Spike-In RNA Variant Control Sets

RNA sequencing is used in many research projects as a powerful tool for transcriptome analysis and also holds promise for expanding into diagnostic and therapeutic applications. To be analytically valid, a laboratory test must deliver accurate information in a reproducible and robust manner. Markers and controls are the only way to unambiguously determine the reliability of your experiment, and Lexogen’s Spike-In RNA Variants (SIRVs) enable the evaluation of both RNA-Seq experiment quality and comparability.

**Advantages**

- SIRVs are universal markers for comparing RNA-Seq experiments on all platforms.
- The spike-in transcripts are compatible with RNA from any organism and RNA input down to single-cell level.
- Evaluate only 1% of your data to draw valid conclusions on the whole data set.

**Validation**

- Validate RNA-Seq workflows and experiments by determining accuracy of gene expression and differential expression measurements.
- Identify error sources and biases.
- Improve workflows from library preparation and sequencing, to data evaluation.
- Measure amounts of RNA classes relative to the input.

**Concordance determination**

- Use SIRV quality control “fingerprints” to measure the concordance of RNA-Seq data.
- Monitor the performance of RNA-Seq workflows over time, at different locations, etc.
- Know and quantify the effects of changes in your workflow on RNA-Seq data.
- Come to an informed decision whether results within or between experiments can be compared.

**Workflow**

The SIRVs are sets of artificial transcripts that comprehensively represent transcriptome complexity and are spiked into the samples (homogenized tissue / cells or purified RNA) to obtain a final NGS read share of 1-2%. Data evaluation of the SIRVs provides quantitative values such as accuracy and precision for the validation of RNA-seq pipelines, and for the determination of “technical noise” specific for each processed sample (Fig. 1). The concordance between experiments is then calculated based on these quality measures and the Coefficient of Deviation (CoD).

**Figure 1 | Using SIRV controls in RNA-Seq.** SIRVs are defined artificial RNA molecules that mimic the main aspects of transcriptome complexity. They are added in minuscule amounts to samples before the library preparation and are processed alongside endogenous RNA. After mapping the reads to the combined genome and SIRVome, the SIRV control data is used to analyze the quality metrics and categorize the experiments. Validation of the RNA-Seq workflow is possible, and biases and “blind spots” are revealed and can be addressed. In addition, this small subset of control data can be searched against a database to identify experiments of high concordance that can then be used for meaningful differential expression analyses.
Design

Modular Concept
The Spike-In RNA Variants (SIRVs), were designed to mimic transcriptome complexity in a condensed manner (Fig. 2) with each module probing a specific component.

Isoform Module
69 artificial transcript variants derived from 7 human model genes comprehensively reflect variations of alternative splicing, alternative transcription start- and end-sites, overlapping genes, and antisense transcripts. With between 6 and 18 transcript variants each, isoform complexity is high to enable in-depth probing of RNA-seq work-flows. Correct as well as (exemplary) insufficient and over-annotations are provided for the testing of workflow robustness towards different transcript annotations.

ERCC Module
The External RNA Controls Consortium (ERCC) has developed 92 artificial transcripts with non-overlapping sequences. Due to their unique sequence identities, the ERCC controls are well suited for measuring technical parameters irrespective of isoforms, and by covering a 2^{20} (10^6) dynamic range ERCC Mix 1 addresses the entire spectrum of transcript concentration complexity. Comparison of the assigned and evaluated reads with known concentrations allows for the assessment of dynamic range, dose response, lower limit of detection, and workflow efficiency.

SIRV-Sets
Three SIRV-Sets are available, each with a different SIRV mix or combination of mixes, addressing different applications (Table 1).

<table>
<thead>
<tr>
<th>SIRV-Set 1</th>
<th>SIRV-Set 2</th>
<th>SIRV-Set 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cat. No</td>
<td>025.03</td>
<td>050.01</td>
</tr>
<tr>
<td>Modules</td>
<td>Isoform</td>
<td>Mixes E0, E1, E2</td>
</tr>
<tr>
<td>ERCC</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Property</td>
<td>Isoform detection and quantification</td>
<td>✔</td>
</tr>
<tr>
<td>Dynamic range</td>
<td>partially applicable</td>
<td>—</td>
</tr>
<tr>
<td>Applications</td>
<td>Pipeline validation</td>
<td>✔</td>
</tr>
<tr>
<td>Sample control</td>
<td>—</td>
<td>✔</td>
</tr>
</tbody>
</table>
**SIRV-Set 1**

The isoform module is available in SIRV-Set 1 in 3 different mixes, termed E0, E1, and E2, with each mix containing all 69 SIRV isoform transcripts (from 7 SIRV genes) but in different concentration ratios (Fig. 3). E0 is ideal for assessing the detection capabilities of a given RNA-Seq workflow, since all 69 transcripts are present in equimolar concentrations, and their detection is not a function of read depth or similar. E1 contains a moderate concentration distribution of transcript variants of a given gene, and E2 represents the natural situation, where a dominant, abundant transcript variant is transcribed from a gene together with (up to 17) other variants present at lower expression levels (down to <1%). The latter situation is challenging for correct transcript determination based on short read assembly but also efficiently tests the linearity and sensitivity of long-read sequencing platforms, and protocols with restricted read-depth.

Figure 3 | Distribution of the 4 SubMixes in the 3 isoform mixes and the resulting intra- and inter-mix ratios. Each SIRV transcript enters the mixes as part of a SubMix, and transcript isoforms of each of the 7 SIRV genes are distributed across all SubMixes. Left, the intra-mix concentration ratios provide three different concentration settings to evaluate accuracy in relative concentration measurements. Right, the preset fold-changes allow for 3 possible inter-mix comparisons to evaluate differential gene expression measurements.

**SIRV-Set 2**

The equimolar Isoform Mix E0 is available on its own as SIRV-Set 2. It is ideally suited for cost-sensitive applications and for RNA-Seq experiments that need to be validated for the detection of a complex mixture of isoforms without applying a high read depth to cover transcripts at different concentrations. SIRV-Set 2 is suitable for the calculation of concordance, since the experimental fingerprints depend solely on SIRV isoform complexity but not on an input concentration differences.

**SIRV-Set 3**

This set contains both the Isoform Mix E0 and the ERCC Mix in equal shares. The mixture of 69 SIRV isoform transcripts and 92 non-overlapping ERCC RNAs addresses the need for complex spike-in RNA controls that cover both a high level of isoform complexity and a large concentration range. Together, they enable more comprehensive quality assessment and monitoring across the whole RNA-Seq workflow, to derive technical details and telling fingerprints for comparing individual samples and experiments. The single-isoform ERCC transcripts cover concentrations of 6 orders of magnitude and are complemented by the equimolar SIRV isoforms (Fig. 4).

Figure 4 | Concentrations and complexity of transcripts in SIRV-Set 3. The isoform module with up to 18 transcripts per gene contains all RNAs at the same molarity (green bar). The single-isoform ERCC module covers concentrations of 6 orders of magnitude (grey bar), which represents the entire dynamic range of naturally occurring transcripts. (a) The amount of attomoles refers to the typical amount that is spiked into 100 ng total RNA.

**Data Analysis**

Data sets of spike-in RNA-Seq experiments are processed uniformly up to the level of mapping, at which point the reads are aligned to a combination of genomic reference and the SIRVome (the artificial "genome" detailing the spike-in sequences and annotations). Standard and custom bioinformatic tools compare the results from measured and expected SIRV read distribution at different levels, from raw read mapping up to transcript identification and quantification.

An exemplary data evaluation workflow has been realized in the freely available “SIRV Suite”, using a collection of published tools and algorithms in the “Galaxy” environment (www.sirvsuite.org). It allows for the design and evaluation of individual SIRV experiments, as well as inter-experimental comparisons. For the evaluation of ERCC reads in RNA-Seq experiments, the NIST provides a software package called the “ERCC dashboard”, and further evaluations are described in publications of the SEQC/MAQC-III Consortium.

Calculated quality measures include accuracy and precision as well as Coefficient of Deviation (CoD), a measure for the deviation between measured and expected coverages (see also Fig. 2). Although it is important to monitor absolute rankings during method development, the crucial parameter for the comparison of experimental data is not the extent of biases in experiments but bias consistency. The differences between experiments can be determined based on the consistent condensed complexity of the SIRVs. This knowledge allows for an informed decision whether data sets can be compared, and enables the setting of a baseline for experiment-inherent deviation. Knowledge of the technical variation between samples is a prerequisite to asking meaningful biological questions.
Published SIRV Applications

- Long-read RNA-seq workflow validation
  “By benchmarking our experimental and computational pipelines on ONT MinION data derived from a mix of synthetic transcripts, we showed that our approach identifies the location of transcription start and end sites as well as splice sites in a genome.”
  “Importantly, despite the high error rates of direct RNA nanopore sequencing, especially on homopolymers like poly(A) stretches, most 3’ ends of synthetic RNAs with encoded poly(A) tails correctly mapped within 25 nt after sequencing.”

- Direct RNA-Seq
  “We evaluated the unambiguous detection of splice variants using Lexogen’s Spike-in RNA Variant Control Mixes (SIRVs).”
  “The long reads generated by direct RNA sequencing by nanopores should allow straightforward detection of splice variants. We investigated this using Lexogen’s SIRV panel, and were able to detect the majority of splice variants from the panel [...]”

- Evaluation of data analysis tools
  “We evaluated minimap2 on SIRV control data [...] where the truth is known.”
  “The results from SpaRC show comparable results with slightly improved recovery of the synthetic spike-in transcripts (more true positives) and slightly reduced artifacts (fewer false positives).”
  “We evaluated FLAIR [...] using both Lexogen Spike-in RNA Variants (SIRVs) sequenced with nanopore 2D technology [...]”

- Platform comparison
  “We evaluated the performance of PacBio, ONT, Hybrid-Seq and Illumina data on isoform quantification, using the gold standard SIRVs.”
  “For comparison, we utilize RNA spike-ins, including External RNA Controls Consortium (ERCC)- and Spike-in RNA Variants (SIRVs). [...] we are the first to compare the quantification obtained with ONT cDNA-Bulk, RNA-Seq and Illumina cDNA-Seq. RT-PCR samples were sampled from brain and liver tissues of mice and were mixed with Lexogen’s Spike-in RNA Variants (SIRVs) as a control for quantification of RT-PCR.”

- Comparison of single-cell RNA-seq protocols
  “This experiment provided quantitative evidence that mRNA splice-form variation can be inferred at the single-cell level when the appropriate protocol is used.”

- Normalization of single-cell RNA-Seq data
  “Our results indicate that spike-in normalization is reliable enough for routine use in scRNA-seq data analyses.”

- Data set concordance
  “By these means SIRV controls increase the comparability within and between sequencing experiments at the transcript isoform level.”
  “The gene count matrix was normalized with voom, using normalization factors computed with the TMM method from the reads aligning to the control RNAs.”
  “For mouse libraries, Lexogen SIRVome was independently aligned and quantified for qualitative assessment of library concordance.”

- Reviews
  “The development of spiked RNA spike-ins, which emulate the complex exon and intron architecture of human genes, has allowed further assessments of alternative splicing and transcript assembly using RNA-seq. [...]”
  “The company’s modules let labs separate various types of sensitivity-detection applications from ones in which isoform detection is key.”

Ordering Information
Catalog Number:
025.03 (SIRV-Set 1 (Iso Mix EO, E1, E2))
050.01 (SIRV-Set 2 (Iso Mix EO), 1 vial)
050.03 (SIRV-Set 2 (Iso Mix EO), 3 vials)
051.01 (SIRV-Set 3 (Iso Mix EO/ERCC), 1 vial)
051.03 (SIRV-Set 3 (Iso Mix EO/ERCC) 3 vials)

References