INTERNATIONAL STANDARD

ISO 20397-2

First edition
2021-02

Biotechnology — Massively parallel sequencing —

Part 2: Quality evaluation of sequencing data

Biotechnologie — Séquençage massivement parallèle —

Partie 2: Évaluation de la qualité des données de séquençage

Reference number
ISO 20397-2:2021(E)

© ISO 2021



COPYRIGHT PROTECTED DOCUMENT

© ISO 2021

All rights reserved. Unless otherwise specified, or required in the context of its implementation, no part of this publication may be reproduced or utilized otherwise in any form or by any means, electronic or mechanical, including photocopying, or posting on the internet or an intranet, without prior written permission. Permission can be requested from either ISO at the address below or ISO's member body in the country of the requester.

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 ... **iv**
Introduction ... **v**
1 Scope ... **1**
2 Normative references ... **1**
3 Terms and definitions ... **1**
4 Raw data ... **6**
4.1 General ... 6
4.2 Raw data file ... 6
4.3 Quality assessment of raw data ... 6
4.3.1 General ... 6
4.3.2 Basic statistics ... 7
4.3.3 Quality metrics ... 7
4.4 Raw data pre-processing ... 8
5 Sequence alignment and mapping ... **8**
5.1 General ... 8
5.2 Alignment and mapping file format ... 9
5.3 Quality control of sequencing alignment and mapping ... 9
5.3.1 Basic alignment statistics ... 9
5.3.2 Quality indicators ... 10
5.3.3 Methods for alignment and mapping quality assessment ... 11
5.4 Alignment post-processing ... 11
6 Variant calling ... **11**
6.1 General ... 11
6.2 Data file for variant calling ... 11
6.3 Quality metrics in the variant calling ... 12
6.4 Processing of false positive variants ... 12
6.5 Sequence annotation ... 12
7 Validation ... **12**
7.1 General ... 12
7.2 Validation of quality metrics ... 13
8 Documentation ... **14**
Annex A (informative) **Quality metrics for specific example MPS platforms** ... **15**
Annex B (informative) **Coverage and read recommendations by applications** ... **16**
Annex C (informative) **Software for sequence alignment and mapping** ... **18**
Bibliography ... **19**

© ISO 2021 – All rights reserved

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 276, *Biotechnology*

A list of all parts in the ISO 20397 series can be found on the ISO website.

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.

© ISO 2021 – All rights reserved

Introduction

Massively parallel sequencing (MPS) is a high-throughput analytical approach to nucleic acid sequencing utilizing massively parallel processing, that allows whole genomes, transcriptomes and specific nucleic acid targets from different organisms to be investigated in a relatively short time.

MPS is used in many life science disciplines permitting determination and high throughput analysis of millions and thousands of millions of nucleotide bases. The biological variability of deoxyribonucleic and ribonucleic acid polymers from living organisms results in challenges in accurately determining their sequences. The quality of sequence determination by MPS depends on many factors including but not limited to sample quality, library preparation, platform selection, and sequencing data quality.

The analysis of sequencing data poses significant bioinformatics challenges in various areas such as data storage, computation time and variant detection accuracy. One of the major challenges associated with sequencing data that is sometimes easily overlooked is monitoring quality control metrics over all stages of the data processing pipeline. Knowledge of data quality is essential for downstream analysis of sequences. Quality control for nucleic acid sequencing data handling and analysis can be separated into three stages: raw data, alignment and variant calling. This document provides a list of considerations for quality evaluation of MPS sequencing data, and the specific recommendations for different MPS platforms.

© ISO 2021 – All rights reserved

Biotechnology — Massively parallel sequencing —

Part 2: Quality evaluation of sequencing data

1 Scope

This document specifies general requirements and recommendations for quality assessments and control of massively parallel sequencing (MPS) data. It covers post raw data generation procedures, sequencing alignments, and variant calling.

This document also gives general guidelines for validation and documentation of MPS data.

This document does not apply to any processes related to de novo assembly.

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 terminological databases for use in standardization at the following addresses:

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

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

3.1
adapter sequence
adapter
artificial oligonucleotide of a known sequence that can be added to the 3’ or 5’ ends of a nucleic acid fragment

Note 1 to entry: It provides the primer site as well as other necessary sequences for sequencing the insert.

3.2
algorithm
completely determined finite sequence of instructions by which the values of the output variables may be calculated from the values of the input variables

[SOURCE: IEC 60050-351:2013, 351-42-27, modified — The notes were deleted.]

3.3
base calling
computational process in massively parallel sequencing of translating raw electrical signals to nucleotide sequence

Note 1 to entry: Base calling application and algorithm performance is characteristically defined by read and consensus accuracy.

© ISO 2021 – All rights reserved

3.4
bioinformatics pipeline
individual programs, scripts, or pieces of software linked together, where raw data or output from one program is used as input for the next step in data processing

EXAMPLE The output from a base quality trimming program may be used as input to a de-novo assembler.

3.5
capture efficiency
percent of all sequenced or mapped reads that overlap the targeted regions

3.6
coverage
coverage depth
number of times that a given base position is read in a sequencing run

Note 1 to entry: The number of reads that cover a particular position.

3.7
coverage breadth
fraction of the genome in assembled/target genome size in sequencing runs

3.8
cluster density
number of clusters for each tile

Note 1 to entry: The cluster density applied to the *MPS* (3.30) platforms requires an amplification step.

Note 2 to entry: The density of individual sequence clusters, each arising from a single molecule on some sequencing platforms.

Note 3 to entry: Cluster density is usually expressed in thousands per mm^2.

3.9
CCS
circular consensus sequencing
sequencing mode where the insert size is sequenced multiple times in a rolling circle amplification type reaction, leading to high accuracy

Note 1 to entry: In this mode, multiple passes from the same molecule can be used to achieve higher single molecule accuracy.

3.10
coverage range
range of coverage depth across a genome for sequencing runs

3.11
CNV
copy number variation
copy number variant
variation of the number of copies of one or more sections of the DNA present in the genome of an organism

Note 1 to entry: CNVs are insertions, deletions, inversions and duplications containing at least 1 000 bases in length.

3.12
DNA
deoxyribonucleic acid
polymer of deoxyribonucleotides occurring in a double-stranded (dsDNA) or single-stranded (ssDNA) form

[SOURCE: ISO 22174:2005, 3.1.2]

© ISO 2021 – All rights reserved

3.13
deletion
loss of one (or more) nucleotide base pair(s) from a nucleic acid sequence compared to its reference sequence

3.14
duplication level
number of identical repeats for every sequence in a library

Note 1 to entry: The duplication level is usually displayed in a plot showing the relative number of sequences with different degrees of duplication.

3.15
GC content
percentage of guanine and cytosine in one or more nucleic acid sequence(s)

Note 1 to entry: The amount of guanine and cytosine in a polynucleic acid, is usually expressed in mole fraction (or percentage) of total nitrogenous bases. Total nitrogenous bases comprise the total number of nucleotide bases of reads from one or more MPS run.

3.16
gene
sequence of nucleotides in DNA or RNA encoding either an RNA or a protein product

Note 1 to entry: Genes are recognized as the basic unit of heredity.

Note 2 to entry: A gene can consist of non-contiguous nucleic acid segments that are rearranged through a nuclear processing step.

Note 3 to entry: A gene may include or be part of an operon that includes elements for gene expression.

3.17
indel
insertion (3.18) or /and *deletion* (3.13) of nucleotides in genomic DNA

Note 1 to entry: Indels are less than 1 000 bases in length.

3.18
insertion
addition of one (or more) nucleotide base pair(s) into a nucleic acid sequence

[SOURCE: ISO/TS 20428: 2017, 3.19, modified — DNA was replaced by nucleic acid.]

3.19
sequencing
determining the order and the content of nucleotide bases (adenine, guanine, cytosine, thymine, and uracil) of a nucleic acid molecule

Note 1 to entry: A sequence is generally described from the 5’ to 3’ end.

[SOURCE: ISO/TS 17822-1:2020, 3.19, modified — DNA was deleted in the term; DNA was replaced by nucleic acid, and uracil was added in the definition.]

3.20
sequence alignment
arrangement of nucleic acid sequences according to regions of similarity

Note 1 to entry: Sequence alignment may not require a reference genome /reference targeted nucleic acid region and its aim might not produce an assembly.

© ISO 2021 – All rights reserved

3.21
raw data
primary sequencing data produced by a sequencer without involving any software-based pre-filtering for analysis purpose

3.22
RNA
ribonucleic acid
polymer of ribonucleotides occurring in a double-stranded or single-stranded form

Note 1 to entry: Synthesis of proteins in cells is directed by genetic information carried in the sequence of nucleotides in a class of RNA known as messenger RNA (mRNA).

3.23
ribonucleotide
nucleotide containing ribose as its pentose component forming the basic building blocks for RNA

Note 1 to entry: The ribonucleotides consist of adenylate (AMP), guanylate (GMP), cytidylate (CMP), or uridylate (UMP).

3.24
read
sequence read
nucleotide sequence generated by a sequencing device

Note 1 to entry: A read is a deduced sequence of nucleic acid base pairs (or base pairs probabilities) corresponding to all (or part of) a single nucleic acid fragment. Read can be used to refer to as those sequences obtained from MPS experiments.

3.25
read type
category of sequence that depends on how the sequence reading experiment is designed and conducted

EXAMPLE Read type can be single-end, paired-end, mate-paired end, continuous long read, circular consensus.

3.26
reference sequence
nucleic acid sequence used either to align by mapping sequence reads or as the basis for annotations such as genes and sequence variations

3.27
demultiplexing
computational reverse of multiplexing process, mixing two or more samples together such that they can be sequenced in a single run on an MPS instrument

Note 1 to entry: Samples that are to be combined need to be barcoded/indexed prior to being mixed together.

Note 2 to entry: Demultiplexing is a computational algorithm that separates a pool of reads according to their original sample based on the barcode.

3.28
mapping
assembling nucleic acid sequences against an existing backbone (reference) sequence, in order to build a consensus sequence

3.29
mate pairs
mate pair reads
paired-end read which correspond to the ends of a long nucleic acid sequence fragment obtained by shrinking the sample into large chunks (larger than 2 kb or at least 2 kb)

© ISO 2021 – All rights reserved

3.30
MPS
massively parallel sequencing
sequencing technique based on the determination of incremental template based polymerization of many independent DNA molecules simultaneously

Note 1 to entry: Massively parallel sequencing technology can provide millions or billions of short reads per run.

3.31
paired-end reads
sequencing reads from both ends of a DNA fragment

Note 1 to entry: In paired-end sequencing, the instrument sequences both ends of short inserts typically ranging from 200 bps to 800 bps.

3.32
quality score
***Q* score**
Phred quality score
measure of the sequencing quality of a given nucleotide base

Note 1 to entry: Q is defined by the following formula:

$$Q = -10\log10(p)$$

where p is the estimated probability of the base call being wrong.

Note 2 to entry: A quality score of 20 represents an error rate of 1 in 100, with a corresponding call accuracy of 99 %.

Note 3 to entry: Higher quality scores indicate a smaller probability of error. Lower quality scores can result in a significant portion of the reads being unusable. Low quality scores can also indicate false-positive variant calls, resulting in inaccurate conclusions.

3.33
run
single process cycle of the sequencer from initiation until the raw data is obtained

3.34
sequence annotation
process of adding a note of explanation, comment or reference about specific features in a DNA, RNA or protein sequence with descriptive information about structure or function

Note 1 to entry: The process of sequence annotation can be regarded as assigning metadata to the sequence.

3.35
single-end read
sequence read obtained by reading a DNA fragment from one end to the other

3.36
SNV
single nucleotide variant
variation in a single nucleotide of a nucleic acid molecule

3.37
SV
structural variation
region of DNA approximately 1 000 bases or larger in size which can include inversions and balanced translocations or genomic imbalances

Note 1 to entry: Common types of structural variants include copy number variants (deletions, insertions, amplifications, duplications), copy number neutral deletions (loss of heterozygosity), inversions, segmental duplications, and translocations (balanced or imbalanced).

© ISO 2021 – All rights reserved

3.38
subread
fraction of the read that is present in between hairpin adapters

3.39
trimming of raw reads
procedure aimed at removing low quality portions or sequence contaminations while preserving the longest high-quality part of an MPS read

3.40
variation
differences of one or more nucleic acid bases in a sequence with respect to the expected one(s)

3.41
variant calling
process of accurately identifying the variations from sequence data with respect to a reference sequence

3.42
ZMW
zero mode waveguide
optical waveguide that guides light energy into a volume that is small in all dimensions compared to the wavelength of the light

Note 1 to entry: A polymerase is anchored at the bottom of that ZMW and the incorporation of nucleotides is measured by the increase of fluorescence during binding followed by the subsequent reduction after incorporation.

4 Raw data

4.1 General

Each nucleotide in a sequence should be assigned a numerical value (base quality score) that correlates to the inferred accuracy of the base calling process, if applicable.

4.2 Raw data file

Generation of sequence read files should use instrument-specific software and/or instrument-specific pipelines. Monitored physical parameters such as signal to noise ratio shall be documented. These physical parameters should be monitored of during each sequencing experiment.

Sequence read files should be configured in the appropriate file format, containing the compilation of individual sequence reads, each with its own identifier, and an associated base quality score for each nucleotide.

NOTE FASTQ format (or convertible to FASTQ format) can be used as a *de facto* standard format for downstream analysis of the quality of MPS data sets. FASTQ is widely accepted as a cross platform interchange file format.

The output files generated after a sequencing run, and associated quality metrics should be analysed in the downstream bioinformatics pipeline using appropriate software.

4.3 Quality assessment of raw data

4.3.1 General

Quality control indicators can differ depending on the MPS platform, library preparation method, and intended use of the analysis.

© ISO 2021 – All rights reserved