Межфакультетский курс «Биоинформатика»
Факультет биоинженерии и биоинформатики МГУ
осень 2014

Секвенирование

Платформы

Сборка

Сергей Александрович Спирин
15 и 22 октября 2014

История

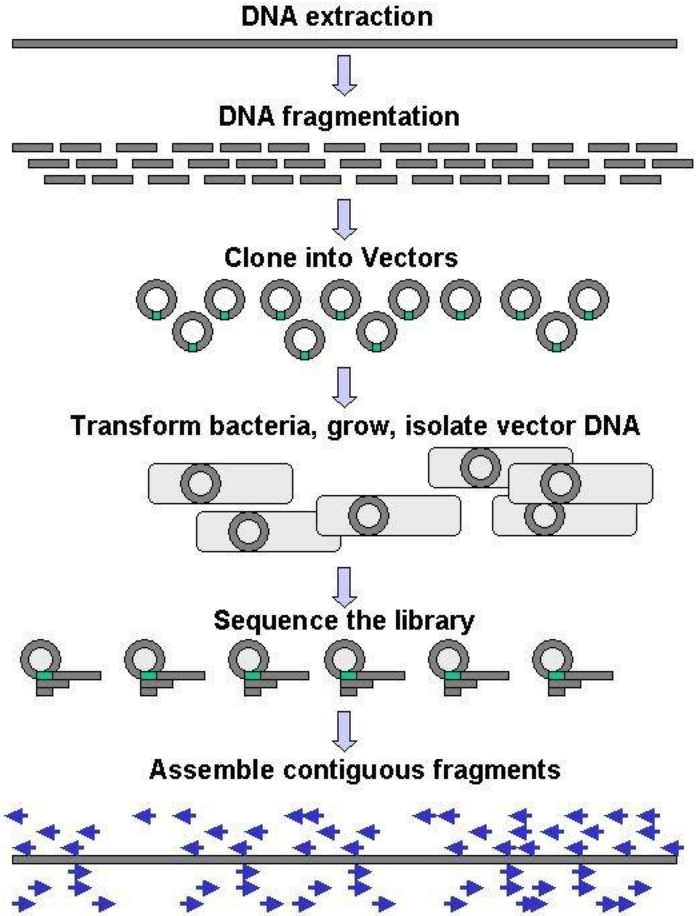
1953: структура ДНК (Уотсон и Крик, Нобелевская премия 1962)
1973: опубликована первая последовательность природной ДНК: 24 п.н. (*lac* оператор, Максам и Гилберт)
1977: опубликованы методы секвенирования Сэнгера и Гилберта (Нобелевская премия 1980)
1982: основан GenBank
1983: разработана полимеразная цепная реакция (ПЦР, PCR)
1987: первый автоматический секвенатор (Applied Biosystems Prism 373)
1995: первый геном бактерии (*Haemophilus influenzae*)
1996: капиллярный секвенатор ABI 310 (основан на методе Сэнгера)
1998: первый геном животного (круглого червя *Caenorhabditis elegans*)
2000: человеческий геном (почти полный)
2005: первый пиросеквенатор 454 Life Sciences (с 2007 – Roche): начало эры NGS
2006: первый секвенатор фирмы Solexa (с 2007 – Illumina)

Для чего

Геномы разных видов (бактерий, животных, растений)
Геномы индивидуумов (изучение индивидуальных различий)
Транскриптомы
Chip-seq и подобные исследования

Секвенирование по Сэнгеру

Этапы:
выделение ДНК
подготовка «библиотеки»
амплификация (клонирование и ПЦР)
секвенирование «мечеными терминаторами»
Характеристики:
время работы несколько суток
длина прочтения («рида») до 1000 п.н.
один рид за раз
ошибки ~0,5%
Повторением части процедуры (ПЦР+секвенирование) можно добиться ридов в несколько тысяч п.н. и почти исключить ошибки.



Секвенирование по Сэнгеру

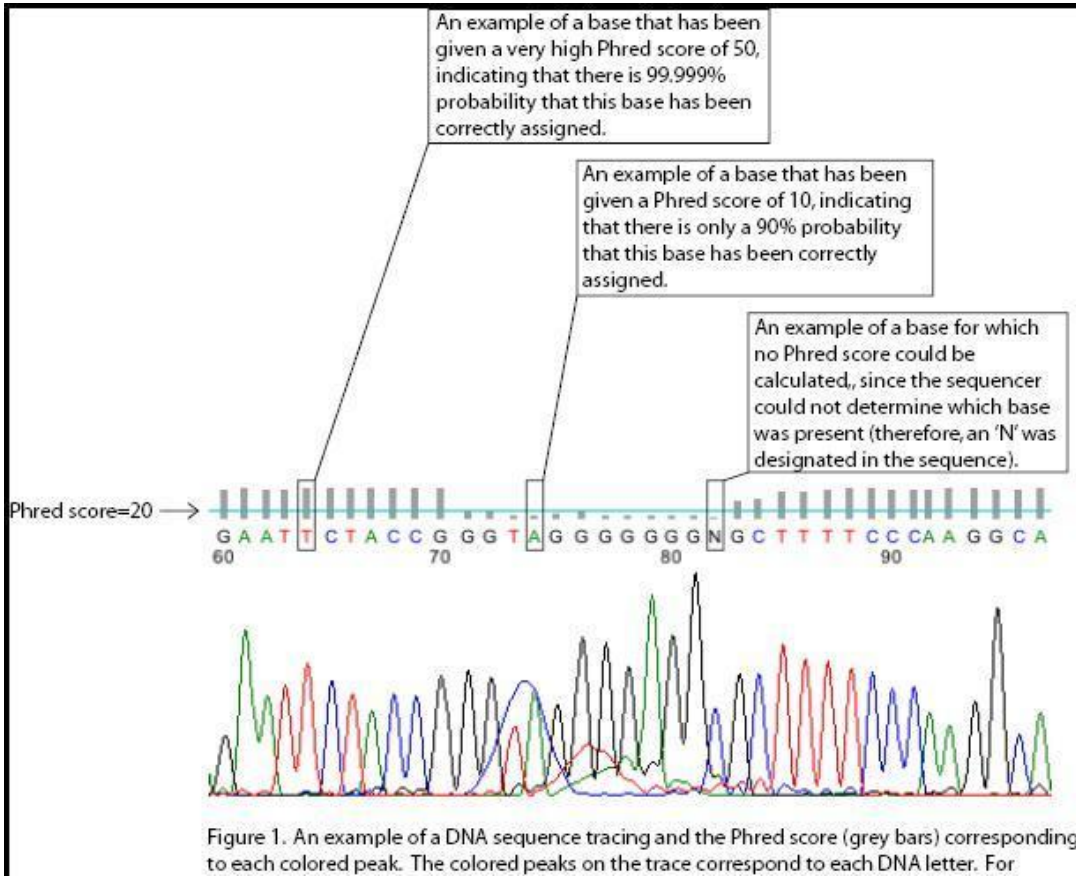
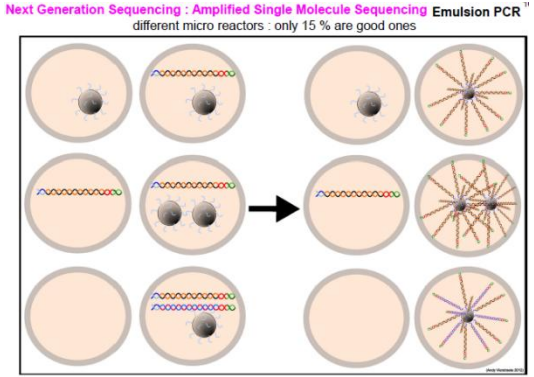


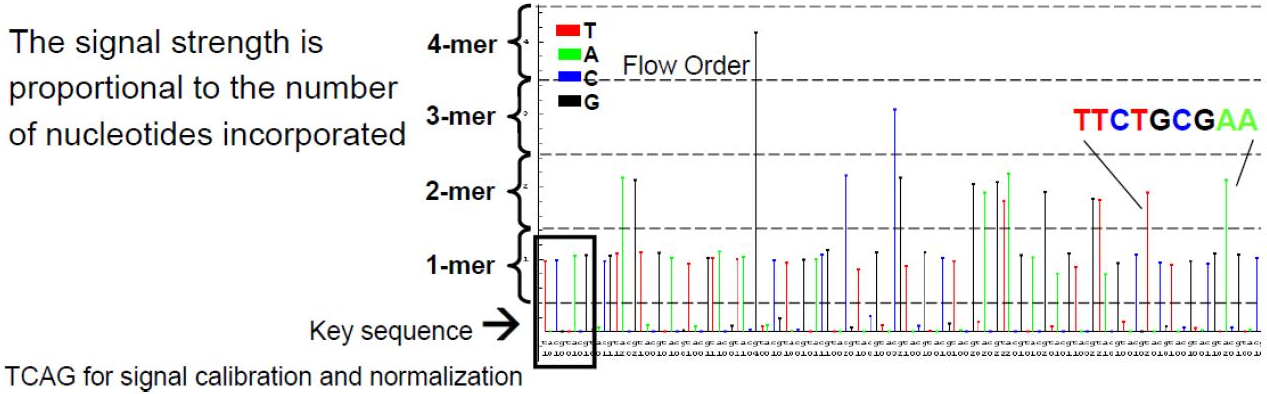
Figure 1. An example of a DNA sequence tracing and the Phred score (grey bars) corresponding to each colored peak. The colored peaks on the trace correspond to each DNA letter. For

http://en.wikipedia.org/wiki/Phred_quality_score

Платформа 454 Life Sciences (Roche)

Этапы:
выделение ДНК
подготовка «библиотеки»
эмульсионный ПЦР
пиросеквенирование
Характеристики:
время работы 24 часа
длина рида 700 п.н.
число ридов 1 млн.
ошибки ~0,003%



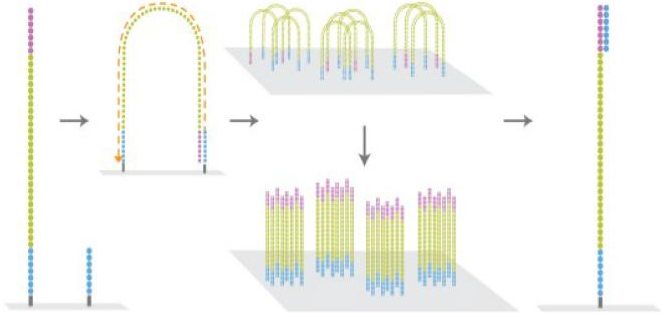


http://users.ugent.be/~avierstr/nextgen/Next_generation_sequencing_web.pdf

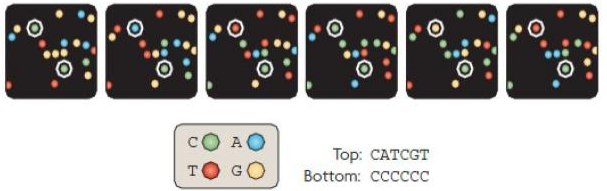
Платформа Illumina (Solexa)

Этапы:
выделение ДНК
подготовка «библиотеки»
ПЦР «мостиками на подложке»
секвенирование «удаляемыми мечеными терминаторами»

Характеристики:
время работы 11 дней
длина рида 100 п.н.
число ридов 3 млрд.
ошибки ~0,1%



4 nucleotides with different dye
flow simultaneous



http://users.ugent.be/~avierstr/nextgen/Next_generation_sequencing_web.pdf

Случайное покрытие

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

В результате риды также представляют собой набор случайных фрагментов заданной длины. В идеальном случае вероятность стать началом рида одинакова для всех позиций в геноме (а на практике это не всегда так).

Секвенирование «третьего поколения»

Ion Torrent: измеряется ток, возникающий при присоединении нуклеотида к растущей цепи (это перспективная технология, но скорее ещё «второго поколения», так как требует амплификации ДНК): риды 200 п.н., 1% ошибок, ~10 млн. ридов, несколько часов.

Helicos: пока риды ~35 п.н., 3% ошибок, 1 млрд. ридов, 8 дней. Появилась в 2009. Читается одна молекула! Тем самым не требуется амплификация. Это очень важно для, например, количественных исследований.

Pacific Bioscience: фиксируется удерживание нового нуклеотида на растущей цепи. Длина рида несколько тысяч п.н.! 70 000 ридов за полчаса, 5% ошибок.

Oxford Nanopore: цепь ДНК просачивается через нанопору, фиксируются характеристики проходящего нуклеотида. Риды длиной в десятки тыс. п.н.!!!
Но пока ~20% ошибок.

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

Сборка на геном

Пусть длина рида 100, размер генома 1 млн п.н. и мы получили 50 000 ридов. Значит, среднее покрытие = 5. Хватит ли этого, чтобы собрать весь геном?
Ответ: вряд ли. Риды 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 рид 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.

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

Сборка *de novo*

Есть два основных типа алгоритмов сборки:

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

Алгоритмы OLC работают непосредственно с ридами.

Алгоритмы, использующие граф де Брайна, сначала составляют список *k*-меров (слов длины *k*, например *k* = 30), встретившихся в ридах.

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

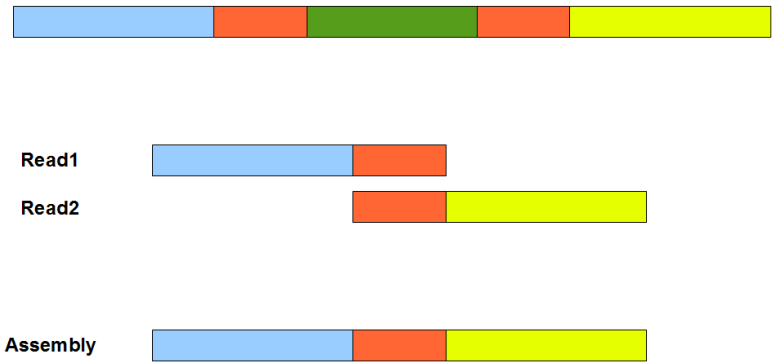
overlap

TTTGGTGCTCTTCGAAAAGGGATCTTCGAGAGAGATCTCGCGATAAGGTTG

GAGAGAGATCTCGCGATAAGGTTGAAGTAGAAAAATGTGTGTGGTGAA

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

Проблема повторов



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

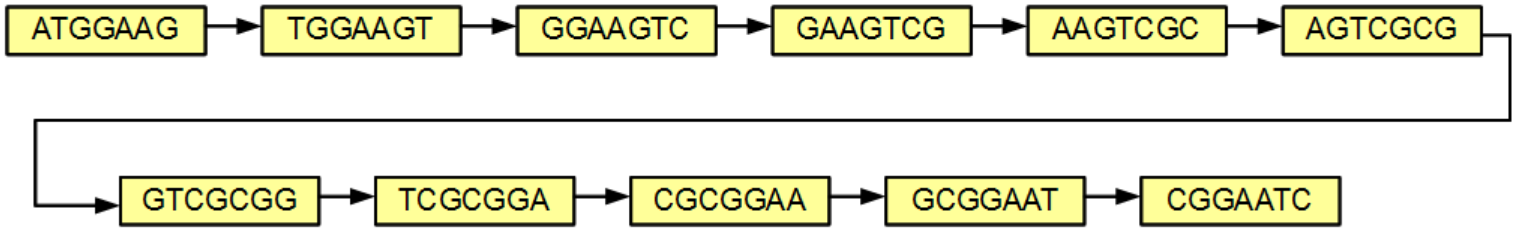
Графы де Брайна

sequence **ATGGAAGTCGCGGAATC**

7mers

ATGGAAG
TGGAAGT
GGAAGTC
GAAGTCG
AAGTCGC
AGTCGCG
GTCGCGG
TCGCGGA
CGCGGAA
GCGGAAT
CGGAATC

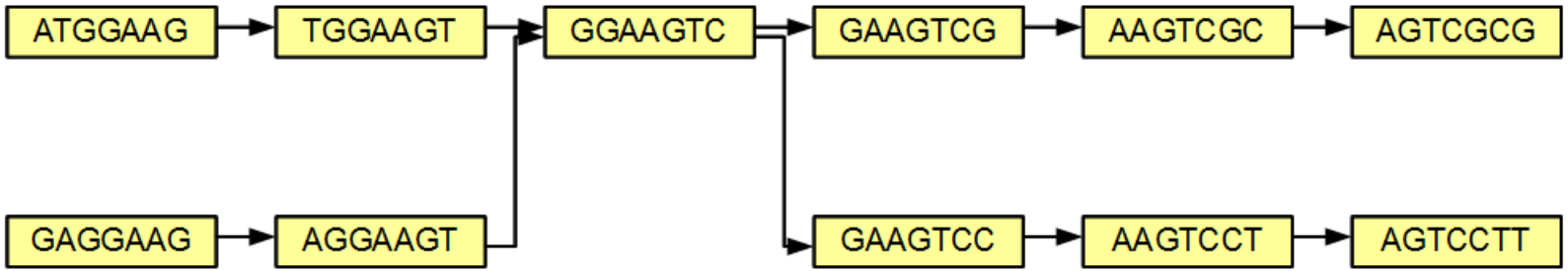
de Bruijn graph



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

Графы де Брайна

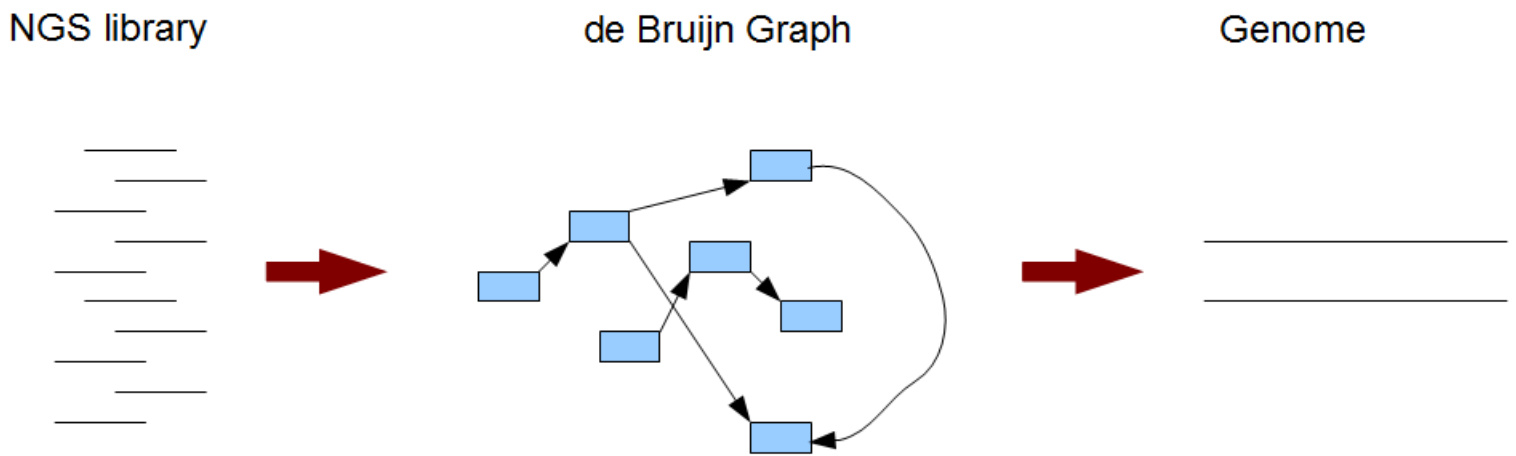
ATGGAAGTCGCG
GAGGAAGTCCTT



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

Графы де Брайна

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



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

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 – особый приём секвенирования, называемый “pair-end read”)

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 – 90% генома).

Эта область биоинформатики очень молода, и удовлетворительные показатели ещё не выработаны!