ФЕДЕРАЛЬНОЕ АГЕНТСТВО

ПО ТЕХНИЧЕСКОМУ РЕГУЛИРОВАНИЮ И МЕТРОЛОГИИ

НАЦИОНАЛЬНЫЙ СТАНДАРТ РОССИЙСКОЙ ФЕДЕРАЦИИ

ГОСТ Р 72598— 2026/ ИСО/TS 24420:2023

Биотехнология

МАССОВОЕ ПАРАЛЛЕЛЬНОЕ СЕКВЕНИРОВАНИЕ ДНК

Общие требования к обработке данных метагеномного секвенирования методом «дробовика»

(ISO/TS 24420:2023, IDT)

Издание официальное

Москва
Российский институт стандартизации
2026

gost.e-ecolog.ru

Предисловие

1 ПОДГОТОВЛЕН Федеральным государственным бюджетным учреждением «Российский институт стандартизации» (ФГБУ «Институт стандартизации») на основе собственного перевода на русский язык англоязычной версии документа, указанного в пункте 4

2 ВНЕСЕН Техническим комитетом по стандартизации ТК 326 «Биотехнологии»

3 УТВЕРЖДЕН И ВВЕДЕН В ДЕЙСТВИЕ Приказом Федерального агентства по техническому регулированию и метрологии от 20 марта 2026 г. № 261-ст

4 Настоящий стандарт идентичен международному документу ISO/TS 24420:2023 «Биотехнология. Массовое параллельное секвенирование ДНК. Общие требования к обработке данных метагеномного секвенирования методом «дробовика» (ISO/TS 24420:2023 «Biotechnology — Massively parallel DNA sequencing — General requirements for data processing of shotgun metagenomic sequences», IDT).

При применении настоящего стандарта рекомендуется использовать вместо ссылочного международного стандарта соответствующий ему национальный стандарт, сведения о котором приведены в дополнительном приложении ДА

5 ВВЕДЕН ВПЕРВЫЕ

6 Некоторые положения международного документа, указанного в пункте 4, могут являться объектом патентных прав. Международная организация по стандартизации (ИСО) не несет ответственности за идентификацию подобных патентных прав

*Правила применения настоящего стандарта установлены в статье 26 Федерального закона от 29 июня 2015 г. № 162-ФЗ «О стандартизации в Российской Федерации». Информация об изменениях к настоящему стандарту публикуется в ежегодном (по состоянию на 1 января текущего года) информационном указателе «Национальные стандарты», а официальный текст изменений и поправок — в ежемесячном информационном указателе «Национальные стандарты». В случае пересмотра (замены) или отмены настоящего стандарта соответствующее уведомление будет опубликовано в ближайшем выпуске ежемесячного информационного указателя «Национальные стандарты». Соответствующая информация, уведомление и тексты размещаются также в информационной системе общего пользования — на официальном сайте Федерального агентства по техническому регулированию и метрологии в сети Интернет (*gost.e-ecolog.ru

© ISO, 2023
© Оформление. ФГБУ «Институт стандартизации», 2026

Настоящий стандарт не может быть полностью или частично воспроизведен, тиражирован и распространен в качестве официального издания без разрешения Федерального агентства по техническому регулированию и метрологии

Содержание

1 Область применения 1
2 Нормативные ссылки 1
3 Термины и определения 1
4 Процесс обработки 3
5 Обработка данных 4
6 Анализ данных 5
7 Архив данных и метаданные 6
Приложение А (справочное) Примеры форматов данных 8
Приложение В (справочное) Каталог элементов данных 13
Приложение ДА (справочное) Сведения о соответствии ссылочного международного стандарта национальному стандарту 15
Библиография 15

Введение

Метагеномное секвенирование методом «дробовика» геномов организмов из сложного образца широко используется в науках о жизни, а также в области клинических исследований (например, анализ сложных заболеваний человека, микроэкология окружающей среды и другие области) с целью получения знаний об их составе и функциях. Данный метод может предоставить важные научные данные для исследований в области естественных наук.

Преимуществом данного метода является его способность выявлять микробное разнообразие и распространенность, обнаруженные в микробных популяциях из различных сред, и определять информацию о последовательности (таксономическая характеристика, функциональная аннотация и сравнительный анализ/метагеномика) для отдельных организмов в этих популяциях. Полученные данные могут быть подвергнуты сравнительному анализу. Массово-параллельное метагеномное секвенирование методом «дробовика» генерирует большой объем данных, содержащих высокую сложность микробных геномов и большое количество неизвестных видов. Важно использовать эффективные процедуры обработки и инструменты контроля качества для данных метагеномного секвенирования методом «дробовика». Стандартизированный формат данных необходим для содействия обмену данными.

Как и любая передовая технология, технологии массового параллельного секвенирования подвержены ошибкам. Важно преодолеть эти недостатки, чтобы обеспечить надежный результат секвенирования и анализа. Настоящий стандарт содержит единые требования для сопоставления, хранения и последующего анализа метагеномных данных, а также руководящие принципы. Он содержит требования и рекомендации для рабочего процесса и процесса метагеномного анализа методом «дробовика», включая контроль качества данных секвенирования и метаданных, а также композиционный и функциональный анализ микробного сообщества. Эти требования и рекомендации могут обеспечить точность данных, полученных в результате метагеномного анализа, устранить потенциальные ошибки и облегчить последующее применение.

НАЦИОНАЛЬНЫЙ СТАНДАРТ РОССИЙСКОЙ ФЕДЕРАЦИИ

Биотехнология

МАССОВОЕ ПАРАЛЛЕЛЬНОЕ СЕКВЕНИРОВАНИЕ ДНК

Общие требования к обработке данных метагеномного секвенирования методом «дробовика»

Biotechnology. Massively parallel DNA sequencing.
General requirements for data processing of shotgun metagenomic sequences

Дата введения — 2026—09—01

1 Область применения

В настоящем стандарте представлен рабочий процесс обработки данных метагеномного секвенирования методом «дробовика» микробиома, полученного от хозяина, и метагеномов окружающей среды.

Настоящий стандарт устанавливает требования к контролю качества обработки данных метагеномного секвенирования методом «дробовика» для массового параллельного секвенирования ДНК.

Настоящий стандарт содержит рекомендации по каталогу данных, архиву данных и метаданным для данных метагеномного секвенирования методом «дробовика».

Настоящий стандарт также рассматривает аспекты, связанные с хранением данных, обменом данными и взаимодействием данных метагеномного секвенирования методом «дробовика».

В настоящем стандарте рассматриваются вопросы обработки и анализа данных метагеномного секвенирования методом «дробовика», но исключается их функциональный анализ.

2 Нормативные ссылки

В настоящем стандарте использована нормативная ссылка на следующий стандарт [для датированной ссылки применяют только указанное издание ссылочного стандарта, для недатированных — последнее (включая все изменения)]:

ISO 20397-1:2022, Biotechnology — Massively parallel sequencing — Part 1: Nucleic acid and library preparation (Биотехнология. Массовое параллельное секвенирование. Часть 1. Подготовка нуклеиновой кислоты и библиотеки)

3 Термины и определения

В настоящем стандарте применены следующие термины с соответствующими определениями.

ИСО и МЭК поддерживают терминологические базы данных, используемые в стандартизации по следующим адресам:

- платформа онлайн-просмотра ИСО: доступна по адресу gost.e-ecolog.ru
- Электропедия МЭК: доступна по адресу gost.e-ecolog.ru

3.1 **значение атрибута** (attribute value): Значение, связанное с экземпляром атрибута.

[ИСО 21962:2003, 1.5.2.3]

3.2 **категория** (category): Набор элементов или концепций, которые имеют общий атрибут или функцию.

Издание официальное

3.3 **классификация** (classification): Исчерпывающий набор взаимоисключающих категорий для агрегации данных на заранее заданном уровне специализации для определенной цели.

[ИСО 17115:2007, 2.7.1]

3.4 **очищенные данные** (clean data): Секвенированные данные, полученные после процедуры предварительной обработки, которая обычно включает несколько этапов обрезки и фильтрации для обеспечения определенных уровней качества (например, базовое качество, последовательности хозяина/контаминанты удалены, линкеры/адаптеры удалены).

3.5 **код** (code): Система правил для преобразования информации, такой как текст, изображения, звуки или электрические, фотонные или магнитные сигналы, в другую форму или представление для облегчения анализа, связи или хранения на носителе данных.

[ИСО 20691:2022, 3.6]

3.6 **кодирование** (encoding): Процесс присвоения кода объектам или концепциям.

3.7 **контиг** (contig): Непрерывная последовательность ДНК, созданная путем сборки перекрывающихся секвенированных фрагментов хромосомы или плазмиды.

3.8 **формат данных** (data format): Организация данных в соответствии с заданными требованиями.

Примечание 1 — Заданные требования обычно создаются для компьютерной обработки.

3.9 **элемент данных** (data element): Одна единица данных, которая в определенном контексте считается неделимой.

[ISO/TS 21089:2018, 3.44]

3.10 **каталог** (directory): Список элементов данных, который дает детализированную информацию, позволяющую отслеживать, идентифицировать и находить связанные данные.

Примечание 1 — Каталог может быть организован в алфавитном, хронологическом или систематическом порядке.

3.11 **идентификатор каталога** (directory identifier): Уникальный независимый от языка знак, присвоенный архивному каталогу в структуре.

3.12 **ген** (gene): Последовательность нуклеотидов в ДНК или РНК, кодирующая РНК либо белковый продукт.

Примечания

1 Гены признаются основной единицей наследственности.

2 Ген может состоять из несмежных сегментов нуклеиновой кислоты, которые перестраиваются на этапе обработки ядра.

3 Ген может включать или быть частью оперона, который содержит элементы для экспрессии гена.

[ИСО 20397-2:2021, 3.16]

3.13 **идентификатор** (identifier): Последовательность символов, способная однозначно идентифицировать то, с чем она связана, в указанном контексте.

[ИСО/МЭК 11179-1:2015, 3.1.3]

3.14 **аналитические данные** (analytical data): Набор элементов для описания качественных или количественных аналитических атрибутов обработанных метагеномных исходных данных.

3.15 **имя** (name): Семантические, естественно-языковые метки, присваиваемые элементам данных, и вариации этих меток, выполняющие различные функции

[ИСО/МЭК 11179-1:2015, 3.43]

3.16 **публичный атрибут** (public attribute): Атрибут, который может иметь одинаковое значение атрибута для разных данных в каталоге

3.17 **показатель качества, показатель Q** (quality score, Q score, Phred score, quality of base calling): Мера вероятности правильного распознавания оснований, обычно выражаемая непосредственно числовым значением.

Примечания

1 Показатель Q определяют по формуле

$$Q = -10\log_{10}(p),$$

где p — предполагаемая вероятность того, что расшифрока оснований будет неверной.

2 Показатель качества, равный 20, представляет собой частоту ошибок 1 из 100 с соответствующей точностью расшифровки 99 %.

3 Показатель качества, равный 30, представляет собой частоту ошибок 1 из 1 000 с соответствующей точностью расшифровки 99,9 %.

4 Более высокие показатели качества указывают на меньшую вероятность ошибки. Более низкие показатели качества могут привести к тому, что значительная часть прочтений окажется непригодной для использования. Низкие показатели качества также могут указывать на ложноположительные варианты расшифровок, что приводит к неточным выводам.

3.18 **исходные данные** (raw data): Первичные данные секвенирования, полученные секвенатором без использования какой-либо предварительной фильтрации на основе программного обеспечения для целей анализа.

[ИСО 20397-2:2021, 3.21]

3.19 **относительная распространенность** (relative abundance): Доля одного микроорганизма, операционной таксономической единицы в общем микробном сообществе определенной среды.

Примечание — Обычно представлена в процентах.

3.20 **требование повторяемости** (repeatability requirement): Требование согласованности при наборе повторяемых условий измерения.

3.21 **скаффолд** (scaffold): Реконструированная геномная последовательность, созданная путем объединения контигов вместе с использованием дополнительной информации об относительном положении и ориентации контигов в геноме.

3.22 **сборка последовательности** (sequence assembly): Обработка, выравнивание и объединение отдельных прочтений секвенирования с целью реконструкции более длинных последовательностей ДНК, целых генов или геномов.

Примечание — К словарной статье: При секвенировании нового генома, где нет референсной последовательности, доступной для выравнивания, прочтения последовательности собираются как контиги, то есть сборка *de-novo*.

3.23 **метагеномное секвенирование методом «дробовика»** (shotgun metagenomics sequencing, shotgun metagenomics): Определение последовательности нуклеотидов геномов нецелевых клеток в сообществах с целью определения состава и функции сообщества.

Примечания

1 Для микробиома метагеномика методом «дробовика» фокусируется на микробных сообществах в определенных средах.

2 Для метагеномного секвенирования методом «дробовика» ДНК извлекается из микробов в образце напрямую без изоляции и культивирования. Затем эта ДНК используется для анализа генетического состава, классификации видов, филогении, функции гена или метаболической сети, или их комбинаций.

3.24 **специализированный атрибут** (specialized attribute): Атрибут, который является уникальным для каждого образца в каталоге.

4 Процесс обработки

Базовый рабочий процесс метагеномики должен включать секвенирование, обработку данных и анализ данных. Обработка данных включает предварительную обработку, контроль качества, сбор данных, профилирование данных и аннотацию, как показано на рисунке 1.

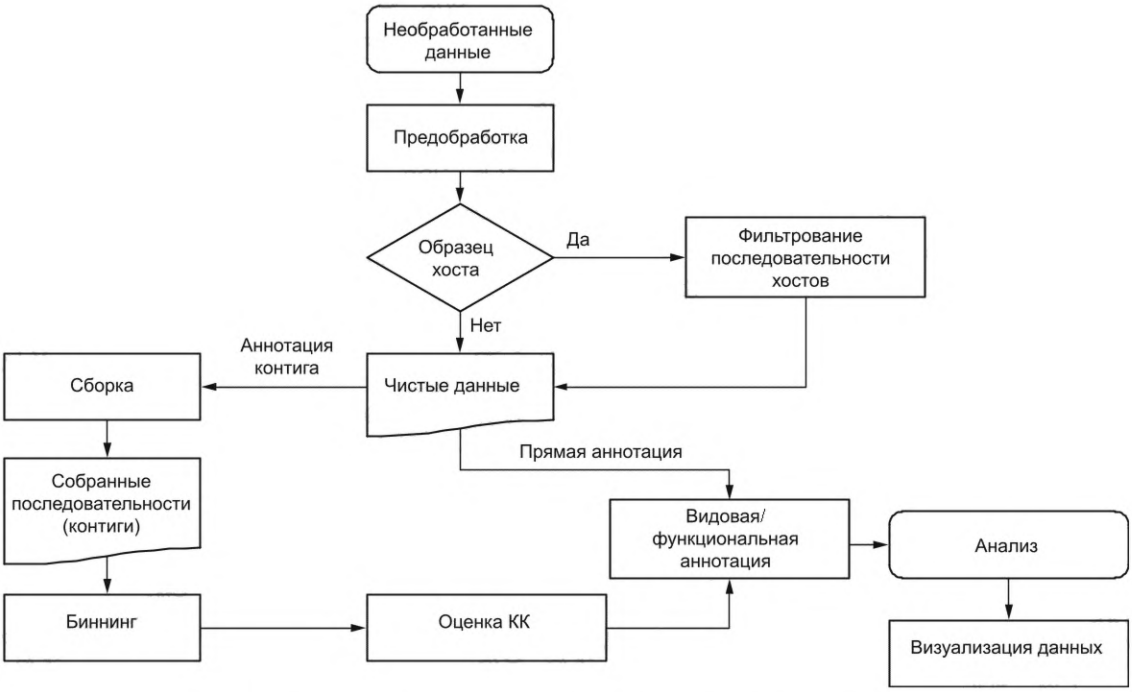

Рисунок 1 — Процесс обработки метагеномных данных

5 Обработка данных

5.1 Требования к оборудованию и программному обеспечению

5.1.1 Необходимо проверить программный конвейер для метагеномной биоинформатики. Приложения для конвейера должны быть заблокированы, включая полный набор инструментов, код, операционную среду и сетевые соединения, которые составляют конвейер, перед его использованием в аналитических целях, таких как оболочка (например, BASH), GNU R и Python. Изменения в любых компонентах конвейера требуют повторной проверки, чтобы гарантировать отсутствие влияния на характеристики производительности конвейера.

5.1.2 Высокопроизводительные вычислительные технологии могут использоваться на любом этапе процесса для обеспечения надлежащего управления и курирования больших коллекций сложных прокариотических и эукариотических геномов, поскольку обработка массивных наборов данных является предпосылкой для аналитики метагеномики NGS.

5.2 Контроль качества последовательности и определение ошибок

5.2.1 Исходные данные метагеномного секвенирования должны первоначально пройти процесс контроля качества (КК), чтобы гарантировать чистый набор данных. Оценка должна соответствовать ИСО 20397-1:2022, раздел 4 и пункт 8.3, а также ИСО 20397-2:2021, пункт 4.3.

5.2.2 Доступные значения качества данных для каждого образца ДНК после секвенирования должны соответствовать следующим требованиям:

a) Q20 ≥ 90 %, более 90 % от значения исходной массы образца должно быть больше 20;

b) Q30 ≥ 80 %, более 80 % от значения исходной массы образца должно быть больше 30.

Вышеуказанные требования применяются только к коротким прочтениям последовательностей ≤ 350 п.о.

5.2.3 Для образцов человека, животного или растения, или всех образцов, хостриды должны быть удалены путем сопоставления с референсом генома человека, животного или растения, таким как UniRef, Единый каталог белков желудочно-кишечного тракта человека. Только чистые данные должны использоваться в дальнейшем биоинформатическом анализе.

5.2.4 Обнаружение и устранение повторов и ошибок секвенирования следует выполнять в качестве первого шага в обработке данных. В процессе устранения должны учитываться следующие факторы и ситуации:

a) несоответствие, вставка или пропуск (инделы) (только при наличии референсного генома) и неопределенные основания (символы N);

b) нераспознаваемая последовательность, которая может быть вызвана, если риды простираются до 3'-конца адаптера, когда целевая последовательность короче;

c) смещения ПЦР при подготовке библиотеки в соответствии с ИСО 20397-1:2022, 5.8.

5.3 Сборка последовательности

5.3.1 Глубина секвенирования должна быть оценена до сборки последовательности, которая должна учитывать сложность образца.

5.3.2 Образцы, не имеющие референсного набора данных генома, такие как образцы почвы или океана, должны использовать сборку последовательности.

5.3.3 Контиги и/или скаффолды, которые напрямую получены из фрагментов последовательности без какой-либо ссылки, следует рассматривать как собирание *de-novo*.

5.3.4 Выбор программного обеспечения для собирания последовательности должен зависеть от относительной важности таких параметров как точность, размер контигов, тип входных данных и доступные вычислительные ресурсы.

5.3.5 Неизбыточный каталог генов может быть получен путем прогнозирования генов из собранных контигов. Для хорошо охарактеризованных микробиомов, например кишечно-инженерных, для быстрой идентификации и количественной оценки данных из метагеномного секвенирования можно использовать надежный каталог генов (например, интегрированный каталог генов (IGC)).

5.3.6 Созданные сборы следует оценивать для определения их качества, например, QUAST.

6 Анализ данных

6.1 Аннотация

6.1.1 Должны быть описаны методы аннотации. Количество выбранных референсных геномов, а также референсные геномы или референсная база данных, используемые для аннотации, должны быть задокументированы.

6.1.2 Методы таксономического профиля следует выбирать в соответствии с данными и потребностями приложения для получения таксономического профиля более высокого уровня (например, вида, рода, отряда, типа), включая метагеномные группы сцепления (MLG), метагеномные кластеры (MGC) или метагеномные виды (MGS).

6.1.3 Профилирование таксономии должно основываться на справочных базах данных, таких как RefSeq полные геномы (RefSeq CG) для микробных видов и базы данных BLAST®[1)] для высококачественных последовательностей нуклеотидов и белков. Точность классификации, скорость и вычислительные требования должны быть приняты во внимание при выборе инструментов таксономической классификации.

6.1.4 Если профиль получен методом собирания последовательности *de-novo*, информация о виде должна быть идентифицирована, когда определяется выравнивание со сходством последовательности более 97 % и охватом более 90 % наиболее связанной справочной базы данных.

6.1.5 Для подходов, основанных на чтении, чтение должно выполнять сопоставление с неизбыточной базой данных последовательностей белков (NR) для данных таксономии (например, BLAST®, Diamond или Last) или маркерных генов после слияния чтений.

6.1.6 Метагеномные профили должны быть аннотированы на различных уровнях в соответствии с референсной аннотацией, т. е. вид, род или выше.

[1)] BLAST® является торговой маркой продукта, поставляемого Национальным центром биотехнологической информации (США). Эта информация предоставляется для удобства пользователей настоящего стандарта и не является одобрением со стороны ИСО названного продукта. Эквивалентные продукты могут быть использованы, если можно продемонстрировать, что они приводят к тем же результатам.

6.2 Расчет относительной распространенности видов

6.2.1 Анализ видов

6.2.1.1 Метод расчета относительной распространенности должен быть определен и реализован для удовлетворения требования повторяемости. Метод расчета относительной распостраненности должен быть задокументирован.

6.2.1.2 Инструменты расчета должны быть выбраны с учетом отражения фактической относительной распостраненности целевой операционной таксономической единицы в образце.

6.2.2 Анализ генов

6.2.2.1 Таблицу распространенности генов можно создать с помощью инструментов на основе выравнивания или методов без выравнивания.

6.2.2.2 Относительное распределение распространенности на уровне генов можно получить путем сравнения очищенных данных с собранным набором генов или соответствующей справочной базой данных.

6.2.2.3 Для получения рабочей таксономической единицы необходимо выполнить суперпозицию относительной распространенности последовательности генов одного и того же вида.

7 Архив данных и метаданные

7.1 Исходные данные

7.1.1 Объем данных секвенирования должен быть оценен для получения насыщенной генной информации в исследованиях ассоциаций по всему метагеному (MWAS).

7.1.2 Независимо от источника образца, формат данных секвенирования должен быть одинаковым для каждой последовательности. Хранимый формат данных последовательности должен быть стандартизован, сохраняя информацию о биологических последовательностях (обычно последовательности нуклеиновых кислот) или качество их секвенирования; ИСО 20691:2022, А.2 включает информацию о форматах данных для методов OMICS, биохимии и молекулярной биологии.

7.2 Аналитические данные секвенирования

7.2.1 Аналитические данные секвенирования должны включать количество прочтений, общую относительную распространенность рода, общую относительную распространенность функции и общую относительную распространенность вида.

7.2.2 Аналитические данные секвенирования должны быть отформатированы в в виде текста и содержать заголовок и строки данных (примеры приведены в приложении А), такие как общая относительная распространенность рода (см. таблицу А.1), общая относительная распространенность вида (см. таблицу А.2), индекс бактериального разнообразия (см. таблицу А.3), индекс энтеротипов (см. таблицу А.4), относительная распространенность бактерий (см. таблицу А.5), общая относительная распространенность генов (см. таблицу А.6). Рекомендуется сжатый формат, например GZ или Zstd.

7.3 Каталог данных и архив

7.3.1 Общие положения

Описание данных и метаданные должны соответствовать ИСО 20691:2022, раздел 4 и раздел 5.

7.3.2 Каталог элементов данных

7.3.2.1 Каталог элементов данных должен включать общедоступные атрибуты и специализированные атрибуты.

7.3.2.2 Публичные атрибуты такие же, как и другие геномные данные, требующие следующих записей: версия, орган регистрации, режим классификации, ответственная организация и организация-отправитель.

7.3.2.3 Специализированные атрибуты должны включать информацию о секвенировании, информацию о биологическом анализе, метагеномную информацию и информацию о контроле качества. Примеры каждого атрибута приведены в приложении В (см. таблицы В.1, В.2 и В.3).

7.3.3 Архивирование данных

Каталог метагеномных данных может быть разделен на исходный каталог данных и каталог результатов анализа. Оба должны быть многоуровневыми каталогами. Каждый уровень каталога должен включать идентификатор каталога, имя каталога, определение каталога и родительский каталог.

7.4 Метаданные

7.4.1 Метаданные могут структурно храниться в виде плоского файла данных во время обработки данных, включая фенотипическую информацию, информацию о контроле качества, преаналитическую информацию (метод сбора, использование стабилизатора(ов), задержки между сбором и обработкой, временные условия хранения, метод выделения ДНК и параметры метода, а кроме того метод контроля качества ДНК и результаты) и т. д.

7.4.2 Должна быть проведена проверка согласованности метаданных и данных секвенирования, чтобы проверить, соответствуют ли образцы данных секвенирования образцам, описанным метаданными, например пропуски данных, дублирования, сопоставление ошибок и маркировка ошибок.

7.4.3 Каталог метагеномных метаданных должен включать каталог информации о последовательностях, биоинформатический аналитический каталог и метагеномный каталог. Эти каталоги должны включать следующую информацию:

a) каталог информации о последовательностях должен включать метаданные, описывающие ход секвенирования, такие как имя списка задач секвенирования, тип секвенирования, имя платформы секвенирования, идентификаторы секвенатора, имя секвенатора и номер чипа;

b) биоинформационный аналитический каталог должен включать метаданные, описывающие программное обеспечение для биоинформатического анализа, например версию и параметры;

c) метагеномный каталог должен включать метаданные, описывающие процесс и результаты метагеномного анализа, например код, имя, значение, длину, описание целостности, тип, род, вид и тип кишечника.

7.4.4 Данные секвенирования должны иметь метаданные для обеспечения прослеживаемости и целостности информации об образце. Информация о метаданных может соответствовать «Минимальной информации о последовательности генома (MIGS)» и «Минимальной информации о последовательности метагенома (MIMS)» [7].

7.4.5 Должна быть проведена проверка для сохранения соответствия между метаданными и данными секвенирования.

Приложение А
(справочное)

Примеры форматов данных

А.1 Общая относительная распространенность рода

А.1.1 Общие положения

Данные могут быть в текстовом формате и отображаться со строкой заголовка и строкой данных. Общая информация об относительном распространении рода в метагеномике содержится в строке данных.

А.1.2 Строка заголовка

Строка заголовка имеет N + 2 (N — количество референсных образцов) фиксированных полей, каждое поле связано с помощью табуляции, как показано ниже:

[1]. # Научное_название
[2]. значение_относительной_распространенности_образца
[3]. значение_относительной_распространенности_референсного_образца_1
……
[N]. значение_относительной_распространенности_референсного_образца_N

А.1.3 Строка данных

Каждая строка имеет N + 2 (N — количество референсных образцов) фиксированных полей, каждое поле связано с помощью табуляции, а нулевое значение каждого поля использует точку «.» для представления. Соответствующая информация о поле выглядит следующим образом.

[1]. научное_название: Название рода бактерий, тип строки и использование базы данных NCBI Taxonomy для описания, см. в gost.e-ecolog.ru

[2]. значение_относительной_распространенности_образца: Общие значения относительной распространенности рода образца, показаны в виде десятичной дроби.

[3]. значение_относительной_распространенности_референсного_образца_1: Общие значения относительной распространенности рода референсного образца 1, показаны в виде десятичной дроби.

……

[N]. значение_относительной_распространенности_референсного_образца_N: Общие значения относительной распространенности рода референсного образца N, показаны в виде десятичной дроби.

А.1.4 Таблица примеров

Таблица А.1 — Примеры общей относительной распространенности рода

Заголовок списка	Пример 1	Пример 2
# научное_название	Abiotrophia	Acetivibrio
значение_относительной_распространенности_образца	3,079 903e-06	4,047 342e-05
значение_относительной_распространенности_референсного_образца_1	3,227 675e-06	1,407 130e-05
……	……	……
значение_относительной_распространенности_референсного_образца_N	0	4,641 824e-06

А.2 Общая относительная распространенность видов

А.2.1 Общие положения

Данные могут быть в текстовом формате и отображаться со строкой заголовка и строкой данных. Общая информация об относительном обилии видов в метагеномике содержится в строке данных.

А.2.2 Строка заголовка

Строка заголовка имеет N + 2 (N — количество референсных образцов) фиксированных полей, каждое поле связано с вкладкой, как показано ниже:

[1]. # научное_название.

[2]. значение_относительной_распространенности_образца.
[3]. значение_относительной_распространенности_референсного_образца_1.
……
[N]. значение_относительной_распространенности_референсного_образца_N.

А.2.3 Строка данных

Каждая строка имеет N + 2 (N — количество референсных образцов) фиксированных полей, каждое поле связано с вкладкой, а нулевое значение каждого поля использует точку «.» для представления. Соответствующая информация о поле выглядит следующим образом:

[1]. научное_название: Название вида бактерий, тип строки и использование базы данных таксономии NCBI для описания, см. в gost.e-ecolog.ru

[2]. значение_относительной_распространенности_образца: Общие относительные значения распространенности видов образца, показаны в виде десятичной дроби.

[3]. значение_относительной_распространенности_референсного_образца_1: Общие значения распространенности видов референтного образца 1, показаны в виде десятичной дроби.

……

[N]. значение_относительной_распространенности_референсного_образца_N: Общие относительные значения распространенности видов референтного образца N, показаны в десятичной дроби.

А.2.4 Таблица примеров

Таблица А.2 — Примеры общей относительной распространенности видов

Заголовок списка	Пример 1	Пример 2
# научное_название	Abiotrophia defectiva	Achromobacter piechaudii
значение_относительной_распространенности_образца	1,163 749 42e-05	0
значение_относительной_распространенности_референсного_образца_1	3,737 392 6e-06	0
……	……	……
значение_относительной_распространенности_референсного_образца_N	2,180 657e-06	1,444 14e-07

А.3 Индекс бактериального разнообразия

А.3.1 Общие положения

Данные могут быть в текстовом формате и отображаться со строкой заголовка и строкой данных. Информация об индексе бактериального разнообразия в метагеномике содержится в строке данных.

А.3.2 Строка заголовка

Строка заголовка имеет 4 фиксированных поля, каждое поле связано с вкладкой, как показано ниже:
[1]. # имя_образца.
[2]. количество_генов.
[3]. индекс_Шеннона.
[4]. соотношение_показателей_разнообразия.

А.3.3 Строка данных

Каждая строка имеет четыре фиксированных поля, каждое поле соединено табуляцией, а нулевое значение каждого поля использует точку «.» для представления. Соответствующая информация о полях следующая:

[1]. # имя_образца: имя образца, тип строки.
[2]. количество_генов: общее количество генов, округленное число.
[3]. индекс_Шеннона: Индекс разнообразия Шеннона, отображается в виде десятичной дроби.
[4]. соотношение_показателей_разнообразия: доля группы ниже индекса Шеннона, отображается в виде десятичной дроби, значения находятся в диапазоне от 0 до 1.

А.3.4 Таблица примеров

Таблица А.3 — Примеры индекса бактериального разнообразия

Заголовок списка	Пример 1	Пример 2
# имя_образца	9 387	9 388
количество_генов	769 324	844 883
индекс_Шеннона	12,380 35	12,396 23
соотношение_показателей_разнообразия	0,965 2	0,970 1

А.4 Индекс энтеротипов

А.4.1 Общие положения

Данные могут быть в текстовом формате и отображаться со строкой заголовка и строкой данных. Информация об энтеротипах в метагеномике содержится в строке данных.

А.4.2 Строка заголовка

Строка заголовка имеет два фиксированных поля, каждое поле соединено табуляцией:

[1]. # имя_образца.

[2]. структура_кишечника.

А.4.3 Строка данных

Каждая строка имеет 2 фиксированных поля, каждое поле соединено табуляцией, а нулевое значение каждого поля использует точку «.» для представления. Соответствующая информация о полях выглядит следующим образом:

[1]. # имя_образца: имя образца, тип строки.

[2]. структура_кишечника: Энтеротип, тип строки.

А.4.4 Таблица примеров

Таблица А.4 — Примеры индекса Энтеротипа

Заголовок списка	Пример 1	Пример 2
# имя_образца	5 002	5 003
структура_кишечника	Prevotella	Bacteroides

А.5 Относительная распространенность бактерий

А.5.1 Общие положения

Данные могут быть в текстовом формате и показаны со строкой заголовка и строкой данных. Информация об относительном содержании бактерий в метагеномике содержится в строке данных.

А.5.2 Строка заголовка

Строка заголовка имеет N + 2 (N — количество референсных образцов) фиксированных полей, каждое поле связано с вкладкой, как показано ниже:

[1]. # название_бактерии.

[2]. значение_относительной_распространенности_образца .

[3]. значение_относительной_распространенности_референсного_образца_1.

……

[N]. значение_относительной_распространенности_референсного_образца_N.

А.5.3 Строка данных

Каждая строка имеет N + 2 (N — количество референсных образцов) фиксированных полей, каждое поле соединено табуляцией, а нулевое значение каждого поля использует точку «.» для представления. Соответствующая информация о поле выглядит следующим образом:

[1]. НАЗВАНИЕ_БАКТЕРИИ: Название бактерий, тип строки.

[2]. значение_относительной_распространенности_образца: Значения относительной распространенности рода бактерий в образце, отображаются в виде десятичной дроби.

[3]. значение_относительной_распространенности_референсного_образца_1: Значения относительной распространенности рода бактерий в референсном 1, отображаются в виде десятичной дроби.

……

[N]. значение_относительной_распространенности_референсного_образца_N: Значения относительной распространенности рода бактерий в референсном образце N, отображаются в виде десятичной дроби.

А.5.4 Таблица примеров

Таблица А.5 — Примеры относительной распространенности бактерий

Заголовок списка	Пример 1	Пример 2
# название_бактерии	Bacteroides uniformis	Bifidobacterium adolescentis
значение_относительной_распространенности_образца	0,003 039 194 6	0,000 125 262 5
значение_относительной_распространенности_референсного_образца_1	0,001 694 538 5	$6{,}003\ 822\ 041 \times e^{-05}$
……	……	……
значение_относительной_распространенности_референсного_образца_N	0,002 179 311 8	$3{,}318\ 616\ 67 \times e^{-05}$

А.6 Общая относительная распространенность генов

А.6.1 Общие положения

Данные могут быть в текстовом формате и отображаться со строкой заголовка и строкой данных. Общая информация об относительном распространении генов в метагеномике содержится в строке данных.

А.6.2 Строка заголовка

Строка заголовка имеет N + 2 (N — количество референсных образцов) фиксированных полей, каждое поле связано с вкладкой, как показано ниже:

[1]. # ID_гена.

[2]. значение_относительной_распространенности_образца.

[3]. значение_относительной_распространенности_референсного_образца_1.

……

[N]. значение_относительной_распространенности_референсного_образца_N.

А.6.3 Строка данных

Каждая строка имеет N + 2 (N — количество референтных образцов) фиксированных полей, каждое поле соединено табуляцией, а нулевое значение каждого поля использует точку «.» для представления. Соответствующая информация о поле выглядит следующим образом.

[1]. ID_ГЕНА: Идентификаторы гена, тип строки, можно использовать идентификатор гена NCBI Entrez, см.в

gost.e-ecolog.ru

[2]. значение_относительной_распространенности_образца: Общие значения относительной распространенности генов образца, показаны в виде десятичной дроби.

[3]. значение_относительной_распространенности_референсного_образца_1: Общие значения относительной распространенности генов референтного образца 1, показаны в виде десятичной дроби.

……

[N]. значение_относительной_распространенности_референсного_образца_N: Общие значения относительной распространенности генов референтного образца N, показаны в виде десятичной дроби.

А.6.4 Таблица примеров

Таблица А.6 — Примеры общей относительной распространенности генов

Заголовок списка	Пример 1	Пример 2
# ID_ГЕНА	25 604	56 324
значение_относительной_распространенности_образца	0,003 039 194 6	0,000 125 262 5
значение_относительной_распространенности_референсного_образца_1	0,001 694 538 5	6,003 822 041 × e^{-05}
.....	.....	.....
значение_относительной_распространенности_референсного_образца_N	0,002 179 311 8	3,318 616 67 × e^{-05}

Приложение В
(справочное)

Каталог элементов данных

Таблица В.1 — Каталог секвенирования информационных элементов

Наименование	Защита	Ед.	Тип	Допустимое значение
Идентификатор списка задач	Нет		S	
Название списка задач	Нет		S	
Тип секвенирования	Нет		S	
Название платформы секвенирования	Нет		S	
Идентификатор секвенатора	Нет		S	
Название секвенатора	Нет		S	
Название библиотеки	Нет		S	
Номер библиотеки	Нет		S	
Стратегия библиотеки	Нет		S	
Чип №	Нет		S	
Дорожка №	Нет		S	
Машина №	Нет		S	
Путь хранения исходных данных секвенирования	Нет		S	
Уникальный номер файлов FASTQ	Нет		S	
Время начала секвенирования	Нет		DT	
Время завершения секвенирования	Нет		DT	
Место завершения	Нет		S	
Примечание — Тип «S» относится к «Строке», «DT» относится к «Дате и времени».				

Таблица В.2 — Каталог анализа элементов биологической информации

Наименование	Защита	Ед.	Тип	Допустимое значение
Путь хранения данных	Нет		S	
Наименование ПО	Нет		S	
Версия ПО	Нет		S	
Параметр ПО	Нет		S	
Примечание — Тип «S» относится к «Строке».				

Таблица В.3 — Каталог метагеномных элементов

Наименование	Защита	Ед.	Тип	Допустимые значения
Метагеномный код	Нет		S	
Название гена	Нет		S	
Значение гена	Нет		N	
Длина гена	Нет		N	
Описание целостности	Нет		S	
Происхождение микробного сообщества	Нет		S	
Классификация филума	Нет		S	
Значение количества микробного филума	Нет		N	
Классификация рода	Нет		S	
Значение микробных видов	Нет		N	
Классификация видов	Нет		S	
Значение количества микробных видов	Нет		N	
Аннотация (аннотация KO)	Нет		S	
Аннотация линейных гомологичных генов	Нет		S	
Частота, с которой ген присутствует во всех образцах	Нет		N	
Частота, с которой ген присутствует во всех особях	Нет		N	
Функциональная классификация аннотации KEGG	Нет		S	
Функциональная классификация аннотации eggNOG	Нет		S	
Оцениваемое происхождение сообщества	Нет		S	
Функция микробного сообщества	Нет		S	
Примечание — Тип «S» относится к «Строке», «N» относится к «Числу».				

Приложение ДА
(справочное)

Сведения о соответствии ссылочного международного стандарта национальному стандарту

Таблица ДА.1

Обозначение ссылочного международного стандарта	Степень соответствия	Обозначение и наименование соответствующего национального стандарта
ISO 20397-1:2022	IDT	ГОСТ Р ИСО 20397-1—2025 «Биотехнология. Массовое параллельное секвенирование. Часть 1. Подготовка нуклеиновой кислоты и библиотеки»
Примечание — В настоящей таблице использовано следующее условное обозначение степени соответствия стандарта: - IDT — идентичный стандарт.		

Библиография

[1] ISO/IEC 11179-1:2015[1)] Information technology — Metadata registries (MDR) — Part 1: Framework (Информационная технология. Регистры метаданных (РМД). Часть 1. Основные положения)

[2] ISO 17115:2007[2)] Health informatics — Vocabulary for terminological systems (Информатика в здравоохранении. Словарь терминологических систем)

[3] ISO 20691:2022 Biotechnology — Requirements for data formatting and description in the life sciences (Биотехнология. Требования к форматированию и описанию данных в области медико-биологических наук)

[4] ISO/TS 21089:2018 Health informatics — Trusted end-to-end information flows (Информатизация здоровья. Потоки конфиденциальной сквозной информации)

[5] ISO 20397-2:2021 Biotechnology — Massively parallel sequencing — Part 2: Quality evaluation of sequencing data (Биотехнология. Массовое параллельное секвенирование. Часть 2. Оценка качества данных секвенирования)

[6] ISO 21962:2003 Space data and information transfer systems — Data entity dictionary specification language (DEDSL) — PVL syntax (Космические системы передачи данных и информации. Язык спецификаций для словаря информационного объекта. Синтаксис PVL)

[7] Field D, Garrity G, Gray T et al. The minimum information about a genome sequence (MIGS) specification. Nat Biotechnol. 2008;26(5):541-7. Nature group. Available at: doi: 10.1038/nbt1360

[8] Simon H. Ye., Katherine J. Siddle, Daniel J. Park et al. Benchmarking Metagenomics Tools for Taxonomic Classification. Cell. 2019; 8;178(4):779-794. Available at: doi: 10.1016/j.cell.2019.07.010.

[9] Amos, G.C.A., Logan A., Anwar S. et al. Developing standards for the microbiome field. Microbiome 8, 98 (2020). Available at: doi: 10.1186/s40168-020-00856-3

[10] Yong-Xin Liu, Yuan Qin, Tong Chen et al. A practical guide to amplicon and metagenomic analysis of microbiome data. Protein Cell. 2021;12(5):315—330. Available at: doi: 10.1007/s13238-020-00724-8

[1)] В настоящее время действует ISO/IEC 11179-1:2023.

[2)] В настоящее время действует ISO 17115:2020 «Health informatics — Representation of categorial structures of terminology (CatStructure)».

УДК 658:5:006.354 ОКС 07.080

Ключевые слова: ДНК, массовое параллельное секвенирование, метагеномное секвенирование, метод «дробовика», формат данных

Редактор *Н.А. Аргунова*
Технический редактор *И.Е. Черепкова*
Корректор *С.И. Фирсова*
Компьютерная верстка *А.Н. Золотаревой*

Сдано в набор 23.03.2026. Подписано в печать 15.04.2026. Формат 60×84⅛. Гарнитура Ариал.
Усл. печ. л. 2,32. Уч.-изд. л. 1,97.

Подготовлено на основе электронной версии, предоставленной разработчиком стандарта

Создано в единичном исполнении в ФГБУ «Институт стандартизации»
для комплектования Федерального информационного фонда стандартов, 117418 Москва, Нахимовский пр-т, д. 31, к. 2.
gost.e-ecolog.ru gost.e-ecolog.ru

gost.e-ecolog.ru