

# Next-generation sequencing

## Высокопроизводительное секвенирование

Лаборатория для исследования древних  
ДНК открыта в СО РАН

«Агентство Химэксперт»



Next-generation sequencing  
Высокопроизводительное  
секвенирование

Лаборатория для исследования древних  
ДНК открыта в СО РАН

«Агентство Химэксперт»



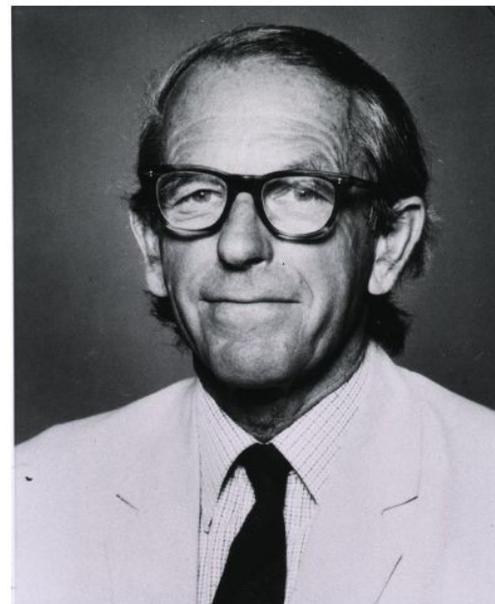
## Введение в высокопроизводительное секвенирование

- Эволюция секвенирования: от Сэнгера к NGS
- Ключевые понятия NGS
- Этапы подготовки и проведения высокопроизводительного секвенирования

«Агентство Химэксперт»



# Поколения секвенирования

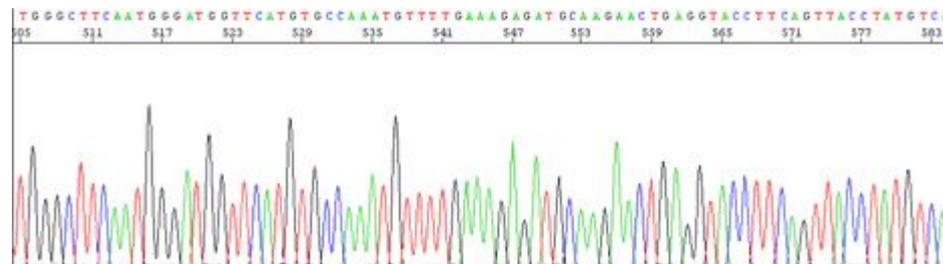
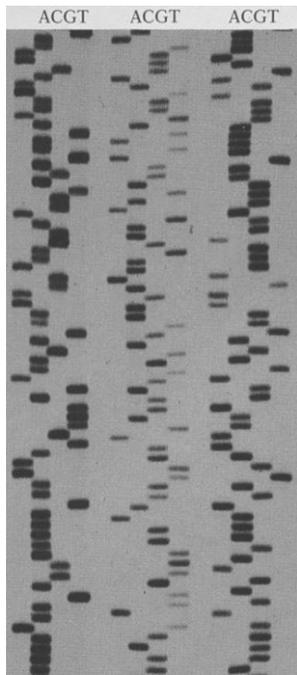


**1918 - 2013**

**Фредерик Сэнгер**

# Поколения секвенирования

## I поколение метод Сэнгера



# Поколения секвенирования

## II поколение

### Next-generation sequencing



Позволяет определять нуклеотидную последовательность принципиально большей общей протяженности за один рабочий цикл по сравнению с методами секвенирования предыдущего поколения (методы Сэнгера и Максама-Гилберта)



# Next-generation sequencing

секвенирование нового поколения

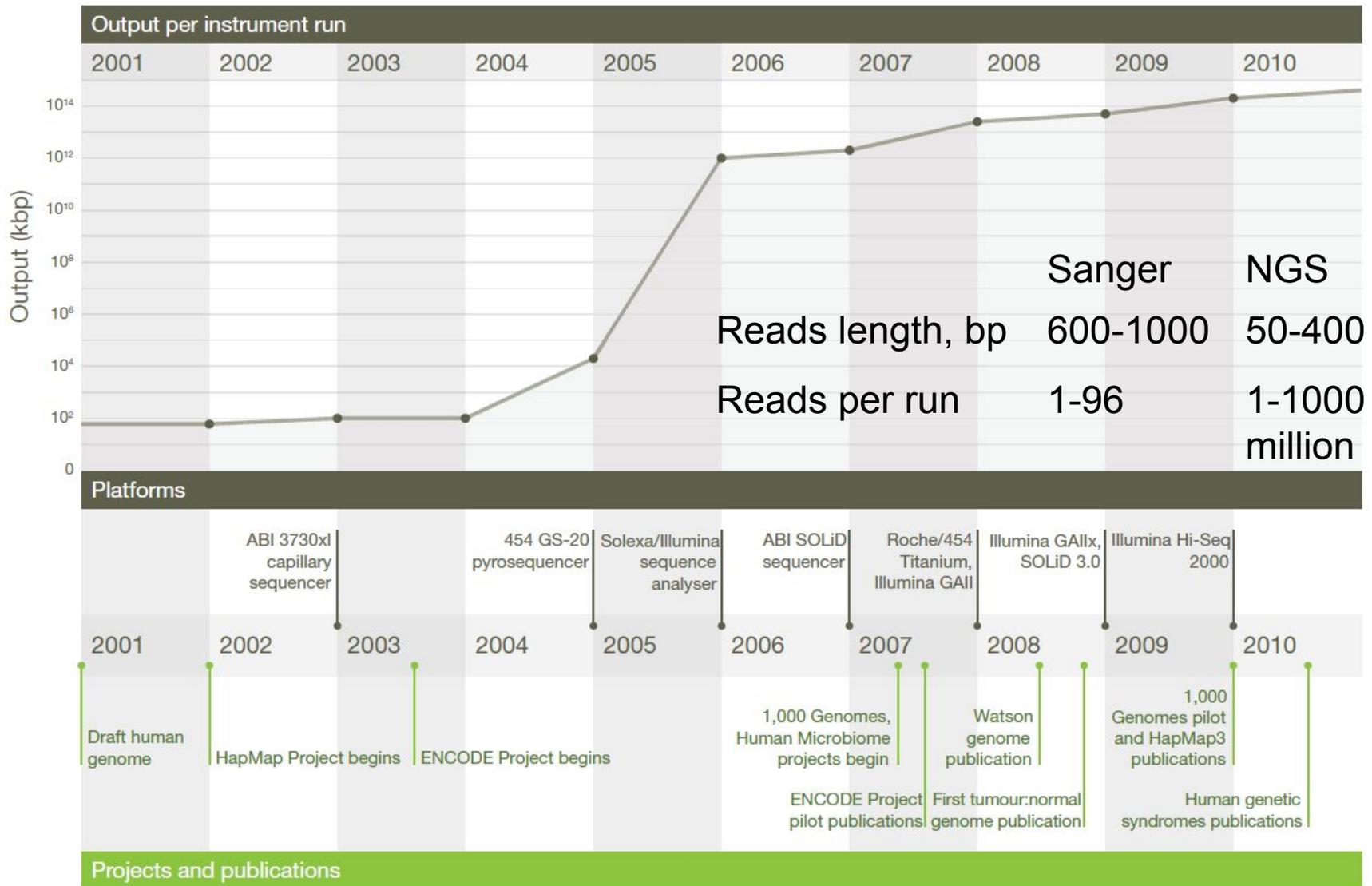
секвенирование следующего поколения

высокопроизводительное секвенирование

- определение нуклеотидной последовательности ДНК, основанная на параллельном «чтении» перекрывающихся фрагментов исходной молекулы.



# Эволюция секвенирования



A decade's perspective on DNA sequencing technology. ER Mardis - Nature, 2011

# Технология NGS позволяет проводить:



- Секвенирование отдельных генов или наборов генов
- Массовое секвенирование ампликонов
- Секвенирование транскриптомов
- Секвенирование экзонов
- Полногеномное секвенирование

# NGS. Общий принцип

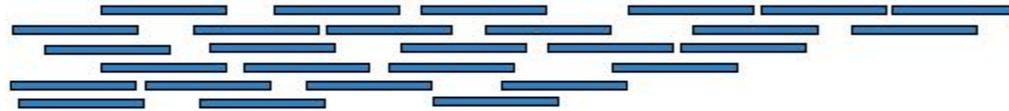
- ДНК нарезается на фрагменты определенной длины
- К ним лигируются адаптеры
- Амплификация каждого отдельного фрагмента в изолированных от других условиях
- Анализ последовательности амплифицированных клонов ДНК

# NGS

**Multiple Copies of a Genome**



**Reads**



**High Coverage**

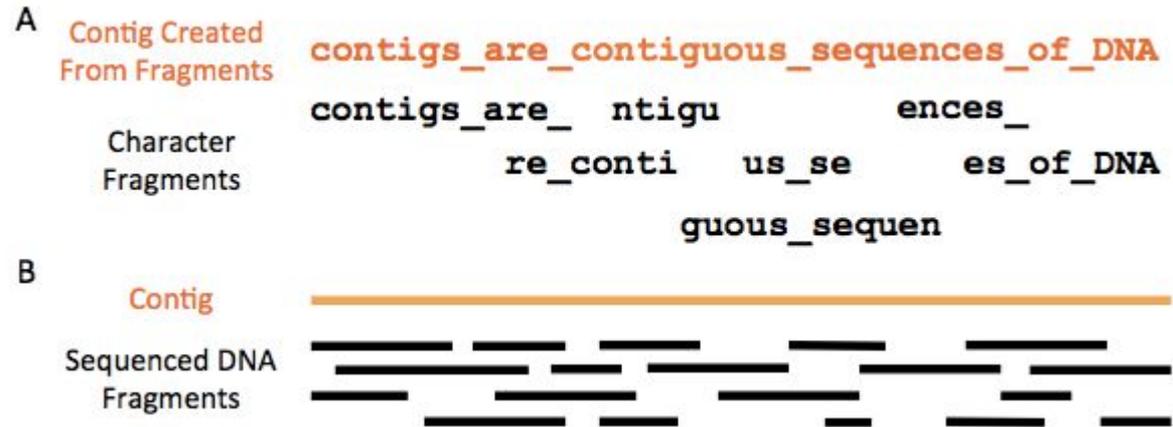
**Low Coverage**



**Consensus Sequence**



# NGS

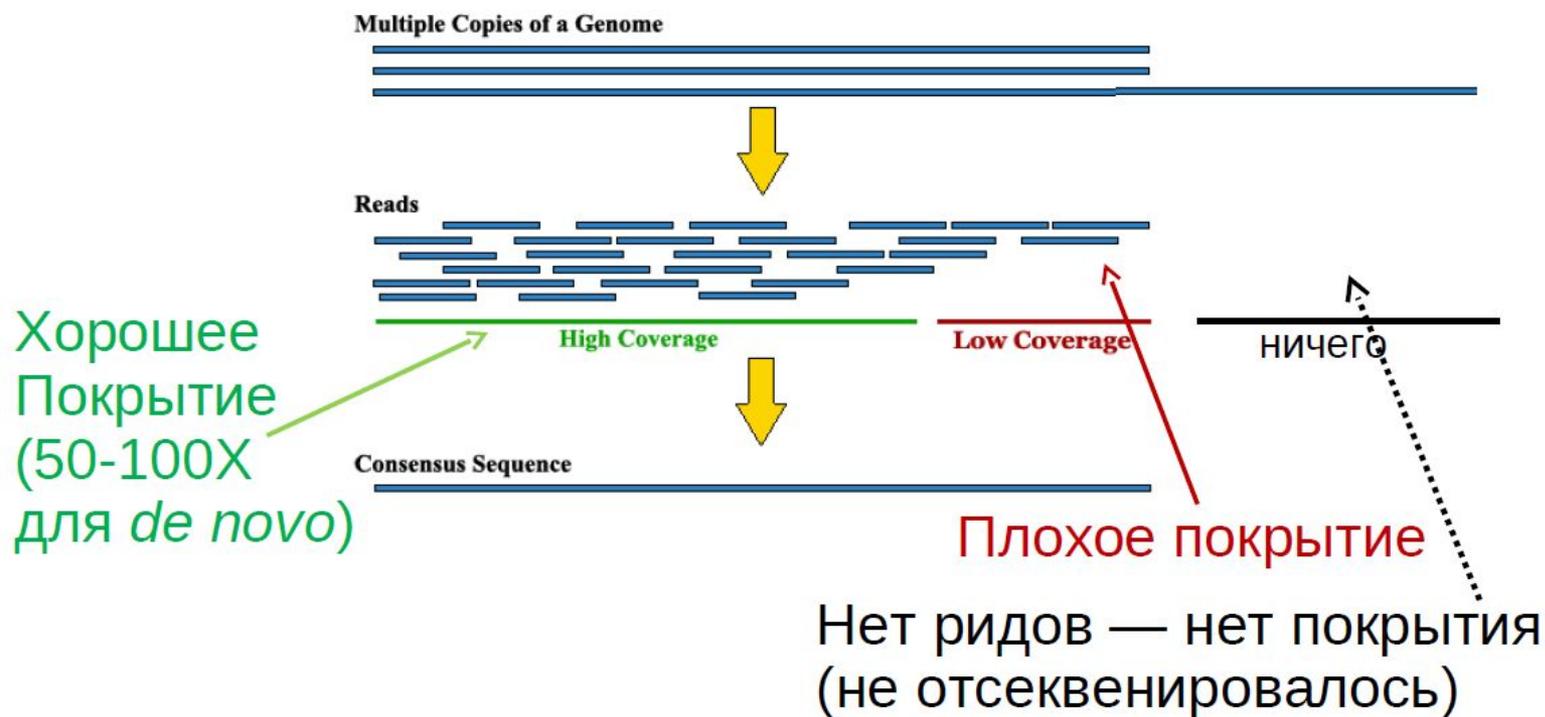


# NGS

	ATGGCATTGCAA
	TGGCATTGCAATTTG
	AGATGGTATTG
Reads	GATGGCATTGCAA
	GCATTGCAATTTGAC
	ATGGCATTGCAATTT
	AGATGGTATTGCAATTTG
Consensus	
Sequence	AGATGGCATTGCAATTTGAC

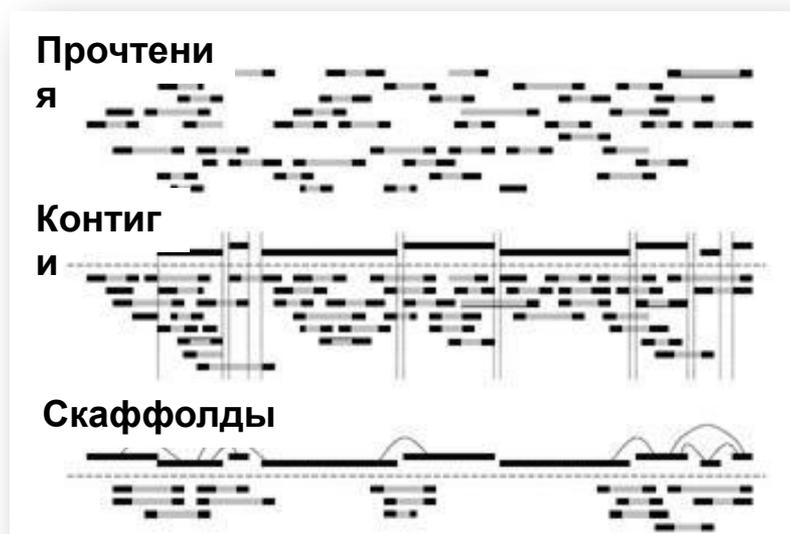
# Что такое покрытие (coverage)

- Это сколько раз в среднем нуклеотид генома покрыт ридами

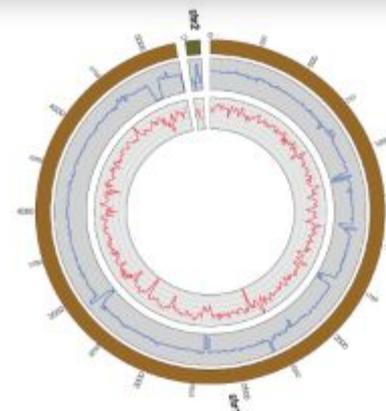


# *de novo* сборка генома

- **Сборка последовательности** - выравнивание и слияние фрагментов более длинной последовательности ДНК для восстановления исходной последовательности
- *de novo* сборка выполняется в случае, если отсутствует референсная последовательность для сравнения
- При гибридной сборке можно использовать элементы близкородственной или частично референсной последовательности



**Собранн  
й  
геном**



Circos plot of a bacterial genome assembly.  
Comparison between a german *E. coli* outbreak strain and reference strain *E. Coli* 559889

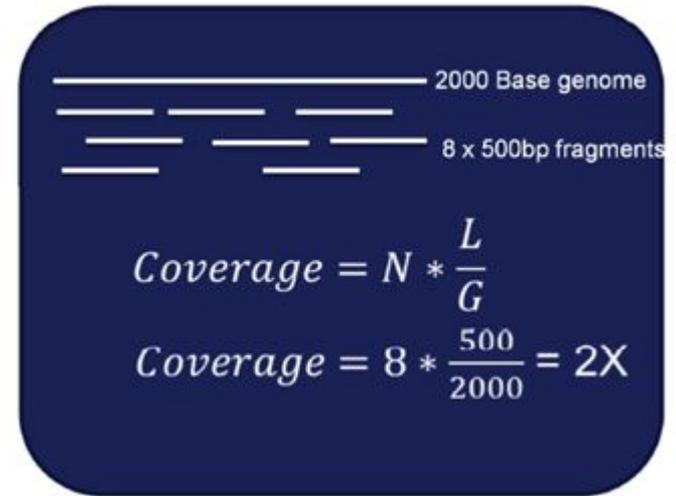


# Пример контига, собранного по данным Torrent Data в программе MIRA

```
Contig Editor: +611979 B66P8:2113:1545
Cons 2  Qual 0  Insert  Edit Modes >>  Cutoffs  Undo  Next Search  Commands >>  Settings >>  Quit  Help >>
191720  191730  191740  191750  191760  191770  191780  191790  191800  191810  191820  191830  191840
+674614 B66P8:1027:158 CTTATTCGCCCCGGAAGCTACCGGTACGGTCTGGTTATTTACAGAAGCAATGCTAAACGCCACGGTATG
-674615 B66P8:1620:205 CTTATTCGCCCCGGAAGCTACCGGTACGGTCTGGTTATTTACAGAAGCAATGCTAAACGCCACGGTATG
+674616 B66P8:570:256 CTTATTCGCCCCGGAAGCTACCGGTACGG
+674617 B66P8:1752:221 CTTATTCGCCCCGGAAGCTACCGGTACGGTCTGGTTATTTACAGAAGCAATGCTAAACGCCACGGTATGGGTTTTGA
-674618 B66P8:1578:124 CTTATTCGCCCCGGAAGCTACCGGTACGGTCTGGTTATTTACAGAAGCAATGCTAAACGCCACGGTATGGGTTTTGAAGGGATGCGCGTTT
-674619 B66P8:2173:208 CTTATTCGCCCCGGAAGCTACCGGTACGGTCTGGTTATTTACAGAAGCAATGCTAAACGCCACGGTATGGGTTTTGA
+674620 B66P8:1479:134 CTTATTCGCCCCGGAAGCTACCGGTACGGTCTGGTTATTTACAGAAGCAATGCTAAACGCCACGGTATGGGTTTTGAAGGGATGCGCGT
+674621 B66P8:2351:882 CTTATTCGCCCCGGAAGCTACCGGTACGGTCTGGTTATTTACAGAAGCAATGCTAAACGCCACGGTATGGGTTTTGAAGGGATGCGCGTTT
+674622 B66P8:633:1989 CTTATTCGCCCCGGAAGCTACCGGTACGGTCTGGTTATTTACAGAAGCAATGCTAAACGCCACGGTATGGGTTTTGAAGGGATGCGCGTTTCCG
+674623 B66P8:1394:643 CTTATTCGCCCCGGAAGCTACCGGTACGGTCTGGTTATTTACAGAAGCAATGCTAAACGCCACGGTATGGGTTTTGAAGG
-674624 B66P8:117:1772 CTTATTCGCCCCGGAAGCTACCGGTACGGTCTGGTTATTTACAGAAGCAATGCTAAACGCCACGGTATGGGTTTTGAAGGGATGCGCGTTTCCGTTTCT
-674625 B66P8:1952:239 TATTCGCCCCGGAAGCTACCGGTACGGTCTGGTTATTTACAGAAGCAATGCTAAACGCCACGGTATGGGTTTTGAAGGGATGCGCGTTTCCGTTTCTGGCT
+674626 B66P8:2002:207 TATTCGCCCCGGAAGCTACCGGTACGGTCTGGTTATTTACAGAAGCAATGCTAAACGCCACGGTATGGGTTTTGAAGGGATGCGCGTTTCCGTTTCTGGCTCCGGCAAC
+674627 B66P8:2022:972 GAAGCTACCGGTACGGTCTGGTTATTTACAGAAGCAATGCTAAACGCCACGGTATGGGTTTTGAAGGGATGCGCGTTTCCGTTTCTGGCTCCGGCAACGTC
-674628 B66P8:1095:435 AAGCTACCGGTACGGTCTGGTTATTTACAGAAGCAATGCTAAACGCCACGGTATGGGTTTTGAAGGGATGCGCGTTTCCGTTTCTGGCTCCGGCAACGTCGCCAGTACG
-674629 B66P8:178:1206 GCTACGGTCTGGTTATTTACAGAAGCAATGCTAAACGCCACGGTATGGGTTTTGAAGGGATGCGCGTTTCCGTTTCTGGCTCCGGCAACGTCGCCAGTACG
-674630 B66P8:1583:149 GCTACGGTCTGGTTATTTACAGAAGCAATGCTAAACGCCACGGTATGGGTTTTGAAGGGATGCGCGTTTCCGTTTCTGGCTCCGGCAACGTCGCCAGTACG
+674631 B66P8:1307:176 TACGGTCTGGTTATTTACAGAAGCAATGCTAAACGCCACGGTATGGGTTTTGAAGGGATGCGCGTTTCCGTTTCTGGCTCCGGCAACGTCGCCAGTACG
+674632 B66P8:2487:969 TACGGTCTGGTTATTTACAGAAGCAATGCTAAACGCCACGGTATGGGTTTTGAAGGGATGCGCGTTTCCGTTTCTGGCTCCGGCAACGTCGCCAGTACG
-674633 B66P8:1177:133 CTGGTTATTTACAGAAGCAATGCTAAACGCCACGGTATGGGTTTTGAAGGGATGCGCGTTTCCGTTTCTGGCTCCGGCAACGTCGCCAGTACG
-674634 B66P8:1181:215 TTATTTACAGAAGCAATGCTAAACGCCACGGTATGGGTTTTGAAGGGATGCGCGTTTCCGTTTCTGGCTCCGGCAACGTCGCCAGTACG
-674635 B66P8:552:1819 TTACAGAAGCAATGCTAAACGCCACGGTATGGGTTTTGAAGGGATGCGCGTTTCCGTTTCTGGCTCCGGCAACGTCGCCAGTACG
-674636 B66P8:387:997 CAGAAGCAATGCTAAACGCCACGGTATGGGTTTTGAAGGGATGCGCGTTTCCGTTTCTGGCTCCGGCAACGTCGCCAGTACG
+674637 B66P8:1759:198 AGAAGCAATGCTAAACGCCACGGTATGGGTTTTGAAGGGATGCGCGTTTCCGTTTCTGGCTCCGGCAACGTCGCCAGTACG
-674638 B66P8:964:217 AATGCTAAACGCCACGGTATGGGTTTTGAAGGGATGCGCGTTTCCGTTTCTGGCTCCGGCAACGTCGCCAGTACG
+674639 B66P8:2001:167 AATGCTAAACGCCACGGTATGGGTTTTGAAGGGATGCGCGTTTCCGTTTCTGGCTCCGGCAACGTCGCCAGTACG
+674640 B66P8:1220:254 GCTAAACGCCACGGTATGGGTTTTGAAGGGATGCGCGTTTCCGTTTCTGGCTCCGGCAACGTCGCCAGTACG
-674641 B66P8:2061:117 CTTATTCGCCCCGGAAGCTACCGGTACGGTCTGGTTATTTACAGAAGCAATGCTAAACGCCACGGTATGGGTTTTGAAGGGATGCGCGTTTCCGTTTCTGGCTCCGGCAACGTCGCCAGTACG
<>  CONSENSUS  -%>
Tag type: SVEC  Direction: -  Comment: **
```

# Что такое покрытие?

- Среднее количество прочтений, содержащих данный нуклеотид в реконструированной последовательности
- Позволяет оценить % покрытия генома
- Высокое покрытие исправляет ошибки при сборке



$N$  = Кол-во прочтений  
 $L$  = средняя длина прочтений  
 $G$  = Длина прочитываемого участка (генома)

Типичное желательное покрытие генома 30x



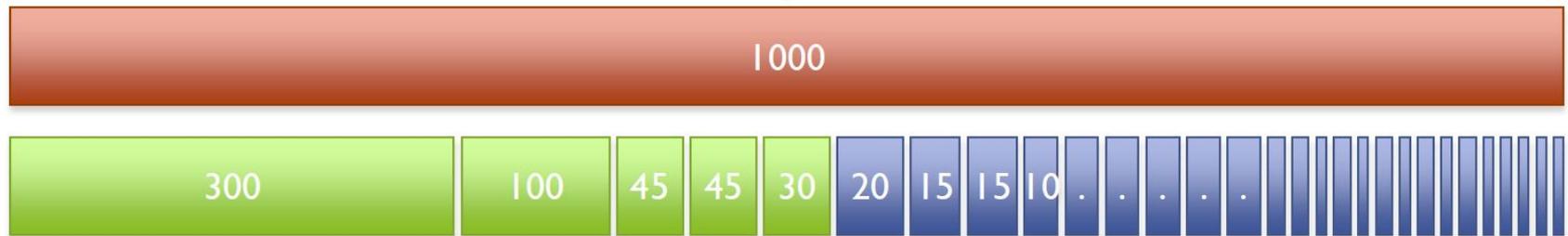
# Что такое N50

Example: 1 Mbp genome

50%



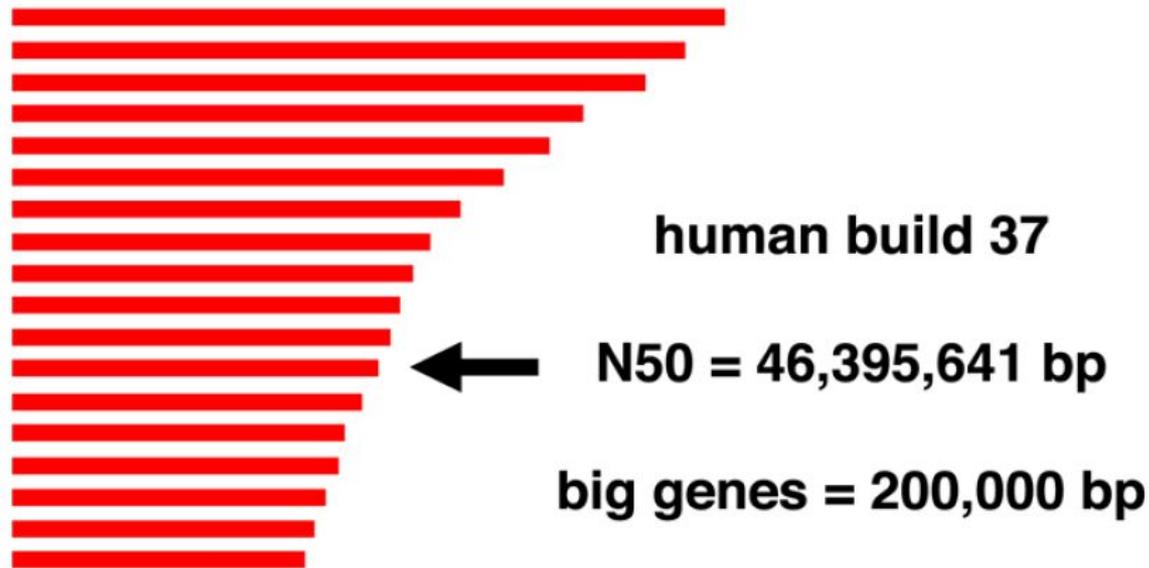
1000



N50 size = 30 kbp

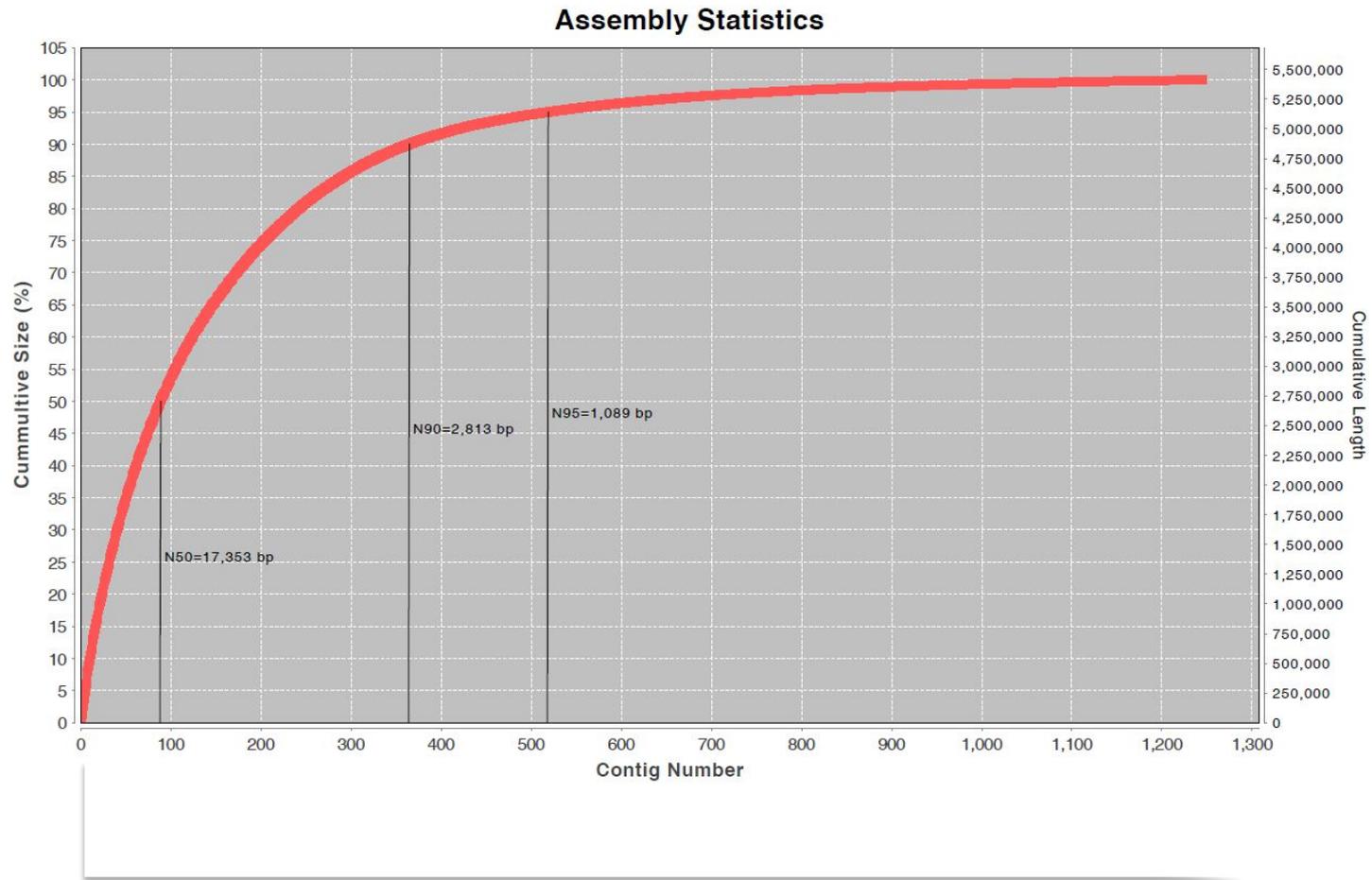
$(300k + 100k + 45k + 45k + 30k = 520k \geq 500kbp)$

# Что такое N50



- N50 показывает качество сборки
- Скаффолды располагают по убыванию длины
- Суммируют длину, начиная с самого большого скаффолда
- На каком скаффолде покроем половину генома?
- ✓ Длина этого скаффолда называется N50.

N50 – это длина контига, при которой контиги такой же или большей длины составляют 50% всех нуклеотидов в сборке



Контиги, отсортированные по  
длине

# Ключевые параметры *de novo* сборки

Параметр	Описание	Пример
N50	<p>«Чем больше, тем лучше»</p> <p>N50 это длина контига, при которой контиги такой же или большей длины составляют 50% всех нуклеотидов в сборке.</p>	<p>N50 = 98.6Kb <i>E. coli</i> 400bp PGM</p>
Число контигов <small>Контиг: набор перекрывающихся фрагментов ДНК, которые в совокупности представляют собой консенсус области ДНК</small>	<p>«Чем меньше, тем лучше»</p> <p>По мере увеличения числа контигов, которые укладываются в сборку, это число уменьшается. В идеале полностью собранный геном без пробелов будет состоять из 1 контига.</p>	<p># Contigs = 158 <i>E. coli</i> 400bp PGM</p>
% покрытия референса	<p>«Чем ближе к 100%, тем лучше»</p> <p>% покрытия референса означает % всех оснований в геноме, которые были покрыты по крайней мере одним прочтением.</p>	<p>% Ref Coverage = 98.15% <i>E. coli</i> 400bp PGM</p>



# Рабочий процесс NGS



# Основные этапы секвенирования



**Приготовление  
библиотеки**



**Подготовка  
матрицы**



**Секвенирование**



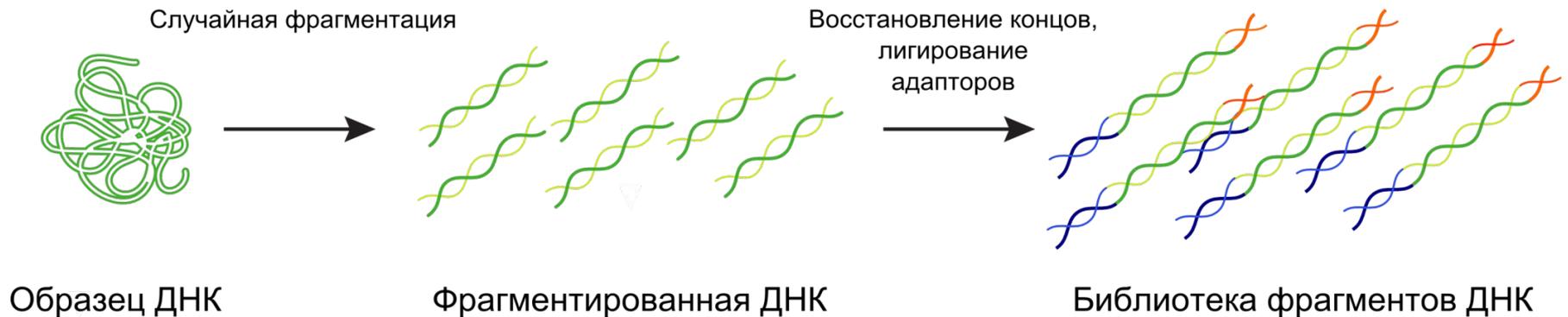
**Анализ  
данных**



# Приготовление библиотеки

## Источники ДНК для приготовления библиотек:

- геномная ДНК (Whole genome);
- часть геномной ДНК (Targeted Enrichment);
- набор ПЦР-продуктов (ампликонов);
- кДНК (RNA-Seq);
- иммунопреципитированный хроматин (ChIP-Seq) и т. д.

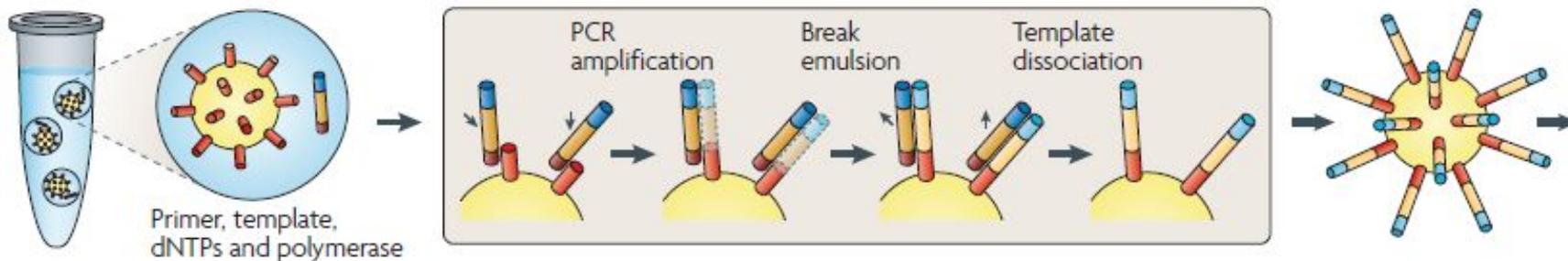




# Подготовка матрицы

Эмульсионная ПЦР и  
автоматизированная система  
обогащения на магнитных  
частицах

Система **Ion OneTouch**  
**2**

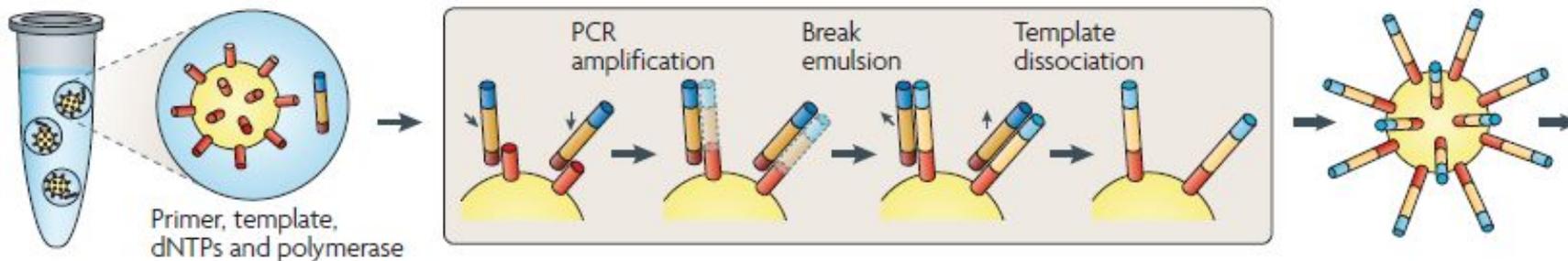




# Подготовка матрицы

- Амплификация библиотеки
- Восстановление и обогащение сфер
- Загрузка чипа

## Станция **Ion Chef**





# Секвенирование

- регистрация локального изменения pH на полупроводниковом микрочипе при последовательном удлинении олигонуклеотидной затравки ДНК-полимеразой



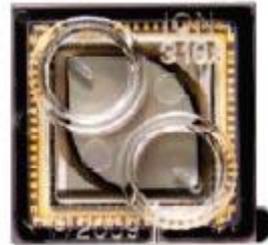
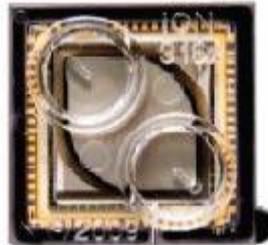
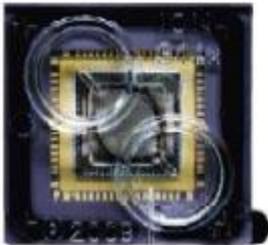
Ion PGM™ Sequencer



Ion Proton™ Sequencer



# Секвенирование



Ion  
314™

Ion  
316™

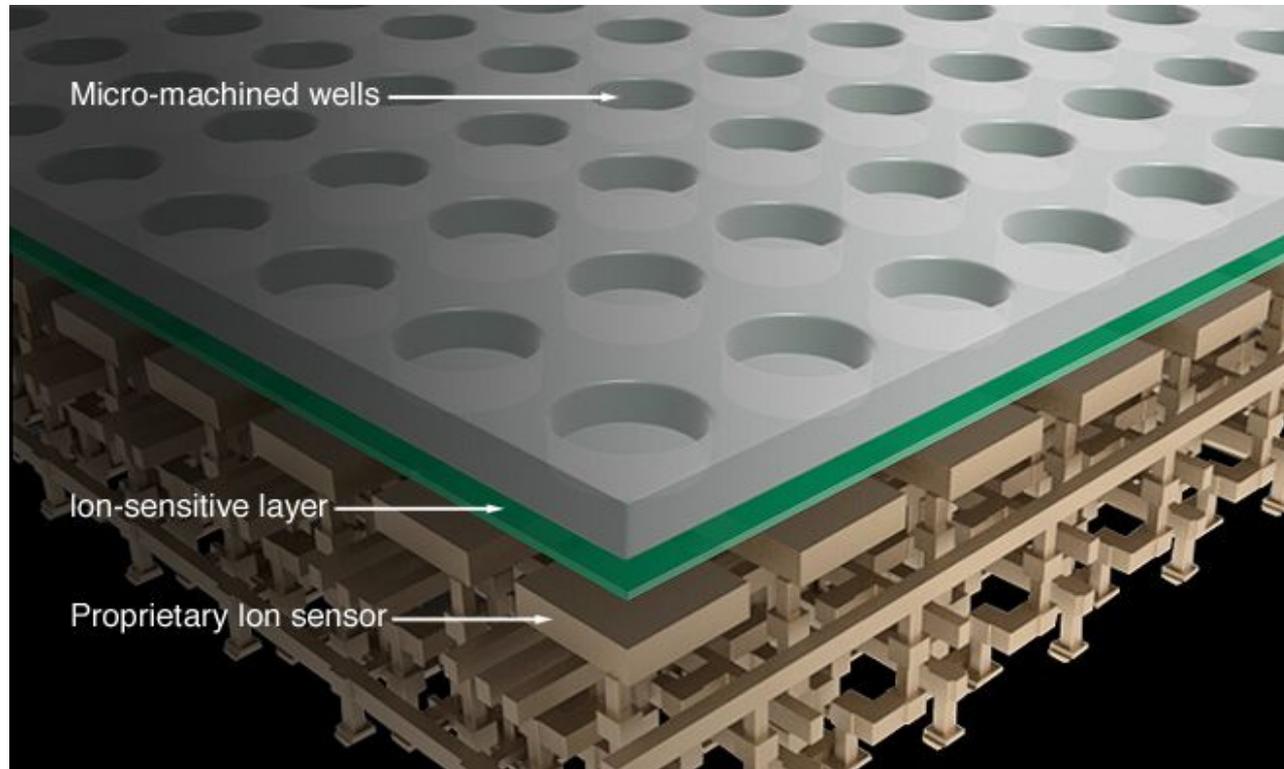
Ion  
318™



P1  
™

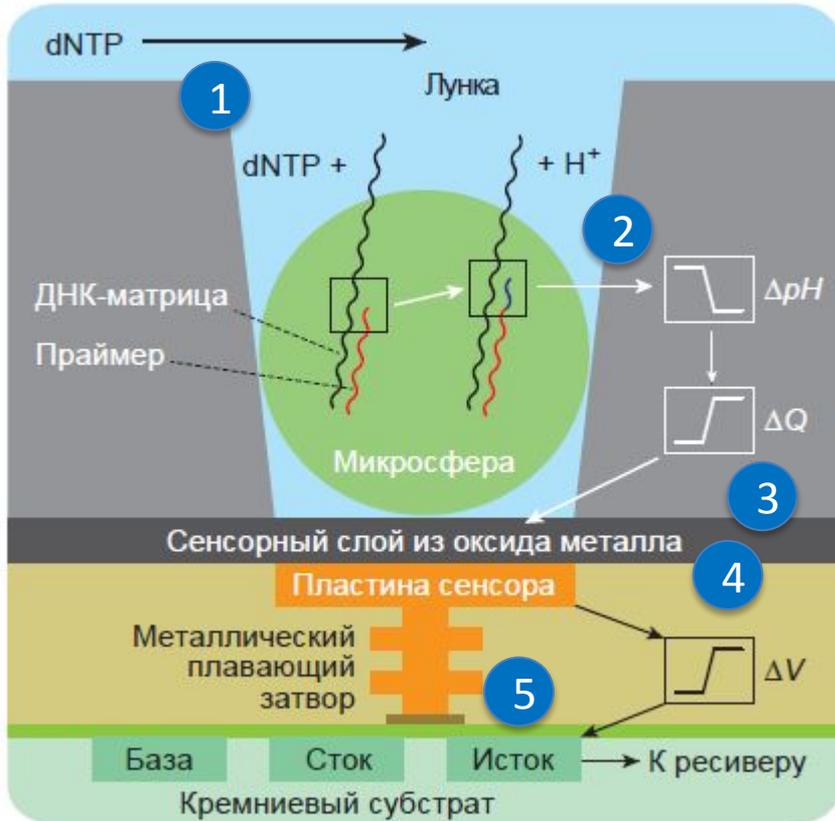


# Ion ЧИП



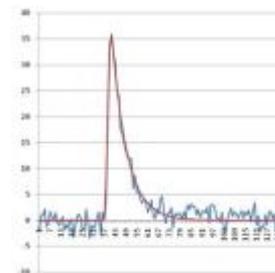


# Детекция на чипе в реальном времени



Схематическое изображение поперечного сечения одной ячейки на чипе

- 1 дНТФ последовательно подаются один за другим на чип
- 2 После встраивания нуклеотида в цепь происходит отщепление протона и изменение pH в ячейке
- 3 Чувствительный слой регистрирует изменение pH
- 4 Сенсор переводит химический сигнал в цифровой
- 5 Изменение напряжения пропорционально количеству встроенных нуклеотидов



Измерение напряжения





# Анализ данных

## Программное обеспечение

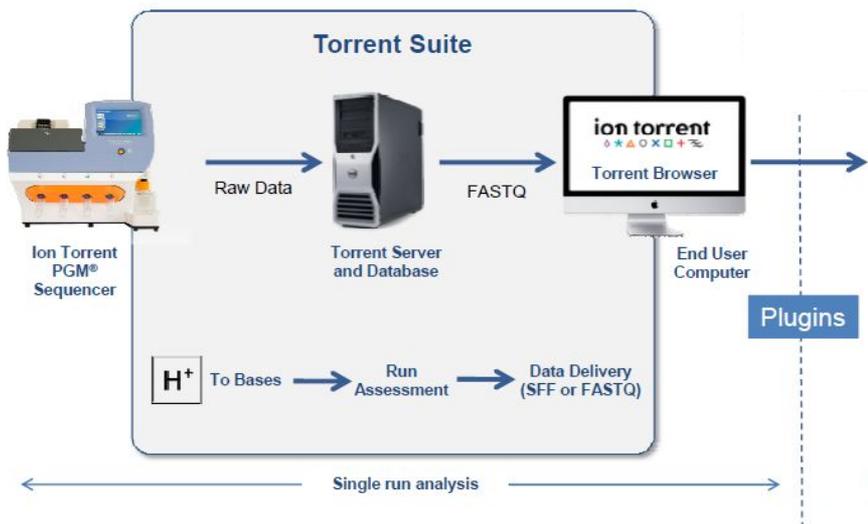
### • Torrent Suite

- Программа для запуска и настройки хода секвенирования.
- Простые опции для показа метрики анализа, доступные для понимания тех, кто не является экспертом в области биоинформатики.
- Точность при определении гомополимеров.

- Различные приложения.
- Отправка данных через «облако» по сети.

### • Ion Reporter

- Анализ данных через «облако», аннотация и составление отчёта о результате.

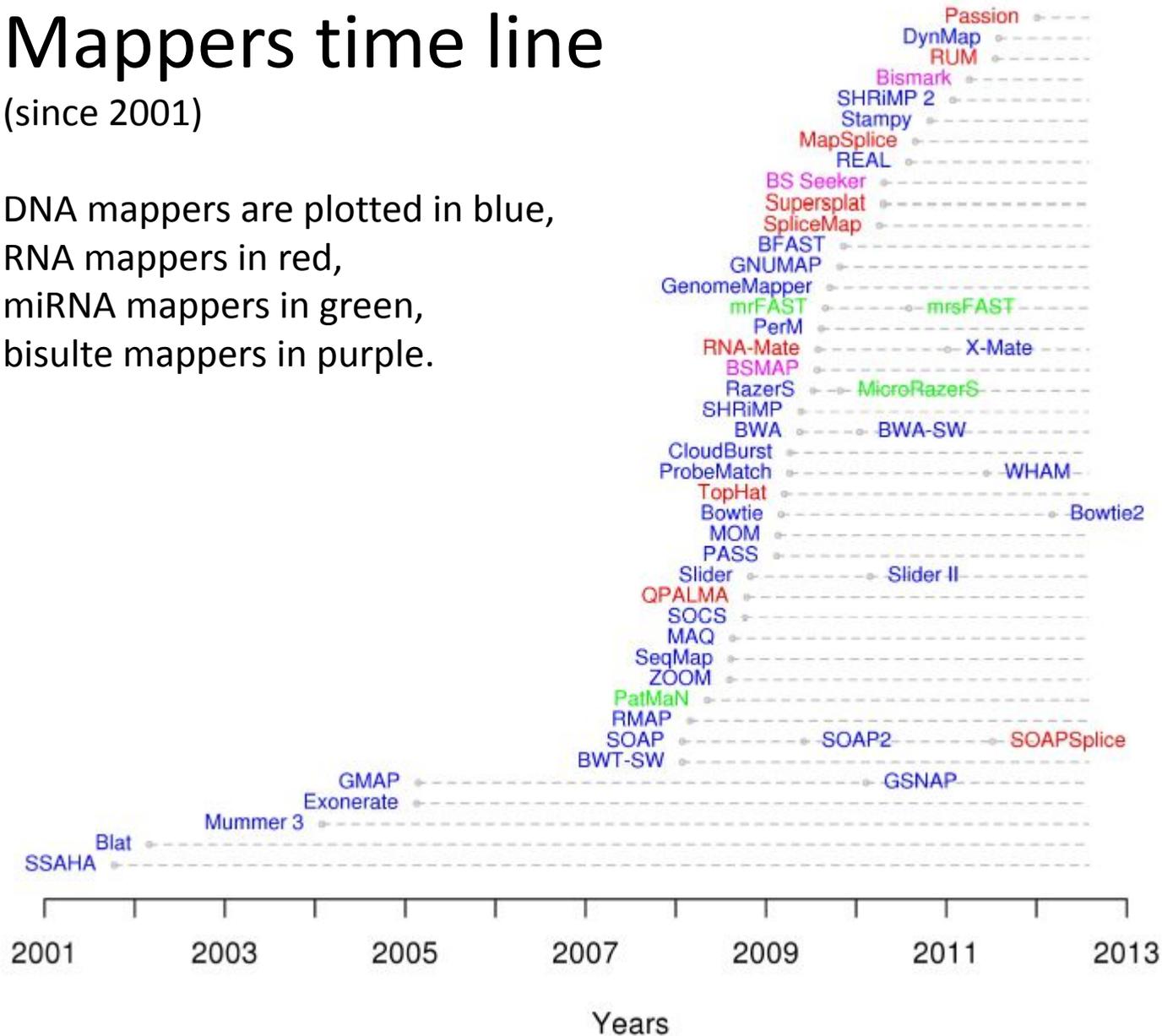




# Mappers time line

(since 2001)

DNA mappers are plotted in blue,  
RNA mappers in red,  
miRNA mappers in green,  
bisulte mappers in purple.



# Масштабное применение NGS

- 1000 Genomes Project: поиск геномных вариаций в ~20 популяциях
- PGP: personal genome project, изучение вклада наследственности и среды в проявление различных признаков; секвенирование геномов добровольцев, согласных предоставить личную информацию
- GEUVADIS Genetic European Variation in Disease: медицинская интерпретация секвенсов данных для разных моно- и полигенных заболеваний; на основе RNA-Seq и ExonSeq
- ESGI: European Sequencing and Genotyping Infrastructure: организация NGS инфраструктуры для европейского научного сообщества
- IHEC: International Human Epigenome Consortium: характеристика эпигенома человека, ~1000 образцов
- **Раковые программы**
  - выявление диагностических маркеров, анализ и выбор лечения
- ICGC: International Cancer Genome Consortium: подпрограммы по разным видам рака (например, рак простаты, детские опухоли мозга)
- Treat 1000: разные виды рака
- OncoTrack: рак прямой кишки
- CAGEKID: рак почки
- Treat 20: меланома



# Геном

## Сиквенс и ресиквенс

- микроорганизмы / малые геномы: самый эффективный подход
- большие геномы: ресиквенс с анализом небольших перестроек был хорош всегда, сиквенс de novo и анализ больших перестроек совершенствуются на глазах
- классификация организмов по молекулярным признакам
- сиквенс определенных областей генома: обогащение: функциональный экзом (~5%) и произвольно выбранных областей

## Геномная архитектура

- позиция нуклеосом
- трёхмерная структура генома: 3C (hypothesis-free) и 5C (высокопроизводительный анализ)

## Эпигенетика

- метилирование: сиквенс после бисульфидной обработки или ChIP-Seq
- возможно, машины третьего поколения смогут определять модификации напрямую

## Взаимодействие нуклеиновых кислот с белками

- ChIP-Seq
- полногеномный неселективный анализ мест посадки белков
- гиперчувствительность к нуклеазам, FAIRE-Seq
- анализ специфичности взаимодействия (напрямую), поиск консенсусных последовательностей

# Транскриптом

## Полнота анализа

- все типы РНК: кодирующие последовательности, малые некодирующие РНК, антисенс-транскрипция и т.д.
- чувствительность: при адекватной организации эксперимента, можно быть уверенным, что замечены все действующие РНК

## RNA-Seq позволяет

- определять как относительный, так и абсолютный уровни экспрессии
- аннотировать новые и уточнять аннотацию известных генов
- смотреть отдельно ядерную, цитоплазматическую, ассоциированную с рибосомами (в цитоплазме и на мембранах) и т.п. РНК
- выявлять присутствие в образце микроорганизмов и вирусов

## Чувствительность

# Сравнительный функциональный анализ

## Сравнение геномов разных видов позволяет

- выявить функциональные элементы генома
- связать эволюционные нововведения с появлением в геноме новых функциональных участков, типа проект «Origins of Multicellularity»
- выявить механизмы и движущие факторы эволюции для разных эпох
- разобраться, что отличает нас от ближайших и дальних родичей

## Внутривидовое сравнение геномов

- причина тонких различий в реакции на факторы окружающей среды (болезни) и поведении
- факторы, определяющие сложные полигенные признаки (рост, вес, предрасположенность к распространенным болезням)
- механизмы рекомбинации
- факторы отбора в «историческое» время
- история расселения человечества

# Медицина

## Микробиология

- выявление патогенов: холера на Гавайях, токсичная E.coli в Германии
- анализ симбиотических организмов: микробиом

## Предрасположенность к болезни → генотип

- GWAS, вовлеченные гены, молекулярные причины болезни, выбор мишени
- редкие болезни, семейный анализ: «быстрый» тест сиквенсом кодирующих последовательностей
- иммунный репертуар В- и Т-клеточных рецепторов: подверженность болезням, ответ на вакцинацию
- неинвазивная и пренатальная диагностика по ДНК в крови

## Болезнь → профиль экспрессии

- классификация опухолей, тонкий и/или ранний диагноз
- спектр мутагенеза – причина опухолеобразования

# Медицина

## Проекты по секвенсу разных типов опухолей

- спектр мутагенеза, «завершить» список онкогенов
- тестирование лекарств на клеточных линиях
- тонкая диагностика и выработка «стандартных» способов лечения
- неинвазивная диагностика (посттreatment): опухолевые ДНК и клетки в крови

# Медицина

## Превентивная медицина

- Болезнь Альцгеймера
- Диабет II типа

## Персональная медицина

- Генетически опосредованная чувствительность к лекарствам (varfarin)