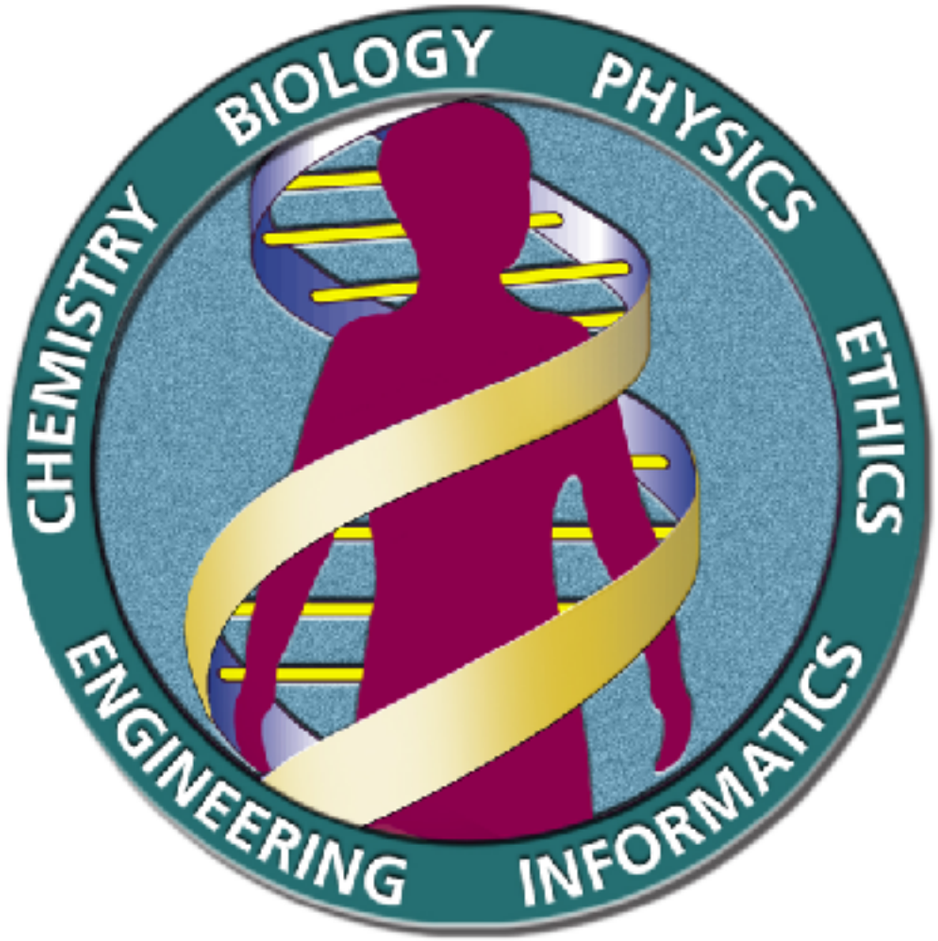
Биоинформатика

Секвенирование нового поколения (***NGS***)

Герасимов Евгений Сергеевич

Истина

ВКонтакте

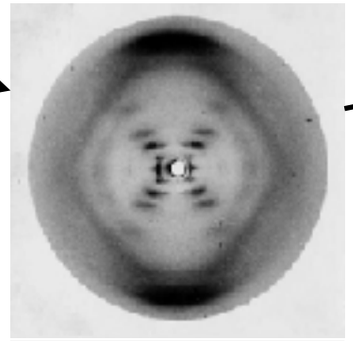


The **Human Genome Project** (HGP) was one of the great feats of exploration in history - **an inward voyage of discovery** rather than an outward exploration of the planet or the cosmos; an international research effort to sequence and map all of the genes - together known as the genome - of members of our species, *Homo sapiens*. Completed in April 2003, the HGP gave us the ability, for the first time, to read nature's complete **genetic blueprint for** building **a human being**.

Проекты-конкуренты

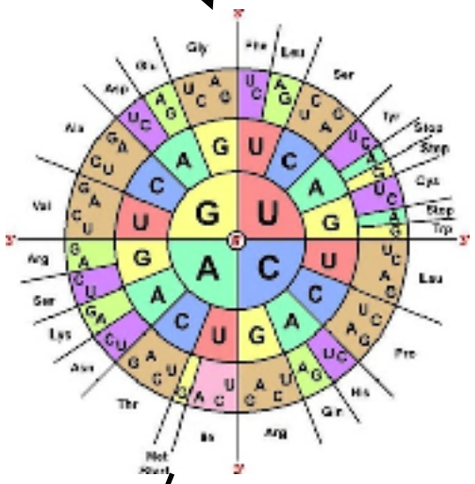
IHGSC **(Фрэнсис Коллинз)**	**Крейг Вентер**
Проект ведется под NIH	Работал в NIH, занимался мРНК, ввел EST
Анонимные доноры, открытость данных	Президент Celera Genomics
Разработали стандарты качества секвенирования	Основатель TIGR
Проект положил начало таким проектам, как ENCODE, 1000 геномов, HapMap	Создатель Синтии (первого биосинтетического организма)
	Создатель первого ассемблера и по сути первого метода NGS
Nature	Science

1865

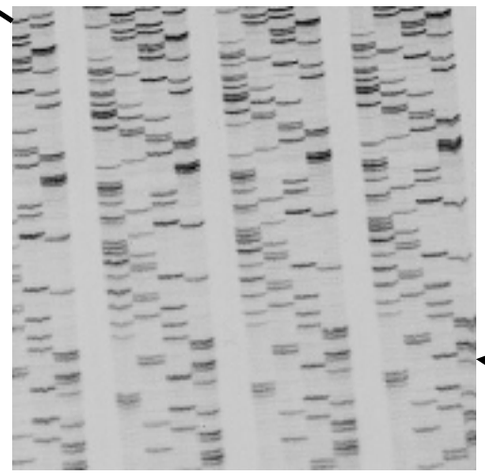


1952

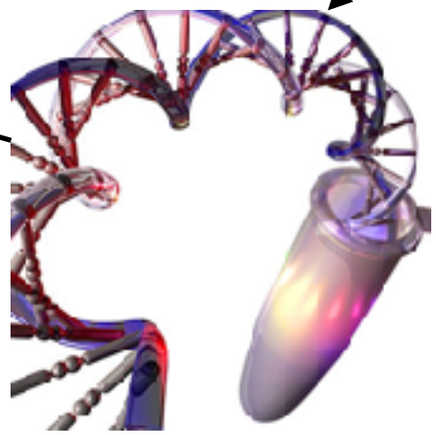
1953



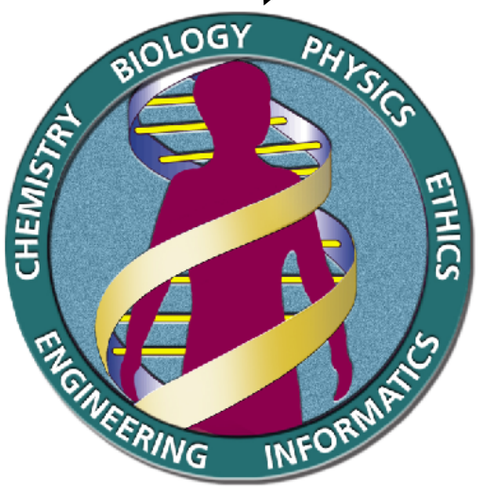
1961



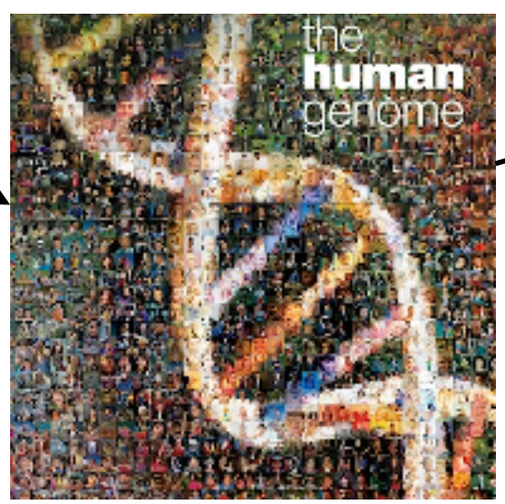
1977



1983



1990



2001

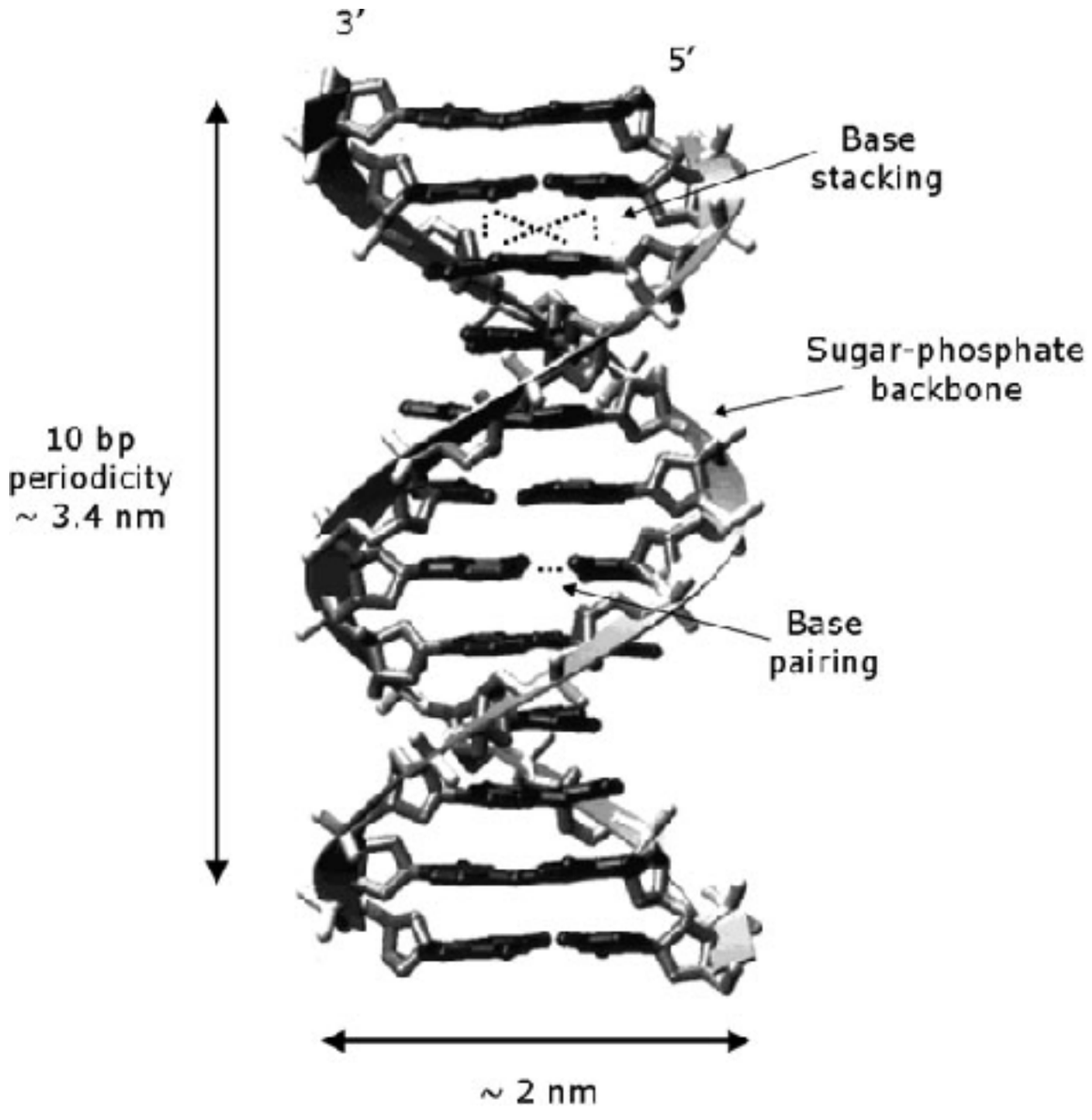
NGS

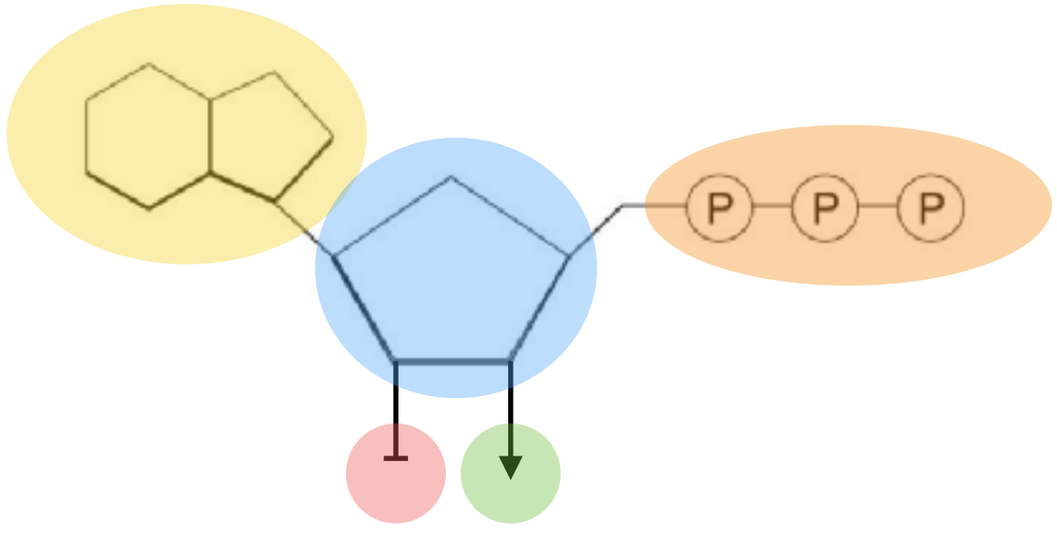
>2010

from https://unlockinglifescode.org/timeline

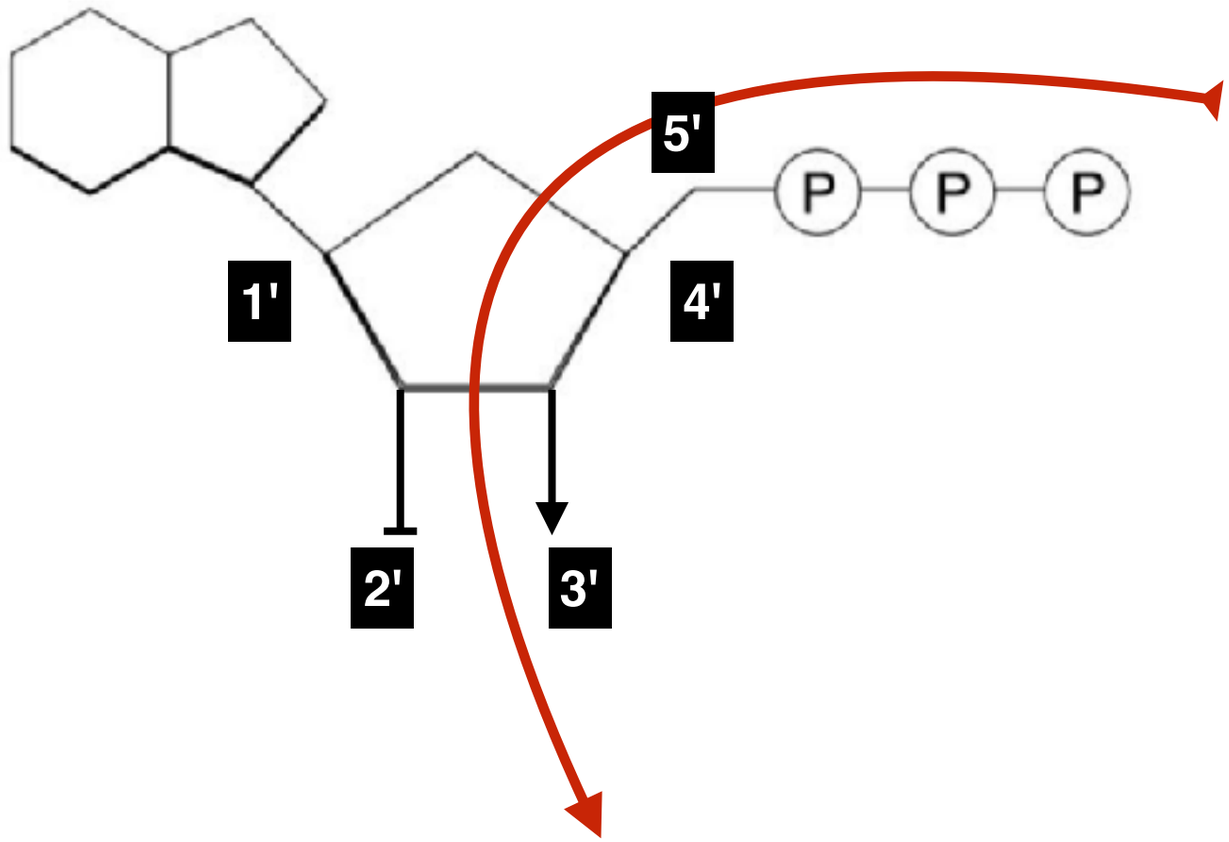
Проект ”Геном человека”

Структура ДНК

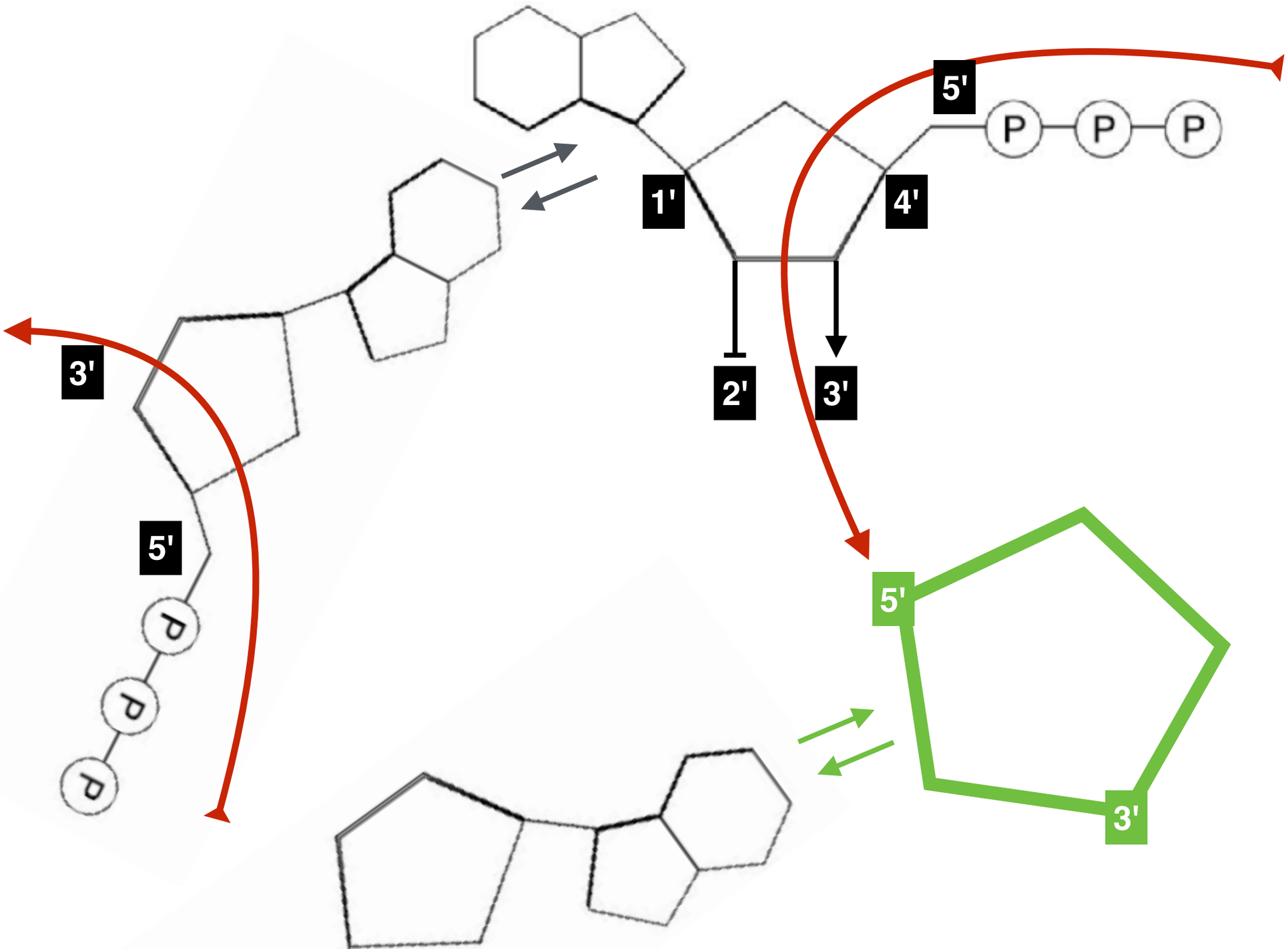




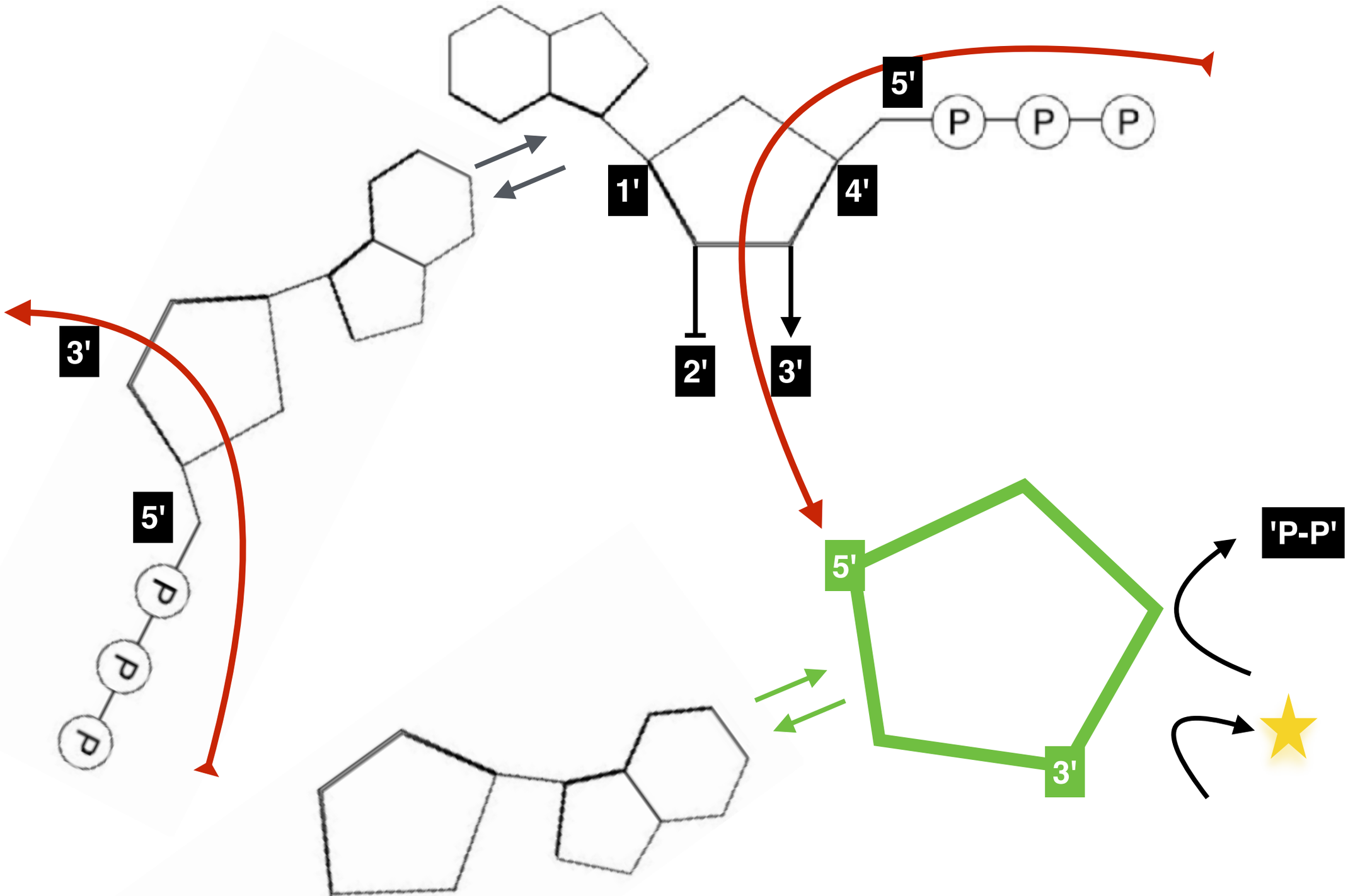
Нуклеотид в цепи



Полимеразная реакция



Включение нуклеотида

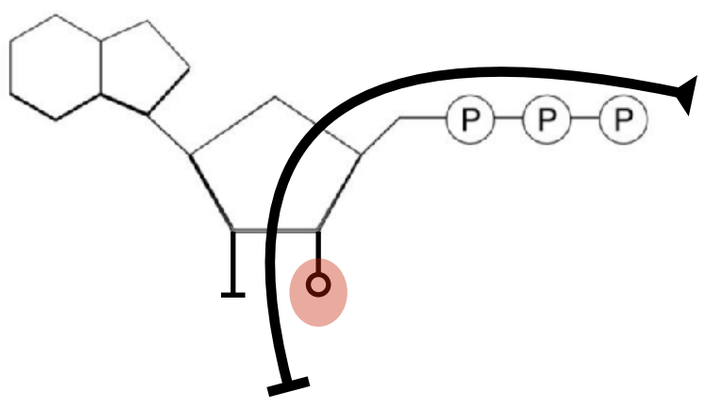


”По Сэнгеру”

ДНК секвенирование первого поколения

- Исторически ***первый*** метод секвенирования
- Как и практически все методы секвенирования основан на активности ДНК-зависимой ***ДНК-полимеразы***
- В качестве способа детекции сигнала использовался метод ***авторадиографии***
- Основан на вероятностном прерывании синтеза цепи в определенной позиции (***ddNTP***)

ddNTP не имеет гидроксила на 3' конце, поэтому продление цепи невозможно



”По Сэнгеру”: принцип метода

праймер **AGTC**

матрица **TCAGATCTAGGTACTG**

ddATP

AGTCTA
AGTCTAGA
AGTCTAGATCCA
AGTCTAGATCCATGA

ddGTP

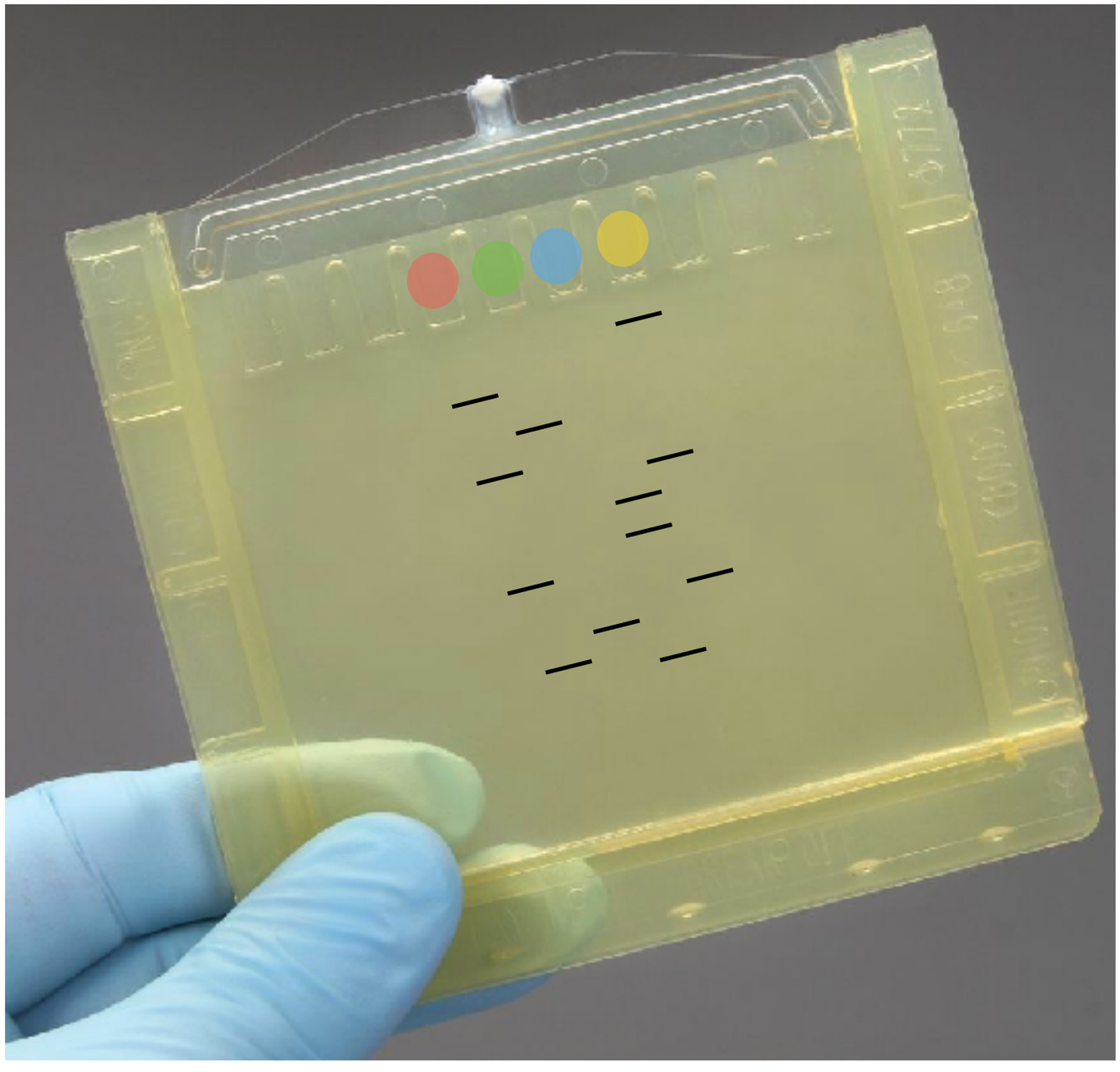
AGTCTAG
AGTCTAGATCCATG

ddCTP

AGTCTAGATC
AGTCTAGATCC
AGTCTAGATCCATGAC

ddTTP

AGTCT
AGTCTAGAT
AGTCTAGATCCAT



FASTA формат

A, G, T, C и другие буквы

```
>KE147310.1 Leishmania tropica L590 unplaced genomic scaffold
TCACAAATAACTGTGTCGGCGCATGGCTGTAAGGCGTGTGCGGCCACGGCTGCCGCGCACCACGCGATGG
GGTCTGTGACAGGGGGCGAGGCAGTGATCTTGTGTTCTGTGGGCGGAATGGCGACATTGAAGGAAAACGT
AATGACGAGCACCTCCCTACATCTTATAGACATACATACACGCAAACCTCAGGTTACGCATTAATGATTC
ACGCACTTCTGCTTCTATAATGTTTTCAGTGCTTATATGTGTGCGTGTTTGGTTGCCAGCGATGTGTGTT
GCCTCTCTTGTGTCTTCTCGTCGTTTCTGTCGTCTCTCGCCGTACTGCTATTGACGAAGACGGATGAAAC
GGTGCCGAAGAGCAAATCGTCCCCTGTTCAAGATCACTGGCTGAAAAGGTCGATAGCTGGAGGGCCGTTT
CGCCGGACTTCGATGCCGCAGTCAGCGCCATCGAGAAAAGATCACCCACCACAGAAGATAAGTCCCCTTT
CCTTGAGAGATGCGAATGGCGGAACTGAGCATCGCGCGCATATTAGAGACCGTCTGCGGCACCAAGCTAC
```

5' AATGACGAGCACCTCCCTACATCTTATAGACATACATACACGCAAACCTCAGGTTACGCATTAATGATTC **3'**

3' TTACTGCTCGTGGAGGGATGTAGAATATCTGTATGTATGTGCGTTTGGAGTCCAATGCGTAATTACTAAG **5'**

reverse & complement

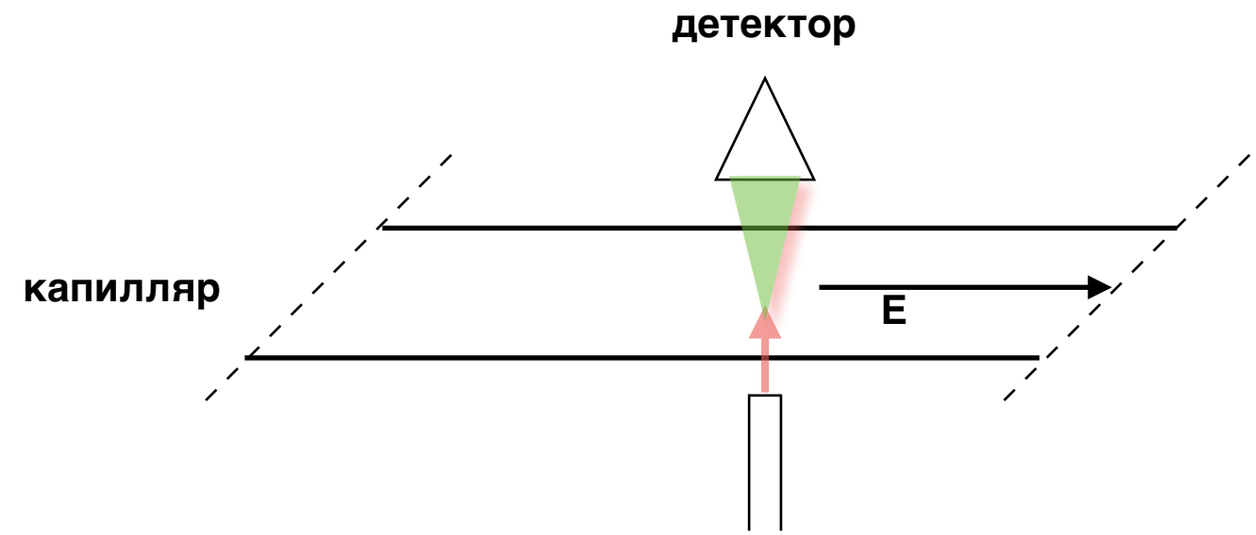
5' GAATCATTAATGCGTAACCTGAGGTTTGCGTGTATGTATGTCTATAAGATGTAGGGAGGTGCTCGTCATT **3'**

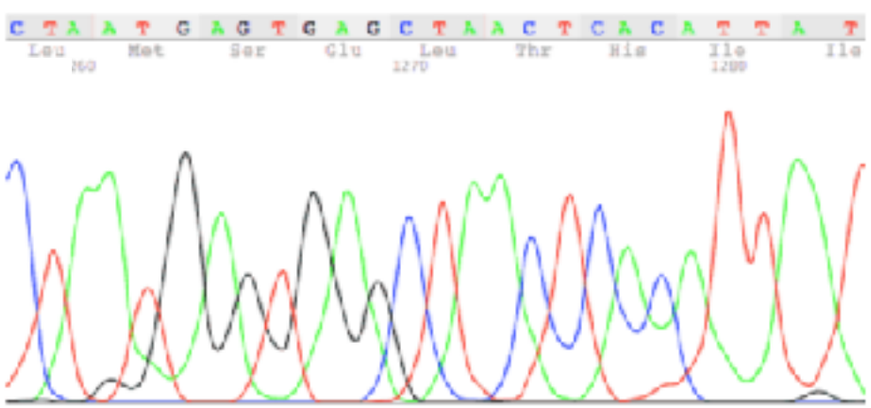
”По Сэнгеру”: 96x

ABI Prism 3100

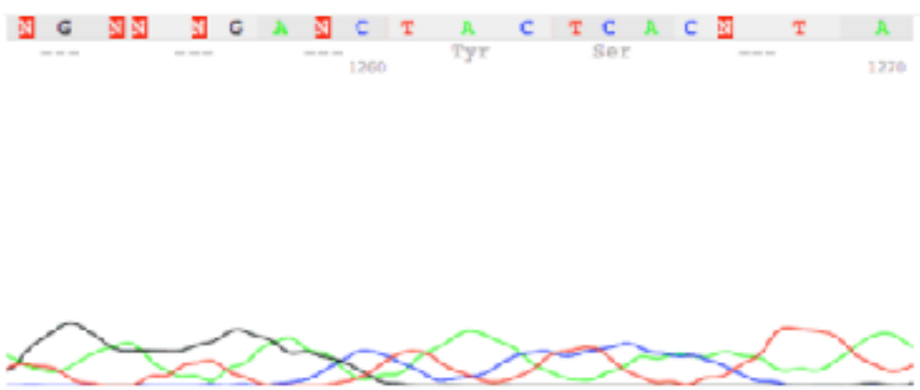
Автоматизация метода секвенирования методом Сэнгера: капиллярные машины.

Качество прочтения: с какой вероятностью нуклеотид в данной позиции прочитан с ошибкой





...



Ошибки секвенирования

Источники ошибок

На этапе секвенирования образца	На этапе пробоподготовки
Обусловлены особенностями измерительной системы прибора и технологиями секвенирования	Вызваны присутствием нецелевых молекул в образце, ошибками ферментов (полимеразы)
Можно выявить или предсказать. На механизме предсказания ошибки основано присвоение качества прочтения основания прибором.	Нельзя выявить на этапе секвенирования Как правило, одни и те же причины вне зависимости от технологии секвенирования

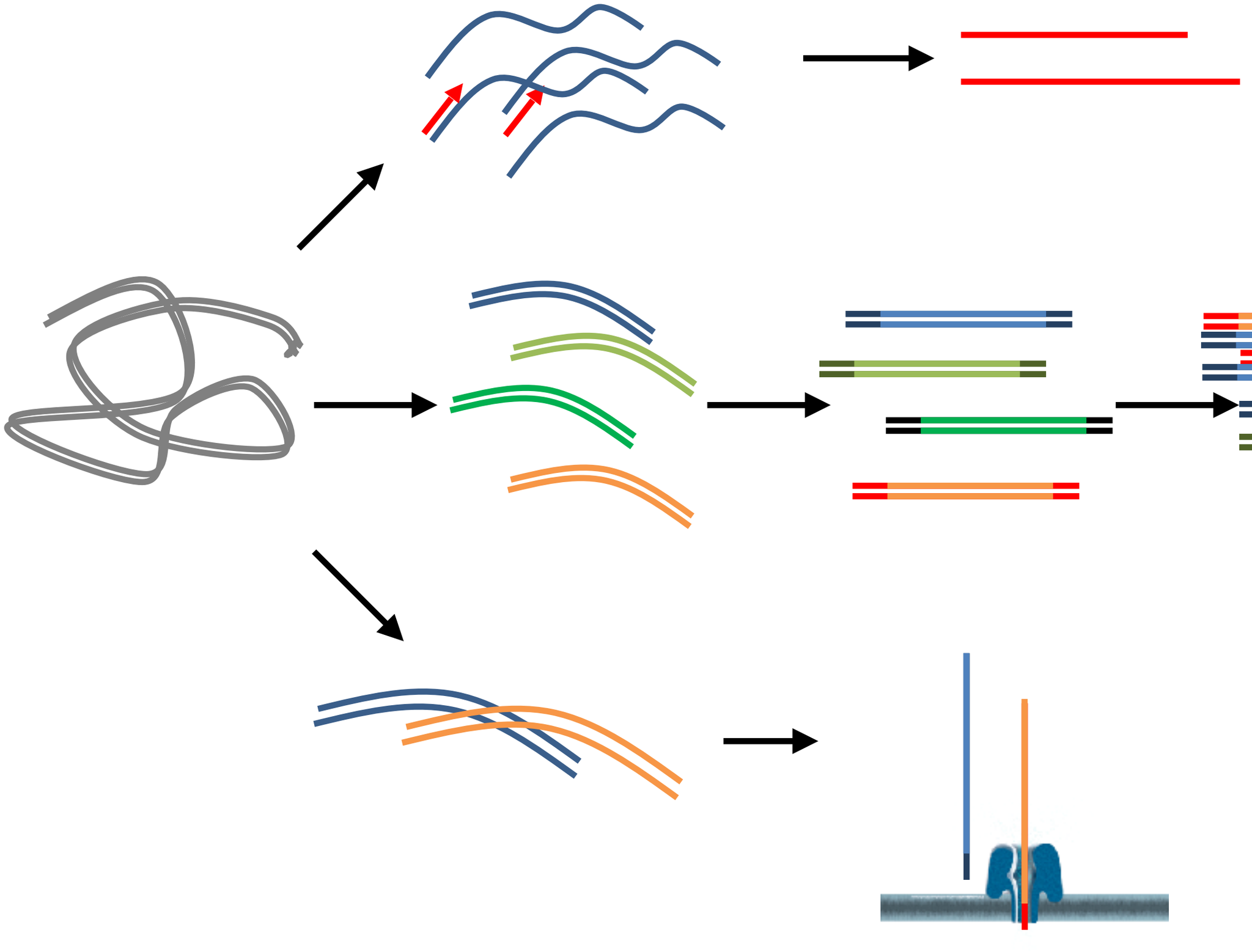
Качество прочтения нуклеотида может быть выражено вероятностью ошибки. Принятый сегодня способ записи ***Phred*** Score (Q).

$$Q = -10 \log_{10} P \quad \Longrightarrow \quad P = 10^{\frac{-Q}{10}}$$

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10000	99.99%
50	1 in 100000	99.999%

Теория поколений

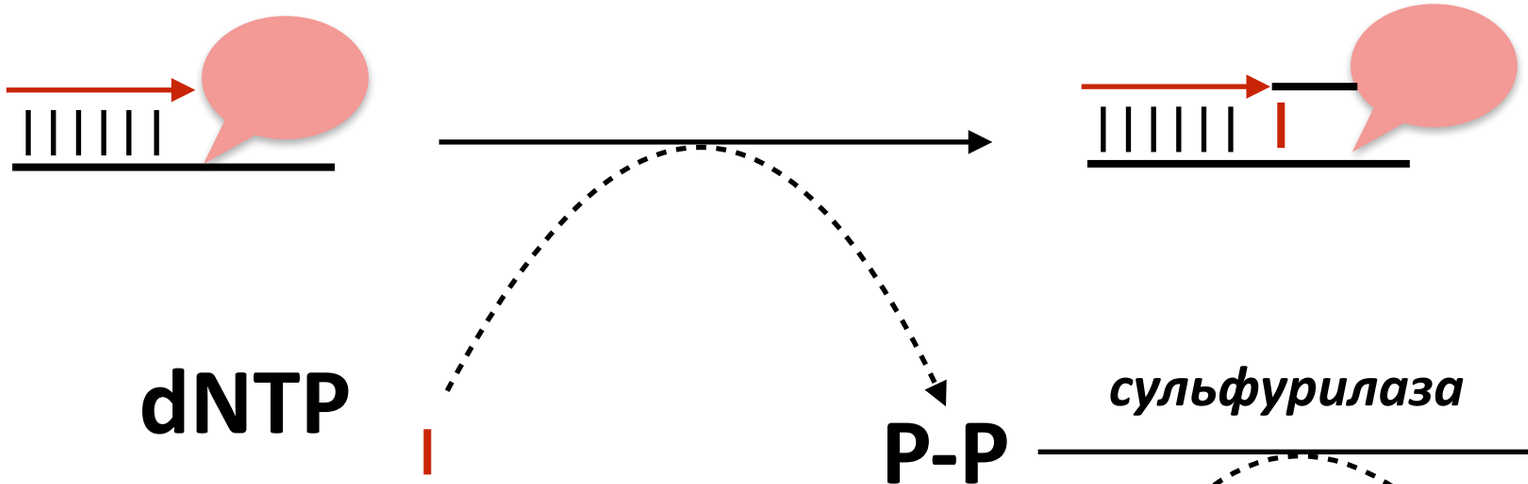
быстрее, больше, дешевле



Пиросеквенирование

"Roche" 454: прочтение с огоньком

ДНК-полимераза
Bacillus stearothermophilus



dNTP

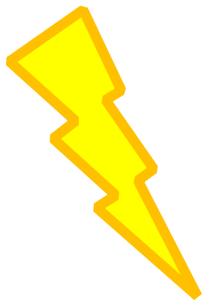
P-P

сульфурилаза

SO_4^{2-}

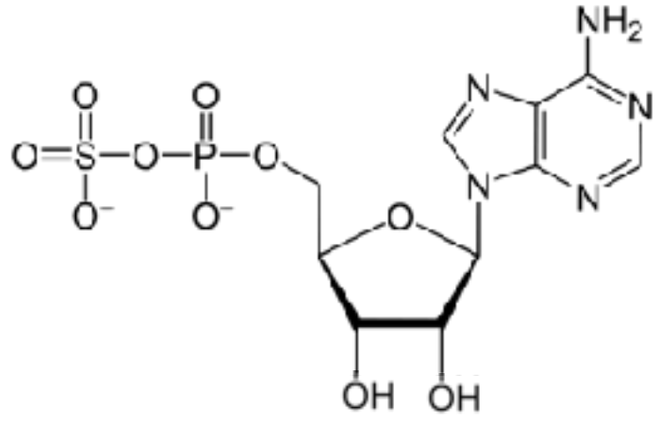
ATP

люцифераза



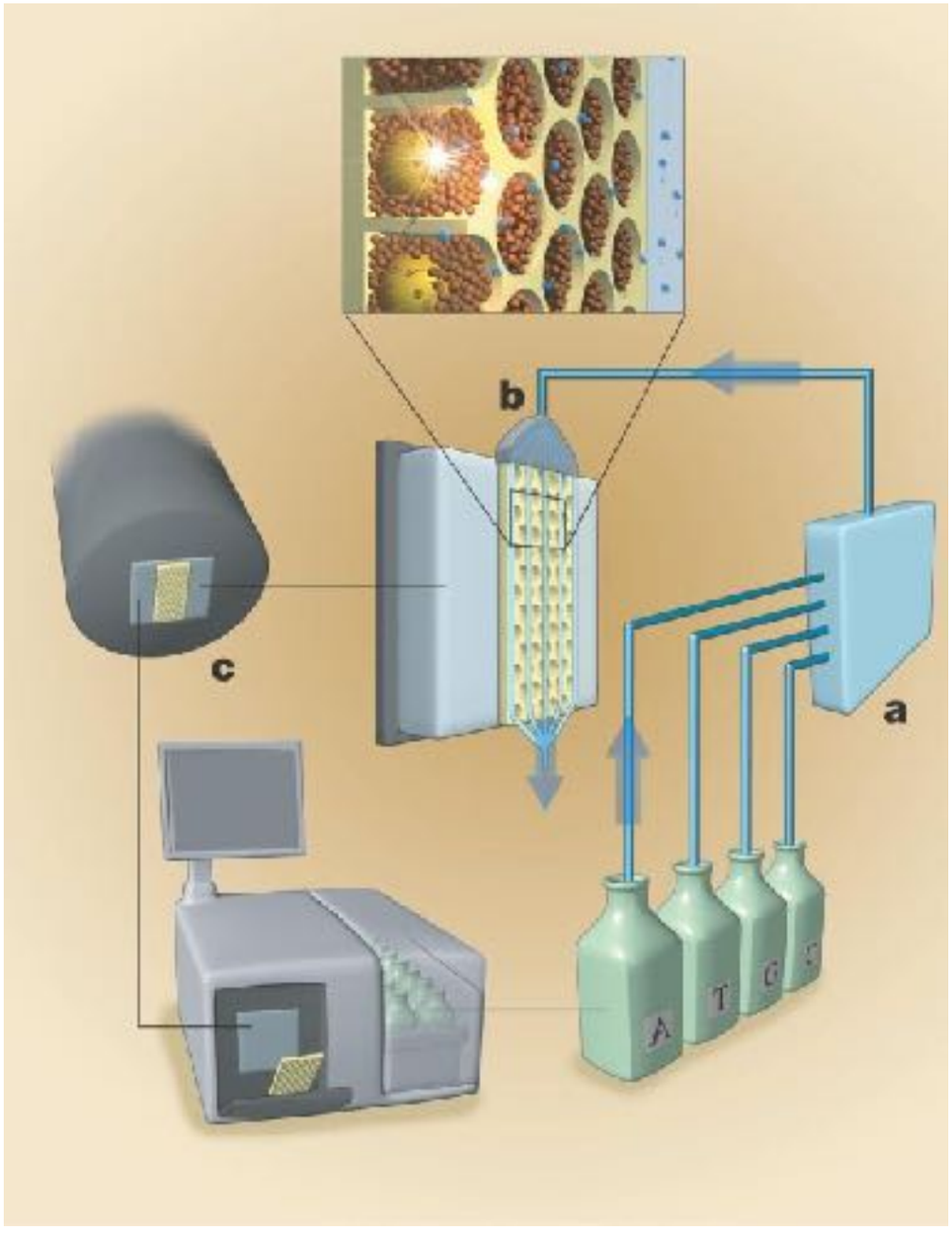
оксилюциферин

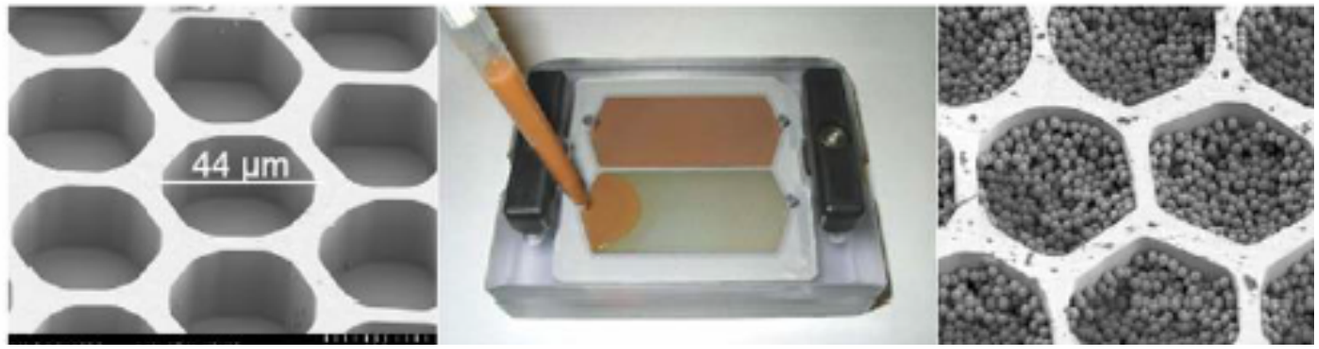
люциферин



аденозин фосфосульфат

Пиросеквенирование

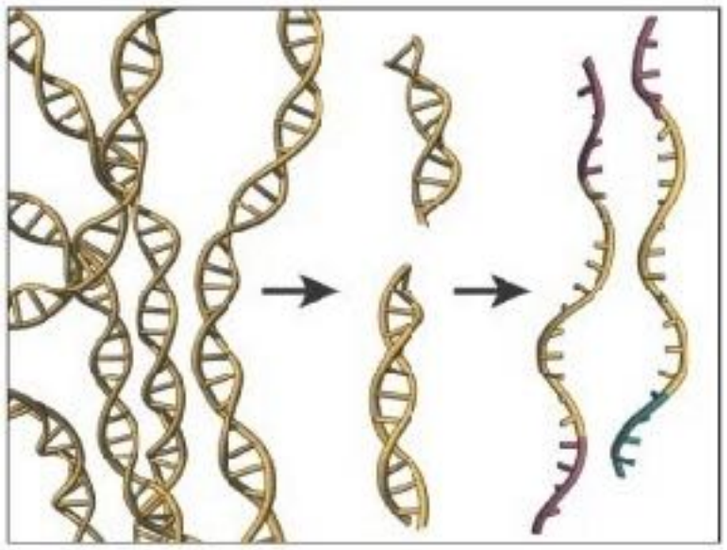


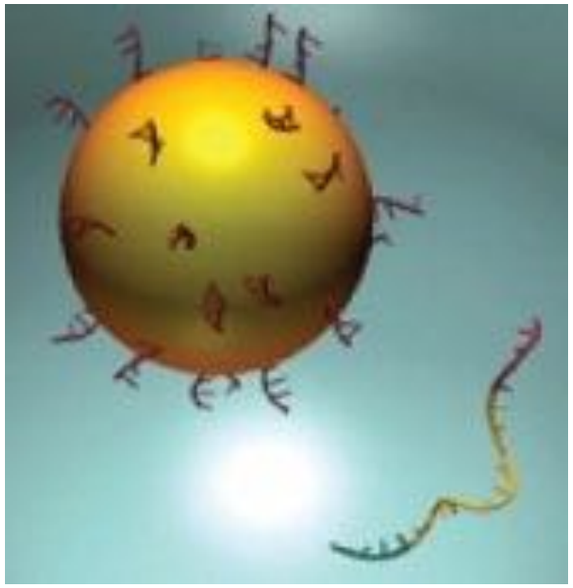


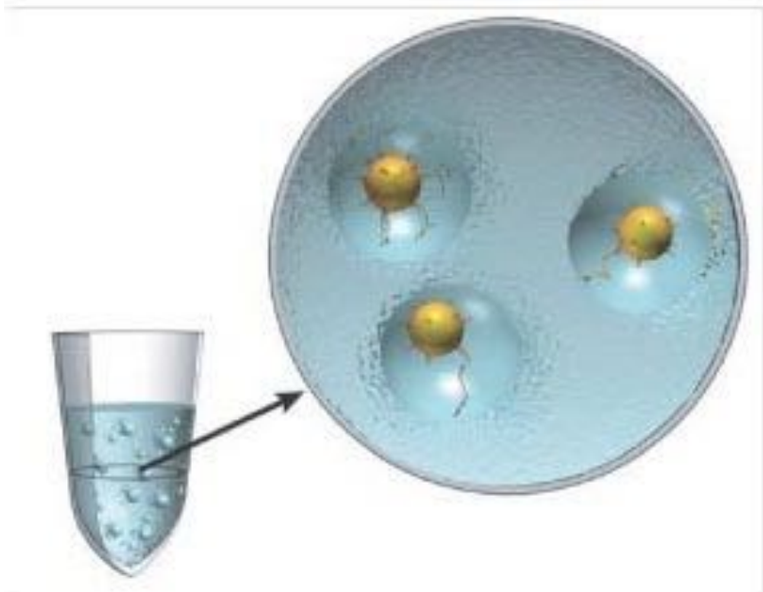
Margulies et al. Nature 2005

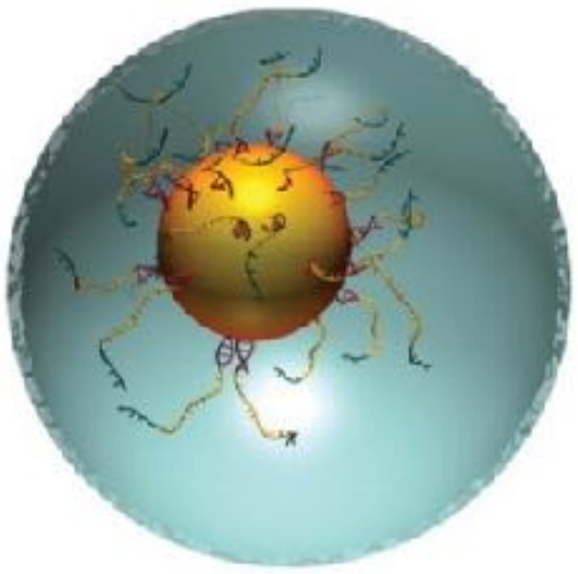
Пиросеквенирование

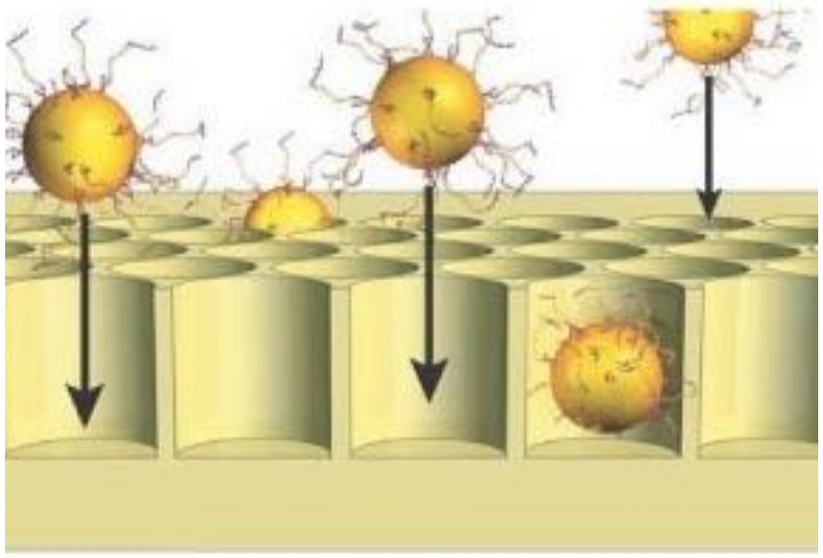
Молекулярные колонии на нанопланетах

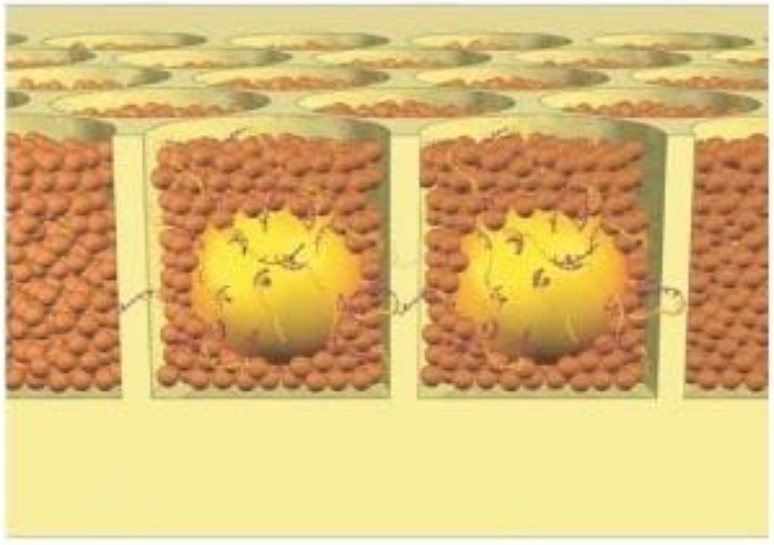








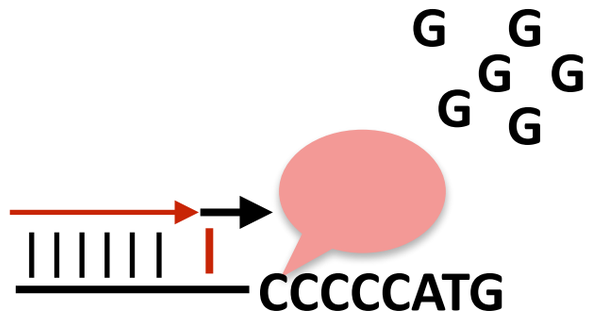




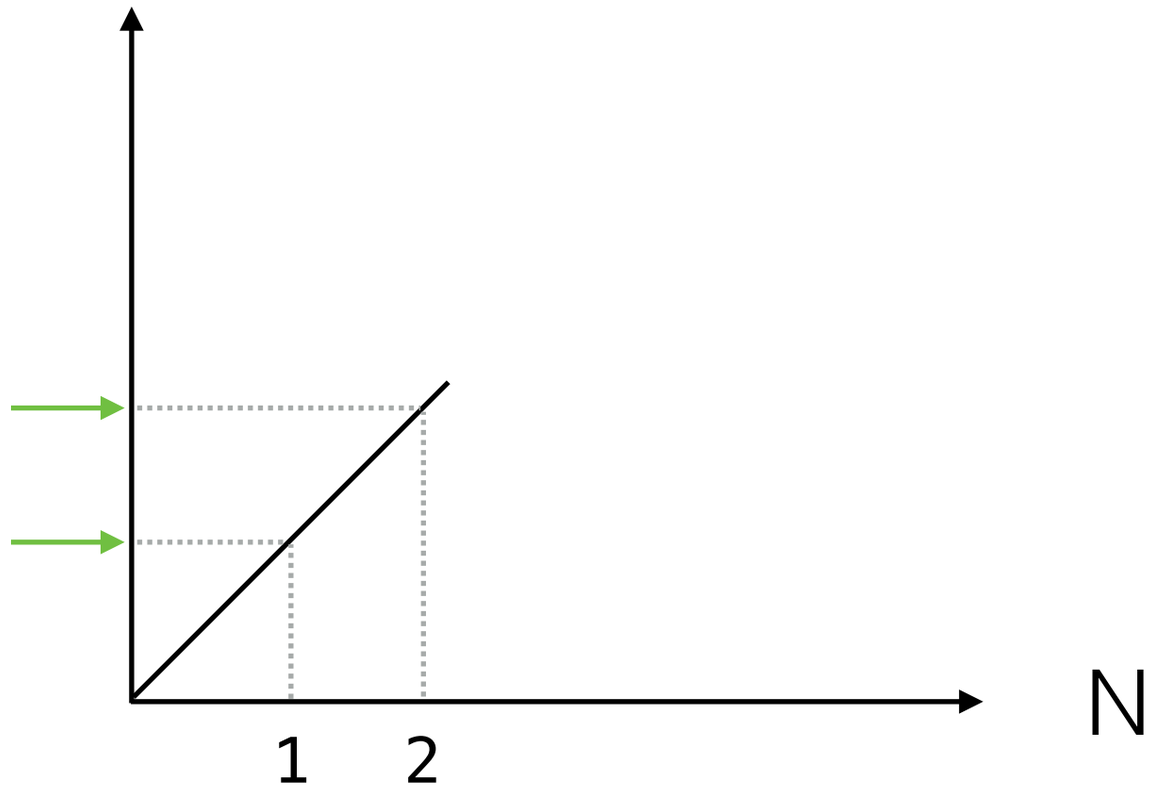
Margulies et al. Nature 2005

Ошибки секвенирования

погрешность измерения

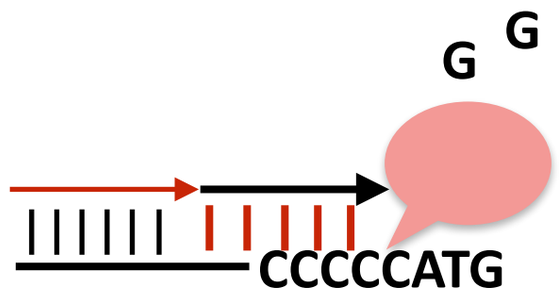


I

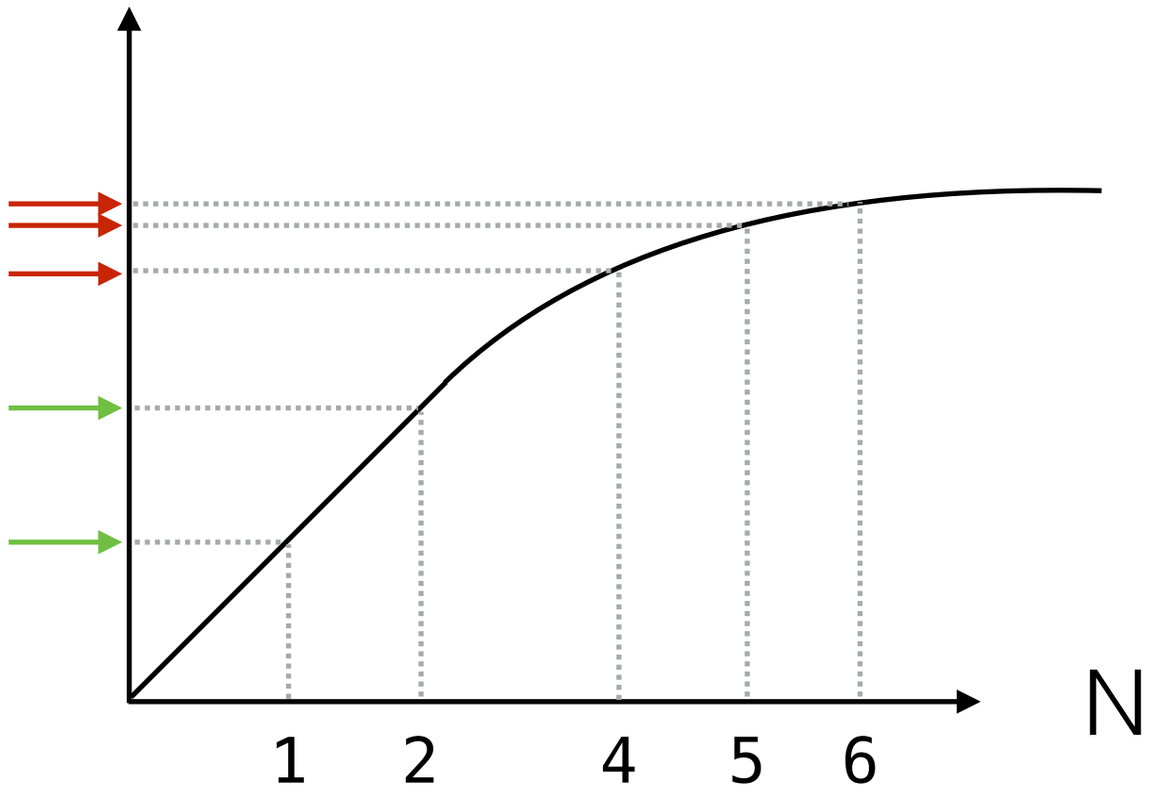


Ошибки секвенирования

погрешность измерения



I



Illumina

"Sequencing-by-Synthesis"

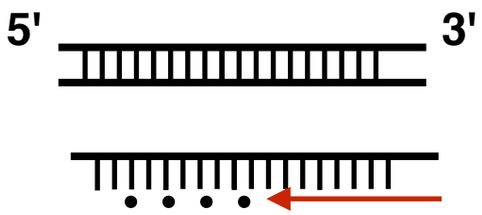
Sample Prep

Cluster Generation

Sequencing

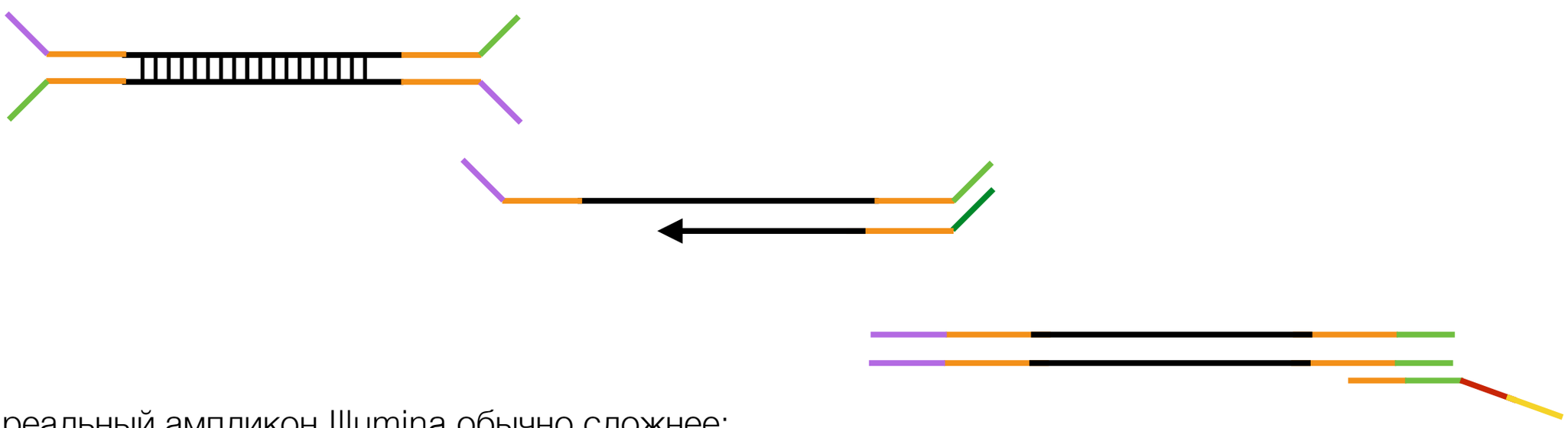
Структура ампликона

Рождение молекулярного шедевра

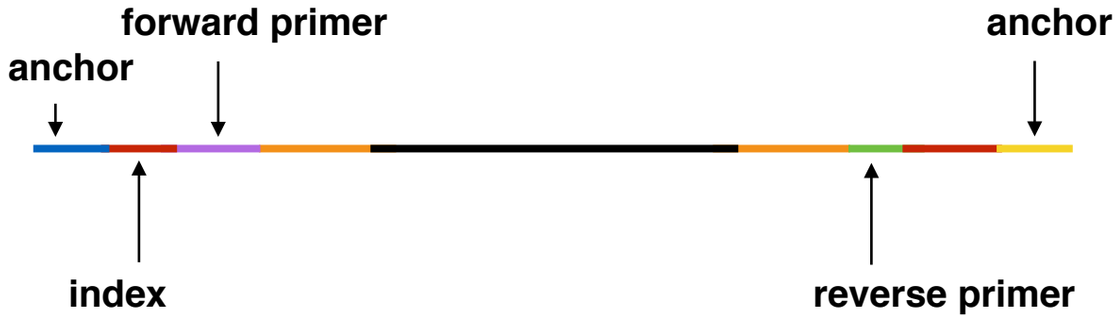


Как добиться того, чтобы молекула ДНК неизвестной последовательности "читалась" лишь с одной определенной стороны?

Лигировать так называемый Y-адаптор амплифицировать с праймера

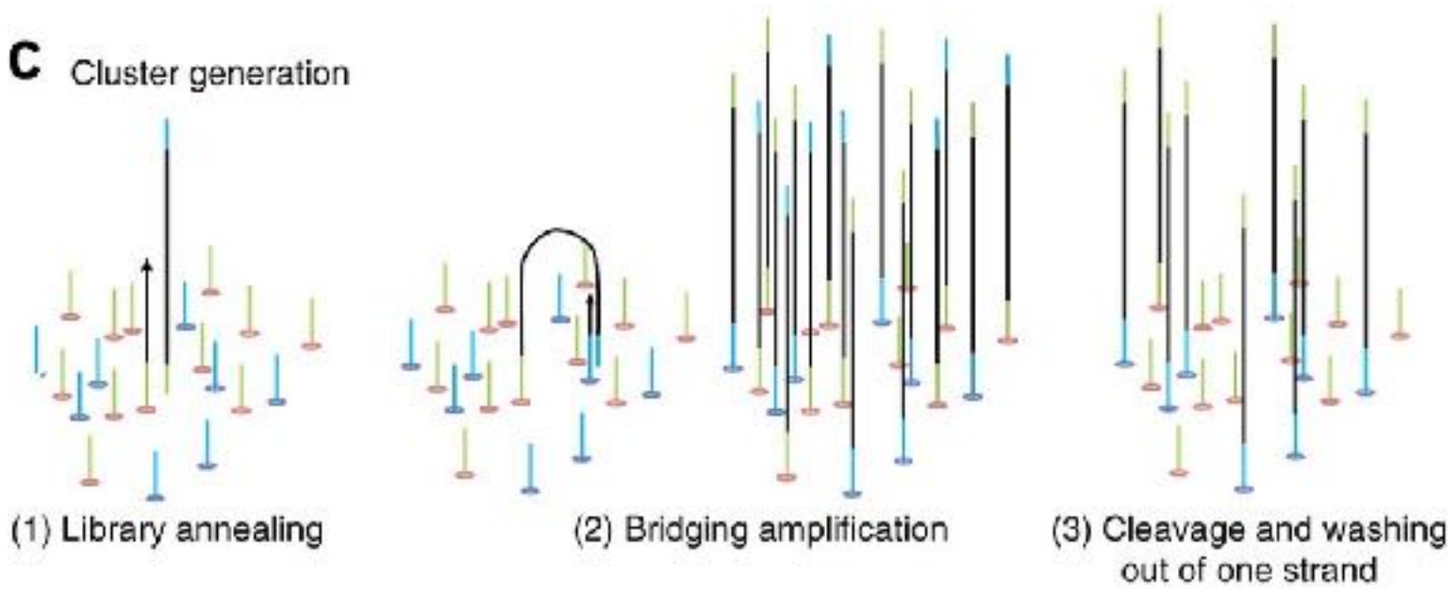


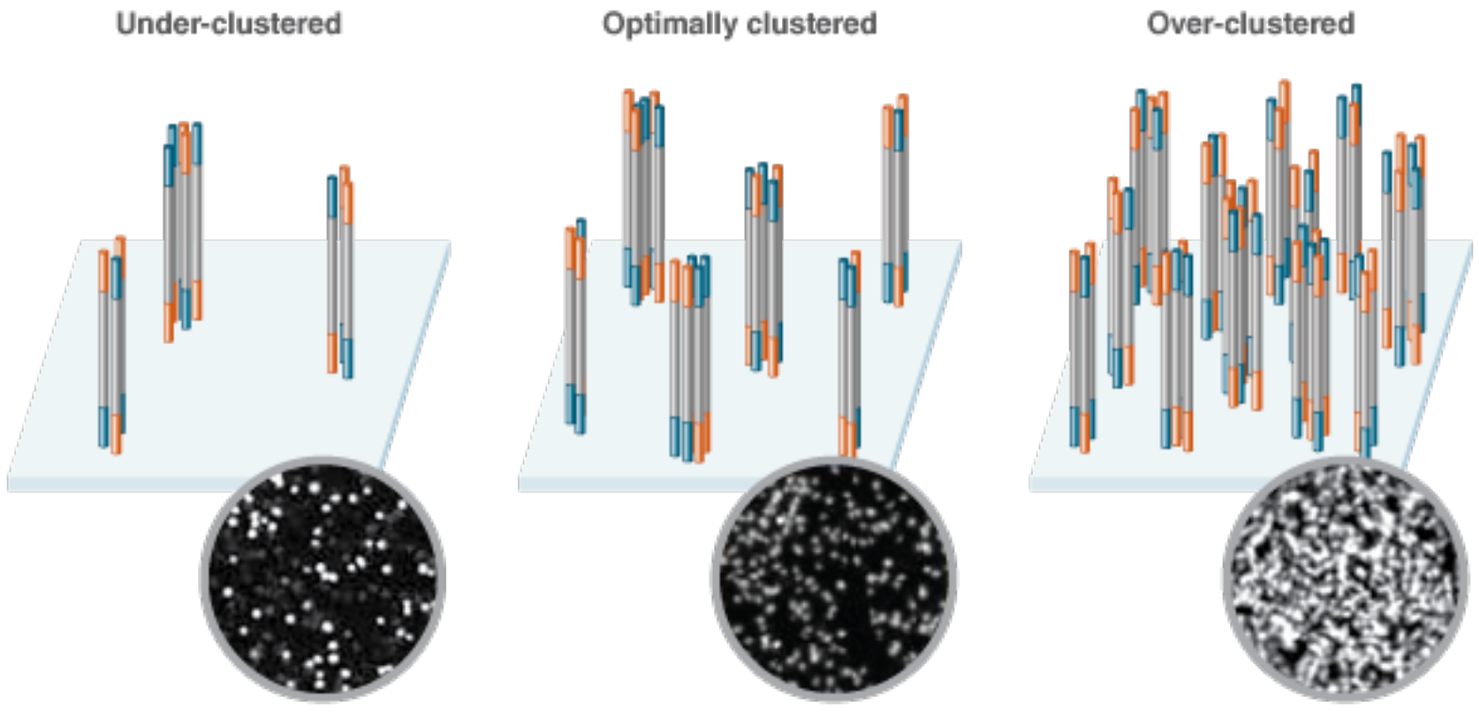
реальный ампликон Illumina обычно сложнее:



Генерация кластеров

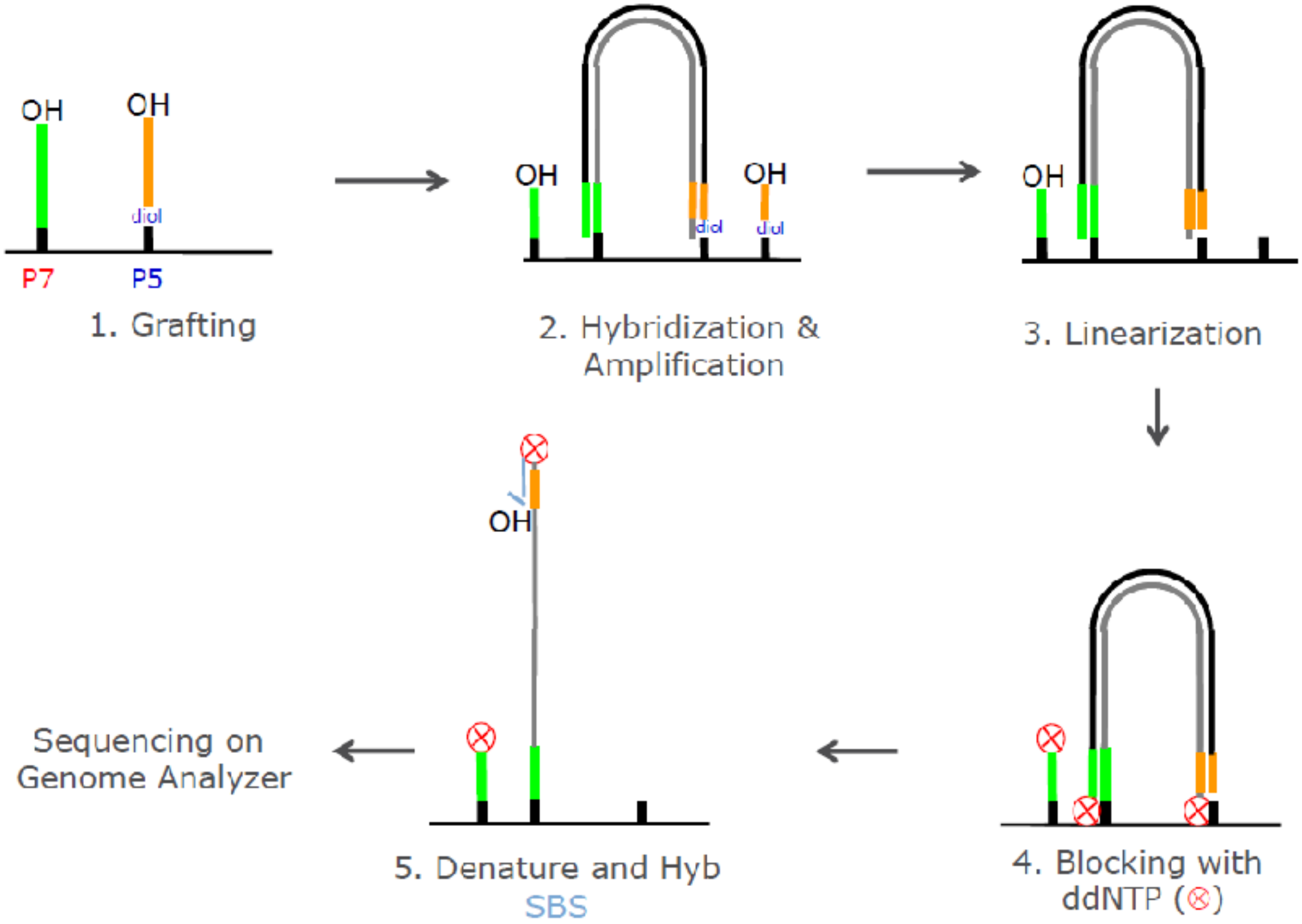
Молекулярные основы ландшафтного дизайна





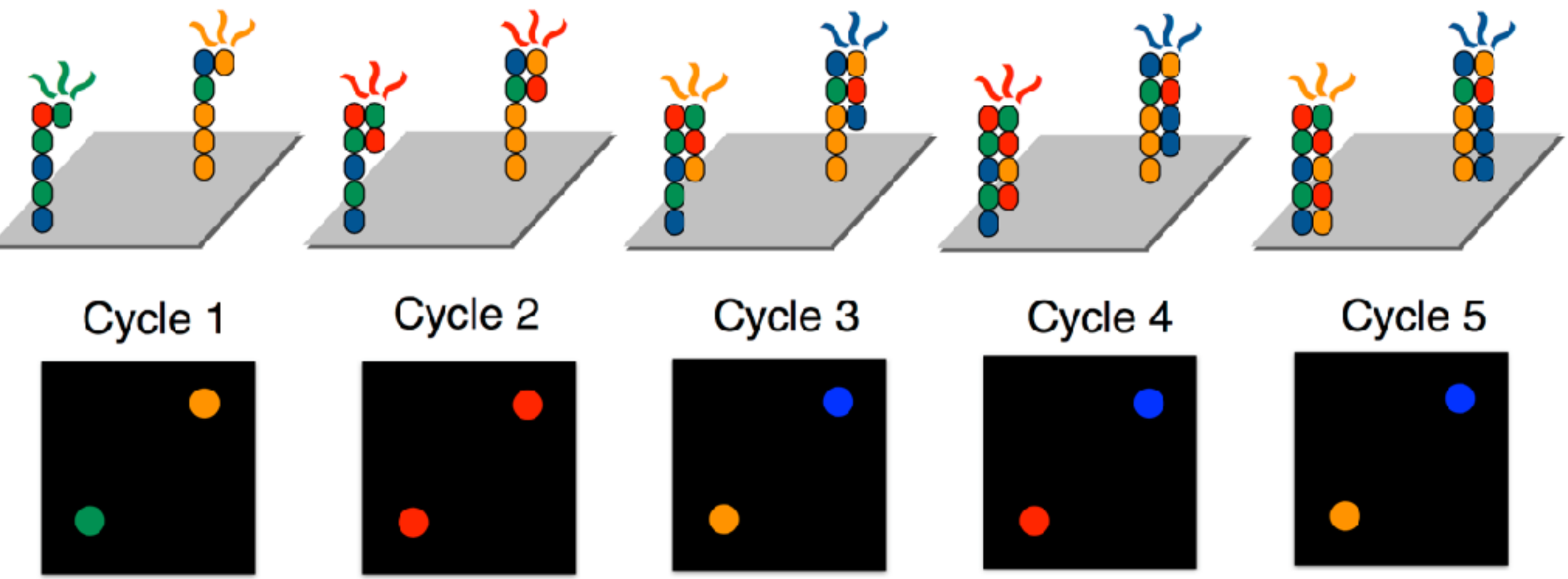
Bridge amplification

and single strand cut



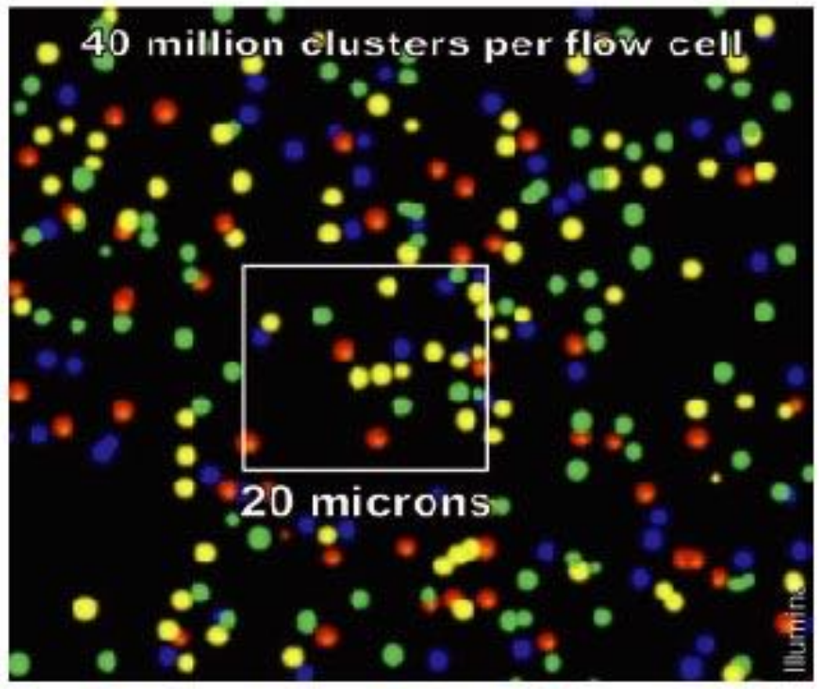
Секвенирование через синтез

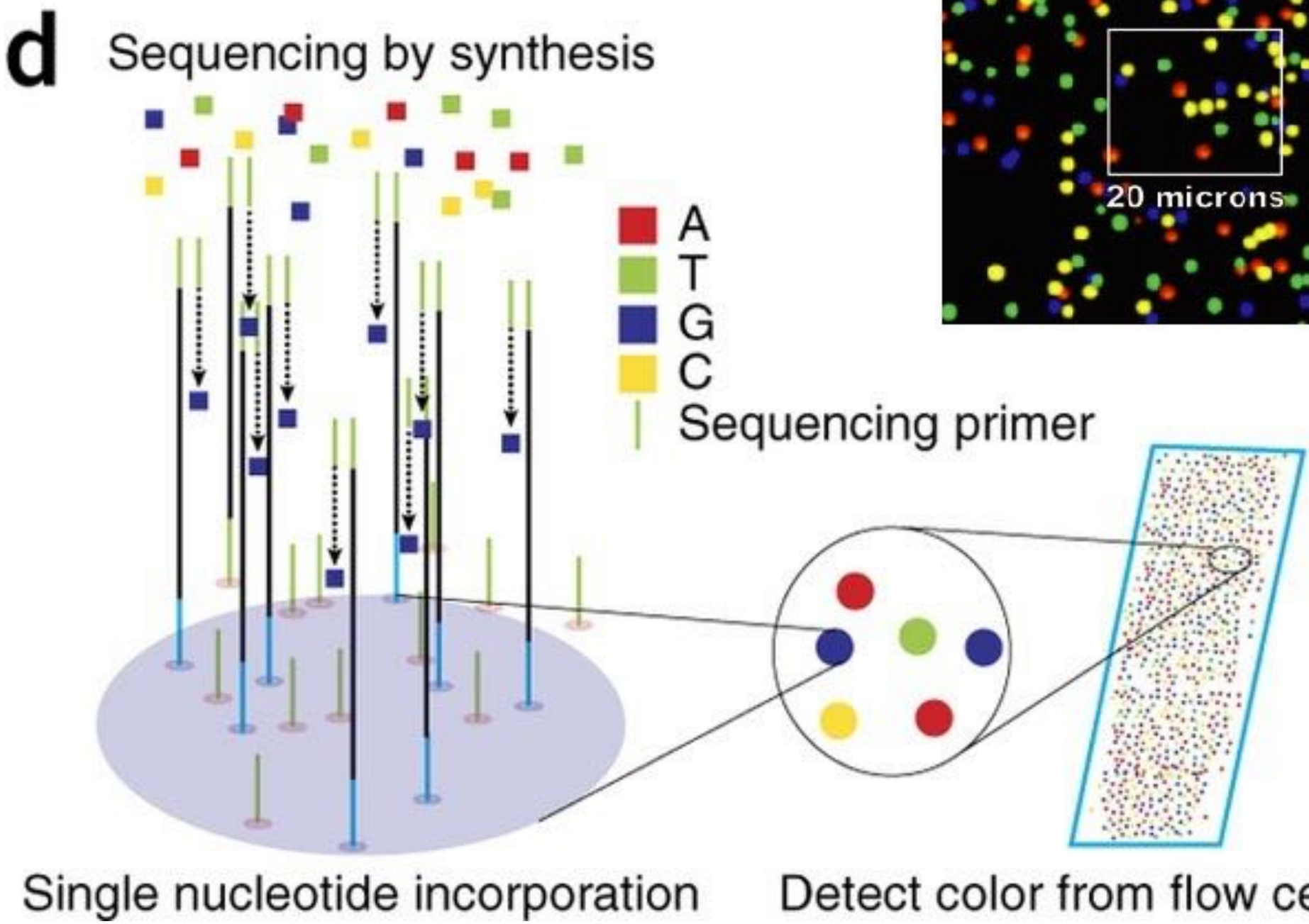
Ёлочка, гори!



Sequencing by synthesis

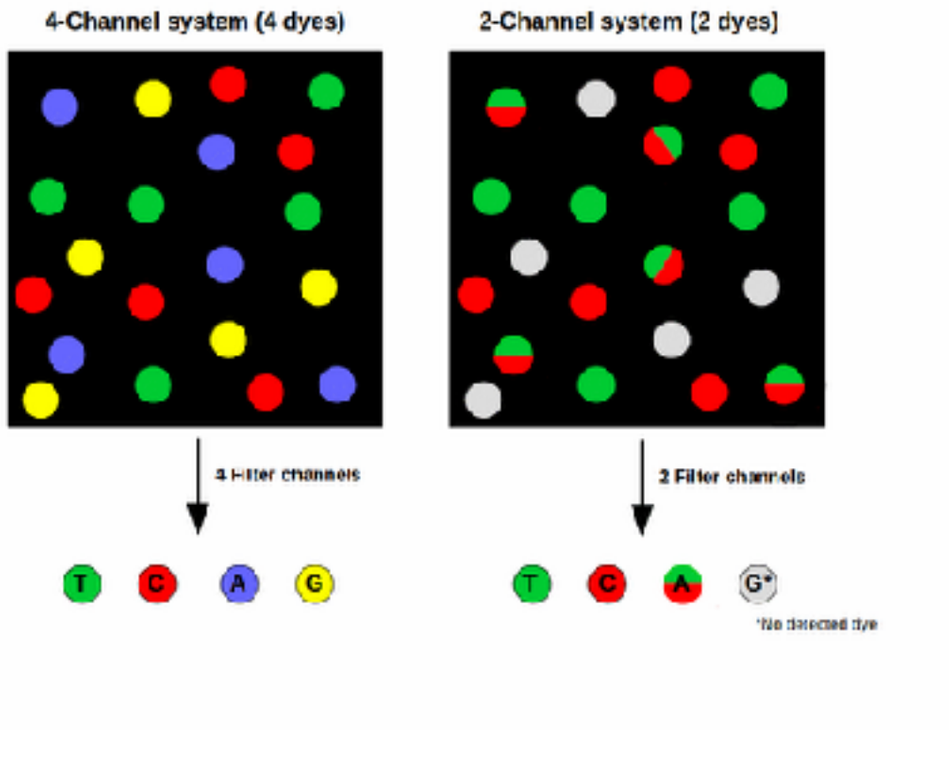
Illumina

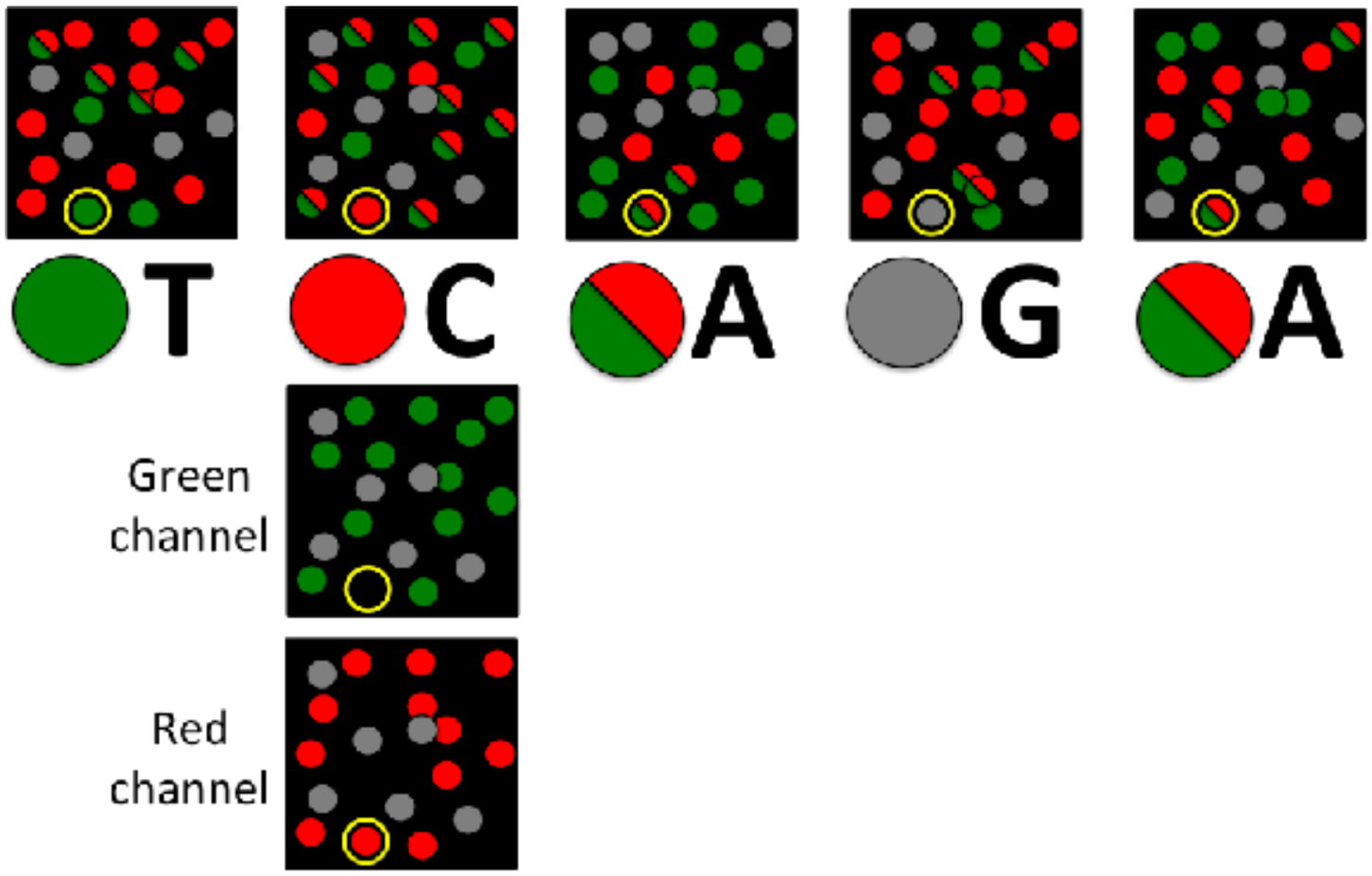


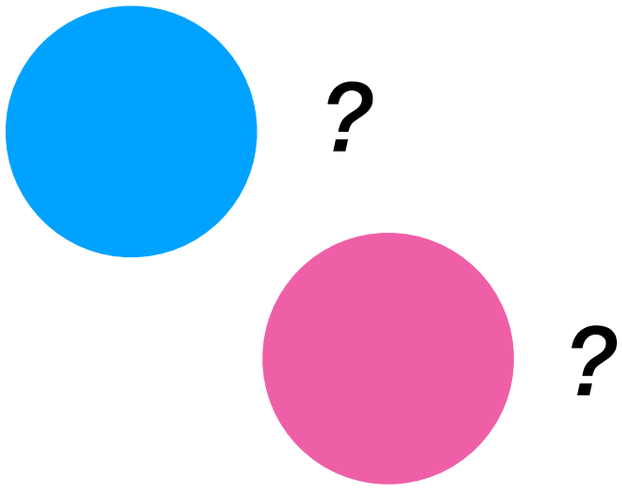


Ошибки секвенирования

Illumina

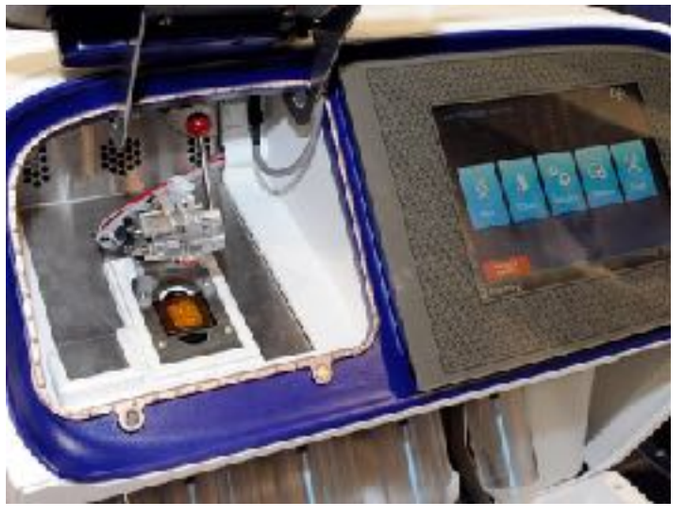




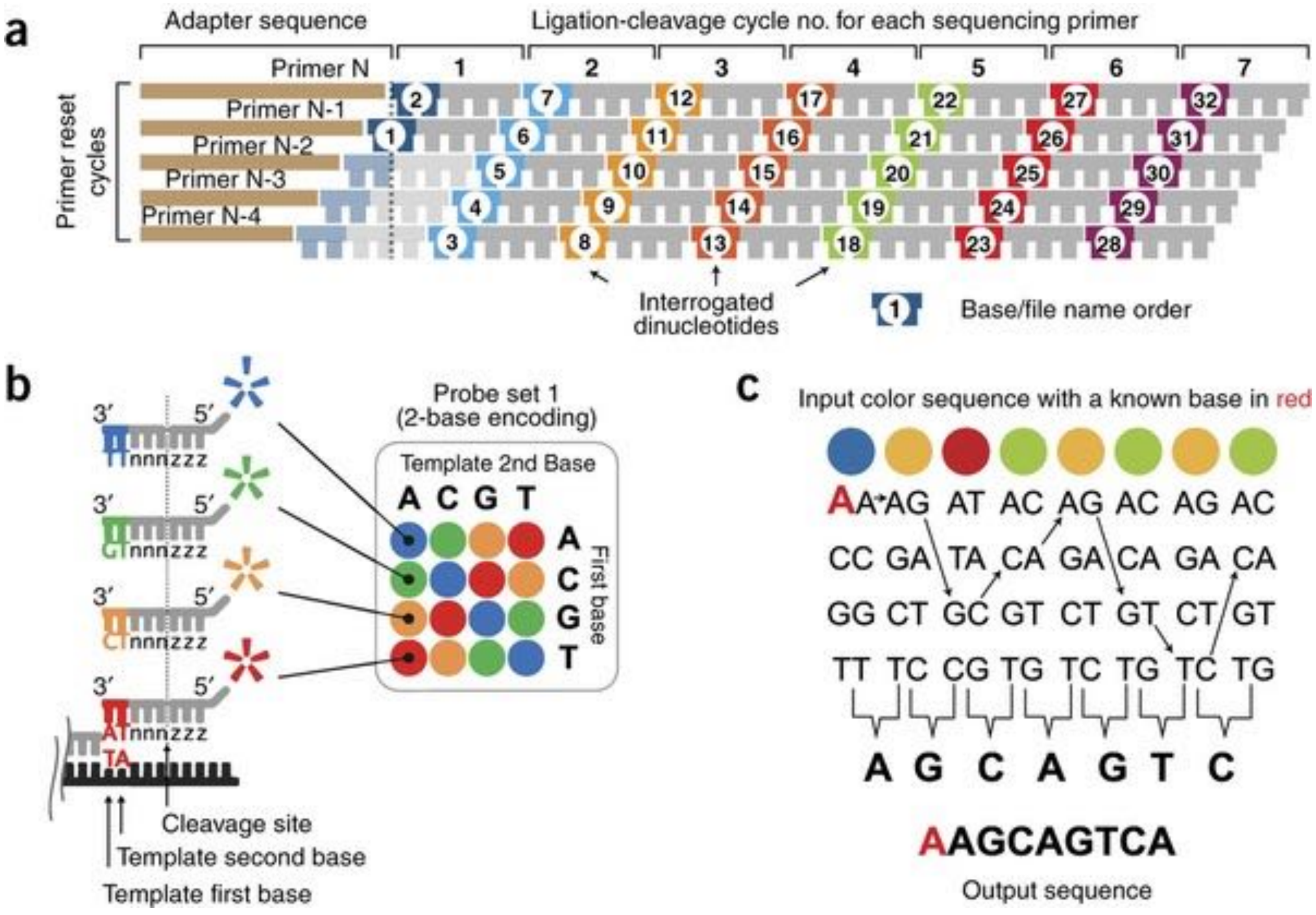


Полупроводниковое секвенирование

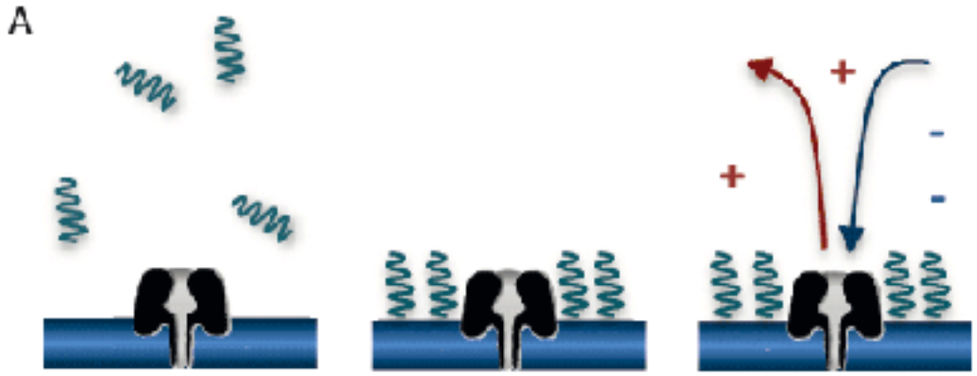
pH метр высокого разрешения

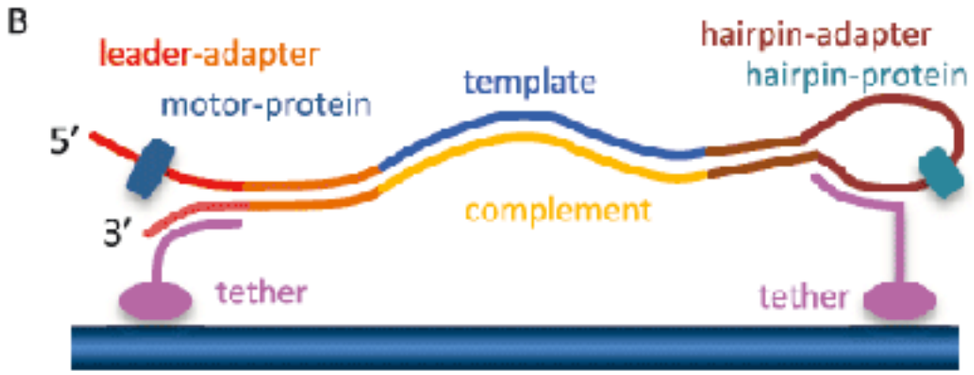


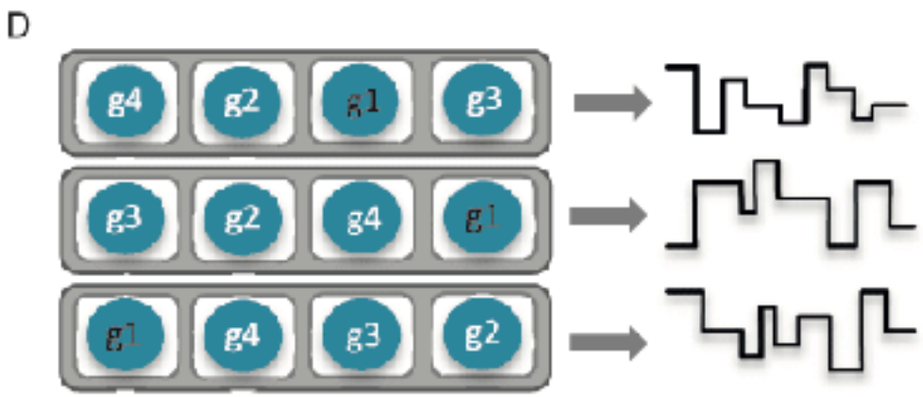
Секвенирование через лигирование

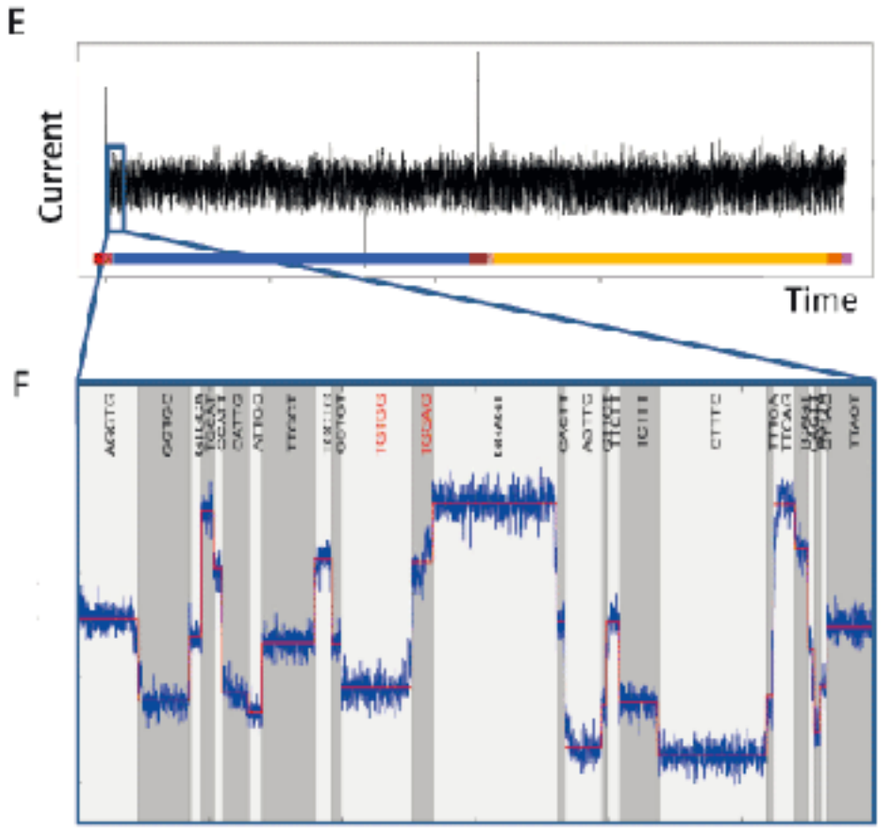


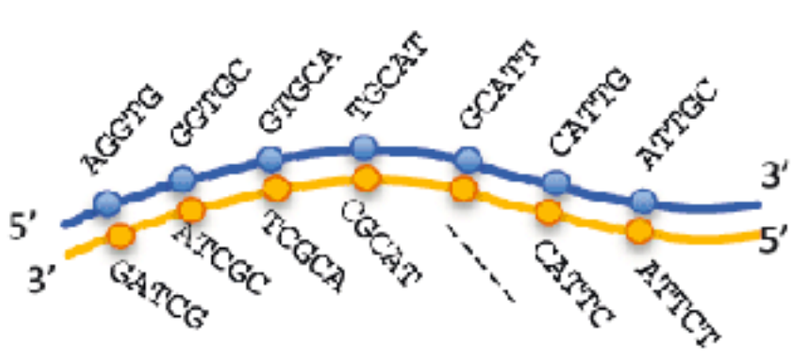
Oxford Nanopore









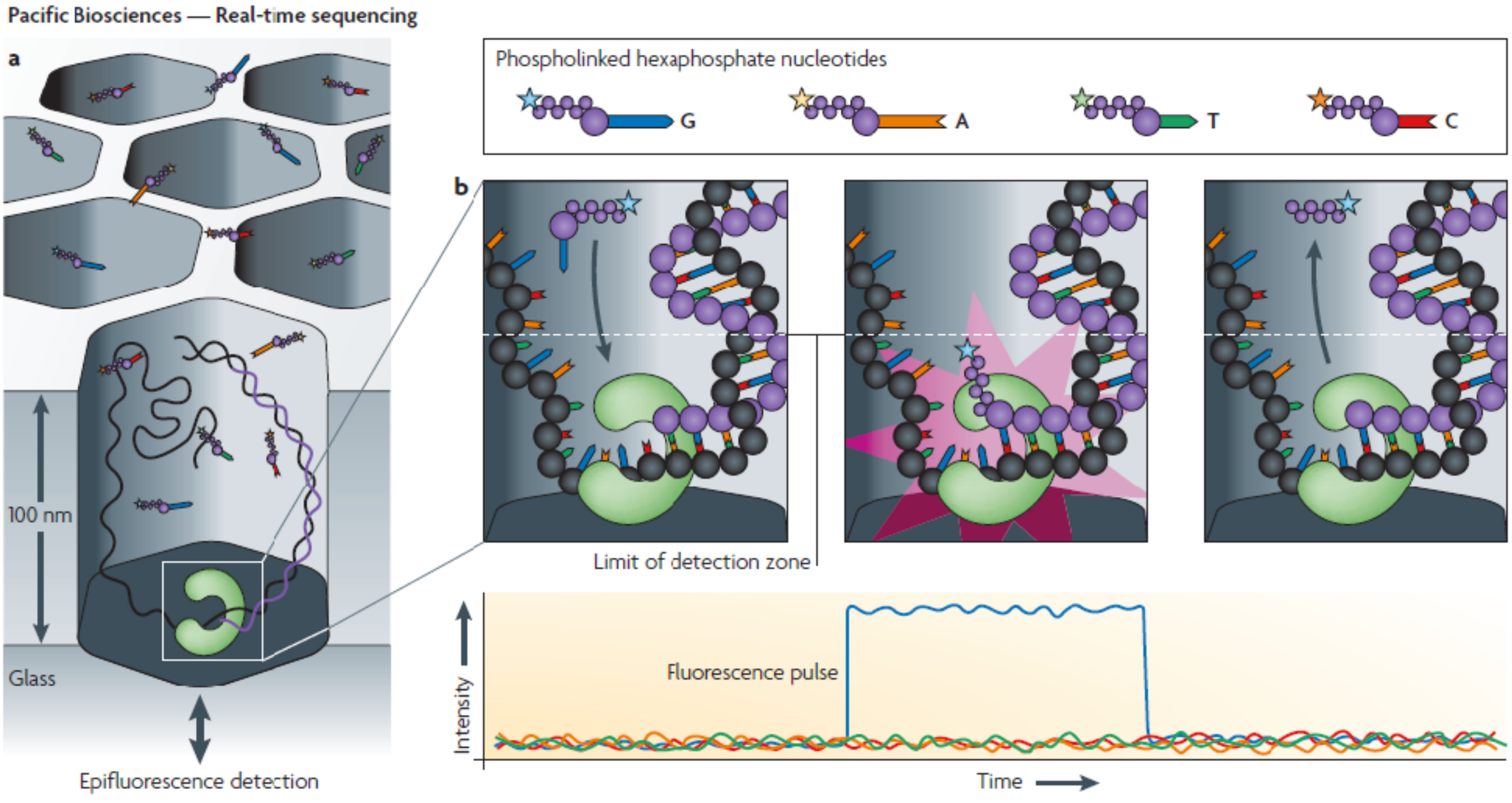


H

```
1D Temp 5′-GTGCATT-G-CTTGAGTTCTTT-CAGTT-A-3′
                ||||  ||  |  ||||||||||||  |||||  |
     2D 5′-CTGCATTCTCCGTGAATTCTTTCGAATTCT-3′
                ||||  ||  |  ||||||||||||  |||||  |
1D Comp 5′-TCGC-TT-TCCGCGAATTCCTT-GAATT-C-3′
```

Pacific Biosystems

Читать молекулы в реальном времени



Сравнение платформ

Тип	454		Ion		Solexa (Illumina)			SOLiD	PacBio	
	Junior	FLX	Torrent	Proton	HiSeq	MiSeq	NextSeq	5500	RSII	Sequel
Длина чтения, bp	400	800	400	200	2x150	2x300	2x150	2x60	10-15kb	10-15kb
Объем данных, Gb	0,035	0,7	2	16	180	15	129	150	1	10
Цена за 1Mbp, $	22	7	0,6	0,02	0,05	0,14	0,03	0,07	0,4	0,09
Цена инструмента, тысяч $	125	500	50	149	740	125	250	600	700	350
Цена за запуск, $	1100	6200	939	1000	6145	1600	4000	10503	400	850
Время работы	10	24	7	4	40	65	29	8 days	0,5-6	0,5-6
Частота ошибок, %	1	1	0,5-2,5	0,5-2,5	0,1	0,1	0,1	--	14	14
Тип ошибок	индели	индели	индели	индели	замены	замены	замены	замены	индели	индели

Спасибо за внимание