LEXOGEN
The RNA Experts

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)

SIRV-Set 4

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

User Guide

Catalog Number:
050 (SIRV-Set 2 (Iso Mix E0))
051 (SIRV-Set 3 (Iso Mix E0/ERCC))
141 (SIRV-Set 4 (Iso Mix E0 / ERCC / long SIRVs))

050UG134V0210

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.

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 products 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 referring to this spike-in mix in a publication, please use "SIRV-Set 2", "SIRV-Set 3", "SIRV-Set 4", "Spike-In RNA Variant Controls with Isoforms", "Spike-In RNA Variant Controls with Isoforms and ERCCs", or "Spike-In RNA Variant Controls with Isoforms, ERCCs, and long SIRVs". Individual transcripts can be referred to as "SIRV101", "ERCC-0025", and "SIRV4001", with the entirety being "SIRV isoforms", "ERCCs", and "long SIRVs". Stating the Catalog Number (Cat. No. 050, 051, or 141) 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 comprise RNA purification, library generation, sequencing itself, and the evaluation of 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 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 (Figure 1).

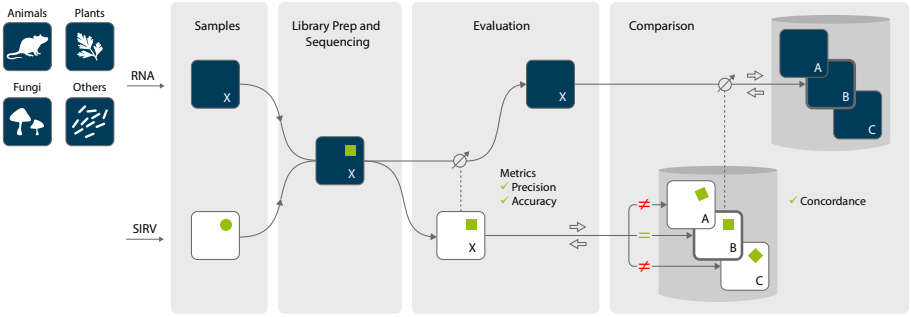

Figure 1. Workflow for using spike-in controls in RNA-Seq. Spike-in controls (SIRVs) are defined synthetic RNA molecules that mimic the main aspects of transcriptome complexity. They are added in minuscule amounts to samples before library preparation to undergo the very same processing steps as the endogenous RNA (green circle and square 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 = , red "does not equal" signs, respectively).

The Spike-In RNA Variants (SIRV) 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 (Figure 2). While the SIRV isoform module is available as a stand-alone module (Cat. No. 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 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 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 (Figure 3). For the sake of simplicity, all these transcriptional variants are referred to as isoforms. Each SIRV gene locus contains between 6 and 18 transcript 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 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 (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.

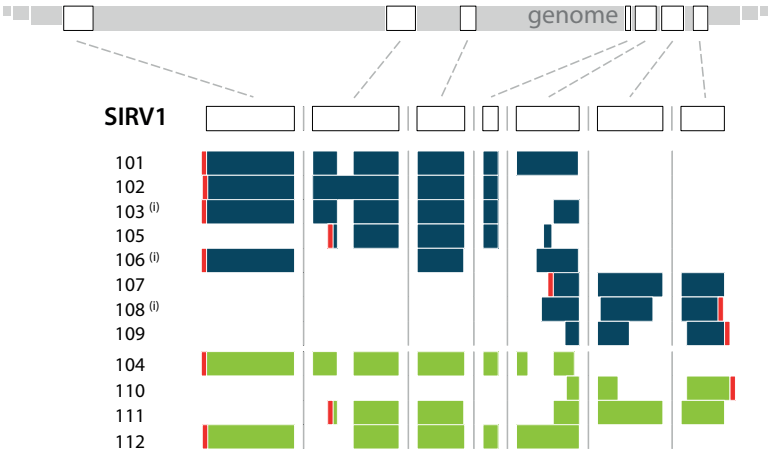

Figure 4. SIRV isoform design exemplified by SIRV1. The SIRV1 locus was derived from the human KLK5 gene, with transcripts added to the Ensembl annotations to generate comprehensive transcriptome complexity. All original and derived gene structures are shown in the Appendix. Transcripts in blue are part of SIRV mixes, 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 gene expression values.

For SIRV-Set 2, SIRV-Set 3, and SIRV-Set 4 the SIRV isoforms are provided in equimolar amounts; hence, their detection is not affected by concentration.

1.3 ERCCs: 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).

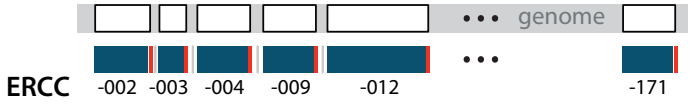

Figure 5. ERCC single-isoform design. ERCC transcripts follow the 1 gene, 1 exon, 1 transcript layout, providing each ERCC transcript with a unique sequence identity. Genes (exons) are shown in white, derived transcripts in blue, and the poly(A) tail is marked in red to indicate transcript 5'-3' orientation. Note that while there are 92 ERCC transcripts in the mix, the RNAs are numbered non-consecutively up to 171.

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 Long SIRVs: Length Complexity

The introduction of third generation sequencing platforms like Pacific Biosciences™ and Oxford Nanopore Technologies™ has significantly increased the available read length, now easily exceeding the average transcript length. The ERCC and SIRV isoform modules are optimzed for assessing RNA abundance and isoform complexity aspects. However, the average ERCC length is 909 nt (max. 2036 nt), the average SIRV isoform length is 1134 nt (max. 2528 nt), and thus spike-in RNA transcripts of both modules are below the average reported length for eukaryotic protein-coding mRNAs (e.g., 3.5 kb for the human transcriptome [7]).
Lexogen has therefore developed "long SIRVs", a module that contains three different transcripts for each of the five length categories 4 kb, 6 kb, 8 kb, 10 kb, and 12 kb (Figure 6). These RNAs cover the length of the majority of cellular transcripts. The sequence of each of these 15 RNAs is unique and does not overlap with any other spike-in or endogenous transcripts (similar to the ERCC module). Therefore, the equimolar long SIRVs are optimal tools to evaluate the transcript length aspect in RNA-Seq workflows. While designed in particular to assess long-read platforms, long SIRVs reveal length dependencies also in short read workflows.

Figure 6. Long SIRVs SIRV4001-SIRV12003. Long SIRVs follow the 1 gene, 1 exon, 1 transcript layout with three genes / transcripts corresponding to each of the five different length categories: 4 kb, 6 kb, 8 kb, 10 kb, and 12kb. 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.

1.5 SIRV Sets

The modular structure of available spike-in 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 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, 3, and 4 are covered in this User Guide. *Refers to number of vials, 1 or 3. The ERCC Module includes ERCC Mix 1[8].

		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 SIRV)

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 three mixes of the isoform module (SIRV Mixes E0, E1, and E2) designed for the validation of concentration measurement (including fold change) with isoform resolution.

SIRV-Set 2, **SIRV-Set 3**, and **SIRV-Set 4** contain one, two, or three modules in one mix. Each transcript is present in one defined concentration: SIRV isoforms and long SIRVs at equimolar ratios, ERCCs 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 (Figure 1). All quality metrics, except fold change measurements, can be determined experimentally for each individual sample.

SIRV-Set 2 (Cat. No. 050) 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 2 (as

well as SIRV-Set 3 and SIRV-Set 4) can be used to determine the performance also of isoform-agnostic workflows, if expression is determined on the gene (and not on the transcript) level.

SIRV-Set 3 (Cat. No. 051) contains the same mix of 69 isoforms (as in SIRV-Set 2) plus the ERCC module consisting of 92 non-overlapping sequences. This set covers both, a high level of isoform complexity and a large concentration range, and enables an even more comprehensive quality monitoring of individual samples. These two dimensions, abundance of the transcripts (x-axis), and concurrence of isoforms per gene (y-axis) are shown in Figure 7.

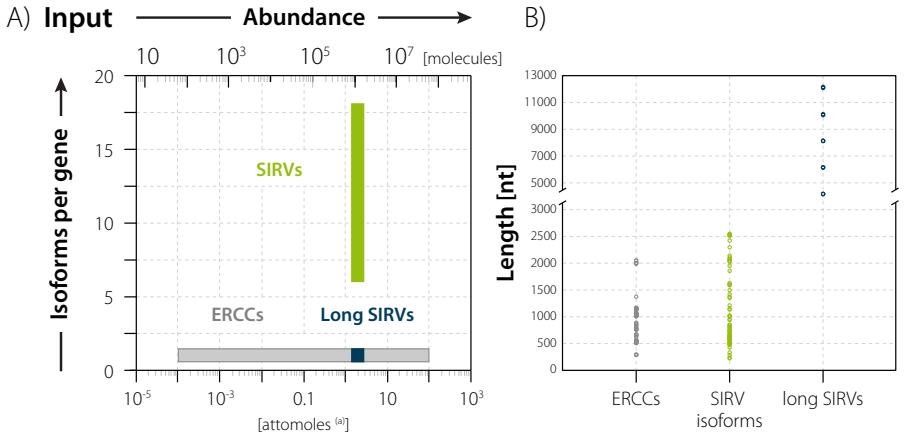

Figure 7. A) Isoform and abundance complexity and B) Length complexity. The SIRV isoform and ERCC transcripts in SIRV-Set 3 control for the two main dimensions of transcriptome complexity: isoforms and abundance. SIRV-Set 4 additionally contains controls for lenght complexity. The isoform module with 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 approx. 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, 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 12kb.

SIRV-Set 4 (Cat. No. 141) contains the long SIRV module with 15 RNAs of 4 - 12 kb length in addition to the 69 isoforms and 92 ERCC transcripts of SIRV-Set 3. This set thereby covers three spike-in aspects: isoform complexity, abundance, and length. As in Set 3, the modules are clearly segregated in its focus with the long SIRVs having non-overlapping sequences and equal concentrations.

2. Kit Components and Storage Conditions

SIRV-Set 2, 3, and 4 are provided in frozen format. Each tube contains 10 µl liquid volume (Figure 8). 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. Inert additives included stabilize the solution(s) sufficiently to undergo one freeze-thaw cycle. We recommend to aliquot the solution upon first time usage.

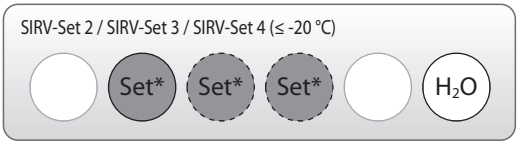

Figure 8. Location of kit components. *Each box of SIRV controls contains one or three (dotted lines) vials of the corresponding SIRV-Set: SIRV-Set 2 ● (Cat. No. 050), SIRV-Set 3 ● (Cat. No. 051), or SIRV-Set 4 ● (Cat. No. 141).

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. The final concentration is 2.52 ng/µl.

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. The final concentration is 3.03 ng/µl.

SIRV-Set 4 combines the equimolar mix of the 69 SIRV isoforms (Iso Mix E0) at 0.6 fmol each, the 92 single-isoform transcripts (ERCCs) ranging from 0.014 amol to 15.0 fmol, and the equimolar mix of 15 long SIRVs at 0.6 fmol each. The total amount of RNA is 102.2 fmol or 53.5 ng per tube. The final concentration is 5.35 ng/µl.

Table 2. Compositions of SIRV-Set 2, SIRV-Set 3, and SIRV-Set 4.

		# transcripts		ng		fmol		fmol / transcript	
Set 2	**Iso Mix E0**	**69**	**100 %**	**25.2**	**100 %**	**69.0**	**100 %**	**1.0**	**1.4 %**
Set 3	Iso Mix E0	69	43 %	15.1	50 %	41.4	44 %	0.6	0.6 %
	ERCC	92	57 %	15.2	50 %	51.8	56 %	7×10^{-6} to 15	8×10^{-6} to 16 %
	Iso Mix E0 / ERCC	**161**	**100 %**	**30.3**	**100 %**	**93.2**	**100 %**		
Set 4	Iso Mix E0	69	39 %	15.1	28 %	41.4	41 %	0.6	0.6 %
	ERCC	92	52 %	15.2	28 %	51.8	51 %	7×10^{-6} to 15	7×10^{-6} to 15 %
	long SIRVs	15	9 %	23.2	43 %	9.0	9 %	0.6	0.6 %
	Iso Mix E0 / ERCC / long SIRVs	**176**	**100 %**	**53.5**	**100 %**	**102.2**	**100 %**		

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 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

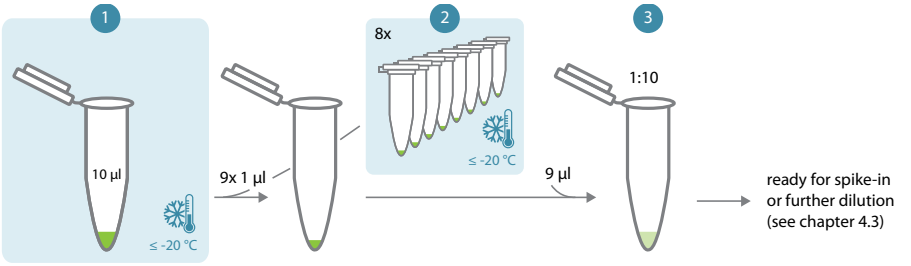

Figure 9. Workflow pictogram illustrating 1 SIRV stock, 2 SIRV aliquoting and storage, and 3 an example dilution for spike-in use. For the handling of SIRVs, strictly follow the guidelines (see 3.1 and 4.1 below). The final dilution depends on the experimental setup and must be calculated beforehand using Eq. 2. This can be achieved in one or several dilution steps.

4.1 Preparation

Unless the entire amount of SIRVs is immediately required, we recommend to aliquot the solution upon first time usage. Further dilutions should then originate from these aliquots. In the depicted workflow, one aliquot will be processed immediately, and the other ones are stored at ≤-20 °C (Figure 9).

ATTENTION: For preparing SIRVs to the required stock concentration apply the following guidelines:

- Any dilutions must be prepared immediately before spiking the SIRVs into RNA samples. Storage and freeze-thawing of diluted SIRVs 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 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, optimal dilutions will typically be 1:100 or higher (see chapter 4.3 "Spiking of RNA Samples").

4.2 Aliquoting and Interim Storage

Each SIRV tube contains exactly 10 µl at a concentration of 2.52 ng/µl (SIRV-Set 2), 3.03 ng/µl (SIRV-Set 3), or 5.35 ng/µl (SIRV-Set 4), respectively. The content of each SIRV tube can be divided into up to 9 aliquots and used for independent experiments.

1. 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 absorbence 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 8 aliquots immediately at ≤-20 °C for later use (Fig. 8).
3. Proceed with the 9th 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 steps:

ATTENTION: SIRVs should ideally be added to total RNA prior to any pre-processing steps (e.g., DNase I treatment, ribosomal RNA depletion, or poly(A) selection). SIRVs can also be mixed with lysis buffers and then added to cell or tissue samples prior to RNA extraction.

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

Example Eq. 1

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

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

F SIRV reads

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) efficiently covers the dynamic range of complex transcriptomes. The $F_{SIRV\ reads}$ value 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 (see Fig. 6), 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 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 ex-

cept 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).

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 1:100 dilution, the concentration reaches 25.2 pg/µl (SIRV-Set 2). Accordingly, by preparing a 1:200 dilution with SIRV-Set 3, the concentration is 15.15 pg/µl, and a 1:300 dilution of SIRV-Set 4 yields a concentration of 17.83 pg/µl. The volume needed to spike-in 30 pg SIRVs from SIRV-Set 2 is 1.19 µl (see below), 1.98 µl from SIRV-Set 3, or 1.68 µl of SIRV-Set 4. Pipetting low volumes is often error-prone. Therefore, higher dilutions and the spike-in of proportionally larger volumes are recommended.

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

Mass versus Molarity: Shot Gun Sequencing and Tag Profiling

Total RNA-Seq and mRNA-Seq: Standard short-read RNA-Seq generates numerous overlapping reads as a function of transcript length (mass). Since the average length of SIRV isoforms and ERCCs is shorter but comparable to the average length of cellular transcripts, Eq. 1 and Eq. 2. work well for SIRV-Set 2 and SIRV-Set 3. In SIRV-Set 3 the reads mapping to the spike-ins will be distributed evenly between SIRV Iso Mix E0 and the ERCC module (Tab. 2). In SIRV-Set 4, the 15 long SIRV transcripts of 4 kb to 12 kb length contribute proportionally more mass and are hence expected to generate 43 % of all spike-in reads, compared to 28% for each of the other two modules. A target share of 1 % SIRV-mapping reads ($F_{SIRV\ reads}$ in Eq. 1) will provide enough coverage for all modules except for very low read depths or when spiked into samples with an unusual high content of polyadenylated RNAs.
In a theoretical mRNA-Seq experiment generating 20 million PE100 reads of which 1% are mapping to SIRV-Set 4, 56 k reads will cover the 69 SIRV isoforms (having a total length of 77 kb), leading to a theoretical average nucleotide coverage of >140-fold. The 15 long SIRVs (total length of 120 kb) receive 86 k reads to obtain the same high coverage of >140-fold. The 92 ERCCs with a total length of 86 kb receive the same read share as the SIRV isoforms (56 k reads), but coverage of the individual ERCC transcripts depends directly on its concentration (see also Fig. 11).

Third generation sequencing & 3' mRNA-Seq: In tag profiling methods such as oligo(dT)-primed long-read sequencing and QuantSeq™ 3' mRNA-Seq, each transcript is represented ideally by a single read. In SIRV-Set 2, all SIRV isoforms are present in equimolar amounts to generate the same number of reads each. In SIRV-Set 3, SIRV Iso Mix E0 and the ERCC module are represented by almost equal molarities (Tab. 2), hence spike-in derived reads will be distributed about evenly between the two modules. In SIRV-Mix 4, the 15 long SIRVs have the same per transcript molarity as the 69 isoforms of SIRV Iso Mix E0. Thereby, the long SIRVs and the SIRV

isoforms will yield sufficient reads on platforms like Oxford Nanopore Technologies™ and Pacific Biosciences™ to test for isoform identification and deviation from the expected equimolar concentration as well as to assess the ability to sequence RNAs full-length up to 12 kb. The 92 ERCCs will generate together 51 % of all spike-in reads, but due to their differences in concentration the individual molarities range over 6 orders of magnitude.
In a setup with 5 million long reads and 2 % SIRV-Set 4 spike-in (based on mass, i.e., 2 % m_{SIRVs} per m_{mRNA}) approximately 100 k reads are assigned to the three modules, with each of the 69 SIRV isoform and 15 long SIRV transcripts destined to be covered by 600 full-length reads. The remaining 92 ERCC transcripts receive 51 k reads in total.

Up- and Down-Scaling

As recommended above, the provided SIRV RNA should be divided into 9 aliquots of 1 µl. This provides enough material to prepare 9 batches of libraries. For example, considering SIRV-Set 3, and an input amount of 100 ng for 1 % SIRV read coverage, 30 pg of SIRV RNA should be spiked into each sample. This requires 1.98 µl of a 1:200 dilution to be added per sample. Assuming 200 µl of the SIRV dilution is prepared, 100 samples can be prepared per 1 µl of provided SIRV RNA. This equates to a total of 900 samples (9 x 100) that can be prepared per tube of SIRV RNA. 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 - 30 pg of total RNA, which would require only 6 - 13 fg of SIRVs (depending on the SIRV-Set). Therefore, each of the 9 aliquots is theoretically sufficient to spike 420,000 cells, which corresponds to several large-scale, single-cell experiments. The number of samples and experiments that can be prepared using the provided SIRV-Set 2, SIRV-Set 3, and SIRV-Set 4 tube volumes can also be derived from the SIRV Calculator Worksheets available from www.lexogen.com/sirvs/download.

4.5 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 and the long SIRVs contain a 30 nt long poly(A) tail each, whereas the single-isoform ERCCs have slightly shorter and variable poly(A) tails of 24 ± 1.05 nt. All of these poly(A) tails allow for poly(A) enrichment and oligo(dT)-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. 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.

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.

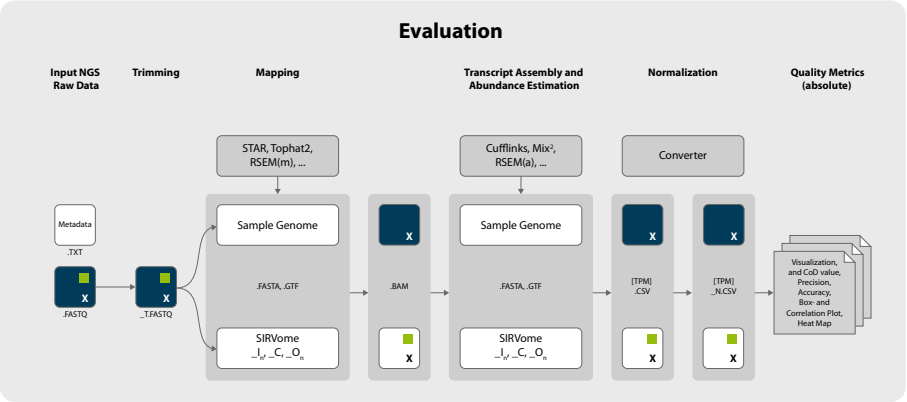

Figure 10. SIRV data evaluation scheme. SIRV reads undergo the very same processing steps as reads derived from the RNA sample. CoD, Coefficient of Deviation.

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 7 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).

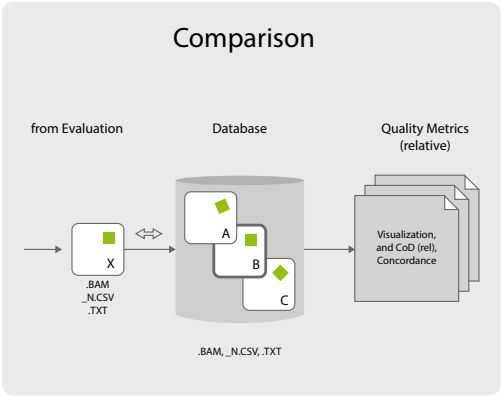

Figure 11. SIRV data comparison scheme. The control data set is used to carry out pairwise comparisons between experiments using the small subset of SIRV control data. Data sets of high concordance can be selected and the extent of expected error rates can be estimated before making a decision about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

5.2 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 12.

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.

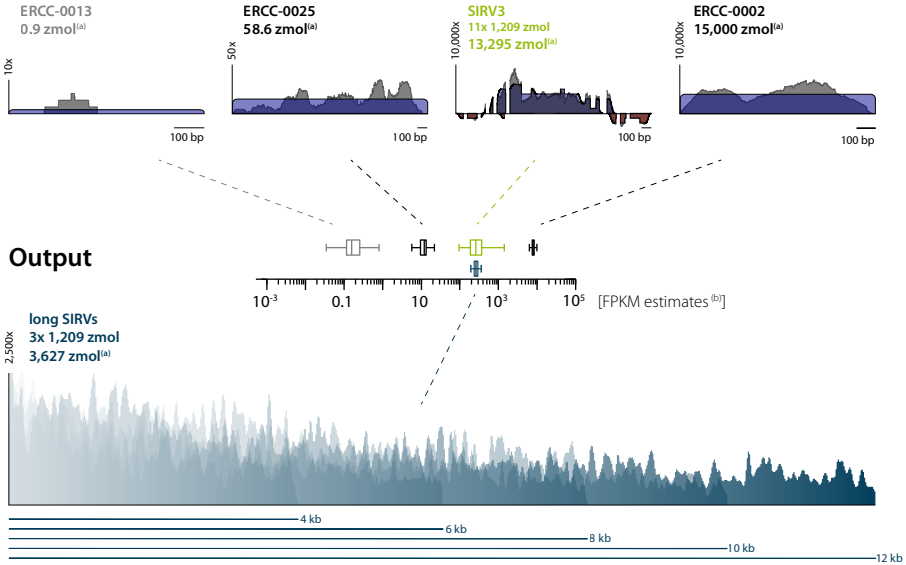

Figure 12. Read coverages of SIRV isoform and single-isoform (ERCC) genes depend on input concentration, library preparation efficiency, biases, and read depth. Quantifying the 92 single-isoform transcripts (ERCCs) depends on the averaged overall coverage but is rather independent of positional coverage 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 gray 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 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.

The ability of a given RNA-Seq workflow to cover transcripts exceeding 2,500 nt is tested by the long SIRV module. For non-poly(A) selective short-read methods (total RNA-Seq), 5' and 3' end representation and continuous, uniform coverage across the whole transcript length are main quality criteria (Figure 12). On long-read platforms, the equal molarities of long SIRVs and SIRV isoforms in SIRV-Set 4 provide for a comprehensive mix to control for isoform detection and quantification, and for the representation of very long transcripts. Methods that prime at the poly(A) tail and produce full-length cDNAs such as Lexogen's TeloPrime™ kit (Cat. No. 013) can be assessed for their full-length reverse transcription efficiency. Since degraded RNA would result in truncated cDNAs, integrity of the SIRV transcripts has to be maintained throughout the

experiment, which is also true for poly(A) selective short-read methods (mRNA-Seq), where RNA degradation is reflected in the ratio of of 5' / 3' end coverage.

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 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.

5.4 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 5.2).

5.6 Use of the Different SIRV Annotations

SIRV reads should be mapped initially using the correct **SIRV_C** annotation (see downloads, Chapter 7). 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. 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.

*) In the isoform and long SIRV 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 the 92 genes of the single-isoform (ERCC) module yields one measure, the sample-specific mean CoD value, which quantifies the coverage uniformity.

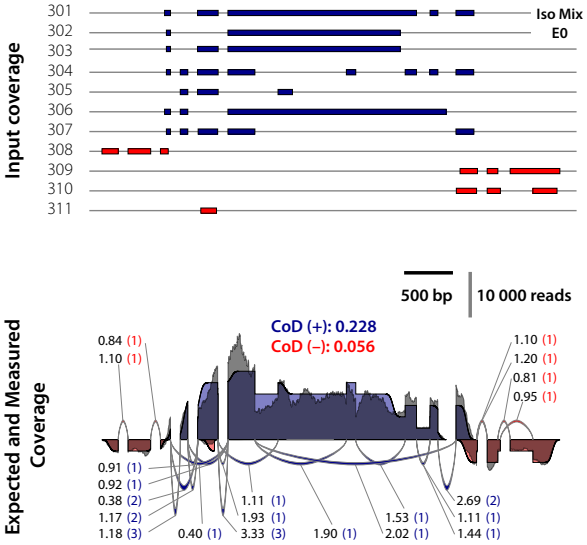

Figure 13. Comparison of the expected and the measured coverages for the SIRV3 locus of the isoform module. Top, individual transcripts of SIRV3 with the exons on the plus strand in **blue** and exons on the minus strand in **red**. Bottom, the expected SIRV3 coverage is shown as transparent **blue** 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.

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 13). 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.
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 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 our FAQ section).

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, 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 data sets are comparable across samples or experiments.

5.9 Recommended Software Packages

SIRV Suite

Lexogen's SIRVsuite - A publicly-available command line tool - can be used to QC an RNA-Seq workflow using Lexogen's SIRVs. The SIRVsuite features 3 main modules:

- **Coverage Module** – compares observed spike-in gene coverage to the expected coverage, which is estimated from transcripts.
- **ERCC Correlation Module** – validates expected vs measured ratios for single-transcript spike-ins in the whole concentration range.
- **SIRV Concentration Module** – detects anomalies of SIRV transcript concentration among samples or the overall spiked-in transcript distribution

To get started with the SIRVsuite, please visit Lexogen's GitHub page: https://github.com/Lexogen-Tools/SIRVsuite.

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 [8].

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.

Please contact support@lexogen.com for further information.

6. Appendix A: 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/download) are required for detailed viewing.

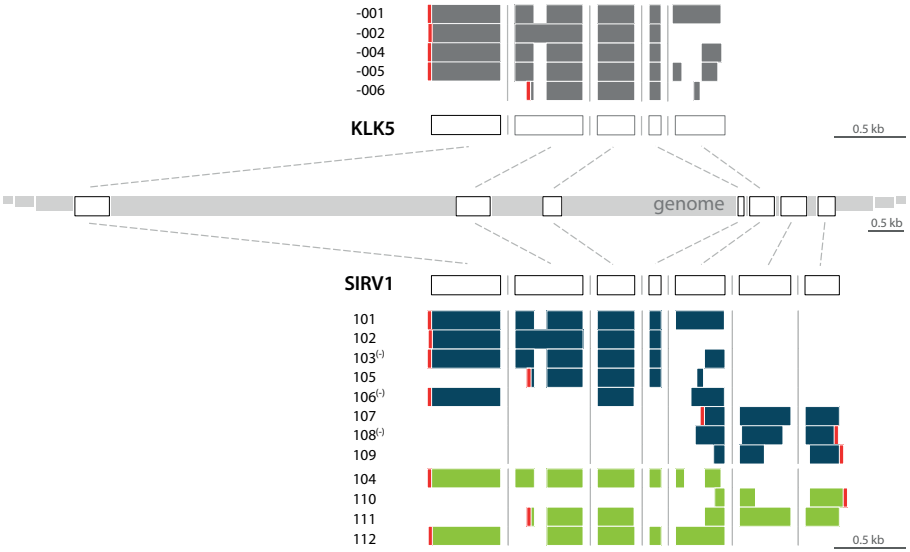

Figure 14. SIRV1 | based on human gene *KLK5*. The human Kallikrein-related peptidase 5 gene was taken as template for SIRV1 locus generation. Its expression is up-regulated by estrogens and progestins, and alternative splicing results in multiple transcript variants encoding the same core protein. The 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**. 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**.

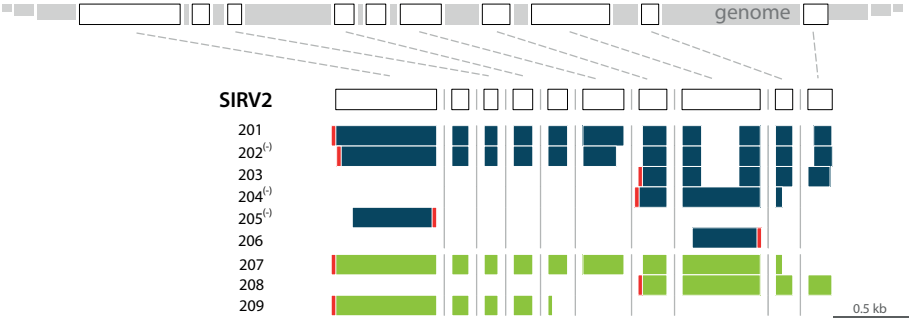

Figure 15. SIRV2 | based on human gene *LDHD* contains 6 transcript variants (shown in **blue**). The transcript variants shown in **green** are additional annotations for alternative evaluation procedures, whereas the SIRVs marked with (-) are missing in the insufficient annotation.

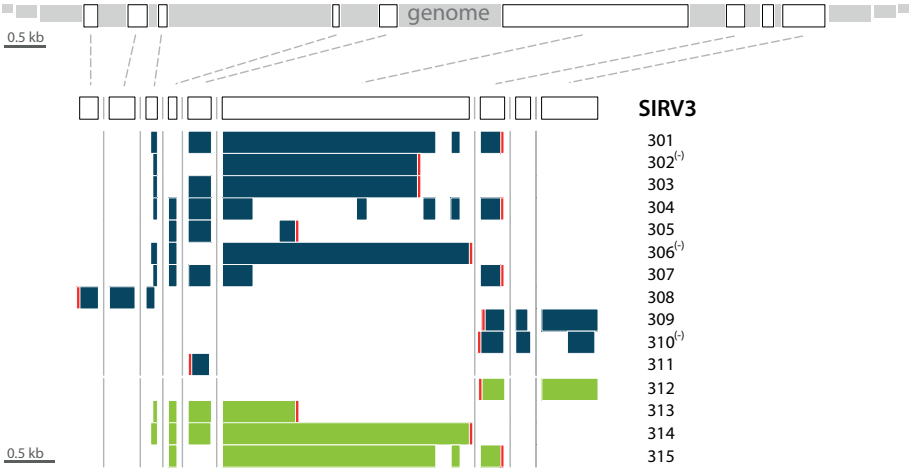

Figure 16. SIRV3 | based on human gene *LGALS17A* contains 11 transcript variants (shown in **blue**). The transcript variants shown in **green** are additional annotations for alternative evaluation procedures, whereas the SIRVs marked with (-) are missing in the insufficient annotation.

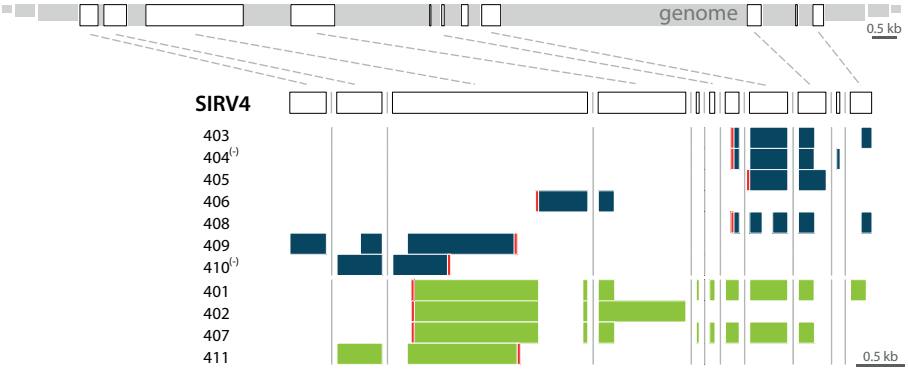

Figure 17. SIRV4 | based on human gene *DAPK3* contains 7 transcript variants (shown in **blue**). The transcript variants shown in **green** are additional annotations for alternative evaluation procedures, whereas the SIRVs marked with (-) are missing in the insufficient annotation.

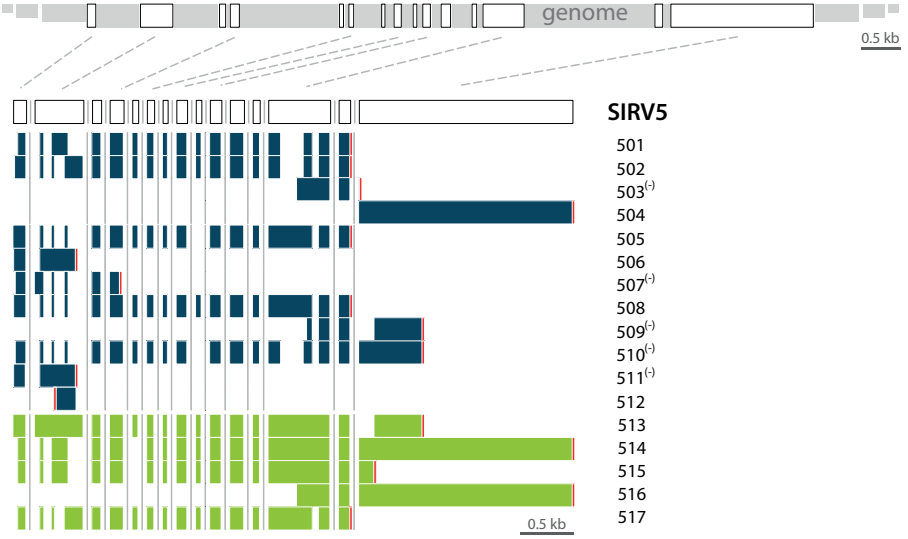

Figure 18. SIRV5 | based on human gene *HAUS5* contains 12 transcript variants (shown in **blue**). The transcript variants shown in **green** are additional annotations for alternative evaluation procedures, whereas the SIRVs marked with (-) are missing in the insufficient annotation.

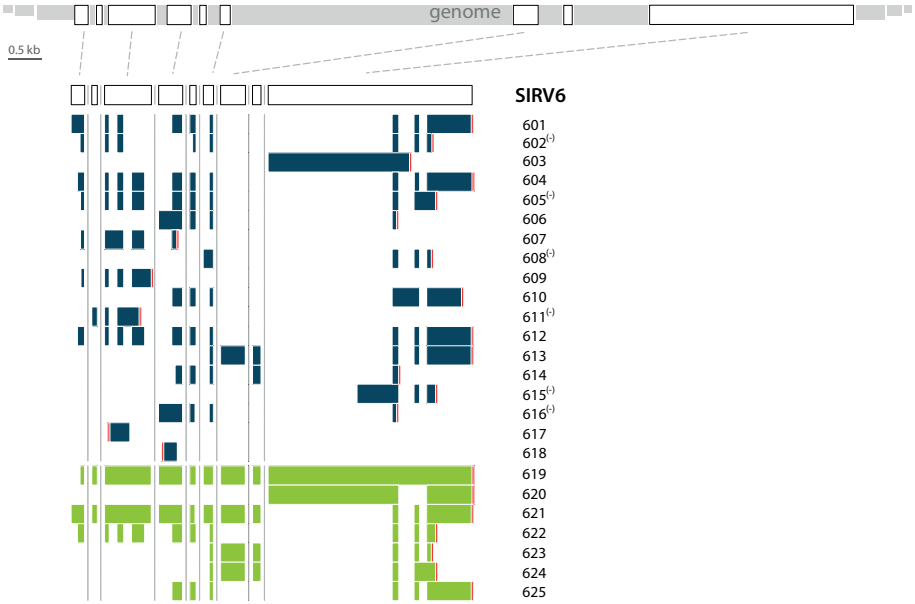

Figure 19. SIRV6 | based on human gene *USF2* contains 18 transcript variants (shown in **blue**). The transcript variants shown in **green** are additional annotations for alternative evaluation procedures, whereas the SIRVs marked with (-) are missing in the insufficient annotation.

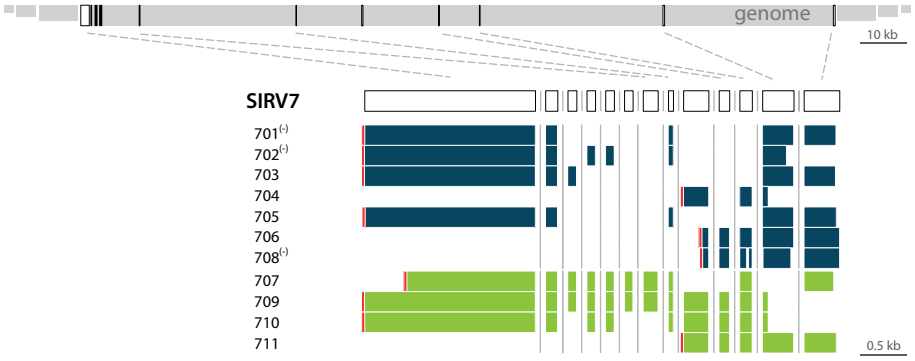

Figure 20. SIRV7 | based on human gene *TESK2* contains 7 transcript variants (shown in **blue**). The transcript variants shown in **green** are additional annotations for alternative evaluation procedures, whereas the SIRVs marked with (-) are missing in the insufficient annotation.

7. Appendix B: Downloads

Sequences, annotations, and concentration tables for each SIRV-Set can be obtained from: www.lexogen.com/sirvs/download.

SIRV-Set Files

For each SIRV-Set a "_sequence-design-overview_" (XLSX) file and a "_Sequences_" ZIP archive 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 (and ERCC for Sets 3 and 4) 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 each SIRV-Set 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: 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. Piovesan, A., et al. Human protein-coding genes and gene feature statistics in 2019. ***BMC Res Notes*** 12, 315 (2019).
8. 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).

9. Revision History

Publication No. / Revision Date	Change	Page
050UG134V0210 Nov. 16, 2021	Updated SIRV spike-in Attention note in section 4.3.	13
	Dilution factors corrected in Eq. 2 example.	15
	Updated SIRV data analysis and downloads information.	25, 30
050UG134V0200 Jul. 23, 2020	Legal terms and conditions statements updated.	2
	Figure 2 replaced.	5
	Storage conditions updated.	11
	Section Detailed Protocol was included and the content of subsections was updated.	11-12
	Recommendations specific to SIRVs provided in dry format removed.	9, 11-12, 14
	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.	
	Details introduced regarding SIRV-Set 4 and long SIRV Module introduced.	
	Update of literature citation section.	2
	Change of SIRV gene to SIRV locus.	5, 6
	New section 1.4 (incl. Figure 6) about Long SIRVs.	7
	Update of SIRV set selection guide to include SIRV-Set 4.	8
	Update of Figure 7 to include length distribution of SIRV isoforms, ERCCs, and long SIRVs.	9
	Figure 8 and referencing text updated to show uniform kit content for SIRV-Set 2, 3, and 4.	10
	Table 2 updated to show detailed composition of SIRV-Set 2, 3, and 4.	10
	New section on mass and molarity.	15
	Figure 12 updated to include long SIRV quantification deviation and total RNA-Seq coverage of the 5 length categories.	19
	Reference to bioinformatic resources for SIRV-Set 4 inserted.	30
	Reference to average length of human mRNAs (Piovesan et al., 2019) inserted.	31
050UG134V0100 Jul. 14, 2017	Initial Release.	

LEXOGEN

Enabling complete transcriptome sequencing

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

SIRVs · Spike-In RNA Variant Controls · User Guide

SIRV-Set 2 (Iso Mix E0)

SIRV-Set 3 (Iso Mix E0 / ERCC)

SIRV-Set 4 (Iso Mix E0 / ERCC / long SIRVs)

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