LEXOGEN
The RNA Experts

SIRVs™
Spike-in RNA Variant Control Mixes

SIRV-Set 1

Spike-In RNA Variant Controls with Isoforms (Iso Mix E0, E1, E2)

User Guide

Catalog Number:
025 (SIRV-Set 1 (Iso Mix E0, E1, E2))

025UG063V0200

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

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 is subject to Lexogen general terms and conditions (https://www.lexogen.com/terms-and-conditions/) and 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 describing a procedure for publication using this product, please refer to it as Lexogen's SIRVs, Spike-In RNA Variants, SIRV Mixes, or Spike-In RNA Variant Control Mixes. Stating the Catalog Number (Cat. No. 025) 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

The RNA Experts

1. Introduction

1.1 Spike-in RNA Controls

RNA sequencing (RNA-Seq) workflows include RNA purification, library generation, sequencing itself, and the evaluation of sequenced fragments. The initial steps impose biases, which data processing algorithms try to compensate for 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 development of new RNA-Seq compatible platforms and protocols has created the need for multifunctional spike-in controls, which are integrated and processed with the samples to enable monitoring and comparison of key performance parameters like sensitivity and input-output correlation as well as the detection and quantification of transcript variants (Fig. 1).

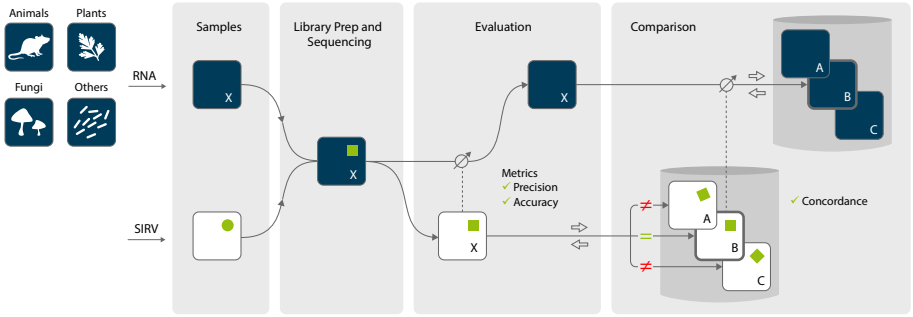

Figure 1. Workflow for using spike-in controls in RNA-Seq. Spike-In RNA Variants (SIRVs) are defined synthetic RNA molecules that mimic the main aspects of transcriptome complexity. They are added in minuscule amounts to samples before library preparation to undergo the very same processing steps as the endogenous RNA (**green** circles and squares in Samples and Library Prep and Sequencing). After mapping the reads to the combined genome, the spike-in data are used to derive quality metrics and to categorize the experiments (see Evaluation). 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 (**green** "equal" and **red** "does not equal" signs, respectively).

The Spike-In RNA Variants (SIRVs) were conceived as a family of modules to offer tailored solutions for the control of RNA-Seq experiments. SIRVs are available as an isoform module, which contains a group of synthetic transcripts that mimic transcriptome complexity, and as a length module to cover transcript lengths of up to 12 kb (Fig. 2). While the SIRV isoform module is available as a stand-alone module (Cat. No. 025 and 050) or mixed with ERCCs to additionally mimic abundance complexity (Cat. No. 051), the long SIRVs module is provided in a mix together with the SIRV isoform module and the ERCC module (Cat. No. 141).

Figure 2. SIRV modules. The SIRV Isoforms, single-isoform transcripts (ERCCs), and long SIRVs are established synthetic RNA molecules that mimic three aspects of transcriptome complexity, isoforms, abundance, and transcript length. The SIRVome is the corresponding artificial reference genome.

1.2 SIRV Isoforms: Isoform Complexity

The SIRV isoform module 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 that mimic the splicing characteristics of 7 human model gene loci, 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 (Fig. 3). For the sake of simplicity, all of these transcriptional variants are referred to as isoforms. Each SIRV genetic locus contains between 6 and 18 transcript isoforms [1].

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 database entries but match eukaryotic gene features in terms of their sequence 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 (Fig. 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.

Between 6 and 18 transcript variants were designed and produced for each of the model genes. The mRNAs range from 191 to 2,528 nt with a GC content of 29.5 - 51.2 %, with the shortest mRNAs being antisense monoexonic transcripts.

	Alternative 1st exon	Start site variation	Alternative 5' splice site	Alternative 3' splice site	Exon skipping	Exon splitting	End site variation	Alternative last exon
SIRV1	5	4	5	2	2	3	4	1
SIRV2	1	3	3	2	0	3	2	2
SIRV3	1	5	5	4	5	4	7	4
SIRV4	4	2	2	4	2	1	5	3
SIRV5	3	9	6	8	5	17	7	7
SIRV6	9	10	7	26	27	28	13	3
SIRV7	2	5	1	1	31	1	4	3

Table 1. Summary of alternative splicing and transcription variations for each SIRV per genetic locus. The occurrences of the different events are counted for each transcript in reference to a hypothetical master transcript of maximal length containing all exon sequences from all transcript variants of a given genetic locus. Therefore, in a formal sense, no intron retention can occur, but this event is defined as exon splitting caused by the introduction of an intron sequence.

Exonic sequences were created from a pool of database-derived genomes, modified to lose identity, whereas intronic sequences were randomly generated, accounting for variable GC content. These SIRV sequences were tested by blasting against the NCBI database on the nucleotide and protein level, whereby no significant similarities were found. SIRV reads of an *in silico* NGS experiment (FLUX generator) map very well to the "SIRVome" but hardly to model genomes (human, mouse, Drosophila, Arabidopsis, ERCC, etc.). Since off-target mapping is *de facto* absent, the artificial SIRV sequences can be used for qualitative and quantitative assessment in the context of known genomic systems and in conjunction with ERCC sequences.

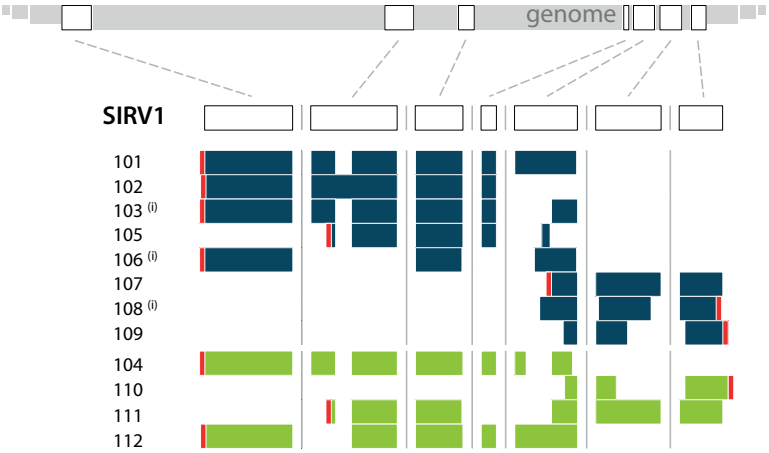

Figure 4. SIRV isoform design exemplified by SIRV1. The SIRV1 locus was derived from the human *KLK5* gene, with transcripts added to the Ensembl annotations to generate comprehensive transcriptome complexity. All original and derived gene structures are shown in the Appendix. Transcripts in **blue** are part of SIRV mixes and transcripts in **green** are only part of an over-annotation reference SIRVome available for pipeline validation. (i) refers to transcripts that are omitted in an incomplete reference 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 locus (gene) expression values.

1.3 SIRV Sets

The modular structure of available SIRV controls (SIRV isoforms, ERCCs, and long SIRVs) enables these to be used in specific combinations to probe the different dimensions of transcriptome complexity. We use the following definitions:

Module Group of spike-in controls that mimic predominantly one aspect of transcriptome complexity.

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

Set Term for the combination of mixes or modules.

The overview of currently available sets are shown in Table 2.

Table 2. 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-Set 1 is covered in this User Guide. *Refers to the number of vials, 1 or 3. The ERCC module includes ERCC Mix 1[3,4].

		SIRV-Set 1	SIRV-Set 2	SIRV-Set 3	SIRV-Set 4
Cat. No.		**025.03**	**050.0***	**051.0***	**141.0***
Module(s)	Isoforms	Isoform Mixes E0, E1, E2	Isoform Mix E0	Isoform Mix E0	Isoform Mix E0
	ERCC	✗	✗	ERCC Mix 1	ERCC Mix 1
	Long SIRVs	✗	✗	✗	long SIRVs
Property	Isoform detection and quantification	✓	✓	✓	✓
	Dynamic range	partially	✗	✓	✓
	Length >2.5 kb	✗	✗	✗	✓
Applications	Pipeline Validation	✓	partially	partially	partially
	Sample Control	✗	✓	✓	✓
Number of spike-in transcripts in each mix		69 (69 isoforms in each mix)	69 (69 isoforms)	161 (69 isoforms, 92 ERCCs)	176 (69 isoforms, 92 ERCCs, 15 long SIRVs)

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 1 (Cat. No. 025) includes the three SIRV Isoform Mixes E0, E1, and E2; with each mix containing all 69 SIRVs but in different concentration ratios. This set is designed for the validation of concentration measurement (including fold change) with isoform resolution.

SIRV-Sets 2, **3**, and **4** contain single mixes, that feature different combinations of the three spike-in modules. SIRV isoforms and long SIRVs are included at equimolar ratios in these mixes; ERCCs are present in a pre-determined concentration gradient. These sets can validate sensitivity, isoform, and length aspects in addition to being spiked into every RNA-Seq sample for controlling the consistency of sample processing and measurement (Table 2). All quality metrics, except fold change measurements, can be determined experimentally for each individual sample. Further details for the use of these SIRV sets can be found in the User Guide for SIRV-Set 2, 3, and 4 (050UG134).

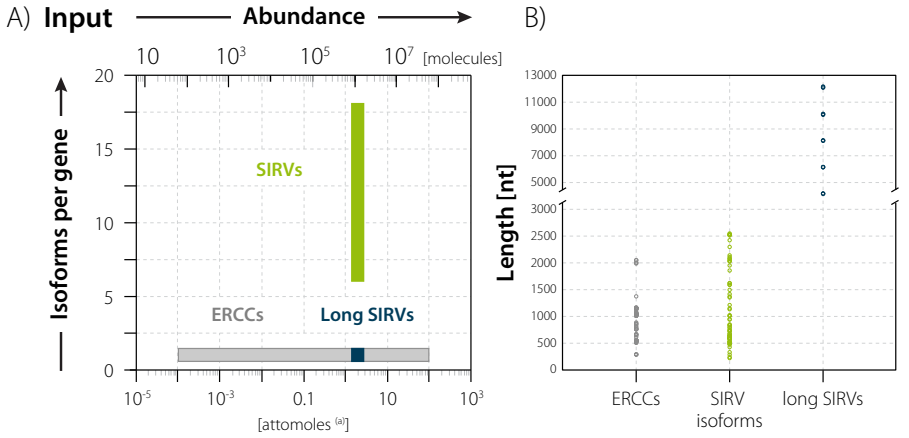

Figure 5. A) Isoform and abundance complexity and B) length complexity. The SIRV isoforms and ERCC transcripts in SIRV-Set 3 control for the two main dimensions of transcriptome complexity: isoforms and abundance. SIRV-Set 4 additionally contains controls for length complexity. The isoform module with 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 (**gray 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 approximately 1 % of the mRNA-Seq reads. Long SIRVs (**blue bar**) contain 1 transcript per gene and are present at equimolar concentrations in SIRV-Set 4. B) Transcripts of the ERCC module range up to 2 kb in length, and the ones of the SIRV isoform module up to 2.5 kb. The long SIRV module contains three transcripts in each of the length categories 4 kb, 6 kb, 8 kb, 10 kb, and 12 kb.

1.4 SIRV Mixes E0, E1, and E2

The isoform module available in SIRV-Set 1 (Cat. No. 025) is provided in 3 SIRV mixes: E0, E1, and E2, with each mix containing all 69 SIRV isoform transcripts but in different concentration ratios. The RNA purity and individual concentrations are measured by absorbance spectroscopy (Nanodrop by Thermo Scientific) and verified by capillary electrophoresis (Bioanalyzer by Agilent Technologies), which is also used to determine the molecular integrity. The concentration ratios between transcripts within each mix are as follows:

E0	identical (1:1)
E1	cover approximately one order of magnitude (up to 1:8)
E2	extend over more than two orders of magnitude (up to 1:128, Fig. 6)

The total molarity of each mix is close to 69.5 fmol/µl. The concentration of each mix is 25.3 ± 0.1 ng/µl (see Table 3).

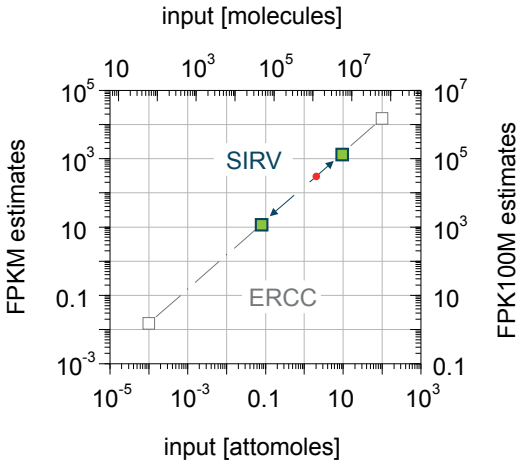

Figure 6. Dynamic concentration range of SIRVs. The maximal dynamic range is given by the SIRV transcripts of Isoform Mix E2 with minimal and maximal concentrations (**green** boxes). These are depicted in an input-output correlation based on the known input and a FPKM (fragments per kilobase of exon per million fragments mapped) estimation for SIRV mixes in reference RNA background samples which have been prepared according to the suggested mass ratios in Example Eq. 1. As reference, the concentration range of the ERCC mix is also shown (**gray boxes**). In contrast, in SIRV Isoform Mix E0, all concentrations are identical which is shown by the **red** dot. The FPKM value is an estimate of the number of reads to be expected for a 1 kb long transcript at 1 M reads, and the FPK100M corresponds to the number of reads to be expected for a 1 kb long transcript in an experiment analyzing 100 M reads.

Each SIRV isoform mix (E0, E1, and E2) is composed of 4 individual submixes. Each SIRV transcript is included in one of 4 submixes. The 4 submixes are then combined in 3 different relative ratios within each mix, as shown in Figure 7. The concentration of each SIRV within each submix is equimolar, and consequently, the relative concentration ratios of SIRVs within the same submix, are preserved in the E0, E1, and E2 mixes (see Appendix C, p.30).

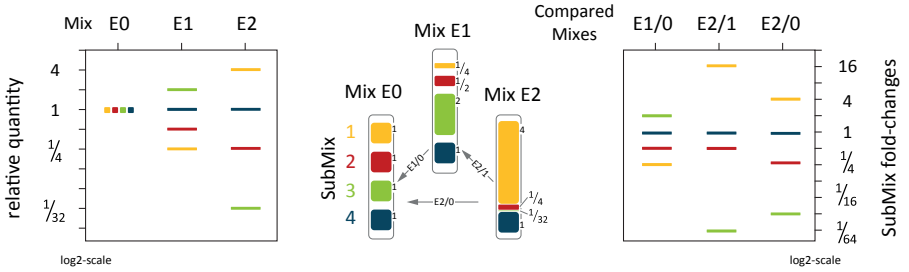

Figure 7. Graphical representation of the submix (1 - 4) distribution in the 3 SIRV mixes and the resulting intra- and inter-mix ratios. The 4 submixes are represented by different colors and contain between 12 and 21 SIRVs to keep the total molarity and weight of the 3 mixes evenly balanced at 69.5 fmol/µl and 25.3 ng/µl. Left, the intra-mix concentration ratios provide three different concentration settings to evaluate the accuracy in relative concentration measurements. Right, the preset fold changes allow for 3 possible inter-mix comparisons to evaluate differential gene expression measurements. For further details see Appendix C, p.30

The assignment of SIRVs to the 4 submixes was optimized in such a way that the final total mass as well as the final total molarity in all 3 mixes are the same. Because of the various transcript lengths, the number of transcripts within each submix is not equal. Each submix contains between 12 and 21 SIRVs. The distribution of transcript variants is as diverse as possible so each SIRV genetic locus is represented in each submix with at least one transcript variant (details in Appendix A, p.25). The *a priori* knowledge of SIRV abundance in the SIRV Isoform Mixes E0, E1, and E2 enables the assessment of differential gene expression (DE) measurement accuracy, based on transcript variant identification, quantification, and variance of technical repeats. The SIRVs from submix 4 are always present at the same concentration and serve as controls for false positive detection in DE evaluation pipelines.

NOTE: Although the 1/64-fold change provides for the most distinct DE value, the SIRV submix 3 concentrations in SIRV Isoform Mix E2 are the lowest in the entire sample set and they are the hardest to determine correctly at low read depths.

2. Kit Components and Storage

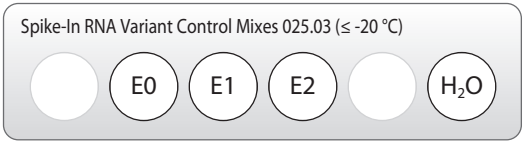

Figure 8. Location of the kit components (Cat. No. 025.03).

Each SIRVs box contains 3 tubes labeled E0, E1, and E2. Each mix contains all 69 SIRV transcripts. The total molarity of the mixes is close to 69.5 fmol/µl, and the concentrations are set to 25.3 ± 0.1 ng/µl. Small variations are the result of setting the molarities and molar ratios of individual SIRVs in well-spaced values within the dual system;. The precise values are shown in Table 3. The individual concentrations of each SIRV in all mixes can be obtained from www.lexogen.com/sirvs/download.

The tubes must be stored at, or below, -20 °C. We recommend to aliquot the solution upon first time usage to avoid freeze-thaw cycles. Follow guidelines in chapter 4.

Table 3. Content of the tubes.

		Concentration	
Component	**Content [µl]***	**[fmol/µl]**	**[ng/µl]**
Mix E0	4	69.0	25.2
Mix E1	4	68.5	25.2
Mix E2	4	70.8	25.4

(*) Each tube is filled with nominal 4.4 µl to ensure safe aliquoting of 4 × 1 µl.

Inert additives are included in the solution that stabilize SIRVs sufficiently to enable one freeze-thaw cycle. Unless the entire amount of SIRVs is immediately required, we recommend to aliquot the solution upon first time usage. You can draw 4 times 1 µl. Additional aliquots not required immediately should be further stored at or below -20 °C.

3. General

3.1 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 can lead to fragmentation and loss of SIRVs.

3.2 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).

3.3 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.

4. Detailed Protocol

4.1 Preparation

The number of reactions possible per mix, depends on the spike-in amount required per sample. The provided SIRV stock solution (1 µl aliquots) should first be stepwise diluted for use at the optimal working concentration. The optimal working concentration of SIRVs will depend on the exact experimental conditions, and can be determined using our SIRV-Set 1 Spike-in Calculator Worksheet is available via https://www.lexogen.com/sirvs/download (see also chapter 4.4 for equations).

ATTENTION: SIRV dilutions are recommended for single-use only. Storage and freeze-thawing of diluted SIRVs should be avoided as it contributes significantly to alteration of the RNA integrity and concentration.

When diluting SIRVs to the required working concentration, please observe these important guidelines:

- Always spin down tubes before use.
- Any dilutions must be prepared immediately before adding the SIRVs to your RNA samples.
- Plan the dilution of the SIRVs and the amount to spike-in as one continuous workflow. Minimize the time RNA is kept at low concentrations (minutes instead of hours).
- Work with ice-cold solutions on a cool block (at 0 - 5 °C) or on ice. Do not use cool blocks at temperatures below 0° C.
- For dilutions, use RNase-free buffers. Recommended buffers are sodium citrate at pH 6.4 or Tris-EDTA at pH 7.0. **Divalent cations should be avoided as buffer components!**
- 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.
- 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.
- 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).
- Depending on the amount of RNA that is targeted by the SIRV spike-in, the dilution should ideally be at least 1:100 or higher.

4.2 Aliquoting and Interim Storage

The contents of each SIRV tube can be divided into 4 x 1 µl aliquots, for use in independent experiments.

1	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.
2	Freeze the remaining 3 aliquots immediately at ≤-20 °C for later use.
3	Proceed with one 1 µl aliquot and perform subsequent dilution step(s) in short succession.

4.3 Spiking of RNA Samples

The workflow is easily adjusted to any type and amount of RNA sample and consists of 3 main steps:

1	In the formulae below (see 4.4, Eq. 1 and 2), 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.

4.4 Determining the Amount of SIRVs to Spike-in

The SIRV mixes can be used with crude cell extract, homogenized cells or tissues, purified total RNA, rRNA depleted RNA, or poly(A) enriched RNA. The spike-in ratios have to be chosen in concordance with the desired final SIRV content (i.e., final SIRV read share). The final results depend on several factors such as the expression state of cells, the quality and integrity of the RNA, as well as the kind of NGS library preparation. The accuracy of spike-in experiments depends on correct volumetric dilution series, thorough mixing, and careful handling of dilutions.

The SIRV mixes in their current format cover a smaller concentration range compared to the ERCC mixes, with 2 versus 6 orders of magnitude (Figure 8). Higher transcript coverage rates will increase the chance of correctly distinguishing variants. Whereas one read can be sufficient to map an ERCC sequence, the high sequence identity of the isoforms of a given SIRV gene will require significantly more reads. Lower spike-in ratios and/or lower read depths can always be simulated by downsampling to estimate how well a certain sequencing pipeline can cope with lower coverages.

All SIRV mixes have nearly identical total mass concentrations as well as highly similar molar concentrations to allow for an easy first design and for a comparative evaluation of experiments.

Equations Eq. 1 and Eq. 2 (below) can be used to derive the appropriate SIRV amount for a given sample, and to determine the required dilution factor and pipetting volume.

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 to which SIRV RNA will be added.
C_{SIRV}	concentration of SIRVs at the suitable dilution.
V_{SIRV}	volume to be used in the spike-in procedure.

These equations can be performed using Lexogen's SIRV Calculation Worksheets, available from www.lexogen.com/sirvs/download/.
The following examples show-the use of these equations, which can easily be adjusted to other-RNA-Seq experiments.

Example Eq. 1

Assuming a starting input amount of 100 ng of UHRR total RNA, the mass of SIRVs (m_{SIRV}) to be used in one spike-in experiment is estimated by multiplying the final fraction of desired SIRV reads ($F_{SIRV\ reads}$) by the targeted RNA fraction ($F_{target\ RNA}$) and the total RNA input mass (m_{RNA}).
In this example:

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

F SIRV reads

The final fraction of desired SIRV reads ($F_{SIRV\ reads}$) is usually 0.01 (or 1 %). At this final read share, standard short-read workflows will resolve most (but not all) SIRV isoforms at the transcript level, sufficient for the calculation of RNA-Seq quality parameters (see also chapter 5.2).
However, the $F_{SIRV\ reads}$ can be adjusted depending on the intended application. Larger spike-in ratios may suit workflow validation, or long-read sequencing applications where higher SIRV read percentages are desired. Lower spike-in ratios are recommended when target transcripts cover a lower abundance range, samples have low RNA complexity, or input RNA is derived from highly degraded or modified sample types (e.g., Formalin-Fixed Paraffin Embedded tissues, FFPE).

F target RNA

The fraction of the targeted RNA ($F_{target\ RNA}$) depends on sample type, RNA integrity, and the experimental design. 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 mRNA content of Human Brain Reference RNA (HBRR, Ambion) is approximately 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 3' tag-based methods).

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 (following the guidelines in chapter 4.1), the concentration reaches 25.2 pg/µl (SIRV Isoform Mix E0), and the volume needed to spike-in 30 pg SIRVs from SIRV Isoform Mix E0 is then 1.19 µl.

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

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

4.5 Considerations for Library Preparation and Sequencing

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 each contain a 30 nt long poly(A) tail (compared to a slightly shorter and variable poly(A) tail of 24 ± 1.05 nt for ERCC transcripts). All of these poly(A) tails allow for poly(A) enrichment and oligo(dT)-priming. SIRVs do not have a 5'-cap structure (5'-m7G) but a 5' triphosphate end and are resistant to 5'- 3' exonucleases. Therefore, the use of SIRVs for cap-specific cDNA preparation methods will result in incomplete library generation from SIRV RNA. SIRVs are also not recommended for any exon-capture, or target-capture applications unless probes specific to the SIRV transcripts will be included in the target capture probeset. For further questions on suitable applications, please contact support@lexogen.com.

RNA-Seq libraries should be sequenced with sufficient read depth to overcome certain coverage thresholds outlined in Figure 8. Different lower read depth thresholds need to be considered when using full-length, single molecule sequencing, or (3') tag sequencing methods. In these methods each transcript or amplicon is represented at most by one single read and not by numerous reads as a function of transcript length. In these cases, the molar and not the mass ratio is of relevance.

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 9.

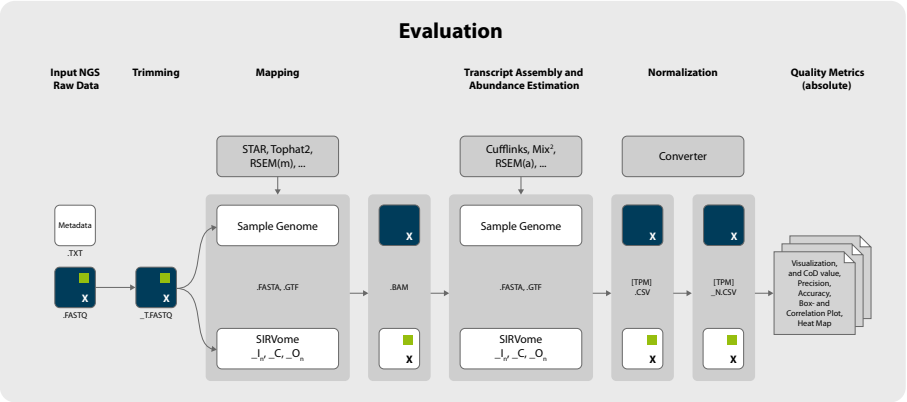

Figure 9. SIRV data evaluation scheme. SIRV reads undergo the very same processing steps as reads derived from the RNA sample. CoD, Coefficient of Deviation.

Stages of Data Evaluation

1. All reads are quality- and barcode-trimmed and then mapped to a reference combining sample genome and SIRVome (see Appendix B, p.29 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 10).

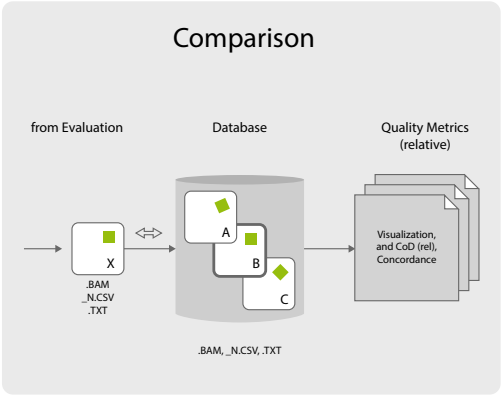

Figure 10. SIRV data comparison scheme. The control data set is used to carry out pairwise comparisons between experiments using the small subset of SIRV control data. Data sets of high concordance can be selected and the extent of expected error rates can be estimated before making a decision about comparing the complete data sets.

5.2 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[1].

The **precision** (random error) in quantifying 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.

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 single-isoform ERCCs. While SIRV Isoform Mixes E1 and E2 provide the SIRV isoforms in concentration ranges of 1 and 2 magnitudes, respectively (see Figure 8).-The task of identifying a given isoform is not confounded by differing input concentrations in SIRV Isoform Mix E0. The gene coverage of the isoform module in relation to the single-isoform ERCC module and long SIRVs is shown in an exemplary series in Figure 11.

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

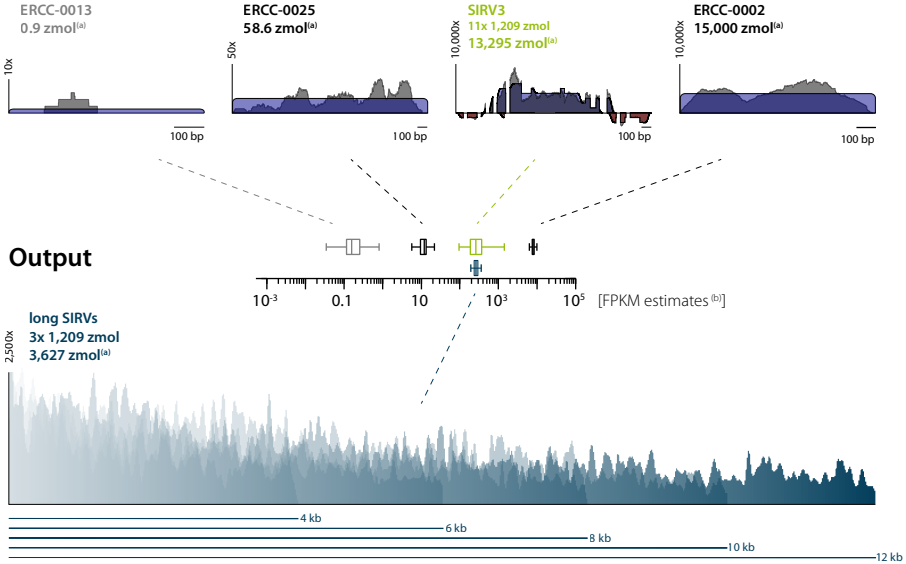

Figure 11. Read coverages of SIRV isoform and single-isoform modules (ERCC, long SIRVs) depend on input concentration, library preparation efficiency, biases, and read depth. Three ERCC examples of different input abundances are shown. With 6 - 18 isoforms mapping to the 7 SIRV genes, NGS read assignment 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** represent areas the expected coverage in the antisense direction. The **gray** areas show exemplary coverages from one stranded library preparation that has been sequenced in paired-end mode. [a] The number of zettamoles (zmol) 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 long SIRVs are present at concentrations identical to the SIRV isoforms, but quantification is not affected by isoform complexity, resulting in a smaller standard error. The lower panel shows the read coverage of all long SIRVs as a result of sequencing CORALL™ Total RNA-Seq libraries. Coverages are averaged for each length category.

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.

5.3 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 are 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 precise and accurate pipetting during SIRV dilution and spike-in procedures.

5.4 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.

5.5 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 5.7).

5.6 Use of the Different SIRV Annotations

SIRV reads should be mapped initially using the correct **SIRV_C** annotation (see Appendix B, p.29 for downloads). However, when variant identification and quantification is the goal, 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. While ERCCs and long SIRVs are annotated correctly, 25 of the 69 SIRV isoforms are missing. 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 SIRV isoforms are annotated than are actually contained in the samples. This reflects i) situations where transcript variants were discovered in other tissues or ii) in the same tissue but at different developmental stages, iii) the occurence of falsely annotated variants, and iv) the annotation of 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 and long SIRVs modules.

5.7 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 (see chapter 5.3).

*) In the isoform and long SIRVs modules, 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, where applicable, 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 single-isoform modules (ERCCs and long SIRVs), where accuracy depends mainly on the mean read counts per transcript length.

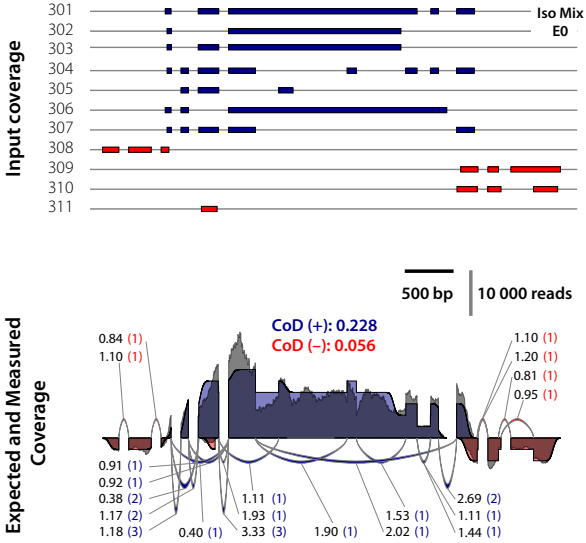

Figure 12. Comparison of the expected and the measured coverages for the SIRV3 locus of the isoform module. Top, individual transcripts of SIRV3 with the exons on the plus strand in **blue** and exons on the minus strand in **red**. Bottom, the expected SIRV3 coverage is shown as transparent **blue** and **red** areas superimposed over the measured transcript coverage after read mapping (shown in **gray**), 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. 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.
In SIRV Isoform Mix E0, the input concentrations of the SIRV isoforms are identical, hence 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 genetic loci, 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 assessed for the isoform module and the single-isoform (ERCC and long SIRVs) modules, 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 Isoform Mixes E0, E1, or

E2 for comparison, or the two Ambion™ ERCC ExFold RNA Spike-In Mixes 1 and 2 (Thermo Fisher Scientific, not included in Lexogen's SIRV-Sets)[4]. 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 1, 2, 3, or 4, 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.

5.8 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 datasets are comparable across samples or experiments.

6. Appendix A: SIRV Isoform 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 (via www.lexogen.com/sirvs/download) are required for detailed viewing.

The individual SIRVs are present in predefined amounts in mixes E0, E1, and E2. Their relative ratios are given alongside of the variant structure in each figure. The SIRV concentrations within a given mix are either equal (E0), differ up to 8-fold (E1), or up to 128-fold (E2).

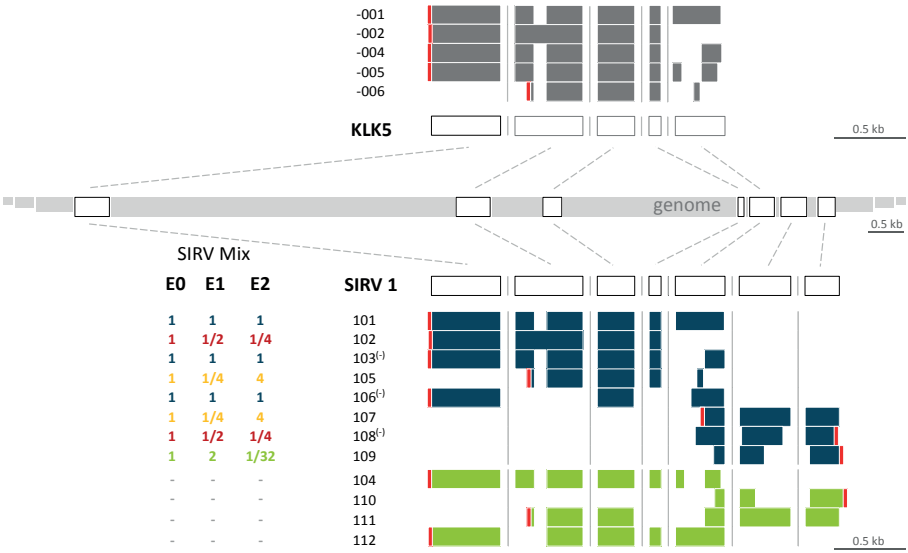

Figure 13. SIRV 1 | based on human gene *KLK5*. The human Kallikrein-related peptidase 5 gene was taken as template for SIRV1 genetic locus generation. Its expression is up-regulated by estrogens and progestins, and alternative splicing results in multiple transcript variants encoding the same core protein. 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 **gray**. SIRV 1 contains 8 real transcript variants (shown in **blue**) present in the mixes in the respective relative ratios as shown in the table to the left. 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**.

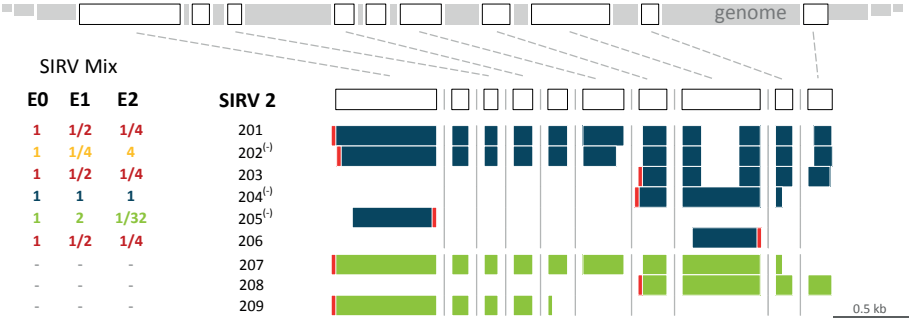

Figure 14. SIRV 2 | based on human gene *LDHD* contains 6 transcript variants (shown in **blue**) which are present in the mixes in the respective relative ratios as shown in the table to the left. The transcript variants shown in **green** are additional annotations for alternative evaluation procedures, whereas the SIRVs marked with (-) are missing in the insufficient annotation.

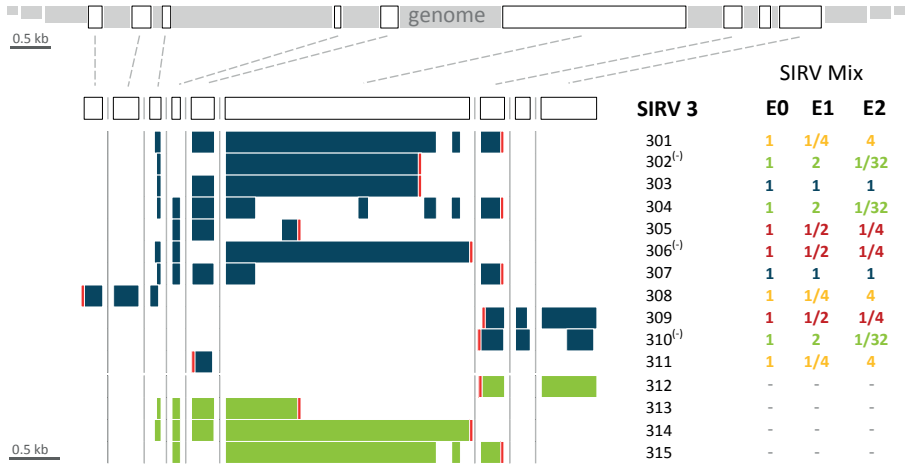

Figure 15. SIRV 3 | based on human gene *LGALS17A* contains 11 transcript variants (shown in **blue**) which are present in the mixes in the respective relative ratios as shown in the table to the right. The transcript variants shown in **green** are additional annotations for alternative evaluation procedures, whereas the SIRVs marked with (-) are missing in the insufficient annotation.

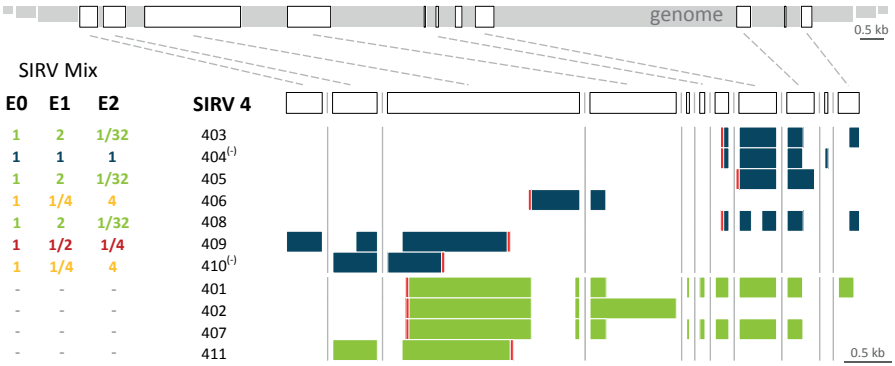

Figure 16. SIRV 4 | based on human gene *DAPK3* contains 7 transcript variants (shown in **blue**) which are present in the mixes in the respective relative ratios as shown in the table to the left. The transcript variants shown in **green** are additional annotations for alternative evaluation procedures, whereas the SIRVs marked with (-) are missing in the insufficient annotation.

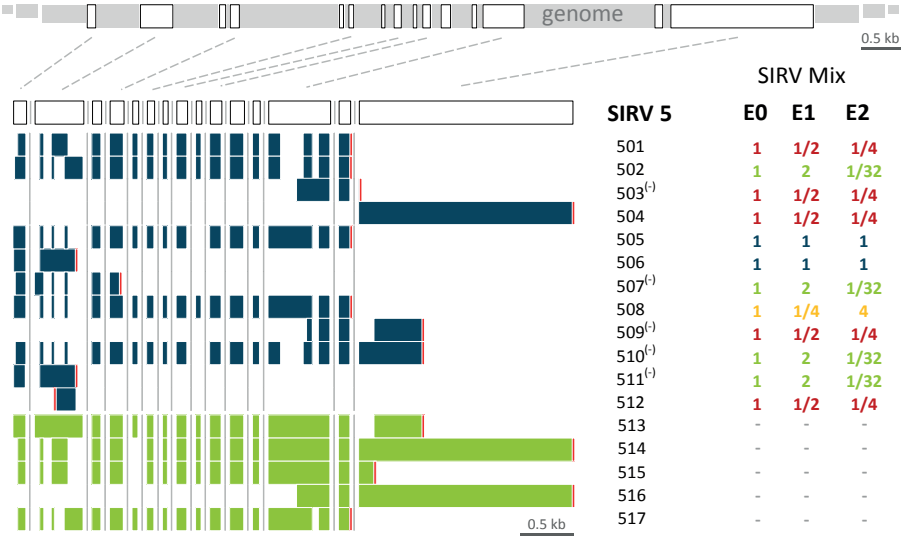

Figure 17. SIRV 5 | based on human gene *HAUS5* contains 12 transcript variants (shown in **blue**) which are present in the mixes in the respective relative ratios as shown in the table to the right. The transcript variants shown in **green** are additional annotations for alternative evaluation procedures, whereas the SIRVs marked with (-) are missing in the insufficient annotation.

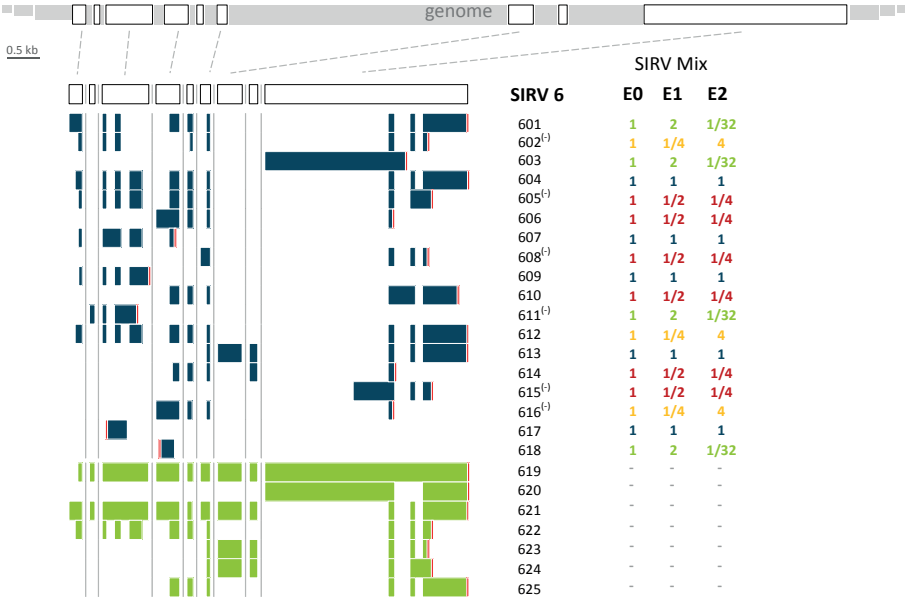

Figure 18. SIRV 6 | based on human gene *USF2* contains 18 transcript variants (shown in **blue**) which are present in the mixes in the respective relative ratios as shown in the table to the right. The transcript variants shown in **green** are additional annotations for alternative evaluation procedures, whereas the SIRVs marked with (-) are missing in the insufficient annotation.

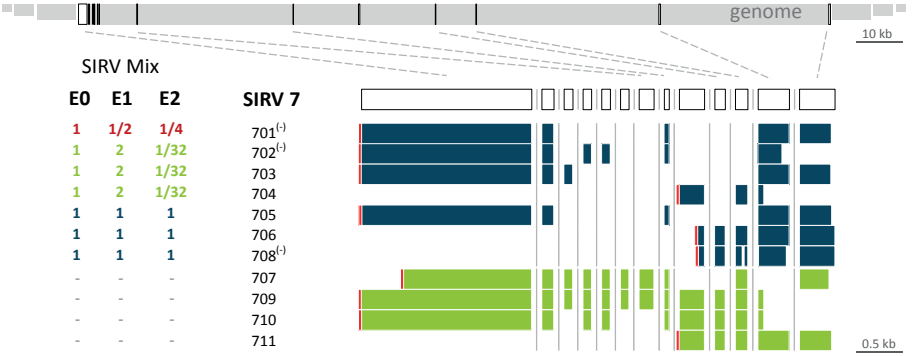

Figure 19. SIRV 7 | based on human gene *TESK2* contains 7 transcript variants (shown in **blue**) which are present in the mixes in the respective relative ratios as shown in the table to the left. The transcript variants shown in **green** are additional annotations for alternative evaluation procedures, whereas the SIRVs marked with (-) are missing in the insufficient annotation.

7. Appendix B: Downloads

Sequences, annotations, and concentration tables can be obtained from: www.lexogen.com/sirvs/download. The "_sequence-design-overview_" (XLSX) file and "_Sequences_" ZIP archive files are provided under the "SIRV annotations, concentrations and sequence features" section. The overview XLSX file contains the transcript concentration tables and detailed statistical information for each SIRV. The ZIP archive contains all of the sequence and reference annotation files for the SIRVs in GTF and FASTA formats. A readme file is also included that outlines the files included in each ZIP archive.

The FASTA files and corresponding GTF files for SIRV-Set 1 are provided as either a continuous SIRVome, or as multi-fasta files treating each gene / locus as individual entities. The FASTA and GTF files can be included into data analysis pipelines similar to the inclusion of additional single or multiple synthetic chromosomes. FASTA files contain the complete exon and intron sequences flanked by 1 kb of upstream and 1 kb of downstream sequence. GTF files contain information about the variant structures.

ATTENTION: There are specific file sets and annotations for different lot numbers. Please check the lot number to identify the correct annotation to use. Please use lot-specific file sets when applicable, otherwise use the "Norm" file sets. For some lot numbers, an additional "Amendment" (PDF) file is provided. These files indicate any deviations in the SIRV composition present within this lot. For any questions regarding SIRV lot amendments please contact support@lexogen.com.

Additional SIRV Reference Annotations

The individual SIRV-Set file sets include the **correct** reference annotation for all SIRV isoforms. However, to additional **insufficient** and **over-annotated** reference GTF files are also available via www.lexogen.com/sirvs/download, under the "Additional annotations" section. The annotation type is defined in the filename as follows:

- **_C:** Contains the **correct** annotation of all SIRV isoforms.
- **_I:** Contains an **insufficient** annotation. Here, some SIRV isoforms that are actually present in the mixes are not annotated in the reference.
- **_O:** Contains a representative of a possible **over-annotation**. Additional SIRV isoforms are annotated, that are not present in the mixes.

SIRV spike-in calculation worksheets are also available for download from: www.lexogen.com/sirvs/download. These enable calculation of the amount of SIRV RNA required to spike-in per sample, as well as the optimal working concentration and number of samples that can be prepared using each tube of SIRVs.

8. Appendix C: Quality Parameters

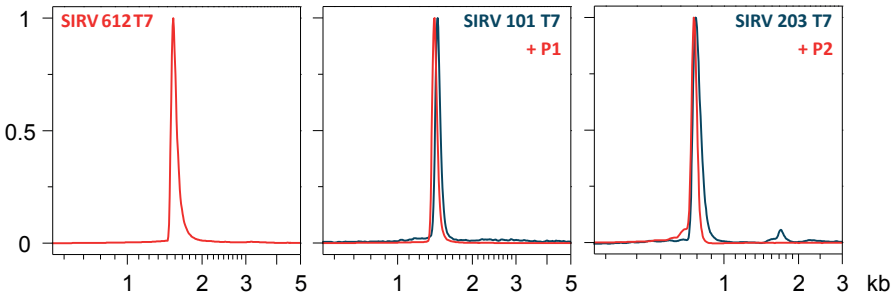

Figure 20. Examples for T7 transcription and purification of SIRVs. Left, T7 transcription of SIRV612 produced RNA of almost uniform correct length with 4/92/4 % in the pre-, main-, and post-peak fractions. Middle, SIRV101 T7 transcription shared a purity of 11/83/6 %, which could be increased to 4/95/1 % using purification method P1. Right, SIRV203 T7 transcription displayed a purity of 3/88/9 % with distinct longer sequence artifacts, which could be removed by purification method P2 to obtain the final product with 11/89/0 % . All SIRVs contain ≥85 % in the main peak. All % are w/w.

Purification of T7 Transcription Products

T7 transcription-produced SIRV RNAs are of high purity and high, although varying, integrity as determined by RNA length evaluation using capillary electrophoresis (Bioanalyzer RNA 6000 Pico Chip, Agilent). A series of optimized, tailored methods are applied to purify full-length RNAs with minimal amounts of side products. Examples are shown in Figure 20.

Determination of SIRV Integrity

Within limits, Bioanalyzer traces are good measurements for the integrity of SIRVs. Using a high resolution inspection of the pre-peak, main-peak, and post-peak, areas are quantified to be:

pre-peak	7.36 ± 3.43 %,
main-peak	90.31 ± 3.72 %,
post-peak	2.36 ± 3.04 % (all w/w),

The values given are the mean of the relative fractional mass content ± standard deviation.

The manufacturer's specification for the RNA 6000 Pico Chip Kit are 20 % CV for reproducibility of quantitation, 30 % CV for quantitation accuracy.

Quantification of SIRVs

SIRV solutions are measured by absorbance spectroscopy (Nanodrop, Thermo Scientific) and stock solution concentrations are adjusted to ≥50 ng/µl. The ratios of absorbance at 260 nm to 280 nm and 260 nm to 230 nm indicate the highest RNA purity.

$A_{260\ nm/280\ nm}$ 2.14 ± 0.12,
$A_{260\ nm/230\ nm}$ 2.17 ± 0.20

Nanodrop measurements allow for precise RNA quantification. Error, according to the manufacturer's specification, is ± 2 ng/µl for nucleic acid samples ≤100 ng/µl. The relative error for quantification of the final SIRV stock solutions, with concentration measurements near 50 ng/µl, is ± 4 %.

The molarity of each solution is calculated from the base distribution of the SIRV sequences according to:

MW [g/mol] =A*329.2 + U*306.2 + C*305.2 + G*345.2 + 159

9. Appendix D: Mixing Scheme

PreMixes

Eight premixes were designed to contain 6 - 11 SIRV transcripts in equimolar ratios. Their length distribution allows for unique identification in Bioanalyzer traces (Figure 21A). Bioanalyzer analysis also monitors the occurrence and the integrity of SIRVs in the 4 submixes and final E0, E1, and E2 mixes (Figures 21B and C). Although the Bioanalyzer traces do not allow for absolute quantitation, they were used to follow the relative compound distribution and consistency of the mixing procedure.

The accurate volumetric preparation of the 8 premixes was controlled by Nanodrop concentration measurements with a deviation of 0.002 % ± 3.4 % (maximal 7.6 %) from the calculated target concentrations. The mixing of the volumes was further monitored by weighing on an analytical balance, which showed a deviation of 1.8 % ± 0.65 % (maximal 2.5 %).

SubMixes

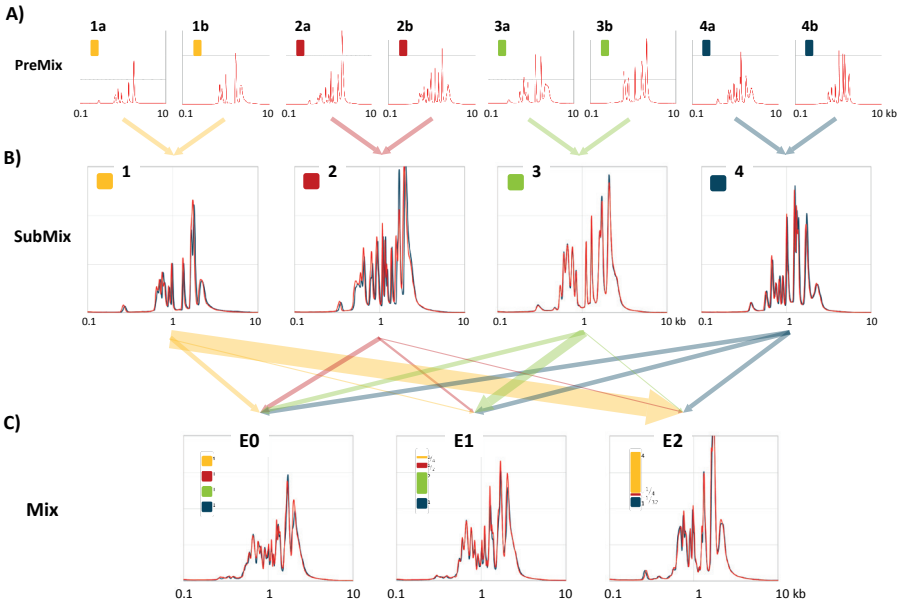

Figure 21. SIRV mixing scheme to obtain mixes E0, E1, and E2. A) The 8 premixes contain between 6 and 11 SIRVs which are different in length so that the SIRVs can be unambiguously identified in the Bioanalyzer traces. B) Two premixes were combined in equal ratios to yield four submixes in total. C) These, in turn, were combined in defined ratios (see Fig. 6) to obtain the final mixes E0, E1, and E2. Measured traces are shown in **red**, and traces computed from the premix traces to validate submixes and final mixes are shown in **blue**.

10. Appendix E: References

1. Paul, L. *et al.* SIRVs: Spike-In RNA Variants as External Isoform Controls in RNA-Sequencing. ***bioRxiv*** DOI: 10.1101/080747.
2. External RNA Controls Consortium. Proposed methods for testing and selecting the ERCC external RNA controls. ***BMC Genomics*** 6, 150 (2005).
3. Baker, S. C. *et al.* The External RNA Controls Consortium: a progress report. ***Nature Methods*** 2, 731–734 (2005).
4. 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).

11. Revision History

Publication No. / Revision Date	Change	Page
025UG063V0200 Jul. 6, 2021	Legal terms and conditions statements updated.	2
	Revised spike-in calculation equations and preparation steps (chapter 4).	13 - 16
	Revised SIRV download information (Appendix B).	29
	Addition of Table 2: SIRV set selection guide.	7
	Addition of Figures, 1, 5, 9 - 12, and update of Figure 2 (previously Figure 3).	4 - 22
	Product name updated to SIRV-Set 1 (Iso Mix E0, E1, E2).	
	Contact email updated to support@lexogen.com throughout.	
	Chapters and Appendices were reordered for Chemical Safety, MSDS, Applications, Analysis, Downloads, and References. References to SIRV Suite, materials and equipment, and support were revised or removed.	
	Change of SIRV gene to SIRV (genetic) locus throughout.	
025UG063V0111 Mar. 28, 2017	Kit content updated.	5
	Spike-in Data Evaluation using the SIRV Suite added.	22
025UG063V0110 Sep. 04, 2015	Product Release.	
025UI063V0100 Jun. 03, 2015	Initial Release. First release of the documentation together with the FASTA and GTF sequence file package, *name*_150601a.*extension*.	

12. Notes

LEXOGEN
The RNA Experts

Associated Products:

015 (QuantSeq 3′ mRNA-Seq Library Prep Kit for Illumina (FWD))
016 (QuantSeq 3′ mRNA-Seq Library Prep Kit for Illumina (REV) with Custom Sequencing Primer)
050, 051, 141 (SIRVs Spike-in RNA Variant Control Mixes)
095 (CORALL Total RNA-Seq Library Prep Kit)
113 - 115, 129 - 131 (QuantSeq 3′ mRNA-Seq Library Prep Kit FWD with UDI 12 nt Set A1, A2, A3, A4, A1-A4, or B1)
117 - 119, 132 - 134 (CORALL Total RNA-Seq Library Prep Kit with UDI 12 nt A1, A2, A3, A4, A1-A4, or B1)
125 - 127, 144, 145 (RiboCop rRNA Depletion Kits)
146, 147 (CORALL Total RNA-Seq Library Prep Kits with RiboCop)
157 (Poly(A) RNA Selection Kit V1.5)
158 - 163 (CORALL mRNA-Seq Library Prep Kit with UDI 12 nt Set A1, A2, A3, A4, A1-A4, or B1)

SIRV-Set 1 Spike-In RNA Variant Controls with Isoforms (Iso Mix E0, E1, E2) • 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: support@lexogen.com

Lexogen, Inc.
51 Autumn Pond Park
Greenland, NH 03840, USA
Telephone: +1-603-431-4300
Fax: +1-603-431-4333
www.lexogen.com