



НАУЧНО-ИССЛЕДОВАТЕЛЬСКИЙ ИНСТИТУТ
СИСТЕМНОЙ БИОЛОГИИ И МЕДИЦИНЫ
РОСПОТРЕБНАДЗОРА

СОВРЕМЕННЫЕ ТЕХНОЛОГИИ СЕКВЕНИРОВАНИЯ

ПРИНЦИПЫ ПОДГОТОВКИ БИБЛИОТЕК ДЛЯ ВЫСОКОПРОИЗВОДИТЕЛЬНОГО СЕКВЕНИРОВАНИЯ

*Анна Сергеевна СПЕРАНСКАЯ
к.б.н., зав. лаб. мультиомиксных
исследований
НИИ СБМ Роспотребнадзора*

illumina



NextSeq



MiSeq



华大基因
BGI



Thermo
Fisher
SCIENTIFIC

СБМ
НАУЧНО-ИССЛЕДОВАТЕЛЬСКИЙ ИНСТИТУТ
СИСТЕМНОЙ БИОЛОГИИ И МЕДИЦИНЫ
РОСПОТРЕБНАДЗОРА



Ion S5 + Ion Chef



Oxford
NANOPORE
Technologies



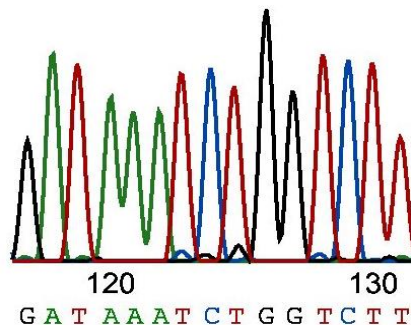
QIAGEN

СЕКВЕНИРОВАНИЕ ПЕРВОГО, ВТОРОГО И ТРЕТЬЕГО ПОКОЛЕНИЯ

ПЕРВОЕ ПОКОЛЕНИЕ секвенирование по Сэнгеру	ВТОРОЕ ПОКОЛЕНИЕ Иллюмина, MGI, полупроводниковое секвенирование	ТРЕТЬЕ ПОКОЛЕНИЕ ONT, PacBio
Для одного образца анализируется один фрагмент генома во время сиквенса	Одновременно анализируется много фрагментов генома (но можно и один)	Одновременно анализируется много фрагментов генома (но можно и один)
Короткая длина анализируемой молекулы (до 1000 п.о.)	Короткая длина анализируемых фрагментов (до 600 п.о.)	Большая длина анализируемых фрагментов (больше 10 тыс п.о.), но можно анализировать и короткие
Чтение образца только в одну сторону	Чтение образца в одну или в две стороны	
Одна пробирка – один образец	В одной пробирке может быть много образцов (но может быть и один)	В одной пробирке может быть много образцов (но может быть и один)

СЕКВЕНИРОВАНИЕ ПЕРВОГО, ВТОРОГО И ТРЕТЬЕГО ПОКОЛЕНИЯ: ТЕХНОЛОГИИ

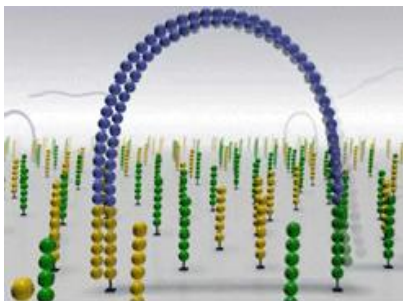
I



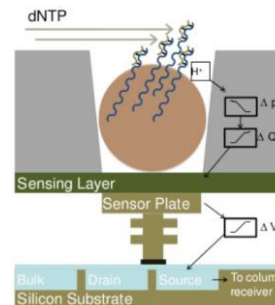
Секвенирование методом Сэнгера

Единственный секвенатор российского производства, Нанофор

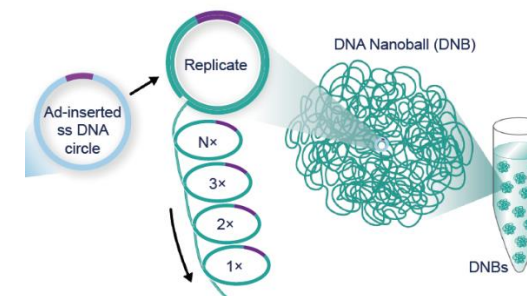
II



Иллюмина (Illumina, США)

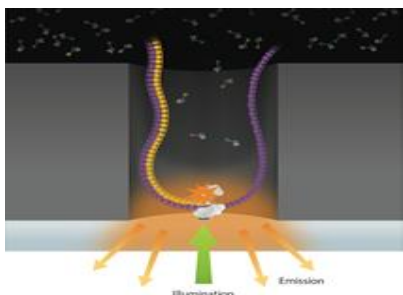


Полупроводниковые (ThermoFischer, США)



DNBSeq (MGI, Китай)

III

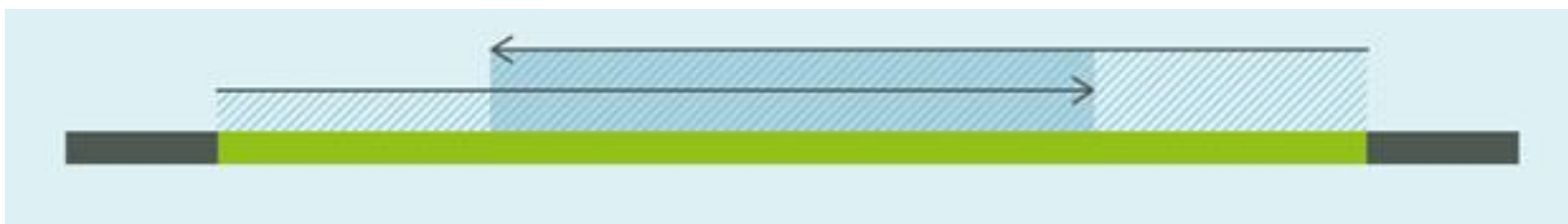
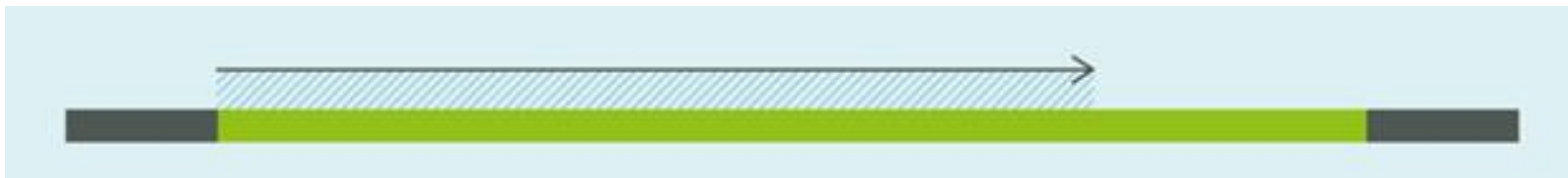


Пакбио или SMRT (PacBio, США)



Нанопоровые (ONT, Великобритания)

ОДНО- И ПАРНОКОНЦЕВОЕ СЕКВЕНИРОВАНИЕ

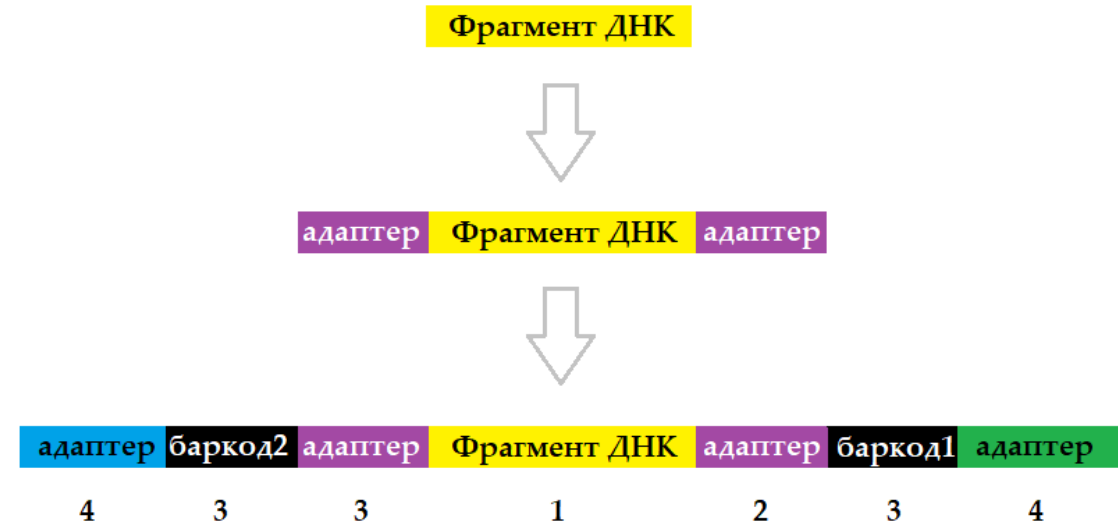


ОБЩИЕ КЛЮЧЕВЫЕ ОСОБЕННОСТИ СЕКВЕНИРОВАНИЯ НА ПЛАТФОРМАХ II И III ПОКОЛЕНИЙ

- ❑ Длина сиквенса (прочтения), как правило, меньше длины генома анализируемого образца, поэтому сначала надо получить много фрагментов ДНК анализируемого образца с подходящей для секвенатора длиной.
- ❑ Поскольку одновременно анализируется много образцов, значит, нужно их как-то различать на молекулярном уровне (поэтому образцы помечаются молекулярными баркодами).
- ❑ Если для одного образца одновременно секвенируется очень много фрагментов ДНК, значит, в результате для каждого образца получается очень много прочтений, из которых в конечном итоге надо собирать исходную последовательность анализируемого образца (следовательно, нужны методы сборки и анализа).



- a. Изготовление препарата, содержащего фрагменты ДНК, которые будут секвенироваться
- b. Пришивка адаптеров, которые затем позволят провести баркодирование
- c. Баркодирование и пришивка адаптеров, специфичных для конкретной секвенирующей платформы/метода

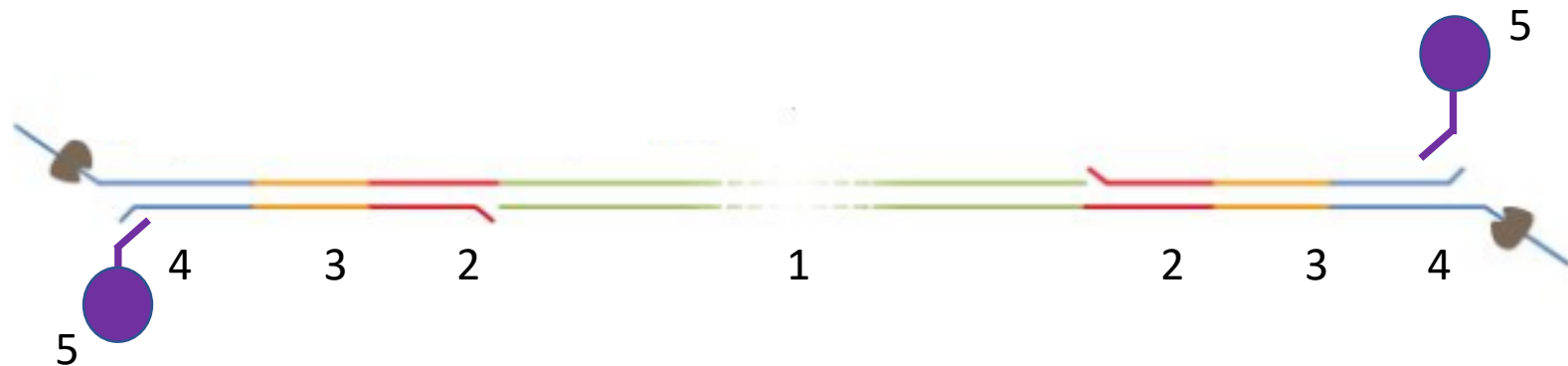


1. Секвенируемый фрагмент
2. Адаптер (может включать последовательность, комплементарную праймеру, с которого начинается секвенирование)
3. Индекс (или баркод), короткая последовательность из 8-10 нуклеотидов
4. Специфичная для секвенирующей платформы последовательность нуклеотидов

УСТРОЙСТВО ФРАГМЕНТА ДНК ДЛЯ СЕКВЕНИРОВАНИЯ НА OFROD NANOPORE, 1D

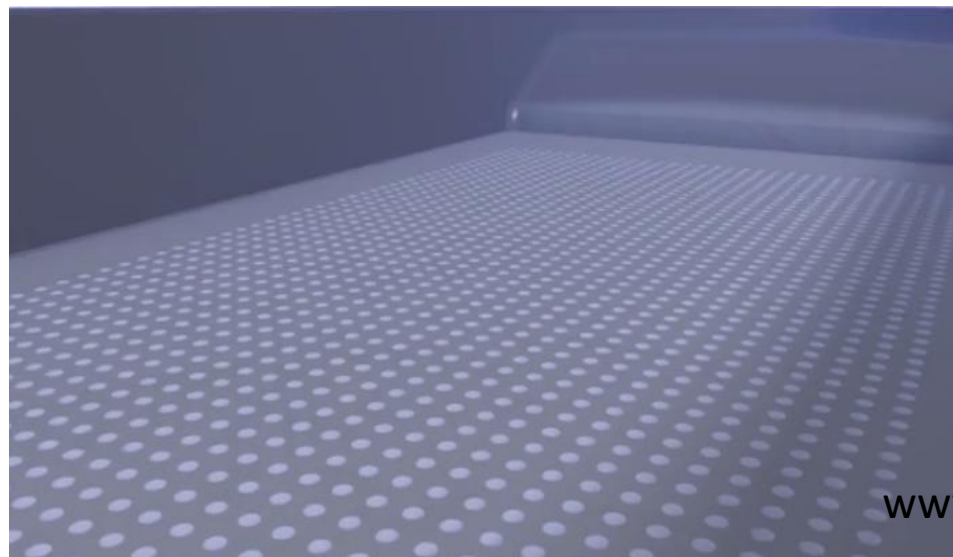
Содержит:

1. Секвенируемый фрагмент
2. Адаптер
3. Баркод
4. Адаптер, содержащий моторный белок

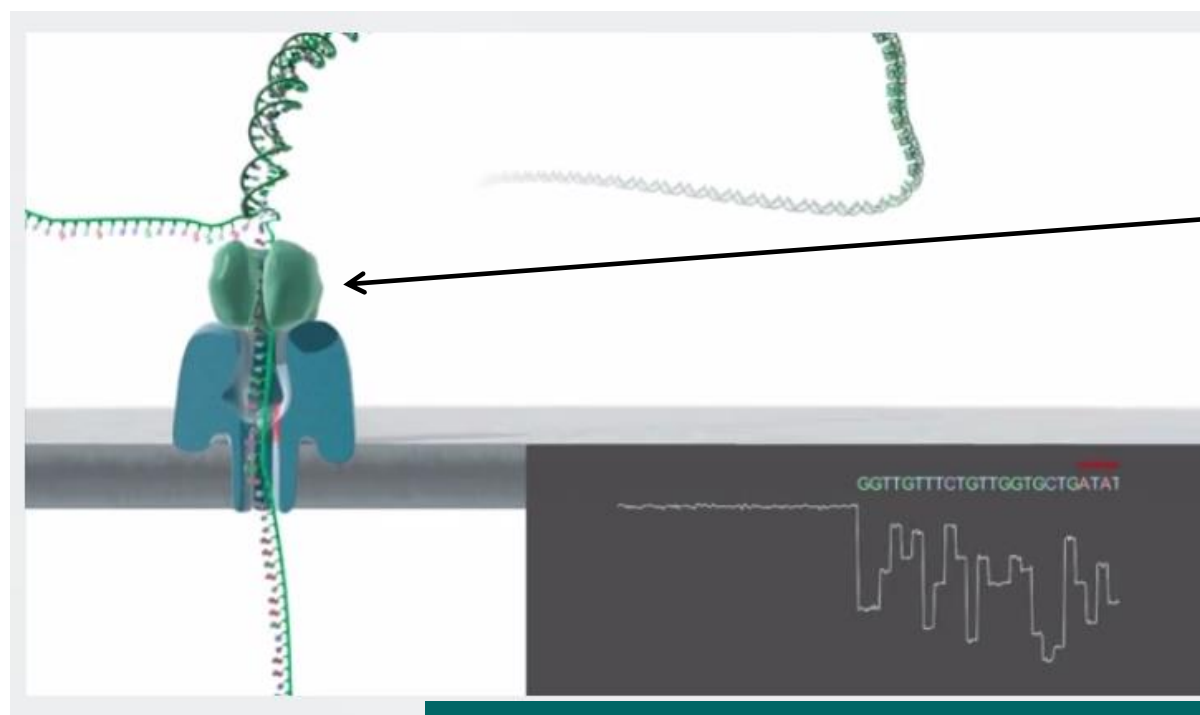
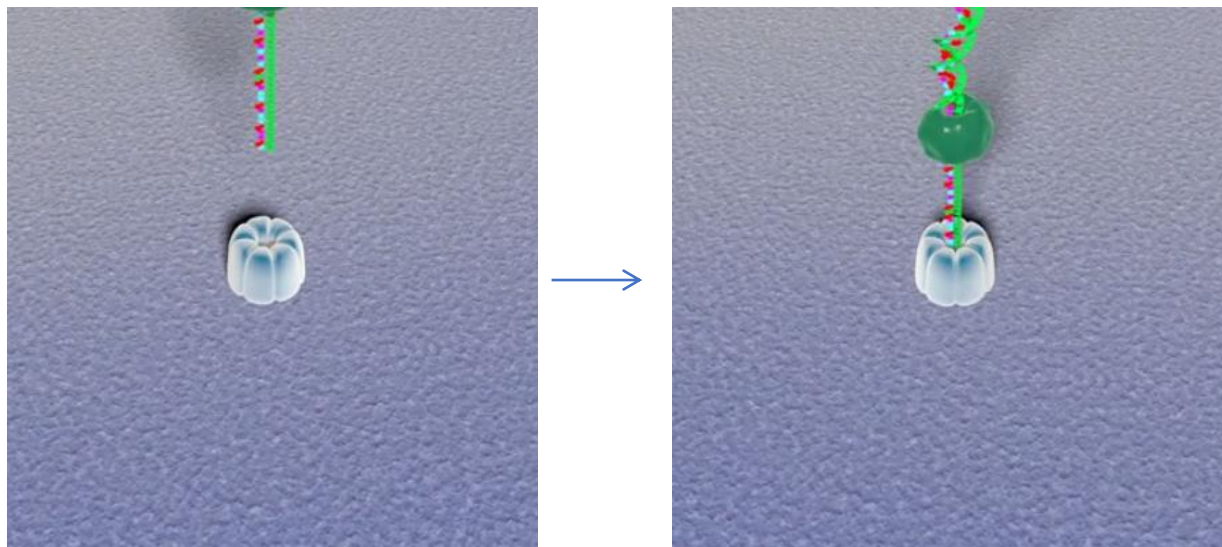


5. Непосредственно перед секвенированием добавляется тетер, который связывает фрагмент библиотеки с мембраной

ПРОТОЧНЫЕ ЯЧЕЙКИ ДЛЯ MinION ИЛИ GridION (OXFORD NANOPORE TECHNOLOGIES)



www.nanoporetech.com



Моторный белок

www.nanoporetech.com

Конечная цель пробоподготовки – получить набор фрагментов ДНК необходимой длины, содержащих на концах подходящие для установленного в лаборатории секвенатора «служебные» последовательности (адаптеры).

ДЛЯ ОБОГАЩЕНИЯ ОБРАЗЦА НУКЛЕИНОВЫМИ КИСЛОТАМИ SARS-CoV-2 ИСПОЛЬЗУЮТ РАЗНЫЕ СТРАТЕГИИ

Амплификация генома путем наработки большого количества коротких ампликонов, 300-400 п.о.

Панель **ARTIC**, позволяющая амплифицировать полный геном SARS-CoV-2, используя 98 пар праймеров, Oxford Nanopore

<https://artic.network/ncov-2019>



Амплификация генома путем наработки маленького количества ампликонов длиной 1800-2100 п.о.

Протокол **Midnight**, размер амплифицируемых фрагментов 12000 п.о., Oxford Nanopore

Панель **SCV-2000bp**, размер амплифицируемых фрагментов 1800-2000 п.о.

<https://www.protocols.io/view/protocol-for-scv-2000bp-a-primer-panel-for-sars-co-bn77mhrn.html>

ПОДГОТОВКА ОБРАЗЦОВ К СЕКВЕНИРОВАНИЮ (при использовании протокола MIDNIGHT АМПЛИФИЦИРУЮТСЯ ФРАГМЕНТЫ 1200 П.О., В СУММЕ ПЕРЕКРЫВАЮЩИЕ ВЕСЬ ГЕНОМ)

- Экстракция РНК

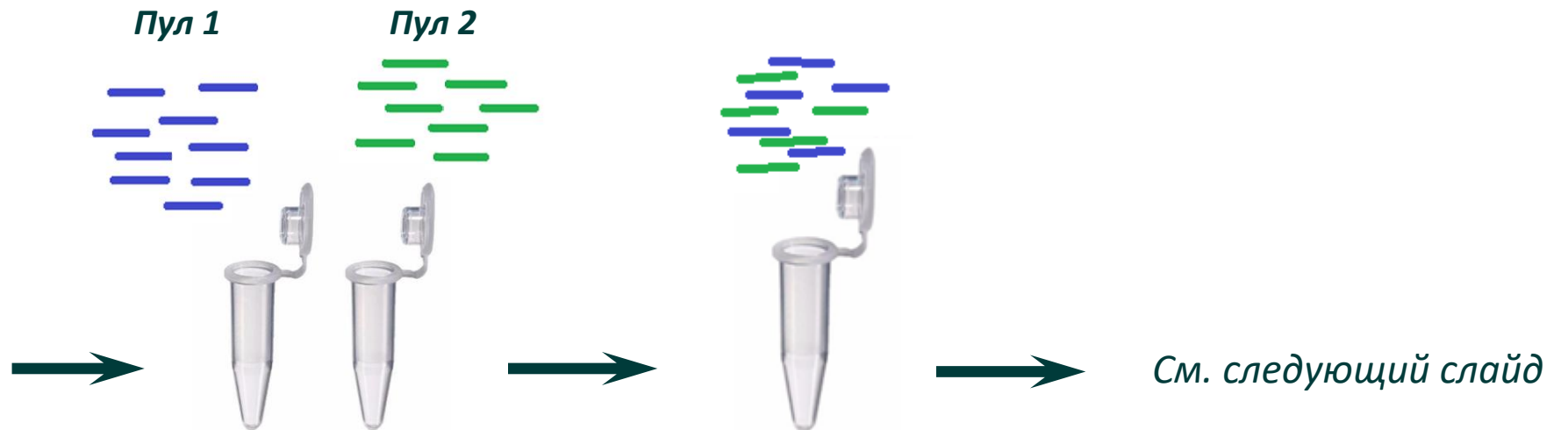
- Реверсия

- Обогащение образца (амплификация генома SARS-CoV-2)

- Подготовка библиотек

- Реагенты и ячейки для секвенирования

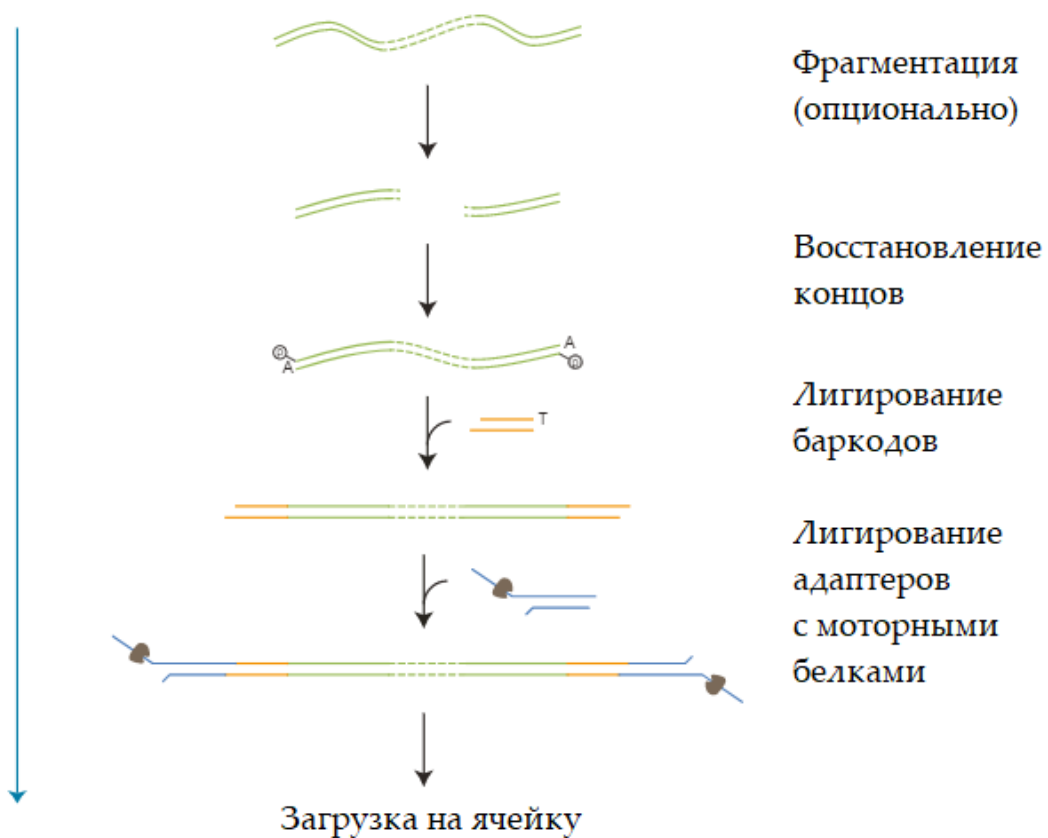
← Реагенты, входящие в состав набора
Midnight



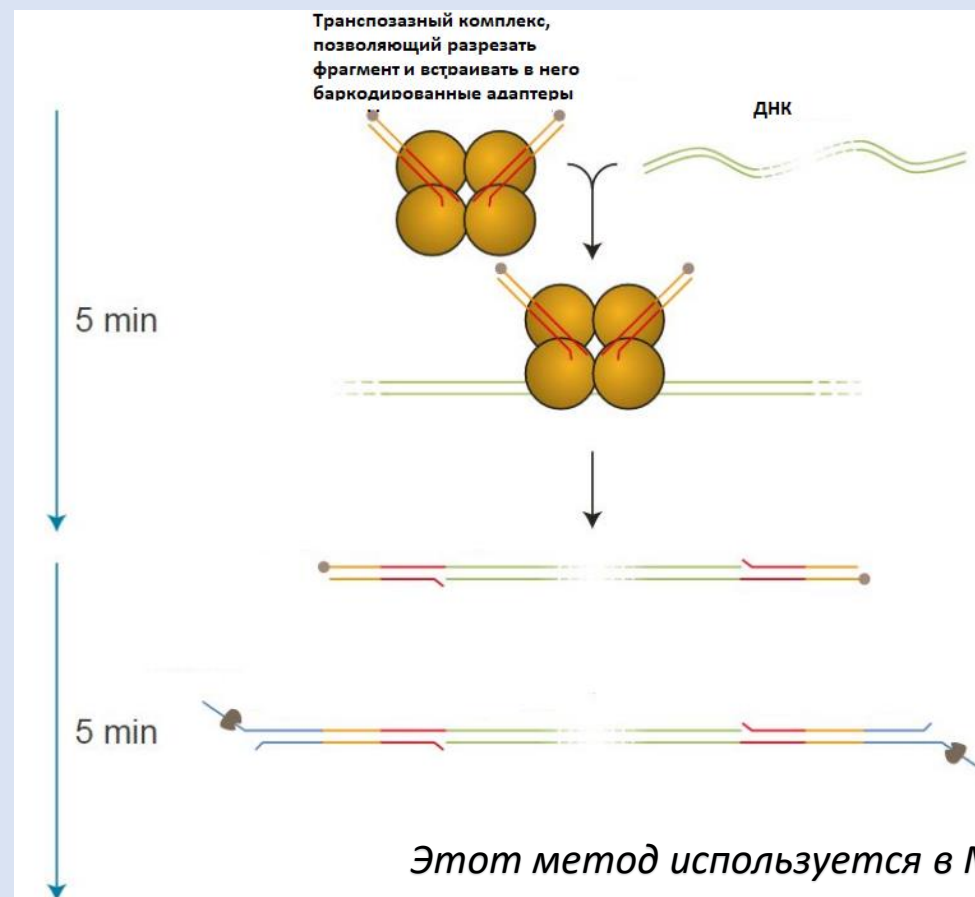
МЕТОДЫ ПРИШИВКИ АДАПТЕРОВ

Лигирование

Высокомолекулярная геномная ДНК или Ампликоны



Встраивание с помощью транспозазы

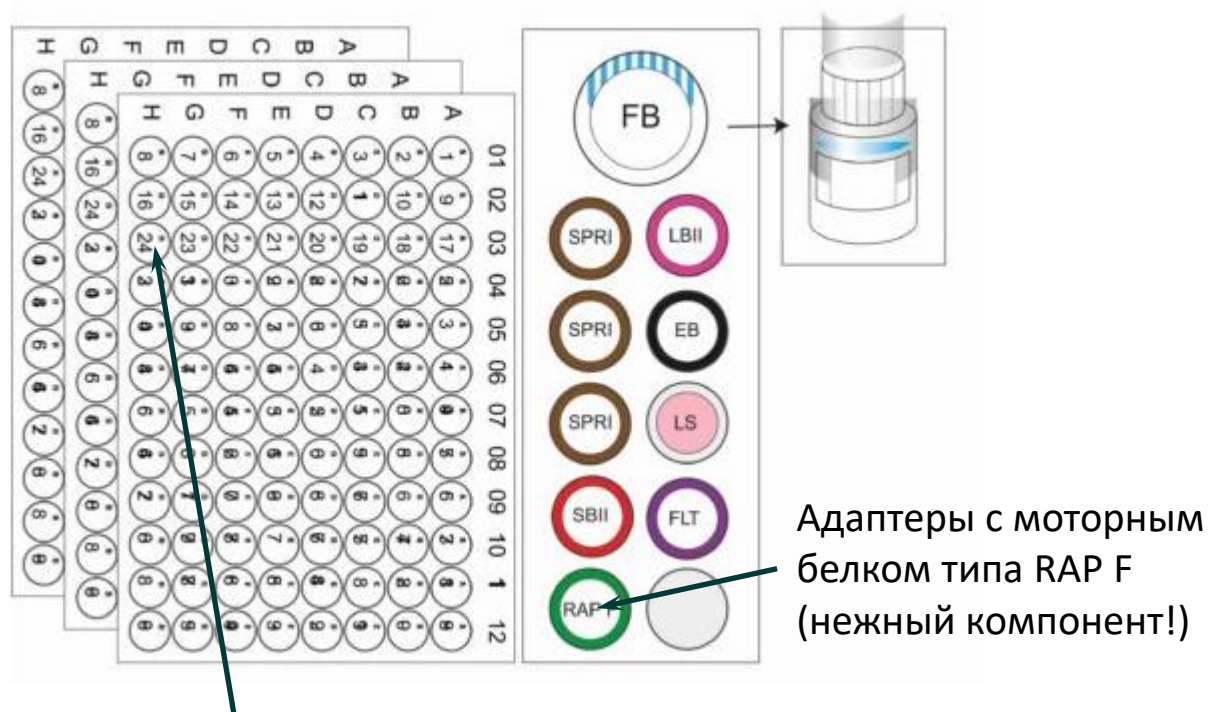


ONT ПРОИЗВОДИТ РАЗНЫЕ БАРКОДЫ ДЛЯ БАРКОДИРОВАНИЯ БИБЛИОТЕК И РАЗНЫЕ АДАПТЕРЫ, СОДЕРЖАЩИЕ МОТОРНЫЕ БЕЛКИ

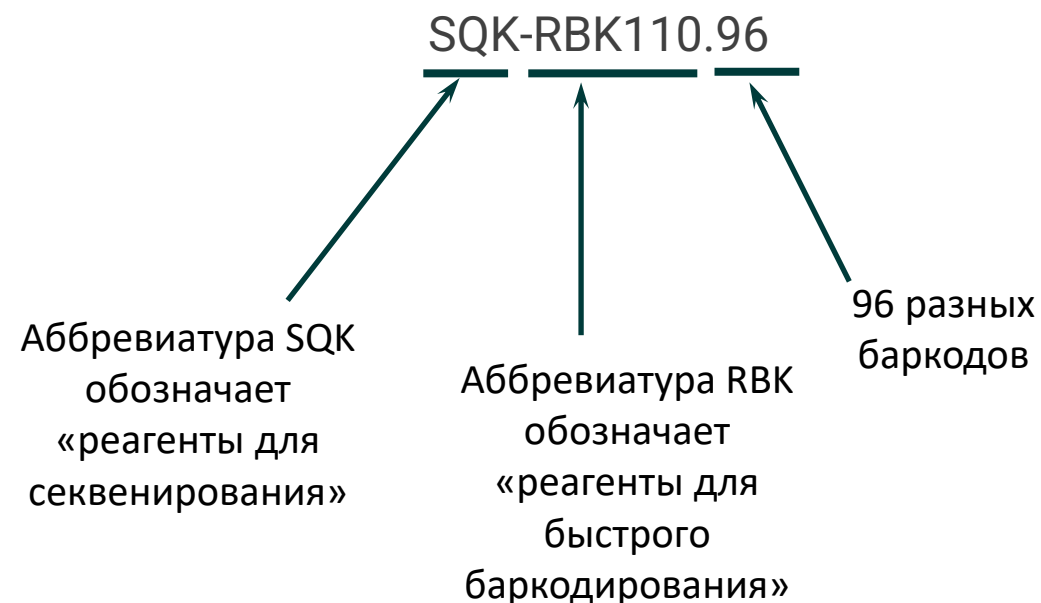
В состав набора Midnigth входят:

- реагенты для быстрого баркодирования RBK110 (от Rapid Barcoding Kit);
- адаптеры с моторными белками (типа RAP F);
- буфер и дополнительные реагенты для активации ячейки и секвенирования, а также для загрузки библиотеки на ячейку.

Перечисленные реагенты также можно купить в виде отдельного набора с каталожным номером



Баркоды (Внимание! Риск
контаминации при повторном
использовании)





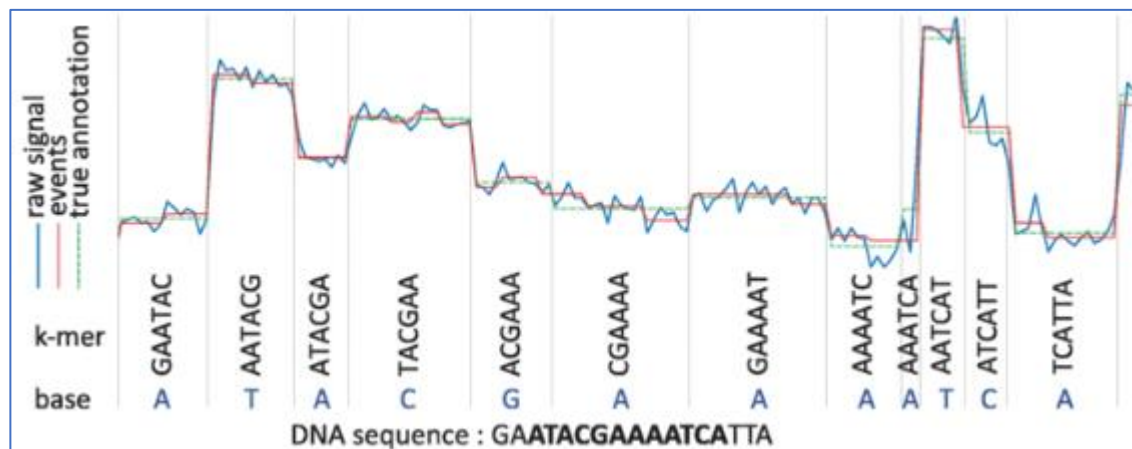
Oxford Nanopore
выпускает наборы
с 96 баркодами

**96 образцов → одна проточная
ячейка**



ПРИБОРЫ ВСЕХ ТИПОВ ИЗНАЧАЛЬНО ЗАПИСЫВАЮТ «СЫРУЮ» ИНФОРМАЦИЮ О СИГНАЛАХ, ПОСТУПАЮЩИХ С ПРИБОРА В ПРОЦЕССЕ СИКВЕНСА, И ВЫДАЮТ ЕЕ ПОЛЬЗОВАТЕЛЮ РАЗЛИЧНЫХ «ПЕРВИЧНЫХ» ФОРМАТАХ

- ☐ Сэнгер - *.abi
- ☐ Иллюмина - *.bcl
- ☐ Полупроводниковые секвенаторы, пиросеквенаторы - *.sff
- ☐ **Оксфорд Нанопор - *.fast5**



«Сырой» сигнал нанопорового секвенатора.

Рисунок из статьи «GPU accelerated adaptive banded event alignment for rapid comparative nanopore signal analysis»
August 2020, BMC Bioinformatics 21(1) DOI:10.1186/s12859-020-03697-x

*.fast5

*.fastq

Бейсколлинг – термин, означающий распознавание нуклеотидов по исходному сигналу.

Процесс бейсколлинга осуществляет специальное программное обеспечение (поставляемое вместе с прибором).

Phred Quality Score (параметр Q)	Вероятность ошибочного распознавания нуклеотида	Корректность бейсколлинга
10	1 на 10	90%
20	1 на 100	99%
30	1 на 1000	99,9%
40	1 на 10000	99,99%

ТОЧНОСТЬ СЕКВЕНАТОРОВ:

- ← ONT (нанопоровые)
- ← Ion Torrent
- ← Illumina и DNBSeg (MGI)
- ← PacBio

ДЛЯ ЗАПИСИ КАЧЕСТВА В ФАЙЛ С ПОСЛЕДОВАТЕЛЬНОСТЬЮ НУКЛЕОТИДОВ ИСПОЛЬЗУЮТ ASCII КОД

ASCII_BASE=33 Illumina, Ion Torrent, PacBio and Sanger

Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII
0	1.00000	33 !	11	0.07943	44 ,	22	0.00631	55 7	33	0.00050	66 B
1	0.79433	34 "	12	0.06310	45 -	23	0.00501	56 8	34	0.00040	67 C
2	0.63096	35 #	13	0.05012	46 .	24	0.00398	57 9	35	0.00032	68 D
3	0.50119	36 \$	14	0.03981	47 /	25	0.00316	58 :	36	0.00025	69 E
4	0.39811	37 %	15	0.03162	48 0	26	0.00251	59 ;	37	0.00020	70 F
5	0.31623	38 &	16	0.02512	49 1	27	0.00200	60 <	38	0.00016	71 G
6	0.25119	39 '	17	0.01995	50 2	28	0.00158	61 =	39	0.00013	72 H
7	0.19953	40 (	18	0.01585	51 3	29	0.00126	62 >	40	0.00010	73 I
8	0.15849	41)	19	0.01259	52 4	30	0.00100	63 ?	41	0.00008	74 J
9	0.12589	42 *	20	0.01000	53 5	31	0.00079	64 @	42	0.00006	75 K
10	0.10000	43 +	21	0.00794	54 6	32	0.00063	65 A			

ASCII_BASE=64 Old Illumina

Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII
0	1.00000	64 @	11	0.07943	75 K	22	0.00631	86 V	33	0.00050	97 a
1	0.79433	65 A	12	0.06310	76 L	23	0.00501	87 W	34	0.00040	98 b
2	0.63096	66 B	13	0.05012	77 M	24	0.00398	88 X	35	0.00032	99 c
3	0.50119	67 C	14	0.03981	78 N	25	0.00316	89 Y	36	0.00025	100 d
4	0.39811	68 D	15	0.03162	79 O	26	0.00251	90 Z	37	0.00020	101 e
5	0.31623	69 E	16	0.02512	80 P	27	0.00200	91 [	38	0.00016	102 f
6	0.25119	70 F	17	0.01995	81 Q	28	0.00158	92 \	39	0.00013	103 g
7	0.19953	71 G	18	0.01585	82 R	29	0.00126	93]	40	0.00010	104 h
8	0.15849	72 H	19	0.01259	83 S	30	0.00100	94 ^	41	0.00008	105 i
9	0.12589	73 I	20	0.01000	84 T	31	0.00079	95 _	42	0.00006	106 j
10	0.10000	74 J	21	0.00794	85 U	32	0.00063	96 `			

ФОРМАТЫ ФАЙЛОВ *.fasta и *.fastq

Формат файла *.fasta (он же *.fas, он же *.fa)

```
>HWI-M01460:53:000000000-A7F3Y:1:1101:19488:1313 1:N:0:TGACCA
GATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAATATCGTATGCCGTCTTCTGCTTGAAAAAAAAAATACCTTAATGATGTAAACAATTCAGAGATGAGGAAAGTCAA
```

Формат файла *.fastq

```
@HWI-M01460:53:000000000-A7F3Y:1:1101:19488:1313 1:N:0:TGACCA
GATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAATATCGTATGCCGTCTTCTGCTTGAAAAAAAAAATACCTTAATGATGTAAACAATTCAGAGATGAGGAAAGTCAA
+
9CC,CEGDGGGGF8FEEGC@FDFF,CCC@FC,CA8EFG,,6C;;6,C,@C8@,,CEEE9,C,FECD:+@#####
```


ИСХОДНЫЙ ГЕНОМ (В ПРОБИРКЕ МНОГО КОПИЙ)

Мне нравится, что Вы больны не мной. Мне нравится, что я больна не Вами.

Мне нравится, что Вы больны не мной. Мне нравится, что я больна не Вами.

Мне нравится, что Вы больны не мной. Мне нравится, что я больна не Вами.

ФРАГМЕНТЫ В ПРОБИРКЕ ПОСЛЕ ФРАГМЕНТАЦИИ

Вы больны не мной,

нравится, что Вы

Вы больны не мной,

я больна не Вами

я больна не Вами

Мне

мне нравится

У больнУ не УУУУ

нравится, что Вы

мне нравится, что

Мне нравится, что

что я больна

Мне нравится, что

не Вами

больны не мной,

Мне нравится, что

Мне нравится, что

нравится, что Вы

Вы больны не мной,

Вы больны не мной,

я больна не Вами

я больна не Вами

Мне

мне нравится

У больнУ не УУУУ

мне нравится, что

нравится, что Вы

что я больна

не Вами

больны не мной,

Мне

Мне нравится, что

Мне нравится, что

нравится, что Вы

нравится, что Вы

Вы **больны** не мной,

больны не мной,

У **больн**У не **УУУУ**

мне нравится

мне нравится, что

что я больна

я **больна** не Вами

я **больна** не Вами

не Вами

The diagram shows a central red question mark (?) with three arrows pointing to different sentence structures:

- Top-left: "Мне нравится, что Вы больны не мной, мне нравится" (I like that you are sick not of me, I like)
- Top-right: "я больна не Вами" (I am sick not of you)
- Bottom-left: "Мне нравится, что Вы больны не мной, мне нравится" (I like that you are sick not of me, I like)

Below the central question mark, the text "У больнУ не УУУУ" is written, with "У" in red and "УУУУ" in green.

- ❑ **Килобаза** или **т.п.о.** (kb или kbp) — тысяча пар оснований;
- ❑ **Мегабаза** или **млн п.о.** (Mb или Mbp) — миллион пар оснований;
- ❑ **Гигабаза** или **млрд п. о.** (Gb или Gbp) — миллиард пар оснований.

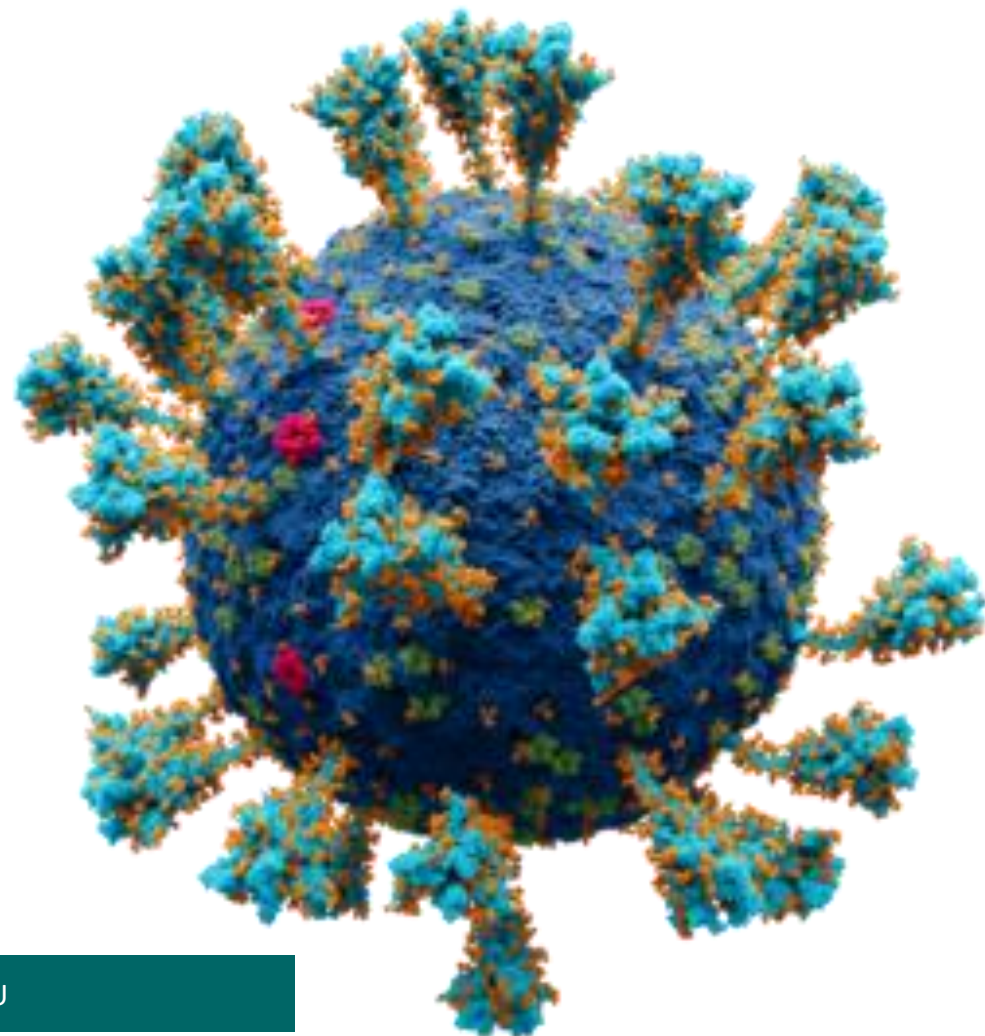
СКОЛЬКО НУЖНО ИНФОРМАЦИИ, ЧТОБЫ УСПЕШНО СОБРАТЬ ОДИН ГЕНОМ SARS-CoV-2

Размер генома около 30 000 п.о. (30 kВ).

Если брать в качестве требования среднее покрытие 1000.

Тогда необходимо получить $30\,000 \times 1000 = 30\,000\,000$ (30 Mb).

Для 96 образцов это $30\text{ Mb} \times 96 =$ около 3000 Mb (3-6 Gb).



ОБЩАЯ СХЕМА ПОЛУЧЕНИЯ ИНФОРМАЦИИ О ГЕНОМЕ С ПОМОЩЬЮ NGS

(синим выделен вариант, который используется при секвенировании SARS-CoV-2 методом Midnighth)

ПОДГОТОВКА БИБЛИОТЕК

1. Выделение нуклеиновых кислот, ДНК или РНК (с последующей реверсией)
2. Фрагментация (если далее материал будет обогащаться путем амплификации, тогда этот этап пропускается)
3. Обогащение образца нуклеиновыми кислотами нужного организма: амплификация генома нужного организма или зонды
4. Пришивка адаптеров (используются разные методы: лигирование/включение адаптеров с помощью ПЦР/транспозаза)
5. Баркодирование библиотек, подготовленных из разных образцов, разными молекулярными баркодами
6. Смешивание нескольких библиотек в один пул

СЕКВЕНИРОВАНИЕ

1. Определение концентрации и качества полученной библиотеки.
2. Загрузка пула библиотек в прибор
3. Получение первичной информации и запись этой информации в виде текстового файла (выполняет ПО прибора)
4. Бейсколинг: интерпретация первичной информации, «превращение» в последовательность нуклеотидов с присвоением каждому нуклеотиду значения качества
5. Демультиплексирование

БИОИНФОРМАТИЧЕСКИЙ АНАЛИЗ

(для сборки геномов ряда организмов существуют готовые решения, в том числе, для SARS-CoV-2)

1. Сборка генома
2. Определение % покрытия генома
3. Определение глубины покрытия генома