



National Standard of the People's Republic of China

GB/T XXXXX—XXXX

Single-molecule gene sequencing—
Part 1: Terms
单分子基因测序 第 1 部分：术语

Draft for approval

Issue date: Implementation date:

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

Foreword

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

This document is Part 1 of GB/T ****—**** “Single-molecule gene sequencing”. The following parts of GB/T ****—**** have been published:

——Part 1: Terms.

Certain contents of this document may involve patents. The issuing authority assumes no responsibility for identifying such patent.

This document is proposed by the National Medical Products Administration.

This document is under the jurisdiction of the Clinical Laboratory Testing and Invitro Diagnostic Test Systems (SAC/TC 136).

This document was drafted by: National Institutes for Food and Drug Control, BGI Research-Shenzhen, Beijing Qitan Technology Co., Ltd., Beijing Berry Genomics Co., Ltd., Beijing PolySeq Technology Co., Ltd., Anxuyuan Biotechnology (Shenzhen) Co., Ltd., Geneus Technologies (Chengdu) Co., Ltd., Zhejiang Digena Diagnostic Technology Co., Ltd., Science and Technology Research Center of China Customs, Beijing Institute of Medical Device Testing (Beijing Medical Biological Protective Equipment Inspection and Research Center), Hunan Institute for Drug Control, Beijing GrandOmics Biosciences Co., Ltd., Oxford Nanopore Technologies (Shanghai) Co., Ltd., Pacific Biosciences (Shanghai) Biotechnology Co., Ltd.

Main drafters of this document: Huang Jie, Zeng Tao, Dong Yuliang, Lin Jianxun, Mao Aiping, Liu Kesheng, Dong Lizeng, Su Yunpeng, Guo Huimin, Wang Yikai, Ren Yujie, Wang Ruixia, Liu Yuanyuan, Wang Yang, Pu Zijing, Xi Rui, Li Lili, Zhang Wenxin

INTRODUCTION

GB/T **** — specifies terms, quality evaluation of sequencers, sample processing specifications, data quality control requirements, analysis workflows, and formats, and software requirements for single-molecule gene sequencing. The standard is structured into five parts.

- — Part 1: Terms. To Define terms and definitions of single-molecule gene sequencing, and specify key metrical and technical terms for single-molecule gene sequencing featured with continuous sequencing characteristics.
- — Part 2: Quality evaluation of sequencers. To specify requirements and test methods for the quality evaluation of single-molecule gene sequencers.
- — Part 3: Sample processing specifications and quality requirements. To define sample types and technical requirements for sample processing compatible with single-molecule gene sequencing, and specify quality criteria and evaluation methods for sample.
- — Part 4: Data quality control requirements, analysis workflows, and formats. To define the terms and definitions for the quality of single-molecule gene sequencing data, specify quality criteria and evaluation methods for data, and describe the analysis pipelines for converting sequencing data to base sequence, hierarchical directory formats of sequencing read, and general formats of data.
- — Part 5: Software. To specify architecture, general requirements, specialized requirements, and validation methods for bioinformatics software utilized to analyze single-molecule gene sequencing data, as well as hardware configuration requirements, interface specifications and so forth for sequencing software.

Single-molecule gene sequencing—Part 1: Terms

1 Scope

This document defines terms and definitions for single-molecule gene sequencing.

This document applies to single-molecule gene sequencing technologies featured with continuous sequencing, including: single-molecule real-time fluorescent sequencing, single-molecule nanopore strand sequencing, single-molecule nanopore tag sequencing, etc.

The metrical and technical terms in this document are not compatible with:

- First generation of gene sequencing employing Sanger sequencing as the main technical principle;
- Massively parallel high-throughput gene sequencing employing semiconductor sequencing, reversible end termination sequencing, combined probe-anchored ligation sequencing, combined probe-anchored polymerization sequencing, pyrophosphate sequencing, and other technologies as the main technical principles;
- Single-molecule gene sequencing technology that utilizes multiple discrete steps for non-continuous sequencing;
- Synthetic single-molecule long-read gene sequencing technology utilizing cobarcoding for concatenation of short reads.

2 Normative References

This document cites no normative reference.

3 Terms and definitions

3.1 General terms

3.1.1

genome

complete set of all genetic information of an organism.

[Source: GB/T 30989—2014, 3.2]

3.1.2

gene

sequence of nucleotides encoding a specific functional product (e.g., a protein or RNA molecule), functioning as the fundamental unit of genetic information.

[Source: GB/T 30989—2014, 3.1, modified]

3.1.3

nucleic acid

macromolecule that is the medium for genetic information or acts as an agent in expressing the information.

Note to entry: There exist two types of nucleic acid, deoxyribonucleic acid (3.1.4) and ribonucleic acid (3.1.5).

[Source: ISO 17822:2020, 3.32, modified]

3.1.4

deoxyribonucleic acid; DNA

polymer of deoxyribonucleotides occurring in a double-stranded (dsDNA) or single-stranded (ssDNA) form.

[Source: ISO 17822:2020, 3.17]

3.1.5

ribonucleic acid; RNA

polymer of ribonucleotides occurring in a double-stranded or single-stranded form.

[Source: ISO 17822:2020, 3.42]

3.1.6

base

nitrogen-containing heterocyclic organic compound which is the primary building block of purines and pyrimidines, functioning as the "letters" of the genetic code.

Note to entry: In DNA, the primary bases are adenine (A), guanine (G), cytosine (C), and thymine (T). In RNA, the primary bases are adenine (A), guanine (G), cytosine (C), and uracil (U).

[Source: GB/T 30989—2014, 3.16]

3.1.7

base sequence

string of characters representing the arrangement of bases (3.1.6) in a sequencing read.

Note to entry: Bases in a sequence are represented by uppercase or lowercase letters, with undetermined bases denoted by "N" or "n".

[Source: GB/T 35890—2018, 3.6, modified]

3.1.8

gene sequencing

measurement of the different types of bases in a nucleic acid (3.1.3), that is, determination of the combination or order of bases (e.g., adenine (A), guanine (G), cytosine (C), thymine (T), or uracil (U)), as well as base modification information.

[Source: YY/T 1723—2020, 3.1, modified]

3.1.9

library

sequencing library

nucleic acid fragments that are used as sequencing templates, characterized by a specific size range and typically containing adapters and/or identifiers recognised for sequence specific priming, sequence capture, and/or identification of specific extracts.

[Source: ISO 20397—1:2022, 3.5, modified]

3.1.10

base calling

process of converting electrical signals, optical signals, or other signals generated during sequencing into base sequence (3.1.7) information.

3.1.11

quality of base calling

measure of the probability that a base (3.1.6) is correctly recognized, typically expressed numerically.

The relationship between quality of base calling and error rate of base calling can be expressed by formula (1):

$$Q = -10 \lg P \dots\dots\dots (1)$$

In the formula:

Q —quality of base calling;

P —error rate of base calling.

[Source: GB/T 30989—2014, 3.29, modified]

3.1.12

single-molecule gene sequencing

determination of base sequences (3.1.7) continuously in nucleic acids (3.1.3) at the single-molecule level.

Note 1 to entry: This single-molecule long-read sequencing method typically relies on optical or electrical signals for base identification, distinguishing itself from amplification-dependent approaches, e.g., Sanger sequencing, massively parallel high-throughput sequencing, as well as synthetic single-molecule sequencing by means of concatenation of short reads based on cobarcoding.

Note 2 to entry: Depending on the platform and library preparation, single-molecule gene sequencing enables direct measurement of tens of thousands to millions of continuous bases.

3.1.13

direct sequencing

sequencing method that reads the sequencing signals of the raw template strands (including modified bases) without amplification or conversion of the single-molecule sequencing library (3.1.17) during sequencing.

Note to entry: A technical feature of single-molecule gene sequencing (3.1.12).

3.1.14

direct RNA sequencing

sequencing method that directly reads the sequencing signals of the raw RNA template strands (including modified bases) after library preparation.

3.1.15

real-time sequencing

sequencing method that performs continuous sequencing reactions, base calling (3.1.10), and sequence output synchronously at the single-molecule level.

Note to entry: A technical feature of single-molecule gene sequencing (3.1.12).

3.1.16

direct sequencing for epigenetic modifications

sequencing method that directly detects epigenetic modifications on target nucleic acid (3.1.3) without chemical or biological conversion pretreatment of the nucleic acids (3.1.3).

Note to entry: Epigenetic modifications include 5mC (5-methylcytosine), 5hmC (5-hydroxymethylcytosine), and 6mA (6-methyladenine) on DNA, as well as m5C and m6A on RNA molecules.

3.1.17

single-molecule sequencing library

structurally tailored nucleic acid (3.1.3) prepared for single-molecule gene sequencing (3.1.12), typically formed by coupling target nucleic acids with platform-specific adapters via enzymatic or chemical methods.

Note to entry: Categorized into single-stranded circular DNA, double-stranded circular DNA, and double-stranded linear DNA libraries based on the topology of the libraries..

3.1.18

single-molecule consensus sequence

single high-confidence base sequence derived from integrating multiple copies, repeated reads, or complementary strands of a target region/fragment during single-molecule gene sequencing (3.1.12).

3.2 Metrical terms

3.2.1

throughput of sequencing

number of measured reads containing sequence information or number of measured DNA (3.1.4) and RNA (3.1.4) bases obtained in a single sequencing run.

Note to entry: Throughput metrics must specify units, e.g. reads or bases).

[Source: GB/T 30989—2014, 3.21, modified]

3.2.2

throughput of sequencing per flow cell

throughput of sequencing (3.2.1) achieved on a single sequencing flow cell.

3.2.3

throughput of sequencing per unit time

throughput of sequencing (3.2.1) achieved per unit time during a sequencing run.

Note to entry: Unit time, usually refers to hours (h) or minutes (min).

3.2.4

read length of sequencing

length of quality passed sequencing reads obtained in a single run, typically denoted as the number of bases..

Note to entry: Key metrics for single-molecule gene sequencing include maximum read length, average read length, read length N50, etc.

[Source: GB/T 30989—2014, 3.24, modified]

3.2.5

maximum read length

length of the longest quality passed sequencing read obtained in a single run, denoted as the number of bases.

3.2.6

average read length

length obtained by dividing the total number of bases of all quality passed sequencing reads obtained in a single run by the number of reads, denoted as the number of bases.

3.2.7

read length N50

length of the read last added when cumulative counts of bases reach or exceed 50% of the total number of bases of all reads, under the circumstances that quality passed sequencing reads are sorted by length from longest to shortest, denoted as the number of bases.

3.2.8

median read length

length of the read last tallied when cumulative counts of reads reach or exceed 50% of the total number of reads, under the circumstances that quality passed sequencing reads are sorted by length from longest to shortest, denoted as the number of bases.

3.2.9

accuracy of sequencing

concordance between raw reads or processed single-molecule consensus sequences [3.1.18] and the reference sequence for quality passed sequencing reads obtained in a single run.

Accuracy is calculated using formula (2):

$$Acc = M / (M + X + I + D) \dots\dots\dots (2)$$

In the formula:

Acc—accuracy of sequencing;

M—number of bases correctly mapped;

X—number of substitution errors;

I—number of insertion errors;

D—number of deletion errors.

Note to entry: Used for evaluation of a single read.

3.2.10

average accuracy

overall concordance between quality passed sequencing reads (raw or processed) and the reference sequence in a single run.

Note to entry: Used for statistical evaluation of the entire data set.

3.2.11

median accuracy

accuracy of the read last tallied when cumulative counts of reads reach or exceed 50% of the total number of reads, under the circumstances that quality passed sequencing reads are sorted by accuracy of sequencing (3.2.9) from highest to lowest.

3.2.12

modal accuracy

peak accuracy value in a histogram of accuracy of sequencing (3.2.9) generated with all quality passed sequencing reads obtained in a single run.

3.2.13

consensus accuracy

concordance between a consensus sequence (generated by multi-sequence alignment correction) and the reference sequence.

Note 1 to entry: Calculated similarly to accuracy of sequencing (3.2.9), denoted as the ratio of correctly aligned bases in the consensus sequence to the total bases in the target region.

Note 2 to entry: Used for the evaluation of consensus sequences.

3.2.14

single-pass accuracy of sequencing

concordance between uncorrected raw reads and the reference sequence.

Note to entry: Used for the evaluation of a single read.

3.2.15

single-molecule accuracy of sequencing

concordance between the reference sequence and the sequence derived from a single template molecule after base calling and correction (e.g., via multiple copy sequencing, circular consensus sequencing [3.3.6], or duplex sequencing [3.3.5]).

Note to entry: For platforms performing single-pass sequencing, this equals single-pass accuracy.

For platforms using circular consensus sequencing, this equals circular consensus accuracy. For duplex sequencing platforms, this equals intra-molecular complementary strand consensus accuracy.

3.3 Technical Terms

3.3.1

single-molecule real-time fluorescent sequencing

sequencing method where a single nucleic acid (3.1.3) is immobilized at the bottom of a zero-mode waveguide (ZMW) (3.3.9) and captured by a polymerase. During primer extension, newly incorporated bases generate continuous pulsed fluorescent signals (3.3.8), which are translated into base sequences and modification information via base calling.

3.3.2

single-molecule nanopore strand sequencing

sequencing method where a single nucleic acid (3.1.3) is captured by a nanopore (3.3.13). Bases pass through the nanopore sequentially under the control of a motor protein (3.3.15), generating time-dependent electrical signal changes that are converted into base sequences and modification information via base calling.

3.3.3

single-molecule nanopore tag sequencing

sequencing method where a single nucleic acid (3.1.3) is captured by a polymerase attached to a nanopore (3.3.13). During primer extension, tagged nucleotides interact with the nanopore (via translocation or other mechanisms), producing characteristic current signals that are translated into base sequences and modification information via base calling.

3.3.4

single-molecule sequencing by synthesis

sequencing method that identifies base sequences by detecting fluorescently labeled or tag labeled nucleotides or derivatives as they are sequentially incorporated into a newly synthesized nucleic acid strand.

3.3.5

duplex sequencing

sequencing method where both the template strand and its complementary strand are sequenced and base called sequentially.

Note to entry: Complementary strand sequencing improves accuracy by cross-validating the template strand sequence.

3.3.6

circular consensus sequencing

sequencing method that generates subreads (3.3.7) by repeatedly sequencing a circularized template molecule, in which the subreads are aligned and corrected to produce a high-confidence consensus sequence.

Note to entry: Applied to single-molecule real-time fluorescent sequencing (3.3.1) and single-molecule nanopore strand sequencing (3.3.2).

3.3.7

subread

sequence read generated in one round of circular consensus sequencing (3.3.6).

Note: One round refers to the sequencing of a full template strand or complementary strand. For example, in dumbbell-shaped libraries, one rolling circle replication cycle produces two rounds (template and complementary strand reads).

3.3.8

pulsed fluorescent signal

transient fluorescent signal emitted during single-molecule real-time fluorescent sequencing (3.3.1) when a fluorophore-labeled nucleotide enters the detection zone and is incorporated into the primer strand mediated by a polymerase.

3.3.9

zero - mode waveguide; ZMW

nanophotonic device with a tens-of-nanometer diameter pore structure for detecting optical signals emitted in a confined, ultra tiny volume.

Note to entry: The excitation is confined to the vicinity of the polymerase at the bottom of the pore owing to the subwavelength diameter, thus enabling monitoring of single-molecule synthesis fluorescent events, so as to detect optical signals with specific wavelengths while minimizing background fluorescence.

3.3.10

dumbbell-shaped library

sequencing library formed by linking both ends of a double-stranded template molecule with hairpin adapters, creating a circular "dumbbell" structure

Note to entry: Compatible with multiple single-molecule sequencing technologies, including single-molecule real-time fluorescent sequencing (3.3.1) and single-molecule nanopore tag sequencing (3.3.3).

3.3.11

kinetics of sequencing

method for analyzing nucleic acid (3.1.3) modifications by modeling differences in pulse width and inter-pulse duration of signals generated during nucleotide incorporation, which vary based on base types and modifications.

3.3.12

tagged nucleotide

modified nucleotide with a unique tag molecule attached to its phosphate group, used in single-molecule nanopore tag sequencing (3.3.3).

Note to entry: Different nucleotides carry distinct tag.

3.3.13

nanopore

nanoscale pore with a transmembrane channel, functioning as a transducer to encode base sequences into electrical signals.

Note to entry: Usually categorized into biological protein pores and solid-state pores.

3.3.14

nanopore sequencing flow cell

complete sequencing unit integrating nanopore sensors, signal detection components, and fluidic structures.

3.3.15

motor protein

protein used in single-molecule nanopore strand sequencing (3.3.2) to control the translocation speed of nucleic acids (3.1.3).

Note to entry: Usually helicases, also can be polymerases or other types of proteins.

3.3.16

adaptive sampling

targeted sequencing method that supports sequencers to select or reject sequences based on preset conditions during real-time sequencing.

Note to entry: This technology takes advantage of real-time base calling and accepts or rejects DNA fragments for further sequencing based on their initial sequence, to achieve specific sequencing goals or to improve the quality of sequencing data.

Bibliography

- [1] GB/T 30989—2014 Technical regulation of high-throughput gene sequencing
- [2] GB/T 35890—2018 Technical specification of high throughput sequencing data format
- [3] YY/T 1723—2020 High-throughput gene sequencer
- [4] ISO 17822:2020 In vitro diagnostic test systems — Nucleic acid amplification-based examination procedures for detection and identification of microbial pathogens — Laboratory quality practice guide
- [5] ISO 20397-1:2022 Biotechnology — Massively parallel sequencing — Part 1: Nucleic acid and library preparation
- [6] Eid J., Fehr A., Gray J. et al. Real-time DNA sequencing from single polymerase molecules. *Science* 2009 Jan, 233 pp.133-138
- [7] Wenger A.M., Peluso P., Rowell W.J. et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat. Biotechnol.* 2019 Oct, 37 pp.1155-1162
- [8] Wang Y.H., Zhao Y., Bollas A. et al. Nanopore sequencing technology, bioinformatics and applications. *Nat. Biotechnol.* 2021 Nov, 39 pp.1348-1365
- [9] Hook P.W., Timp W. Beyond assembly: the increasing flexibility of single-molecule sequencing technology. *Nat. Rev. Genet.* 2023 Sep, 24 pp.627-641
- [10] Fuller C.W., Kumar S., Porel M. et al. Real-time single-molecule electronic DNA sequencing by synthesis using polymer-tagged nucleotides on a nanopore array. *Proc. Natl. Acad. Sci. U.S.A.* 2016 May, 113 pp.5233-5238

Indexing

A

average accuracy.....	3.2.10
accuracy of sequencing.....	3.2.9
average read length accuracy of sequencing.....	3.2.6
adaptive sampling.....	3.3.16

B

base.....	3.1.6
base calling.....	3.1.10
base sequence.....	3.1.7

C

circular consensus sequencing.....	3.3.6
consensus accuracy.....	3.2.13

D

deoxyribonucleic acid.....	3.1.4
dumbbell-shaped library.....	3.3.10
direct RNA sequencing.....	3.1.14
direct sequencing.....	3.1.13
duplex sequencing.....	3.3.5
direct sequencing for epigenetic modifications.....	3.1.16

G

genome.....	3.1.1
gene.....	3.1.2
gene sequencing.....	3.1.8

K

kinetics of sequencing.....	3.3.11
-----------------------------	--------

M

median accuracy	3.2.11
modal accuracy.....	3.2.12
motor protein.....	3.3.15
maximum read length.....	3.2.5
median read length.....	3.2.8

N

nucleic acid.....	3.1.3
nanopore.....	3.3.13

nanopore sequencing flow cell.....	3.3.14
P	
pulsed fluorescent signal.....	3.3.8
Q	
quality of base calling.....	3.1.11
R	
ribonucleic acid.....	3.1.5
read length N50.....	3.2.7
read length of sequencing.....	3.2.4
real-time sequencing.....	3.1.15
S	
single-pass accuracy of sequencing.....	3.2.14
single-molecule accuracy of sequencing.....	3.2.15
single-molecule consensus sequence.....	3.1.18
single-molecule gene sequencing.....	3.1.12
sequencing library.....	3.1.9
single-molecule nanopore strand sequencing.....	3.3.2
single-molecule nanopore tag sequencing.....	3.3.3
single-molecule real-time fluorescent sequencing.....	3.3.1
single-molecule sequencing by synthesis.....	3.3.4
single-molecule sequencing library.....	3.1.17
subread.....	3.3.7
T	
tagged nucleotide.....	3.3.12
throughput of sequencing.....	3.2.1
throughput of sequencing per flow cell.....	3.2.2
throughput of sequencing per unit time.....	3.2.3
Z	
zero-mode waveguide.....	3.3.9