LEXOGEN

Enabling complete transcriptome sequencing

025PF068V0100

lubio science

Your distributor in Switzerland

LubioScience GmbH
Baumackerstrasse 24
8050 Zürich
Phone 041 417 02 80 info@lubio.ch
Fax 041 417 02 89 www.lubio.ch

Spike-in RNA Variants Mixes (SIRVs)

- External RNA controls for the validation of mRNA-Seq quantification pipelines
- Transcript variants comprehensively addressing alternative splicing, alternative transcription start- and end-sites, overlapping genes, and antisense transcripts
- 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 essential in RNA-Seq experiments to assess workflow and platform properties (Fig. 1).[1]

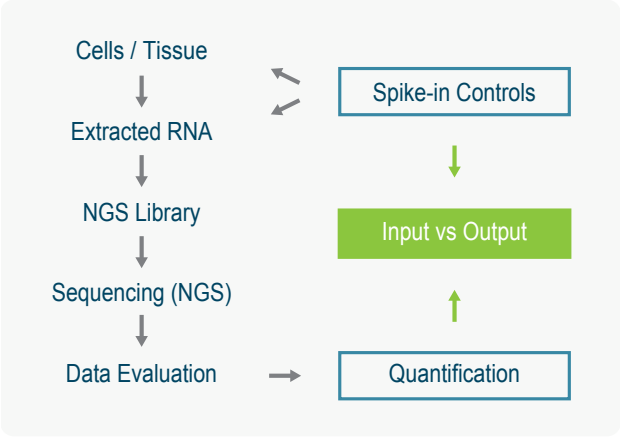

Figure 1 | Spike-in controls in the evaluation of RNA sequencing workflows. Controls can be spiked into the homogenized cells / tissue or the extracted RNA.

However, existing external RNA controls are generally mono-exonic and non-variant, significantly limiting their ability to reflect the true nature of eukaryotic transcriptomes, which are characterized by extensive alternative splicing, alternative and antisense transcription, overlapping genes, and rare events like the formation of fusion genes. This complexity prevents a straight-forward bioinformatics analysis. Accurate gene expression measurements require the reconstitution of transcript variants, which is complicated by incomplete and changing genome annotations and difficulties in sequence assembly. Progress in RNA preparation, library generation, sequencing and bioinformatics algorithms have improved the determination of isoforms and their expression. However, the performance of these methods cannot be assessed adequately without known transcript spike-in controls of representative complexity. Furthermore, only an evaluation of RNA-Seq biases can make data sets comparable across samples and experiments.

Transcript Variants

The Spike-in RNA Variant Mixes (SIRVs) offer a unique solution to this problem. Seven genes were designed, based on human gene structures and with 6 - 18 transcript variants each. These comprehensively address alternative splicing, alternative transcription start and end sites, overlapping genes, and antisense transcription (Fig. 2).

Figure 2 | Exon structures of transcript variants of 1 out of 7 SIRV genes. Transcripts in blue are part of the SIRV mixes, transcripts in green are only part of an over-annotation, and (i) refers to transcripts omitted in an incomplete annotation. The polyadenylated 3' end is marked in red, indicating sense and antisense orientations.

Accurate Production and Mixing of SIRV Components

The RNAs were produced to the highest 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.

SIRV Sequence Features

The spike-in RNAs conform to the GT-AG exon-intron junction rule and comprise an A(30) tail, enabling oligodT based selection and priming (mRNA-Seq), in addition to total RNA analysis (RNA-Seq).

The SIRV sequences show no significant similarity to nucleotide and protein entries in the NCBI database, confirming their suitability for qualitative and quantitative assessment in the context of essentially all known genomic systems and also in conjunction with existing external RNA controls.

Three SIRV Mixes for the Assessment of Differential Gene Expression on the Transcript Level

Transcript variants from the same gene are allocated across 4 SubMixes, enabling the creation of final mixes with variants at equimolar level (Mix E0), with concentrations differing up to 1:8 fold (Mix E1) or up to 1:128 fold (Mix E2), as depicted in Fig. 3. Mix E0 contains all SIRV transcripts equally 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 widely differing abundances of individual transcript isoforms.

The comparison of the pre-set SubMix fold-changes with the actual quantifications of the SIRVs in Mixes E0, E1, and E2 reveals the accuracy of differential gene expression quantification on the transcript isoform level (Fig. 3). For example, the calculation of true positive (TP) and false positive (FP) rates enables the evaluation of the diagnostic performance of a differential expression analysis by comparing AUC values (Area Under the TP vs. FP Curve).

SIRVs RNA-Seq Experiment Set-up and Evaluation

In a typical SIRVs experiment set-up, either the cell homogenate or purified RNA samples are spiked with one of the 3 SIRV Mixes. By spiking already the cell lysis homogenate the RNA extraction workflow can be assessed for biases as well. By spiking the purified RNA the share of reads allocated to the SIRVs (typically 2 - 3%, depending on RNA integrity, mRNA content and RNA-Seq type) can be controlled.

For workflow validation, only one SIRV mix (E0, E1, or E2) can be used. However, for differential gene expression assessment two or three different SIRV Mixes have to be matched with the RNA samples.

After RNA extraction, NGS library preparation, and high-throughput sequencing the reads mapping to the "SIRV genome" are evaluated separately. A bioinformatics analysis based on the correct SIRVs annotation can assess for input-output correlation, sensitivity, etc. However, the most accurate and reproducible assessment can be realized by matching the preset differential expression values or fold-changes with the measured ones.

A follow-up, in-depth investigation into the cause of deviations can be performed by comparing the SIRVs annotation with the transcript hypotheses formed by the data evaluation algorithms and by evaluating "SIRVome" coverage and mapped junctions in a genome browser.

Coping with Different SIRV Annotations

In virtually all RNA-Seq experiments, the transcript annotations available will not match the transcript variants actually present in the sample. To enable an investigation of this scenario, an insufficient SIRVs annotation table is also provided. 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 SIRV variants that are not part of the real sample.

The Spike-in RNA Variant Mixes are the perfect means to evaluate RNA-Seq workflows that are based on transcript variant quantification in general and the associated bioinformatics algorithms in particular. Knowing the biases introduced in isoform quantification enables to judge whether data sets are comparable across samples or experiments.

1. Munro, SA *et al.* (2014) *Nat. Comm.* 5:5125, doi:10.1038/ncomms6125

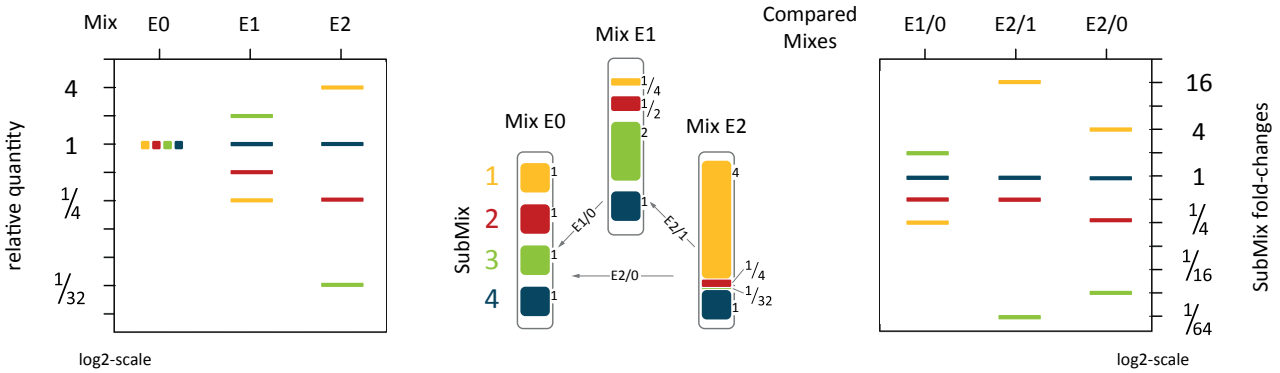

Figure 3 | Graphical representation of the SubMix (1-4) distribution in the 3 SIRV Mixes and the resulting intra- and inter-mix ratios. The 4 SubMixes are represented by different colours and contain between 12 and 21 SIRVs to keep the total molarity and weight of the mixes evenly balanced at 69.5 fmol/µl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

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

SIRVs™
Spike-in RNA Variant Control Mixes