Факультет биоинженерии и биоинформатики МГУ
II курс

Сборка чтений

С.А. Спирин
13 декабря 2022

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).

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)

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, λ – среднее покрытие (в нашем случае $\lambda = 5$).

Значит, вероятность того, что на нуклеотид нее 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 (аналогично BLAST) и суффиксные деревья (преобразование Барроуза – 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
- Граф де Брёйна (de Bruijn graph)

Алгоритмы OLC работают непосредственно с прочтениями.

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

Недостатки:

теряется часть информации

Достоинства:

сильно экономится память (большинство *k*-меров встречается во многих ридах)
упрощается работа с повторяющимися участками
есть возможность отсеивать ошибки уже на начальной стадии

Алгоритмы сборки OLC

Программы: Phrap, Cap3, Tigr, …

Read1 - TTTGGTGCTCTTCGAAAAGGGATCTTCGAGAGAGATCTCGCGATAAGGTTG

Read2 - GAGAGAGATCTCGCGATAAGGTTGAAGTAGAAAAATGTGTGTGGTGAA

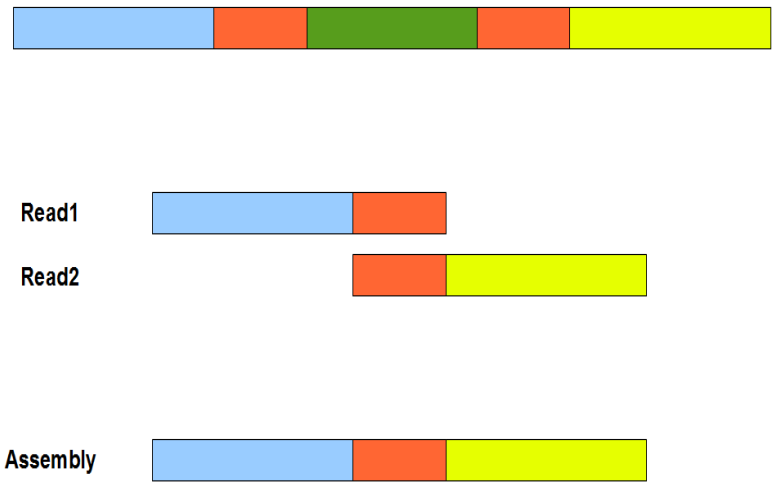
overlap

TTTGGTGCTCTTCGAAAAGGGATCTTCGAGAGAGATCTCGCGATAAGGTTG

GAGAGAGATCTCGCGATAAGGTTGAAGTAGAAAAATGTGTGTGGTGAA

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

Проблема повторов



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

Из-за повторов алгоритм OLC может собрать неверный контиг

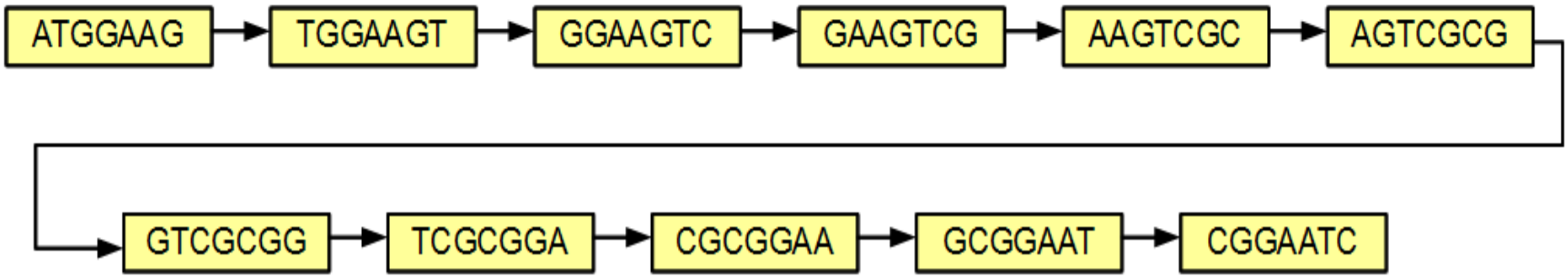
Графы де Брёйна

sequence **ATGGAAGTCGCGGAATC**

7mers

ATGGAAG
TGGAAGT
GGAAGTC
GAAGTCG
AAGTCGC
AGTCGCG
GTCGCGG
TCGCGGA
CGCGGAA
GCGGAAT
CGGAATC

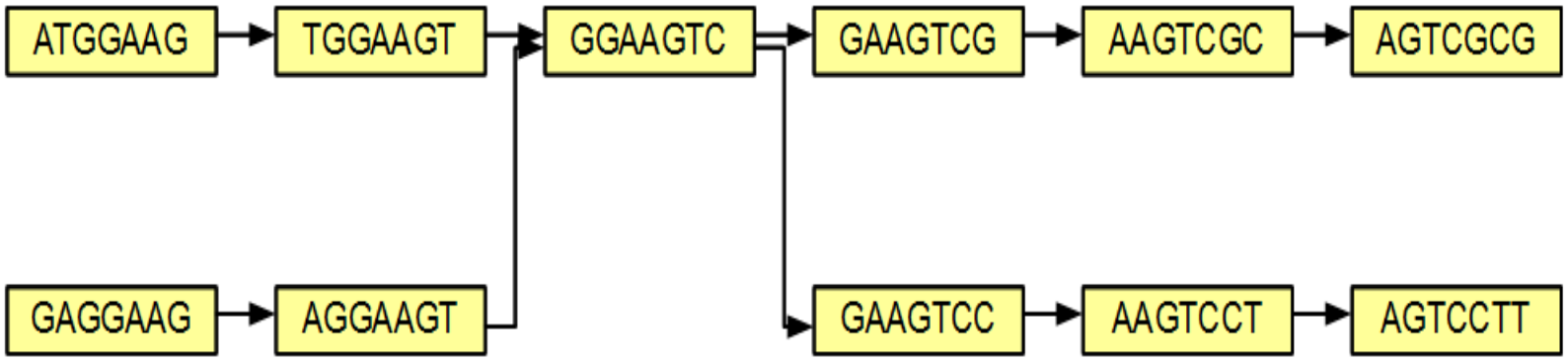
de Bruijn graph



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

Графы де Брёйна

ATGGAAGTCGCG
GAGGAAGTCCTT

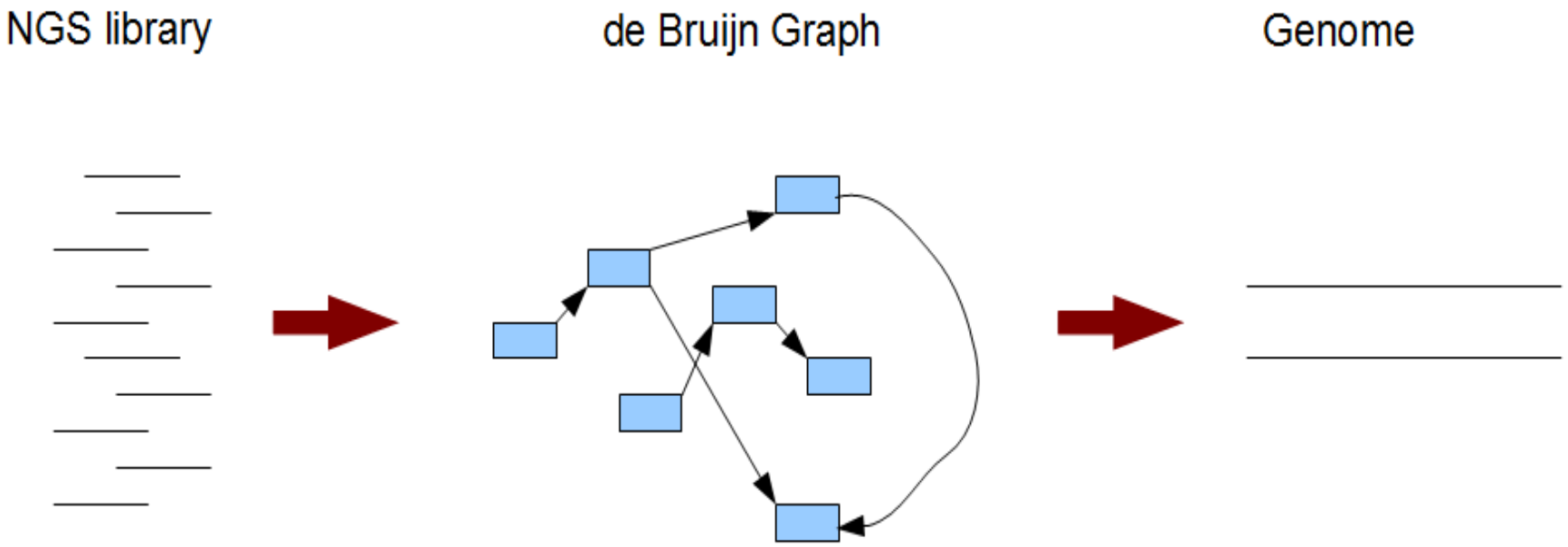


http://www.homolog.us/Tutorials/index.php?p=1.4&s=1

Оборвав контиг в точке неопределённости, можно гарантировать, что неверных контигов не будет

Графы де Брёйна

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



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

Pair-end reads и mate pair reads

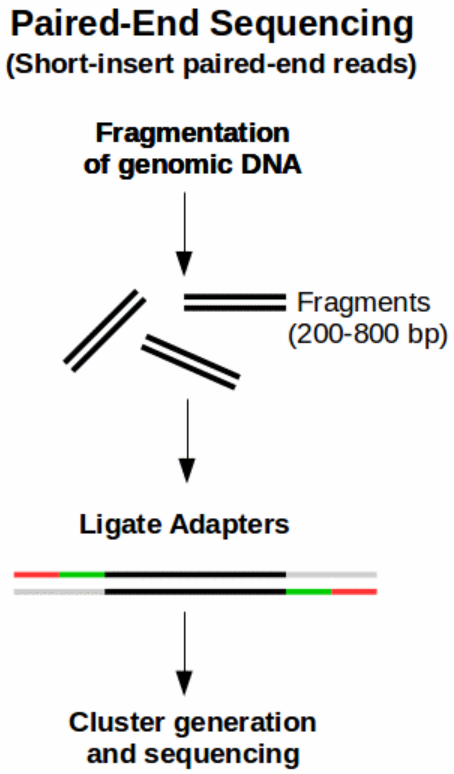
Технология *Illumina* предполагает чтение заданного числа (например, 100) нуклеотидов с двух концов случайного фрагмента генома небольшой (200–600 п.н.) длины.

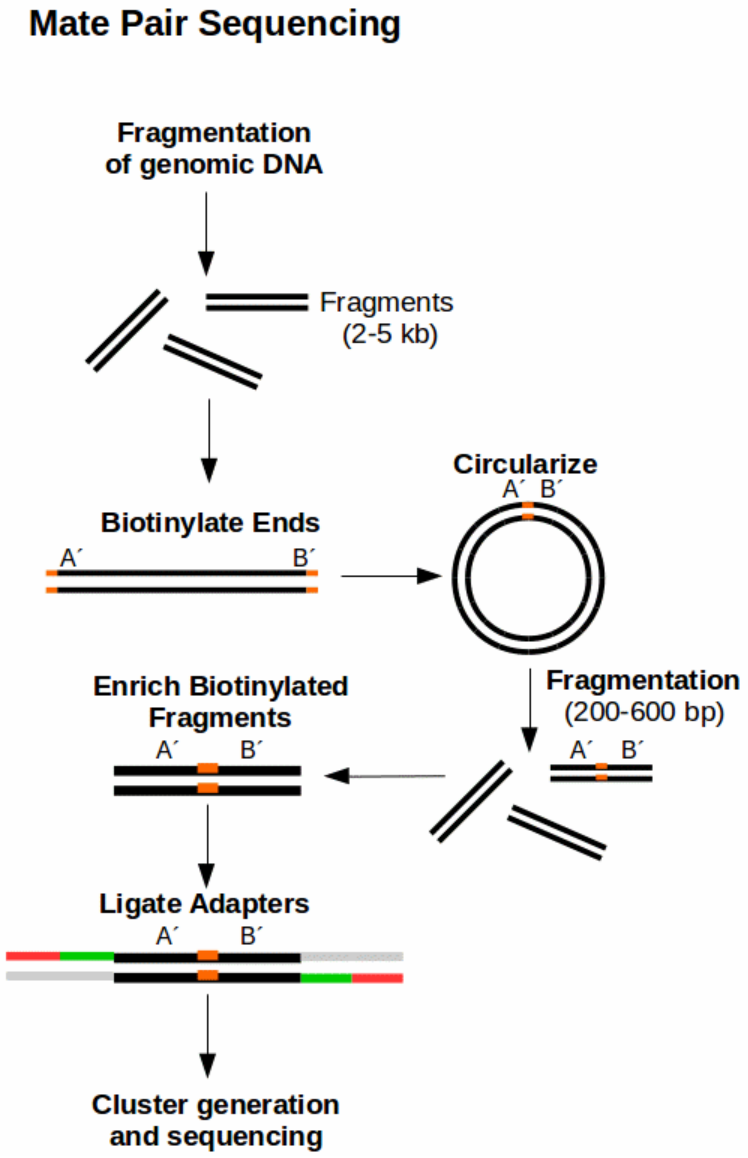
В выходном файле последовательности концов одного и того же фрагмента тем или иным способом ссылаются друг на друга. Это и есть парноконцевые чтения (**pair-end reads**).

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 тем или иным способом учитывают «парность» прочтений.





https://www.ecseq.com/support/ngs/what-is-mate-pair-sequencing-useful-for

Результат сборки

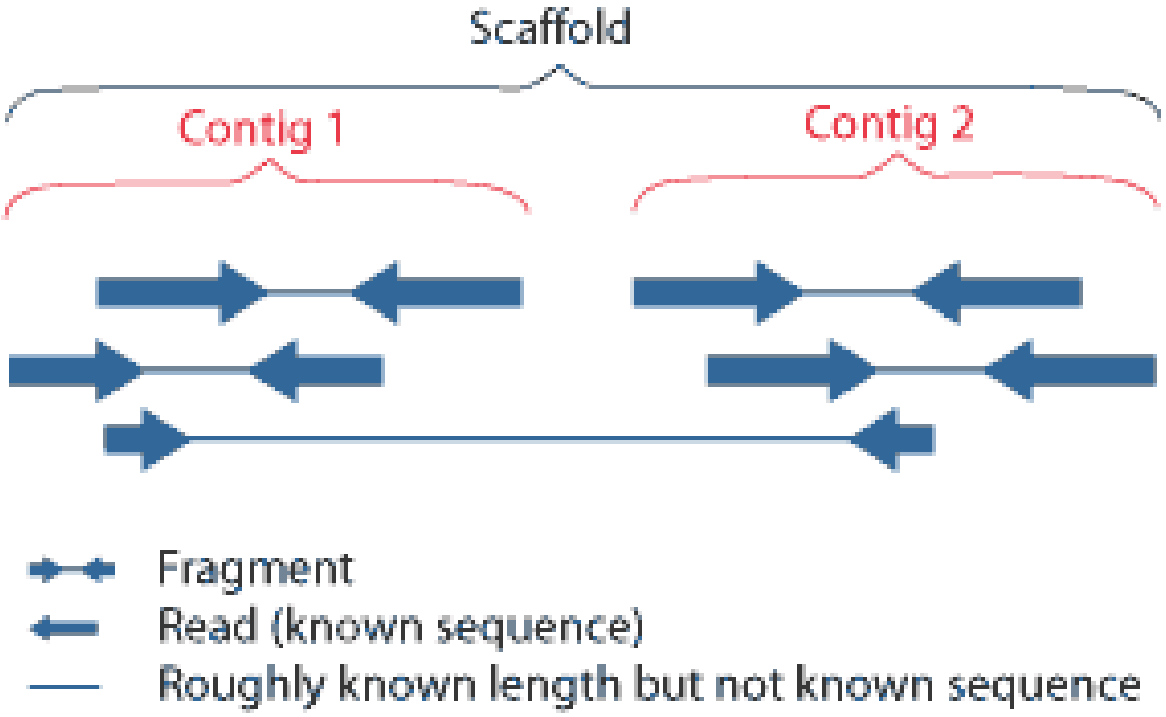
Результат – так называемые «контиги», то есть непрерывные участки генома.

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 «скаффолды» – последовательность контигов, между которыми остаются неизвестные участки
(источник такой информации – парноконцевые чтения).

Контиги и скаффолды



https://en.wikipedia.org/wiki/Scaffolding_(bioinformatics)#/media/File:PET_contig_scaffold.png

Результат сборки

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

Контигов больше: для человеческого генома их 996, для мышиного 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.

Изредка 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 ≥ N50

Показатели качества сборки

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

Геном	N50 (bp)	L50
Homo sapiens	57879411	18
Mus musculus	59462871	15
Equus caballus	1502753	462