

QuantSeq 3' mRNA-Seq with integrated automatic data analysis - a complete workflow for user-friendly and cost-efficient gene expression profiling

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 hrs from high and low quality RNA (incl. FFPE). 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. Each QuantSeq kit also includes a complimentary automated differential gene expression data analysis option. QuantSeq is the RNA sample preparation method of choice for accurate and affordable gene expression measurement and is the best alternative to using gene expression arrays.

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's unique method reduces the expenditures in all these areas.

Sample Preparation. QuantSeq is a fast and simple protocol that generates NGS libraries of sequences close to the 3' end of poly(A) RNAs within 4.5 hrs with just 2 hrs 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. QuantSeq's high strand specificity (>99.9 %) enables the discovery and quantification of antisense transcripts and overlapping genes. QuantSeq data analysis pipeline has been automated on the Bluebee® Genomics Analysis Platform, providing a fast and user friendly data processing tool.

The QuantSeq Workflow. Library generation begins with reverse transcription using an oligodT primer (Fig. 1A). Following first strand synthesis, the RNA template is removed and second strand synthesis initiated by random priming. Illumina- or Ion Torrent-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 9,216 different i5 / i7 index combinations compatible with both single-read and paired-end sequencing reagents. The insert size is optimized for short reads (e.g., SR50 or SR100), however in-protocol options allow library size to be adapted for longer read lengths. Ion Torrent libraries can be multiplexed using 48 indices.

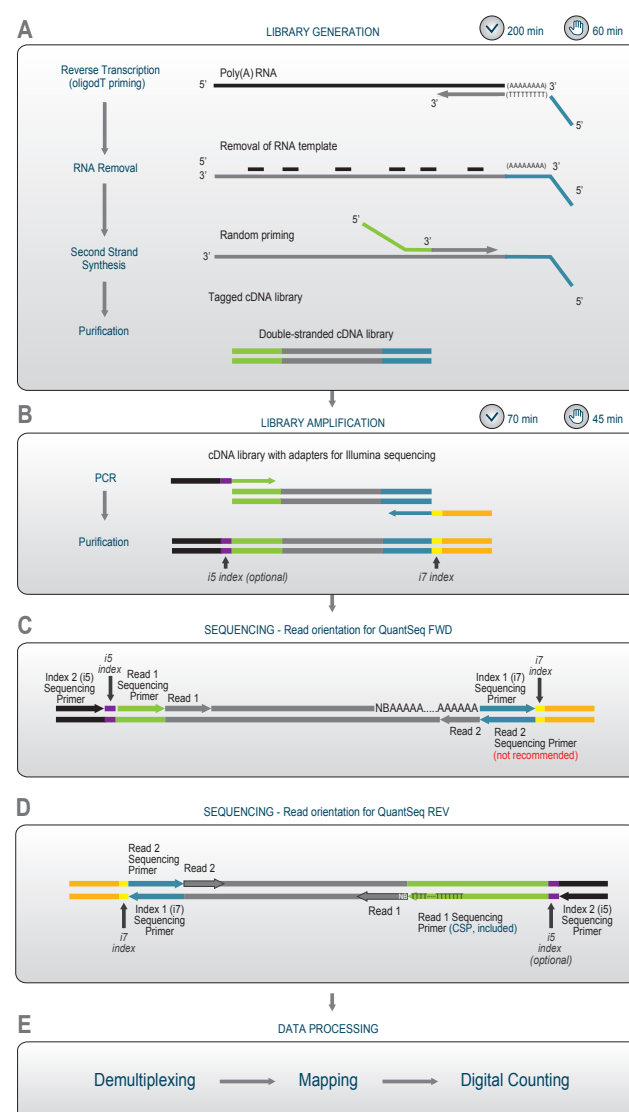


Figure 1 | Schematic overview of the QuantSeq FWD library preparation workflow. For QuantSeq REV the position of adapters for Read 1 (green) and Read 2 (blue) are switched.

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, A₁₋₃: SEQC mixture of Universal Human Reference RNA (UHRR) and External RNA Controls Consortium (ERCC) ExFold Spike-In Mix 1, B₁₋₃: SEQC mixture of Human Brain Reference RNA (HBRR) and ExFold Spike-In Mix 2. ^aCommon 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. ^bFor 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 A₁₋₃ and B₁₋₃. ^cStrandedness was calculated on ERCC reads.

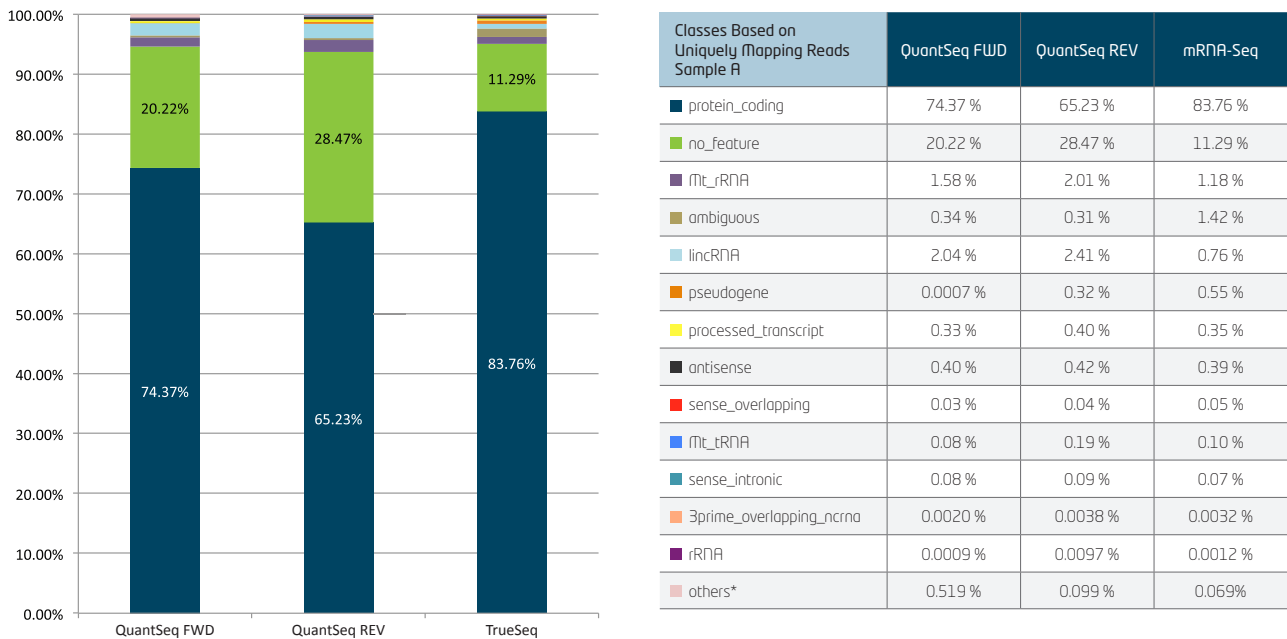


Figure 2 | Gene and transcript biotypes. Uniquely mapped reads from QuantSeq FWD, QuantSeq REV, and mRNA-Seq libraries were assigned to biotypes based on the Ensembl annotation. * Includes miRNA, non-coding RNA, snRNA, snoRNA, IG and TR genes, and ncRNA-related pseudogenes.

QuantSeq is available in two versions 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 (Fig. 1C). 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. 1D). 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 out¹.

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 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 A₁₋₃ and B₁₋₃, 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 a published Association of Biomolecular Resource Facilities (ABRF) NGS study². 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 A₁₋₃ and B₁₋₃, to obtain single-read data comparable to the QuantSeq REV data.

QuantSeq FWD data was generated from in house prepared UHRR reference RNA spiked with ERCC RNA Spike-In Mix 1.

We pooled the 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 QuantSeq 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. As QuantSeq generates only one fragment per transcript, length normalization is not required, and gene expression quantification is read-count based (Fig. 1E). QuantSeq showed comparable mapping statistics to mRNA-Seq data, but with superior strandedness (Tab. 1). Mapped reads were further categorized with HTSeq-count (Fig. 2).

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’ end⁴. 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 mRNA⁵. 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 transcriptome⁶. 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, all ERCC reads were down-sampled to identical ERCC read numbers. These subsets of ERCC reads were processed with routines embedded in the ERCC dashboard³. 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 ex-

periments focusing on differential gene expression³. The ability of a method to measure differentials can be evaluated using the predetermined fold changes between ERCC ExFold RNA Spike-In 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.

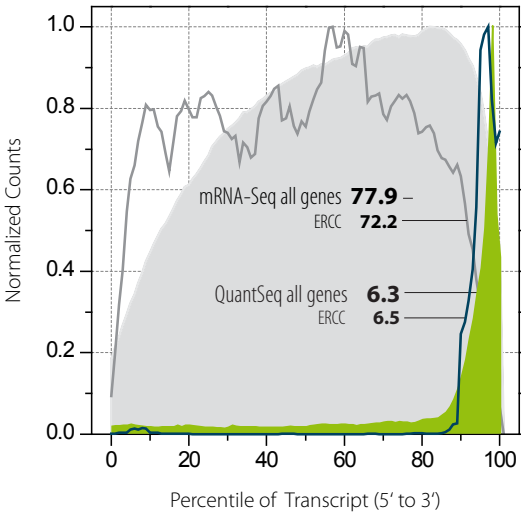


Figure 3 | Coverage versus normalized transcript length in QuantSeq REV and standard mRNA-Seq. RSeQC-derived coverage is plotted for all transcripts (areas) and ERCC spike-in control mix only (lines) for QuantSeq (colored) and mRNA-Seq (gray). Numbers give the AUC (Area Under the Curve) values as a measure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

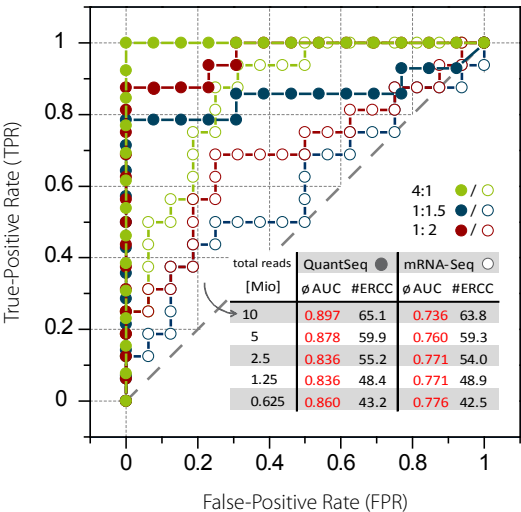


Figure 4 | Differential gene expression performance of QuantSeq REV and mRNA-Seq. The predetermined fold changes (4:1, 1:1.5, 1:2) between ERCC ExFold Spike-In Mix 1 and 2 were used to assess True and False Positive Rates (TPRs and FPRs). Optimal detection of differential gene expression is indicated by a maximum AUC (Area Under the Curve) value of 1. 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 A₁₋₃ and B₁₋₃ each are presented in the insert table.

QuantSeq Yields Good Correlation Between High and Low Quality (FFPE) Samples

The suitability of QuantSeq when using highly degraded samples (e.g., formalin-fixed, paraffin-embedded (FFPE) material) was evaluated by comparing two samples derived from one source but of different RNA qualities.

A xenograft of the MOLP-8 human tumor cell line was split into two pieces, which were subsequently processed either as fresh frozen cryo-block or embedded FFPE material, leading to different RNA qualities from the same original sample. To determine RNA quality often the RIN (RNA Integrity Number) is used, with an RIN of >8 indicating high RNA quality. For heavily degraded samples this is not a sensitive measure, and hence the DV₂₀₀ value (distribution value of RNA fragments >200 nucleotides) should rather be used. Low RNA integrity correlates with low DV₂₀₀ values.

After RNA extraction, the FFPE sample had a DV₂₀₀ of 87 % (RIN of 2.8), while the Cryo sample yielded a RIN of 8.3. The libraries were generated with the QuantSeq 3' mRNA-Seq FWD kit using 50 ng total RNA input. QuantSeq libraries were also successfully generated of FFPE-derived RNA with a DV₂₀₀ of down to 23 % (data not shown). For the FFPE sample the protocol recommendations for low quality RNA input were followed, for the Cryo sample the standard protocol was applied. The libraries were sequenced on a HiSeq® 2500 instrument at 1x 50 bp read length (Fig. 5).

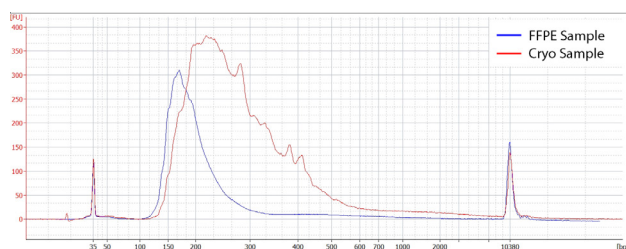


Figure 5 | Bioanalyzer 2100 HS DNA chip traces of QuantSeq 3' mRNA-Seq FWD prepared libraries using FFPE (blue) or Cryo RNA input (red). For the degraded input, the resulting library shows a smooth size distributing with no visible linker-linker by-products but only a shift towards shorter fragments. Average library size is 204 bp (FFPE) and 286 bp, respectively.

Plotting the relative coverage across the normalized transcript length shows that coverage is focused on the transcripts' 3' end, independent of the input RNA quality (Fig. 6). However, as QuantSeq FWD libraries are sequenced towards the poly(A) site coverage is dependent on library size and sequencing length.

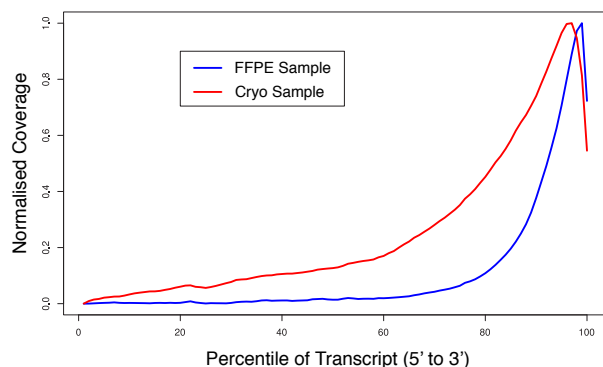


Figure 6 | QuantSeq read coverage versus normalized transcript length of NGS libraries derived from FFPE RNA (blue) and cryo-preserved RNA (red).

The FFPE libraries were significantly shorter than the Cryo sample, therefore the ends of the transcripts were reached more frequently, which is reflected in the coverage plots.

Gene expression correlation between libraries derived from FFPE and cryo-preserved RNA is high (R^2 0.86) and indicates that QuantSeq performs consistently well on samples of different RNA quality (Fig. 7).

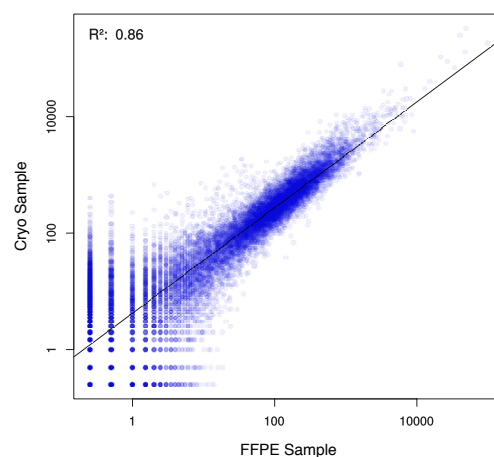


Figure 7 | Correlation of gene counts of FFPE and cryo samples.

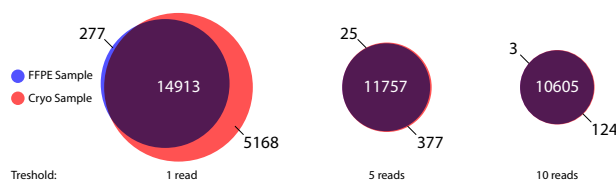


Figure 8 | Venn diagrams of genes detected by QuantSeq at a uniform read depth of 2.5 M reads in FFPE and cryo samples with 1, 5, and 10 reads/gene thresholds.

QuantSeq has a very high sensitivity, at 26.5 M reads, 25,842 genes were detected in the intact Cryo RNA sample (data not shown). At uniformly 2.5 M reads, 20,081 genes were detected in the Cryo sample with at least 1 read, compared to 15,190 genes in the FFPE samples, which represents a 24 % difference (Fig. 8). However, increasing the detection level to 5 or 10 reads / gene reduces the difference to 3 % and 1 %, respectively. These alignments show that QuantSeq reliably detects gene expression in both cryo-preserved and FFPE samples; the difference in detection of lowly expressed genes is due to the higher susceptibility of their low copy transcripts to degradation during FFPE treatment, storage, and recovery.

QuantSeq is based on oligodT priming during the reverse transcription and only generates one fragment per transcript. This enables accurate gene expression quantification independent of the RNA quality (including FFPE samples). Standard mRNA-Seq protocols aim to cover the whole transcript, but will result in a heavy 3' bias when used on degraded RNA input. Therefore, QuantSeq 3' mRNA-Seq is an efficient tool to generate NGS libraries from low quality samples compared to other mRNA-Seq protocols using poly(A) selection.

Automated Data Analysis

Free access to the QuantSeq data analysis pipeline hosted by the Bluebee® Genomics Analysis Platform is included with all QuantSeq 3' mRNA-Seq Library Prep Kits (FWD and REV) for Illumina. It enables fully automated data analysis without bioinformatics involvement. After uploading the sample sequencing data, the analysis will take less than 10 minutes/sample (batch processing is possible). The pipeline performs trimming, quality control, alignment, and read counting from fastqc input (see Fig. 9). Report I contains the mapping statistics and read counts, which can be further processed using the differential expression (DE) pipeline. Report II results in a table with differentially expressed genes (ENSEMBL IDs) and the corresponding DE values as well as a heat map of the top 20 expressed genes and genes with the highest absolute log fold change, an MA plot, and a Principal Component Analysis (PCA) plot. All results including aligned read .bam files are available for download.

In addition, the differential expression analysis pipeline can be run up to 500 times with each code, enabling fully-flexible, pair-wise comparisons to identify differentially-expressed genes.

QuantSeq data analysis is also available on Partek® Flow® (license required). Users can import raw sequencing files directly into Partek® Flow® and automate their analysis. Customizable data analysis options are also available.

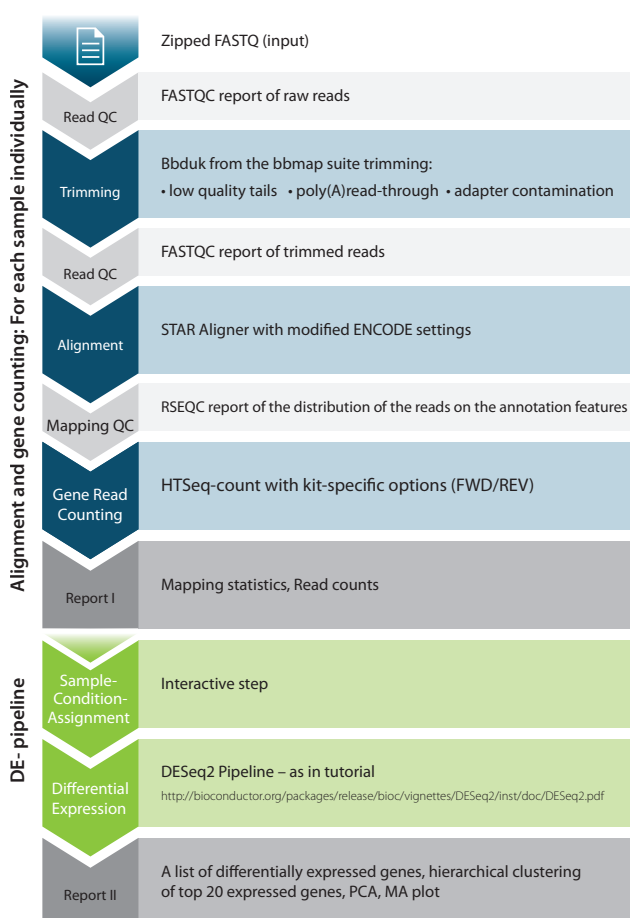


Figure 9 | Automated data analysis pipeline on the Bluebee® Genomic Analysis Platform.

New Applications for QuantSeq

Globin Block - QuantSeq for Blood RNA Without Any Additional Steps

Lexogen's Globin Block Modules (Cat. No. 070, 071) for QuantSeq prevent the generation of library fragments from globin mRNAs (*HBA1*, *HBA2*, *HBB*) during the library prep itself. Lower input amounts starting from 50 ng of total RNA from blood can be used, with no additional pre-processing or protocol steps required. The Globin Blockers bind to globin first strand cDNA and prevent the generation of double-stranded cDNA from globin mRNAs during second strand synthesis. The module is compatible with QuantSeq 3' mRNA-Seq Library Prep Kits for Illumina (FWD, Cat. No. 015, and REV, Cat. No. 016), and is intended for the preparation of libraries from blood RNA samples for human (*Homo sapiens*) and pig (*Sus scrofa*). Contact us at info@lexogen.com for use with other species.

Libraries prepared with Globin Block Modules show significant reduction of total globin mapped read percentages, compared to libraries prepared with standard QuantSeq (Fig. 10). Total globin mapped read percentages dropped to as low as 0.7 % for leukocyte-enriched blood and 9.7 % for whole blood in +Globin Block libraries.

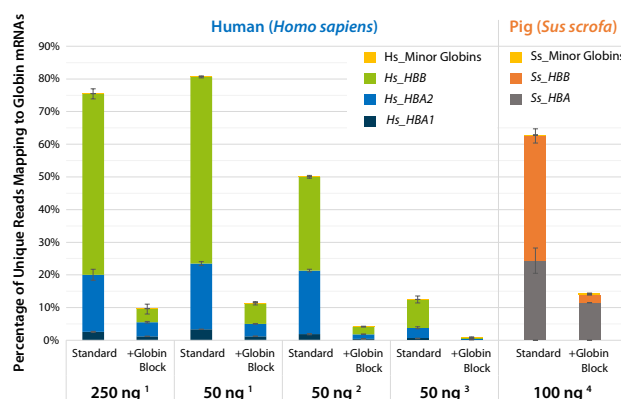


Figure 10 | Percentage of reads uniquely mapping to human and pig globin mRNAs. Libraries were prepared from whole blood RNA with the Standard QuantSeq FWD protocol, versus QuantSeq +Globin Block. ¹ SPLIT RNA Extraction Kit without red blood cell lysis (Lexogen), ² PAXgene® Blood RNA Kit (Qiagen, includes red blood cell lysis), ³ SPLIT RNA Extraction Kit with red blood cell lysis, ⁴ Preserved Blood RNA Purification Kit I (Norgen Biotek, without red blood cell lysis).

Unique Dual Indexing - Enhanced Multiplexing Capacity

The i5 Dual Indexing Add-on Kits for Illumina (Cat. No. 047) contain up to 96 perfectly balanced i5 indices (5001-5096) for dual indexing of QuantSeq and SENSE libraries. Each of the 96 i5 (Index 2) indices can be introduced at the PCR step and combined with any of the 96 i7 indices (Index 1), enabling up to 9,216 different i5/i7 index combinations, or up to 96 unique dual indexing barcoding options. Thus, index hopping can be detected and quantified, while avoiding index mis-assignment.

The online Index Balance Checker tool can be used to determine the optimal combination of indices (i5 and i7) to use for multiplexed sequencing runs (www.lexogen.com/support-tools/index-balance-checker).

Unique Molecular Identifiers (UMI) - Improved Quantification Accuracy

The UMI Second Strand Synthesis Module for QuantSeq FWD (Illumina, Read 1) (Cat. No. 081) contains random primers for second strand synthesis that include 6 nt unique molecular identifiers (UMIs) (see Fig. 11). This tagging of individual transcripts prior to library amplification enables the identification of PCR duplicates and elimination of amplification bias, allowing for unbiased gene expression analysis.

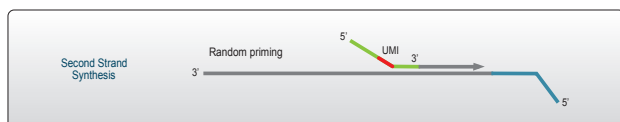


Figure 11 | UMIs are added during the second strand synthesis step of the QuantSeq workflow.

The UMI is read-out at the beginning of Read 1, meaning libraries are ideal for single-read sequencing with read length of 75 bp or longer.

Tools for analysis of UMI diversity and de-duplication of sequencing data using UMIs are available from Lexogen. All the reads with the same genomic location and UMI are collapsed into a single representative read. These unique reads can then be used for accurate estimation of transcript abundance.

QuantSeq Automation - High Throughput Library Prep on Multiple Liquid Handling Platforms

Lexogen offers autoQuantSeq, the automated version of the QuantSeq 3' mRNA-Seq Library Prep protocol. Generate Illumina-compatible 3' mRNA-Seq libraries on an automated platform for high throughput, scalable gene expression profiling with up to 9,216 i5/i7 index combinations for maximal multiplexing capacity. Automation saves hands-on time, maximizes throughput, and avoids pipetting and sample tracking errors.

AutoQuantSeq is currently available for the following platforms: PerkinElmer Sciclone NGS and Zephyr liquid handlers, Hamilton MicroLab STAR Workstations, the Agilent NGS Workstation (Bravo), Eppendorf epMotion 5075, and Beckman Coulter Biomek (FX[®] and i7). For automation of QuantSeq on other liquid handling platforms or for support with setting up your automated protocol, please contact us at info@lexogen.com.

QuantSeq Services

Lexogen provides a fully integrated Service workflow for generation of QuantSeq 3' mRNA-Seq libraries, multiplexed sequencing, and data analysis.



Sample

The QuantSeq Service bundle requires purified total RNA samples as input and is available for starting amounts as low as 100 pg. QuantSeq is suitable for (reproducible) library generation from high, and low quality RNA, including FFPE samples. For a variety of sample types Lexogen also offers RNA isolation using the SPLIT RNA Extraction Kit. Contact us for RNA and sample submission details.



Library Preparation & Automation

The QuantSeq 3' mRNA-Seq Library Prep automation on the best-in-class liquid handler allows for simultaneous preparation of large sample numbers with superb workflow consistency.



Sequencing

High multiplexing capabilities are a native feature of QuantSeq and form the basis for this economically competitive solution. Sequencing of libraries to an average depth of just a few million reads is sufficient to obtain reliable gene expression quantification. Sequencing is routinely performed on an Illumina NextSeq[®] 500.



Data Analysis

A clear-cut bioinformatics analysis solution rounds off the workflow. An automated data analysis pipeline provides streamlined analysis throughput, including demultiplexing, read quality control, trimming and filtering procedures, mapping, read counting, and differential gene expression (DGE).



Report

Data analysis results and raw sequencing data are hosted on our sftp server for convenient download upon completion of the project. In addition, a standardized report for individual samples, including differential gene expression quantification is provided for download. Please contact us for an example report at services@lexogen.com.

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

LISTE DES PRODUITS

NOUS CONTACTER



Service client - commande : commande@ozyne.fr
Service technique :
Réactifs : 01 34 60 60 24 - tech@ozyne.fr
Instrumentation : 01 30 85 92 88 - instrum@ozyne.fr

