

ICS 07.080
CCS A 40



**NATIONAL STANDARD OF THE
PEOPLE'S REPUBLIC OF CHINA**

GB/T 43584.2-2023

**Biotechnology - Massively parallel sequencing - Part 2: Quality
evaluation of sequencing data**

生物技术 大规模并行测序 第2部分：测序数据的质量评估

(English Translation)

Issued on: December 28, 2023

Implemented on: December 28, 2023

**Issued by: State Administration for Market Regulation, China National
Standardization Administration**

Table of Contents

Preface	
Introduction	
1 Scope	
2 Normative references	
3 Terms and Definitions	
4 Raw data	
4.1 General	
4.2 Raw data files	
4.3 Quality assessment of raw data	
4.4 Raw data preprocessing	
5 Sequence alignment and positioning	
5.1 General	
5.2 Sequence alignment and positioning file format	
5.3 Quality control of sequence alignment and mapping	
5.4 Post-processing	
6 Variant identification	
6.1 General	
6.2 Data files for variant calling	
6.3 Quality indicators for variant calling	
6.4 False positive mutation processing	
6.5 Sequence annotation	
7 Verify	
7.1 General	
7.2 Quality indicator verification	
8 File	
Appendix A (Informative) Quality Indicators for Specific MPS Platform Examples	
Appendix B (informative) Coverage and recommended reading order by application	
Appendix C (Informative) Sequence Alignment and Positioning Software	
Reference	

Foreword

This document complies with the provisions of GB/T 1.1-2020 "Standardization Work Guidelines Part 1: Structure and Drafting Rules of Standardization Documents" Drafting:

This document is Part 2 of GB/T 43584 "Massively Parallel Sequencing in Biotechnology":

GB/T 43584 has published the following parts:

---Part 2: Quality assessment of sequencing data:

This document is equivalent to ISO 20397-2:2021 "Massively parallel sequencing in biotechnology Part 2: Quality assessment of sequencing data" estimate":

Please note that some content in this document may be subject to patents: The publisher of this document assumes no responsibility for identifying patents:

Introduction

Massively parallel sequencing (MPS) is a high-throughput analysis method for nucleic acid sequencing that utilizes massively parallel processing:

Study the whole genome, transcriptome and specific target nucleic acids of different organisms in a relatively short period of time:

MPS has been used in many life science fields and can perform determination and high-throughput analysis of millions or even tens of millions of nucleotide bases: born Biological variation in DNA and RNA polymers within the body poses challenges for accurate sequence determination: Measured by MPS, sequence Column quality depends on many factors, including but not limited to sample quality, library preparation, platform selection, and sequencing data quality:

GB/T 43584 is intended to consist of the following parts:

---Part 1: Nucleic acid and library preparation: Part 1 mainly provides basic research and aims to provide the prerequisites for sequencing and data generation:

General guidelines and considerations for library preparation and library quality assessment:

---Part 2: Quality assessment of sequencing data: Part 2 carries out specific operations and data quality control based on Part 1 and provides Part 3 provides the basis for the research:

---Part 3: General requirements and guidelines for metagenomics: Part 3 includes Part 1 and Part 2, and stipulates macrogenes Omics guidelines from sample preparation, generation and analysis of sequencing data:

Sequencing data analysis poses major challenges to bioinformatics in many areas such as data storage, computing time, and mutation detection accuracy:

war: One of the major challenges associated with sequencing data is monitoring quality control metrics at all stages of the data processing pipeline, which can easily be overlooked:

Understanding data quality is critical to downstream sequence analysis: Quality control of nucleic acid sequencing data processing and analysis can be divided into three stages: original Data, alignment and variant identification: This document provides precautions for MPS sequencing data quality assessment, as well as instructions for different MPS platforms:

The station provides specific suggestions:

Massively parallel sequencing in biotechnology Part 2: Quality assessment of sequencing data

1 Scope

This document clarifies the overall requirements and recommendations for quality assessment of massively parallel sequencing data: Covers the post-generation of raw data Programs, sequence alignment and variant calling:

This document provides general guidance for the validation and archiving of massively parallel sequencing (MPS) data:

This document does not apply to any processing associated with de novo assembly:

2 Normative reference documents

This document has no normative references:

3 Terms and definitions

The following terms and definitions apply to this document:

adapter sequence adapter an artificial oligonucleotide of known sequence that can be attached to the 3' or 5' end of a nucleic acid fragment:

NOTE: It provides primer sites and other necessary sequences required for sequencing insert fragments:

Algorithm

A fully deterministic finite sequence of instructions by which the value of an output variable can be calculated from the value of an input variable:

[Source: IEC 60050-351:2013,351-42-27, with modifications] 3.3 base calling A computational process that converts raw electrical signals from massively parallel sequencing into nucleotide sequences:

NOTE: Base calling applications and algorithm performance are determined by read sequence and consensus sequence accuracy:

The integration and sequential execution of a program, script, or software in which raw data or the output of one program is used as the next input for each step:

Example: The output of a base quality shearing program can be used as input to a de novo assembly program:

captureefficiency

The measured target region sequence as a percentage of all sequenced sequences or reference sequences:

coverage

coveragedepth in a sequencing run, the number of times each base at a specified position is sequenced: