LEXOGEN

Enabling complete transcriptome sequencing

SIRVs™

Spike-In RNA Variant Controls

SIRV-Set 2

Spike-In RNA Variant Controls with Isoforms
(Iso Mix E0)

SIRV-Set 3

Spike-In RNA Variant Controls with Isoforms and ERCCs
(Iso Mix E0 / ERCC)

User Guide

Catalog Number:
050 (SIRV-Set 2 (Iso Mix E0))
051 (SIRV-Set 3 (Iso Mix E0/ERCC))

050UG134V0100

LEXOGEN

Enabling complete transcriptome sequencing

ATTENTION: Changed storage and usage

050UI149V0100

Changed storage and usage conditions for SIRV-Set 2 and SIRV-Set 3!

ATTENTION: The Spike-In RNA Variant Controls SIRV-Set 2 and 3 are currently provided in frozen format only (from December 2017). No resolubilization is required. Storage is at, or below, -20 °C. Detailed protocol adaptations for storage and usage see below.

Detailed protocol changes for SIRV-Sets 2 and 3 ordered from December 2017:

Kit Components and Storage Conditions (p.10):

- The current SIRV-Set 2 and 3 are provided in frozen format only. The tube(s) must be stored at, or below, -20 °C. Freeze/thaw cycles should be avoided, as these contribute significantly to alteration of the RNA integrity and concentration. The contained additives stabilize the solution(s) sufficiently to undergo one freeze-thaw cycle. We recommend to aliquot the solution upon first time usage.

SIRV Solubilization, Dilution, Aliquoting, and Interim Storage (p.12-13):

- **Skip "Resolubilization" (p.12, upper description):** The current SIRV-Set 2 and 3 are already provided in 10 µl liquid volume at a concentration of 2.52 ng/µl and 3.03 ng/µl for SIRV-Set 2 and 3 respectively. Do not add additional nuclease-free water. They are delivered frozen and must not be resolubilized. Resolubilization is only applicable for the dried format.
- **Proceed with "Dilution" (p.12, lower description):** The delivered solution(s) correspond to a 1:10 dilution. The contained additives stabilize this 1:10 dilution sufficiently for the solution to undergo one freeze-thaw cycle. We recommend to aliquot the solution upon first time usage.

If you have any questions, please contact us at support@lexogen.com or +43 1 3451212 41.

FOR RESEARCH USE ONLY. NOT INTENDED FOR DIAGNOSTIC OR THERAPEUTIC USE.
INFORMATION IN THIS DOCUMENT IS SUBJECT TO CHANGE WITHOUT NOTICE.
Lexogen does not assume any responsibility for errors that may appear in this document.

PATENTS AND TRADEMARKS
The SIRVs are covered by issued and/or pending patents. SIRV is a trademark of Lexogen. Lexogen is a registered trademark (EU, CH, USA).

Agilent is a registered trademark of Agilent Technologies Inc., Ambion is a registered trademark of Life Technologies Corporation, RNaseZap is a registered trademark of Ambion, Inc., and RNasin is a trademark of Promega Corporation. All other brands and names contained in this user information are the property of their respective owners.

The use of ERCC in the product name does not constitute an affiliation or sponsorship by the External RNA Controls Consortium.

Lexogen does not assume responsibility for violations or patent infringements that may occur with the use of its products.

LIABILITY AND LIMITED USE LABEL LICENSE: RESEARCH USE ONLY
This document is proprietary to Lexogen. The SIRV mixes are intended for use in research and development only. They need to be handled by qualified and experienced personnel to ensure safety and proper use. Lexogen does not assume liability for any damage caused by the improper use or the failure to read and explicitly follow this user guide. Furthermore, Lexogen does not assume warranty for merchantability or suitability of the product for a particular purpose. The purchase of the product does not convey the right to resell, distribute, further sublicense, repackage, or modify the product or any of its components. This document and its contents shall not be used or distributed for any other purpose and/or otherwise communicated, disclosed, or reproduced in any way without the prior written consent of Lexogen.

For information on purchasing additional rights or a license for use other than research, please contact Lexogen.

WARRANTY
Lexogen is committed to providing excellent products. Lexogen warrants that the product performs to the standards described in this user guide up to the expiration date. Should this product fail to meet these standards due to any reason other than misuse, improper handling, or storage, Lexogen will replace the product free of charge or issue a credit for the purchase price. Lexogen does not provide any warranty if product components are replaced with substitutes. Under no circumstances shall the liability of this warranty exceed the purchase price of this product. We reserve the right to change, alter, or modify any product without notice to enhance its performance.

LITERATURE CITATION

When referring to this spike-in mix in a publication, please use "SIRV-Set 2" or "SIRV-Set 3", "Spike-In RNA Variant Controls with Isoforms" or "Spike-In RNA Variant Controls with Isoforms and ERCCs". Individual transcripts can be referred to as "SIRV101" and "ERCC-0025" with the entirety being "SIRV isoforms" and "ERCCs". Stating the Catalog Number (Cat. No. 050.0x or 051.0x) and the Lot Number (on the tube label) in the Materials and Methods section uniquely identifies the SIRV product you are using.

CONTACT INFORMATION

Lexogen GmbH
Campus Vienna Biocenter 5
1030 Vienna, Austria
www.lexogen.com
E-mail: info@lexogen.com

Support
E-mail: support@lexogen.com
Tel. +43 (0) 1 3451212-41
Fax. +43 (0) 1 3451212-99

Table of Contents

LEXOGEN

Enabling complete transcriptome sequencing

1. Introduction

1.1 Spike-in RNA Controls

RNA sequencing (RNA-Seq) workflows comprise RNA purification, library generation, the sequencing itself, and the evaluation of the sequenced fragments. The initial steps impose biases for which the data processing algorithms try to compensate afterwards. Key tasks for data evaluation algorithms are the concordant assignment of fragments to the transcript variants, robustness towards annotation flaws, and the subsequent deduction of the corresponding abundance values. Unless the quality of all individual processing steps can be unequivocally determined, subsequent comparisons of experimental data remain ambiguous. The proliferation of different RNA-Seq platforms and protocols has created the need for multifunctional spike-in controls, which are integrated and processed with the samples to enable the monitoring and comparing of key performance parameters like sensitivity and input-output correlation as well as the detection and quantification of transcript variants (Figure 1).

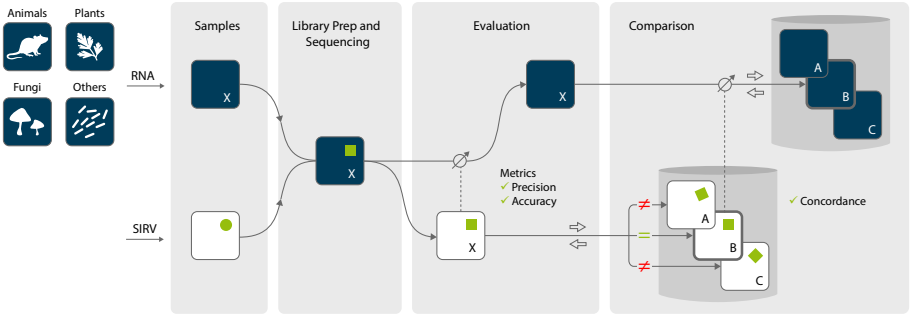


Figure 1. Workflow for using spike-in controls in RNA-Seq. Spike-in controls are defined synthetic RNA molecules that mimic the main aspects of transcriptome complexity. They are added in minuscule amounts to samples before library preparation to undergo the very same processing steps as the endogenous RNA. After mapping the reads to the combined genome, the spike-in data are used to derive quality metrics and to categorize the experiments. The dotted lines show the decision-making processes of deciding i) if the complete data set is worthy of further processing (or if an experiment needs to be repeated), and ii) which data sets have concordance that will permit meaningful comparison of the full data sets with each other.

The Spike-In RNA Variants (SIRV) were conceived as a family of modules to offer tailored solutions for the control of RNA measurements (Figure 2). Each module contains a group of synthetic transcripts, which mimic predominantly only one aspect of transcriptome complexity to facilitate systematic analysis.

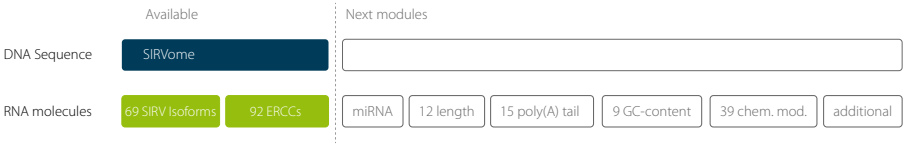


Figure 2. SIRV modules. The SIRV isoforms and single-isoform transcripts (ERCCs) are established synthetic RNA molecules that mimic the two major aspects of transcriptome complexity, isoforms and abundance. Additional modules are outlined and will be added to the SIRV family at availability. SIRVome is the corresponding artificial reference genome.

1.2 Isoform Complexity

The isoform module of the Spike-In RNA Variants was developed to validate the performance of isoform-specific RNA-Seq workflows and to serve as a control for the comparison of RNA-Seq experiments and individual sample preparations. It is a set of 69 artificial transcript variants derived from 7 human model genes, which were complemented by additional isoforms and transcription variants to comprehensively reflect variations of alternative splicing, alternative transcription start- and end-sites, overlapping genes, and antisense transcripts (Figure 3), to contain finally between 6 and 18 transcript isoforms per gene. For the sake of simplicity, we will refer to all these transcriptional variants just as isoforms.

Figure 3. SIRV isoform design. SIRV isoforms mimic human model genes to represent in their entirety all main aspects of alternative splicing and transcription in numerous repeats and variations. The transcript isoforms are shown aligned to a *master gene* (top line), and hence there can be no *intron retention* event. Therefore, the opposite is described here as *exon splitting*. The sequences themselves have no significant similarities to any known data base entries but match eukaryotic gene features in terms of their makeup and exon-intron structure. A5SS and A3SS, alternative 5'/3' splice sites; MXE, mutually exclusive exons.

Considerations for coping with non-ideal transcript annotations were incorporated in the SIRV isoform design (Figure 4). Exemplary insufficient and over-annotations are provided in addition to the correct reference SIRVome to enable the testing of Next Generation Sequencing (NGS) data evaluation algorithms for their robustness towards realistic, imperfect annotations.

Figure 4. SIRV isoform design exemplified by SIRV1. The SIRV1 gene was derived from the human KLK5 gene, with transcripts added to the Ensembl annotations to generate comprehensive transcriptome complexity. All original and derived gene structures are shown in the Appendix. Transcripts in **blue** are part of SIRV mixes, transcripts in **green** are only part of an over-annotation. (i) Refers to transcripts that are omitted in an incomplete annotation. Exons of the master gene structure are shown in white, and the 3' poly(A) tail is marked in **red** to indicate transcript 5'-3' orientation.

The SIRV isoforms enable the measurement of quality metrics such as precision and accuracy of entire workflows including mapping, isoform assembly, and quantification, to rank concordance and comparability of individual experiments at isoform resolution. Summing isoform read counts yields the corresponding SIRV gene expression values.

Because in SIRV-Set 2 and SIRV-Set 3 these transcripts are provided in equimolar amounts, their detection is not biased by different concentrations.

1.3 Abundance Complexity

The ERCC RNA spike-in controls module was developed by the External RNA Controls Consortium (ERCC)[1,2] and provides a set of 92 artificial transcripts with non-overlapping, mono-exonic, single-isoform sequences (Figure 5).

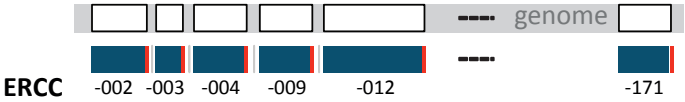


Figure 5. ERCC single-isoform design. ERCC transcripts follow the 1 gene, 1 exon, 1 transcript layout, providing each ERCC transcript with a unique sequence identity. Genes (exons) are shown in white, derived transcripts in **blue**, and the poly(A) tail is marked in **red** to indicate transcript 5'-3' orientation.

The single-isoform ERCC mix enables a straight-forward assessment of dose response, as well as the definition of the lower limit of detection and the assessment of workflow efficiency[3–6].

Because the assignment of all uniquely mapping reads is unambiguous not only on the gene but also on the transcript level, their detection and the derived input-output correlation is not influenced by any isoform complexity.

1.4 SIRV Sets

The modular structure of the SIRVs (Figure 2) allows for combining these synthetic transcripts in specific combinations to probe the different dimensions of transcriptome complexity either isolated or in combination. We use the following definitions:

Module Group of SIRVs which mimic predominantly one aspect of transcriptome complexity,

Mix SIRVs of the same module which are combined in precise defined molarity,

Set Term for the union of mixes.

The currently available sets are show in the overview in Table 1.

Table 1. SIRV set selection guide for choosing suitable controls to either validate different quality metrics of RNA-seq pipelines or to monitor the concordance of measuring individual samples. SIRV-Sets 2 and 3 are covered in this User Guide. Letter "x" refers to number of vials, 1 or 3. The ERCC Module is represented by ERCC Mix 1[7].

		SIRV-Set 1	SIRV-Set 2	SIRV-Set 3
Cat. No		025.03	050.0x	051.0x
Module(s)	Isoforms	Isoform Mixes E0, E1, E2	Isoform Mix E0	Isoform Mix E0
	ERCC	✖	✖	ERCC Mix 1
Property	Isoform detection and quantification	✔	✔	✔
	Dynamic range	partially	✖	✔
Applications	Pipeline Validation	✔	partially	partially
	Sample Control	✖	✔	✔

SIRV-Set 1 (Cat. No. 025.03) includes three mixes of the isoform module designed for the validation of concentration measurement (including fold change), with isoform resolution.

Validation is the process of assessing the reliability of a method, either of the entire RNA-Seq pipeline or steps thereof. The fragile nature of RNA, the transcriptome complexity, and the large number of different RNA-Seq workflows result in inherently high variability. Workflow-validation is crucial as a proof-of-concept. However, it cannot assure the faultless processing of each individual sample, which requires spike-in controls in every sample.

SIRV-Set 2 and SIRV-Set 3 contain one or two modules in only one mix. Hence, each SIRV transcript is present in one defined concentration. These sets are designed to be spiked into every RNA-Seq sample for controlling the consistency of sample processing and measurement (Figure 1). All quality metrics, except fold change measurements, can be determined experimentally for each individual sample.

SIRV-Set 2 (Cat. No. 050.0x) contains 69 isoform sequences in equimolar ratios, which originate from 7 genes to probe the boundaries of resolving isoform complexity. Resolution depends on coverage biases and the ability of the data analysis to account for these biases.

SIRV-Set 3 (Cat. No. 051.0x) contains the same mix of 69 isoforms and 92 non-overlapping sequences covering both a high level of isoform complexity and a large concentration range. This set enables an even more comprehensive quality monitoring of individual samples. The two dimensions, abundance of the transcripts (x-axis) and concurrence of isoforms per gene (y-axis), covered by this SIRV set are shown in Figure 6.

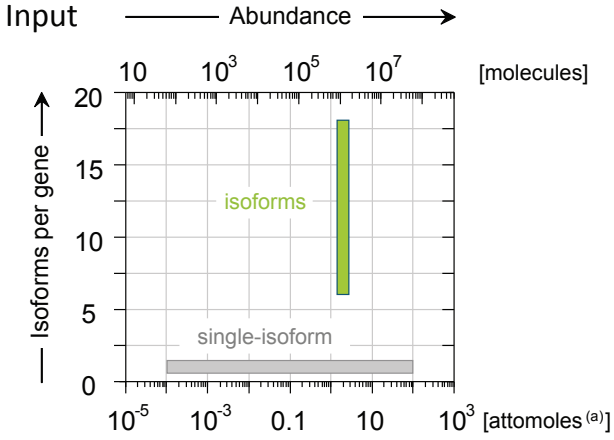


Figure 6. The SIRV isoform and single-isoform (ERCCs) transcripts in SIRV-Set 3 control for the two main dimensions of transcriptome complexity: isoforms and abundance. The isoform module with 69 transcripts from 7 genes contains all species at the same molarity (**green bar**). The single-isoform module with 92 ERCC transcripts spans a concentration range of 6 orders of magnitude (**grey bar**), which is sufficient to cover 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 with the aim to obtain approx. 1 % of the mRNA-Seq reads.

1.5 Main Aspects of SIRV Data Evaluation

The SIRVs are processed alongside endogenous RNA. The condensed representative complexity of the SIRVs senses quality parameters of the entire RNA-Seq experiment in each controlled sample.

The **precision** (random error) in quantifying single-isoform (ERCC) transcripts in RNA-seq experiments is method-, concentration-, and read depth-dependent with reads being typically Poisson distributed. The **accuracy** (systematic error) depends on biases introduced by the respective methods. The single-isoform (ERCC) module covers a wide concentration range of 6 orders of magnitude to probe all technical parameters related to transcript abundance.

In contrast, meaningful isoform detection and quantification, which goes beyond mere statistical probabilities of assigning read counts to all available annotations, requires sufficient coverage of specific sequences. Therefore, the isoform spike-ins are provided at a concentration in the upper range of the ERCCs. Thereby, the task of identifying a given isoform is not confounded by differing input concentrations. The gene coverage of the isoform module in relation to the single-isoform module is shown in an exemplary series in Figure 7.

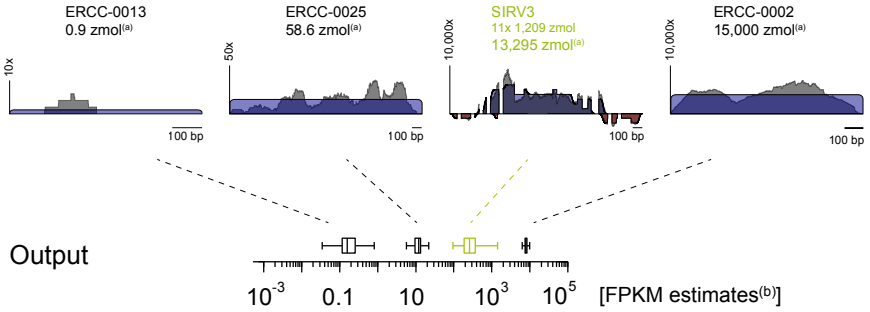


Figure 7. Read coverages of SIRV isoform and single-isoform (ERCC) genes depend on input concentration, library preparation efficiency, biases, and read depth. Quantifying the 92 single-isoform transcripts (ERCCs) depends on the averaged overall coverage but is rather independent of positional coverage fluctuations. Three ERCC examples of different input abundance are shown. With 6-18 isoforms mapping to the 7 SIRV genes, NGS read assignment and subsequent isoform quantification is much more challenging and depends strongly on coverage uniformity. One gene with 11 SIRV isoforms is shown alongside the ERCCs. The **blue** areas represent the expected coverage in the sense direction, and the **red** areas the expected coverage in the antisense direction. The grey areas show exemplary coverages from one stranded library preparation that has been sequenced in paired end mode. (a) The number of zettamoles refers to the total amount per SIRV-Set 3 vial. (b) Reflects the FPKM bandwidth of the controls when those occupy around 1 % of the reads in an mRNA-Seq experiment.

The quantification of SIRV isoforms remains challenging on both short-read platforms (mostly due to alignment issues and coverage biases) and long-read platforms (e.g., because of per base error, low read numbers, and amplification bias). This implies that the precision in the quantification of transcripts from genes with multiple isoforms is often significantly lower than for single-isoform genes at similar input concentrations.

Depending on both the abundance complexity and the isoform complexity, random errors define the lower boundaries of confidence intervals, which estimate the distribution of endogenous RNA measurements. Further, they allow for calculating the lower limit of detection for differential expression, either by applying simplified mathematical models, or by tracing the concentration region of interest by down-sampling and reassigning isoform reads.

2. Kit Components and Storage Conditions

The SIRV controls are supplied in a stabilized dried format. The tubes are distributed in a cardboard box and can be stored in this package at ambient laboratory room temperatures of 20 to 25 °C. The SIRV controls should not be exposed to higher temperatures (exceeding 40 °C) for an extended time.
Each box of SIRVs contains either 1 or 3 tubes as well as molecular biology-grade nuclease free-water for resolubilization and subsequent dilution.

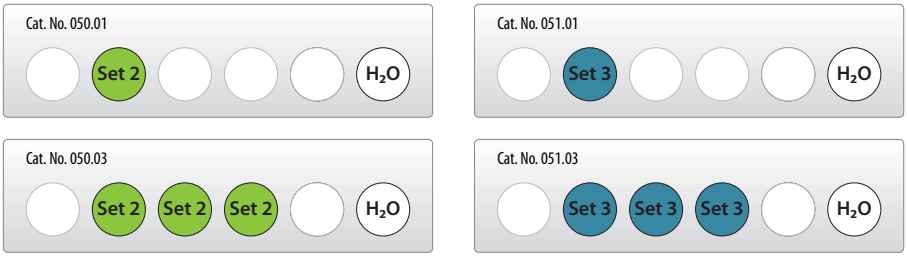


Figure 8. Location of kit components.

SIRV-Set 2 contains an equimolar mix of the 69 SIRV isoforms (Iso Mix E0) at 1.0 fmol each. The total amount of RNA is 69.0 fmol corresponding to 25.2 ng.

SIRV-Set 3 combines the equimolar mix of the 69 SIRV isoforms (Iso Mix E0) at 0.6 fmol each and 92 single-isoform transcripts (ERCCs) ranging from 0.014 amol to 15.0 fmol. The total amount of RNA is 93.2 fmol or 30.3 ng per tube.

Table 2. Content of tubes with dried SIRV RNA.

		Amount	
Component	**Content**	**fmol**	**ng**
Set 2	Iso Mix E0	69.0	25.2
Set 3	Iso Mix E0 / ERCC	93.2	30.3

3. Materials and Equipment Required

The SIRVs should be resuspended and diluted with either, RNase-free molecular-biology grade water (supplied with the kit) or RNA-compatible buffers, e.g., sodium citrate, pH 6.4, or Tris-EDTA, pH 7.0. Divalent cations catalyze RNA fragmentation and must be avoided as dilution buffer components.

4. Application

4.1 RNA Handling Guidelines

- RNases are ubiquitous, and special care should be taken throughout the procedure to avoid RNase contamination.
- It is important that the solutions as well as all materials that come into contact with the SIRVs are absolutely RNase-free. Working with SIRVs requires decontaminated pipettes. The use of barrier pipette tips is advised. Use a sterile and RNase-free workstation or laminar flow hood if available. Please note that RNases may still be present on sterile surfaces and that autoclaving does not completely eliminate RNase contamination. Before starting to work with SIRVs, clean your work space, pipettes, and other equipment with RNase removal spray (such as RNaseZap, Ambion Inc.) as per the manufacturer's instructions. **ATTENTION:** Do not forget to rinse off any RNaseZap residue with RNase-free water after usage! Residues of RNaseZap may damage the RNA.
- Protect all reagents and your RNA samples from RNases on your skin by wearing a clean lab coat and fresh gloves. Change gloves after making contact with equipment or surfaces outside of the RNase-free zone.
- Avoid speaking above opened tubes. Keep reagents closed when not in use to avoid airborne RNase contamination.
- Use commercial ribonuclease inhibitors (i.e., RNasin, Promega Corp.) to maintain RNA integrity when storing samples. SIRV mixes contain RNasin.
- All disposables that come into contact with SIRVs must have a low binding capacity for nucleic acids. This concerns vials, microtubes, plates, and pipette tips.
- When working with SIRVs in solution, freeze-thaw cycles must be minimized for the concentrated stock solutions and should be avoided for diluted aliquots. Although the samples contain RNasin and are provided in a stabilizing buffer, hydrolysis, oxidation, and adsorption lead to fragmentation and loss of SIRVs.

4.2 SIRV Solubilization, Dilution, Aliquoting, and Interim Storage

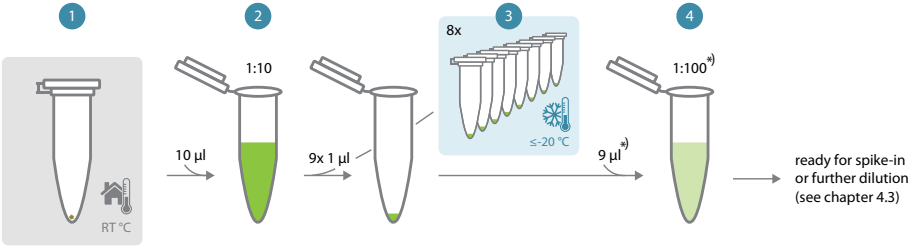


Figure 9. Workflow pictogram illustrating 1 SIRV storage, 2 resolubilization / dilution, 3 aliquoting, and further interim storage. For the handling of SIRVs strictly follow the guidelines of chapter 4.1. 4 The final dilution depends on the experimental setup and must be calculated beforehand using Eq. 2 (p.14). This can be achieved in one or several dilution steps. *) Illustration shows only one example.

Resolubilization: SIRVs are supplied in a stabilized dried format and must be stored in the supplied lightproof cardboard box, or other lightproof equivalent storage container. Store SIRVs at ambient laboratory room temperatures of around 20 to 25 °C. To resolubilize the SIRVs perform the following four steps:

1. Add 10 µl of molecular biology-grade water to one tube.
2. Incubate the tube at room temperature (20 - 25 °C) for 10 minutes. **ATTENTION:** Do not attempt to recover SIRVs on cooling blocks or ice.
3. Pipette the solution gently 10 times up and down. Alternatively, tubes can be vortexed gently for 10 seconds at low speed to avoid wetting more surface than necessary. Centrifuge briefly to collect the entire sample.
4. Place the tube onto a cooling block (ca. 0 - 5 °C) or on ice.

The SIRVs are now resolubilized.

Dilution: Resolubilizing the SIRVs in 10 µl H_2O sets the concentration in Set-2 to 2.52 ng/µl and in Set-3 to 3.03 ng/µl. In comparison to SIRVs provided in solution (SIRV-Set 1, Cat. No. 025.03) these concentrations correspond to a 1:10 dilution (Figure 9, step 2). From here proceed to aliquoting, unless the entire amount of SIRVs is immediately required. In the depicted workflow, one aliquot will be processed immediately, the other ones are stored at ≤-20 °C.

For diluting SIRVs to the required stock concentration apply the following guidelines.

- Any dilutions higher than 1:10 must be prepared immediately before spiking the SIRVs into RNA samples. Storage and freeze-thawing should be avoided as it contributes significantly to alteration of the RNA integrity and concentration.
- Plan the dilution of the SIRVs and the spike-in of the RNA as one continuous workflow to minimize the time RNA is kept at low concentrations (minutes instead of hours).
- Work with ice-cold solutions on a cool block (at ca. 0 - 5 °C) or on ice. Do not use cool blocks at temperatures below 0° C.

- Special care must be taken when pipetting small volumes in the range of 1 µl. Pipettes in combination with the tips must first be correctly calibrated using H_2O. The pipetting must be carried out very precisely by applying the recommended pipetting technique for the pipettes in use (as per the manufacturer's instructions).
- For dilutions, use RNase-free buffers. Recommended are sodium citrate at pH 6.4 or Tris-EDTA at pH 7.0.
- Avoid pipetting volumes below 1 µl and always use larger volumes (e.g., total volume 100 µl) in the dilution series to minimize the relative error.
- We recommend performing iterative dilutions. During each dilution step, gentle but thorough mixing is essential. To mix, pipette at least 90 % of the entire volume gently up and down approximately 10 times. Alternatively, tubes can be gently vortexed for 10 seconds at low speed to avoid wetting more surface than necessary. Centrifuge briefly afterwards to collect the entire sample.
- Depending on the amount of RNA that is targeted by the SIRV spike-in, the dilution should be at least 1:100 or higher (see chapter 4.3 "Spiking of RNA Samples").

Aliquoting and Interim Storage: After SIRV resolubilization each SIRV tube contains exactly 10 µl at a concentration of 2.52 ng/µl (SIRV-Set 2) or 3.03 ng/µl (SIRV-Set 3), respectively, which is defined as a nominal 1:10 dilution. The contained additives stabilize this 1:10 dilution sufficiently for the solution to undergo one freeze-thaw cycle. We recommend to aliquot the solution directly after SIRV resolubilization.

- The 10 µl are sufficient to draw 9x 1 µl aliquots. Based on practical considerations the remaining volume is often less than 1 µl. The 1 µl aliquots must be pipetted into low adsorbence tubes and tightly sealed. Ensure that the 1 µl remains at the very bottom of the tube and is not displaced by electrostatic force. If required, spin down the solution.
- Freeze 8 aliquots immediately at ≤-20 °C for later use (Figure 9).
- Proceed with the 9. aliquot and perform subsequent dilution step(s) in short succession.

4.3 Spiking of RNA Samples

The content of each SIRV tube can be divided into up to 9 aliquots (see chapter 4.2) and used for as many independent experiments. In the following, we explain in one example how a successful spike-in experiment is carried out from this point on. The workflow is very easily adjusted to any type and amount of RNA sample and consists of 3 steps:

1. In the formulas below, enter all known variables to estimate the amount of SIRVs to be used per sample.
2. Prepare a suitable dilution that can be pipetted with high accuracy.
3. Spike in the estimated amount of SIRVs to the RNA sample.

Determining the Amount of SIRVs for the Spike-in Experiment

The equations Eq. 1 and Eq. 2 are used in the planning of the spike-in experiment:

Eq. 1	$m_{SIRV} = F_{SIRV\ reads} \times F_{target\ RNA} \times m_{RNA\ input}$
Eq. 2	$V_{SIRV} = \frac{m_{SIRV}}{C_{SIRV}} = \frac{F_{SIRV\ reads} \times F_{target\ RNA} \times m_{RNA\ input}}{C_{SIRV}}$

m_{SIRV}	mass of SIRVs to be used in a spike-in experiment per sample.
$F_{SIRV\ reads}$	fraction of desired SIRV reads.
$F_{target\ RNA}$	fraction of the RNA targeted in the RNA-Seq experiment.
$m_{RNA\ input}$	mass of RNA input per sample.
C_{SIRV}	concentration of SIRVs at the suitable dilution.
V_{SIRV}	volume to be used in the spike-in procedure.

The following example shows the use of these equations that can be easily adjusted to similar RNA-Seq experiments.

Example Eq. 1

The final fraction of desired SIRV reads ($F_{SIRV\ reads}$) is usually 0.01 (or 1 %). At 0.01 the dynamic range of the non-isoform module (ERCC) covers efficiently the dynamic range of complex transcriptomes. However, for certain applications like the spotlight validation of workflows larger fractions can be targeted. Vice versa, if the isoform module needs to be spiked-in at lower ratios, e.g., to cover a lower abundant region of interest (compare Figure 6), then the $F_{SIRV\ reads}$ value can be reduced accordingly.

Universal Human Reference RNA (UHRR, Agilent Technologies), for example, contains approximately 0.03 (or 3 %) mRNA, measured as a proportion of the total RNA. The fraction of the targeted RNA ($F_{target\ RNA}$) depends on sample type, RNA integrity, and the experimental design. The mRNA content of Human Brain Reference RNA (HBRR, Ambion) is approx. 1/3rd lower and counts for 0.02 (or 2%) of the total RNA. In contrast, if the targeted RNA is not only mRNA but all RNA except ribosomal RNA (corresponding to the ribo-depleted fraction) the fraction ($F_{target\ RNA}$) usually exceeds 0.04 (or 4 %). If certain highly abundant mRNAs are depleted from the mRNA fraction, (e.g., globin RNA in blood samples), then the fraction of remaining mRNA decreases accordingly. Poly(A) selective methods are also sensitive to RNA integrity (except tag-based methods).

The mass of SIRVs m_{SIRV} to be used in one spike-in experiment, can be estimated by multiplying $F_{SIRV\ reads}$ by the targeted RNA fraction $F_{target\ RNA}$ and $m_{RNA\ input}$. In this example:

$$m_{SIRV} = 0.01\ (F_{SIRV\ reads}) \times 0.03\ (F_{target\ RNA}) \times 100\ ng\ (m_{RNA\ input}) = 0.03\ ng\ (30\ pg)$$

Example Eq. 2

The required volume (V_{SIRV}) depends on the concentration of the SIRV solution (C_{SIRV}). The final dilution must be chosen in such way that all pipetting steps can be carried out as precisely as possible. By preparing a 1:1,000 dilution, the concentration reaches 25.2 pg/µl (SIRV-Set 2). Accordingly, by preparing a 1:2,000 dilution with SIRV-Set 3, the concentration is 15.15 pg/µl. The volume needed to spike-in 30 pg SIRVs from SIRV-Set 2 is 1.19 µl (see below), or 1.98 µl from SIRV-Set 3.

$$V_{SIRV} = 30\ pg\ (m_{SIRV}) / 25.2\ pg/\mu l\ (C_{SIRV}) = 1.19\ \mu l$$

Pipetting low volumes is often error-prone. Therefore, higher dilutions and the spike-in of proportionally larger volumes are recommended.

NOTE: When applying full-length single-molecule sequencing or tag-sequencing methods, each transcript is represented ideally by a single read, not by numerous overlapping reads as a function of transcript length (mass). In these cases, the molar ratio instead of the mass ratio would have to be considered. However, because the lengths of the SIRVs are very similar to the endogenous RNAs (median of isoform module is 1.1 kb, and of the single-isoform ERCC module 1 kb) Eq. 2 can be used likewise.

Up- and Down-Scaling

As recommended above, the resolubilized SIRV stock should be divided into 9 aliquots of 1 µl. This provides enough material to produce 9x 100 µl of a 1:1,000 dilution. According to the assumptions above, this provides enough SIRVs to spike in 9x 84 samples of 100 ng total RNA input (756 samples in total).

At the lower end of RNA input amounts, the SIRVs can also be used to control single-cell experiments. On average, a single cell contains between 10 and 30 pg of total RNA, which would require only 6 fg of SIRVs. Therefore, each of the 9 aliquots is theoretically sufficient to spike 420,000 cells, which corresponds to several large-scale single-cell experiments.

4.4 Considerations for Library Preparations

The SIRV transcripts behave in an identical way to mRNA in most aspects of any RNA-Seq library preparation. SIRVs have no sequence homology to rRNA and are therefore not targeted by rRNA directed depletion methods. The SIRV isoforms contain a 30 nt long poly(A) tail and the single-isoform ERCCs have slightly shorter and variable poly(A) tails of 24 ± 1.05 nt, which allows for poly(A) enrichment and oligodT-priming. SIRVs do not have a 5'-cap structure (5'-m^7G) but a 5' triphosphate end and are resistant to 5'- 3' exonucleases. Therefore, the use of SIRVs for cap-specific cDNA preparation methods is not feasible.

5. Analysis of Sequencing Data

5.1 Data Evaluation Overview

Although there are numerous possibilities for in-depth evaluation of SIRV data, the basic routine follows a simple workflow as depicted in Figure 10.

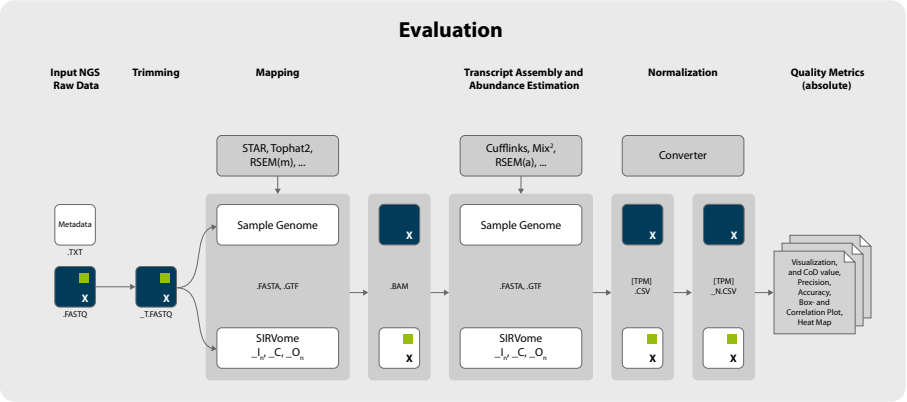


Figure 10. SIRV data evaluation scheme. SIRV reads undergo the very same processing steps as reads derived from the RNA sample. CoD, Coefficient of Deviation.

Stages of Data Evaluation:

1. All reads are quality- and barcode-trimmed and then mapped to a reference combining sample genome and SIRVome (see chapter 6 for downloads). Alternatively, a *de novo* mapper can be applied if required.
2. At the level of the BAM files, the reads are allocated to the endogenous RNA, the SIRV controls, and the non-mapping reads.
3. The mapped reads are processed by transcript assemblers and quantification algorithms.
4. Some assemblers tend to occasionally produce abundance value outliers that do not obey plausible read distributions. Therefore, sanity checks are highly recommended, which can command normalization afterwards.
5. Absolute quality metrics are calculated based on the comparison of the SIRV measures with the known input and provide unique quality control signatures for the sample.
6. Finally, sample-specific unique quality control signatures can be compared to calculate the relative quality metrics (Figure 11).

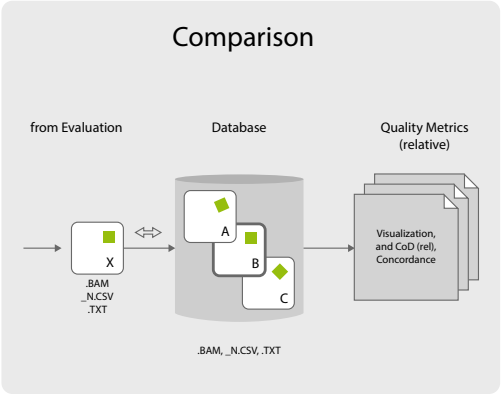


Figure 11. SIRV data comparison scheme. The control data set is used to carry out pairwise comparisons between experiments using the small subset of SIRV control data. Data sets of high concordance can be selected and the extent of expected error rates can be estimated before making a decision about comparing the complete data sets.

5.2 Read Mapping and Calculating the Mass Ratios

After barcode- and quality-trimming, the reads are mapped to the respective genome(s) and the synthetic SIRVome. The share of SIRVome reads is set in relation to its expected mass or molar ratios. For all library preparations that aim to cover the length of RNA molecules with reads, the proportion of SIRV reads obeys the input mass ratio. For library preparations that either tag or independently count RNA molecules, the share of SIRV reads should be compared to the molar input ratio.

From the ratio between the number of reads mapping to the endogenous RNA and the SIRVs, the content of the target RNA (e.g., mRNA or ribo-depleted RNA) in the spiked input can be calculated (Eq. 3).

Eq. 3

$$F_{\text{target RNA measured}} = \frac{F_{\text{target RNA assumed}} \times F_{\text{SIRV reads targeted}}}{F_{\text{SIRV reads measured}}}$$

For example, when 3 % mRNA content was assumed and 1 % SIRV reads targeted by the spike-in but actually 1.5 % SIRV reads measured, then the mRNA fraction in the sample was only 2 %. This can be interpreted as a metabolic state. However, this can also indicate that the endogenous mRNA was partially degraded. Note, that this calculation assumes accordingly precise and accurate pipetting.

5.3 Transcript Assembly, Abundance Estimation, and Calculating the Molar Ratios

In short-read NGS experiments transcript assembly algorithms must be applied to calculate abundance values whereas single-molecule and tag-sequencing technologies allow for direct counting.

5.4 Normalization

The correction of sample-specific biases is important for the subsequent interpretation of differential expression (DE) analyses. Varying RNA sample background, mRNA content, RNA quality and integrity, and variations in depletion and/or mRNA enrichment procedures influence the SIRV content in sequenced libraries.

The bias correction is important for normalization of abundances beyond relative normalization procedures. However, a careful and quantitatively precise spiking procedure at the start of the workflow is a prerequisite for accurate quantification. All measures and subsequent normalizations need to be set in context with obvious experimental variables including the achievable pipetting accuracy when operating in tiny volume scales.

SIRV abundance values can be normalized such that the measured and the expected sum of molecules for each SIRV are equal. In doing so, the comparison of relative and absolute concentration measures are uncoupled. Absolute read counts are used separately in the read count statistics to measure, e.g., mRNA content or technical variability (see chapter 5.2).

5.5 Coping with Different SIRV Annotations

SIRV reads should be mapped initially using the correct **SIRV_C** annotation (see downloads, Chapter 6). However, the mapping should be repeated using different annotations such as the provided annotations SIRV_I and SIRV_O, which mimic different annotation situations.

The under-annotated version **SIRV_I** (insufficient) can be used to assess the ability of a pipeline to detect new transcript variants. This mapping experiment shows how reads of non-annotated but sequenced SIRVs are spuriously distributed to the annotated subset skewing the quantification. The degree of variation in the derived concentrations provides an additional measure for the robustness of the RNA-Seq pipeline.

The over-annotated version **SIRV_O** refers to a third situation. Here, more SIRVs are annotated than are actually contained in the samples, for example if transcript variants were discovered in other tissues, in the same tissue but at different developmental stages, were falsely annotated, or are relics of earlier experiments, for which the high number of variants with the typical length of cloned ESTs are examples. In this setup, reads can be assigned to SIRV variants which are not part of the real sample. The degree and robustness of correct SIRVome detection in this setting

is another measure for the pipeline performance, and the share of false positives (FP) can be estimated also for the endogenous RNA.

The different annotations are provided for the SIRV isoform module but can be extended to i) develop further variations for the isoform module and ii) to design alternative annotations for the single-isoform ERCC module.

5.6 Quality Metrics

mRNA Content

Based on the assumption that the endogenous RNA and the spike-in controls are proportionally targeted by the library preparation method, the relative mass partition between controls and endogenous RNA allows for calculating the relative amounts of respective endogenous RNA fractions, e.g., all polyadenylated RNA. Here, the extrapolation of the input amounts to the output read ratio depends on the mRNA content, integrity of the input RNA, the relative recovery efficiencies of controls compared to the mRNA*), and the variability of spiking a sample with controls.

*) In the isoform module the length of the poly(A) tail comprises 30 adenosines, and in the single-isoform ERCC module 24 ± 1.05 adenosines. This is sufficient for the majority of poly(A) selective methods, hence the recovery efficiencies are identical to the polyadenylated mRNA percentage, but needs to be considered for differences observed when changing library preparation protocols.

Coefficient of Deviation (CoD)

Because the ground truth of the complex input is known, detailed target-performance comparisons of the read alignments can be performed. NGS workflow-specific read start-site distributions cause systematic lower coverages of transcript start- and end-sites. However, these systematic biases are accompanied by a variety of biases, that introduce severe local deviations from the expected ideal coverage. To obtain a comparative measure, gene-specific coefficients of deviation (CoD) can be calculated. The mean of CoD values from all 7 genes of the isoform module and the 92 genes of the single-isoform (ERCC) module yields one measure, the sample-specific mean CoD value, which quantifies the coverage uniformity.

CoDs describe the often-hidden biases in sequencing data, predominantly caused by non-homogeneous library preparation but also by subsequent sequencing and mapping. The coverage target-performance comparisons highlight the inherent difficulties in deconvoluting read distributions to correctly identify transcript variants and determine concentrations (Figure 12). Logically, the consequences of the coverage quality influence transcript quantification of the isoform module more than the single-isoform (ERCC) module, where accuracy depends mainly on the mean read counts per transcript length.

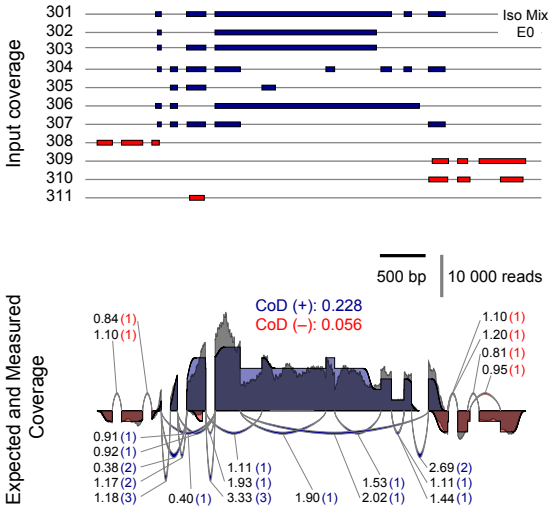


Figure 12. Comparison of the expected and the measured coverages for the SIRV3 locus of the isoform module. Top, individual transcripts of SIRV3 with the exons on the plus strand in **blue** and exons on the minus strand in **red**. Bottom, the expected SIRV3 coverage is shown as transparent **blue** and **red** areas superimposed over the measured transcript coverage after read mapping (shown in **grey**), in which the terminal sites have been modelled by a transient error function. The measured coverages and number of splice junction reads were normalized to obtain identical areas under the curves and identical sums of all junctions for the expected and measured data. The measured splice junction reads are shown by the numbers before the brackets, while the expected values are shown inside the brackets. The CoD values are given for the plus and minus strand in the respective colors. The figure is drawn in the Compact Coverage Visualizations (CCV) format. Intron sequences shared by all transcripts are reduced to small gaps of the same length focusing the visualization on relevant sequences.

The CoD does not allow to distinguish between periodicity and randomness in the biases nor does it forecast how well a data evaluation pipeline can subsequently cope with bias contributions. Nevertheless, smaller CoD values are expected to correlate with a simpler and less error-prone data evaluation. The CoD values can be taken as a first, indicative measure to characterize the mapped data and to compare data sets for similarity right up to this point in the workflow.

Input-Output Correlations

Any calculated abundances can be compared to the known input amounts. Input-output correlations should be calculated in logarithmic space as the set concentration range of the single-isoform (ERCC) module spans 6 orders of magnitude. By these means the relative deviation of low, medium, and highly abundant transcripts are treated equally. The Pearson product-moment correlation coefficient, Pearson's r, should approach 1.

Because the input concentrations of the isoform module are identical, a simple measure of the variance is already sufficient and should approach 0. The distribution of errors (variance) with respect to the individual variants and in the context with competing sequences within the respective genes provides insights into the strengths and weaknesses of the sequencing pipelines.

Precision

Precision measures the scatter of calculated abundance values. Using the technical replicates of identical samples as well as the spike-ins from the entire experiment, the relative standard deviation (RSD) or coefficient of variation (CV) of log2-fold changes (LFC) between the measured and the expected values can be calculated for each SIRV transcript. The overall precision is the mean of all standard deviations of all SIRV RSDs, and can be divided into the precision based on the isoform module and the single-isoform (ERCC) module, respectively. The precision can also be calculated for a certain concentration range, only to reduce the influence of low abundant species with much more scattered abundance values.

Precision can also be determined using the RSD values of endogenous RNA in the concentration range of interest, which depends on the availability of technical replicates.

Accuracy

Accuracy measures the deviation of the calculated abundance values from the expected values and can only be measured using known controls. The accuracy is the median of all LFC moduli. LFC moduli consider relative increases and decreases across the probed concentration range. The accuracy shows the average fold deviation between measured and expected values. Although median, mean, and standard deviation of the LFC moduli describe the distribution of error values, the median is the most robust value against the extent of outliers that can shift when changing certain threshold settings.

The accuracy can be visualized by detailed heat maps, in which each SIRV RNA in the context of competing transcripts can be inspected. Heat maps show the abundances as LFC relative to the expected values. A LFC window of ± 0.11 presents the SIRV confidence interval as a result of the currently achievable accuracy in producing the SIRV mixtures (read more about producing SIRV mixes in the FAQ section on our homepage).

Identifying Detection Limits for Differentially Expressed Transcripts

The experimental analysis of fold change detection as a function of transcript abundance and isoform complexity, requires control results from several defined x-fold ratios that are spread across a wide concentration range. Such data can be obtained by using different mixtures (e.g., from the isoform module SIRV-Set 1 the Isoform Mixes E0, E1, and E2 (Cat. No. 025.03), or from the single-isoform (ERCC) module, the two Ambion™ ERCC ExFold RNA Spike-In Mixes). The combination of different mixes is applicable for pipeline validation experiments, but not for controlling individual sample processing. When using identical controls of the present SIRV-Set 2 or 3, the Analysis of Variance (ANOVA) provides measures for the dispersion of the gene expression measurements as a function of abundance and isoform complexity. Based on exemplary dispersions of the SIRVs the lower boundary for significant fold change measurements can be calculated.

5.7 Experiment Comparisons

CoD, precision, and accuracy are independent quality metrics for the description of NGS pipelines during validation experiments and the characterization of individual experiments. These quality metrics are derived by comparing the experimental results to the expected outcome. Importantly, not only do differences in the RNA input determine the experimental outcome but also any change in the data generation and evaluation pipeline.

While 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 the bias consistency. A head-to-head comparison determines the difference between experiments based on the consistent condensed complexity of the SIRVs. Experiments can be compared pairwise or within entire databases.

The following comparison values can be calculated:

Pairwise Coefficient of Deviation

Similar to the CoD value for one experiment, the CoD can be calculated by comparing the normalized coverages of experiments N1 and N2. Identical biases lead to small values approaching zero in an ideal case.

Concordance

The concordance is the median of all LFC moduli calculated for SIRVs in two experiments, which is essentially the relative accuracy measure calculated by comparing two experiments to each other. High concordances are represented by small values.

Knowing the biases introduced in isoform and single-isoform quantification allows for evaluating whether data sets are comparable across samples or experiments.

5.8 Recommended Software Packages

SIRV Suite

For the evaluation of data for isoform module (Iso Mix E0), the SIRV Suite brings together NGS data that includes SIRV isoform spike-ins, annotations, and data evaluation. The suite can be accessed for free at www.sirvsuite.org. It is hosted on the Galaxy platform (https://usegalaxy.org/). The tools are also available upon request at sirvsuite-support@lexogen.com as CLI (command line interface) and AMI (Amazon Machine Image).

The **SIRV Suite Evaluator** generates the experiment quality metrics and visualizations, which include:

- The ratio between expected and measured reads relative to the reads from the endogenous RNA, which is interpreted as, e.g., mRNA content or experimental variability.

- Compact Coverage Visualization (CCV) graphs, which provide an overview of the normalized expected and measured coverages.
- Calculated Coefficients of Deviation (CoD) as a measure for the deviation between measured and expected coverages.
- Table with normalized counts of identified vs. expected telling junction reads.

Based on the scaled TPM values the Evaluator further calculates:

- Precision, as the mean of all relative SIRV standard deviations (RSD).
- Accuracy, as the median of all log2-fold difference moduli between the measured and the expected relative SIRV concentrations.

The performance diagnostic of measuring control ratios requires the input of different concentration mixtures, e.g., Iso Mix E0, E1, and E2 (Cat. No. 025.03, SIRV-Set 1), to as many different samples and is not applicable when using data from SIRV-Set 2 or 3.

ERCC Dashboard

For the evaluation of reads from the single-isoform module (ERCC) the NIST (National Institute of Standards and Technology) provides a software package called the ERCC dashboard at http://bioconductor.org/packages/release/bioc/html/erccdashboard.html [7].

The ERCC dashboard calculates the following quality metrics based on ERCC Mix 1:

- Estimated mRNA fraction differences for the pair of samples using replicate data.
- Log2-normalized ERCC counts vs. Log2-ERCC spike amount.
- Signal-abundance plot to evaluate dynamic range.

The performance diagnostic of measuring control ratios requires the input of different ERCC concentration mixtures of Ambion™ ERCC ExFold RNA Spike-In Mixes, to as many different samples and is not applicable when using the data from SIRV-Set 3.

6. Downloads

Sequences, annotations, and concentration tables can be obtained from: www.lexogen.com/sirvs/downloads. The FASTA files and corresponding GTF files for SIRV-Set 2 and 3 are provided as either a continuous SIRVome, or as multi-fasta file treating each gene as individual entity. The FASTA and GTF files are included into data analysis pipelines similar to the inclusion of additional single or multiple synthetic chromosomes. The FASTA files contain the complete exon and intron sequences flanked by 1 kb of upstream and 1 kb of downstream sequence. The GTF files contain information about the variant structures.

The following three annotations identified by the letters, C, I, or O in the GTF file name are provided.

_C
Contains the correct annotation of all 69 SIRV isoforms. These SIRVs are shown in blue in the SIRV alignment views in the Appendix.

_I
Contains an insufficient annotation. Here, some SIRV isoforms that are actually present in the mixes are not annotated. These missing annotations are marked with a superscript (-) in the respective Appendix figures.

_O
Contains a representative of a possible over-annotation. Additional SIRV isoforms are annotated, that are not present in the mixes. These transcript variants are shown in green in the respective Appendix figures.

SIRV-Set 2
Sequences and annotations in FASTA and GTF format for SIRV Set 2 (SIRV Isoform Mix E0) are all included in SIRV_Set2_Sequences_170612a (ZIP). The concentration tables together with sequences and detailed statistical information are given in SIRV_Set2_sequence-design-overview-170612a (XLSX).

SIRV-Set 3
Sequences and annotations in FASTA and GTF format for SIRV Set 3 (SIRV Isoform Mix E0 / ERCC) are all included in SIRV_Set3_Sequences_170612a (ZIP). The concentration tables together with sequences and detailed statistical information are given in SIRV_Set3_sequence-design-overview-170612a (XLSX).

ATTENTION: Please access lot-specific download links when available, otherwise access the "Norm" download links.

7. Support

For the latest information and SIRVs-related questions, please write to info@lexogen.com or call Lexogen Support at +43(0) 1 345 1212-41.

8. Safety

Chemical Safety

Follow general safety guidelines for chemical usage, storage, and waste disposal. Minimize contact with chemicals. Wear appropriate personal protective equipment such as gloves and lab coat when handling chemicals. Comply with the RNA handling guidelines when working with SIRVs (see chapter 4.1).

MSDS

SIRV mixes are not a hazardous substance, mixture, or preparation according to EC regulation No. 1272/2008, EC directives 67/548/EEC or 1999/45/EC.

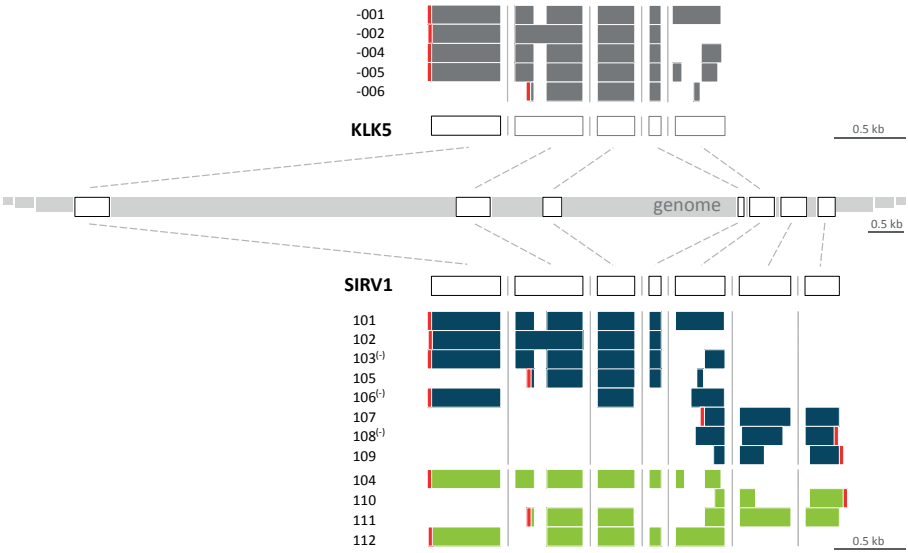
9. References

1. External RNA Controls Consortium. Proposed methods for testing and selecting the ERCC external RNA controls. ***BMC Genomics*** 6, 150 (2005).
2. Baker, S. C. *et al.* The External RNA Controls Consortium: a progress report. ***Nature Methods*** 2, 731–734 (2005).
3. Li, S. *et al.* Multi-platform assessment of transcriptome profiling using RNA-seq in the ABRF next-generation sequencing study. ***Nature Biotechnology*** 32, 915–925 (2014).
4. 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).
5. Xu, J. *et al.* Cross-platform ultradeep transcriptomic profiling of human reference RNA samples by RNA-Seq. ***Scientific Data*** 1, 140020 (2014).
6. Li, S. *et al.* Detecting and correcting systematic variation in large-scale RNA sequencing data. ***Nature Biotechnology*** 32, 888–895 (2014).
7. 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).

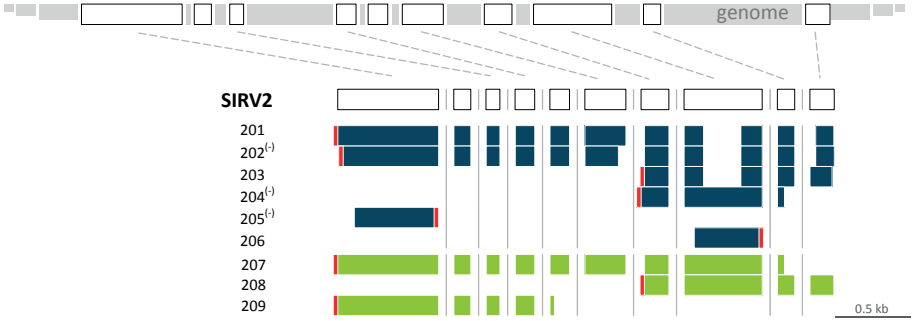
10. Appendix

SIRV Isoforms Alignment View

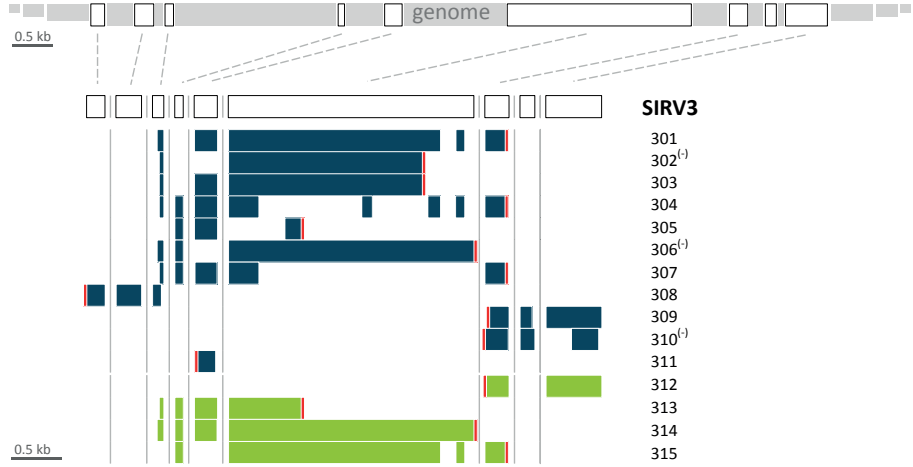
The individual transcript variants of the isoform module are schematically drawn in the condensed intron-exon format (see below) allowing for an overview of the complexity of transcript variants. However, minor start- and end-site variations that differ by just a few nucleotides are not visible in this representation. The spreadsheet summaries or FASTA and GTF files (downloads at www.lexogen.com/sirvs/downloads) are required for detailed viewing.



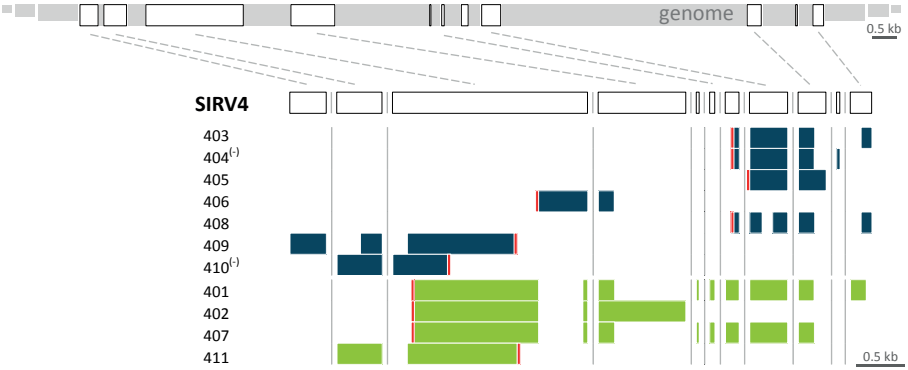
SIRV1 | based on human gene *KLK5*. The human Kallikrein-related peptidase 5 gene was taken as template for SIRV1 gene generation. Its expression is up-regulated by estrogens and progestins, and alternative splicing results in multiple transcript variants encoding the same core protein. The current Ensembl annotation (GRCh38.p2) contains 5 transcript variants, *KLK5-1, 2,* and *4-6*. Its condensed exon-intron structure is shown in the upper section in **grey**. SIRV1 contains 8 real transcript variants (shown in **blue**) present in the mixes. SIRVs marked with a superscript (-) are omitted in the insufficient annotation (SIRV_I). The transcript variants shown in **green** are additional annotations, part of the over-annotation (SIRV_O). The transcript orientations are indicated by the relative position of the poly(A) tail marked in **red**.



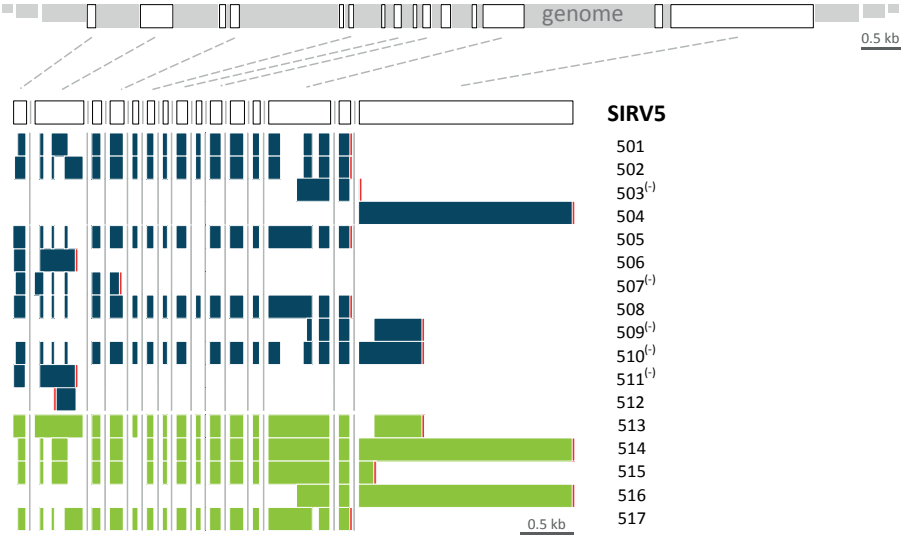
SIRV2 | based on human gene *LDHD* contains 6 transcript variants (shown in **blue**). The transcript variants shown in **green** are additional annotations for alternative evaluation procedures, whereas the SIRVs marked with (-) are missing in the insufficient annotation.



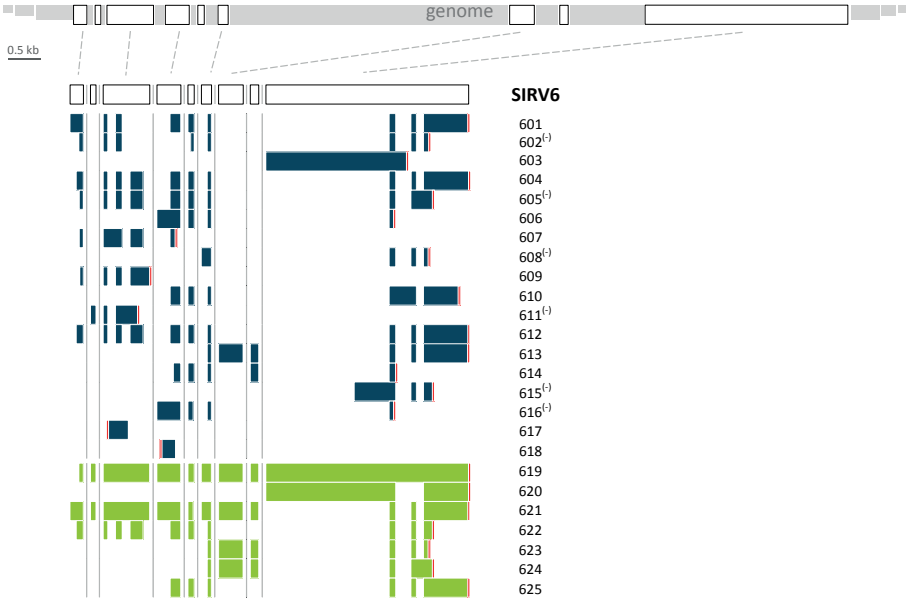
SIRV3 | based on human gene *LGALS17A* contains 11 transcript variants (shown in **blue**). The transcript variants shown in **green** are additional annotations for alternative evaluation procedures, whereas the SIRVs marked with (-) are missing in the insufficient annotation.



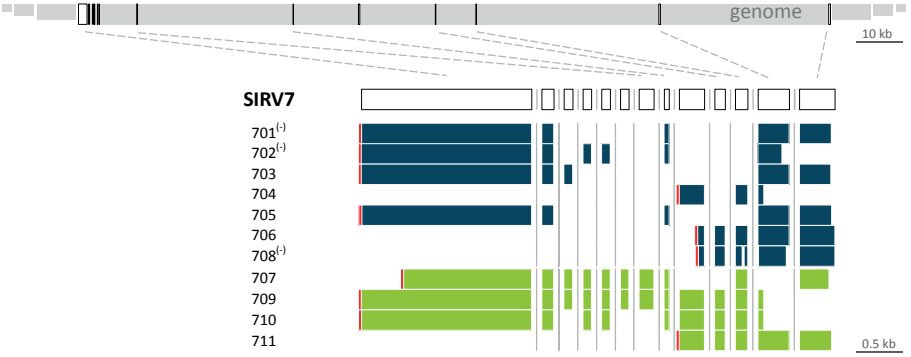
SIRV4 | based on human gene *DAPK3* contains 7 transcript variants (shown in **blue**). The transcript variants shown in **green** are additional annotations for alternative evaluation procedures, whereas the SIRVs marked with (-) are missing in the insufficient annotation.



SIRV5 | based on human gene *HAUS5* contains 12 transcript variants (shown in **blue**). The transcript variants shown in **green** are additional annotations for alternative evaluation procedures, whereas the SIRVs marked with (-) are missing in the insufficient annotation.



SIRV6 | based on human gene *USF2* contains 18 transcript variants (shown in **blue**). The transcript variants shown in **green** are additional annotations for alternative evaluation procedures, whereas the SIRVs marked with (-) are missing in the insufficient annotation.



SIRV7 | based on human gene *TESK2* contains 7 transcript variants (shown in **blue**). The transcript variants shown in **green** are additional annotations for alternative evaluation procedures, whereas the SIRVs marked with (-) are missing in the insufficient annotation.

11. Revision History

Publication No. / Revision Date	Change	Page
050UG134V0100 Jul. 14, 2017	Initial Release.	

Notes

LEXOGEN

Enabling complete transcriptome sequencing

SIRVs · Spike-In RNA Variant Controls
SIRV-Set 2 (Iso Mix E0) and Set 3 (Iso Mix E0 / ERCC)
User Guide

Lexogen GmbH
Campus Vienna Biocenter 5
1030 Vienna, Austria
Telephone: +43 (0) 1 345 1212-41
Fax: +43 (0) 1 345 1212-99
E-mail: info@lexogen.com