TECHNICAL SPECIFICATION

ISO/TS 22692

First edition
2020-10

Genomics informatic — Quality control metrics for DNA sequencing

Informatique génomique — Mesures de contrôle de la qualité pour le séquençage de l'ADN

Reference number
ISO/TS 22692:2020(E)

© ISO 2020

COPYRIGHT PROTECTED DOCUMENT

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

The rapid progress in Next Generation Sequencing (NGS) technology has drastically reduced the cost and time for genomic analysis. A number of research institutions, corporations, and government agencies are competitively collecting a large volume of genomic data through multi-national, multi-institutional projects such as "DiscovEHR"[9], "gnomAD"[10] and "UK Biobank"[11]. The demand for sharing of "high quality" genomic data is growing because large-scale reference data is required for reliable detection of mutation for both industrial and clinical applications.

However, the quality of available genomic data is less than desirable. To establish consistent quality control metrics, details of each stage of NGS process need to be recorded, shared and standardized (processes and data elements collected and coded for each stage and sub-stage). These processes include sample preparation, library preparation, sequencing, and data processing, among others, as shown in Figure 1.

Sample Preparation	Library Preparation	Sequencing	Data Processing
• Sample sequencing type	• DNA extraction method	• Sequencing information	• Data quality measurement
• Sample information	• DNA quality	• Running quality information	• Sequencing alignment
	• Library construction		• Variant calling
			• Variant filtering and annotation
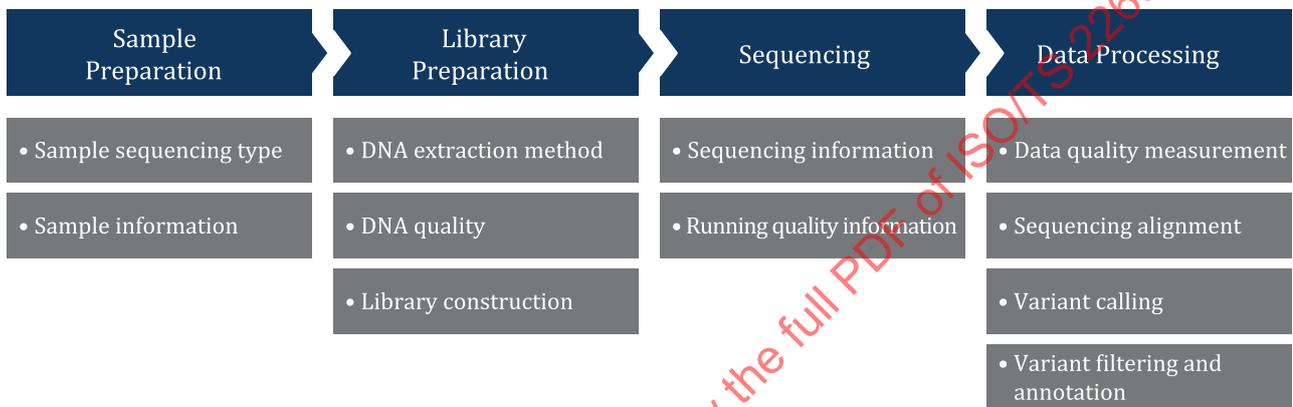
Figure 1 — NGS process

Genomics informatic — Quality control metrics for DNA sequencing

1 Scope

This document identifies quality metrics for the detection of DNA variants using next generation sequencing (NGS) technology. It also defines the data types, relationships, optionality, cardinalities and terminology bindings of the data.

This document provides a basis for sharing and for the application of "high quality" genomic data and contributes to the realization of the precision medicine and the development of relevant industries.

This document is intended to serve as a catalogue of sequencing data elements used to address quality metrics for various clinical, industrial and commercial applications. The exchange of these data allows researchers, commercial entities, and regulatory bodies to assess for the purpose of selective utilization of the data by setting application-specific quality criteria

This document is not intended for

— sequencing methods other than NGS, such as the Sanger sequencing,

— targets other than genome, such as transcriptome or proteome, or

— specimens of species other than humans.

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
copy number variation
CNV
variation (3.18) in the number of copies of one or more sections of the *DNA* (3.3)

[SOURCE: ISO/TS 20428:2017, 3.7]

3.2
deletion
contiguous removal of one or more bases from a genomic sequence

[SOURCE: ISO/IEC 23092-2:2019, 3.4]

3.3
DNA
deoxyribonucleic acid
molecule that exists in nuclei and in mitochondria of human cells and is composed of a linear array of 4 bases (Adenine: A, Thymine: T, Guanine: G and Cytosine: C)

[SOURCE: ISO 18074:2015, 4.1, modified — Note 1 to entry deleted.]

3.4
DNA sequencing
determining the order of nucleotide bases (adenine, guanine, cytosine and thymine) in a molecule of *DNA* (3.3)

Note 1 to entry: Sequence is generally described from the 5' end.

[SOURCE: ISO/TS 17822-1:2014, 3.20]

3.5
exome
part of the genome formed by exons

[SOURCE: ISO/TS 20428:2017, 3.13]

3.6
FASTA
genomic information representation that includes a name and a nucleotide sequence for each *sequence read* (3.17)

[SOURCE: ISO/IEC 23092-2:2019, 3.7, modified]

3.7
FASTQ
genomic information representation that includes *FASTA* (3.6) and quality values

[SOURCE: ISO/IEC 23092-2:2019, 3.8]

3.8
gene
basic unit of hereditary information composed of chains of nucleotide base pairs in specific sequences that encodes a protein or protein subunit

[SOURCE: ISO 11238:2018, 3.29]

3.9
germline
series of germ cells each descended or developed from earlier cells in the series, regarded as continuing through successive generations of an organism

[SOURCE: ISO/TS 20428:2017, 3.17]

3.10
indel
insertion (3.11) or/and *deletion* (3.2)

[SOURCE: ISO/TS 20428:2017, 3.18]

3.11
insertion
contiguous addition of one or more bases into a genomic sequence

[SOURCE: ISO/IEC 23092-2:2019, 3.18]

3.12
large indel
insertion ([3.11](#)) or *deletion* ([3.2](#)) up to around 1 kb

[SOURCE: ISO/TS 20428:2017, 3.21]

3.13
nucleotide
monomer of a nucleic acid polymer such as *DNA* ([3.3](#)) or RNA

Note 1 to entry: Nucleotides are denoted as letters ('A' for adenine; 'C' for cytosine; 'G' for guanine; 'T' for thymine which only occurs in DNA; and 'U' for uracil which only occurs in RNA). The chemical formula for a specific DNA or RNA molecule is given by the sequence of its nucleotides, which can be represented as a string over the alphabet ('A', 'C', 'G', 'T') in the case of DNA, and a string over the alphabet ('A', 'C', 'G', 'U') in the case of RNA. Bases with unknown molecular composition are denoted with 'N'.

[SOURCE: ISO/IEC 23092-2:2019, 3.20]

3.14
polymerase chain reaction
PCR
in vitro enzymatic technique to increase the number of copies of a specific DNA fragment by several orders of magnitude

[SOURCE: ISO 16577:2016, 3.148]

3.15
quality score
Phred quality score
Q score
quality measure used to assess the accuracy of a sequencing reaction

Note 1 to entry: This quality measure indicates the probability that a given base is called incorrectly by the sequencer. Phred scores are on a logarithmic scale. Therefore, if Phred assigns a Q score of 30 (Q30) to a base, this is equivalent to the probability of an incorrect base call 1 in 1 000 times. A lower base call accuracy of 99 % (Q20) will have an incorrect base call probability of 1 in 100, meaning that every 100 base pairs sequencing read will likely contain an error.

[SOURCE: ISO 21286:2019, 3.4]

3.16
reference sequence
nucleic acid sequence with biological relevance

Note 1 to entry: Each reference sequence is indexed by a one-dimensional integer coordinate system whereby each integer within range identifies a single nucleotide. Coordinate values can only be equal to or larger than zero. The coordinate system in the context of this standard is zero-based (i.e. the first nucleotide has coordinate 0 and it is said to be at position 0) and linearly increasing within the string from left to right.

[SOURCE: ISO/IEC 23092-1:2019, 3.22]

3.17
sequence read
read
fragmented nucleotide sequences that are used to reconstruct the original sequence for next generation sequencing technologies

[SOURCE: ISO/TS 20428:2017, 3.26]

3.18
sequence variation
DNA sequence variation
variation
differences of DNA sequence among individuals in a population

Note 1 to entry: Variation implies *copy number variation* ([3.1](#)), *deletion* ([3.2](#)), *insertion* ([3.11](#)), *indel* ([3.10](#)), *small indel* ([3.20](#)), *large indel* ([3.12](#)), or *single nucleotide variant* ([3.19](#)).

[SOURCE: ISO/TS 20428:2017, 3.30]

3.19
single nucleotide variant
SNV
DNA sequence variation ([3.18](#)) that occurs when a single nucleotide, A, T, C, or G, in the genome (or other target sequence) differs between templates

[SOURCE: ISO 20395:2019, 3.35]

3.20
small indel
insertion ([3.11](#)) or *deletion* ([3.2](#)) of 2 nucleotides to 100 nucleotides

[SOURCE: ISO/TS 20428:2017, 3.32]

3.21
specimen
biospecimen
biological specimen
sample of tissue, body fluid, food, or other substance that is collected or acquired to support the assessment, diagnosis, treatment, mitigation or prevention of a disease, disorder or abnormal physical state, or its symptoms

[SOURCE: ISO/TS 20428:2017, 3.34]

3.22
targeted sequencing
disease-targeted gene panel
technique used for sequencing only selected/targeted genomic regions of interest from a DNA sample

Note 1 to entry: For further details, see Reference [[12](#)].

3.23
whole exome sequencing
WES
technique for sequencing the *exomes* ([3.5](#)) of the protein-coding *genes* ([3.8](#)) in a genome

[SOURCE: ISO/TS 20428:2017, 3.38]

3.24
whole genome sequencing
WGS
technique that determines the complete DNA sequence of an organism's genome at a single time

[SOURCE: ISO/TS 20428:2017, 3.39]

4   Abbreviated terms

BAM   Binary Alignment/Map

BED   Browser Extensible Data

FMA   Foundational Model of Anatomy

HGNC   HUGO Gene Nomenclature Committee

HUGO   Human Genome Organization

NGS   Next Generation Sequencing

RefSeq   NCBI Reference Sequences

SAM   Sequence Alignment/Map

SPREC   Standard PREanalytical Code

VCF   Variant Call Format

5   Quality control metrics for sample preparation

5.1   General

As the first step in NGS workflow, sample preparation is the process to assess and review the submitted sample before being accepted for sequencing. The amount of information in sequencing data is quite different according to sequencing types (whole genome sequencing, whole exome sequencing, and targeted sequencing) and specimen types (blood, surgical tissue, biopsy, etc.), so quality control metrics for sample preparation would be used to select appropriate comparison data in analysis process.

5.2   Sample sequencing type

5.2.1   Sequencing type

Sequencing types are three levels of analysis via NGS: WGS, WES, and targeted sequencing[12]. WGS covers all regions of the genome: both coding and noncoding regions. WES covers all coding regions, which is estimated to comprise 1% to 2 % of the genome, yet contains around 85 % of recognized disease-related mutations[13]. Targeted sequencing interrogates known disease-associated genes, which is focusing on a limited set of genes to allow greater depth of coverage.

5.2.2   Target gene

Sets of genes or gene regions for targeted sequencing. Target gene is represented in accordance with HGNC and RefSeq.

EXAMPLE   HGNC approved symbol: BRCA1, HGNC ID: HGNC:1100, RefSeq: NM_007294.

NOTE   Target gene is used only for WGS.

5.3   Sample information

5.3.1   Specimen type

Type of specimen (e.g. whole blood, cell, urine, fresh cell & tissue) with related data during sample collection (e.g. biopsy, surgical excision, EDTA, heparin), processing (e.g. formalin, centrifugation),

and storage (e.g. paraffin block, cryotube). Specimen type is represented in accordance with SPREC of International Society for Biological and Environmental Repositories[14].

EXAMPLE   BLD (Blood), CEN (Fresh cells from nonblood specimen type), URN (Urine).

5.3.2   Sampling date

Sampling date is the date when the specimen was acquired. It is represented in accordance with the ISO 8601 series.

EXAMPLE   2020-01-01 (ISO 8601 (all parts) basic format: YYYY-MM-DD).

5.3.3   Specimen origin

Specimen origin refers to the anatomical site from which the specimen was acquired. The anatomical site is represented by SNOMED CT®[1)], FMA or other well-known ontologies.

EXAMPLE   SNOMED CT® fully specified name: Lung structure (body structure), concept identifier: 39607008; FMA preferred name: Lung, FMAID: 7195.

5.4   Summary of sample preparation related metrics

The summary of quality control metrics for sample preparation shall be in accordance with Table 1. Examples for the metrics are provided in Annex A.

Table 1 — Quality control metrics for sample preparation

Category	Metrics	Value representation	Optionality
Sample sequencing type	Sequencing type	—	Mandatory
	Target gene	HGNC, RefSeq	Mandatory [a]
Sample information	Specimen type	SPREC	Mandatory
	Sampling date	ISO 8601 (all parts)	Mandatory
	Specimen origin	SNOMED CT®, FMA, other ontologies	Mandatory [a]
[a] There might be cases where there is no relevant data.			

6   Quality control metrics for library preparation

6.1   General

Library preparation is the process of generating DNA fragments to meet the input requirement of sequencing platforms. In general, this workflow involves DNA extraction, DNA quality measurement and library construction. Depending on the purposes of sequencing, library construction methods can vary and sequencing results will be affected. Therefore, for quality control of the data, the conditions used in each library preparation step shall be specified.

6.2   DNA extraction method

6.2.1   DNA extraction kit

Product, product's version and manufacturer's name of the DNA extraction kit.

---

1)   SNOMED CT® is the registered trademark of the International Health Terminology Standards Development Organisation (IHTSDO). This information is given for the convenience of users of this document and does not constitute an endorsement by ISO.

6.3 DNA quality

6.3.1 General

Information about whether DNA quality assessment was performed and the results.

6.3.2 DNA purity

The ratio of optical density (OD 260/280, OD 260/230) is used as an indicator of the sample purity.

6.3.3 DNA integrity

The absence of both single or double strand breaks and absence of nucleotide modifications in the DNA[15], which could be represented by the size and distribution of DNA fragments.

EXAMPLE   DNA median size: 9269 bp; DIN: 9.7.

NOTE   Agilent®[2)] 2200 TapeStation®[3)] provides a numerical assessment of gDNA integrity referred as the DNA Integrity Number (DIN)[16].

6.4 Library construction

6.4.1 Library input amount

Initial input DNA amount for library construction.

6.4.2 Library insert size

Library insert size used to construct the library, which can be confirmed by quality control at the desired size.

6.4.3 Library construction kit

Product, product's version and manufacturer's name of the library construction kit.

6.5 Summary of library preparation related metrics

The summary of quality control metrics for library preparation shall be in accordance with Table 2. Examples for the metrics are provided in Annex A.

Table 2 — Quality control metrics for library preparation

Category	Metrics	Value	Performance criteria	Optionality
DNA extraction method	DNA extraction kit	Text	—	Mandatory [a]
DNA quality	DNA purity	Numeric	Min, Max	Mandatory [a]
	DNA integrity	Numeric	Min	Mandatory [a]
Library construction	Library input amount	Numeric	Min	Mandatory [a]
	Library insert size	Numeric	Min, Max	Mandatory [a]
	Library construction kit	Text	—	Mandatory [a]
[a] There might be cases where there is no relevant data.				

---

2)   Agilent® is the registered trademark of Agilent Technologies, Inc. This information is given for the convenience of users of this document and does not constitute an endorsement by ISO.

3)   TapeStation® is the registered trademark of Agilent Technologies, Inc. This information is given for the convenience of users of this document and does not constitute an endorsement by ISO.

7 Quality control metrics for sequencing

7.1 General

Sequencing is the process to run NGS platform with pre-processed DNA fragments. Currently available sequencing platforms have different chemistries for sequencing that include sequencing by synthesis (Illumina®[4] NGS platforms) and semi-conductor-based sequencing (ThermoFisher's Ion systems), as well as different detection methods. Despite the similarities in technical performance, differences between platforms exist, notably different DNA input requirements, different cost of reagents, run time, read length, and cost per sample. These differences have implications to the instruments' capacity to handle low-quality samples, capability of detecting insertions/deletions, and sample throughput[17]. Targeted sequencing technology is used for testing in a part of genomic region, specified with BED format.

7.2 Sequencing information

7.2.1 Sequencing instrument

Sequencing device for NGS.

EXAMPLE   Illumina® HiSeq®[5] 2500, NovaSeq™[6], NextSeq®[7] 550; THERMO SCIENTIFIC®[8] Ion PGM™[9].

7.2.2 Read length

Read length in bp (base pair).

7.2.3 Sequencing direction

Single-end sequencing (reading in one direction in sequencing) or paired-end sequencing (reading in both directions in sequencing).

7.2.4 Running mode

Sequencing instrument's options for faster run time or high-throughput run mode.

EXAMPLE   Rapid Run Mode, High Output Run Mode in Illumina® platform.

---

4)   Illumina® is the registered trademark of Illumina, Inc. This information is given for the convenience of users of this document and does not constitute an endorsement by ISO.

5)   HiSeq® is the registered trademark of Illumina, Inc. This information is given for the convenience of users of this document and does not constitute an endorsement by ISO.

6)   NovaSeq™ is the trademark of Illumina, Inc. This information is given for the convenience of users of this document and does not constitute an endorsement by ISO.

7)   NextSeq® is the registered trademark of Illumina, Inc. This information is given for the convenience of users of this document and does not constitute an endorsement by ISO.

8)   THERMO SCIENTIFIC® is the registered trademark of Thermo Fisher Scientific Inc. This information is given for the convenience of users of this document and does not constitute an endorsement by ISO.

9)   Ion PGM™ is the trademark of Thermo Fisher Scientific Inc. This information is given for the convenience of users of this document and does not constitute an endorsement by ISO.

7.3   Running quality information

7.3.1   Error rate

A measure of misread nucleotide proportion of monitoring material, whose sequences are provided by the manufacturer (e.g. phiX in Illumina® platform).

NOTE   Illumina® phiX library is derived from the small, well-characterized genome.

7.3.2   Percent data quality >Q30

The Q30 rate is the proportion of Phred quality score over 30 for each base in the sequencing read of the sequencing instrument. A Phred quality score is a measure of the quality of the identification of the nucleobases generated by automated DNA sequencing[18][19]. If Phred assigns a quality score of 30 to a base, the chances that this base is called incorrectly are 1 in 1 000.

7.4   Summary of sequencing related metrics

The summary of quality control metrics for sequencing shall be in accordance with Table 3. Examples for the metrics are provided in Annex A.

Table 3 — Quality control metrics for sequencing

Category	Metrics	Value	Performance criteria	Optionality
Sequencing information	Sequencing instrument	Text	—	Mandatory
	Read length	Numeric	—	Mandatory a
	Sequencing direction	Text	—	Mandatory
	Running mode	Text	—	Optional
Running quality information	Error rate	Numeric	Max	Optional
	Percent data quality >Q30	Numeric	Min	Mandatory a
a There might be cases where there is no relevant data.				

8   Quality control metrics for data processing

8.1   General

NGS data processing is the analysis of data derived from sequencing platforms and is entirely dependent on computing resources. This workflow consists of various processes, such as read alignment, deduplication, variant calling etc. There are many tools available for the purpose of sequencing. Sometimes, in-house tools are generated and used as needed. It is necessary to provide several bioinformatics quality control metrics for the final data to be used for detection of variation.

8.2   Data quality measurement

8.2.1   Total reads

The number of sequencing reads in the FASTQ file.

8.2.2   Mean coverage

The mean number of sequencing depths aligned at each base of the target area. With hybridization-based capture and whole-genome sequencing, PCR duplicates should be removed from coverage calculations[20].

8.2.3   Uniformity

The proportion of base, among the target area, whose number of sequencing read is over x % of the mean coverage and greater than 20 %.

8.2.4   Duplication rate

Proportion of reads that were aligned at the exact same position among the total sequencing reads of the BAM file.

8.2.5   On-target rate

The proportion of reads being on the target site relative to total reads.

8.2.6   Q30 rate

The Q30 rate is the proportion of Phred quality score over 30 for each base in the sequencing read of the alignment file in SAM, BAM, or CRAM format. A Phred quality score is a measure of the quality of the identification of the nucleobases generated by automated DNA sequencing[18][19]. If Phred assigns a quality score of 30 to a base, the chances that this base is called incorrectly are 1 in 1 000.

8.3   Sequencing alignment

8.3.1   Mapping algorithm

Mapping algorithm used in read mapping step.

EXAMPLE   BWA (v0.7.17), Bowtie2 (v2.4.1).

8.3.2   Local realignment software and version

Software and version used in local realignment.

EXAMPLE   GATK (v4.1.2.0).

8.4   Variant calling

8.4.1   Variant calling software and version

Software and version used in variant calling.

EXAMPLE   MuTect (v1.1.4), LoFreq (v0.6.1), Strelka (v2.9.10).

8.4.2   Variant call quality score

Variant call quality score is an estimate of how confident the caller correctly identified that a given genome position displays variation in at least one sample[21]. It is generated during the variant calling step and it is a requisite component of the VCF, which specifies the format of a text file used in bioinformatics for storing gene sequence variations.

8.4.3   Allelic read percentage & ratio

Allelic read percentages, including percentage of different variant types (e.g., heterozygous calls, indels, nonsense variants), and portion and ratios of base substitutions.

8.5 Variant filtering and annotation

8.5.1 General

Variant filtering is the process by which variants representing false-positive artifacts of the NGS method are flagged or filtered from the original VCF file on the basis of several sequence alignment and variant calling associated metadata[20].

8.5.2 Germline filter criteria

Criteria used to distinguish germline mutation from detected mutations.

8.5.3 Mutation and annotation database

The database is used for identifying clinically relevant mutation and includes information such as minor allele frequency in the population, predictions of variant effects on protein function or splicing[22][23].

EXAMPLE   The Human Gene Mutation Database (HGMD), ClinVar and other disease- or locus-specific mutation databases.

8.6 Summary of data processing related metrics

The summary of quality control metrics for data processing shall be in accordance with Table 4. Examples for the metrics are provided in Annex A.

Table 4 — Quality control metrics for data processing

Category	Metrics	Value	Performance criteria	Optionality
Data quality measurement	Total reads	Numeric	Min	Mandatory a
	Mean coverage	Numeric	Min	Mandatory a
	Uniformity	Numeric	Min	Mandatory a
	Duplication rate	Numeric	Max	Optional
	On-target rate	Numeric	Min	Mandatory a
	Q30 rate	Numeric	Min	Mandatory a
Sequence alignment	Mapping algorithm	Text	—	Mandatory a
	Local realignment software and version	Text	—	Mandatory a
Variant calling	Variant calling software and version	Text	—	Mandatory a
	Variant call quality score	Numeric	Min	Optional
	Allelic read percentage & ratio	Numeric	Min	Optional
Variant filtering and annotation	Germline filter criteria	Text	—	Mandatory a
	Mutation and annotation database	Text	—	Mandatory a
a There might be cases where there is no relevant data.				