Технологии секвенирования

Кирилл Григорьев
Caribbean Genome Center

О чем вообще речь

ДНК –

Геном –

Секвенирование –

О чем вообще речь

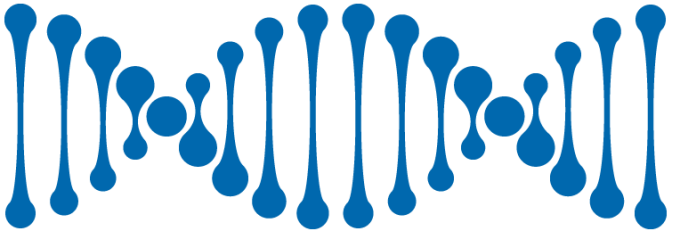
ДНК – макромолекула, обеспечивающая хранение, передачу из поколения в поколение и реализацию генетической программы развития и функционирования живых организмов

Геном – совокупность наследственного материала, заключенного в клетке организма; содержит биологическую информацию, необходимую для его построения и поддержания. В большинстве организмов записан в ДНК.

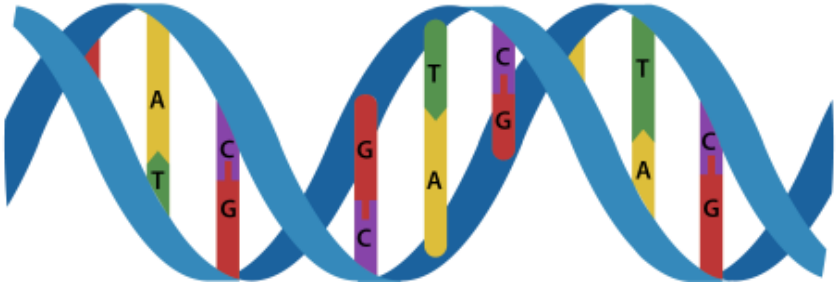
Секвенирование – определение последовательности (нуклеотидной, белковой).

О чем вообще речь

ДНК –

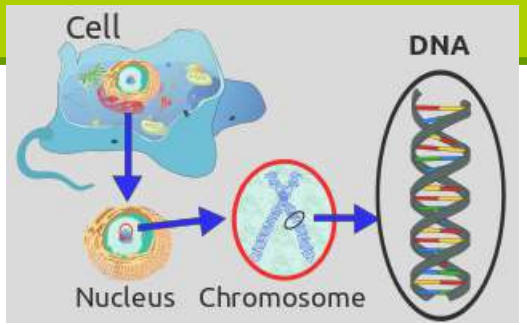


Геном –

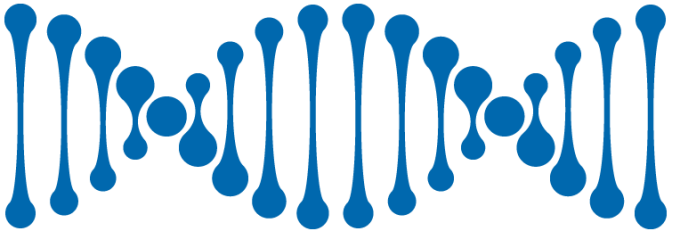


Секвенирование –

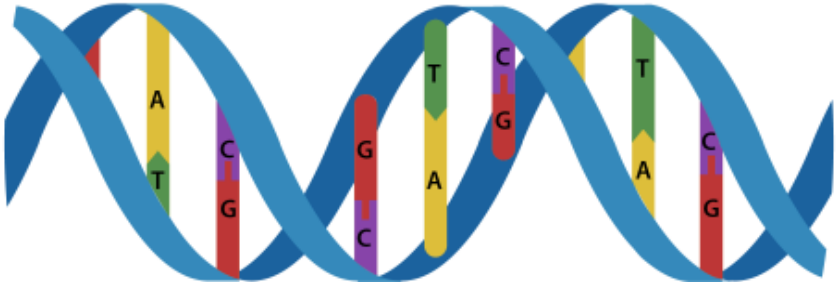
О чем вообще речь



ДНК –



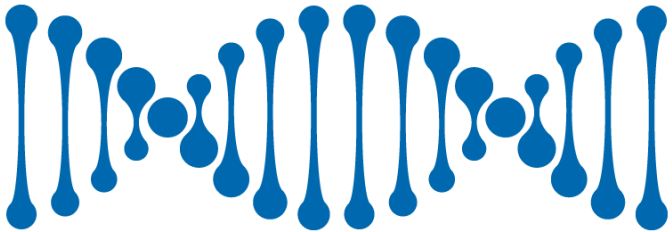
Геном –

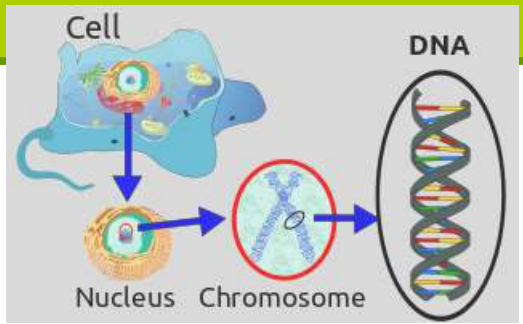


Секвенирование –

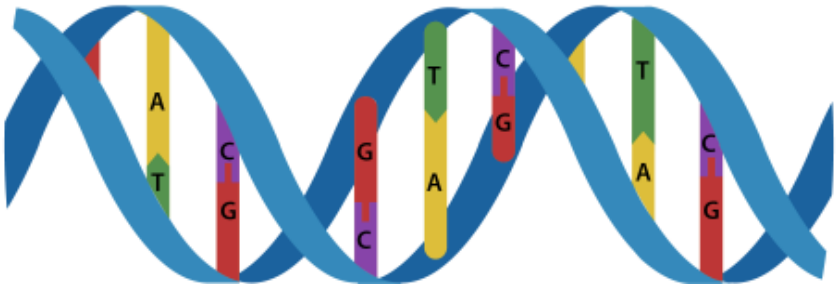
О чем вообще речь

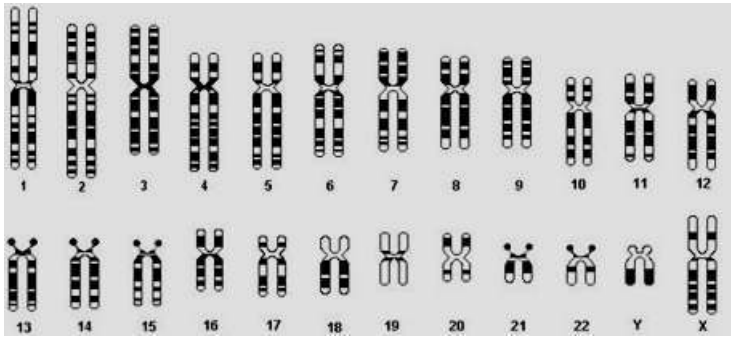
ДНК –





Геном –





Секвенирование –

Цели и средства: например...

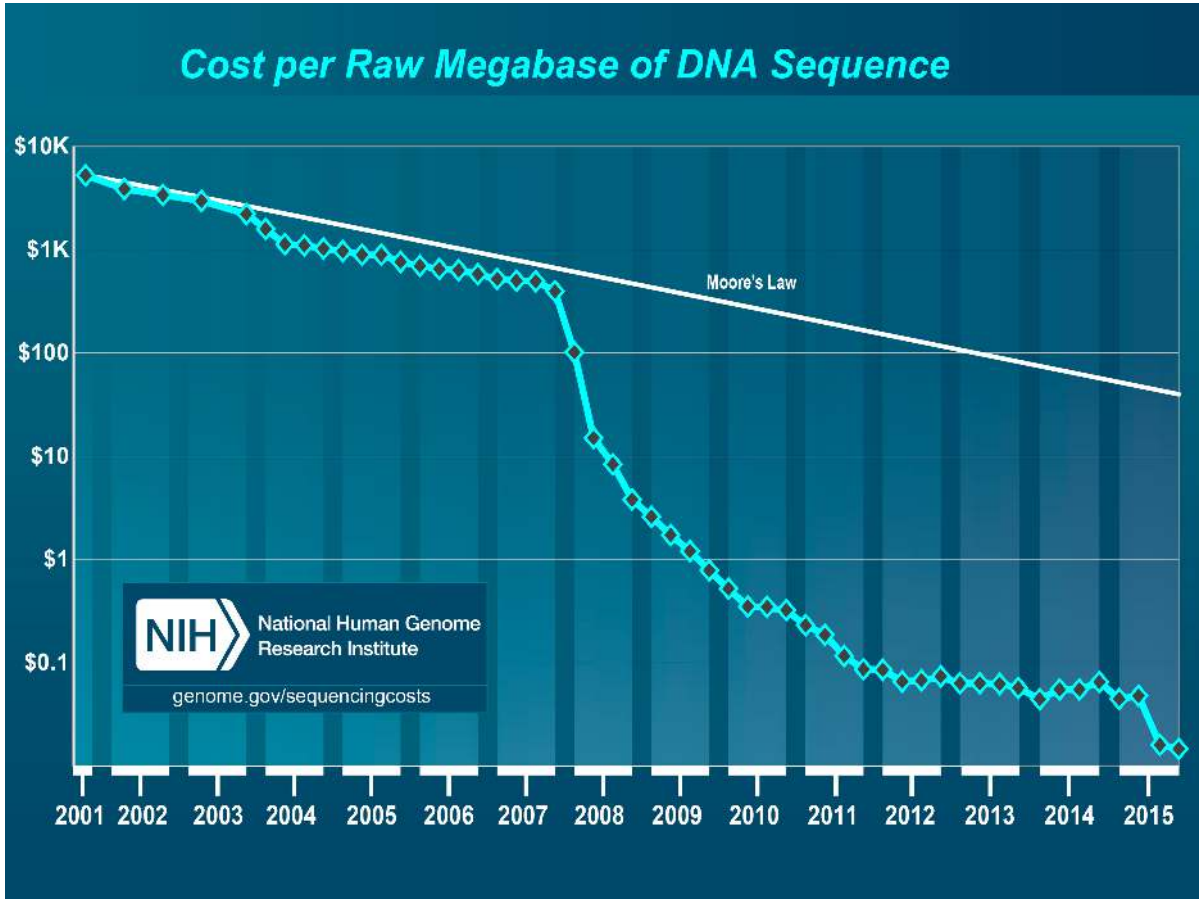
- Секвенирование отдельных генов
- Сборка:
  - вирусных, бактериальных геномов
  - эукариотических геномов
  - экзомов
  - ...
- Поиск genomic rearrangements
- Genotyping by sequencing
- RNA-seq
- Исследование профилей метилирования
- ...

Цели и средства: например...

- Секвенирование отдельных генов
- Сборка:
  - вирусных, бактериальных геномов
  - эукариотических геномов
  - экзомов
  - ...
- Поиск genomic rearrangements
- Genotyping by sequencing
- RNA-seq
- Исследование профилей метилирования
- ...

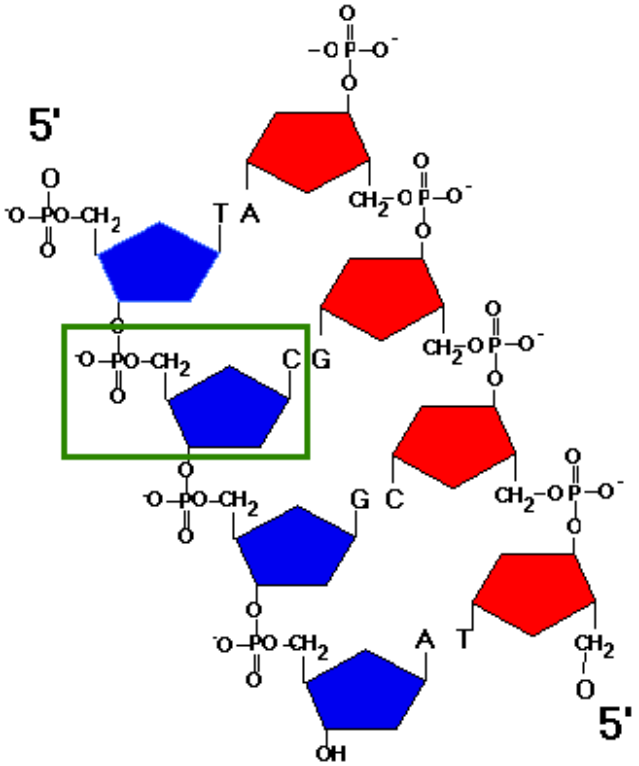
- Вопросы эволюционной биологии: аннотация генов, демографическая история...

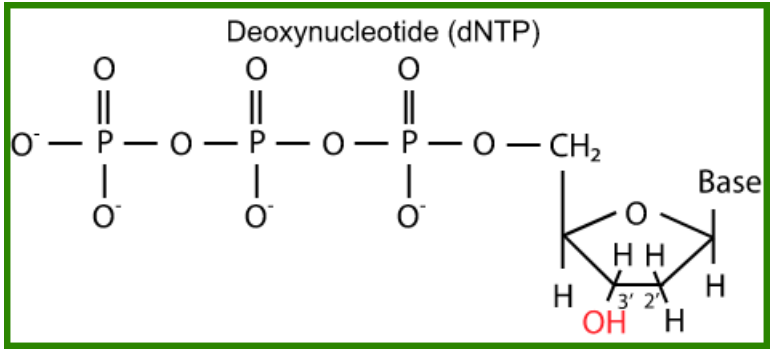
График, который все вставляют в свои презентации



Секвенирование первого поколения: **Sanger sequencing**

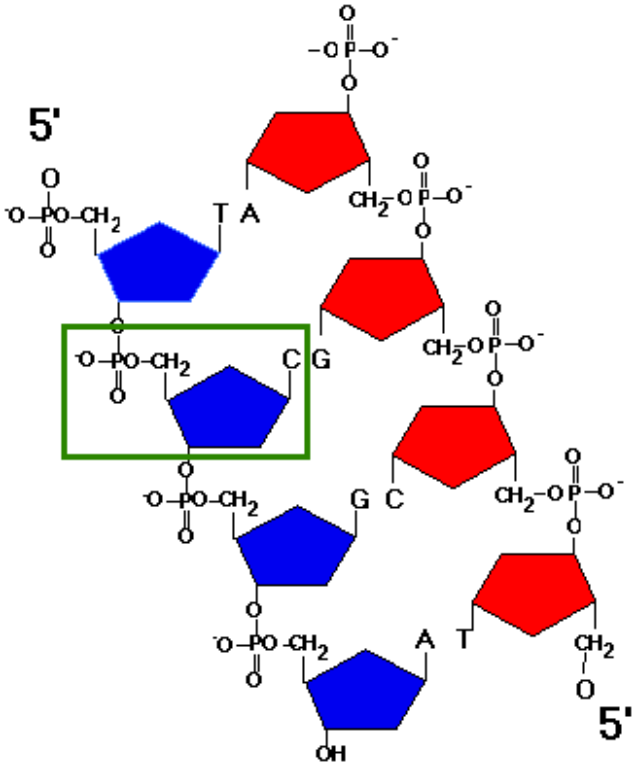
a.k.a. the **chain termination** method

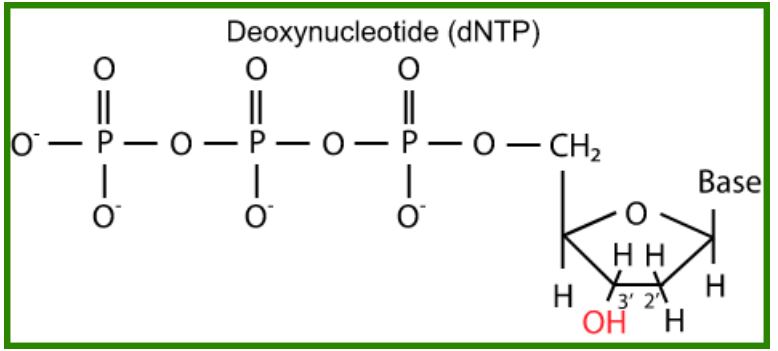


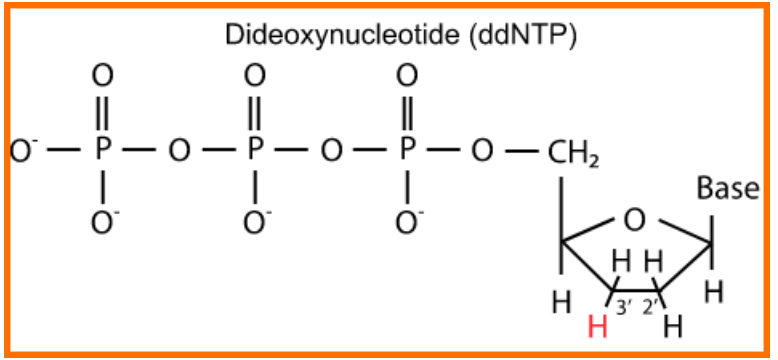


Секвенирование первого поколения: **Sanger sequencing**

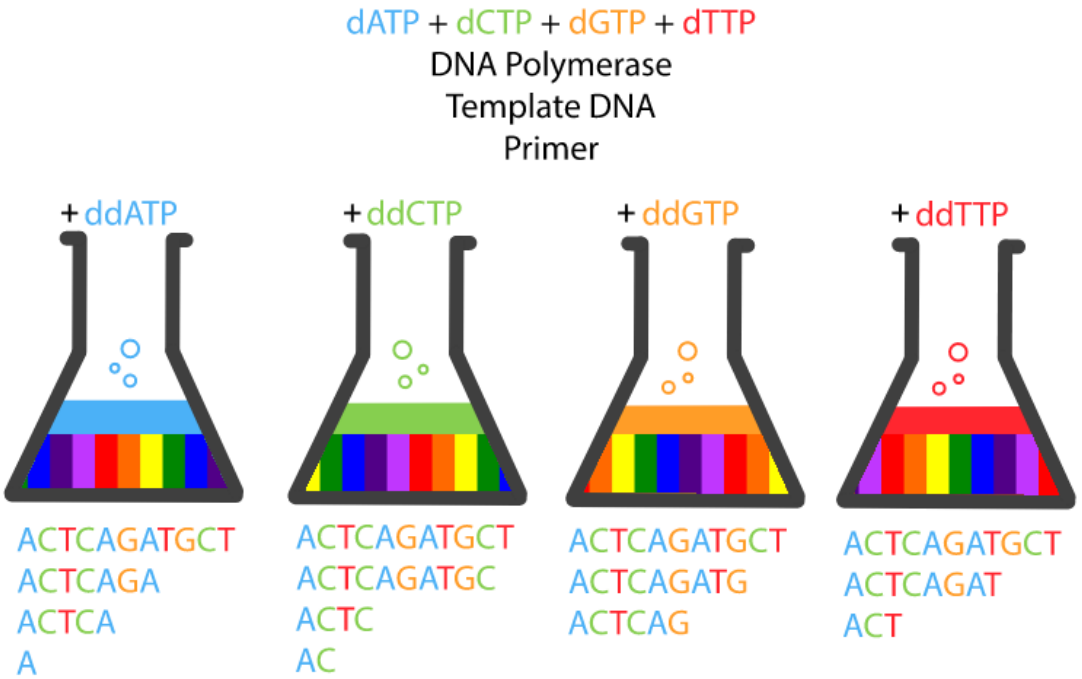
a.k.a. the **chain termination** method

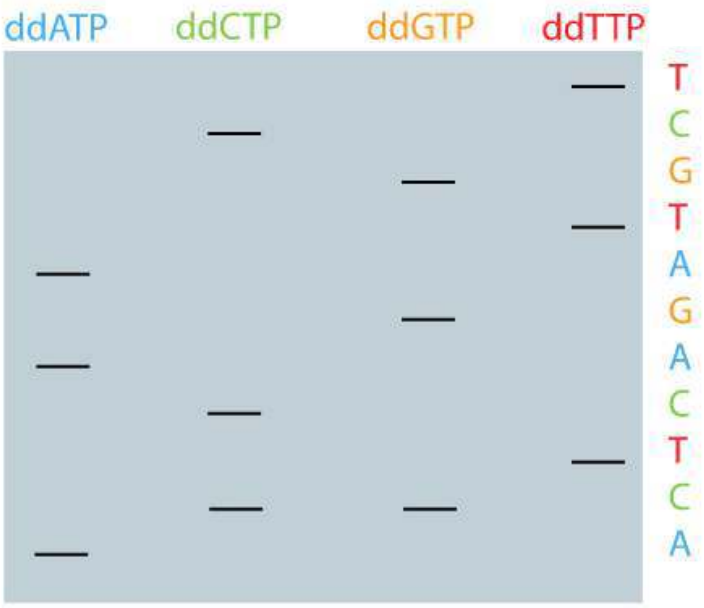






Секвенирование первого поколения: Sanger sequencing





+ Высокая точность (≥99.999%)
+ Длина рида 300-1000 bp
- Высокая стоимость (сотни долларов за Mbp)

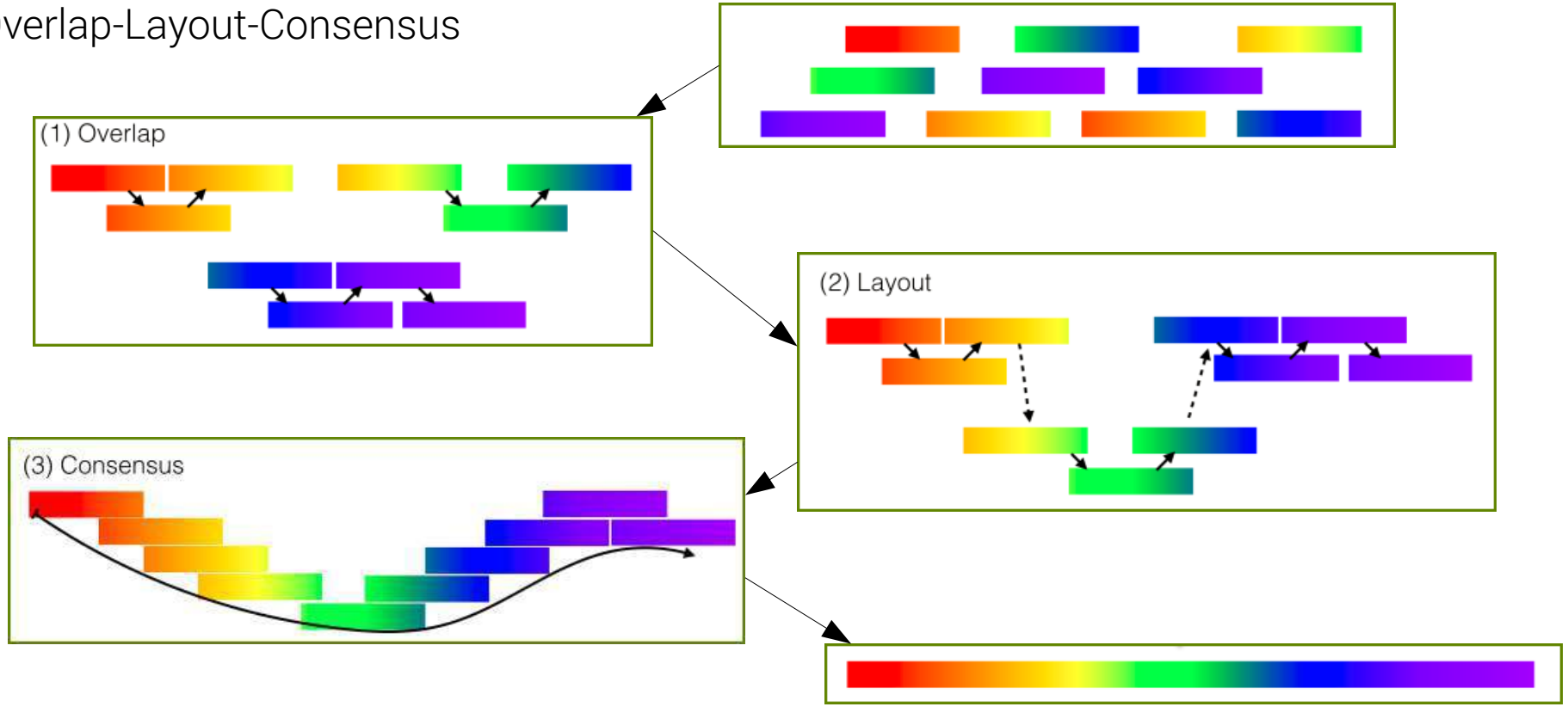
Секвенирование первого поколения: **применения**

- Секвенирование одиночных генов
- Валидация результатов, полученных с помощью секвенирований следующих поколений
- Секвенирование мелких геномов (вирусов, бактерий)
- …

- Когда-то крупные геномы было секвенировать нечем, кроме Сэнгера: the Human Genome Project

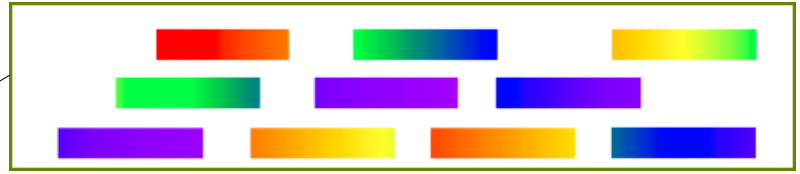
Секвенирование первого поколения: **сборка генома**

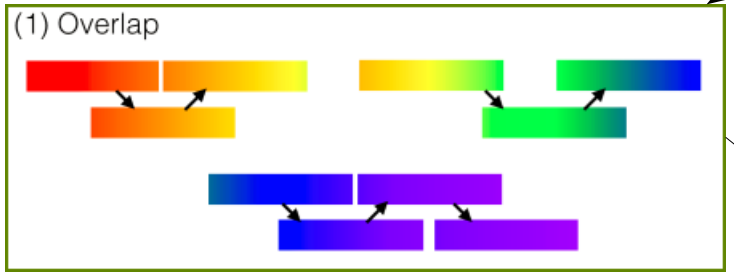
Overlap-Layout-Consensus

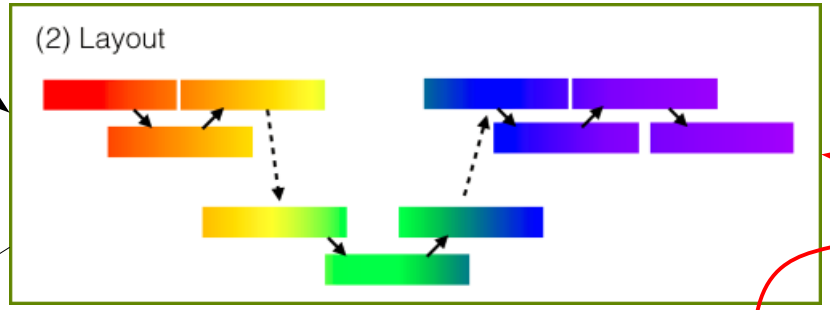


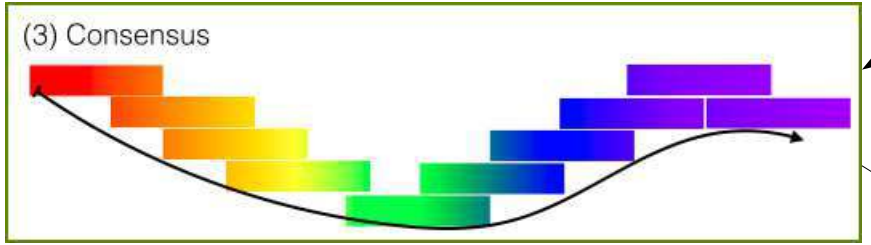
Секвенирование первого поколения: **сборка генома**

Overlap-Layout-Consensus



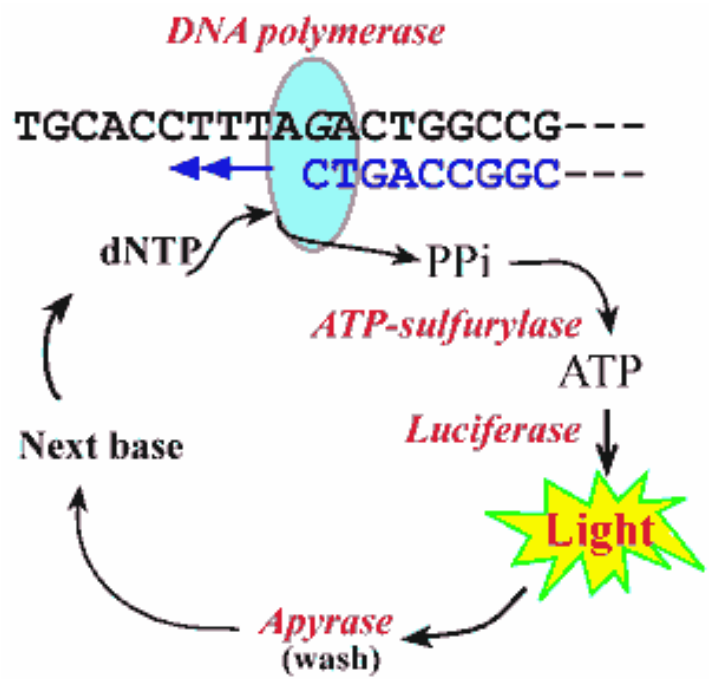


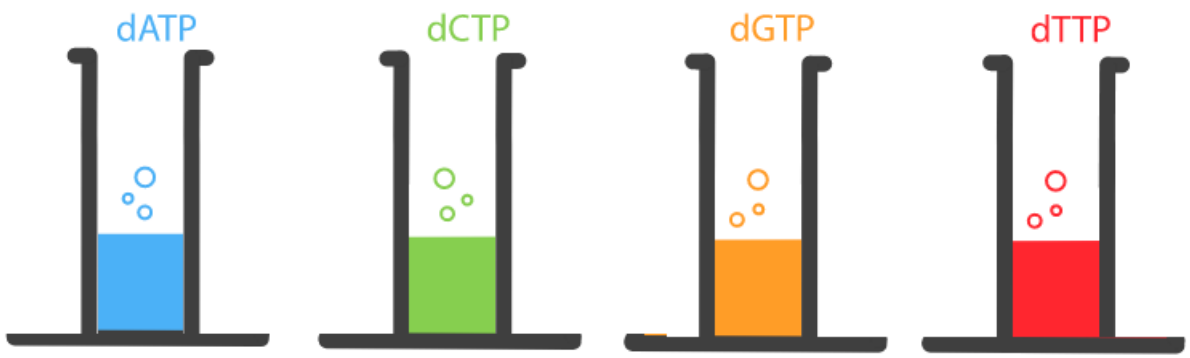


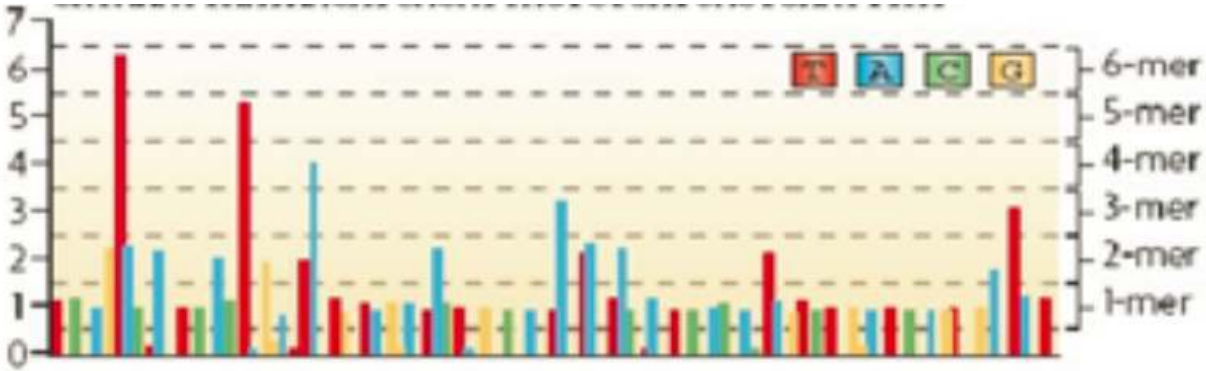


Гамильтонов путь :(

Секвенирование ~~нового~~ второго поколения (NGS): **пиросеквенирование**





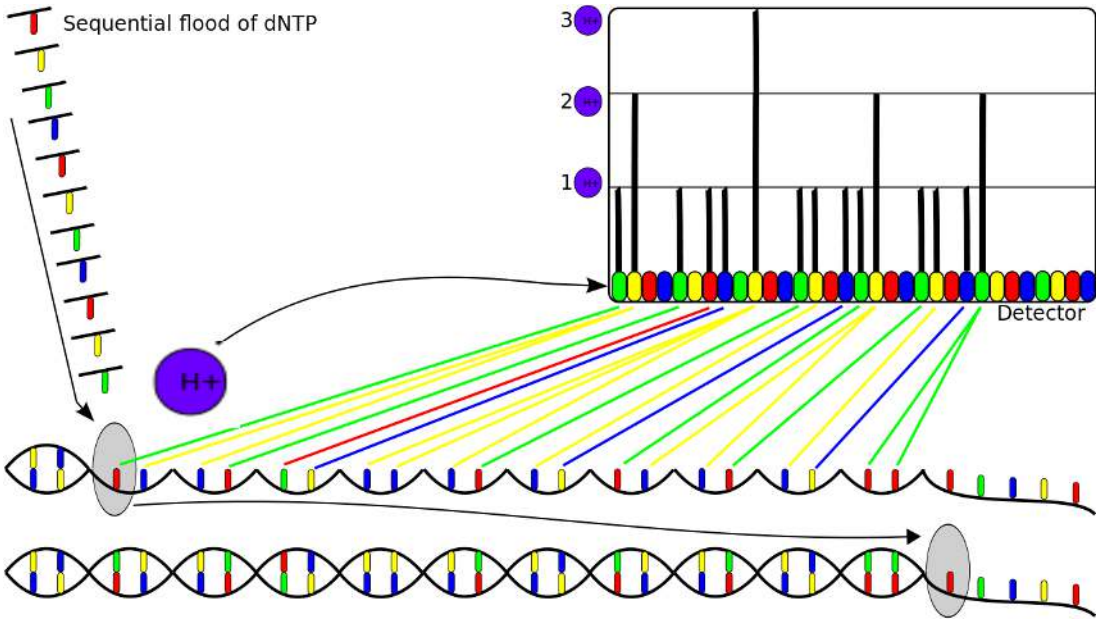


.. Риды длиной до 500 bp

- Точность ~99%, ошибки в гомополимерах

+ Намного дешевле (несколько долларов / десятки центов за Mbp)

Секвенирование ~~нового~~ второго поколения (NGS): ионное полупроводниковое

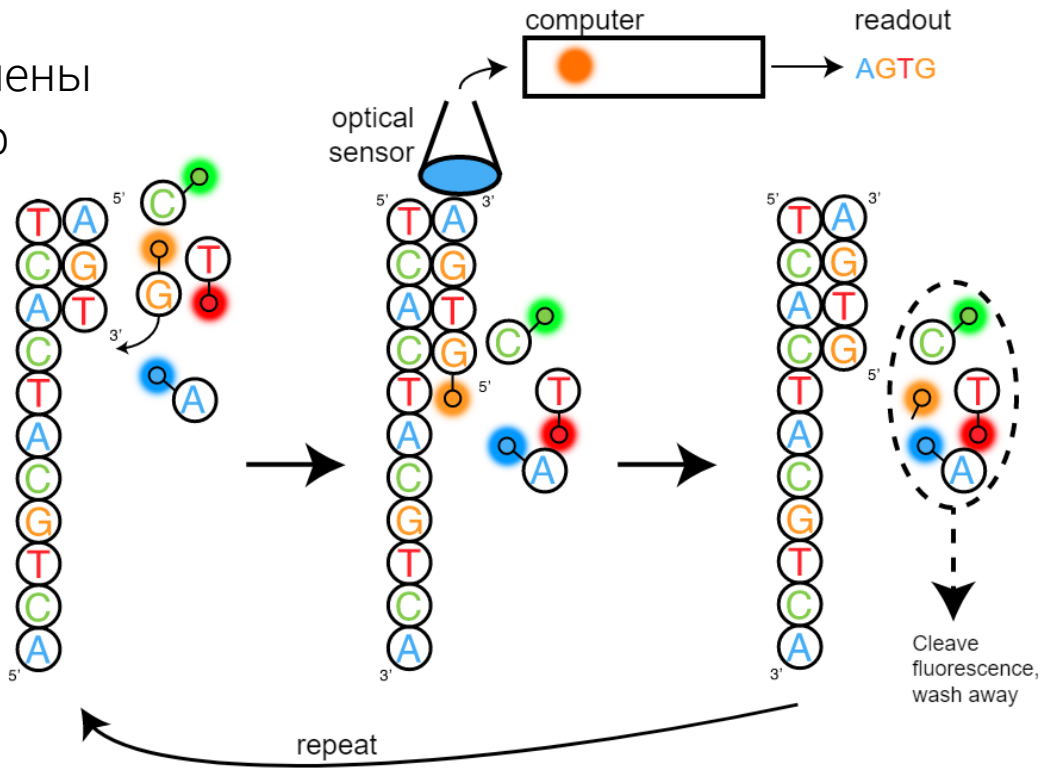


- Риды длиной до 400 bp
- Точность ≥99%, ошибки в гомополимерах

+ Еще дешевле (десятки центов за Mbp)

Секвенирование ~~нового~~ второго поколения (NGS): Illumina

- Длина ридов: 100~250 bp
.. Точность: 99.5~99.9%, ошибки замены
+ Смехотворная цена: центы за Mbp



Секвенирование ~~нового~~ второго поколения (NGS): коррекция ошибок

Для устранения ошибок замены:

- Строим базу данных k-меров
- Идем по ридам и стараемся, чтобы частота k-меров менялась плавно

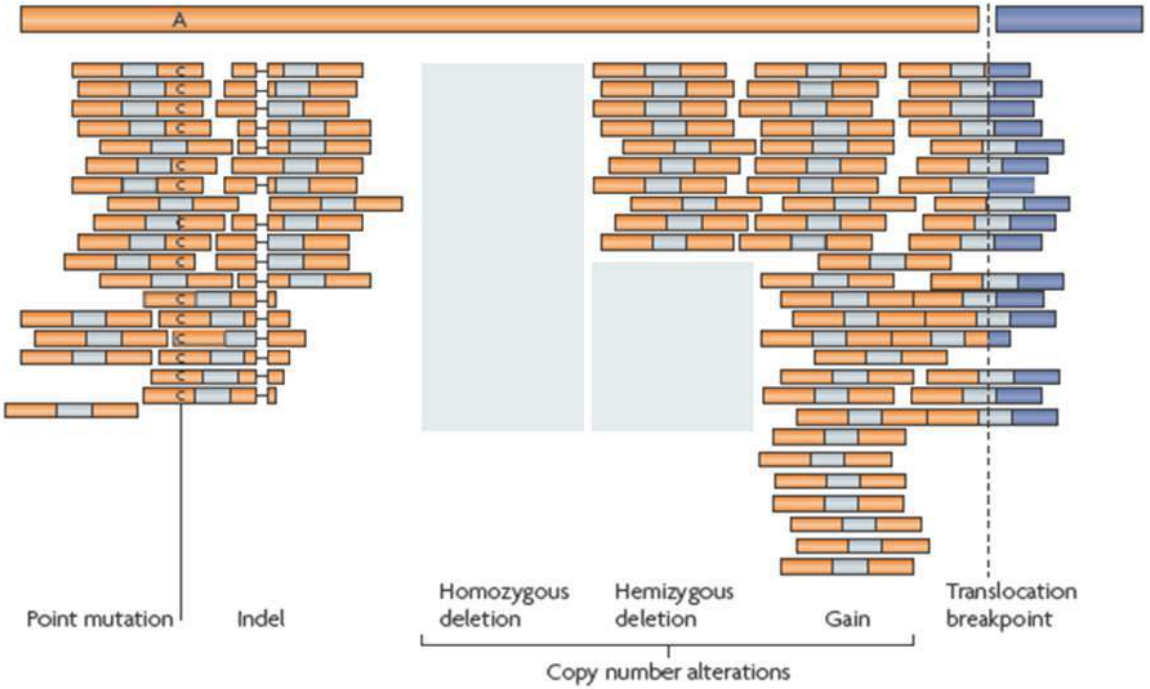
Алгоритм просто и емко описан в статье про QuorUM.
Отлично работает на данных Illumina.

Ошибки в гомополимерах нарушают плавность, но модификации k-mer counting алгоритмов существуют: Pollux.

Также есть IonHammer, основанный на байесовских моделях. Написан Антоном Коробейниковым.

Секвенирование ~~нового~~ второго поколения (NGS): **применения**

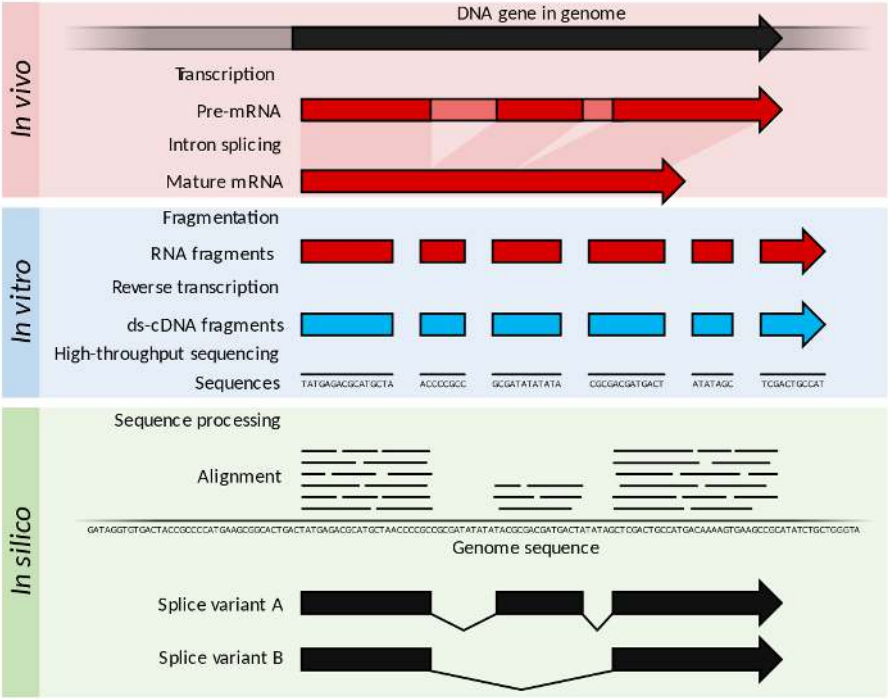
Ресеквенирование



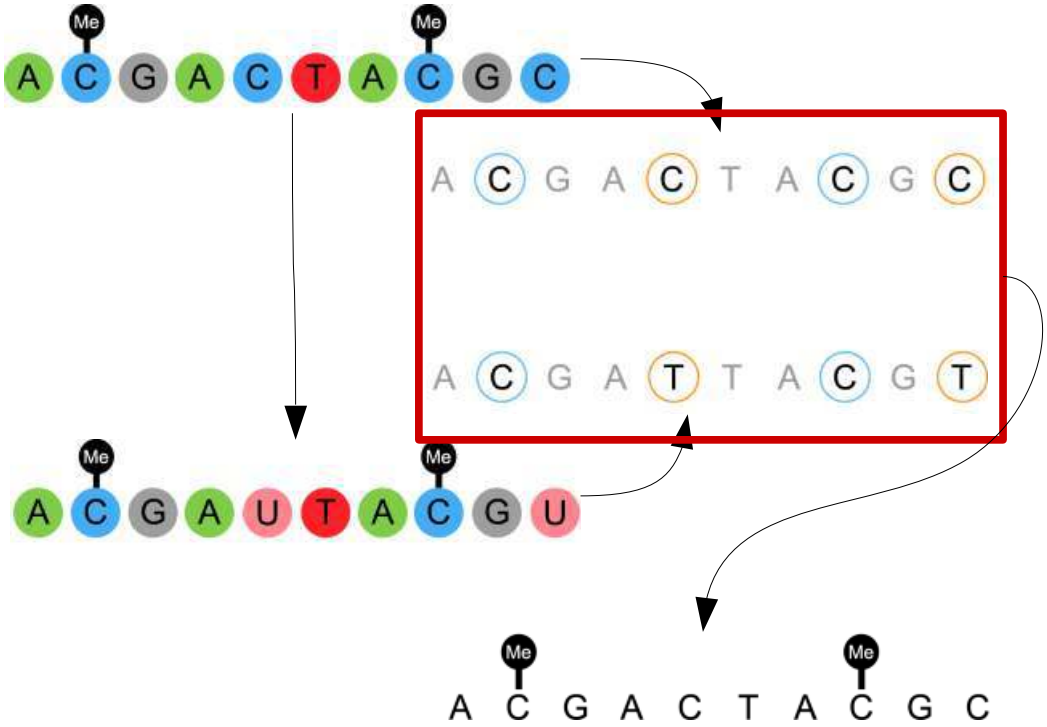
Таргетное секвенирование, экзомное и т.п.

Секвенирование ~~нового~~ второго поколения (NGS): **применения**

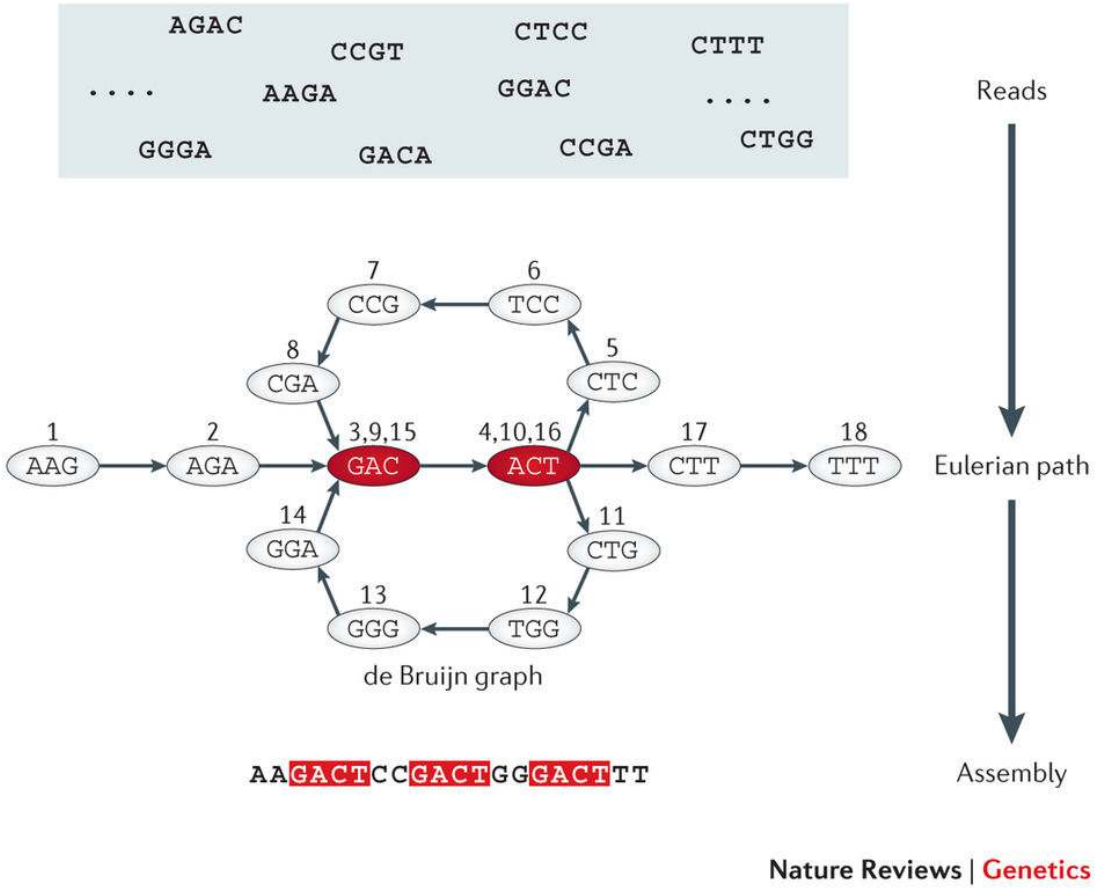
RNA-seq



Бисульфитное секвенирование



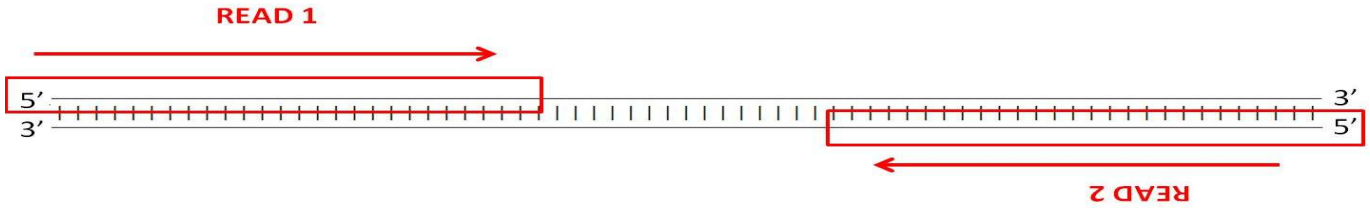
Секвенирование ~~нового~~ второго поколения (NGS): **сборка генома**



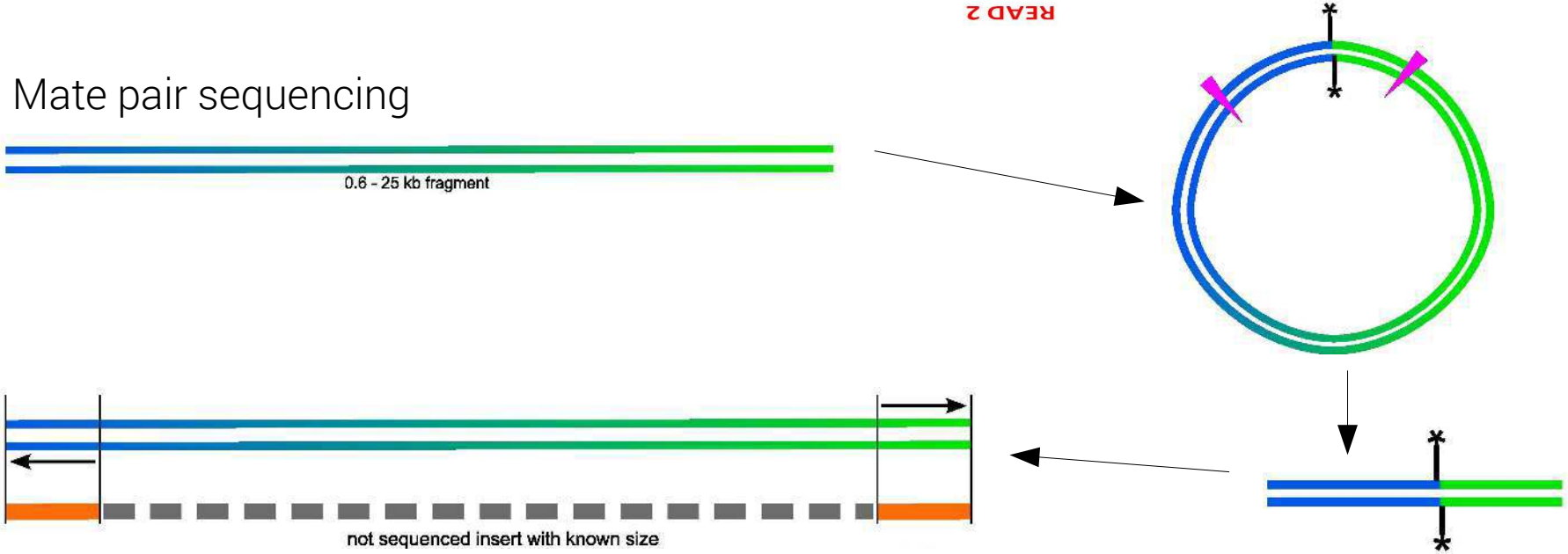
Nature Reviews | Genetics

Секвенирование ~~нового~~ второго поколения (NGS): PE и MP

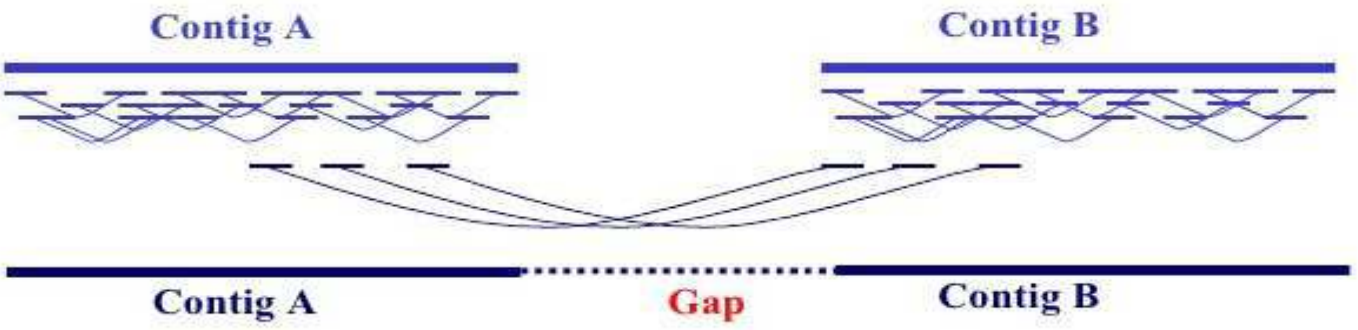
Paired end sequencing



Mate pair sequencing



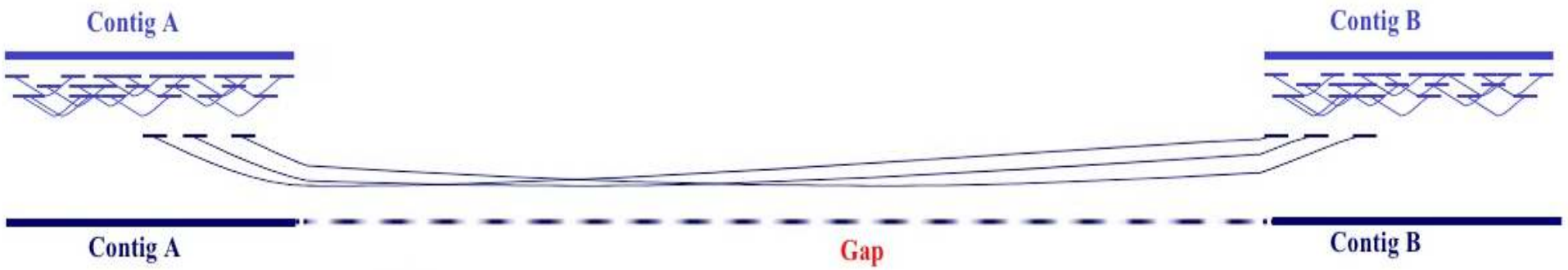
Секвенирование ~~нового~~ второго поколения (NGS): scaffolding

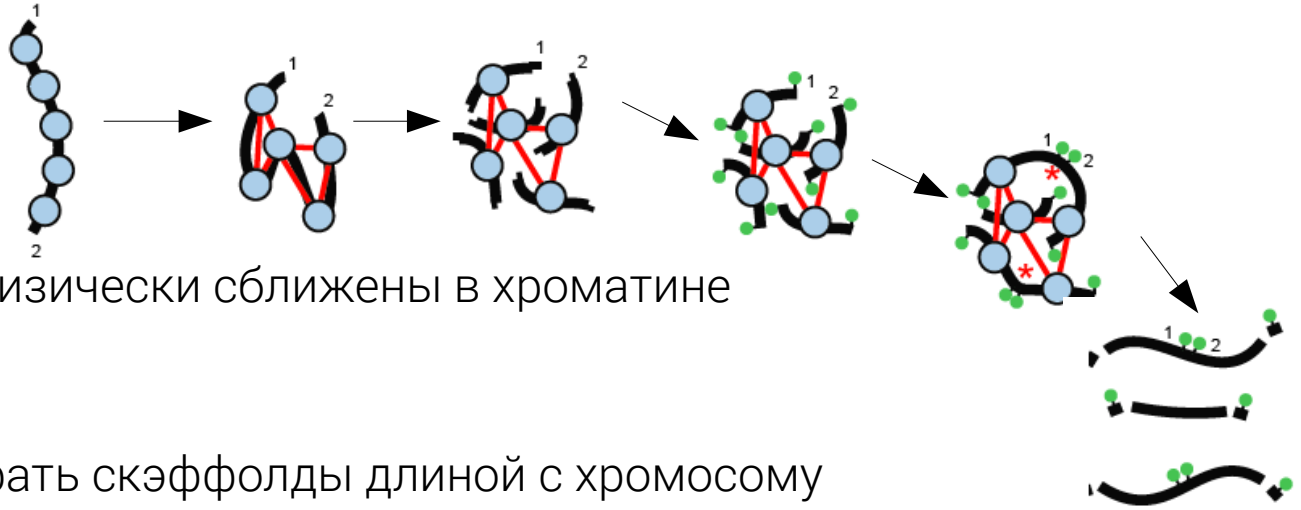


Гэпы размерами:

- 100 - 200 bp (при использовании PE)
- 600 - 20 000 bp (при использовании MP)

Секвенирование ~~нового~~ второго поколения (NGS): Hi-C (Dovetail)



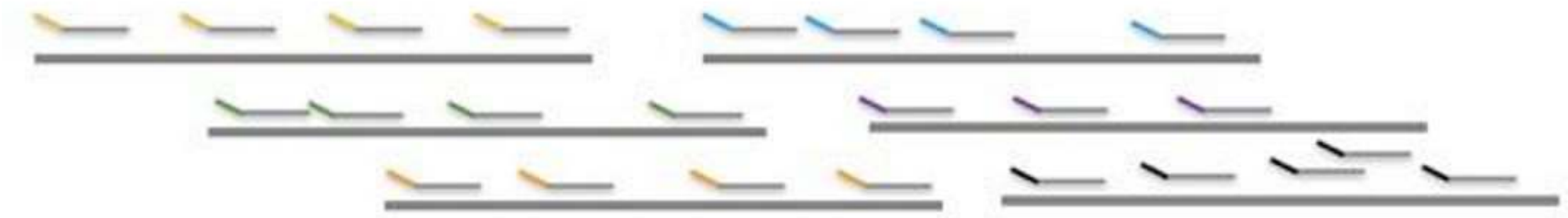


Использует тот факт, что
далекие участки генома физически сближены в хроматине
Гэпы размерами:

- порядка 100 000 bp

Практически можно собирать скэффолды длиной с хромосому

Секвенирование ~~нового~~ второго поколения (NGS): **linked reads (10X Genomics)**



NGS-риды с баркодами:
Все риды с одного участка ~50 Kbp длины имеют общий баркод.

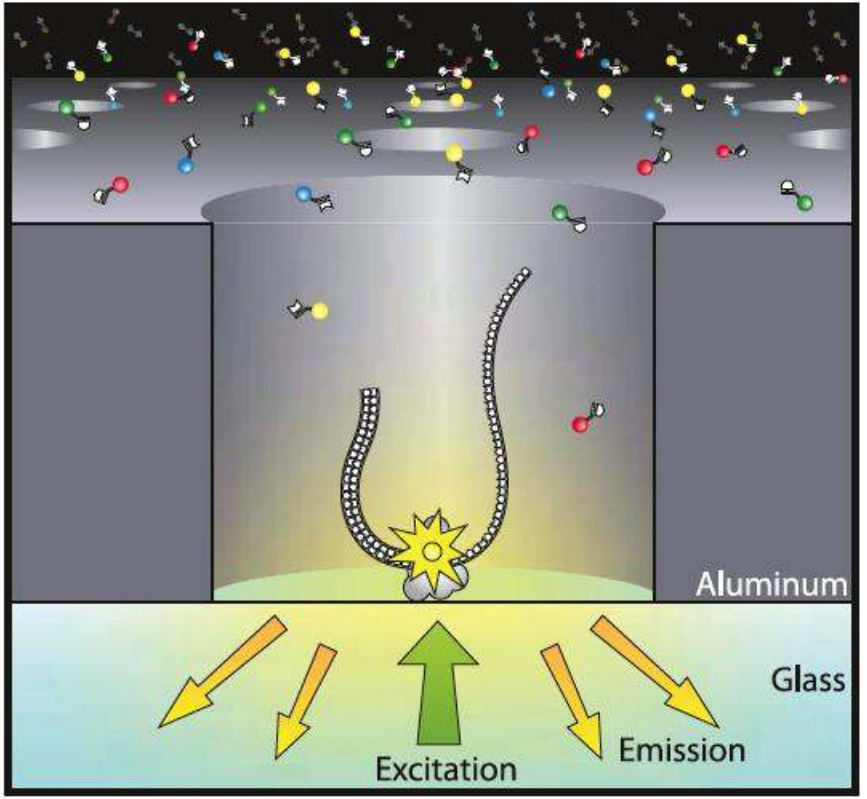
Можно:

- повышать качество *de novo* сборки Illumina
- искать структурные вариации; например, для изучения рака

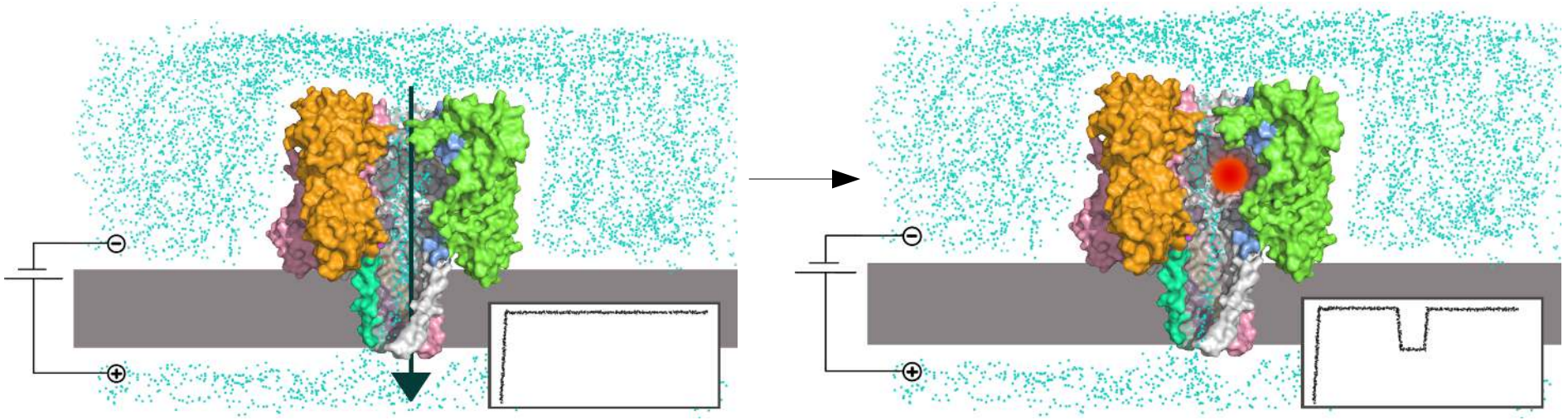
Секвенирование третьего поколения: PacBio

SMRT == single molecule real time

+ Длина ридов 10-50 Kbp
+ RT → можно определять метилирование
- Дороже, чем второе поколение
- Точность ~87%, ошибки вставки/делеции



Секвенирование третьего (четвертого??) поколения: Oxford Nanopore



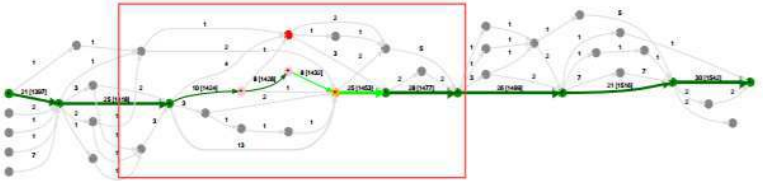
+ Может умещаться в кармане (одна из моделей — MinION)
+ Длина ридов якобы до 150 Kbp
+ Спектр сигналов → можно определять метилирование
- Дороже второго поколения (MinION доллары, PromethION десятки центов за Mbp)
- Точность тоже около 87% (еще недавно была 70%), ошибки делеций

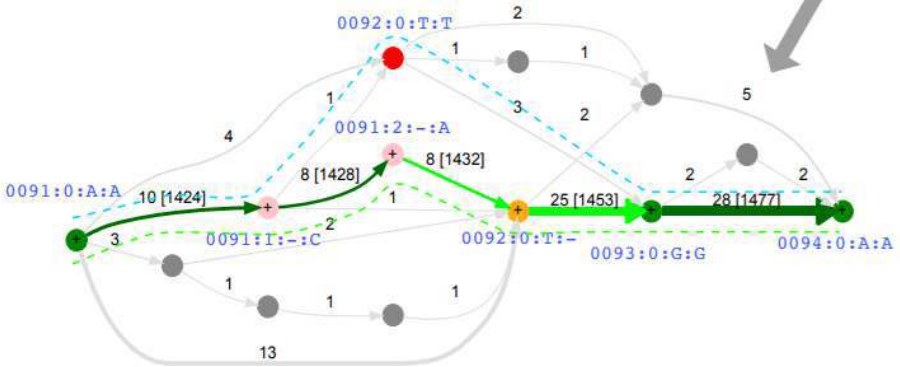
Секвенирование третьего поколения: **коррекция ошибок**

Canu, FALCON-sense

```
    p =                          p = 0123455678
    d =                          d = 0000001000
 seed = ATATTA-GGC            seed = ATATTA-GGC
read 1 = ATAT-ACGGC          read 1 = ATAT-ACGGC

    p =                          p = 011234555678
    d =                          d = 001000012000
 seed = AT-ATTA--GGC          seed = AT-ATTA--GGC
read 2 = ATCAT--CCGGC        read 2 = ATCAT--CCGGC
```





Seed Sequence Path	Consensus Sequence Path
0091:0:A:A	0091:0:A:A
0092:0:T:T	0091:1:-:C
0093:0:G:G	0091:2:-:A
0094:0:A:A	0092:0:T:-
	0093:0:G:G
	0094:0:A:A

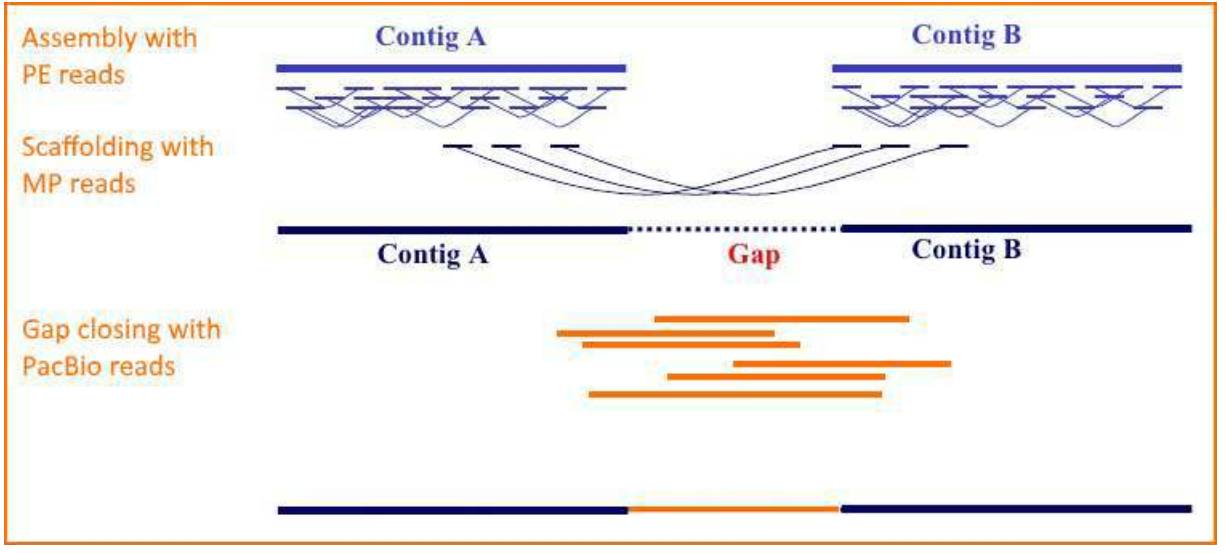
Seed Sequence		Consensus Sequence
ATGA	→	ACAGA

In this example, "T" insertion error corrected, and "CA" missing error recovered.

Секвенирование третьего поколения: scaffolding / gap closing

Gap Closing:

- PBJelly
- GMcloser

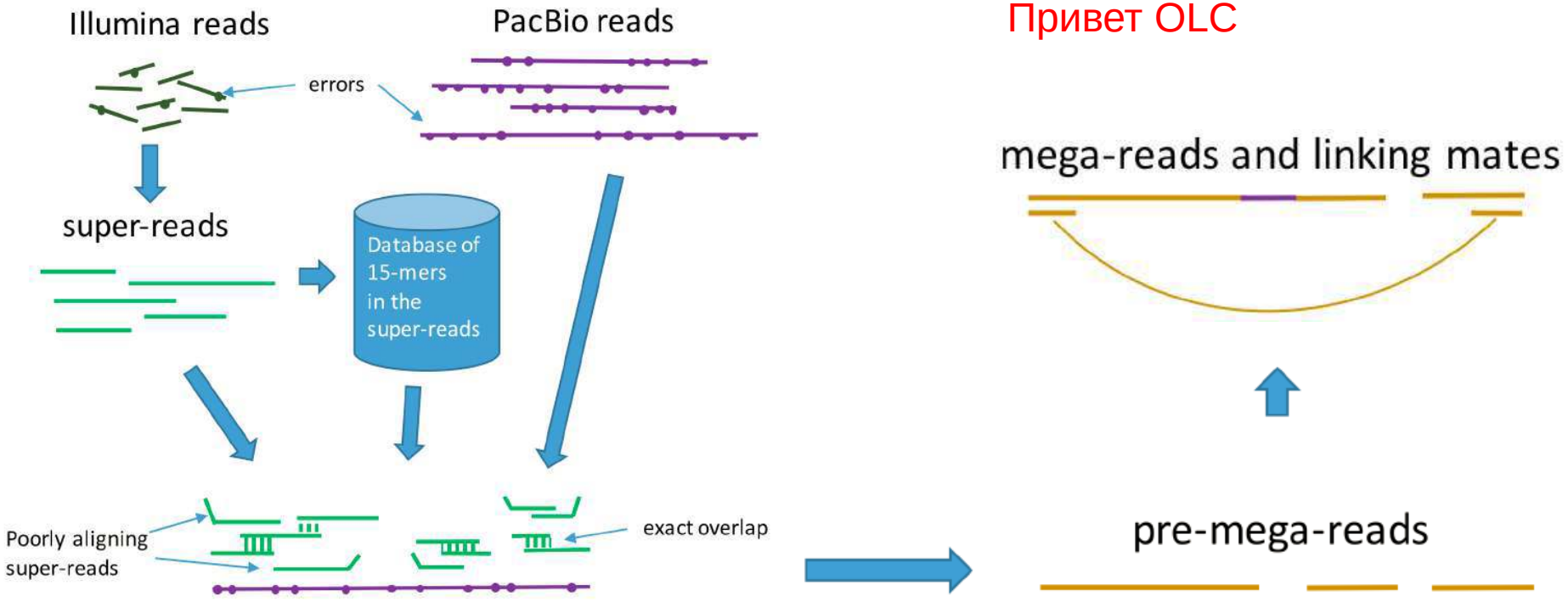


Scaffolding: теоретически возможно, однако предпочтение отдается гибридным методам (см. далее)

Секвенирование третьего поколения: **гибридная сборка**

MaSuRCA (Maryland Super-Read Celera Assembler)

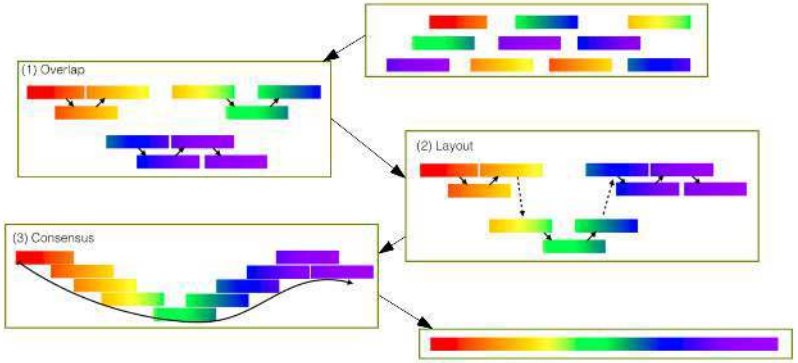
Привет OLC



Секвенирование третьего поколения: **сборка генома из длинных ридов**

С такой длиной и характером ошибок графы де Брюйна не работают.

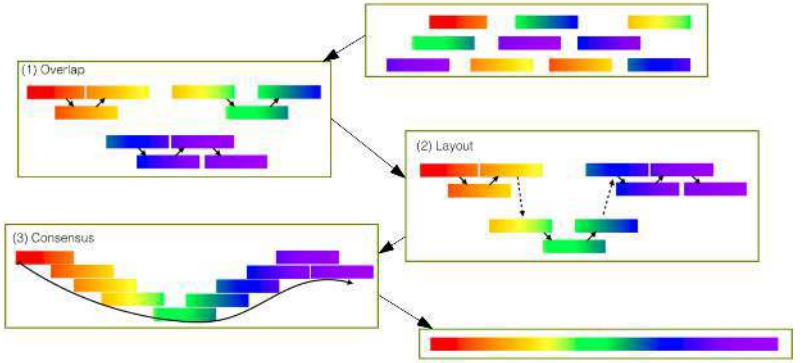
Модификации Overlap-Layout-Consensus с кучами новых хаков для увеличения скорости — работают



Секвенирование третьего поколения: **сборка генома из длинных ридов**

С такой длиной и характером ошибок графы де Брюйна не работают.

Модификации Overlap-Layout-Consensus с кучами новых хаков для увеличения скорости — работают:



- Canu —
  работает уже с 20~30 покрытием,
  скалируется до геномов длиной в гигабазы
- HINGE —
  собирает бактериальные геномы из покрытия от 30x
- RACON —
  пока что, похоже, самый быстрый, черная магия, но нужно от 50x

Секвенирование в 2017 году: **сборка больших геномов**

1. Секвенирование PacBio или ONT, желательно от 30x и выше
2. Коррекция входных данных (например*, в Canu есть соответствующий модуль)
3. **Сборка** откорректированных данных (например*, тем же Canu)
4. **Полировка** сборки ридами Illumina (например*, с помощью Pilon)
5. **Long-range scaffolding** с помощью Hi-C (Dovetail)

* например, на всё существуют тысячи разных тулов

Секвенирование в 2017 году: **сборка больших геномов**

1. Секвенирование PacBio или ONT, желательно от 30x и выше
2. Коррекция входных данных (например*, в Canu есть соответствующий модуль)
3. **Сборка** откорректированных данных (например*, тем же Canu)
4. **Полировка** сборки ридами Illumina (например*, с помощью Pilon)
5. **Long-range scaffolding** с помощью Hi-C (Dovetail)

* например, на всё существуют тысячи разных тулов

Например, HINGE отказывается от исходной коррекции:
overlaps удается находить даже при уровне ошибок 15%

Например, RACON утверждает, что он сам себе и консенсус, и корректор.

Секвенирование третьего поколения: **другие применения**

- RNA-seq: позволяет прочитать сплайс-вариант целиком, а не собирать его по кускам
- Поиск вариантов: как минимум для маленьких геномов или при таргетном секвенировании
- Поиск структурных вариаций
- Профили метилирования (см. выше)
- ...

Технологии секвенирования

Спасибо за внимание