LEXOGEN

Enabling complete transcriptome sequencing

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).

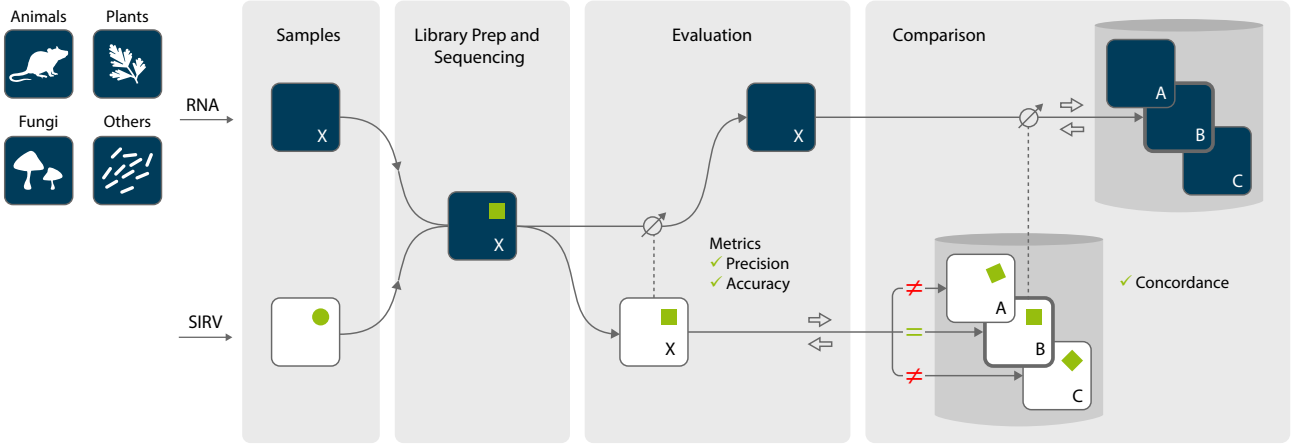

Figure 1 | Using SIRV controls in RNA-Seq. SIRVs are defined artificial RNA molecules that mimic the main aspects of transcriptome complexity. They are added in minuscule amounts to samples before the library preparation and are processed alongside endogenous RNA. After mapping the reads to the combined 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 workflows [1]. Correct as well as (exemplary) insufficient and over-annotations are provided for the testing of workflow robustness towards different transcript annotations [2].

The SIRV isoform module is available in the form of three mixes, with equimolar concentrations of all SIRV transcripts in mix E0, and molar ratios at magnitudes 1 (up to 8-fold) and 2 (up to 128-fold) in mixes E1 and E2, respectively.

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 [3, 4]. 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.

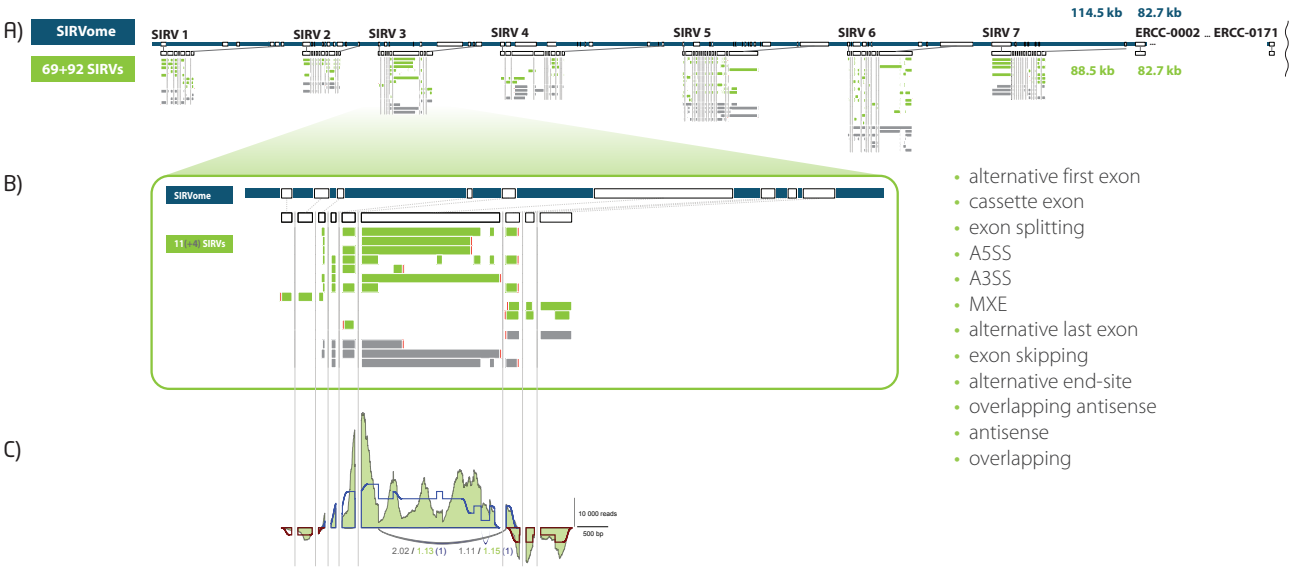

Figure 2 | SIRVs design overview. SIRV 1 to SIRV 7 mimic human model genes to comprehensively represent all main aspects of alternative splicing and transcription in numerous repeats and variations. **(A)** The 7 SIRV isoform genes and the 92 ERCC genes depicted as an artificial chromosome, the "SIRVome". **(B)** Enlargement of SIRV 3 which provides 11 transcript variants (shown in green); transcript variants shown in grey are additional annotations for alternative evaluation procedures. **(C)** The known transcript isoform concentrations in the SIRV mixes allow for comparison of the expected gene and exon junction coverage (blue and red lines for the plus and minus strand, respectively) with experimentally derived read coverages (green areas).

SIRV-Sets

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

Table 1 | SIRV-Set selection guide. SIRV-Set 1 (Cat. No 025.03) contains the isoform mixes E0, E1, and E2 of the isoform module, SIRV-Set 2 (Cat. No 050.01 and 050.03) provides the isoform Mix E0 only, whereas SIRV-Set 3 (Cat. No 051.01 and 051.03) contains SIRV Isoform Mix E0 in a mixture with the ERCCs. ✔ : applicable, — : not applicable, and partially applicable (or parts of the sets applicable).

		SIRV-Set 1	SIRV-Set 2	SIRV-Set 3
Cat. No		025.03	050.0x	051.0x
Modules	Isoform	Mixes E0, E1, E2	Mix E0	Mix E0
	ERCC	—	—	Mix 1
Property	Isoform detection and quantification	✔	✔	✔
	Dynamic range	partially applicable	—	✔
Applications	Pipeline validation	✔	partially applicable	partially applicable
	Sample control	—	✔	✔

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 already contains a moderate concentration distribution of transcript variants of a given gene, and E2 represents the natural situation, whereby 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 already 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.

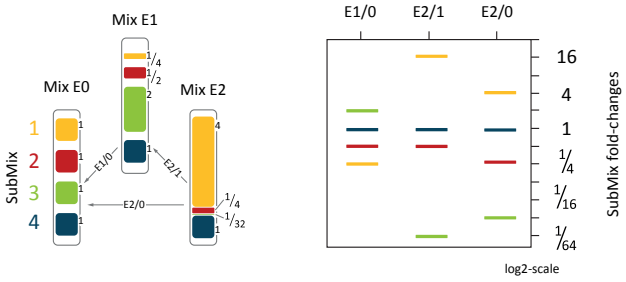

Figure 3 | Distribution of the 4 SubMixes in the 3 isoform mixes and the resulting intra- and inter-mix ratios. Each SIRV transcript enters the mixes as part of a SubMix, and transcript isoforms of each of the 7 SIRV genes are distributed across all SubMixes. Left, the intra-mix concentration ratios 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 very suitable for the calculation of concordance, since the experimental fingerprints depend solely on SIRV isoform complexity but not on 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).

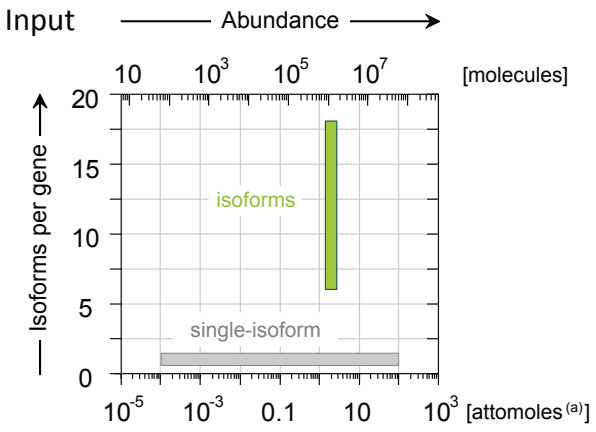

Figure 4 | Concentrations and complexity of transcripts in SIRV-Set 3. The isoform module with up to 18 transcripts per gene contains all RNAs at the same molarity (green bar). The single-isoform ERCC module covers concentrations of 6 orders of magnitude (grey bar), which represents the entire dynamic 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" [5], and further evaluations are described in publications of the SEQC/MAQC-III Consortium [6].

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.

Novel SIRV RNA Delivery Format

SIRV-Set 2 and SIRV-Set 3 are delivered in a dried, stable format, and the resuspended RNAs are stable for an extended period of time under recommended storage conditions. Together, this allows for efficient use of the SIRVs in multiple experiments conducted at different times.

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." [7]
- **Platform comparison:** "We evaluated the performance of PacBio, ONT, Hybrid-Seq and Illumina data on isoform quantification, using the gold standard SIRVs." [2]
- **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." [8]
- **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." [9]
- **Data set concordance:** "By these means SIRV controls increase the comparability within and between sequencing experiments at the transcript isoform level." [1]
- **Direct RNA-Seq:** "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. [...] However, many RNAs in the SIRV panel are highly similar, resulting in some mismapping. [...] We anticipate that improvements in the accuracy of direct RNA sequencing will enhance our ability to map highly similar splice variants and will decrease bias further still." [10]
- **Reviews:** "The development of spliced 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." [11] | "New spike-in RNAs derived from human sequences are more representative of mammalian transcripts, and may alleviate some of these issues [...]" [12]

References

[1] Paul, L. et al. SIRVs: Spike-In RNA Variants as External Isoform Controls in RNA-Sequencing. *bioRxiv* (2016).

[2] Weirather, J. L. et al. Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis. *F1000Res 6*, 100 (2017).

[3] Baker, S. C. et al. The External RNA Controls Consortium: a progress report. *Nature Methods 2*, 731–734 (2005).

[4] External RNA Controls Consortium. Proposed methods for testing and selecting the ERCC external RNA controls. *BMC Genomics 6*, 150 (2005).

[5] Munro, S. A. et al. Assessing technical performance in differential gene expression experiments with external spike-in RNA control ratio mixtures. *Nature Communications 5*, 5125 (2014).

[6] SEQC/MAQC-III Consortium. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. *Nature Biotechnology 32*, 903–914 (2014).

[7] Byrne, A. et al. Nanopore long-read RNAseq reveals widespread transcriptional variation among the surface receptors of individual B cells. *Nature Communications 8*, 16027 (2017).

[8] Svensson, V. et al. Power analysis of single-cell RNA-sequencing experiments. *Nature Methods 14*, 381–387 (2017).

[9] Lun, A. T. L., Calero-Nieto, F. J., Haim-Vilmovsky, L., Gottgens, B. & Marioni, J. C. Assessing The Reliability Of Spike-In Normalization For Analyses Of Single-Cell RNA Sequencing Data (2017).

[10] Oxford Nanopore Technologies. Nanopores allow direct sequencing of full-length RNA strands and modified RNA Nucleotides. Available at https://nanoporetech.com/index.php/publications/nanopores-allow-direct-sequencing-full-length-rna-strands-and-modified-rna-nucleotides.

[11] Hardwick, S. A., Deveson, I. W. & Mercer, T. R. Reference standards for next-generation sequencing. *Nature Reviews, Genetics* (2017).

[12] Andrews, T. S. & Hemberg, M. Identifying cell populations with scRNASeq. *Molecular Aspects of Medicine* (2017).

Ordering Information

Catalog Number:

025.03 (SIRV-Set 1 (Iso Mix E0, E1, E2))
050.01 (SIRV-Set 2 (Iso Mix E0), 1 vial)
050.03 (SIRV-Set 2 (Iso Mix E0), 3 vials)
051.01 (SIRV-Set 3 (Iso Mix E0/ERCC) 1 vial)
051.03 (SIRV-Set 3 (Iso Mix E0/ERCC) 3 vials)

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

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

SIRVs™
Spike-in RNA Variant Control Mixes