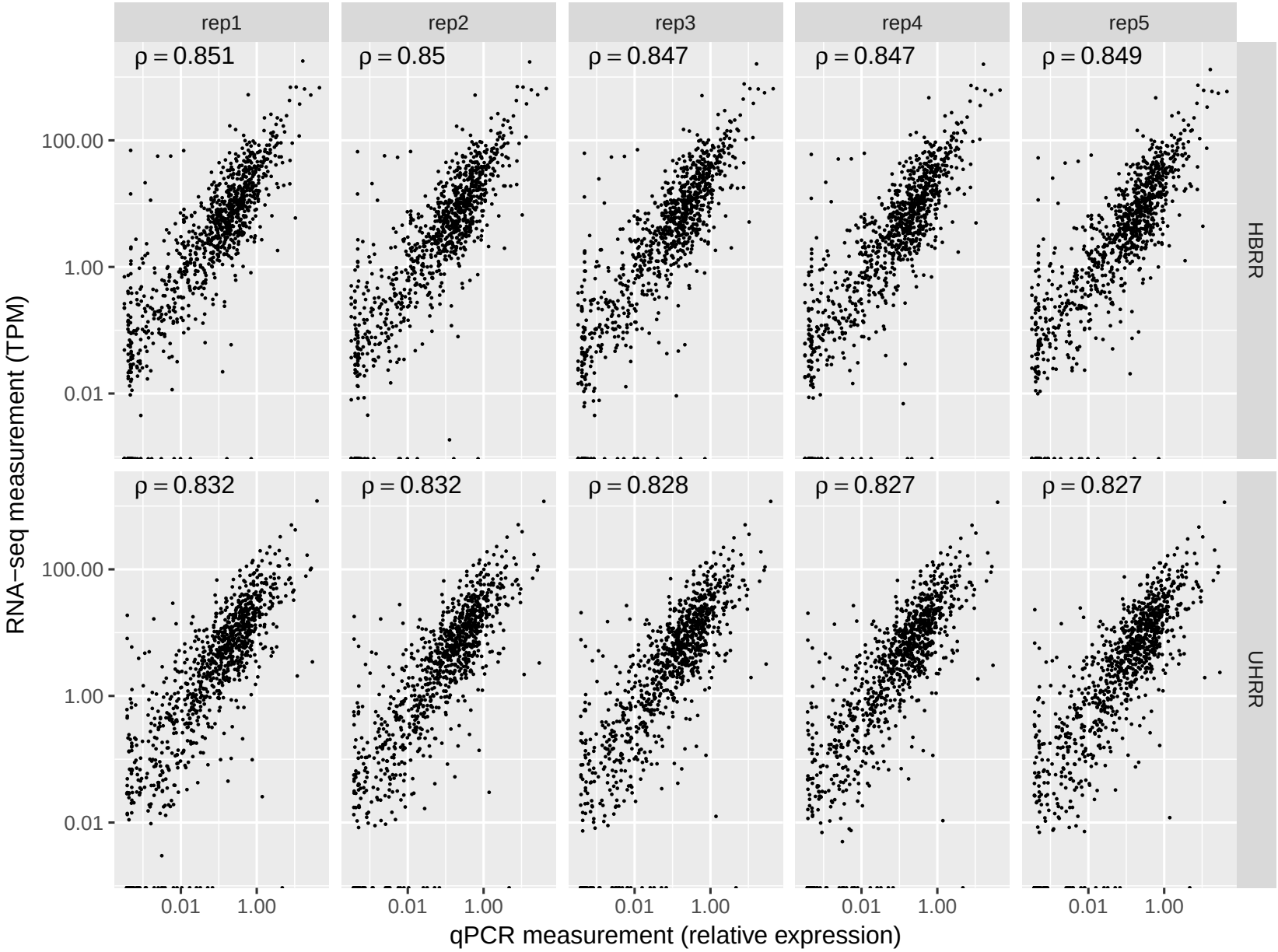


Figure S14: Correspondence of RNA-seq results with TaqMan qPCR measurements.

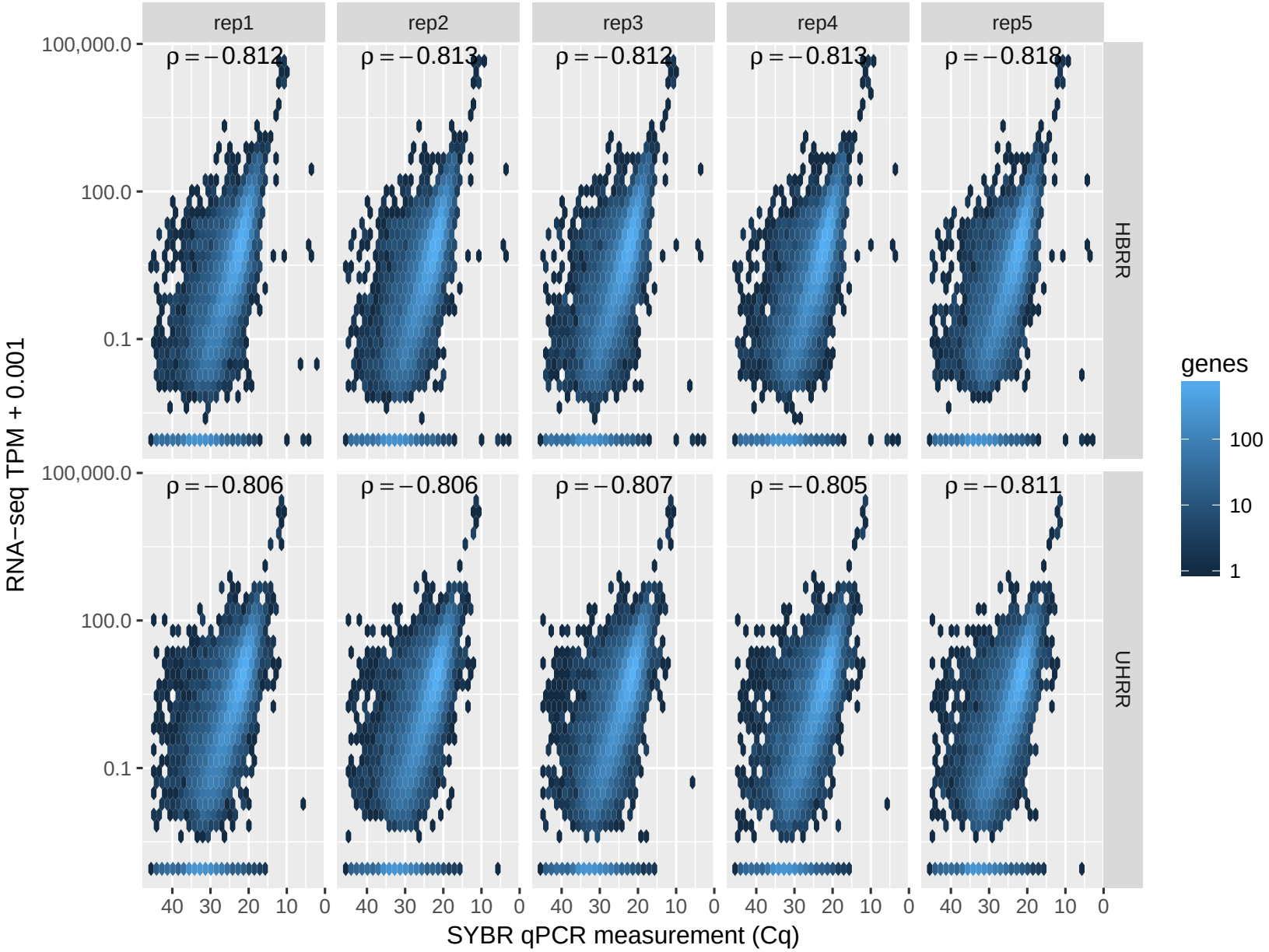


Figure S15: Correspondence of RNA-seq results with SYBR qPCR measurements.

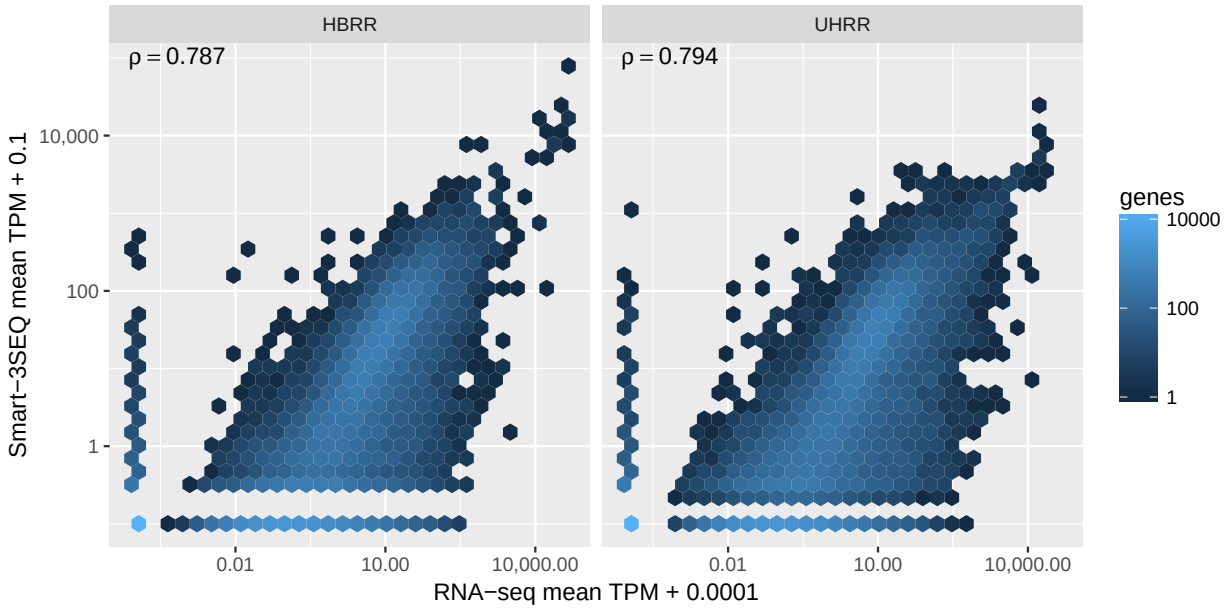


Figure S16: Correspondence of Smart-3SEQ results with RNA-seq results (means of all replicates). All annotated genes are shown.

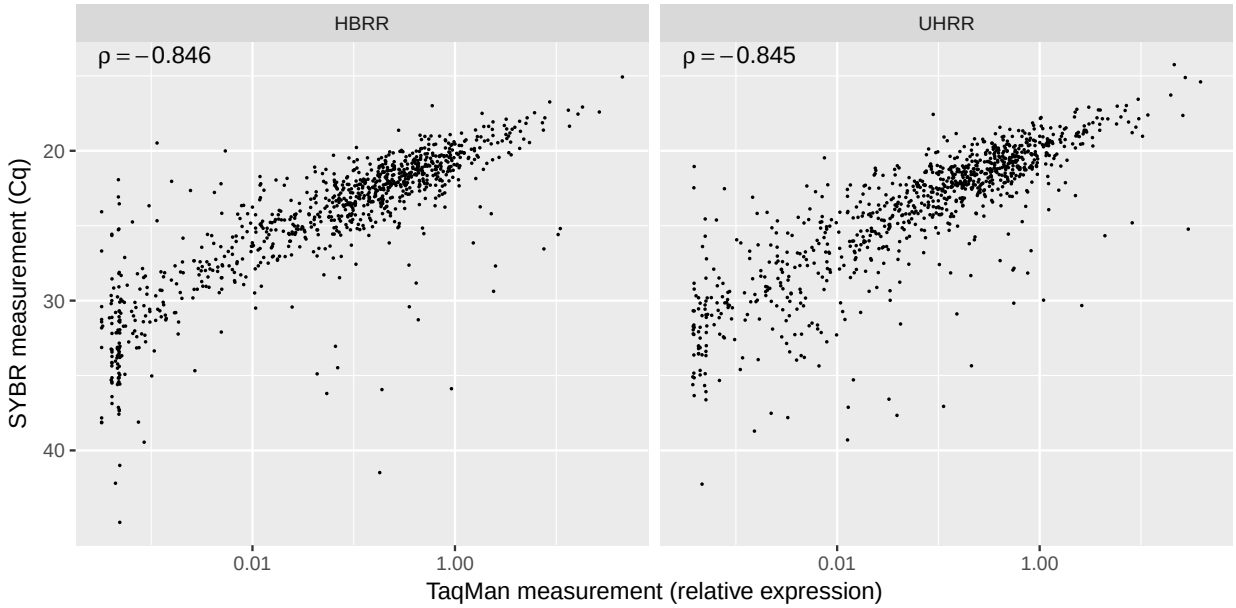


Figure S17: Correspondence of the two different qPCR measurements. Each point is a single gene with available data from both qPCR platforms.

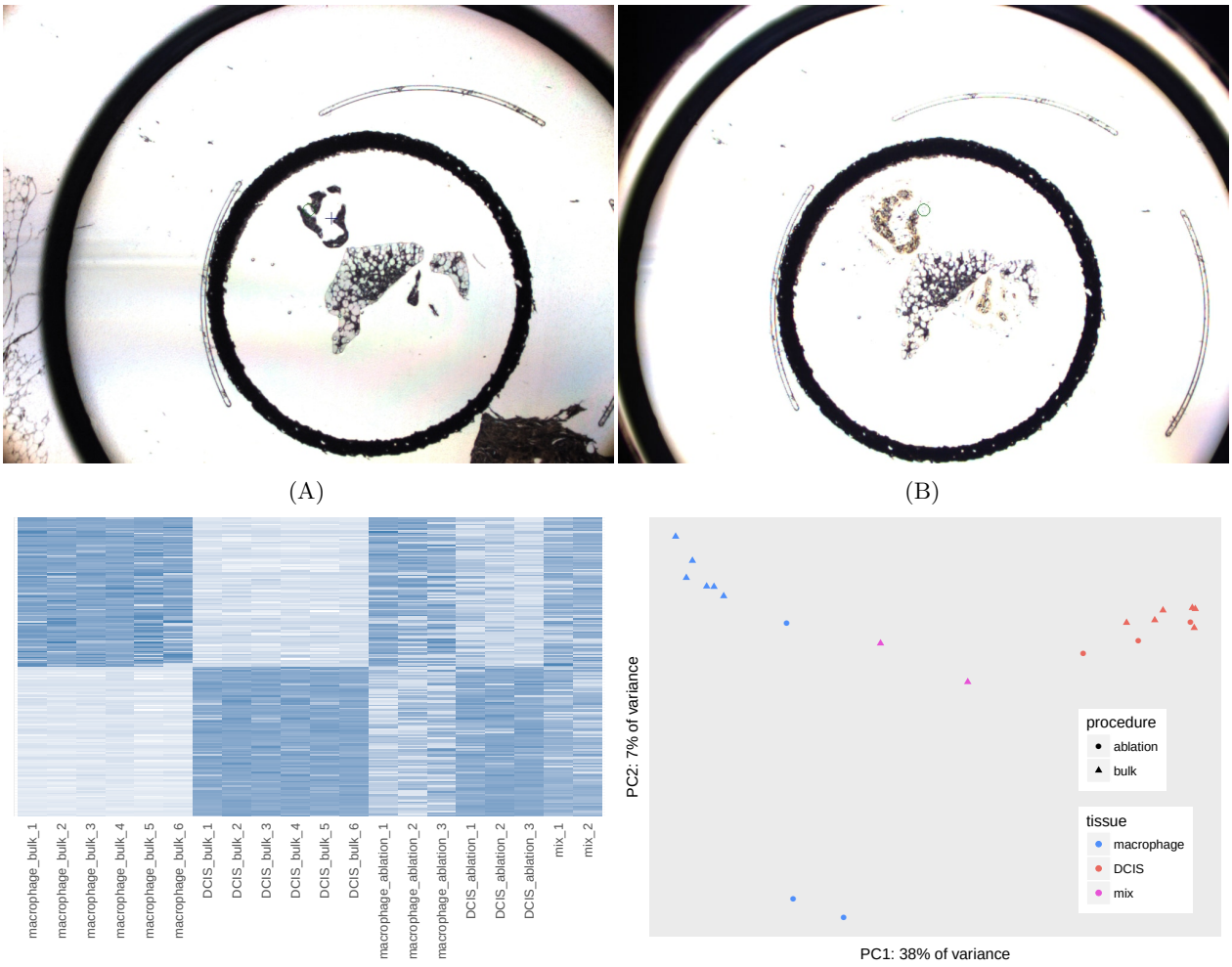
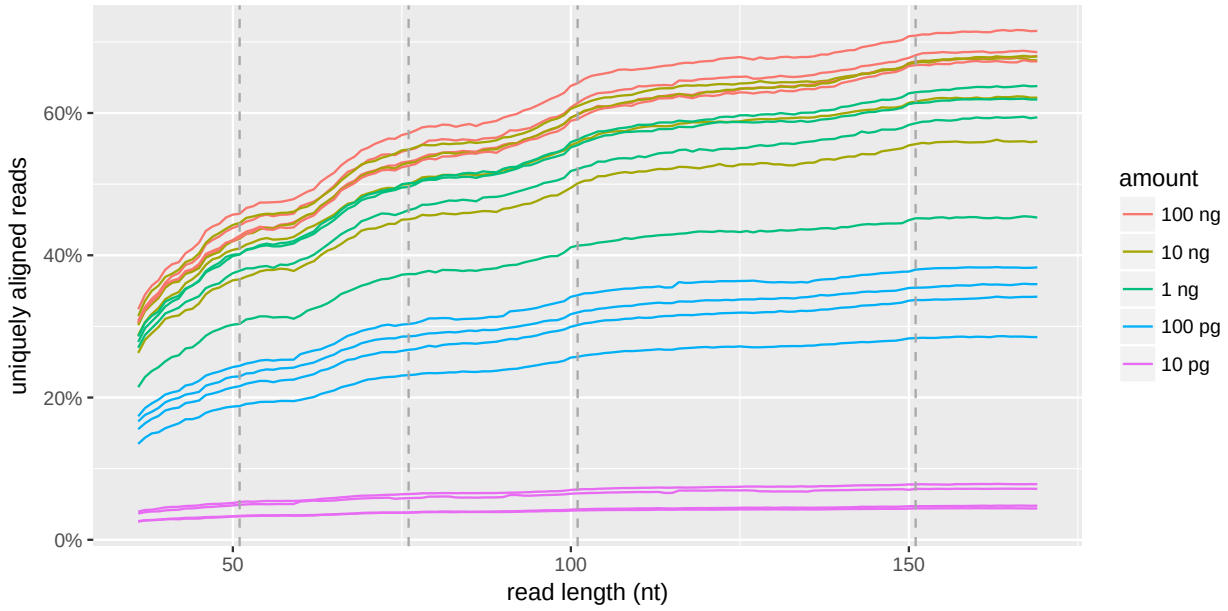
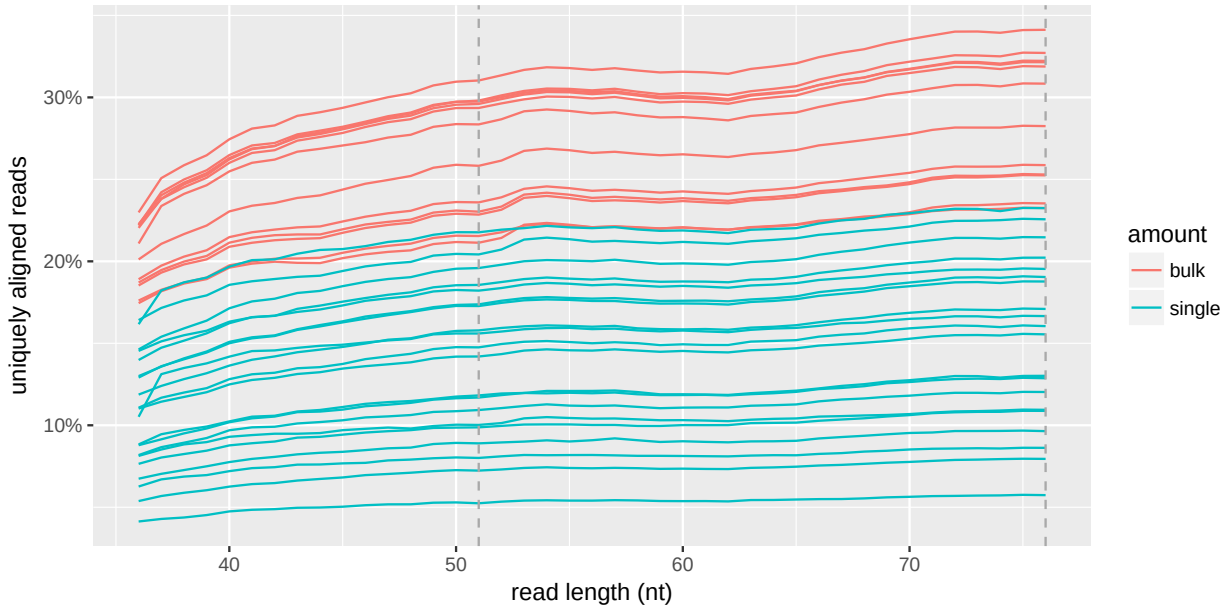


Figure S18: Validation of the laser ablation method. A: Example of mixed DCIS regions and macrophage regions on the same LCM cap. B: DCIS tissue ablated by destruction with the UV laser. C: Expression (regularized log read count, normalized by row) of the 100 genes with the greatest enrichment in bulk macrophage relative to bulk DCIS and the 100 genes with the opposite enrichment, all significant at $p_{adj} < 0.05$. D: Principal components analysis of all genes.

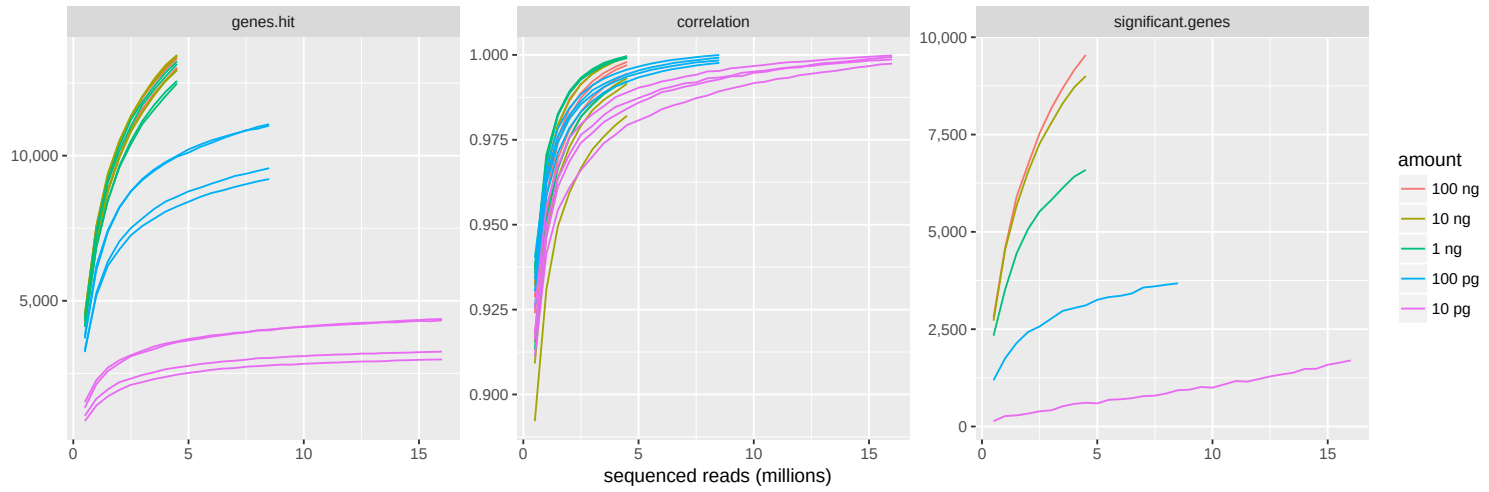


(A)

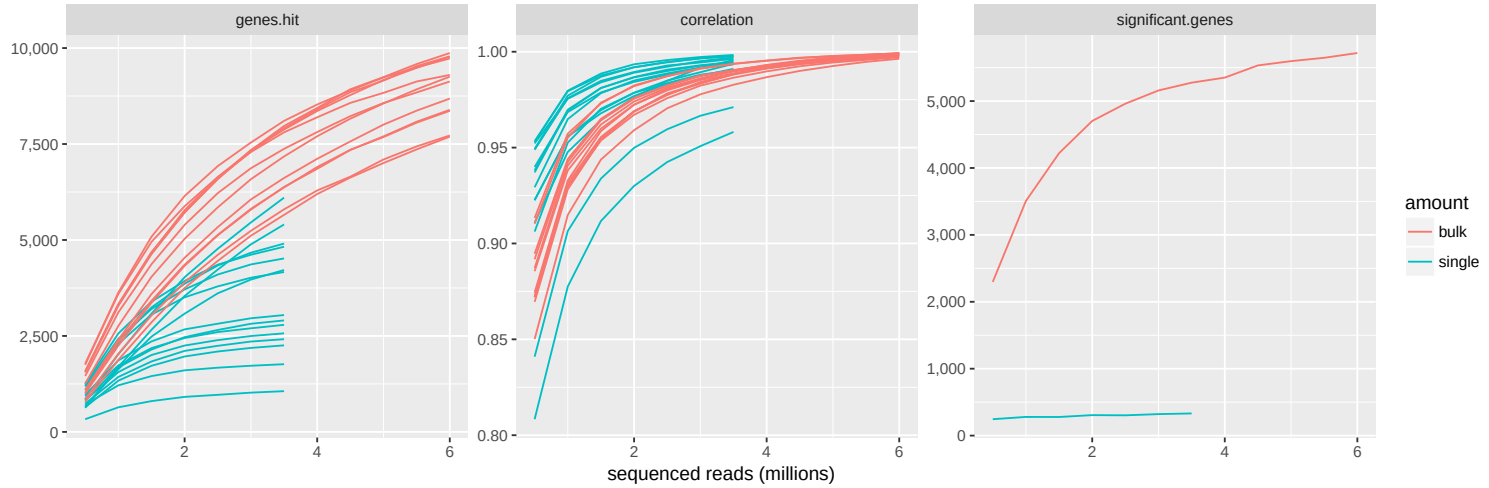


(B)

Figure S19: Diminishing returns from increased read lengths. Results are simulated by truncating reads to the specified length and rerunning the alignment pipeline. Each line traces results from a single library. Commonly used lengths are noted with dashed lines. A: Dilution series of human reference RNAs (high-quality RNA; long fragments), sequenced with long reads on an Illumina MiSeq. B: Bulk tissue and single cells from LCM on FFPE tissue (degraded RNA; short fragments).



(A)



(B)

Figure S20: Diminishing returns from increased sequencing depth. Results are simulated by binomial subsampling of the gene-aligned read counts. `genes.hit`: number of genes with at least 10 reads aligned. `correlation`: Pearson correlation of gene-expression values (as $\log_{10}(c+1)$ for read count c) between subsampled data and original. `significant.genes`: number of genes with $p_{\text{adjusted}} < 0.05$ for significant differential expression between biological categories. A: Dilution series of human reference RNAs (HBRR vs. UHRR); 2 library replicates per condition at each dilution. B: Bulk tissue and single cells from LCM on FFPE tissue (macrophage vs. DCIS; single DCIS-labeled cells lacking *ERBB2* amplification not included); 6 dissection replicates per condition for bulk and 10 vs. 5 single cells.