

Факультет биоинженерии и биоинформатики МГУ  
II курс

# Сборка чтений

С.А. Спирин  
9 декабря 2025

# Проблема сборки

## Сборка на уже известный геном

(например, чтобы изучать различия между ДНК разных людей)

## Сборка *de novo*

(например, хотим изучать геном вида, чей геном пока не секвенирован)

# Случайное покрытие

Все платформы «второго поколения» включают подготовку **случайных** фрагментов генома и их амплификацию (размножение).

В результате полученные чтения (они же прочтения, они же риды) также представляют собой набор случайных фрагментов заданной длины.

В идеальном случае вероятность стать началом чтения одинакова для всех позиций в геноме (а на практике это не всегда так).

# Сборка на геном

Пусть длина чтения 100, размер генома 1 млн п.н. и мы получили 50 000 чтений. Значит, среднее покрытие = 5. Хватит ли этого, чтобы собрать весь геном?

# Сборка на геном

Пусть длина чтения 100, размер генома 1 млн п.н. и мы получили 50 000 чтений. Значит, среднее покрытие = 5. Хватит ли этого, чтобы собрать весь геном?

Количество чтений, покрывающих данный нуклеотид, распределено по Пуассону:

$$P(k) = \exp(-\lambda) \lambda^k / k!$$

где  $k$  – число чтений,  $\lambda$  – среднее покрытие (в нашем случае  $\lambda = 5$ ).

Значит, вероятность того, что на нуклеотид не попадёт **ни одного** чтения, равна  $P(0) = \exp(-\lambda)$ . При  $\lambda = 5$  эта вероятность равна  $1/\exp(5) \approx 1/148$ .

# Сборка на геном

Пусть длина чтения 100, размер генома 1 млн п.н. и мы получили 50 000 чтений. Значит, среднее покрытие = 5. Хватит ли этого, чтобы собрать весь геном?

**Ответ:** вряд ли. Чтения ложатся случайно, примерно каждый 150-ый нуклеотид ими не покроется. То есть почти наверняка более 6 000 нуклеотидов не будет покрыто, и при самой идеальной сборке получится не целый геном, а много кусков, разделённых непокрытыми участками.

При таком размере генома нужно не менее чем 15-кратное среднее покрытие, чтобы можно было рассчитывать собрать геном полностью!

Ещё проблема – повторы. Не всегда чтение однозначно «ложится» на геном.

Третья проблема – время (при большом покрытии большого генома)

# Сборка на геном

Главная проблема, решаемая разработчиками алгоритмов – время.  
Два основных подхода: хэш-таблицы и суффиксные деревья  
(преобразование Берроуза – Уилера).

Имеется несколько десятков программ, часть из них платные, часть –  
свободно распространяемые.

*Это вы уже знаете :)*

# Сборка *de novo*

Есть два основных типа алгоритмов сборки:

- OLC = overlap-layout-consensus
- de Bruijn graph

Алгоритмы OLC работают непосредственно с чтениями.

Алгоритмы, использующие граф де Брёйна, сначала составляют список  $k$ -меров (слов длины  $k$ , например  $k = 31$ ), встретившихся в чтениях.

**Недостатки:**

теряется часть информации

**Достоинства:**

сильно экономится память (большинство  $k$ -меров встречается во многих ридсах)

упрощается работа с повторяющимися участками

есть возможность отсеивать ошибки уже на начальной стадии

*NB:* любой алгоритм сборки дополняет входной набор последовательностей (чтений или  $k$ -меров) комплементарными



# Алгоритмы сборки OLC

Программы: Phrap, Cap3, Tigr, ...

**Read1** - TTTGGTGCTCTTCGAAAAGGGATCTTCGAGAGAGATCTCGCGATAAGGTTG

**Read2** - GAGAGAGATCTCGCGATAAGGTTGAAGTAGAAAAATGTGTGTGGTGAA

**overlap**

TTTGGTGCTCTTCGAAAAGGGATCTTC**GAGAGAGATCTCGCGATAAGGTTG**

**GAGAGAGATCTCGCGATAAGGTTG**AAGTAGAAAAATGTGTGTGGTGAA

<http://www.homolog.us/Tutorials/Tut-Img/Set1/fig2.png>

# Проблема повторов



Read1



Read2



Assembly



<http://www.homolog.us/Tutorials/Tut-Img/Set1/fig3.png>

# Графы де Брёйна

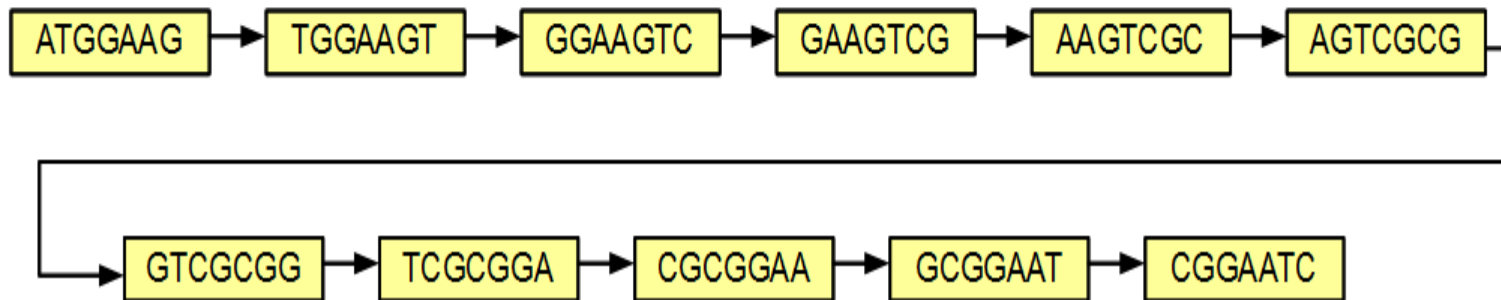
sequence

ATGGAAGTCGCGGAATC

7mers

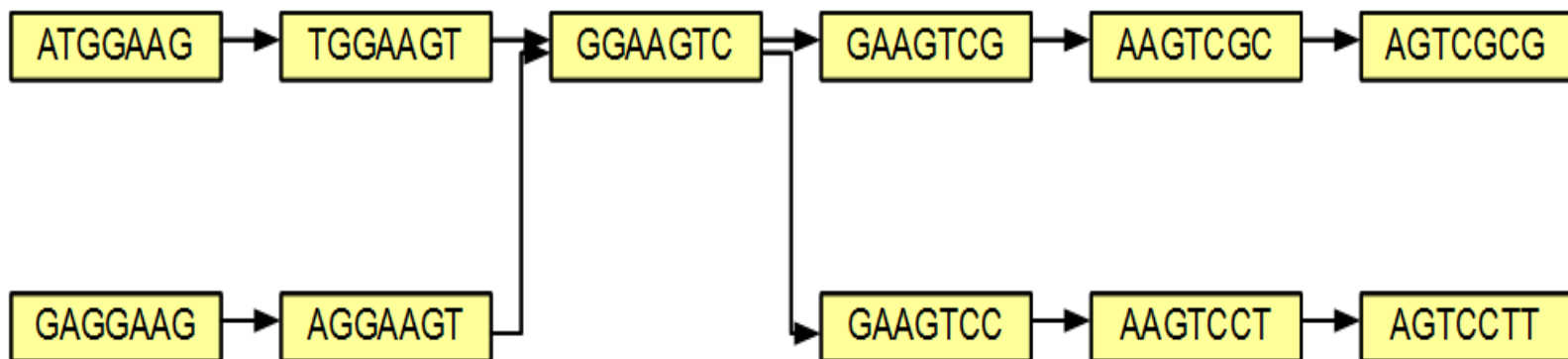
ATGGAAG  
TGGAAGT  
GGAAGTC  
GAAGTCG  
AAGTCGC  
AGTCGCG  
GTCGCGG  
TCGCGGA  
CGCGGAA  
GCGGAAT  
CGGAATC

de Bruijn graph



# Графы де Брёйна

ATGGAAGTCGCG  
GAGGAAGTCCTT



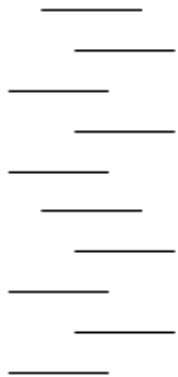
# Вопрос

Почему длина  $k$ -мера (т.е., число  $k$ ),  
применяемого при создании графа де Брёйна,  
должна быть нечётной?

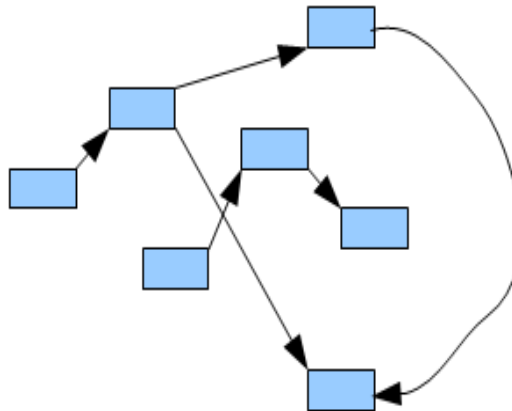
# Графы де Брёйна

Десятки программ: SPAdes, Velvet, ABySS, Trinity, Oases, SOAPdenovo, ...

NGS library



de Bruijn Graph



Genome



<http://www.homolog.us/Tutorials/index.php?p=1.4&s=1>

# Pair-end reads и mate pair reads

Технология Illumina предполагает чтение заданного числа (например, 100) нуклеотидов с двух концов случайного фрагмента генома небольшой (200–600 п.н.) длины.

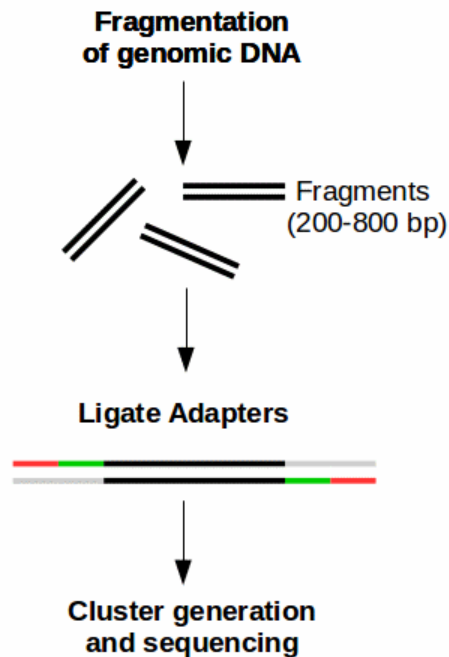
В выходном файле последовательности концов одного и того же фрагмента некоторым способом ссылаются друг на друга. Это и есть парноконцевые чтения (pair-end reads).

Имеется особый способ приготовления библиотеки для секвенирования, при котором концы секвенируемых фрагментов в геноме удалены друг от на большее расстояние (2–5 тысяч п.н.).

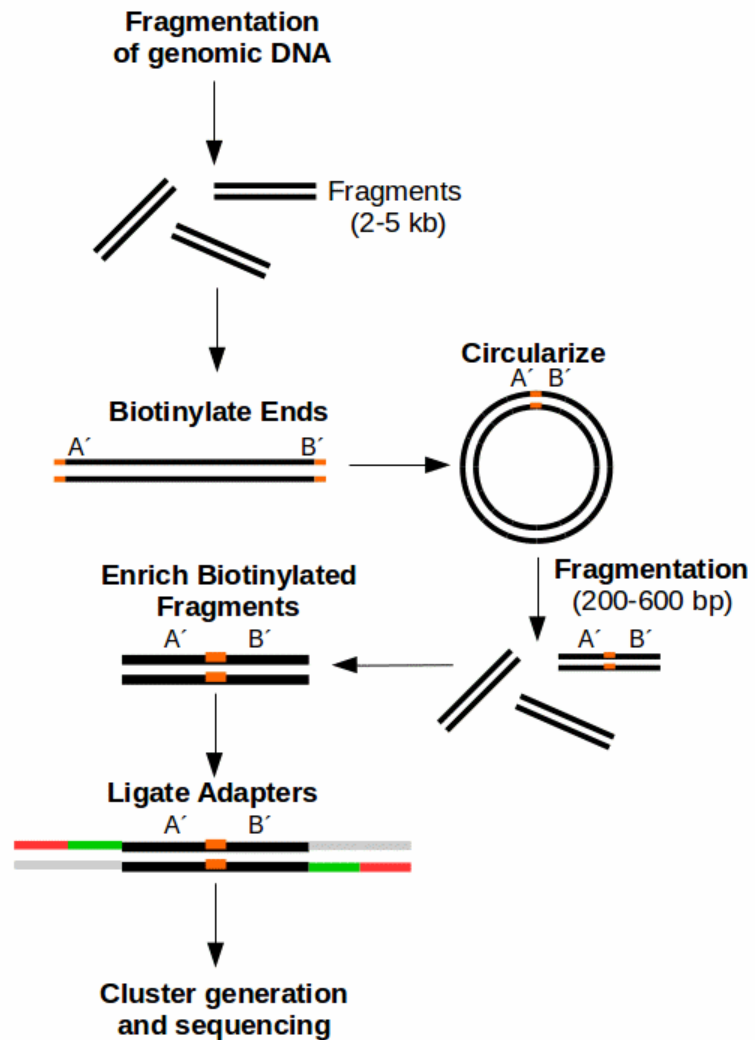
При таком способе секвенирования парноконцевые чтения называются «встречноконцевыми» (mate pair reads).

Большинство программ сборки могут учитывать «парность» чтений.

## Paired-End Sequencing (Short-insert paired-end reads)



## Mate Pair Sequencing





# Результат сборки

Результат – так называемые «контиги» (contigs), то есть непрерывные участки генома, однозначно выводимые из чтений

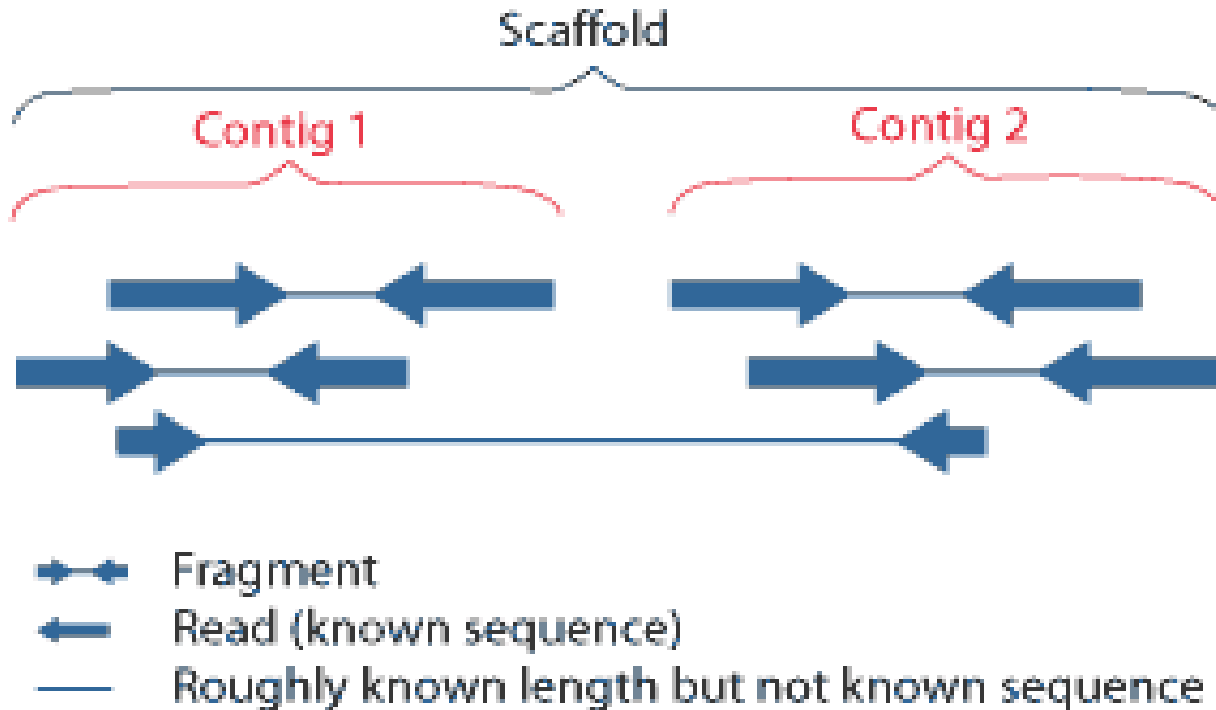
Для прокариот часто удаётся собрать весь геном (но редко «полностью автоматически» – обычно нужны дополнительные усилия, например секвенирование плохо покрытых участков по Сэнгеру).

Для эукариот, как правило, «геномом» объявляется свалка контигов, тем или иным способом приписанных к известным хромосомам.

Кроме контигов, бывают ещё «скэффолды» (scaffolds).

Скэффолд – это последовательность контигов, про которые известно, что они соответствуют близким участкам одной и той же молекулы ДНК, но между которыми остаются неизвестные участки. Источник такой информации – парноконцевые чтения.

# Контиги и скэффолды



[https://en.wikipedia.org/wiki/Scaffolding\\_\(bioinformatics\)#/media/File:PET\\_contig\\_scaffold.png](https://en.wikipedia.org/wiki/Scaffolding_(bioinformatics)#/media/File:PET_contig_scaffold.png)

LOCUS MU862503 33801 bp DNA linear CON 17-NOV-2023  
 DEFINITION UNVERIFIED\_CONTAM: Salix arbutifolia isolate SAR-KD-M1 unplaced  
 genomic scaffold scaffold2257, whole genome shotgun sequence.  
 ACCESSION MU862503 AWYI01000000  
 VERSION MU862503.1  
 DBLINK BioProject: PRJNA211611  
 BioSample: SAMN10688929  
 KEYWORDS WGS; UNVERIFIED; UNVERIFIED\_CONTAMINANT.  
 SOURCE Salix arbutifolia  
 ORGANISM Salix arbutifolia  
 Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;  
 Spermatophyta; Magnoliopsida; eudicotyledons; Gunneridae;  
 Pentapetales; rosids; fabids; Malpighiales; Salicaceae; Saliceae;  
 Salix.  
 REFERENCE 1 (bases 1 to 33801)  
 AUTHORS Zhang,J., Wang,Y., Zeng,Y., He,C., Rao,G. and Wang,Z.  
 TITLE The genome of Salix arbutifolia, an important member of the  
 Salicaceae family  
 JOURNAL Unpublished  
 REFERENCE 2 (bases 1 to 33801)  
 AUTHORS Zhang,J., Wang,Y., Zeng,Y., He,C., Rao,G. and Wang,Z.  
 TITLE Direct Submission  
 JOURNAL Submitted (18-SEP-2013) Research Institute of Forestry, Chinese  
 Academy of Forestry, No. 2 Dongxianfu, Xiangshan Road, Haidian  
 District, Beijing, Beijing 100091, China  
 COMMENT GenBank staff has noted that the sequence(s) may be contaminated.  
  
 ##Genome-Assembly-Data-START##  
 Assembly Method :: SOAPdenovo v. 2.01  
 Genome Coverage :: 166x  
 Sequencing Technology :: Illumina HiSeq 2000  
 ##Genome-Assembly-Data-END##  
 FEATURES Location/Qualifiers  
 source 1..33801  
 /organism="Salix arbutifolia"  
 /mol\_type="genomic DNA"  
 /submitter\_seqid="scaffold2257"  
 /isolate="SAR-KD-M1"  
 /db\_xref="taxon:75699"  
 /chromosome="Unknown"  
 /country="China: Kuandian, Dandong, Liaoning Province"  
 CONTIG join(AWYI01042789.1:1..1286,gap(6516),AWYI01042790.1:1..4274,  
 gap(3069),AWYI01042791.1:1..18656)

//

# Результат сборки

Например, т.н. «референсная» версия генома человека (GRCh38.p14, февраль 2022) состоит из 470 скэффолдов, генома домового мыши (GRCm39) – из 101 скэффолда, а генома лошади (EquCab3.0) – из 4700 скэффолдов.

Контигов больше: для человеческого генома их 996, для мышинного 305, для лошадиного 10986.

# Показатели качества сборки

Самый популярный – N50.

Это наибольшее число такое, что контигами длины  $> N50$  покрыто 50% генома.

*При этом чаще всего за длину генома принимают суммарную длину контигов.*

Используется также N90 (аналогично – наименьшая длина контига из минимального набора, покрывающего 90% генома).

Есть ещё показатели L50 и L90 (минимальное **число** контигов, покрывающих, соответственно, 50% и 90% генома).

*То есть минимальный набор, покрывающий 50% генома, состоит из L50 контигов, чья длина  $\geq N50$*

# Показатели качества сборки

Например, для человеческого, мышиноного и лошадиного референсных геномов показатели такие:

Геном	N50 (bp)			L50
<i>Homo sapiens</i>	57	879	411	18
<i>Mus musculus</i>	59	462	871	15
<i>Equus caballus</i>	1	502	753	462