LEXOGEN

Enabling complete transcriptome sequencing

SiRVs™

Spike-in RNA Variant Control Mixes

Spike-in RNA Variant Control Mixes E0, E1, and E2

User Guide

Catalog Number:
025.03 (Spike-in RNA Variant Control Mixes)

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, RNasin is a trademark of Promega Corporation, Microsoft Excel is a registered trademark of Microsoft Corporation in the United States and/or other countries.

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 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 and warrants that the product performs to the standards described in this user guide up to the expiration date.

LITERATURE CITATION

When describing a procedure for publication using this product, please refer to it as the SIRVs, Spike-In RNA Variants, SIRV Mixes, or Spike-In RNA Variant Control Mixes.

We reserve the right to change, alter, or modify any product without notice to enhance its performance.

CONTACT INFORMATION

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

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

Table of Contents

LEXOGEN

Enabling complete transcriptome sequencing

1. Introduction

RNA sequencing (RNA-Seq) workflows comprise RNA purification, library generation, the sequencing itself, and the evaluation of the sequenced fragments. The initial steps impose biases for which the data processing algorithms try to compensate. Significant issues are the concordant assignment of fragments to the original transcript variants, the robustness towards annotation flaws, and the subsequent deduction of the correct abundances. As long as the quality of all individual processing steps cannot be unambiguously determined subsequent comparisons of experimental data, in particular between different data sets, remain fuzzy.

To date, comparisons within and between workflows have been carried out in a series of exemplary studies which have been initiated by the FDA Sequencing Quality Control (SEQC) Consortium and the Association of Biomolecular Resource Facilities (ABRF). Here, up to four reference RNA samples (SEQC samples A to D from the US Food and Drug Administration, FDA) were processed using four different treatments (poly(A) selected, ribo-depleted, size selected, and degraded RNA) and examined on five NGS platforms (Illumina HiSeq, Life Technologies Personal Genome Machine and Proton, Roche 454 GS FLX, and Pacific Biosciences RS) [Li, 2014]. The proliferation of different RNA-Seq platforms and protocols has created a need for well-defined reference material to compare the different performance characteristics. The reference RNA samples contain Universal Human Reference RNA (Agilent Technologies), Human Brain Reference RNA (Ambion, Life Technologies), both of relative robust but unknown transcript diversity, and a first set of spike-in controls. This first set of artificial RNA controls was developed by the External RNA Controls Consortium (ERCC) which led to 92 ERCC Spike-in Controls. These constructs are available from Ambion (Life Technologies) as spike-in control for non-comparative experiments and as set of 2 controls for assessing differential expression. Comparisons of the evaluated Spike-in Mix reads with known concentrations allow to assess dynamic range, dose response, lower limit of detection and efficiency, and fold-change response of RNA sequencing pipelines within the boundaries of the complexity of the monoexonic, non-overlapping RNA sequences [Munro, 2014]. The ERCC Spike-in Controls contain no transcript variants. Therefore, one of the main challenges of sequencing complex transcriptomes – to distinguish splice variants – could not yet be evaluated.

Now, the Spike-in RNA Variants, SIRVs, provide for the first time a comprehensive set of transcript variants to validate the performance of isoform-specific RNA-Seq workflows, and to serve as a control and anchor set for the comparison of RNA-Seq experiments. The SIRV sequences enable bioinformatic algorithms to accurately map, assemble, and quantify isoforms, and provide the means for the validation of these algorithms in conjunction with the preceding RNA-Seq pipelines. The SIRVs are tightly controlled for consistency and comparability, and provide reliable measures for detection of the sources of errors. SIRV Mixes can be used as single spike-ins, or as a combination of two or three different Mixes for the assessment of differential gene expression.

Li, S et al. (2014) *Nature Biotechnology* 32, 888–895;   Munro, SA et al. (2014) Nat. Comm. 5:5125, doi:10.1038/ncomms6125;

2. Kit Components and Storage

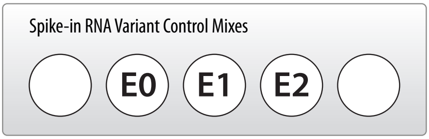


Figure 1 | Location of the kit components (Cat. No. 025.03)**.**

Each SIRVs box contains 3 tubes labeled E0, E1 or 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 1. The individual concentrations of each SIRV in all Mixes can be obtained from the SIRVs download section at **www.lexogen.com/sirvs/#sirvsdownload**.

The tubes must be stored at, or below, -20°C. Freeze/thaw cycles have to be minimized for the stock solutions and should be avoided for diluted aliquots. For further information read Chapter 6.1.

Table 1 | 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 the save taking of 3 × 1 µl.

The number of reactions depends on the spike-in amount required. You can draw 4 times 1 µl. This 1 µl should then be stepwise diluted to 1:1000, of which for a typical experiment using 100 ng total RNA input (e.g., spiking of Human Brain Reference RNA (HBRR)) 3.6 µl are required for an rRNA depletion experiment, respectively 2.4 µl for an mRNA-Seq experiment.
NOTE: We do not recommend to keep the dilution for very long as the diluted RNA solutions are increasingly unstable.

3. Materials and Equipment Required

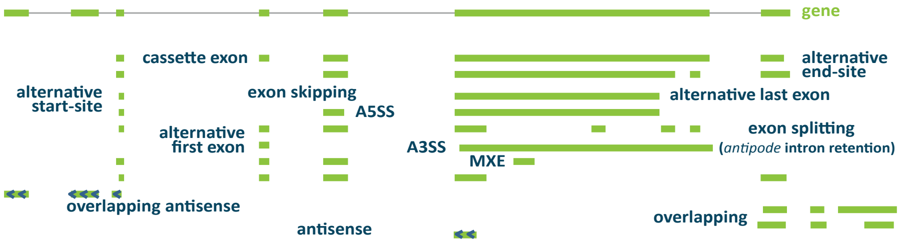
For the dilution of SIRVs either RNase-free molecular biology grade water or RNA-compatible buffers, e.g., sodium citrate, pH 6.4, or Tris-EDTA, pH 7.0, can be used. **Divalent cations should be**

avoided as buffer components! It is important that the solutions as well as all materials which-come into contact with the SIRVs are absolutely RNase-free. Further, all materials which come into contact with the SIRVs must have a low binding capacity for nucleic acids. This concerns vials, microtubes, well plates, and pipette tips. It is preferable to use barrier pipette tips. Working with SIRVs requires decontaminated pipettes and, preferentially, a sterile environment.

4. SIRVs

SIRVs consist of 69 artificial transcript variants which mimic 7 human model genes extended by exemplary isoforms to reflect comprehensively variations of alternative splicing, alternative transcription start- and end-sites, overlapping genes, and antisense transcription (Figure 2). The SIRV sequences conform to the canonical GT-AG exon-intron junction rule (96.9 % of all junctions) with few exceptions including the less frequently occurring natural variations GC-AG (1.7 %) and AT-AC (0.6 %), as well as two non-canonical splice sites with CT-AG and CT-AC (0.4 % each).

- ❑ **Take natural gene structure** and **annotated transcript variants**
- ❑ **Shorten transcript length** to a maximum of 2.5 kb
- ❑ Fill gene structure with **heterologous sequence**



- ❑ Duplicate and modify to add **alternative splicing variants**
- ❑ Add **transcription start-site and end-site** variants
- ❑ Observe intron donor-acceptor site rule
- ❑ Add **antisense** and **overlapping** variants

Figure 2 | SIRVs design overview. Seven artificial SIRV genes mimic human model genes to represent in their entirety all main aspects of alternative splicing in numerous repeats and variations. The sequences have no significant similarities to any known sequences but obey all common design features. Alternative splicing as stated or abbreviated with A5SS and A3SS, alternative 5'/3' splice sites; MXE, mutually exclusive exons.

Between 6 and 18 transcript variants were designed and produced for each of the model genes. The mRNAs range from 191 to 2528 nt with a GC content of 29.5 - 51.2 %, with the shortest mRNAs being antisense monoexonic transcripts.

Table 2 | Summary of alternative splice variations for each SIRV per gene. 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 gene. 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 (cf. Figure1).

	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

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 on the 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.

The SIRVs were produced by *in vitro* transcription from synthetic genes. Constructs were designed for each of the sequences that comprise 5' to 3' (a) a unique restriction site, immediately upstream of (b) a T7 RNA polymerase promoter, whose 3' G is the first nucleotide of (c) the SIRV sequence, seamlessly followed by (d) a A_{30} tail that is fused with (e) an exclusive 2nd restriction site. The complete gene cassettes were synthesized, cloned into a vector, and Sanger-sequenced (Figure 3).

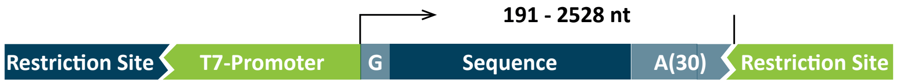


Figure 3 | SIRV production by T7 transcription. Transcription occurred from sequence-verified, linearized plasmid templates. All SIRVs start with a G and end with a poly(A)-tail containing 30 adenosines.

All SIRVs were produced by T7 transcription which partially generated RNA of varying integrity. A series of tailored methods was applied to purify full-length SIRV RNAs with a minimal amount of any side products, because high molecular integrity is a prerequisite for the definition of transcript variant controls. Details to be found in Chapter 7.2.

5. SIRV Mixes E0, E1, and E2

The set of SIRV mixtures consists of the 3 SIRV Mixes, E0, E1, and E2, with each Mix containing all 69 SIRVs but in different concentration ratios. The RNA purity and individual concentrations were measured by absorbance spectroscopy (Nanodrop by Thermo Scientific) and verified by capillary electrophoresis (Bioanalyzer by Agilent Technologies), which was also used to determine the molecular integrity. The concentration ratios in

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

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 1).

Each SIRV transcript enters the Mixes as part of one of 4 SubMixes. The 4 SubMixes are arranged in 3 different combinations according to defined ratios as shown in Figure 4. The concentration of the SIRVs within each SubMix is equimolar, and consequently, the relative concentration ratios of SIRVs which descend from the same SubMix remain identical throughout all final Mixes (see chapter 7.3, p.20).

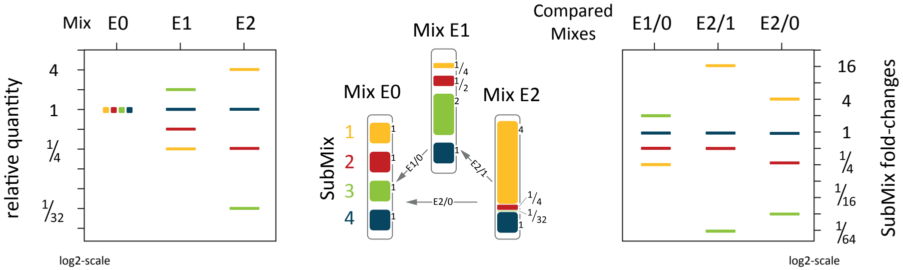


Figure 4 | Graphical representation of the SubMix (1-4) distribution in the 3 SIRV Mixes and the resulting intra- and inter-mix ratios. The 4 SubMixes are represented by different colours and contain between 12 and 21 SIRVs to keep the total molarity and weight of the mixes evenly balanced at 69.5 fmol/µl 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 also Chapter 7.3.

The assignment of the SIRVs to the 4 SubMixes was optimized as such that the final total mass as well as the final total molarity in all 3 Mixes are the same. Because of the various lengths each SubMix contains not the same but between 12 and 21 SIRVs. The distribution of transcript variants is as diverse as possible so each SIRV gene is present in each SubMix with at least one transcript variant (details in chapter 7.1) .

The *a priori* knowledge of SIRV abundance in the Mixes E0, E1, and E2 allows to assess the correctness of differential gene expression (DE) measurements based on transcript variant identification, quantification and variance of technical repeats . The SIRVs from SubMix 4, shown in blue in Figure 3, are always present at the same concentration and serve as false positive control in DE evaluation pipelines. The most challenging ratios are set by one 2-fold and vice versa by two 1/2-fold changes in concentration. Although the 1/64-fold change provides for the most distinct DE value, the SIRV SubMix 3 concentrations in Mix E2 are the lowest in the entire sample set and they are the hardest to determine correctly at low read depths.

6. Application

The SIRVs are intended to be used as spike-ins in RNA-Seq experiments to validate entire RNA-Seq workflows under authentic conditions. The measurement of SIRVs and the comparison to the known input help to determine the accuracy of the pipelines and to identify sources of error.

6.1 Spiking of RNA Samples

SIRV Transcript Stability

Freeze/thaw cycles have to be minimized for the 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. Lexogen delivers the SIRVs at sufficiently high concentrations of 25.3 ± 0.1 ng/µl to reduce these adverse effects. In all cases, it is obligatory to adhere to the high experimental standards required for handling RNA.

RNA Handling Guidelines

- RNases are ubiquitous, and special care should be taken throughout the procedure to avoid RNase contamination.
- Use commercial ribonuclease inhibitors (i.e., RNasin, Promega Corp.) to maintain RNA integrity when storing samples. SIRV Mixes contain RNasin.
- 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 a library preparation, clean your work space, pipettes, and other equipment with RNase removal spray (such as RNaseZap, Ambion Inc.) as per the manufacturer's instructions.
- 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.

Preparation of Samples

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. To provide a first orientation of the possible concentration ranges the values in Table 3 contain estimates of relevant RNA fractions found in total RNA for two RNA-Seq workflows, one starting with rRNA-depleted RNA, the other with poly(A)-selected RNA. Volumes and the amounts of molecules for the SIRV Mixes are given as reference, and in case of co-spiking, for the ERCC Mixes from Ambion (Life Technologies). The ratios have to be understood only as guidelines for designing the appropriate spike-in experiments.

	total RNA		100		ng
		average rRNA depleted total RNA (r-RNA)		3	% (equiv. est. ng)
		average mRNA content (polyA RNA)	2		% (equiv. est. ng)
🟩	SIRV Mix	volume	2.4	3.6	µl of 1 : 1 000 dilution
		amount	0.06	0.09	ng
		final content	2.83		%
		min amount	$0.63 \cdot 10^{-3}$ $37.6 \cdot 10^{3}$	$0.94 \cdot 10^{-3}$ $56.5 \cdot 10^{3}$	attomole molecules
		rel. NGS concentration	10.4		FPKM
		max amount	8 $4.8 \cdot 10^{6}$	12 $7.2 \cdot 10^{6}$	attomoles molecules
		rel. NGS concentration	1333		FPKM
□	ERCC Mix	volume	2	3	µl of 1 : 1 000 dilution
		amount	0.06	0.09	ng
		final content	2.83		%
		min amount	$0.29 \cdot 10^{-6}$ 17	$0.43 \cdot 10^{-6}$ 26	attomole molecules
		observed NGS concentration	0.01		FPKM
		max amount	60 $36.1 \cdot 10^{6}$	90 $54.2 \cdot 10^{6}$	attomoles molecules
		observed NGS concentration	10 000		FPKM
	entire NGS relevant RNA amount		2.12	3.18	ng

Table 3 | Mixing table for spiking total RNA, poly(A) selected RNA (left column), or rRNA-depleted total RNA (right column) with any of the SIRV Mixes. The minimum and maximum numbers correspond exemplarily to the lower and upper expectation figures when using SIRV Mix E2. The numbers for spiking an ERCC Mix are shown as comparison. The RNA content and subsequent spike-in ratios after poly(A) selection or rRNA-depletion of 100 ng total RNA is estimated based on an average percentage of the RNA classes which can vary in different samples. Dilutions can be made using RNase-free molecular biology water, and all RNA-compatible buffer solutions. FPKM, fragments per kilobase of exon per million fragments mapped.

The final results depend on several factors like the expression state of cells, the quality and integrity of the RNA, as well as the kind of NGS library preparation. The accuracy of the spike-in experiments depends on correct volumetric dilution series, thorough mixing and careful handling of dilutions. Diluted RNA solutions are increasingly unstable. Freeze-thaw cycles must be minimized. We recommend to freshly prepare adequate dilutions of the SIRVs, e.g., stepwise a final dilution of 1:1000, for accurate pipetting of the Spike-in Mixes.

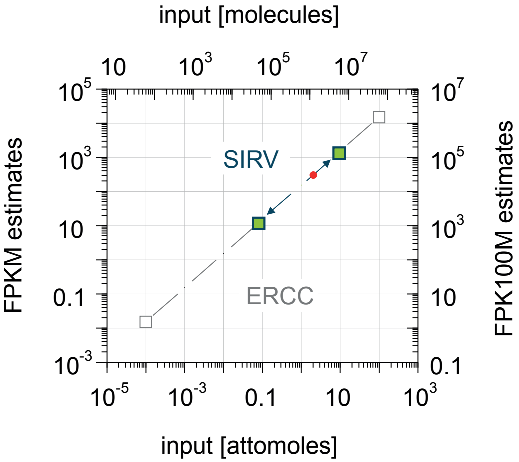


Figure 5 | Dynamic concentration range of SIRVs. The maximal dynamic range is given by the SIRV transcripts of Mix **E2** with minimal and maximal concentrations (green boxes). These are depicted in an input-output correlation based on the known input and a FPKM (Fragments Per Kilobase of exon per Million fragments mapped) estimation for SIRV Mixes in reference RNA background samples which have been prepared according to the suggested mass ratios in Table 3. As reference, the concentration range of the ERCC mix is also shown (grey boxes). In contrast, in SIRV Mix **E0** all concentrations are identical which is shown by the red dot.
The FPKM value is an estimate for 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.

The SIRV Mixes in their current format cover a smaller concentration range than the ERCC Mixes with 2 versus 6 orders of magnitude (Figure 5). Higher transcript coverage rates will increase the chance to correctly distinguish 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.

6.2 Library Preparation and Sequencing

Considerations for Library Preparations

The SIRV transcripts behave identical to mRNA in most aspects of any RNA-Seq library preparation. SIRVs have no sequence homology to rRNA and are therefore not targeted by any rRNA directed depletion method. SIRVs also comprise a 30 nt long poly(A)-tail, which allows poly(A)-enrichment and oligodT-priming.

SIRVs do not have a 5′-cap structure (5′-m^{7}G) but a 5' triphosphate end. Cap-specific cDNA preparation methods are not feasible.

Sequencing

RNA-Seq libraries have to be sequenced with sufficient read depth to overcome certain coverage thresholds outlined Table 3 and Figure 5 (see comment in figure legend).

Different lower read depth thresholds need to be considered when using full-length single molecule sequencing or tag sequencing methods. In these methods each transcript or amplicon is represented at most by one single read and not by numerous reads as function of transcript length. In these cases, the molar and not the mass ratio is of relevance. 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 an comparative evaluation of experiments.

6.3 Analysis of Sequencing Data

Files for Analysis of SIRV Data

Different annotations are provided as a starting point for unified approaches to measure the performance of RNA-Seq experiments (www.lexogen.com/sirvs/#sirvsdownload). The FASTA file, **SIRV_150601a.fasta**, contains the complete exon and intron sequence together with a 1 kb long upstream and 1 kb long downstream sequence. The GTF files contain information about the variant structures. The following variations are provided:

SIRV_C_150601a.gtf contains the correct annotation of all 69 SIRVs that are in the Mixes E0, E1, and E2. These SIRVs are shown in blue in the SIRV figures of Appendix 7.1.

SIRV_I_150601a.gtf is one of several possibilities of an insufficient annotation. Here, some SIRVs which are actually present in the mixes are not annotated. These missing annotations are marked with an superscript (-)in the respective appendix figures.

SIRV_O_150601a.gtf is one of an endless number of possible over-annotations. Additional SIRVs are annotated, which are not present in the Mixes. These transcript variants are shown in green in respective appendix figures.

The possibilities of data evaluation using the SIRVs are manifold. The following proposal outlines the basic procedures which have to be performed for evaluating the performance of RNA-Seq pipelines.

Read Mapping

Barcodes provide the information for the demultiplexing step. After barcode and quality trimming the reads must be mapped to the respective genome, SIRVome, and where applicable ERCC sequences. All reads which map to the SIRVome can be filtered and treated separately.

The share of reads mapping to the SIRVome provides a first indication on the variability of the spike-in procedure. The SIRV content must be in relation to its expected mass or molar proportion as outlined in Table 3. For library preparations which aim to cover the length of RNA molecules and lead to measure such as FPKM the proportion of SIRV reads must obey the mass ratio while for library preparations which either tag or independently count RNA molecules the SIRV reads must obey the molar ratio.

Normalization

The correction of sample-specific biases is important for differential expression (DE) analyses. Varying RNA sample background, mRNA content and integrity, and variations in depletion and/or mRNA enrichment procedures lead to different SIRV Mix contents in the sequenced libraries. The mRNA content of total RNA samples can vary by a factor exceeding 2.5. The correction for such biases is important for the correct testing of differential expression, and subsequently normalizing the DE measurements in RNA samples themselves.

The offset factor is a measure of the RNA class distribution and can be used for SIRV control-based normalization. For an accurate quantification, however, a careful and quantitatively precise spiking procedure at the start of the workflow is a prerequisite. All measures and subsequent normalizations need to be set into context with obvious experimental variables like the achievable pipetting accuracy when operating in tiny volumes scales.

Input-Output Correlation

The assignment of SIRV reads has to be performed against one of the SIRV annotations. The abundances are calculated based on the read assignments, and have to be related to the known input amounts. Input-output correlations should be calculated in linear space as the set concentration range spans only 2 orders of magnitude. The Pearson product-moment correlation coefficient, Pearson's *r*, should approach 1. Because the input concentrations of the E0 sample are identical a simple measure of the variance is already a sufficient measure and should approach 0. The distribution of errors with respect to the individual variants and in the context to any competing sequences within the respective gene provide good indications for strengths and weaknesses of the sequencing pipelines.

Alternatively, concentrations can be visualized and calculated in log-space, in particular when including the entire background of the sample RNA into the correlation plots. All concentration values which have been given extreme low probabilities, e.g., SIRV FPKM values below 10^{-3}

(in particular, when using the reference annotation SIRV_O) should be set to a certain threshold, e.g., said 10^{-3}, before calculating correlation results like r-value as otherwise extreme distortions of the main correlation due to arbitrary low figures can be expected.

The equimolarity in SIRV Mix E0 allows for calculating variance as a significant quality measure. This approach can be applied also to the other Mixes because the 12 to 21 transcripts within each of the 4 SubMixes are at the same concentration. For each SIRV Mix the quality of the sequencing pipeline can be demonstrated as a set of 4 mean values together with the corresponding variances. It also provides the basis for a simplified correlation plot which consists of 4 data point clusters (compare Fig. 4).

Differential Expression

The most accurate and reproducible assessment can be realized by determining differential expression values or fold-changes. As the Mixes were prepared by precise volumetric combination of 4 SubMixes, the differentials are unaffected by other quality measures like the full-length integrity of the SIRVs. The fold-changes allow to calculate a number of parameters like the true and false positive rates, TP and FP, in calling differential expression. The Area Under the TP vs. FP Curve, AUC, can be taken as measure for the diagnostic performance in differential expression analysis.

Coping with Different SIRV Annotations

SIRV reads might be initially mapped using the correct SIRV_C annotation. The mapping should be repeated using different annotations like the also provided SIRV_I and SIRV_O.

The version SIRV_I (insufficient under-annotation) allows to judge the ability of a pipeline to detect new transcript variants. The experiment shows how reads of non-annotated 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 reflects a third situation. Here, more SIRVs are annotated than actually contained in the samples. The annotation comprises transcript variants which were discovered e.g., in other tissues, in the same tissue but at different developmental stages, which have been falsely annotated, or are relicts of earlier experiments, for which the high number of variants with the typical length of cloned ESTs are typical examples. In this setup, reads can be assigned to SIRV variants which are actually not part of the real sample. The degree and robustness of correct SIRVome detection in this setting is another measure for the pipeline performance.

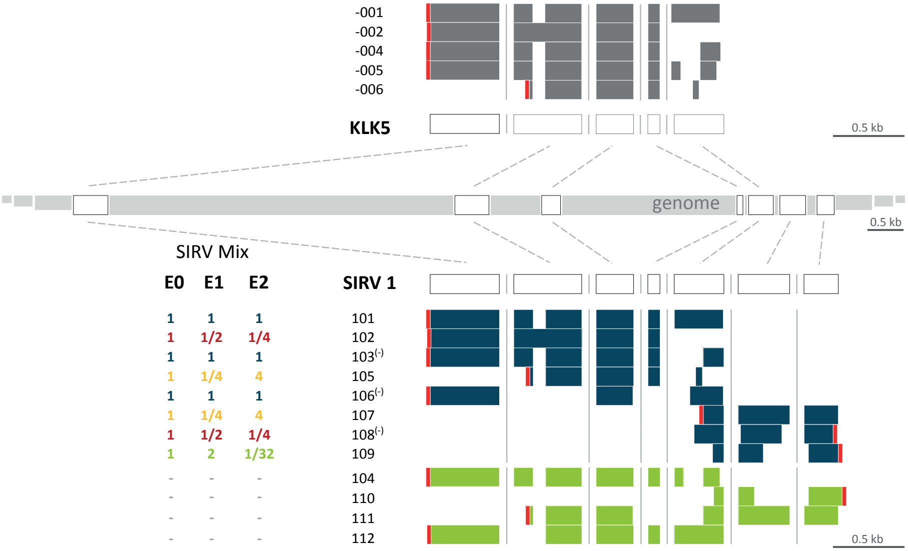
Knowing the biases introduced in isoform quantification enables to judge whether data sets are comparable across samples or experiments.

7. Appendix

7.1 SIRVs Alignment View

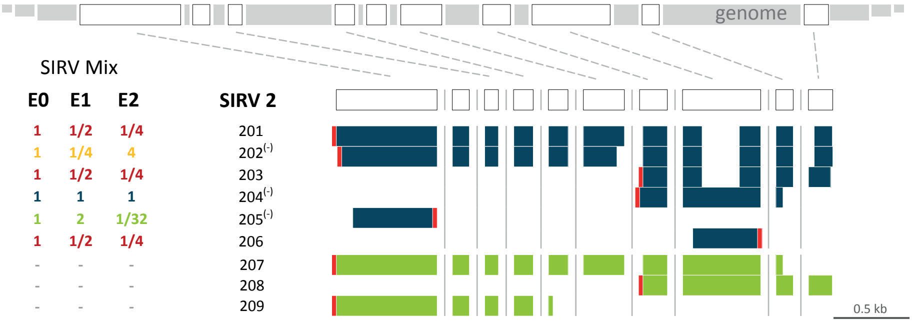
The individual transcript variants are schematically drawn in the condensed intron-exon format (see below) allowing to obtain an overview of the complexity of the transcript variants. However, minor start and end site variations which differ by just a few nucleotides are not visible in this representation.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

The individual SIRVs are present in predefined amounts in the different 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 (Mix E0), differ up to 8-fold (Mix E1) or up to 128-fold (Mix E2).



	SIRV Mix		
SIRV 1	E0	E1	E2
101	1	1	1
102	1	1/2	1/4
103$^{(-)}$	1	1	1
105	1	1/4	4
106$^{(-)}$	1	1	1
107	1	1/4	4
108$^{(-)}$	1	1/2	1/4
109	1	2	1/32
104	-	-	-
110	-	-	-
111	-	-	-
112	-	-	-

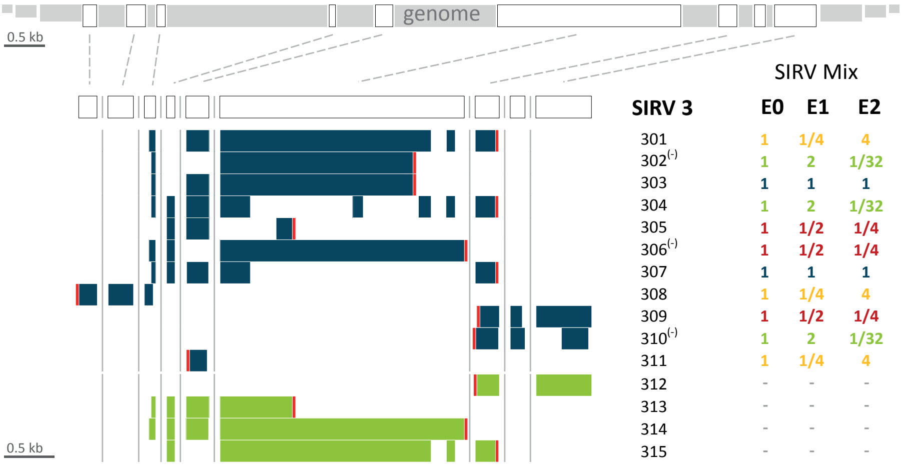
SIRV 1 | based on human gene KLK5. The human Kallikrein-related peptidase 5 gene was taken as template for SIRV1 gene generation. Its expression is up-regulated by estrogens and progestins, and alternative splicing results in multiple transcript variants encoding the same core protein. The current Ensembl annotation (GRCh38.p2) contains 5 transcript variants, KLK5-1, 2, and 4-6. Its condensed exon-intron structure is shown in upper section in **grey**. SIRV 1 contains 8 realized transcript variants (shown in **blue**) which are present in the mixes in the respective relative ratios as shown in the table to the left. SIRVs marked with 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**.



SIRV Mix

E0	E1	E2	SIRV 2
1	1/2	1/4	201
1	1/4	4	$202^{(-)}$
1	1/2	1/4	203
1	1	1	$204^{(-)}$
1	2	1/32	$205^{(-)}$
1	1/2	1/4	206
-	-	-	207
-	-	-	208
-	-	-	209

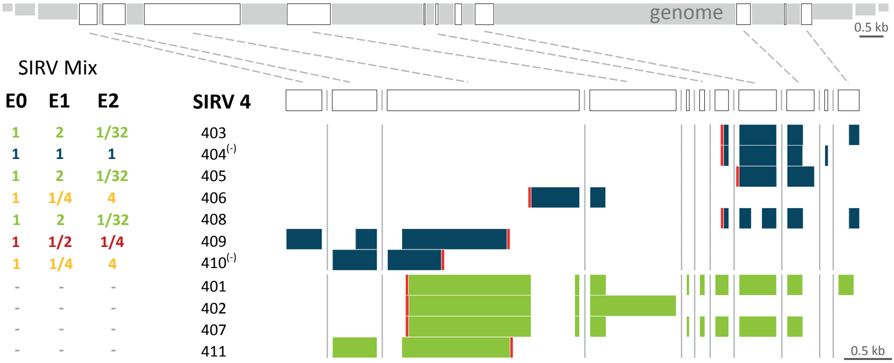
SIRV 2 | based on human gene LDHD contains 6 transcript variants (shown in blue) which are present in relative mixing ratios as shown in the table to the left. The transcript variants shown in green are additional annotations for alternative evaluation procedures, whereas the SIRVs marked with $^{(-)}$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



SIRV Mix

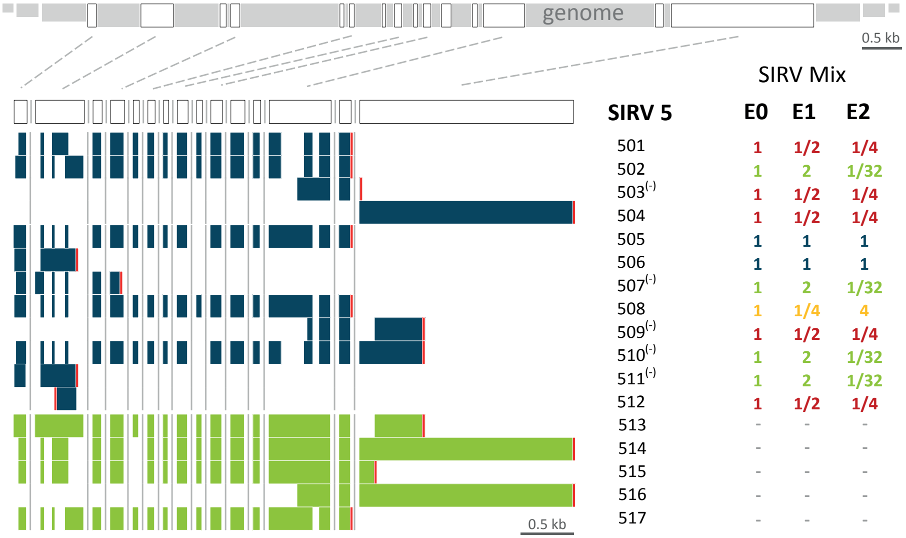
SIRV 3	E0	E1	E2
301	1	1/4	4
$302^{(-)}$	1	2	1/32
303	1	1	1
304	1	2	1/32
305	1	1/2	1/4
$306^{(-)}$	1	1/2	1/4
307	1	1	1
308	1	1/4	4
309	1	1/2	1/4
$310^{(-)}$	1	2	1/32
311	1	1/4	4
312	-	-	-
313	-	-	-
314	-	-	-
315	-	-	-

SIRV 3 | based on human gene LGALS17A contains 11 transcript variants (shown in blue) which are present in relative mixing 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.



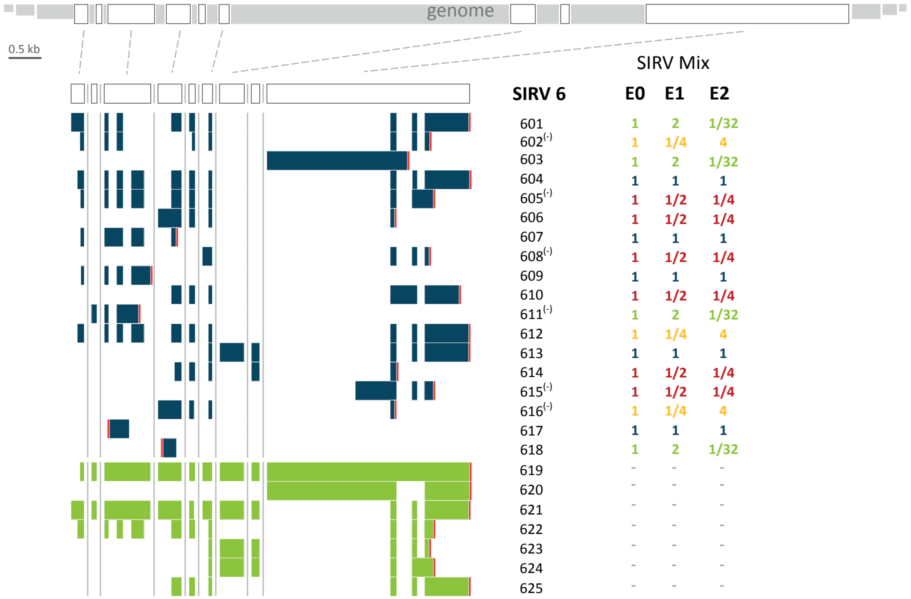
SIRV Mix			
E0	E1	E2	SIRV 4
1	2	1/32	403
1	1	1	404(-)
1	2	1/32	405
1	1/4	4	406
1	2	1/32	408
1	1/2	1/4	409
1	1/4	4	410(-)
-	-	-	401
-	-	-	402
-	-	-	407
-	-	-	411

SIRV 4 | based on human gene DAPK3 contains 7 transcript variants (shown in blue) which are present in relative mixing 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.



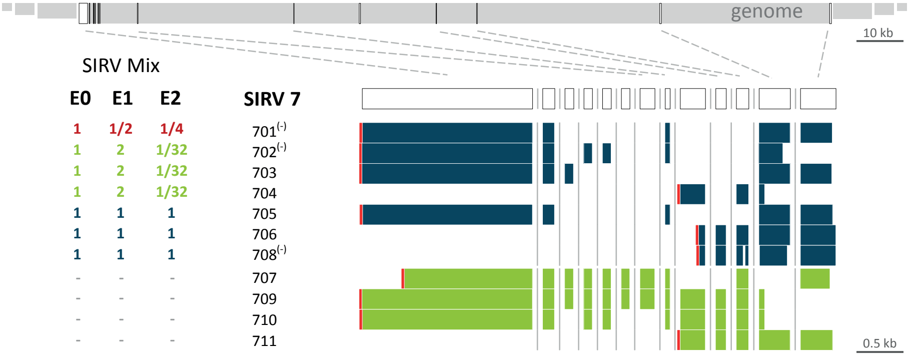
SIRV 5	SIRV Mix		
	E0	E1	E2
501	1	1/2	1/4
502	1	2	1/32
503(-)	1	1/2	1/4
504	1	1/2	1/4
505	1	1	1
506	1	1	1
507(-)	1	2	1/32
508	1	1/4	4
509(-)	1	1/2	1/4
510(-)	1	2	1/32
511(-)	1	2	1/32
512	1	1/2	1/4
513	-	-	-
514	-	-	-
515	-	-	-
516	-	-	-
517	-	-	-

SIRV 5 | based on human gene HAUS5 contains 12 transcript variants (shown in blue) which are present in relative mixing 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.



SIRV 6	SIRV Mix E0	SIRV Mix E1	SIRV Mix E2
601	1	2	1/32
602$^{(-)}$	1	1/4	4
603	1	2	1/32
604	1	1	1
605$^{(-)}$	1	1/2	1/4
606	1	1/2	1/4
607	1	1	1
608$^{(-)}$	1	1/2	1/4
609	1	1	1
610	1	1/2	1/4
611$^{(-)}$	1	2	1/32
612	1	1/4	4
613	1	1	1
614	1	1/2	1/4
615$^{(-)}$	1	1/2	1/4
616$^{(-)}$	1	1/4	4
617	1	1	1
618	1	2	1/32
619	-	-	-
620	-	-	-
621	-	-	-
622	-	-	-
623	-	-	-
624	-	-	-
625	-	-	-

SIRV 6 | based on human gene USF2 contains 18 transcript variants (shown in blue) which are present in relative mixing 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.



SIRV 7	SIRV Mix E0	SIRV Mix E1	SIRV Mix E2
701$^{(-)}$	1	1/2	1/4
702$^{(-)}$	1	2	1/32
703	1	2	1/32
704	1	2	1/32
705	1	1	1
706	1	1	1
708$^{(-)}$	1	1	1
707	-	-	-
709	-	-	-
710	-	-	-
711	-	-	-

SIRV 7 | based on human gene TESK2 contains 7 transcript variants (shown in blue) which are present in relative mixing 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.

7.2 Quality Parameters

Purification of T7 Transcription Products

The T7 transcription produced SIRV RNAs of high purity and high but varying integrity as determined by RNA length evaluation using capillary electrophoresis (Bioanalyzer RNA 6000 Pico Chip, Agilent). A series of optimized, tailored methods was applied to purify full-length RNAs with a minimal amount of side products. Examples are shown in Figure 6.

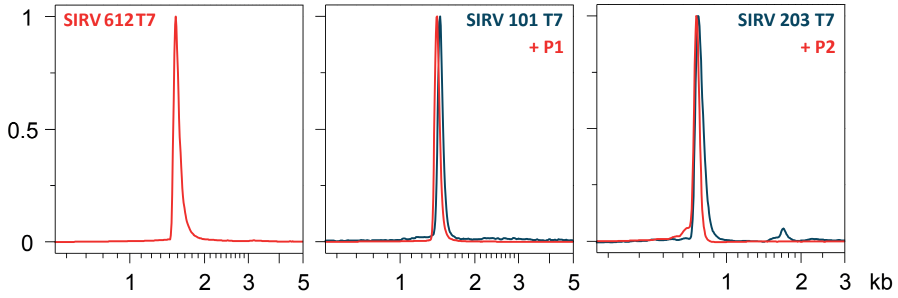


Figure 6 | Examples for T7 transcription and purification of SIRVs. Left, T7 transcription of SIRV 612 produced RNA of almost uniform correct length with 4/92/4 % in the pre-/ main- / and post-peak fractions. Middle, SIRV 101 T7 transcription shared a purity of 11/83/6 %, which could be increased to 4/95/1 % using purification method P1. Right, SIRV 203 T7 transcription displayed a purity of 3/88/9 % with distinct longer sequence artefacts 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.

Determination of the SIRV Integrity

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

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

with the values given as mean of the relative fractional mass content ± standard deviation.

The manufacturer's specification for the RNA 6000 Nano LabChip kit are 10 % CV for reproducibility of quantitation, 20 % CV for ladder quantitation accuracy and >20 % for the RNA sample quantitation accuracy.

Quantification of the SIRVs

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

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

The Nanodrop allows 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 the quantification of the final SIRV stock solution concentration measurement near 50 ng/µl is ± 4 %.

The molarity of each solution was calculated based on the base distribution of the SIRV sequences according to:

$$\text{MW [g/mol]} = A*329.2+U*306.2+C*305.2+G*345.2+159$$

7.3 Mixing Scheme

PreMixes

8 PreMixes were designed to contain 6 - 11 SIRV transcripts in equimolar ratios. Their length distribution allows for a unique identification in Bioanalyzer traces as shown in Figure 7A to monitor the occurrence and the integrity of the SIRVs in the PreMixes and subsequent Mixes (Figure 7B, 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

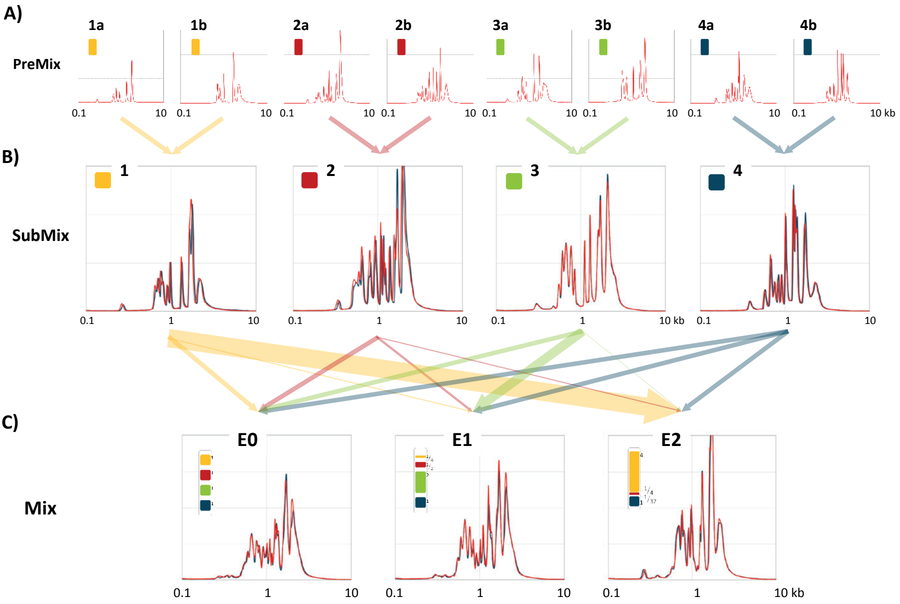


Figure 7 | SIRV mixing scheme to obtain Mixes E0, E1, and E2. A), the 8 PreMixes contain between 6 and 11 SIRVs which are different in length so that the SIRVs can be unambiguously identified in the Bioanalyzer traces. B) Two PreMixes each were combined in equal ratios to yield four SubMixes in total. C) These, in turn, were combined in defined ratios (see Fig. 4) to obtain the final Mixes E0, E1 and E2. Measured traces are shown in red, traces computed from the PreMix traces to validate SubMixes and final Mixes are shown in blue.

The 8 PreMixes were combined pairwise to give 4 SubMixes. The mixing process was quality monitored by electrophoresis as shown in Figure 7B. The volumetric preparation of the 4 SubMixes was controlled by Nanodrop concentration measurements (deviation of 0.8 % ± 2.5 % , maximal 4.5 %).

Final Mixes E0, E1, and E2

The 4 SubMixes were combined to Final Mixes with defined volumetric ratios, the monitoring of the mixing process by electrophoresis is shown in Figure 7C. Nanodrop concentration measurements showed a deviation of 5.1 % ± 3.3 % (maximal 8.6 %) from the calculated target concentrations.

Within very narrow margins all Bioanalyzer traces of Mixes resemble the sum of their respective Pre- and SubMix constituents (Figure 7). The relative peak shapes and positions are a reliable quantitative monitoring tool for the SIRV Mixes.

7.4 Downloads

Sequences

The SIRV sequences and the SIRVome sequences (including introns as well as 1 kb up- and downstream sequences) are available in FASTA format. The annotations are held in the GTF format, in which SIRV_C_150601a.gtf comprises the correct annotation of all physically present SIRVs. Optionally, insufficient under-annotation is provided in SIRV_I_150601a.gtf, and one over-annotation with about 30 % more transcript variants can be found in SIRV_O_150601a.gtf. All files can be accessed at:

www.lexogen.com/sirvs/#sirvsdownload

Concentration Ratios

SIRV concentrations in each of the Final Mixes can be downloaded in Excel table format.

7.5 Support

For the latest information and SIRVs-related questions write to:

info@lexogen.com

or call Lexogen Support at +43(0) 1 345 1212-41.

7.6 Safety

Chemical Safety

Follow general safety guidelines for chemical usage, storage, and waste disposal. Minimize contact with chemicals. Wear appropriate personal protective equipment such as gloves and lab coat when handling chemicals. Comply with the RNA handling guidelines when working with SIRVs (see chapter 6.1).

MSDS

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

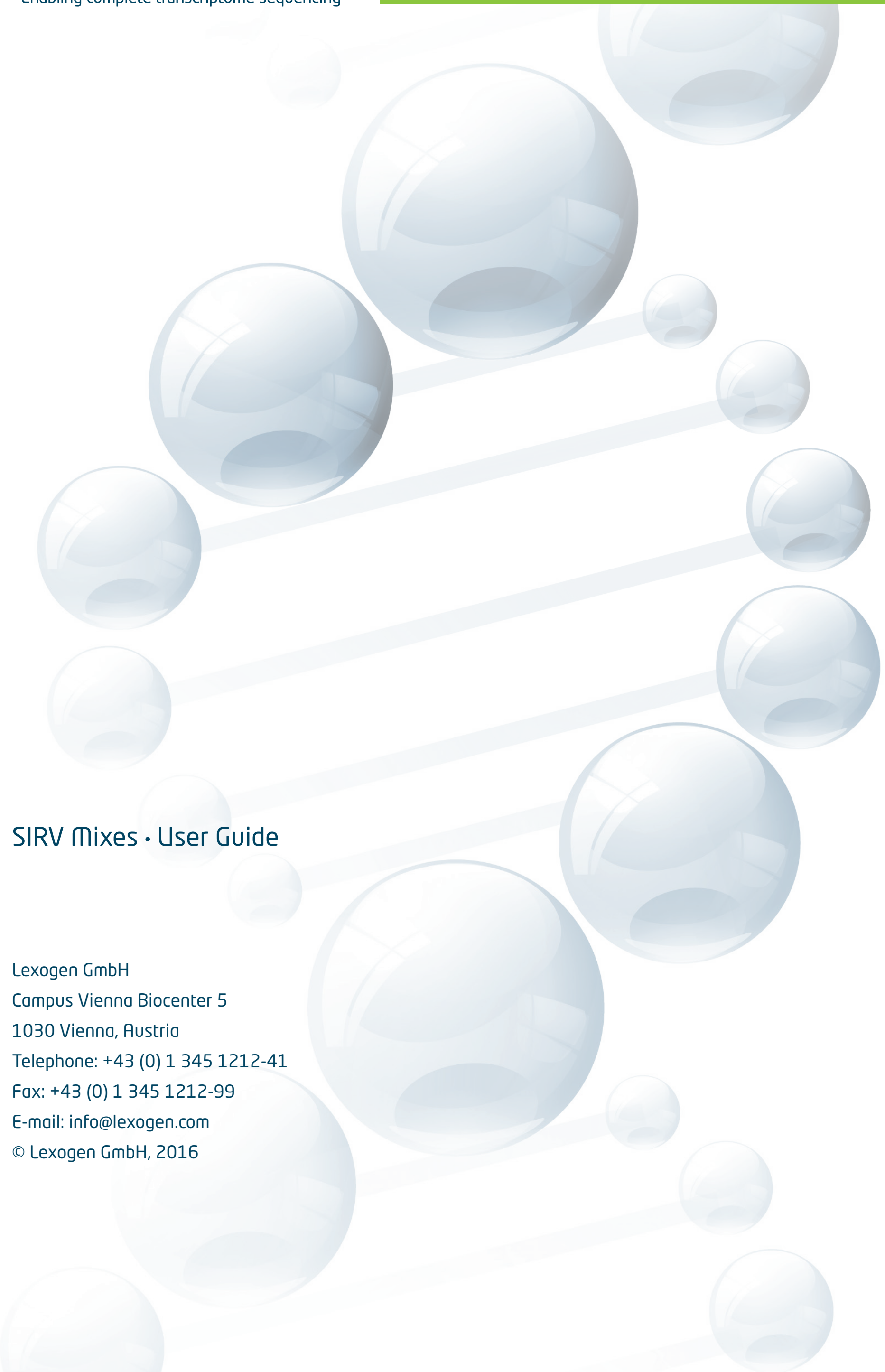
7.7 Certificates of Analysis

Certificates of Analysis provide quality control and product qualification information. They are available for the product lot numbers as stated on the tubes from our download section (www.lexogen.com/sirvs/#sirvsdownload).

8. Revision History

Publication No.	Change	Page
025UG063V0110	Product Release 2015-09-04,	
025UI063V0100	Initial Release 2015-06-03, first release of the documentation together with the FASTA and GTF sequence file package, *name*_150601a.*extension*,	

Enabling complete transcriptome sequencing



SIRV Mixes • User Guide

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