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

identification, higher total number of reads shall be able to adequately cover not just the exons but also exon–exon junctions.

6.2.2 Read length

Average read length shall be ≥35 bp.

6.2.3 Base call quality

A quality score (Q-score) is an indication of the probability of an error occurring during base calling. A higher quality score means a more reliable base call. A Q-score of 40, 30, 20, and 10 corresponds to an error probability of a base call of 0,000 1 (1 in 10,000), 0,001 (1 in 1,000), 0,01 (1 in 100), and 0,1 (1 in 10), respectively. Quality of the signal-to-noise ratio shall be monitored by examining the quality scores and quality of signal-to-noise ratio across a read for each sequencing run. Informatics filters shall be established to eliminate any reads with raw base calls lower than the established quality score threshold.

6.2.4 GC content

GC content shall be monitored with every run to detect changes in test performance. The expected GC content for the sequenced data shall be an approximation of the GC content for the reference sequence.

NOTE In a normal random library, GC content usually is roughly a normal distribution, where the central peak corresponds to the overall GC content of the underlying genome.

6.2.5 Overrepresented sequence

A sequencing library shall contain a diverse set of RNA sequences. A single sequence making up more than 0,1 % of the total reads is considered as an overrepresented sequence.

A single sequence well overrepresented in the data set can indicate either its high biological significance or its contamination.

6.2.6 Adapter residue

Adapters shall be ligated to every single DNA molecule during library preparation. Adapter residue is the undesired sequencing of partial or complete adapter sequences.

If an adapter sequence presents in more than 5 % of all reads, adapter trimmers shall be used to remove this sort of data, or an aligner algorithm supporting partial alignment shall be used.

6.3 Alignment

6.3.1 Alignment ratio

Over 70 % of reads from a regular RNA-seq based on sequencing-by-synthesis technology are expected to align onto the human genome.

6.3.2 Gene body coverage uniformity

Ideally, gene body coverage is uniform from the 3'- to 5'-end of the gene. Coverage profile is the most intuitive way to check uniformity.

NOTE When the coverage uniformity profile is not even, it is an indication of poor quality of library preparation, bias in PCR amplification, or RNA degradation.

6.3.3 Strand specificity

When performing strand-oriented sequencing, investigators shall only obtain the reads from the strand from which the RNA was originally transcribed. The specificity of a strand-specific protocol can be calculated by comparing the strand of reads to the strand of a reference gene.

EXAMPLE If there is only one gene located in the forward strand in a particular region, one expects the majority (>99 %) of reads aligned to this region will be forward reads.

If a non-strand-specific protocol is used, then the reads representing sense and anti-sense RNA shall be roughly 50–50.

6.3.4 Insert size

The distribution of insert size shall follow a non-normal distribution with a peak equal to the targeted size and a long right tail. Insert size is variable across library construction protocols.

Summary statistics of the insert size shall be computed using the median or peak, rather than the mean.

6.3.5 Mismatch

Mismatch types and rates are variable across sequencing platforms. When the ratio of mismatches is much higher than expected, then this is an indication of low quality of base calls, and potentially there can be errors.

6.3.6 Contamination from other sources

A qualified sample shall not be contaminated with materials from other species in excess of 5 %. Contamination from other sources can be detected by aligning unaligned reads to a collection of reference genomes of other candidate species. A small percentage (e.g. <5 %) of detected exogenous RNA usually does not affect follow-up data analysis. A large percentage (e.g. 10 %) of detected exogenous RNA is an indication that the sample is contaminated by RNA from other species.

6.4 Expression

6.4.1 Expression distribution

The magnitude of expression can be visualized in either box plots or density plots. Outlier samples that are not modelled (e.g. by exclusion in extreme cases) can cause inaccurate differential expression analysis.

Box plot shows how the expression values are distributed for each sample. The distributions need to be similar for the different samples to be comparable, i.e. the standard deviation of the expression values from one sample shall not be statistically different from that of other samples. If this is not the case, the data shall be normalized. Expression distributions can also be summarized by means of a histogram or density plot.

6.4.2 Expressed genes

The number of expressed genes per sample is highly dependent on two criteria:

(1) the total number of reads sequenced and aligned to genes;

(2) the threshold used for detection of gene expression.

It is assumed that the number of expressed genes shall be roughly equal (with less than 10 % difference) for all samples within the study within the same phenotype category group (i.e. controls).

NOTE The number of expressed genes is also protocol dependent, and ribosomal depletion protocols can detect more genes than polyA selection protocols.

6.4.3 Saturation

Reads shall be resampled into 5, 10, ..., 100 % of the total aligned reads and expression values shall be subsequently recalculated using each subset. For a particular transcript, expression values can vary at the beginning with very small sample sizes, but finally could reach a plateau. If sequencing depth is saturated, the estimated expression values will be stationary or reproducible.

6.4.4 Reproducibility

Reproducibility among replicates and for possible batch effects[20] shall be checked. If gene expression differences exist among experimental conditions, it shall be expected that biological replicates of the same condition will cluster together in a principal component analysis (PCA) or hierarchical clustering plot. The power of the RNA-seq data to correctly separate biologically different groups of samples shall be carefully examined and can be expressed as the ratio of inter-sample-group distance over intra-replicate distance. Batch effects obscure the resolution of biologically different types of samples by gene-expression data.

NOTE High correlation coefficient between replicates of the same sample does not automatically lead to clear separation of biologically different groups of samples. Quality metrics based on replicates of the same sample (such as correlation coefficients or coefficient of variation) shall not be applied as the sole metrics to estimate reproducibility of RNA-seq data.

6.5 Differentially expressed genes

Specific gene sets derived from expression experiments can be proposed as genomic biomarkers for a specific endpoint in a defined context. Such specific gene/transcript sets shall be reproduced upon review if the analysis protocol is identical to that reported.

To determine which genes are in fact differentially expressed, a number of factors need to be considered that can have confounding effects:

a) rejection criteria for low-count and (or) extremely high-count genes;

b) rejection criteria for outlier samples;

c) platform-specific normalization protocols;

d) data analysis protocol for gene annotation databases;

e) data analysis protocol for selecting differentially expressed genes.

There is no consensus currently regarding the appropriate choices for each of these factors. The settings and version numbers of programs used, as well as operators and other metadata shall be recorded carefully, and the repetition of important analyses using more than one software package shall be considered. A pilot study shall be conducted to choose the optimal method, based on samples to which RNA spike-in has been added in known quantities or a set of DEGs with an independent platform such as quantitative RT-PCR or droplet digital PCR (ddPCR).

RNA-seq experiments shall have a minimum of three biological replicates when sample availability is not limiting to allow all of the differential expression methods to leverage reproducibility between replicates.

To the extent that genomic biomarker sets become part of a decision-making process in drug development or therapeutic applications, transfer of genomic biomarker sets from one platform to other platforms, either other high-throughput platforms or low-throughput platforms, shall be attempted only after these differentially expressed genes are claimed to be sensitive, specific, and reproducible.

6.6 Biological interpretation of differentially expressed genes

General procedure of biological interpretation of differentially expressed genes can be described in two major steps: data support (annotation database) and data mining (algorithm and statistics).

Biological interpretation requires the availability of sufficient functional annotation data for the transcriptome under study. Multiple annotation resources shall be exploited to annotate data in different aspects.

A number of analysis algorithms and statistics are available and are being further developed to help the biological interpretation. But there is no consensus currently regarding the appropriate choices. The settings and version numbers of programs and databases used shall be carefully documented. The biological significance of gene sets shall be accompanied by a standard set of information that will enable recapitulation of the analysis and assessment of the validity of the interpretation by reviewers.

6.7 Sample certificate of origin

The accurate correspondence between the experimental conditions or sources of the samples and the eventual data shall be checked and maintained. Meta data ranging from original sample ID, demographic information, and the information related to sample processing and data generation and analysis shall be accurately recorded to avoid mislabelling.

EXAMPLE The concordance between the sex or tissue-of-origin of the samples predicted from gene-expression data and that reported in the meta data about the samples can be indicative of sample certificate of origin.

6.8 Quality control of batch effects

Quality control of batch effects shall be exercised to identify any and all systematic biases, especially in large datasets, which can result in misleading interpretations during data analysis especially when they are confounded with critical study factors. Data shall be checked for systematic biases from quality parameters in 6.1 to 6.3.

Effective monitoring, identification, and correction of batch effects is critically important for longitudinal large cohort studies involving gene-expression profiling, and could be accomplished by profiling the same reference materials concurrently with study samples during different batches.

7 Spike-in controls

Spike-in control is carried along the whole process and undergoes the same handling steps as the investigated sample, from initial quantification to final downstream processing, and hence reflects performance of the sample quality. If a sequence error is observed in the spike-in control, the same error can occur in all likelihood in the main sample.

RNA spike-ins allow one to determine if an RNA-seq assay accurately represents the composition of known input and to derive standard calibration curves that relate read counts to RNA concentration in the studied sample [15][21][22]. In addition, fixed controls of known exogenous sequences should be used to measure sequencing error rates, coverage biases, and other variables listed in Clause 6 that affect downstream analysis.

8 Proficiency testing

8.1 General

Proficiency testing, an assessment of the overall competence through inter-laboratory comparisons, shall be used to establish and monitor the quality of laboratory tests. Laboratory proficiency can be monitored through a number of approaches including the use of reference materials. See details at ISO/IEC 17043.