Биоинформатика (1)
основные понятия, установка и менеджмент программ

Аспирант АБИБ ЮФУ
2 года обучения
Дёмин К. А.

Ростов-на-Дону, 2024

Понятия, которые необходимо изучить

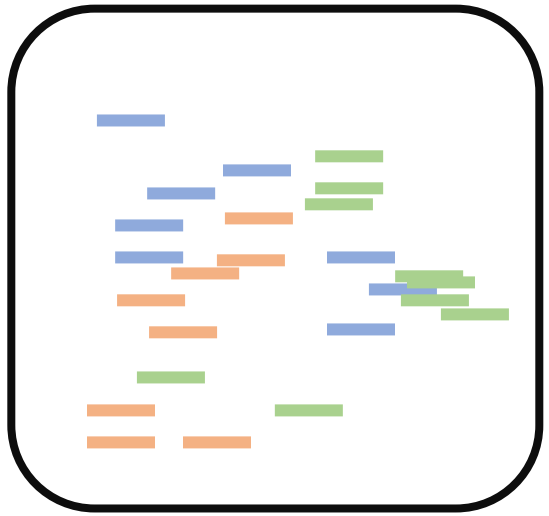
- Рид, контиг, скаффолд, бин, MAG, геном,
- Секвенирование и секвенаторы
- Ампликонные методы, шотган-методы, методы получения длинных прочтений
- Выравнивание (alignment), покрытие (coverage), сборка (assembly), биннинг (binning), картирование (mapping)

1. Рид (read): Короткая последовательность нуклеотидов, полученная в результате секвенирования ДНК или РНК. Риды являются исходными данными для сборки генома.
2. Сборка (assembly): Процесс объединения коротких ридов в более длинные последовательности (контиги и скаффолды) для реконструкции исходного генома. Сборка генома может быть сложной задачей из-за наличия повторов и ошибок секвенирования.
3. Контиг (contig): Непрерывная последовательность нуклеотидов, полученная путем объединения перекрывающихся ридов в процессе сборки генома. Контиги представляют собой более длинные фрагменты генома по сравнению с ридами.
4. Скаффолд (scaffold): Упорядоченная последовательность контигов, разделенных промежутками (гэпами), которые представляют собой неизвестные участки генома. Скаффолды получаются путем объединения контигов с использованием дополнительной информации, такой как парные риды или длинные риды.
5. Биннинг (binning): Процесс группировки контигов или скаффолдов в бины на основе сходства их композиционных характеристик (например, частоты k-меров) и геномных свойств (например, покрытия). Биннинг используется для разделения метагеномных данных на отдельные геномы.
6. Бин (bin): Группа контигов или скаффолдов, которые предположительно принадлежат одному организму или группе близкородственных организмов. Бины формируются в процессе биннинга на основе сходства композиционных и геномных характеристик.
7. MAG (Metagenome-Assembled Genome): Геном, реконструированный из метагеномных данных путем сборки и биннинга. MAG представляет собой геном отдельного организма, извлеченный из метагенома.
8. Геном (genome): Полная генетическая информация организма, содержащаяся в его ДНК или РНК. Геном включает в себя все гены и некодирующие последовательности.
9. Выравнивание (alignment): Процесс сравнения и сопоставления двух или более последовательностей нуклеотидов или аминокислот для определения сходства и различий между ними. Выравнивание позволяет идентифицировать гомологичные участки и изучать эволюционные связи.
10. Покрытие (coverage): Показатель, отражающий количество ридов, которые приходятся на каждую позицию в геноме. Более высокое покрытие означает большую надежность сборки и точность определения последовательности.
11. Картирование (mapping): Процесс выравнивания ридов на референсный геном или сборку для определения их положения и идентификации различий (например, мутаций или структурных вариантов). Картирование позволяет изучать геномные вариации и проводить анализ экспрессии генов.

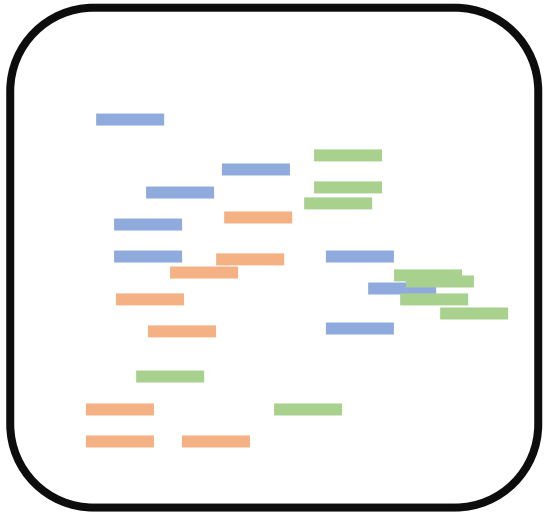
Секвенирование и виды метагеномов

- Ампликонный метагеном (нужна ПЦР) (Ion-Torrent, Illumina, MGI, PacBio, Nanopore)
- Шотган-метагеном (Illumina, MGI)
- Метагеном длинных прочтений (Nanopore, PacBio)

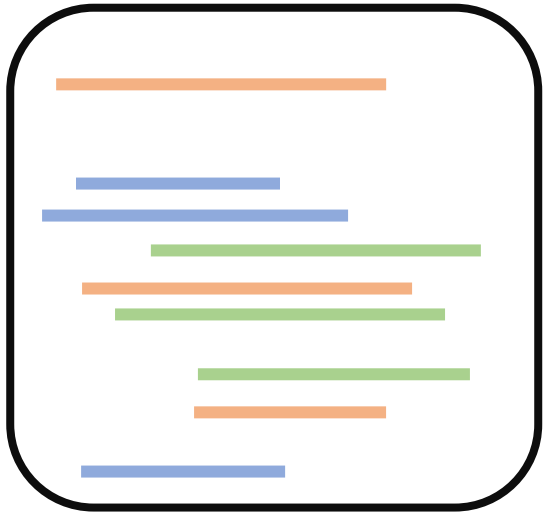
Короткие или длинные прочтения всех копий одного гена

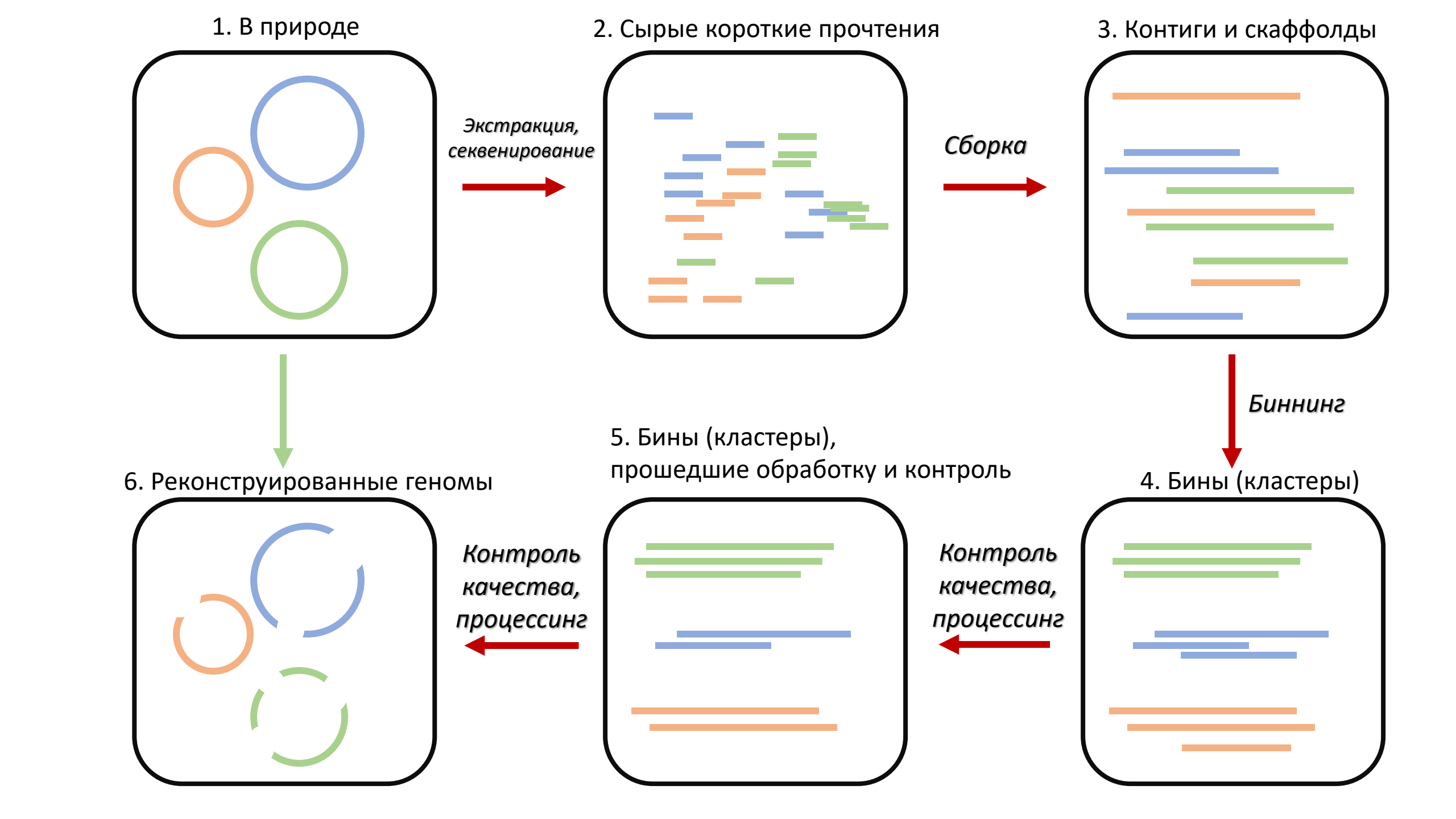


Короткие прочтения всей ДНК



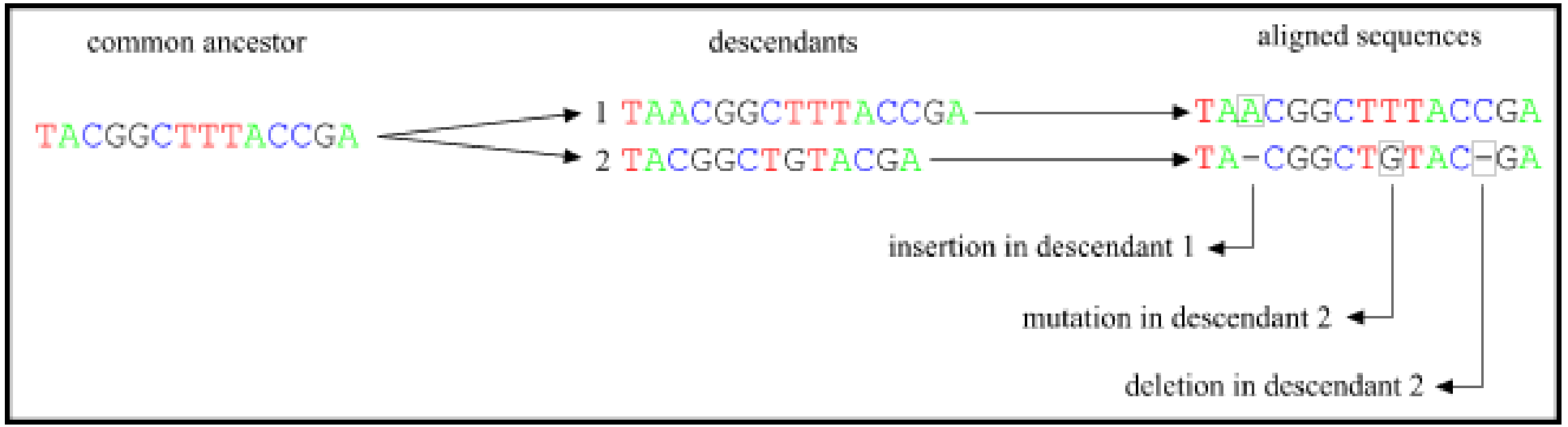
Длинные прочтения всей ДНК

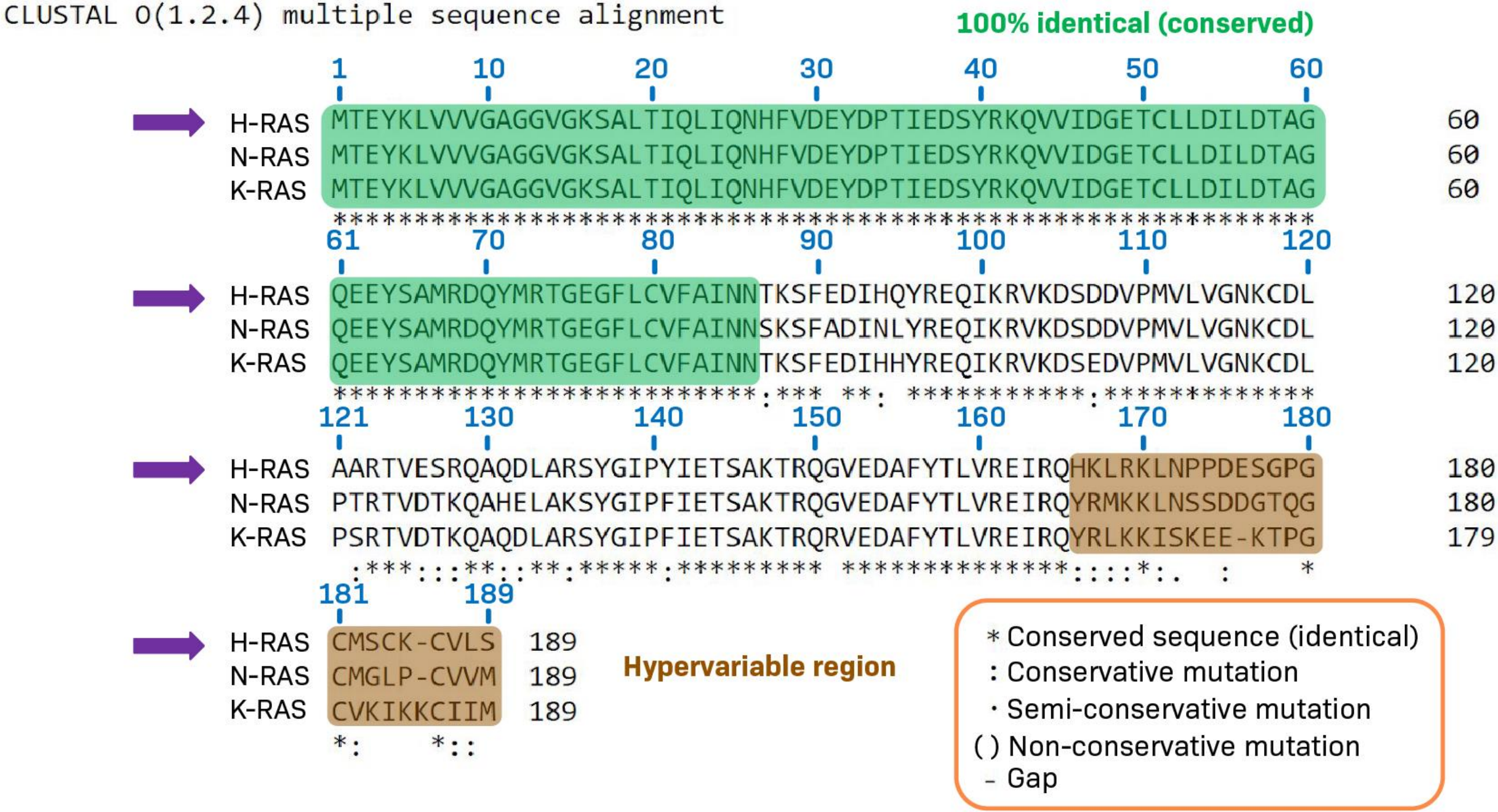



1. В природе
2. Сырые короткие прочтения
3. Контиги и скаффолды
Экстракция,
секвенирование
Сборка
Биннинг
4. Бины (кластеры)
5. Бины (кластеры),
прошедшие обработку и контроль
6. Реконструированные геномы
Контроль
качества,
процессинг
Контроль
качества,
процессинг

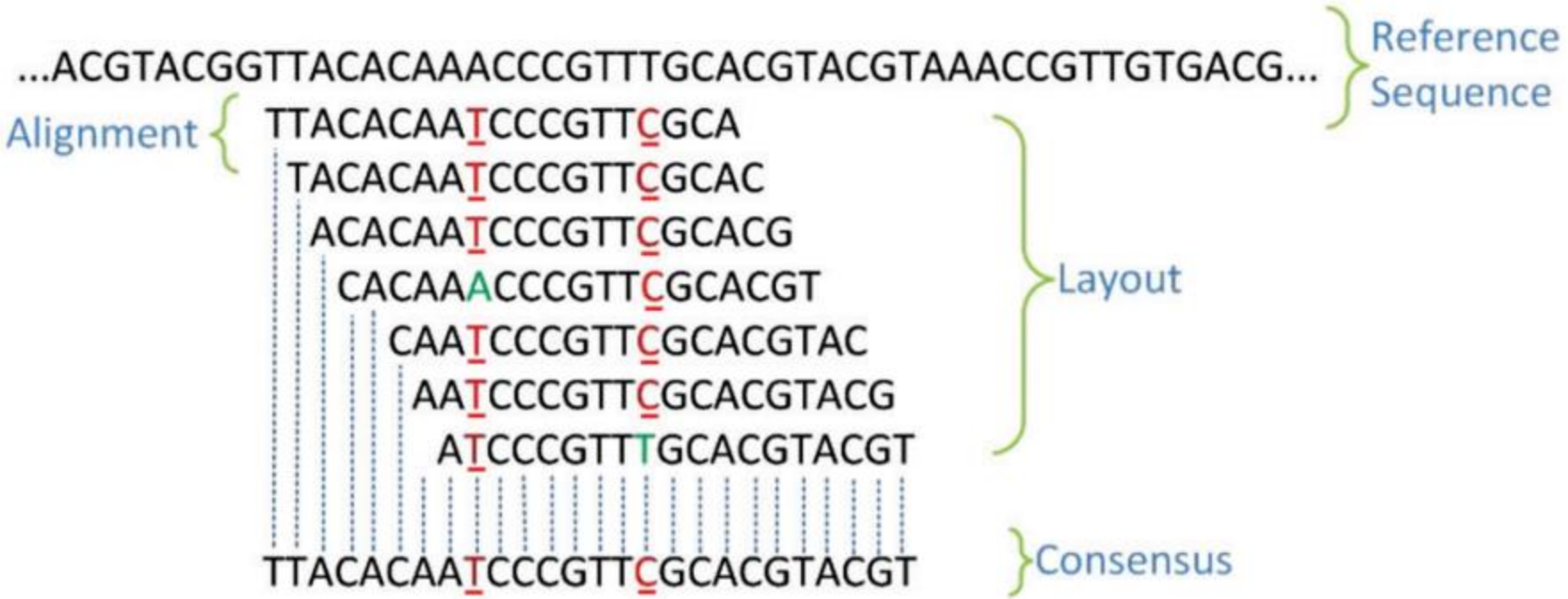

1. В природе
2. Сырые короткие прочтения
3. Контиги и скаффолды
Экстракция, секвенирование
Сборка
Длинные прочтения
Биннинг
4. Бины (кластеры)
5. Бины (кластеры), прошедшие обработку и контроль
6. Реконструированные геномы
Контроль качества, процессинг
Контроль качества, процессинг

Выравнивание




CLUSTAL O(1.2.4) multiple sequence alignment
100% identical (conserved)
1 10 20 30 40 50 60
H-RAS MTEYKLVVVGAGGVGKSALTIQLIQNHFVDEYDPTIEDSYRKQVVIDGETCLLDILDTAG 60
N-RAS MTEYKLVVVGAGGVGKSALTIQLIQNHFVDEYDPTIEDSYRKQVVIDGETCLLDILDTAG 60
K-RAS MTEYKLVVVGAGGVGKSALTIQLIQNHFVDEYDPTIEDSYRKQVVIDGETCLLDILDTAG 60
61 70 80 90 100 110 120
H-RAS QEEYSAMRDQYMRTGEGFLCVFAINNTKSFEDIHQYREQIKRVKDSDDVPMVLVGNKCDL 120
N-RAS QEEYSAMRDQYMRTGEGFLCVFAINNSKSFADINLYREQIKRVKDSDDVPMVLVGNKCDL 120
K-RAS QEEYSAMRDQYMRTGEGFLCVFAINNTKSFEDIHHYREQIKRVKDSEDVPMVLVGNKCDL 120
121 130 140 150 160 170 180
H-RAS AARTVESRQAQDLARSYGIPYIETSAKTRQGVEDAFYTLVREIRQHKLRKLNPPDESGPG 180
N-RAS PTRTVDTKQAHELAKSYGIPFIETSAKTRQGVEDAFYTLVREIRQYRMKKLNSSDDGTQG 180
K-RAS PSRTVDTKQAQDLARSYGIPFIETSAKTRQRVEDAFYTLVREIRQYRLKKISKEE-KTPG 179
181 189
H-RAS CMSCK-CVLS 189
N-RAS CMGLP-CVVM 189
K-RAS CVKIKKCIIM 189
Hypervariable region
* Conserved sequence (identical)
: Conservative mutation
· Semi-conservative mutation
( ) Non-conservative mutation
- Gap

Покрытие контигов



Алгоритмы выравнивания

1. Needleman-Wunsch Algorithm: A dynamic programming algorithm for global pairwise alignment.
2. Smith-Waterman Algorithm: A dynamic programming algorithm for local pairwise alignment.
3. BLAST (Basic Local Alignment Search Tool): A heuristic algorithm for finding local similarities between sequences.
4. Progressive Alignment: A method for multiple sequence alignment that builds a guide tree and aligns sequences in a stepwise manner (e.g., Clustal).
5. Iterative Refinement: A method for improving multiple sequence alignments by repeatedly dividing sequences into subgroups, realigning them, and combining the results (e.g., MUSCLE, MAFFT).

Файлы и их форматы в биоинформатике

1) .txt – text
2) .csv – *comma-separated value*
3) .tsv – *tab-separated value*
4) .fasta biologic sequence
5) .fastq – *fasta with quality scores*
6) .faa – fasta aminoacids
7) .fna – fasta nucleic acids
8) .cds – conding sequences
9) .sam – sequence alignment map
10) .bam – binary alignment map
11) .gff – general feature format
12) .bed – browser extensible data

Базовые программы и уровни сложности

Уровень 1: Сырые прочтения из секвенатора

Уровень 2: Длинные фрагменты собранные из сырых прочтений

Уровень 3: Длинные фрагменты, сгруппированные по схожести

Уровень 4: Реконструированные геномы

Уровень 5: Микробное сообщество

1. FastQC, Seqkit, Trimmomatic
2. Kraken2/Centrifuge/Kaiju
3. Samtools
4. BWA/BWA2/Bowtie2/minimap2
5. Spades, MEGAHIT
6. CoverM
7. Maxbin2, Metabat2, CONCOCT
8. DAS_Tool, Meta_WRAP, MAG_Purify
9. Prodigal, Emapper, Mafft, CheckM, GTDB-tk, IQ-TREE

1. FastQC, Trimmomatic	Контроль качества
2. Seqkit, Samtools	Манипуляция сиквенсами
3. Kraken2/Centrifuge/Kaiju	Классификация прочтений
4. BWA/BWA2/Bowtie2/minimap2	Множественное выравнивание
5. Spades, MEGAHIT	Сборка контигов из прочтений
6. CoverM	Оценка покрытия контигов прочтениями
7. Maxbin2, Metabat2, CONCOCT	Кластеризация контигов (*биннинг*)
8. DAS_Tool, Meta_WRAP,	Мастер-пакеты с множеством функций
9. MAG_Purify	Чистка бинов
10. CheckM	Оценка качества сборки бинов
11. Prodigal,	Предсказание генов,
12. Emapper,	Аннотирование генов,
13. GTDB-tk	Таксономическая классификация геномов
14. Mafft, IQ-TREE	Построение филогенетических древ

Установка программ

1. Установка через встроенный магазин приложений
2. Базовый менеджер программ и пакетов линукса (Apt, Yum)
3. Build-from-source
4. pip
5. Менеджеры микроокружений

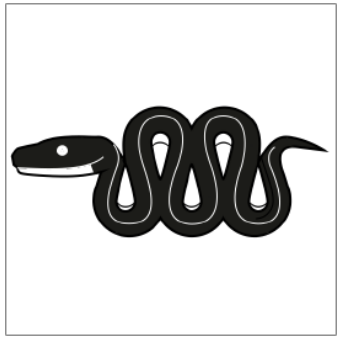
Менеджеры микроокружений

Docker

Singularity

Conda (bioconda)

Mamba (bioconda)



Mamba-микроокружения