LEXOGEN

Enabling complete transcriptome sequencing

QuantSeq 3′ mRNA sequencing for RNA quantification

QuantSeq provides an easy protocol to generate highly strand-specific Next-Generation Sequencing (NGS) libraries close to the 3′ end of polyadenylated RNAs within 4.5 h. Only one fragment per transcript is generated, directly linking the number of reads mapping to a gene to its expression. QuantSeq reduces data analysis time and enables a higher level of multiplexing per run. QuantSeq is the RNA sample preparation method of choice for accurate and affordable gene expression measurement.

With the rapid development of NGS technologies, RNA-Seq has become the new standard for transcriptome analysis. Although the price per base has been substantially reduced, sample preparation, sequencing, and data processing are major cost factors in high-throughput screenings. QuantSeq reduces the expenditures in these areas.

Sample Preparation. QuantSeq is a fast and easy protocol that generates NGS libraries of sequences close to the 3′ end of polyadenylated RNAs within 4.5 h with just 2 h of hands-on time. The kit requires only 0.1 – 2000 ng of total RNA input without the need for poly(A) enrichment or ribosomal RNA depletion. Because of its focus on the 3′ end, QuantSeq is also highly suitable for formalin-fixed, paraffin-embedded (FFPE) samples.

Sequencing. QuantSeq generates only one fragment per transcript, and the number of reads mapped to a given gene is proportional to its expression. No complicated coverage-based quantification is required. Fewer reads are necessary for determining unambiguous gene-expression values, allowing a higher level of multiplexing.

Data Processing. Most sequences will originate from the last exon and the 3′ untranslated region (3′ UTR) containing only few splice junctions, dramatically reducing mapping time (6 samples in 35 min; for details see experiment below). QuantSeq's high strand specificity (>99.9 %) enables the discovery and quantification of antisense transcripts and overlapping genes.

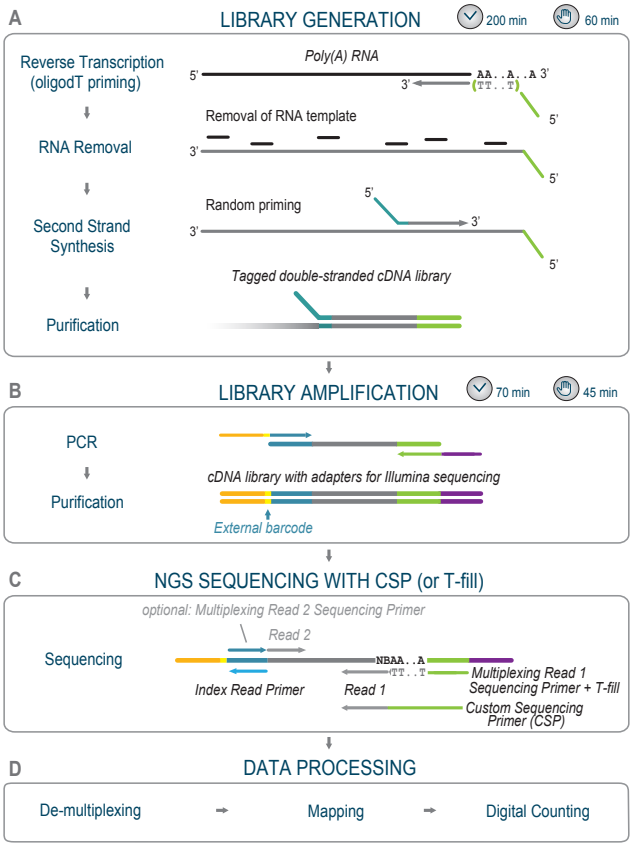


Figure 1 | The QuantSeq REV workflow. (a) Library generation and (b) amplification. (c) Sequencing using a Custom Sequencing Primer (CSP, included in QuantSeq 016) or T-fill reaction, both represented by T*. (d) Data processing.

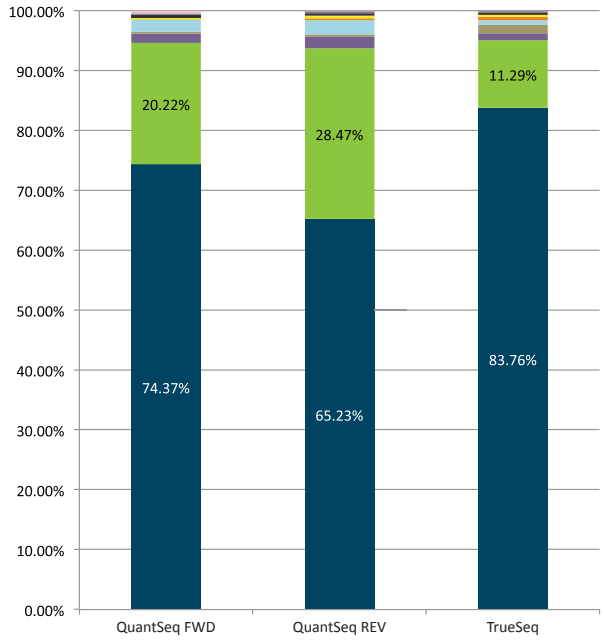
The QuantSeq Workflow

Library generation is initiated by oligodT priming (Fig. 1a), and no prior poly(A) enrichment or ribosomal RNA depletion is required. First-strand synthesis and RNA removal is followed by random-primed synthesis of the complementary strand (second strand synthesis). Illumina- or IonTorrent-specific linker sequences are introduced by the primers. The resulting double-stranded cDNA is purified with magnetic beads, rendering the protocol compatible with automation. Library PCR amplification then introduces the complete sequences required for cluster generation (Fig. 1b). Illumina libraries can be multiplexed with up to 96 external barcodes and are compatible with both single-end and paired-end sequencing reagents. The insert size is optimized for short reads (e.g., SR50 or SR100) while maintaining suitability for longer read lengths, however in-protocol options allow to adapt the library size also for longer read lengths. IonTorrent libraries can be multiplexed using 48 in-line barcodes.

Table 1 | Mapping statistics. Values depicted are averages from triplicates and given in percentage of all reads and percentage of uniquely mapping reads.

	QuantSeq FWD	QuantSeq REV		mRNA-Seq	
	A^*_{1-2}	A_{1-3}	B_{1-3}	A_{1-3}	B_{1-3}
total reads	6,181,833	21,938,757	12,829,269	10,069,397	11,902,438
% mapping reads	87.7 %	91.0 %	87.8 %	95.7 %	96.7 %
% uniquely mapping reads	74.6 %	57.2 % [a]	59.8 % [a]	86.4 %	89.1 %
% ERCCs	1.5 %	4.2 % [b]	3.9 % [b]	0.7 %	1.0 %
% Strandedness [c]	99.9 %	99.9 %	99.9 %	93.4 %	97.8 %

A^*_{1-2}: Universal Human Reference RNA + ERCC RNA Spike-In Mix 1 prepared in house, A_{1-3}: SEQC mixture of Universal Human Reference RNA (UHRR) and External RNA Controls Consortium (ERCC) ExFold Spike-In Mix 1. B_{1-3}: SEQC mixture of Human Brain Reference RNA (HBRR) and ExFold Spike-In Mix 2. [a] Common sequence motifs of the polyadenylation signal and the upstream sequence element limit the variability in the 3' region, thereby reducing the number of uniquely assignable reads in QuantSeq REV. [b] For further analysis, the number of ERCC reads was down-sampled to the common absolute denominator of 0.7 % and 1.0 % as seen in mRNA-Seq A_{1-3} and B_{1-3}. [c] Strandedness was calculated on ERCC reads.



Classes Based On Uniquely Mapping Reads Sample A	QuantSeq FWD	QuantSeq REV	TrueSeq
protein_coding	74.37 %	65.23 %	83.76 %
no_feature	20.22 %	28.47 %	11.29 %
Mt_rRNA	1.58 %	2.01 %	1.18 %
ambiguous	0.34 %	0.31 %	1.42 %
lincRNA	2.04 %	2.41 %	0.76 %
pseudogene	0.0007 %	0.32 %	0.55 %
processed_transcript	0.33 %	0.40 %	0.35 %
antisense	0.40 %	0.42 %	0.39 %
sense_overlapping	0.03 %	0.04 %	0.05 %
Mt_tRNA	0.08 %	0.19 %	0.10 %
sense_intronic	0.08 %	0.09 %	0.07 %
3prime_overlapping_ncrna	0.0020 %	0.0038 %	0.0032 %
rRNA	0.0009 %	0.0097 %	0.0012 %
others*	0.519 %	0.099 %	0.069%

Figure 2 | Gene and transcript biotypes. Uniquely mapped reads from QuantSeq FWD, QuantSeq REV, and mRNA-Seq libraries were assigned to biotypes based on the Ensembl annotation. * Includes miRNA, non-coding RNA, snRNA, snoRNA, IG and TR genes, and ncRNA-related pseudogenes.

QuantSeq is available in two editions with different read orientations. QuantSeq Forward (FWD, Cat. No. 015 for Illumina and Cat. No. 012 for IonTorrent), generates reads toward the poly(A) tail that correspond to the mRNA sequence during Read 1 sequencing. Longer reads may be required if the exact 3' end of the mRNA is of particular interest. QuantSeq Reverse (REV, Cat. No. 016 for Illumina only), generates reads corresponding to the cDNA sequence during Read 1 sequencing (Fig. 1c). Here, a Custom Sequencing Primer (CSP, included in the kit) is used that covers the oligodT stretch to achieve cluster calling on Illumina sequencers, which require a random base distribution within the first sequenced bases. Alternatively, a T-fill reaction can be carried out[1].

Comparison Between QuantSeq and Standard mRNA Sequencing

QuantSeq enables upscaling in multiplexing RNA-Seq experiments, rendering it highly suitable for differential gene expression analysis. Here we present a comparison between QuantSeq and a standard mRNA-Seq protocol, focusing on differential gene expression metrics. We performed QuantSeq REV library preparations (Cat. No. 016.24) on U.S. Food and Drug Administration (FDA) Sequencing Quality Control (SEQC) standard samples A and B in technical triplicates. Sample A is a mixture of Universal Human Reference RNA (UHRR) and External RNA Controls Consortium (ERCC) ExFold Spike-In Mix 1. Sample B is a mixture of Human Brain Reference RNA (HBRR) and ExFold Spike-In Mix 2 (we received SEQC samples A and B from the FDA prepared according to the FDA/National Center for Toxicological Research SEQC RNA Sample Preparation and Testing SOP_20110804). After T-fill, these 6 libraries, referred to as QuantSeq REV A_{1-3} and B_{1-3}, were sequenced in one Illumina HiSeq 2000 lane yielding 150 M single reads of 50 bp (SR50). Residual adapter sequences were removed, and the trimmed pass-filter reads were down-sampled to 10 M each to be comparable with an mRNA-Seq NGS experiment derived from the identical RNA input material. The mRNA-Seq data sets were made available by a laboratory that participated in the recently published Association of Biomolecular Resource Facilities (ABRF) NGS study[2]. In that study, the researchers performed a stranded RNA-Seq library preparation with poly(A) enrichment in 2 technical triplicates, obtaining 50 bp paired-end (PE50) reads on an Illumina HiSeq 2000 (ref. 2; from the GSE48035 data set samples SRR903178–80 from GSM1166109 and SRR903210–12 from GSM1166113 were used in this comparison). We discarded Read 2 in those 6 data sets, referred to as mRNA-Seq A_{1-3} and B_{1-3}, to obtain single-read data comparable to the QuantSeq REV data.

We pooled the 6 mRNA-Seq data sets and aligned them to the GRCh 37.73 genome assembly including ERCC sequences using a splice-junction mapper, TopHat2, which required 2 h 50 min. In contrast, the pooled 6 QuantSeq REV data sets were aligned in only 35 min using the short read aligner Bowtie2 on the same computer system. For gene expression quantification, standard mRNA-Seq relies on length normalization of the number of fragments per kilobase of exon per million fragments (FPKM) mapped, which depends on the correctness of read-to-transcript assignments carried out by Cufflinks. As QuantSeq generates only one fragment per transcript, length normalization is not required, and gene expression quantification is read-count based (Fig. 1d). Mapped reads were further categorized with HTSeq-count (Fig. 2). QuantSeq FWD data was generated from in house prepared UHRR reference RNA spiked with ERCC RNA Spike-In Mix 1.

QuantSeq reads map to intergenic "no_feature" regions to a higher degree than mRNA-Seq reads. This is largely due to the incompleteness of gene annotations at the 3' end[4]. In a recent publication, Schwalb and colleagues demonstrated that transcription termination sites (TTS) were located within a termination window that extends from the last annotated polyadenylation site to an "ultimate TTS". This termination window had a median width of ~3300 bp and could be up to 10 kbp wide. On average they detected four additional TTS per mRNA[5]. QuantSeq reads mapping correctly to such unannotated 3' UTR regions and poly(A) sites do not count towards the "protein_coding" category. This effect is less pronounced for QuantSeq FWD reads since they map more proximal in the 3' UTRs, and these regions might be part of an existing gene annotation. QuantSeq REV reads, however, extend from the nucleotide immediately upstream of the poly(A) tail towards the mRNA's 5' end, and this distal mapping site might be downstream of annotated TTS. Correction of an incomplete annotation (also using mRNA-Seq coverage data) or a sensible 3' extension of the gene definition can address this discrepancy between the available annotation and the real, sequenced transcriptome[6]. Incorrect 3' annotations also cause gene expression based on mRNA-Seq data to be calculated wrongly, since they are based on transcript length normalization.

Data sets were evaluated for ERCC spike-in abundances. To allow a direct comparison between mRNA-Seq and QuantSeq REV, all ERCC reads were down-sampled to identical ERCC read numbers. These subsets of ERCC reads were processed with routines embedded in the ERCC dashboard[3]. One major benefit of QuantSeq can be visualized by plotting the relative coverage across the normalized transcript length (Fig. 3). Standard mRNA-Seq distributes reads across the entire length of transcripts with underrepresentation of 3' and 5' ends, whereas QuantSeq covers the very 3' end of transcripts. In fact, for gene expression and differential expression analysis, one read per transcript is sufficient. The additional sequencing space gained by focusing on the 3' end can be used for a higher degree of multiplexing. In the present example, standard mRNA-Seq has a 12.4-fold higher relative sequence coverage (area under the curve (AUC) ratio for all genes (Fig. 3)), which in turn presents the maximal possible reduction in read depth when using QuantSeq while still determining gene expression accurately.

We compared the results from the QuantSeq and mRNA-Seq experiments focusing on differential gene expression[3]. The ability of a method to measure differentials can be evaluated using the predetermined fold changes between ERCC spike-in control mixes 1 and 2. When plotting the true-positive rate versus the false-positive rate, the AUC is a measure for the correct detection of differential gene expression (Fig. 4). The maximum mean AUC value, corresponding to optimal differential detection, is 1. When the number of reads is down-sampled from 10 M to 0.625 M, standard mRNA-Seq obtains mean AUC values of around 0.776 only, whereas QuantSeq maintains very high AUC values of around 0.860, although similar total numbers of ERCC spike-in RNAs were detected by both methods during the course of down-sampling.

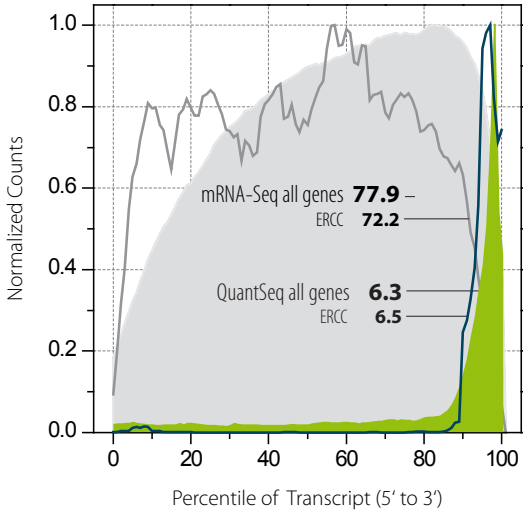


Figure 3 | Coverage versus normalized transcript length in QuantSeq REV and standard mRNA-seq. RSeQC-derived coverage is plotted for all transcripts (areas) and ERCC spike-in control mix only (lines) for QuantSeq (colored) and mRNA-seq (gray). Numbers give the AUC (area under the curve) values as a measure for sequence coverage.

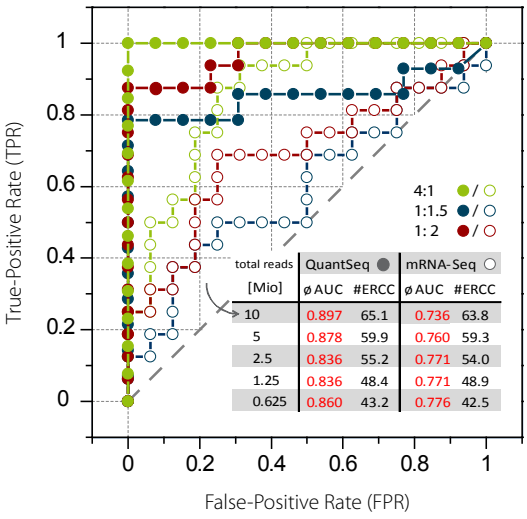


Figure 4 | Differential gene expression performance of QuantSeq REV and mRNA-Seq. The predetermined fold changes (4:1 ●/○, 1:1.5 ●/○, 1:2 ●/○) between ERCC ExFold Spike-In Mix 1 and 2 were used to assess TPRs and FPRs. The receiver operator characteristic– derived AUC value is one measure for the correct detection of differential gene expression. AUC values were assessed together with the number of ERCC RNAs detected (#ERCC) for reads down-sampled from 10 M to 0.625 M. The averaged values of the 6 samples A_{1-3} and B_{1-3} each are presented in the insert table.

In order to quantify QuantSeq's gene count accuracy, the input-output correlation was assessed. Exemplarily, for sample A's three technical replicates, the ERCC spike-in transcripts' average read count was plotted against the input molecules. The numbers of molecules were computed by their concentration and the molecular weight. A linear model in log-log space exhibits a high correlation between the number of reads (output) and the number of input molecules (Fig. 5). Additionally, the Spearman correlation was computed in log-log space, yielding a correlation coefficient of 0.986. For sample B's replicates a R^2 of 0.973 with a slope of 0.983 could be achieved (data not shown); the Spearman correlation coefficient was 0.986.

QuantSeq shows a high input-output correlation for both ERCC ExFold Spike-In Mix 1 and 2 with a very high accuracy in gene expression, in terms of a linear model as well as for Spearman correlation.

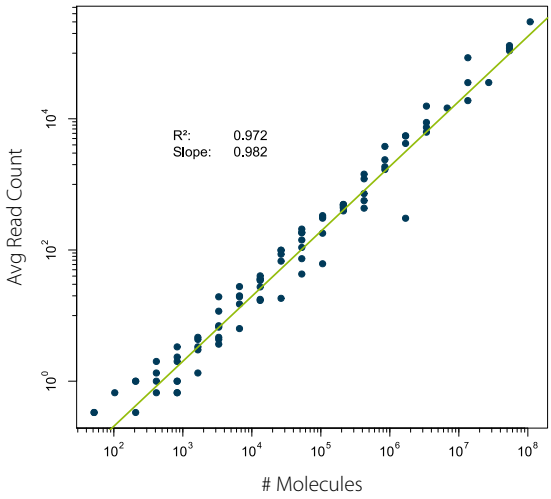


Figure 5 | Input-output correlation of QuantSeq's FDA A samples. The number of molecules was computed according to the ERCC manual, the read counts were assessed with HTSeq-count.

Conclusions

QuantSeq is a robust and simple mRNA sequencing method. It increases the precision in gene expression measurements as only one fragment per transcript is generated. At lower read depths, such focus on the 3' end results in higher stability of differential gene expression measurements. QuantSeq is ideal for increasing the degree of multiplexing in NGS gene expression experiments and is the method of choice for accurately determining gene expression at the lowest cost.

Addendum

QuantSeq is one kit of a series of transcriptome analysis kits provided by Lexogen. For a highly efficient extraction of either total RNA or split fractions of large and small RNA, we offer the SPLIT RNA Extraction kit (Cat. No. 008). RiboCop (Cat. No. 037) is offered for rRNA depletion and a Poly(A) RNA Selection Kit (Cat. No. 039) for mRNA enrichment.

Complementary to the 3' end–focused QuantSeq kit, the SENSE Total RNA-Seq library preparation kit (Cat. No. 009, 042) and the SENSE mRNA-Seq library preparation kit (versions available for Illumina (Cat. No. 001), Ion Torrent (Cat. No. 006), and SOLiD (Cat.No. 004)) provide transcript body–covering RNA-Seq libraries of superior strand specificity (>99.9 %) in less than 5 h.

For targeted sequencing approaches Lexogen offers the QuantSeq-Flex kit (Cat. No. 033, 034, 035) with flexible modules for First Strand and/or Second Strand Synthesis allowing the use of custom primers.

The TeloPrime Full-Length cDNA Amplification Kit (Cat. No. 013) generates full-length cDNA libraries with precisely tagged start and end sites enabling promoter and polyadenylation analysis, splice variant determination, probe generation, etc.

Lexogen now also offers SIRVs (Spike-in RNA Variant Control Mixes, Cat. No. 025.03), which provide for the first time a comprehensive set of transcript variants to validate the performance of isoform-specific RNA-Seq workflows, and to serve as a control and anchor set for the comparison of RNA-Seq experiments.

1. Wilkening, S. et al. An efficient method for genome-wide polyadenylation site mapping and RNA quantification. Nucleic Acids Res. 41, e65 (2013).
2. Li, S. et al. Multi-platform assessment of transcriptome profiling using RNA-seq in the ABRF next-generation sequencing study. Nat. Biotechnol. 32, 915–925 (2014).
3. Munro, S.A. et al. Assessing technical performance in differential gene expression experiments with external spike-in RNA control ratio mixtures. Nat.Commun. 5, 5125 (2014).
4. Schurch N.J. et al. Improved Annotation of 3' Untranslated Regions and Complex Loci by Combination of Strand-Specific Direct RNA Sequencing, RNA-Seq and ESTs. PLoS ONE. 9(4): e94270 (2014)
5. Schwalb, B. et al. TT-seq maps the human transient transcriptome. Science. 352, 1225-1228 (2016)
6. Abasht, B. WEBINAR: Gene Expression Analysis Using 3'-RNA Sequencing (2016). Retrieved from www.labroots.com/ms/webinar/gene-expression-analysis-using-rna-sequencing

Lexogen GmbH · Campus Vienna Biocenter 5 · 1030 Vienna · Austria

Find more about QuantSeq at www.lexogen.com.
Contact us at info@lexogen.com or +43 1 345 1212-41.

QUANT™ SEQ
Sequencing that counts