ICS 11. 100.10
C44

National Standard of the People's Republic of China

GB/T XXXX—XXXX

The Data Quality Evaluation Method of Human Whole Genome Sequencing

人全基因组高通量测序数据质量评价方法

(English Translation)

Issue date: XXXX-XX-XX Implementation date: XXXX-XX-XX

Issued by State Administration for Market Regulation
Standardization Administration of the People’s Republic of China

Foreword

SAC/TC 136 is in charge of this English translation. In case of any doubt about the contents of English translation, the Chinese original shall be considered authoritative.

This document is drafted in accordance with the rules given in GB/T 1.1—2020 *Directives for standardization - Part 1: Rules for the structure and drafting of standardizing documents.*

Attention is drawn to the possibility that some of the contents of this standard may be the subject of patent rights. The issuing body of this document shall not be held responsible for identifying any or all such patent rights.

This document was proposed by National Medical Products Administration.

This document was prepared by SAC/TC 136.

The Data Quality Evaluation Method of Human Whole Genome Sequencing

1. Scope

This document specifies the terminology and definitions, quality requirements, and evaluation methods involved in the quality assessment of human whole-genome high-throughput sequencing data.

This document is applicable to the data quality evaluation of whole genome sequencing of human genome DNA samples using high-throughput sequencing technology.

This document is not applicable to Sanger sequencing and single molecule sequencing, human de novo sequencing, human haplotype sequencing, human tumor tissue sequencing, nor does it is applicable to sequencing of animals, plants, viruses, bacteria, parasites, and other organisms present in human samples.

2. Normative references

The following documents are referred to in the text in such a way that some or all of their content constitutes requirements of this document. For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments) applies.

GB/T 29859-2013, *Bioinformatics terms*

GB/T 30989-2014, *Technical regulation for high-throughput gene sequencing*

GB/T 35537-2017, *Requirements of the high-throughput gene sequencing result evaluation*

GB/T 35890-2018, *Technical specification of high throughput sequencing data format*

YY/T 1723-2020, *High-throughput gene sequencer*

GA/T 1693-2020,*Forensic sciences—Specifications for second generation sequencing-based DNA examination*

3. Terms and definitions

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

3.1 High-throughput sequencing or massively parallel sequencing

A sequencing technology characterized by the ability to parallel sequence tens of millions to billions of nucleic acid molecule sequences at once and generally shorter read lengths.

[Amend GB/T 35890-2018, definition 3.1.]

3.2 Human whole genome sequencing

A method for analysis the entire genome sequence of an individual or a population.

Note: Includes the nucleic acid sequences of 23 pairs of human chromosomes, mitochondrial nucleic acid sequences.

3.3 Adapter

Artificially synthesized DNA fragment with known sequences used for labeling and anchoring the sequences to be sequenced.

[Amend GA/T 1693-2020, definition 3.9.]

3.4 PCR-free library

A sequencing library directly constructed without relying on the PCR amplification process.

Note: Polymerase chain reaction (PCR) is an in vitro enzyme reaction technique that exponentially amplifies specific DNA fragments through DNA polymerase reaction to increase their copy number by several orders of magnitude.

3.5 PCR library

In contrast to 3.4, a sequencing library constructed after a certain number of PCR amplification cycles.

3.6 Barcode or index

A characteristic segment of deoxyribonucleotides that serves as a unique identifier for recognizing the specific source of a sample during multiplexed sequencing.

[Amend GA/T 1693-2020, definition 3.10.]

3.7 Percentage of base call quality

The percentage of bases in sequencing data with a base-calling quality score above a specified threshold, typically denoted as Q20, Q30, and so on.

3.8 GC content

The percentage of the sum of guanine and cytosine in the sequencing fragment base to the total number of purine bases (adenine and guanine) and pyrimidine bases (thymine and cytosine).

3.9 Data filtering

The process of removing low-quality, N bases, adapter contamination, and any other reads that not meet the requirements for downstream analysis from the original sequencing reads.

3.10 Sequencing raw base

The total number of original sequencing bases without data filtering, commonly referred to as raw data.

3.11 Human reference genome sequence

The available publicly released reference genome sequence of human, such as hs37d5, hg38, hg19, etc.

3.12 Mapping reads rate

The percentage of reads that match the human reference genome sequence out of the effective sequencing reads.

3.13 Effective sequencing depth

The average sequencing depth of a whole genome sequence obtained after data filtering and duplication reads masking.

Note: The calculation formula is as follows:

Effective sequencing depth= total number of base pairs (bp) matched the reference genome sequence without N-base regions/total size of the reference genome sequence without N-base regions

The unit of measurement is ×, that is, the multiplier.

3.14 Base mismatch rate

The percentage of bases that are inconsistent with the reference genome sequence divided by the total number of bases aligned to the reference genome sequence.

3.15 Duplicate rate

The percentage of the total number of reads aligned to the reference sequence in terms of position, direction, and base sequence, divided by the total number of reads aligned to the reference sequence.

3.16 Coverage rate of sequencing at least 20×

After sequence alignment, the percentage of non-N bases in the reference genome that are covered by at least 20 times of the sequencing reads in the total number of non-N bases.

3.17 Insertion and deletion, Indel

Gene mutations in genomic DNA with nucleotide insertion/deletion fragments of length equal to or less than 50bp.

3.18 Structural variation， SV

The large segment deletions, insertions, duplications, inversions, and translocations, with segment lengths greater than 50bp.

3.19 Standard variation dataset

A high-confidence standard variation dataset constructed based one human genome reference material.

3.20 Precision

The percentage of detected true genetic variation sites to all detected genetic variation sites.

Note: The calculation formula can be found in formula (1)

$$\text{Precision}=TP/(TP+FP)\times 100\% \text{---}(1)$$

In the formula:

TP - True Positive, refers to the number of detection gene mutation sites that are consistent with the standard mutation dataset gene mutation sites.

FP - False Positive, refers to the number of detected gene mutation sites that are inconsistent with the standard mutation dataset.

3.21 Sensitivity

Also known as recall, it refers to the percentage of detected true gene variant sites out of all gene variant sites to be detected.

Note : The calculation formula can be found in formula (2)

$$\text{Sensitivity}=TP/(TP+FN)\ \times 100\% \text{ ---}(2)$$

In the formula:

FN-False Negatives, refers to the number of gene variant sites included in the standard variation dataset but not detected.

4. Quality requirements

4.1 Sample requirements

4.1.1 Sample type

The sequencing sample type is human genome DNA. The main sources of DNA are human whole blood, blood leukocyte layer, saliva, amniotic fluid, oral swabs, normal tissues, and normal cell lines.

This document uses human genome standard DNA material extracted from normal human cell lines as the standard reference to conduct a more comprehensive quality evaluation and unify the quantitative evaluation method indicators.

NOTE The information on human genome standard DNA involved in this document is shown in Appendix A.

4.1.2 DNA Sample quality

4.1.2.1 Integrity of DNA sample

Genomic DNA is required to be intact, with no significant degradation. Usually, 1-2% agarose gel electrophoresis is used, where genomic bands should be concentrated at 23kb and above, without significant smearing or diffusion.

4.1.2.2 Purity of DNA sample

The DNA sample is visibly transparent and clear, without obvious viscosity, without obvious pigments or impurities.

No RNA or protein residues in the DNA sample. Typically measured by a UV spectrophotometer, represented by OD values (A260/280, A260/230).

4.1.2.3 Volume, Concentration and Total amount of DNA sample

The volume, concentration and total amount of the DNA sample should meet the requirements of the library preparation kit.

4.1.2.4 Components of DNA sample dissolution buffer

The components of the DNA sample dissolution buffer (including the chemical reagent formula of the dissolution buffer, pH value, etc.) should meet the requirements of the library preparation kit.

4.2 Library requirements

4.2.1 Library types

Typically divided into two types: without barcode and with barcode.

Based on whether the library construction process relies on PCR amplification, distinguishing between PCR libraries and PCR-free libraries.

4.2.2 Library quality

4.2.2.1 Size and distribution of library

Should comply with the requirements of the library preparation kit and the sequencer’ s manual, and there should be no contamination with adapter or primer dimers.

4.2.2.2 Concentration, Volume, and Total amount of the library

Should comply with the requirements of the library preparation kit and the sequencer’ s manual.

4.3 Sequencing quality requirements

4.3.1 Sequencing read length and Sequencing type

Select the sequencing read length according to the requirements of human whole genome high-throughput sequencing applications, which is typically PE150.

4.3.2 Sequencing raw base

Should comply with the requirements of the sequencer’ s manual.

4.3.3 Percentage of base call quality

Should comply with the requirements of the sequencer’ s manual, typically with a Q30 score of no less than 85%.

4.3.4 Barcode/Index split rate

When sequencing is performed with barcode library, the split rate comply with the requirements of the sequencer ’ s manual.

4.4 Quality requirements for single-sample sequencing data

4.4.1 Genome alignment rate

Should meet the requirements for human whole genome high-throughput sequencing.

For conventional sample types, such as those derived from cell lines, whole blood, buffy coat, tissues, etc., the standard should not be less than 99%.

For other sample types, such as saliva, oral swabs, amniotic fluid, etc., the standard should not be less than 70%.

4.4.2 GC content

Should meet the requirements for human whole genome high-throughput sequencing. Typically, it should be between 39% and 43%.

NOTE If GC content deviates significantly from the recommended threshold during the establishment of laboratory methods, while the standard material values are normal, it is necessary to investigate the sample source or sample extraction method for potential issues.

4.4.3 Sequencing depth

Should meet the requirements for human whole genome high-throughput sequencing.

Typically, it should be not less than 40x.

4.4.4 Coverage rate of sequencing at least 20 ×

Should meet the requirements for human whole genome high-throughput sequencing.

Typically, it should be not less than 95%.

4.4.5 Duplication rate

Should meet the requirements for human whole genome high-throughput sequencing.

Typically, it should be not less than 30%.

4.4.6 Base mismatch rate

Should meet the requirements for human whole genome high-throughput sequencing.

Typically, it should be not less than 1%.

4.4.7 Specific region sequencing coverage

Should meet the requirements for human whole genome high-throughput sequencing.

Typically, the 10x sequencing coverage for a specific region should be no less than 90%.

NOTE A specific region refers to selected representative gene or genes, such as FOXE3, SMN1, SMN2, FMR1, G6PD, DUOX2, GJB2, PAH, ETFDH, MMACHC, SLC25A13, GCDH, etc.

4.5 Quality requirements for single-sample variant detection

4.5.1 Quality requirements of SNP detection

4.5.1.1 Total SNP numbers

Should meet the requirements for human whole genome high-throughput sequencing.

Typically, it should fall within the threshold range under the corresponding analysis pipeline and standard variation dataset.

4.5.1.2 SNP precision and sensitivity

Should meet the requirements for human whole genome high-throughput sequencing.

When evaluating with a human whole genome standard reference material and accompanying standard variant dataset, the SNP accuracy should be no less than 99%, and the SNP sensitivity should be no less than 98%.

4.5.2 Quality requirements of Indel detection

4.5.2.1 Total indel numbers

Should meet the requirements for human whole genome high-throughput sequencing.

Typically, it should fall within the threshold range under the corresponding analysis pipeline and standard variation dataset.

4.5.2.2 Indel precision and sensitivity

Should meet the requirements for human whole genome high-throughput sequencing.

When evaluating with a human whole genome standard reference material and accompanying standard variant dataset, the Indel accuracy should be no less than 90%, and the Indel sensitivity should be no less than 90%.

NOTE Typically, the precision and sensitivity of indel detection in PCR-free libraries are higher than those in PCR libraries.

4.5.3 Quality requirements of SV detection

4.5.3.1 Total SV numbers

Should meet the requirements for human whole genome high-throughput sequencing.

Typically, it should fall within the threshold range under the corresponding analysis pipeline and standard variation dataset.

4.5.3.2 SV precision and sensitivity

Should meet the requirements for human whole genome high-throughput sequencing.

When evaluating with a human whole genome standard reference material and accompanying standard variant dataset, the Indel accuracy should be no less than 85%, and the Indel sensitivity should be no less than 50%.

5. Evaluation method

5.1 Sample preparation

Human genomic DNA should be used, and the sample quality should comply with the requirements of section 4.1

5.2 Library preparation

Construct the human whole-genome high-throughput sequencing library or libraries according to the library preparation kit’ s manual, and the library quality should comply with the requirements of section 4.2.

5.3 High-throughput sequencing

Sequence the library on the sequencer according to the sequencer’ s manual, and the quality of the raw data obtained should comply with the requirements of section 4.3.

5.4 Sequencing quality evaluation of single sample

After data filtering the raw sequencing data of a single sample, the data quality should comply with the requirements of section 4.4.

5.5 Variation detection quality evaluation of in single sample

Using the human genome standard reference material and accompanying standard variant dataset, the precision and sensitivity of variant detection should meet the requirements of section 4.5.

Annex A

(informative)

Information of Human Genome Standard Materials

A.1 Overview

This appendix provides the information about the human genome standard materials used for the performance evaluation of the sequencer in Chapter 4 of this document, which is sourced from the ‘National Reference Material for Sequencer Performance Evaluation DNA’ (Reference Material Number: 360070).

A.2 Purpose

The human genome standard reference material used in this document is genomic DNA extracted from immortalized cell lines constructed from the peripheral blood leukocytes of healthy individuals.

A.3 Specification and components

Table A.1 Specifications and composition of the reference material

Number	Name	DNA Source	Cell or Bacterial Source	Reference Genome Sequence Version or Sources
1	Human Whole Genome DNA	NIFDC-HJ cell line	Normal male peripheral blood cell construction of immortal cell line	hs37d5

A.4 Others

The current national reference materials manual may be searched and downloaded from the website of the distribution organization of the national reference. Part of the content of the national reference materials manual may be changed based on the batch of the reference materials.

Bibliography

[1] Hao Bolin. Bioinformatics Manual (Second Edition). Shanghai: Shanghai Science and Technology Press, 2002.

[2] Chen Ming. Bioinformatics (Third Edition). Beijing: Science Press, 2018.

[3] Yang Huanming. Genomics. Beijing: Science Press, 2016.

[4] GB/T 29859-2013 Bioinformatics terms

[5] GB/T 30989-2014 Technical regulation for high-throughput gene sequencing

[6] GB/T 35537-2017 Requirements of the high-throughput gene sequencing result evaluation

[7] GB/T 35890-2018 Technical specification of high throughput sequencing data format

[8] YY/T 1723-2020 High-throughput gene sequencer

[9] GA/T 1693-2020 Forensic sciences—Specifications for second generation sequencing-based DNA examination