Heather D. created dUTP library of the Stratagene mouse RNA with the ERCC spike-in controls from Ambion per the CSH/ENCOD protocol.

Sequenced 2x50 on our trial Illumina run; run without qPCR so we guessed at input based on BA. One whole flow cell generated 7,505,102 paired-end reads; 15,010,204 total reads.

Analysis directory: /home/scott/data/dutp

Mapping: results
1) Mapped reads paired-end with Shrimp 2.2.1 to NCBI 37v64 cDNA + ERCC - 1,415,972 were aligned, 5,297,666 were in the unaligned fastq output.
2) Mapped reads single-end after deleting the first 15 bp (so 35 bp reads) with Shrimp 2.2.1 to NCBI 37v64 cDNA + ERCC - 1,515,008 aligned, 3,341,392 unaligned
3) Mapped reads single-end with BWA to NCBI 37v64 cDNA + ERCC - 374,027 of read 1 aligned, 210,499 of read 2 aligned.
4) Mapped all reads, single-end, with BWA to NCBI 37 (mm9) + ERCC - 1,367,099 of read 1, 722,843 of read 2.
5) Mapped all reads, single-end after deleting the first 15 bp (so 35 bp reads) with BWA to NCBI 37 (mm9) + ERCC - 1,431,590 of read 1, 1,867,737 read 2
6) Mapped all reads NOT mapped in 2) with BWA to NCBI 37 (mm9) + ERCC - of the 3,341,392 reads unaligned to cDNA, 1,771,120 were aligned and 1,570,271 unaligned.

Commands:

1) grep -v '^\@' gmapper.cdna.paired.out | awk '{print $1"_"substr ($10,1,10)}' | sort | uniq | wc -l; wc -l gmapper.cdna.unaligned.fastq
2) wc -l gmapper.cdna.35.aligned.fastq; wc -l gmapper.cdna.35.unaligned.fastq
3) cat r1.cdna.sam | awk '{if ($3!="*"{print $1"_"substr($10,1,10)}}' | sort | uniq | wc -l
4) cat r1.gdna.sam | awk '{if ($3!="*"{print $1"_"substr($10,1,10)}}' | sort | uniq | wc -l
5) cat r1.35.sam | awk '{if ($3!="*"{print $1"_"substr($10,1,10)}}' | sort | uniq | wc -l
6) cat unaligned.cdna.35.sam | awk '{if ($3!="*"{print $1"_"substr($10,1,10)}}' | sort | uniq | wc -l; wc -l unaligned.cdna.gdna.35.fastq