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 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 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.

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.

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.

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]

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.

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)

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).

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.

Sequence results should be interpreted by competent staff. The interpretation should be performed to meet the quality level fitting the intended purpose of the analysis considering a statistically reliable repeat number of reads.

Read processing tools should be applied with consideration for quality assessment and trimming of raw reads.

4.3.2 Basic statistics

Basic statistics shall be recorded, including but not limited to:

a) type of platform;

b) type of read;

c) library preparation kit;

d) read length;

e) number of reads;

f) overall GC content;

g) total sequence length.

4.3.3 Quality metrics

The quality control metrics for raw data assessment can refer to but are not limited to:

a) sequence length distribution;

b) per sequence GC content;

c) quality score;

   1) per base sequence quality;

   2) per sequence quality score;

      NOTE 1 Low-quality scores can indicate increased false-positive variant calls.

   3) all sequences should be flagged as either ‘warn’ or ‘pass’ for per base sequence quality.

d) per base sequence content;

e) acceptability of signal/noise ratio;

f) sequence duplication levels;

g) overrepresented level;

h) cluster density;

i) transition/transversion ratio for whole-exome or whole-genome sequencing or large amplicons sequencing;

g) adaptor rate/adaptor sequence contamination;

k) contaminants (identification, quantification);

l) error rate;

   NOTE 2 This includes homopolymer errors: errors in the number of bases called when a single nucleotide occurs more than once in consecutive order in a sequence.

m) *k*-mer analysis;

NOTE 3 In computational genomics, *k*-mers refer to all the possible subsequences (of length *k*) from a nucleic acid sequence. Overrepresentation of *k*-mers can be analysed to detect potential genome mis-assembly where repeated DNA sequences have possibly been combined.

n) N fragment;

NOTE 4 Number and/or percentage of ambiguous calls.

o) repeat stretch and repeat sequence;

p) nucleotide distribution across cycles.

4.4 Raw data pre-processing

Raw data pre-processing may include but is not limited to the following computational steps, if applicable:

a) removal/trimming of low-quality sequences/bases;

b) demultiplexing;

c) removal of adapters/primers and contamination;

d) error correction;

e) filtration of duplicated read;

f) trimming of reads to fixed length;

g) calling the CCS reads.

When CCS data are being used, the CCS reads should be obtained and filtered prior to downstream analysis.

5 Sequence alignment and mapping

5.1 General

Sequence alignment and mapping strategy should be chosen based on the application.

EXAMPLE There is spliced mapping for RNA and un-spliced mapping for the mapping strategy of RNA sequencing.

Alignment and mapping software and tools can be used for alignment.

Alignment quality can be assessed visually using proper alignment views, and using the information provided in the alignment file.

Examples of the software for sequence alignment and mapping of different applications are described in Annex C.

Reference genomes/reference targeted nucleic acid regions shall be used for mapping and should be carefully chosen depending on experimental design.

NOTE 1 Considerations include the version of the reference genome/reference targeted nucleic acid region, choice of different strains in one organism, and choice of masked, soft-masked or unmasked genomes.

NOTE 2 Open source sequencing alignment and mapping software is available online.

5.2 Alignment and mapping file format

Alignments are always stored in the following file formats.

a) Sequence alignment format (SAM)[17][24].

NOTE 1 SAM is a TAB-delimited text format consisting of a header section, which is optional, and an alignment section. Each alignment line has 11 mandatory fields for essential alignment information such as mapping position and variable number of optional fields for flexible or aligner specific information.

b) Binary alignment format (BAM)[15][17].

NOTE 2 It is the compressed format analogous to the SAM format in binary form.

c) Compressed reference-oriented alignment map (CRAM)[16].

NOTE 3 CRAM is a sequencing read file format that is space efficient by using reference-based compression of sequence data and offers both lossless and lossy modes of compression.

d) Moving pictures experts group for genomics (MPEG-G)[3][4][5][6][7][8].

NOTE 4 MPEG-G is a genomics representation format based on the concept of a *Genomic Record*, a data structure consisting of either a single sequence read, or a paired sequence read, and its associated sequencing and alignment information; it can contain detailed mapping and alignment data, a single or paired read identifier (read name) and quality values. Genomic Records are aggregated and encoded in structures called Access Units. These structures are units of coded genomic information that can be separately accessed and inspected.

NOTE 5 MPEG-G is specified in the ISO/IEC 23092 series.

The alignment file should contain the information about the location, orientation, and quality of each read in the alignment.

Algorithms and tools can be applied to manipulate alignment files depending on their respective applications.

5.3 Quality control of sequencing alignment and mapping

5.3.1 Basic alignment statistics

5.3.1.1 General

Basic alignment or mapping statistic shall be obtained and recorded.

Basic alignment or mapping statistic can differ according to the experiment design and read type.

5.3.1.2 Mapping statistics for single-end reads

a) Total number of reads refers to the number of reads that are mapped to the reference sequence or genome.

b) Unmapped reads refer to the number of reads that have failed to map to the reference sequence or genome.

c) Mapped reads refer to the number of reads aligned with the reference sequence or genome.

d) Uniquely mapped reads refer to the number of reads aligned exactly once to the reference sequence or genome.

NOTE 1 Mapping uniqueness depends on circumstances. Reads that are uniquely mapped based on one set of mapping parameters can be multi-hit reads with another set of mapping parameters.

e) Multi-hit mapped reads refer to the number of reads aligned > 1 time to the reference sequence or genome.

NOTE 2 The multi-hit depends on mapping circumstances.

5.3.1.3 Mapping statistics for paired-end reads

a) Total number of mate pairs refers to the number of paired-end reads mapped to the reference sequence or genome.

b) Mapped mate pairs refer to the number of paired reads where both mates were mapped.

c) Partially mapped mate pairs refer to the number of paired reads where only one mate in the pair was mapped.

d) Unmapped mate pairs refer to the number of paired reads that failed to map to the reference sequence or genome.

e) Improperly mapped mate pairs refer to the number of paired reads where one of the mates was mapped with an unexpected orientation.

NOTE 1 This is also known as discordantly mapped pairs.

f) Properly mapped mate pairs refer to the total number of paired reads where both mates were mapped with the expected orientation.

NOTE 2 This is also known as concordantly mapped pairs.

5.3.1.4 Mapped subread length

The length of the subread alignment to a target reference sequence does not include the adapter sequence.

5.3.2 Quality indicators

The following quality control parameters can apply depending on applications:

a) alignment rate;

NOTE 1 Poor mapping quality can be a result of non-specific amplification, capture of off-target DNA contamination or other reasons.

b) fragment length, or the length of DNA/RNA that is to be sequenced.

c) insert size statistics for paired-end read only is the length of DNA/RNA intended for sequencing between the adaptors.

NOTE 2 The peak of the insert size distribution is used for quality valuation.

d) duplication level for amplicon-based sequencing only;

e) coverage to the intended purpose including coverage depth, breadth, and range;

NOTE 3 Annex B provides a list of recommended coverage for different applications.

f) AT/GC bias;

NOTE 4 This can be assessed by correlating % GC vs. sequencing depth/coverage.

g) mapping quality score;

h) capture efficiency;

NOTE 5 The capture efficiency is the most important quality control parameter for exome sequencing or other target capture-based sequencing.

i) average or median depth, percentage of the genome covered by the sequencing at that depth;

j) number of discordantly mapped pairs;

k) high quality reads aligned;

l) mismatch rate;

m) consensus accuracy;

NOTE 6 The consensus accuracy is based on aligning multiple sequencing reads and subreads together, optionally with a reference sequence.

n) circular consensus accuracy;

NOTE 7 The circular consensus accuracy is based on multiple sequencing passes around a single circular template molecule. This is used in CCS.

o) subread accuracy.

NOTE 8 The post-mapping accuracy of the base calls.

5.3.3 Methods for alignment and mapping quality assessment

A scoring-based approach should be applied to assess alignment quality.

NOTE The selection of a scoring matrix depends on the application.

5.4 Alignment post-processing

Alignment post-processing can include but is not limited to:

a) local realignment around indels or calculation of per-base base alignment;

b) removal of duplications;

c) recalibration of base quality scores;

d) average read length after trimming by base quality.

6 Variant calling

6.1 General

6.1.1 It is well established that there are four main classes of sequence variants (SNV, indels, CNVs, and SVs). Different computational approaches shall be applied for different classes of sequence variants for sensitive and specific identification.

6.1.2 The range of software tools, and the type of validation required, depends on assay design.

6.2 Data file for variant calling

6.2.1 Called variants shall be annotated using an appropriate specification. The specification shall contain meta-information, a header line, and data lines that each contain information about a position in the genome and genotype information on samples for each position.

EXAMPLE 1 The called variants are annotated using the variant calling format (VCF)[31].

EXAMPLE 2 Alternative specifications exist for representing and storing variant calls:

a) Genomic VCF Conventions;

b) The Sequence Ontology Genome Variation Format Version 1.10;

c) The Human Genome Variation Society, Human Genome Variation Society (HGVS) Simple Version 15.11;

d) Global Alliance for Genomics and Health (GA4GH) File Formats.

6.2.2 Variant files shall include both the specification and the version used.

6.2.3 Variant callers should be configured to output; reference, variants, and no-calls, together with local information at least in the area of targeted regions.

6.3 Quality metrics in the variant calling

Quality control metrics should include but are not limited to (as applicable):

a) thresholds for read coverage depth at the variant's position;

b) quality score of variants;

c) strand bias;

d) allelic read percentages;

e) additional specific metrics relating to the accuracy and sensitivity of variant calling that can include but are not limited to:

   1) total number of variants;

   2) number of false positives;

   3) number of false negatives;

   4) number of allele and genotype mismatches;

   5) transition/transversion ratio;

   6) heterozygous/homozygous (het/hom) ratio;

f) cross-sample contamination analysis.

6.4 Processing of false positive variants

False positive variants should be flagged or filtered from the original variant files on the basis of several sequence alignments and variant calling associated quality control metrics.

6.5 Sequence annotation

Variants can be annotated to determine their biological significance and enable functional prioritization and downstream interpretation.

7 Validation

7.1 General

7.1.1 Laboratories offering MPS-based testing should perform an "in-house" bioinformatics pipeline validation.

7.1.2 Performance requirements for the assay shall be established during the validation procedure, and the same specifications shall be used to monitor the performance of the assay each time a sample is processed.

7.1.3 Specific quality control and quality assurance parameters shall be evaluated during validation and used to determine satisfactory performance.

7.1.4 Each laboratory shall define the criteria and means to monitor all quality metrics to ensure optimal analytical performance. Quality metrics used for monitoring should be documented and periodically verified.

See Annex A for recommended quality metrics and their specific values for some platforms.

7.1.5 Laboratories shall include specific measures to ensure that each data file generated in the bioinformatics pipeline maintains its integrity and provides alerts for or prevents the use of data files that have been altered in an unauthorised or unintended manner.

7.1.6 Supplemental validation is required whenever a significant change is made to any component of the bioinformatics pipeline.

7.2 Validation of quality metrics

7.2.1 Validation of analysis shall be performed based on the clarified and documented purpose of the analysis. The intended purpose of the measurement shall be determined and documented.

7.2.2 Laboratories shall establish acceptable raw base call quality score thresholds for the assay during validation.

7.2.3 Pre-processing methods to remove low-quality base calls should be established to reduce the false-positive rate.

7.2.4 The extent of GC bias in all parts of the genome included in the assay should be determined during validation.

7.2.5 Parameters for mapping quality shall be established in a validation plan and should demonstrate that the test only assesses reads that map to the regions targeted by the assay. Steps should be established to filter reads that map to non-targeted regions if applicable.

7.2.6 Coverage shall be defined to achieve adequate sensitivity and specificity in the regions of interest.

7.2.7 Each laboratory shall establish the minimum criteria for the depth of coverage characteristic of a particular region under standard assay conditions depending on the aim of the sequencing run. For a homogeneous sample, the sequence needs to be confirmed; a lower depth is acceptable. In a variant calling process over a region, or a rare sequence in a mixed sample of 1 %, a deep sequencing is needed.

7.2.8 The required level of coverage across the targeted regions shall be defined during the validation stage (coverage range). A recommended range for different applications is described in Annex B.

7.2.9 Acceptable parameters for a maximum duplication rate should be established for each assay.

7.2.10 Filtering of duplicate reads by the analysis pipeline should be established to increase the number of useable sequencing data and prevent skewing of allelic fractions.

7.2.11 Each laboratory shall define the tolerance level for strand bias and outline specific criteria for when alternate testing should be instituted.

7.2.12 Quality metrics can be validated with the assistance of relevant reference standards which have been well characterized, and which have reliable reference sequences to enable accurate alignment, variant calling, etc.

7.2.13 Sanger sequencing to validate the most important binding region is recommended.

8 Documentation

8.1 Laboratories shall document all algorithms, software, and databases used in the analysis, interpretation, and reporting of MPS results. The version of each of these components in the overall bioinformatics pipeline shall be recorded and traceable for each result.

8.2 Laboratories shall document any customisations that vary from the default configuration or should indicate which parameters were customised.

8.3 The reference sequence version number and details should be identified if applicable.

8.4 Laboratories should also document quality control parameters for optimal performance.

EXAMPLE In the primary step, a laboratory would determine acceptable criteria such as the number of reads passing instrument-specified quality filters.

8.5 Laboratories should document the bioinformatics processes that are used for reducing a large variant data set to a list of causal and/or candidate genes and/or variants.

8.6 Evidence of compliance for the defined requirements should be documented.

Annex A
(informative)

Quality metrics for specific example MPS platforms

The following MPS platforms are commonly used for nucleic acid sequencing. Examples of the quality metrics used for quality assessments are presented in Table A.1.

NOTE Human whole genome sequencing is used as an example to provide specific values for each quality metrics.

Table A.1 — Quality metrics for specific MPS platforms

Platform name	Raw file format	Read length	Quality score (H/L)	GC content	Duplication rate	Cluster density	Adaptor rate
illumina®[a] HiSeq 4000	fastq.gz	50 bp to 200 bp	> Q30	39 % to 42 %	< 10 %	5 billion	< 3 %
Thermo Fisher Proton™ [b]	DAT	50 bp to 200 bp	> Q20	39 % to 42 %	NA	60 million to 80 million	< 3 %
BGI[c] /MGI MGISEQ-2000	fastq.gz	50 bp to 200 bp	> Q30	39 % to 42 %	< 5 %	1,5 billion	< 3 %
Oxford Nanopore PromethION®[d]	FAST5	10 kbp to 300 kbp	> Q20	39 % to 42 %	NA	2 560 channels[f]	< 3 %
PacBio® Sequel II[e]	bam	10 kbp to 100 kbp	> Q20	39 % to 42 %	NA	8 million ZMWs[g]	< 3 %

[a] illumina® is a trademark of illumina, Inc biotechnology company. This information is given for the convenience of users of this document and does not constitute an endorsement by ISO of the product name.

[b] Thermo Fisher Proton™ is a trademark of Thermo Fisher Scientific biotechnology company. This information is given for the convenience of users of this document and does not constitute an endorsement by ISO of the product name.

[c] MGI is a tradename of BGI genome sequencing company. This information is given for the convenience of users of this document and does not constitute an endorsement by ISO of the product name.

[d] Oxford Nanopore PromethION® is a trademark of Oxford Nanopore Technologies Limited. This information is given for the convenience of users of this document and does not constitute an endorsement by ISO of the product name.

[e] PacBio Sequel II® is a trademark of Pacific Biosciences biotechnology company. This information is given for the convenience of users of this document and does not constitute an endorsement by ISO of the product name.

[f] Oxford Nanopore is measured by channels.

[g] Pacific Biosciences is measured in ZMWs (Zero-mode waveguide).

Annex B
(informative)

Coverage and read recommendations by applications

Table B.1 presents examples of coverage and read levels by a variety of different sequencing applications.

Table B.1 — Coverage and read recommendation by applications

MPS type	Applications	Recommended coverage (E)	Recommended reads
Whole genome sequencing	Homozygous single nucleotide variants (SNVs) – single nucleotide changes in genes where the alleles are identical.	15 × [a]	—
	Heterozygous SNVs – single nucleotide changes in genes where the alleles are different from each other.	33 ×	—
	Insertion/deletion mutations (INDELS) – mutations in the genome where nucleotides are inserted or removed.	60 ×	—
	Copy number variation (CNV) – variance in the number of copies of a gene between individuals.	1 × to 8 ×	—
Whole exome sequencing	Homozygous SNVs	100 × (3 × local read coverage)[b]	—
	Heterozygous SNVs	100 × (13 × local read coverage)[c]	—
Targeted sequencing	INDELs	Not recommended	—
	SNVs/SVs in targeted regions	1 000 times to 10 000 times	—
RNA sequencing - Transcriptome sequencing	16S rRNA gene[23][24]	—	Minimum 100 per sample
	Differential expression profiling – quantitative measurement of gene expression across multiple genes to examine different levels of expression in the sample.	—	10 million to 25 million
	Alternative splicing – identification of different splice variants from mRNA transcripts.	—	50 million to 100 million (for short read platforms) 2 million to 3 million (for long read platforms)
	Allele specific expression – transcript expression which is affected by a specific gene allele.	—	50 million to 100 million

NOTE 1 Results can be validated by complementary proteomics experiments.

NOTE 2 The recommended coverages are for human samples.

[a] 15 × indicates local identical coverage, it is not the overall average coverage. The number here is just an example.

[b] 100 × is overall average coverage for whole exome sequencing. The 3x local read coverage indicates the local coverage to detect SNPVs. The numbers here are just example.

[c] 100 × is the overall average coverage for whole exome sequencing. The 15 × local read coverage indicates the local coverage to detect SNPVs. The numbers here are just examples.