TECHNICAL SPECIFICATION

ISO/TS 22690

First edition
2021-10

Genomics informatics — Reliability assessment criteria for high-throughput gene-expression data

Informatique génomique — Critères d'évaluation de la fiabilité des données d'expression des gènes à haut débit

Reference number
ISO/TS 22690:2021(E)

ISO copyright office
CP 401 • Ch. de Blandonnet 8
CH-1214 Vernier, Geneva
Phone: +41 22 749 01 11
Email: copyright@iso.org
Website: www.iso.org

Published in Switzerland

Contents

Page

Foreword

ISO (the International Organization for Standardization) is a worldwide federation of national standards bodies (ISO member bodies). The work of preparing International Standards is normally carried out through ISO technical committees. Each member body interested in a subject for which a technical committee has been established has the right to be represented on that committee. International organizations, governmental and non-governmental, in liaison with ISO, also take part in the work. ISO collaborates closely with the International Electrotechnical Commission (IEC) on all matters of electrotechnical standardization.

The procedures used to develop this document and those intended for its further maintenance are described in the ISO/IEC Directives, Part 1. In particular, the different approval criteria needed for the different types of ISO documents should be noted. This document was drafted in accordance with the editorial rules of the ISO/IEC Directives, Part 2 (see www.iso.org/directives).

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. ISO shall not be held responsible for identifying any or all such patent rights. Details of any patent rights identified during the development of the document will be in the Introduction and/or on the ISO list of patent declarations received (see www.iso.org/patents).

Any trade name used in this document is information given for the convenience of users and does not constitute an endorsement.

For an explanation of the voluntary nature of standards, the meaning of ISO specific terms and expressions related to conformity assessment, as well as information about ISO's adherence to the World Trade Organization (WTO) principles in the Technical Barriers to Trade (TBT), see www.iso.org/iso/foreword.html.

This document was prepared by Technical Committee ISO/TC 215, *Health informatics*, Subcommittee SC 1, *Genomics informatics*.

Any feedback or questions on this document should be directed to the user's national standards body. A complete listing of these bodies can be found at www.iso.org/members.html.

Introduction

High-throughput gene-expression profiling, including data generated from microarray, next-generation sequencing, and other forms of high-throughput technologies, is a revolutionary technology for genomic studies. It is a fast-moving field both in terms of innovation in measurement technology as well as advances on the data analysis side. High-throughput expression technology enables us to efficiently study complex biological systems and biological processes, mechanisms of diseases, and strategies for disease prevention and treatment. This technology is currently applied in the biomedical research community and industry, and plays an important role in disease characterization, drug development and precision medicine [1][2][3][4].

Challenges and pitfalls in the generation, analysis, and interpretation of high-throughput expression profiling data need to be addressed within the scientific community. Development of omics-based products that influence or improve patient health has been slower than expected. Studies attempting to reproduce findings of 53 papers in preclinical cancer research confirmed only 6 (11 %) of the results[5]. Misleading papers result in considerable expenditure of time, money and effort by researchers following false trails. This affects companies and investors, presenting yet another barrier for the translation of academic discoveries into new medicines by diverting funds away from real advances [6][7]. Irreproducible or inconsistent results could contribute to patient risk or death. As more and more irreproducible reports occur, some scientific journals reported the issue in 2014 [8][9]. The essential role of reproducibility of scientific research has been widely recognized [10].

There exist different reasons for low reproducibility in omics research. One possible reason is the complexity of omics data. The fact that the size of data is so massive that the manual inspection of data quality and analysis results is often impossible. Thus, quality control processes for high-throughput expression experiments are essential for the improvement of reproducibility of biological results.

The MicroArray and Sequencing Quality Control (MAQC/SEQC) consortia conducted three projects [11][12][13] to assess the reliability and reproducibility of genomics technologies, including microarrays, genome-wide association studies, and next-generation sequencing. This has led to the formation of the Massive Analysis and Quality Control Society (MAQC Society) [23], which is dedicated to quality control and analysis of massive data generated from high-throughput technologies for enhanced reproducibility and reliability [14]. It has provided a collection of quality metrics for expression data evaluation that corresponds to the reliability and reproducibility of high-throughput gene expression data for quality control, including (i) from sample to RNA, (ii) expression profiling, (iii) quality control metrics in RNA-seq, (iv) detecting differentially expressed genes, (v) biological interpretation, and (vi) spike-ins. Similar and complementary efforts have been reported elsewhere [15][16].

High-quality data are the foundation for deriving reliable biological conclusions from a gene-expression study. However, large differences in data quality have been observed in published data sets when the same platform was used by different laboratories. In many cases, poor quality of data was due not to the inherent quality problems of a platform but to the lack of technical proficiency of the laboratory that generated the data. Therefore, proficiency testing, an assessment of the overall competence performed through inter-laboratory comparisons, is introduced in this document to establish and monitor the quality of laboratory tests.

This document can be utilized to (i) enhance community's understanding of technical performance of high-throughput gene expression; (ii) benefit the interoperability of qualified gene-expression data by researchers, commercial entities and regulatory bodies, (iii) improve the application of high-throughput gene expression in industry and clinics, (iv) promote the acceptance of transparent reporting according to the FAIR (findable, accessible, interoperable, and reusable) data principles [17], and (v) contribute to the development of precision medicine.

Genomics informatics — Reliability assessment criteria for high-throughput gene-expression data

1 Scope

This document specifies reliability assessment criteria for high-throughput gene-expression data.

It is applicable to assessing the accuracy, reproducibility, and comparability of gene-expression data that are generated from microarray, next-generation sequencing, and other forms of high-throughput technologies.

This document identifies the quality-related data for the process of the next-generation sequencing of RNA (RNA-seq). The sequencing platform covered by this document is limited to short-read sequencers. The use of RNA-seq for mutation detection and virus identification is outside of the scope of this document.

This document is applicable to human health associated species such as human, cell lines, and preclinical animals. Other biological species are outside the scope of this document.

From a biological point of view, expression profiles of all genetic sequences including genes, transcripts, isoforms, exons, and junctions are within the scope of this document.

2 Normative references

There are no normative references in this document.

3 Terms and definitions

For the purposes of this document, the following terms and definitions apply.

ISO and IEC maintain terminology databases for use in standardization at the following addresses:

— ISO Online browsing platform: available at https://www.iso.org/obp

— IEC Electropedia: available at https://www.electropedia.org/

3.1
adapter
short, chemically synthesized oligonucleotide that can be ligated to the ends of DNA or RNA molecules

3.2
alignment ratio
percentage of total sequenced reads aligned to an intended target region

Note 1 to entry: Alignment ratio is different according to the definition of the gene structure (“annotation”) in that region, and is also affected by the origin of the RNA sample, library preparation, sequencing approach, and aligner.

3.3
batch effect
systematic technical variation in data unrelated to biological factors of interests and caused by processing samples in different batches

3.4
differentially expressed gene
DEG
gene that exhibits difference or change in read counts or expression level between two experimental conditions

3.5
functional annotation
process of attaching biological information to sequences of genes or proteins

3.6
guanine cytosine content
GC content
percentage of nitrogenous bases on a DNA or RNA molecule that are either guanine (G) or cytosine (C)

3.7
gene
specific sequence of nucleotides located on a chromosome as a functional unit of inheritance transferred from a parent to offspring

3.8
gene body coverage
description of the overall reads density over the gene region

3.9
gene expression
process by which information from a gene is used in the synthesis of a functional gene product

3.10
insert size
length of the sequence between the adapters

3.11
isoform
one of many forms a gene sequence transcribed from DNA to RNA

3.12
microarray
multiplex lab-on-a-chip on a solid substrate (usually a glass slide or silicon thin-film cell) that assays large numbers of biological molecules using high-throughput screening through miniaturized, multiplexed and parallel processing and detection methods

3.13
mismatch
erroneous insertion, deletion, or mis-incorporation of bases that can arise during DNA replication and recombination, as well as repairing of some forms of DNA damage

Note 1 to entry: Mismatch also refers to the fact that two sequences do not match exactly.

3.14
outlier
observation that appears to deviate markedly from other observations in a study, resulting from biological or technical differences from the rest of the samples of the study group

3.15
read
sequence of a cluster that is obtained at the end of the sequencing process

3.16
reproducibility
fundamental hallmark of good science describing whether the results obtained from one situation can be reproduced under another situation

3.17
RNA Integrity Number
RIN
number at the scale from 1 to 10 to indicate the integrity of an RNA sample, with a RIN of 1 indicating completely degraded RNA and a RIN of 10 meaning a perfectly intact RNA sample

3.18
RNA sequencing
RNA-seq
high-throughput sequencing technology to reveal the presence and quantity of RNA molecules in a biological sample at a given moment in time

3.19
RNA spike-in
RNA transcript added to an RNA sample for calibrating measurements in a high-throughput experiment, based on its known sequence and abundance

3.20
sequencing depth
number of times a genomic region has been sequenced

3.21
transcriptome
set of all RNA molecules in one cell or a population of cells for a specific developmental stage or physiological condition

4 From sample to RNA

4.1 General

High quality, intact RNA shall be extracted from samples. Store the extracted RNA under conditions recommended by the vendor (e.g. -80 °C or liquid nitrogen) to preserve the integrity of the RNA sample[18][19]. The RIN of the MAQC RNA reference materials [11] has been found to be stable over a period of more than 16 years under -80 °C.

For RNA-based experiments, different vendor-provided RNA library construction kits for RNA-seq or hybridizations for microarrays have different input requirements of RNA quality and quantity. Confirm the required quality and quantity of RNA samples before use in a downstream application.

This document is intended for laboratory samples and can be too strict to be used in the medical field, where sample quality can be compromised. For the handling of clinical samples for RNA analysis, see ISO 20166-1, ISO 20184-1 and ISO 20186-1.

The quality of RNA samples can be monitored in several ways. Considerations for RNA quality include the following metrics.

4.2 RNA integrity

RNA integrity shall meet requirement criteria of downstream experiments. For ribosomal RNA depletion protocols, the RIN shall be at least 3, whereas for polyA selection protocols, the RIN shall be at least 7.

4.3 RNA concentration

RNA concentration shall meet the requirement criteria of downstream experiments. An RNA sample shall be kept at a concentration above 100 ng/ul, whenever possible, to minimize degradation.

4.4 RNA purity

Avoid contamination of RNA samples with other molecules (i.e. proteins, RNA from other samples, and organic compounds). For an RNA sample, the 260/280 ratio of absorbances at 260 nm (RNA) and 280 nm (protein) shall be around 1,8 to 2,0. Similarly, the 260/230 ratio of absorbances at 260 nm (RNA) and 230 nm (organic compounds) for an RNA sample shall be around 1,8 to 2,0.

Regardless of the method(s) chosen to assess RNA quality, it shall be ascertained that the acceptance criteria for the RNA samples are consistently appropriate to yield RNA quality that is suitable for the analytical method selected [18][19]. The selected RNA isolation method shall minimize genomic contamination of the isolated RNA because genomic DNA could negatively affect downstream applications.

5 Expression profiling

The operators shall use validated standard operating procedures (SOPs) addressing all aspects of processing to generate RNA-seq, microarray, or other forms of high-throughput expression data, and all operators shall be fully trained on all protocols prior to initiating the sample.

Equipment shall be on an appropriate maintenance schedule and the laboratory environment shall be maintained in accordance with the manufacturer's recommendations. It is also advisable to establish appropriate maintenance schedules for all equipment, and ensure that the laboratory environment is maintained in accordance with the SOPs. See ISO 15189 and ISO 15190.

6 Quality control metrics in RNA-seq analysis

6.1 General

The analysis of RNA-seq data has as many variations as there are applications of the technology. Typically, basic strategy for regular RNA-seq analysis includes the following steps:

a) to align reads to the genome with a gapped aligner algorithm or to the transcriptome with an ungapped aligner algorithm;

b) to detect and quantify known/novel transcripts with or without an annotation file;

c) to identify lists of differentially expressed genes from the differences between the biological states under investigation, using a variety of statistical and analytical tools; and

d) to interpret the biological meaning of gene expression changes.

According to the above analysis processes, the quality control metrics in RNA-seq analysis can be divided into several levels: sequencing read, alignment, detection and quantification, differentially expressed genes, and biological interpretation of differential gene expression. Evaluation of sample origin or identity and batch effect is also included.

6.2 Sequencing read

6.2.1 Total number of reads

Successful gene-expression profiling can be achieved with levels as small as 10 million reads. For studies that involve investigation of alternative splicing, gene fusion detection and novel transcript