

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

(Draft for approval)

Issue date: 2024-1-24

Implementation date: 2024-XX-XX

Issued by State Administration for Market Regulation

Standardization Administration of the People's Republic of China

Contents

Foreword	1
1. Scope	2
2. Normative references	2
3. Terms and definitions	2
4. Abbreviations	6
5. Quality requirements	7
6. Evaluation method	10

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, definitions, quality requirements, and evaluation methods involved in the quality assessment of human whole genome sequencing data.

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

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

2. Normative references

This document does not contain normative references.

3. Terms and definitions

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

3.1

Human whole genome sequencing

Sequencing the entire genome of different individuals or populations.

Note: It includes the nucleic acid sequences of the 23 pairs of human chromosomes and mitochondrial sequences.

3.2

Library

A group of DNA fragments to be sequenced, each with added adapters (known DNA sequence fragments).

Note: Libraries are usually classified into PCR libraries (constructed through a certain number of PCR amplification cycles) and PCR-free libraries (constructed directly without PCR amplification).

3.3

Barcode split rate or index split rate

The percentage of sequencing reads in mixed sample sequencing that are correctly identified and assigned to their respective barcode or index, relative to the total number of sequencing reads.

3.4

Percentage of base call quality

The percentage of sequencing bases with base recognition quality above the specified threshold to the total number of sequenced bases.

Note: Usually represented as Q20, Q30, etc.

3.5

GC content

The percentage of guanine(G) and cytosine(C) in the sequencing bases to the total number of nucleotide bases, including adenine (A), thymine (T), G and C.

3.6

Data filtering

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

3.7

Sequencing raw base

The total number of original sequencing bases before data filtering.

Note: Also referred to as raw bases.

3.8

Human reference genome sequence

The publicly available and released human reference genome sequence.

Example: hs37d5, hg38, hg19, etc.

3.9

Mapping reads rate

The percentage of sequencing reads aligned to the human reference genome sequence out of the total effective sequencing reads.

3.10

Effective sequencing depth

The average sequencing depth of a whole genome sequencing sample, obtained after data filtering, sequence alignment, and deduplication.

The calculation method of effective sequencing depth can be found in formula (1)

$$ESD = (TBNMGNR - TBNDNR) / TBNGNR \quad \dots\dots\dots (1)$$

In the formula:

ESD——Effective Sequencing Depth, the unit is $\times$.

TBNMGNR——Total Base Number of Mapped to Genome non-N Regions, the unit is bp.

TBNDNR——Total Base Number of Duplicate Sequence Reads, the unit is bp.

TBNGNR —— Total Base Number of the Genome non-N Regions , the unit is bp.

3.11

Base mismatch rate

The percentage of number of bases that do not match the reference genome sequence, divided by the total number of bases aligned to the reference genome sequence.

3.12

Duplication rate

The percentage of number of sequencing reads that have the same position, direction, and base sequences aligned to the reference sequence, divided by the total number of sequencing reads aligned to the reference sequence.

3.13

Coverage rate of sequencing at least 20 $\times$

The percentage of the number of non-N bases on the reference genome sequence covered at least 20 times by sequencing reads after alignment to the reference genome out of the total number of non-N bases.

3.14

Insertion and deletion, InDel

Variations in genomic DNA with nucleotide insertions or deletions of length less than or equal to 50 bp.

3.15

Structural variation, SV

Sequence changes and positional relationship changes in large fragments greater than 50 bp.

Note: Includes deletions, insertions, duplications, inversions, and translocations.

3.16

Standard variation dataset

A high-confidence dataset of variations constructed based on the human genome reference materials.

3.17

Precision

The percentage of the number of detected true variations out of the total number of detected variants.

The calculation method of precision can be found in formula (2)

$$PC=TP/(TP+FP) \times 100\% \quad \text{-----}(2)$$

In the formula:

PC——refers to the precision of variant detection.

TP——refers to the number of detected variations that are consistent with the standard variation dataset variations.

FP——refers to the number of detected variations that are inconsistent with the standard variation dataset variations.

Note: Unlike the general definition of precision in other fields, precision in the sequencing field specifically refers to the accuracy of variant detection results.

[Source: YY/T 1723-2020, 5.4.1.2, modified]

3.18

Sensitivity

The percentage of the number of detected true variants out of the total number of all variants in the standard variation dataset.

The calculation method of sensitivity can be found in formula (3)

$$SE=TP/(TP+FN) \times 100\% \quad \text{-----}(3)$$

In the formula:

SE——refers to the sensitivity of variant detection

TP——refers to the number of detected variations that are consistent with the standard variant dataset variations.

FN——refers to the number of variations included in the standard variant dataset but not detected.

[Source: YY/T 1723-2020, 5.4.1.2, modified]

4. Abbreviations

The following abbreviations apply to this document.

Absorbance Ratio at 260 nm to 280 nm ($A_{260/280}$)

Absorbance Ratio at 260 nm to 280 nm ($A_{260/230}$)

base pair (bp)

Deoxyribonucleic Acid (DNA)

Effective Sequencing Depth (ESD)

Polymerase Chain Reaction (PCR)

False Positive (FP)

False Negative (FN)

Guanine and Cytosine (GC)

kilo base pair (kb)

Optical Density (OD)

Polymerase Chain Reaction-free (PCR-free)

Precision (PC)

Potential of Hydrogen (pH)

Paired-End (PE)

Ribonucleic Acid (RNA)

Sensitivity (SE)

Single Nucleotide Polymorphism (SNP)

Total Base Number of Mapped to Genome non-N Regions (TBNMGNR)

Total Base Number of Duplicate Sequence Reads (TBNSDR)

Total Base Number of the Genome non-N Regions (TBNGNR)

True Positive (TP)

5. Quality requirements

5.1 Sample quality requirements

5.1.1 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. The sample source information is clear, and the samples are properly preserved and transported without obvious contamination.

This document adopts human genomic DNA extracted from normal human cell lines approved and released by national authoritative institution, as the standard reference materials for quality assessment.

NOTE The information on the human genomic DNA standard reference materials involved in this document can be found in Appendix A.

5.1.2 DNA Sample quality

5.1.2.1 Integrity of DNA sample

Genomic DNA should be intact without significant degradation. Typically, 1%~2% agarose gel electrophoresis can be used, with genomic bands concentrated at 23 kb and above, showing no significant smearing or diffusion.

5.1.2.2 Purity of DNA sample

The DNA sample should be transparent and clear, without significant viscosity, pigmentation, or impurities. It should not contain RNA or protein residues. Typically, a ultraviolet spectrophotometer is used to measure the OD values ($A_{260/280}$, $A_{260/230}$). The purity of a DNA sample should meet the requirements of the library preparation reagents.

5.1.2.3 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 reagents.

5.1.2.4 Components of DNA sample dissolution buffer

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

5.2 Library quality requirements

5.2.1 Library type

Library types are typically categorized into two types: libraries without barcodes and those with barcodes. Additionally, libraries can be categorized based on the reliance on PCR amplification during the library

preparation process, distinguishing between PCR libraries and PCR-free libraries.

5.2.2 Library quality

5.2.2.1 Size and distribution of library

The fragment size and distribution of the sheared DNA or/and the library should meet the requirements of the library preparation reagents and the sequencer manual. Additionally, there should be no contamination from adapters or primer dimers.

5.2.2.2 Concentration, Volume, and Total amount of the library

The concentration, volume, and total amount of the library should meet the requirements of the library preparation reagents and the sequencer manual.

5.3 Sequencing quality requirements

5.3.1 Sequencing read length and Sequencing type

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

5.3.2 Sequencing raw base

The number of sequencing raw bases per sequencing slide/chip or per sequencing lane should meet the requirements of the sequencer manual.

5.3.3 Percentage of base call quality

The percentage of base call quality per sequencing slide/chip or per sequencing lane should meet the requirements of the sequencer manual, typically with a Q30 value of no less than 85%.

5.3.4 Barcode/Index split rate

When sequencing with barcode libraries, the barcode/index split rate per sequencing slide/chip or per sequencing lane should meet the requirements of the sequencer manual.

5.4 Quality requirements for single sample sequencing data

5.4.1 Genome alignment rate

The genome alignment rate should meet the requirements of human whole genome sequencing. The genome alignment rate should be no less than 99% for conventional sample types, such as those from cell lines, whole blood, blood leukocyte layer, tissues, etc. For certain sample types, such as saliva, buccal swabs, amniotic fluid, the genome alignment rate should be no less than 70%.

5.4.2 GC content

The GC content should meet the requirements of human whole genome sequencing, typically the GC content of a single sample should be 39%~43%.

NOTE At the initial establishment of laboratory methods, if the GC content of the human genome DNA standard reference materials is normal, but the GC content of the non-standard samples deviates significantly from the above recommended threshold, it is necessary to investigate factors affecting DNA sample quality (including but not limited to sample source or sample extraction methods).

5.4.3 Effective sequencing depth

The effective sequencing depth of a single sample should meet the requirements of human whole genome sequencing, typically being no less than $40\times$.

5.4.4 Coverage rate of sequencing at least $20\times$

The coverage rate of sequencing at least $20\times$ of a single sample should meet the requirements of human whole genome sequencing, typically being no less than 95%.

5.4.5 Duplication rate

The duplication rate of a single sample should meet the requirements of human whole genome sequencing, typically not exceeding 30%.

5.4.6 Base mismatch rate

The base mismatch rate of a single sample should meet the requirements of human whole genome sequencing, typically not exceeding 1%.

5.4.7 Specific region sequencing coverage

The specific region sequencing coverage of a single sample should meet the requirements of human whole genome sequencing, with $10\times$ sequencing coverage of specific regions typically being no less than 90%.

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

5.5 Quality requirements for single sample variant detection

5.5.1 Quality requirements of SNP detection

5.5.1.1 Number of SNPs

The number of detected SNPs of a single sample should meet the requirements of human whole genome sequencing. Typically, it should fall within the threshold range under the corresponding analysis pipelines and standard variant datasets.

5.5.1.2 Precision and sensitivity of SNPs

The precision and sensitivity of SNPs in a single sample should meet the requirements of human whole genome sequencing. When evaluated using human genome DNA standard reference materials and corresponding standard variant sets, the precision should be no less than 99%, and the sensitivity should

be no less than 98%.

5.5.2 Quality requirements of InDel detection

5.5.2.1 Number of InDels

The number of detected InDels in a single sample should meet the requirements of human whole genome sequencing. Typically, it should fall within the threshold range under the corresponding analysis pipelines and standard variant datasets.

5.5.2.2 Precision and sensitivity of InDels

The precision and sensitivity of InDels in a single sample should meet the requirements of human whole genome sequencing. When evaluated using human genome DNA standard reference materials and corresponding standard variant sets, the precision should be no less than 90%, and the sensitivity should be no less than 90%.

5.5.3 Quality requirements of SV detection

5.5.3.1 Number of SVs

The number of detected SVs in a single sample should meet the requirements of human whole genome sequencing. Typically, it should fall within the threshold range under the corresponding analysis pipelines and standard variant datasets.

5.5.3.2 Precision and sensitivity of SVs

The precision and sensitivity of SV in a single sample should meet the requirements of human whole genome sequencing. When evaluated using human genome DNA standard reference materials and corresponding standard variant sets, the precision should be no less than 85%, and the sensitivity should be no less than 50%.

6. Evaluation method

6.1 Sample preparation

Use human genomic DNA and determine whether the sample quality meets the requirements of 5.1.

6.2 Library preparation

Construct human whole genome sequencing libraries according to the instructions of library preparation reagents manual. Determine whether the library quality meets the requirements of 5.2.

6.3 High-throughput sequencing

Sequence the libraries on the sequencer according to the sequencer manual. Determine whether the quality of the raw data meets the requirements of 5.3.

6.4 Sequencing quality evaluation of a single sample

Filter the raw bases of a single sample and determine whether the data quality meets the requirements of 5.4.

6.5 Variation detection and quality evaluation of a single sample

Use human genome standard reference materials and corresponding standard variant datasets to determine whether the precision and sensitivity of variant detection meet the requirements of section 5.5.

Annex A

(informative)

Information of Human Genome Standard Reference Materials

A.1 Overview

This appendix provides information on the human genomic DNA standard reference materials applicable in Chapter 5: Human Whole Genome DNA, Human B-lymphocyte whole genome DNA standard reference materials.

A.2 Purpose

The human genome standard reference materials used in this document are genomic DNA samples extracted from immortalized cell lines constructed from healthy human peripheral blood leukocytes.

A.3 Specification and components

The specifications and composition of human whole genome standard samples are shown in Table A.1.

Table A.1 Specifications and Composition of Human Whole Genome Standard Samples

Number	Name	DNA Source	Cell Source	Human Reference Genome Sequence Source
1	Human Whole Genome DNA	NIFDC-HJ cell line	Immortalized Cell Line Constructed from Normal Male Peripheral Blood Leukocytes	Human Reference Genome hs37d5

A.4 Precautions

The standard materials and standard material manuals issued by current national authoritative institutions can be queried and downloaded from the distribution unit's website. Some contents of the manuals may change according to batches, and the current valid version shall prevail.

Bibliography

- [1] GB/T 29859—2013 Bioinformatics terms
- [2] GB/T 30989—2014 Technical regulation for high-throughput gene sequencing
- [3] GB/T 35537—2017 Requirements of the high-throughput gene sequencing result evaluation
- [4] GB/T 35890—2018 Technical specification of high throughput sequencing data format
- [5] GA/T 1693 — 2020 Forensic sciences — Specifications for second generation sequencing-based DNA examination
- [6] YY/T 1723—2020 High-throughput gene sequencer
- [7] Hao Bolin. Bioinformatics Manual [M]. 2nd Edition. Shanghai: Shanghai Science and Technology Press, 2002.
- [8] Chen Ming. Bioinformatics [M]. 3rd edition. Beijing: Science Press, 2018.
- [9] Yang Huanming. Genomics [M]. Beijing: Science Press, 2016.
- [10] Zook, J.M., McDaniel, J., Olson, N.D. et al. An open resource for accurately benchmarking small variant and reference calls. Nat Biotechnol 37, 561 – 566 (2019).
- [11] Krusche, P., Trigg, L., Boutros, P.C. et al. Best practices for benchmarking germline small-variant calls in human genomes. Nat Biotechnol 37, 555 – 560 (2019).
- [12] Ren, L., Duan, X., Dong, L. et al. Quartet DNA reference materials and datasets for comprehensively evaluating germline variant calling performance. Genome Biol 24, 270 (2023).