LEXOGEN
Enabling complete transcriptome sequencing

Spike-in RNA Variant Control Mixes (SIRVs)

- Transcript variants comprehensively addressing transcriptome complexity (e.g., alternative splicing, alternative transcription start- and end-sites)
- External transcript variant controls for the validation of RNA-Seq pipelines
- Evaluation of differential gene expression quantification on the transcript level
- Control variability across experiments for biases in gene expression quantification

Introduction

Spike-in controls are important in RNA-Seq experiments to assess sequencing technologies, workflows, and laboratories[1]. External controls will become essential when translating RNA sequencing into clinical diagnostics[2]. The application of external controls in RNA-Seq workflows is shown in Fig. 1.

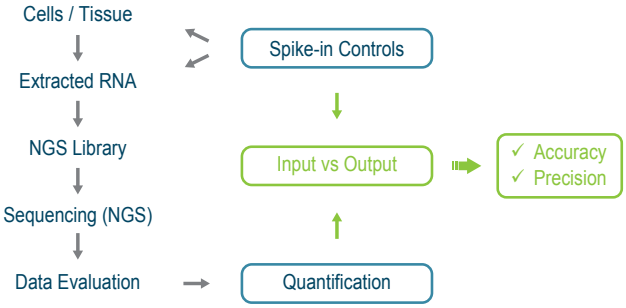

Figure 1 | Spike-in controls in RNA sequencing workflows. Controls can be spiked into a large number of individual samples be it homogenized cells, tissue samples, or the extracted RNA, before being processed together with the endogenous RNA. Data evaluation of the typically just 1% controls allows to determine values for accuracy and precision, hence the technical noise of the experiment. With the extent of the technical variance being deducted, the results from the endogenous RNA can be evaluated regarding biological variance and (differential) gene expression.

In the past, *in-vitro* transcripts (IVTs) were developed as spike-in controls for RNA research by the External RNA Controls Consortium[3]. These „ERCCs" allow to assess dynamic range, dose response, lower limit of detection, and fold-change response of RNA sequencing pipelines within the limitation of the mono-exonic, single-isoform RNA sequences. Because the ERCCs contain no transcript variants, one of the main challenges of sequencing complex transcriptomes - to identify and distinguish splice variants - could not be evaluated until now. Eukaryotic transcriptomes are characterized by extensive alternative splicing, alternative and antisense transcription, overlapping genes, and rare events like the formation of fusion genes. This complexity challenges the bioinformatics analysis of RNA-Seq experiments, in particular of NGS short reads data. The accurate gene expression measurements require the correct reconstitution of transcript variants through the mapping and assigning of reads to incomplete and changing genome annotations. The continuous technological developments in RNA preparation, library generation, sequencing, and bioinformatics algorithms improve the determination of isoforms and their concentrations. However, the performance of these methods cannot be assessed adequately without well-defined transcript spike-in controls that mimic the complexity of transcriptomes. The Spike-in RNA Variant Control Mixes (SIRVs) offer a unique solution to this problem.

SIRV Sequence Design

Seven genes were designed based on human gene structures with 6-18 transcript variants each, 69 transcripts in total. These comprehensively address alternative splicing, alternative transcription start and end sites, overlapping genes, and antisense transcription (example shown in Fig. 2).

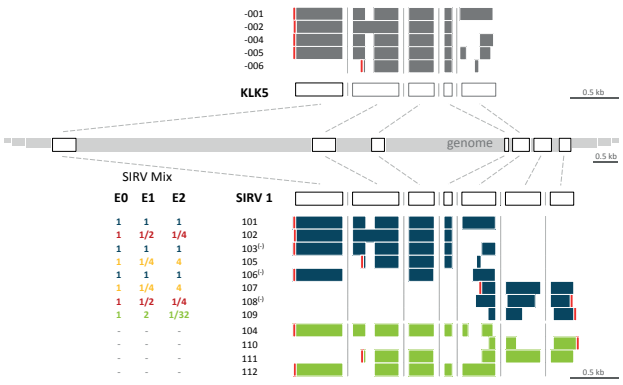

Figure 2 | Exon structures of transcript variants of SIRV1. Gene SIRV1 is derived from the human gene KLK5, and variants were added to comprehensively represent the transcriptome complexity. Transcripts in blue are part of the SIRV mixes, transcripts in green are part of an over-annotation, and (-) refers to transcripts omitted in an incomplete annotation. The polyadenylated 3'-end is marked in red, indicating sense and antisense orientations.

SIRVs range from 191 to 2 528 nt in length ($\bar{x}$, 1 134 nt; $\tilde{x}$, 813 nt), contain a 5'-triphosphate, and an additional 30 nt long poly(A)-tail, enabling oligodT based selection and priming (mRNA-Seq) in addition to other total RNA analysis methods. The GC-content varies between 29.5 and 51.2 % ($\bar{x}$, 43.0 %; $\tilde{x}$, 43.6 %). The exon sequences were created from a pool of database-derived genomes and modified by inverting the sequences to lose identity while maintaining a naturally occurring order in the sequences. Therefore, the artificial SIRV sequences are suitable for non-interfering qualitative and quantitative assessments in known genomic systems and complementary to the ERCC sequences. The splice junctions conform to 96.9 % to the canonical GT-AG exon-intron junction rule with few exceptions harboring the less frequently occurring variations GC-AG (1.7 %) and AT-AC (0.6 %). Two non-canonical splice sites, CT-AG and CT-AC, account for 0.4 % each.

In-vitro Transcript Production and Mixing

The RNAs were produced to highest quality specifications with the purpose to prevent shorter or longer by-products from interfering with the detection of sequence-similar transcript isoforms. All components and the mixing itself were carefully quality-controlled by photometric, weight, and microfluidics analyses (Fig. 3).

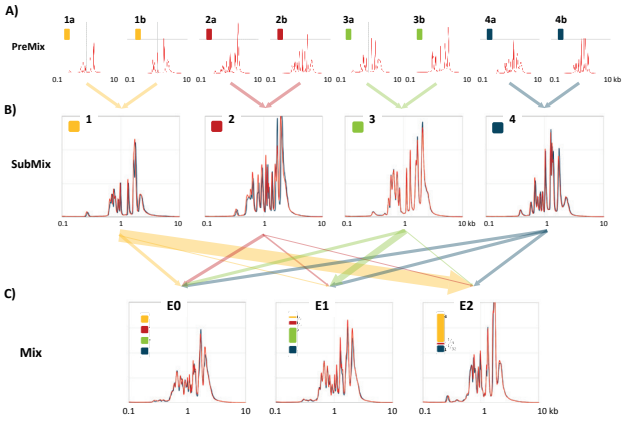

Figure 3 | SIRVs mixing scheme. PreMixes 1a - 4b (containing between 6 and 11 SIRVs that can be unambiguously identified in microfluidic traces) are combined pairwise in equal ratios to yield SubMixes 1 - 4, these are combined in defined ratios (see Fig. 4) to obtain the final Mixes E0, E1, and E2. Measured traces are shown in red, traces computed from the PreMix traces to validate SubMixes and final Mixes are shown in blue.

Transcript variants from the same gene are allocated across SubMixes, enabling the creation of final mixes with variants at equimolar level (Mix E0), with concentrations differing up to 8-fold (Mix E1) or up to 128-fold (Mix E2), as depicted in Fig. 4. Mix E0 contains all SIRV transcripts in equal concentration to directly show up workflow biases in transcript variant detection. Mixes E1 and E2 impose another challenge by mirroring the situation in cells at different expression stages or originating from distinct tissues, whereby their transcriptomes exhibit different relative abundances of individual transcript isoforms.

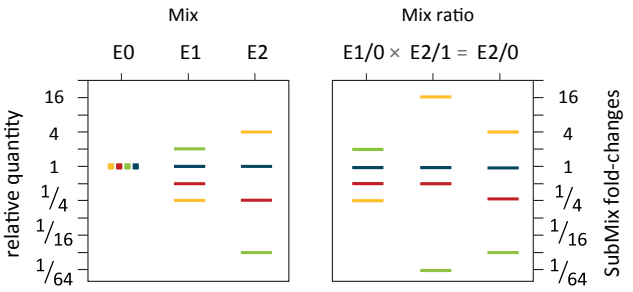

Figure 4 | Graphical representation of SIRV intra- and inter-mix ratios. SubMixes 1 - 4 are represented by different colors in the three SIRV Mixes E0, E1, and E2. **Left**, the intra-mix concentration ratios provide three different concentration settings to evaluate accuracy in relative concentration measurements. **Right**, the preset fold-changes allow for three possible inter-mix comparisons to evaluate differential gene expression measurements.

Preparing RNA-Seq Experiments with SIRVs

RNA-Seq covers a wide range of diverse experimental workflows. Therefore, it is important to plan the control of the experiments with respect to the amounts and the timing of adding SIRV controls. SIRVs can be spiked into cell lysis homogenates to assess the entire RNA extraction and preparation workflow for biases from the start on. Alternatively, SIRVs can be spiked into the purified total RNA or into selected fractions like ribosomal depleted RNA or poly(A)-enriched RNA. The share of reads allocated to the SIRVs provides important information about the RNA samples and the preparation process. The SIRVs User Guide provides several estimates and guidelines on how the samples and SIRV controls need to be combined to aim for target ratios of control reads. As a rule of thumb ribosomal RNA can be expected to comprise between 82 to 90 % of the entire RNA, the mRNA content for approximately 2.5 %[4], while the remaining transcripts are lncRNA, miRNA, siRNA, snoRNA, and other short RNA including tRNA. Hence, 100 ng total RNA will contain approximately 2.5 ng mRNA, and 25 pg SIRVs (1 µl of 1 : 1000 stock dilution) correlate to around 1 % of the mRNA content.

The use of SIRV Mixes E0, E1, and E2, is recommended for the validation of new workflows or when changing existing workflows to assess accuracy and precision in transcript variant detection, and in consequence the performance of differential gene expression measurements. For quality monitoring of individual experiments in established workflows, just one standard SIRV mix can be used. Here, we recommend the use of E0.

After RNA extraction, NGS library preparation and massive parallel sequencing the reads are mapped to the endogenous RNA and the 'SIRVome'. This workflow is evaluated by comparing the SIRV reads and subsequent calculations to the expected values.

Data Evaluation

mRNA Content

During library preparation the controls run alongside endogenous RNA. Because the spike-in amount of the SIRV controls and the amount of total RNA (or cells) is known, the mRNA content can be calculated based on the distribution of reads. The integrity of

the endogenous RNA and the type of NGS library preparation (fragmentation-based versus full-length methods; ribosomal RNA depletion or poly(A) enrichment, etc.) need to be taken into account when interpreting the results.

Coverage Plots and CoD Values

Coverage plots unambiguously show the quality of NGS experiments up to the mapping of reads. To obtain a comparative measure, gene-specific Coefficients of Deviation (CoD) can be calculated which describe the mean difference between measured and ideal coverage.

$$CoD = CoD(L'_{TSS},L'_{TES}) = \sum_{p=TSS1}^{TESe} (c_{ideal,p} - c_{real,p})^2 \div \sum_{p=TSS1}^{TESe} c_{ideal,p}$$

with L', characteristic length of the transient region; TSS, transcription start site; TSS1, first transcription start site; TES, transcription end site; TESe, last transcription end site; p, nucleotide position; c, coverage as the number of readings or base calls at the position p; real (or measured) coverage is scaled so that the integrals of ideal and real coverage are equal.

CoD values are a metric for the often hidden biases in the sequence data predominantly caused by an inhomogeneous library preparation, but also by the subsequent sequencing and mapping. A lower CoD correlates with a better agreement of measured and ideal coverage and eases the load on data evaluation algorithms that have to correct for these biases. CoD values can only be calculated when either a gene has just one transcript (e.g., ERCCs), or if the concentration ratios of the numerous transcript variants are known, for which the SIRVs provide an unique example. Because it is not possible to research thousands of genes over a wide dynamic range in a fast systematic manner, the inspection of the SIRVs provides for the first time a manageable focal point which is representative for the basic performance of the complete experiment. SIRVs enable spotlight inspection at 7 defined loci. Fig. 5 shows by way of example SIRV3 the comparison between two RNA-Seq experiments, LM1 and LM2, which were carried out at different laboratories (L) using the same library preparation kit and the same sequencing platform type but different machine systems (M). The SIRV3 coverage in the E0 sample clearly shows differences between the experiments: CoD_{LM1} for the forward strand is 0.228, and for the reverse strand 0.056, while CoD_{LM2} reaches values of 1.66 and 0.090 respectively. The mean CoD over all seven SIRV genes is for LM1 0.17 ± 0.19 and already 3.3-fold larger for LM2 with 0.56 ± 0.54.

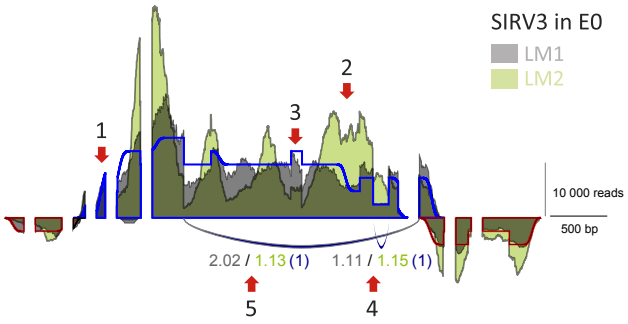

Figure 5 | Comparison of the expected and the measured coverages at the SIRV3 locus in Mix E0 in a condensed visualization with minimized and standardized intron sequences. The expected SIRV3 coverage is shown as superposition of individual transcript coverages in which the terminal sites have been modelled by a transient error function, and are shown by the blue and red traces. The measured coverages for LM1 are shown in grey, and for LM2 in green. The 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. While some sequences are covered as expected (1), others are over- or underestimated (2). The same differences occur in critical areas of telling reads (splice junctions) where prominent steps can be seen although the absolute values can differ significantly (3). These quantitative differences influence the count rates of telling reads which can be close to the expected values (4) or strikingly different (5). Here, the measured splice junction reads are shown by the color-coded numbers before the brackets, while the expected values are shown inside the brackets. LM1,2; experiments were carried out at different laboratories (L) using the same library preparation and the same sequencing platform type but different machine systems (M).

The coverage target-performance comparison highlights the inherent difficulties in deconvoluting read distributions to correctly identify transcript variants and determine concentrations. The distribution of telling reads, splice junctions, and reads towards the termini are references for the assignment of the remaining reads before calculating relative transcript variant abundances. The CoD does not allow to distinguish between periodicity and randomness in the biases nor does it forecast how well a data evaluation strategy can 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 up to this point in the workflow.

At this stage potential experimental errors become evident when the SIRV coverages diverge significantly from any reference experiments. The SIRVs spotlight inspection and CoD summary provide a referenceable starting point to compare the quality of sequencing experiments which would otherwise often be hidden in the mass of data derived from thousands of genes.

Accuracy and Precision

In the final step of the RNA-Seq experiments, data analysis pipelines calculate transcript abundances using the pool of mapped reads. In the first step only the 69 SIRV control transcripts have to be evaluated to obtain comparable quality measures before researching the large number of endogenous transcripts (e.g., 196165 endogenous transcripts are seen in the current human genome annotation GRCh38.p2). The obtained results are relative concentration measures in the format of either Fragments Per Kilobase of exon per Million fragments mapped (FPKM) or Transcripts Per Million (TPM). The concentrations have to be scaled by linear transformation in such way that all SIRVs together reach 69 in E0, 68.5 in E1, and 70.8 in E2, which is the dimensionless value identical to the fmol/µl in the respective SIRV stock solutions. Thereby, the relative quantities of the SIRVs become compatible to the normalized scale that was introduced in Fig. 4. The SIRV results can be matched with the expected values to determine the accuracy of each experiment, or the results of different experiments are compared to measure the precision or concordance of the measurements. Results can be visualized by correlation plots (Fig. 6).

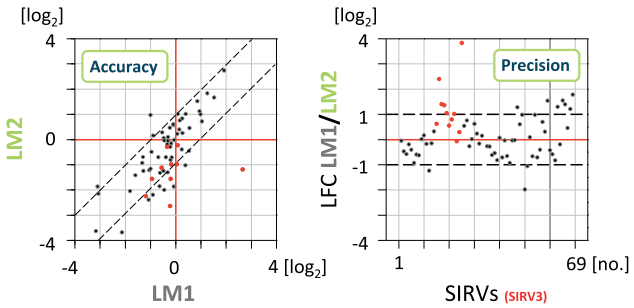

Figure 6 | Comparison of the SIRV concentration measures for E0 as an example. Left, the plot compares the LM1 and LM2 results to the expected value of 1, log_2 of 0, in E0. Accumulation of data points at the center cross or at one of the red lines indicates high accuracy, whereas high precision is signified by data concentration along the diagonal. **Right,** the difference between LM1 and LM2 is shown as log-fold change (LFC). The dashed lines mark the concentration measures with maximal 2-fold differences. The red dots represent the transcripts from SIRV3 which were obtained from the coverages shown in Fig. 5. The other transcripts are shown as black dots. LM1,2; experiments were carried out at different laboratories (L) using the same library preparation and the same sequencing platform type but different machine systems (M).

Boxplots in Fig. 7 exemplarily show for LM1 the concentrations of all SIRV mixes and the derived differentials. Here, the majority of all SubMix mean concentration values stay in good concordance with the expected values. The concentration ratios simulating differential expression measurements are even less prone to systematic offsets and show by trend narrower distributions. However, these clustered correlation figures already highlight the obvious and quite frequent outliers, i.e., transcript variants that are poorly resolved. The detailed analysis can identify particular deficiencies of RNA-Seq workflows which, e.g., have difficulties to resolve 5'-start sites or short antisense transcripts. SIRVs help to improve and optimize RNA-Seq pipelines.

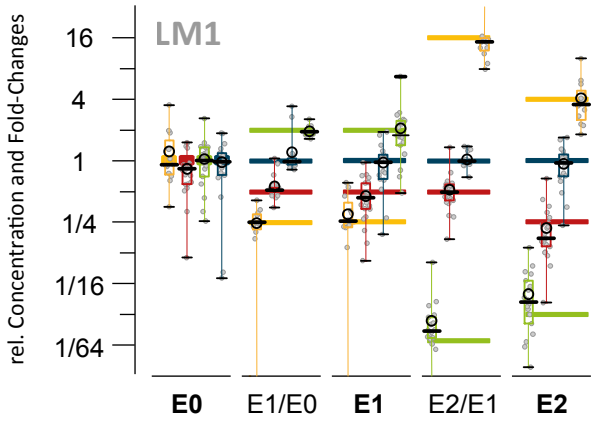

Figure 7 | Box plot overview of calculated concentration values for SIRV mixes and mix ratios. Reads from SIRV Mixes E0, E1, and E2 of experiment LM2 are exemplarily shown in reference to the known inputs (bars in colors correspond to the SIRV SubMixes of Fig. 4). The black circles mark the mean and the bold dashes the medians of the data points which are shown in grey; boxes span the 25 to 75 % region of the data points, and the black whiskers with connecting lines reach up to the min and max values and also indicate outliers outside of the scale of the graph.

The **accuracy** is calculated for each SIRV as the mean of the log-fold changes (LFCs) in all three mixes, and can be determined for individual SIRVs. The mean of all SIRVs together approaches zero due to required normalization when calculating the mRNA content. The **precision** is a measure of how consistent SIRVs are quantified in different samples, and is calculated as the LFC standard deviation. The experiment LM2 as shown in Fig. 7 reached a precision of 1.29, one of the best precision values we have measured so far. The precision is a measure for the technical noise and needs to be related to differential expression measurements when researching biological variance of different samples. The metrics of CoD, accuracy, and precision allow to assess the effect of changes in workflows and individual experiments.

Coping with Different SIRV Annotations

In virtually all RNA-Seq experiments the transcript annotations available will not match the transcript variants present in the samples. To enable an investigation of this scenario, exemplary insufficient SIRV annotations can be investigated. Thereby, it can be judged to which extent reads of non-annotated SIRVs are spuriously distributed to the annotated subset skewing the quantification, and if a pipeline is able to detect new transcript variants. Conversely, by aligning the reads to an over-annotation, a third situation can be evaluated, whereby transcripts might have been falsely annotated or are not expressed in the tissues sampled. This set-up challenges the robustness of a pipeline's performance and evaluates if reads are assigned to SIRVs that are not part of the real sample.

Conclusions

The Spike-in RNA Variant Control Mixes are the perfect means to evaluate complete RNA-Seq workflows for transcript variant quantification. Knowing the biases introduced during the experiments, and foremost the repeatability of such biases, enable to judge whether samples can be compared within experiments, and also between experiments which were performed at different sites, at different times, and/or by using different methods. It is important to quantify the technical noise of RNA-Seq experiments before researching biological variants.

[1] Munro, SA et al. (2014) Nat. Comm. 5:5125, doi:10.1038/ncomms6125
[2] Byron, SA et al. (2016) Nat. Reviews Genetics, doi:10.1038/nrg.2016.10
[3] Baker, SC et al. (2005) Nat Methods 2(10): 731-4.
[4] Shippy, R et al. (2006) Nat Biotechnol. 24(9): 1123-31.

Ordering Information

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

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

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

SIRVs™
Spike-in RNA Variant Control Mixes