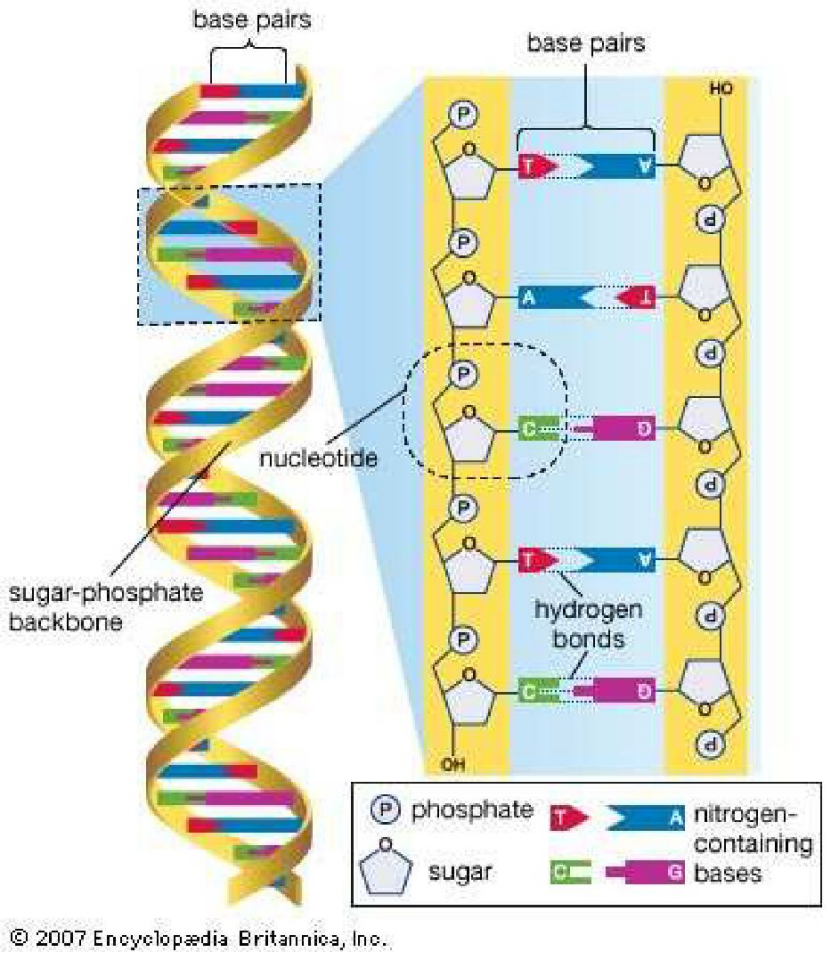
Технология секвенирования генома и сборка генома

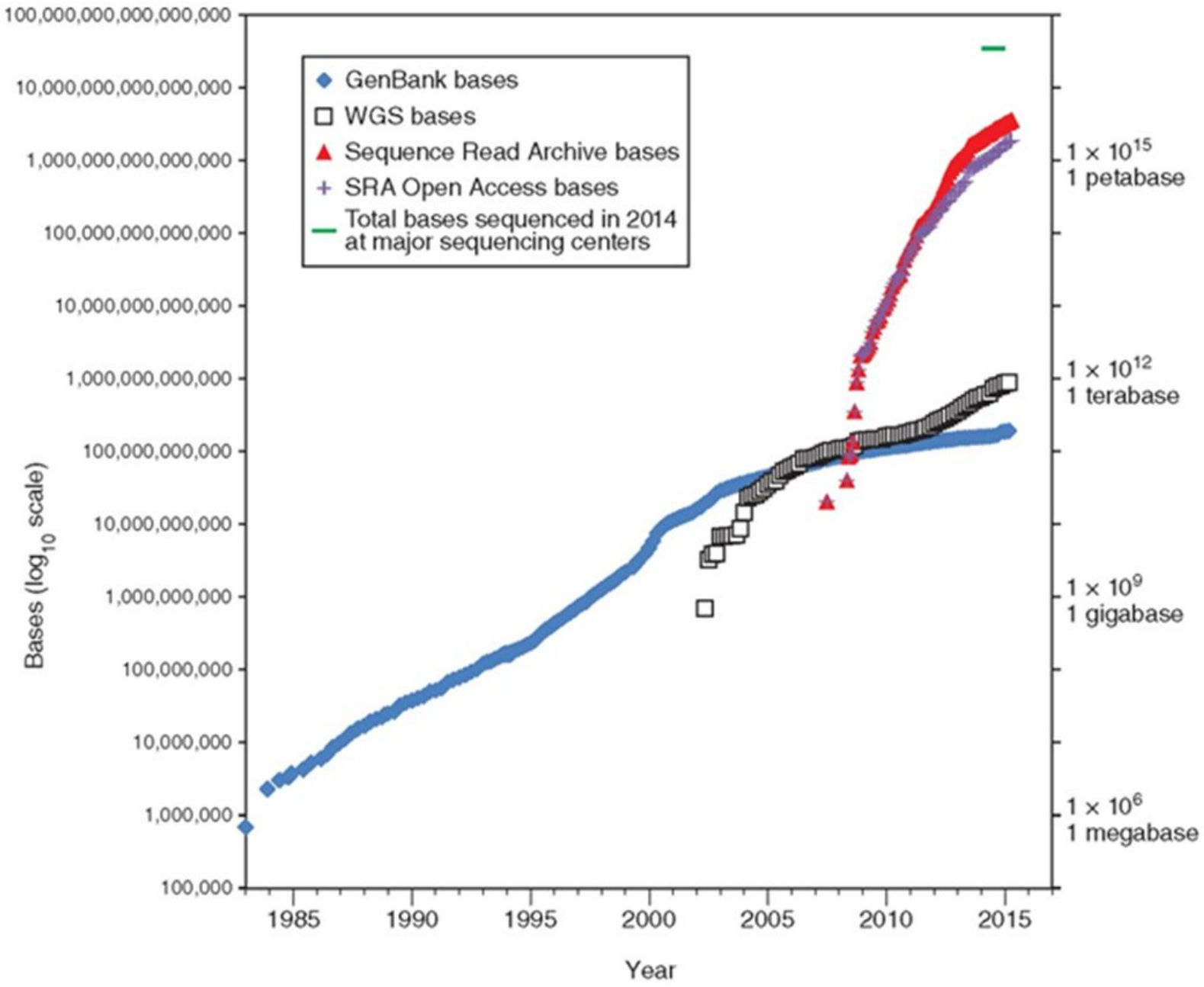
Лекция 8

Многие слайды и материалы используемые в презентации взяты из курса «Введение в биоинформатику» Санкт-Петербурского государственного университета

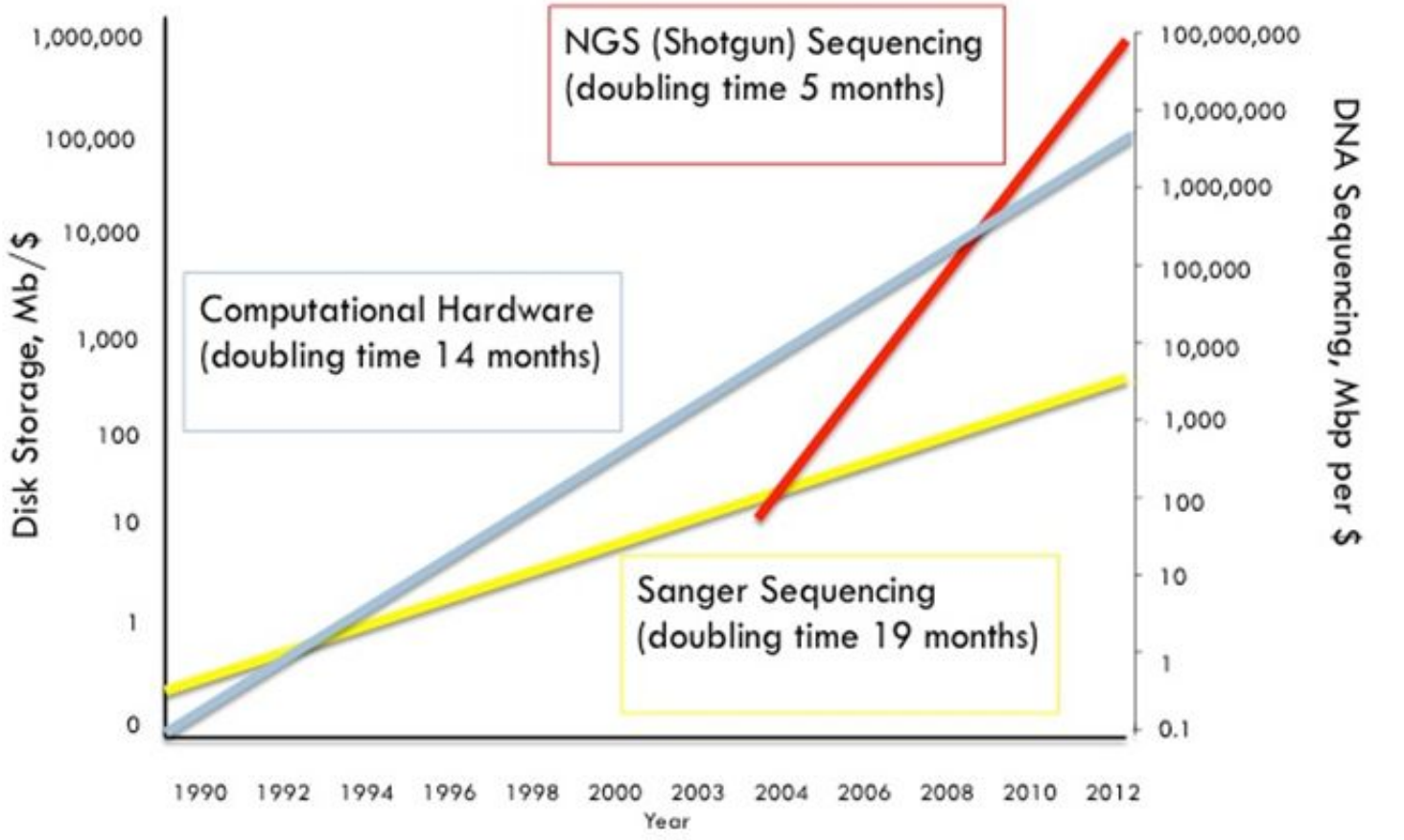
ДНК секвенирование

Подход для определения нуклеотидной последовательности ДНК (дезоксирибонуклеиновой кислоты)




GenBank bases
WGS bases
Sequence Read Archive bases
SRA Open Access bases
Total bases sequenced in 2014 at major sequencing centers
Bases (log10 scale)
100,000,000,000,000,000
10,000,000,000,000,000
1,000,000,000,000,000
100,000,000,000,000
10,000,000,000,000
1,000,000,000,000
100,000,000,000
10,000,000,000
1,000,000,000
100,000,000
10,000,000
1,000,000
100,000
1985
1990
1995
2000
2005
2010
2015
Year
1 × 10^15
1 petabase
1 × 10^12
1 terabase
1 × 10^9
1 gigabase
1 × 10^6
1 megabase

The era of big data in biology



Stein, Genome Biology, 2010

Применение NGS

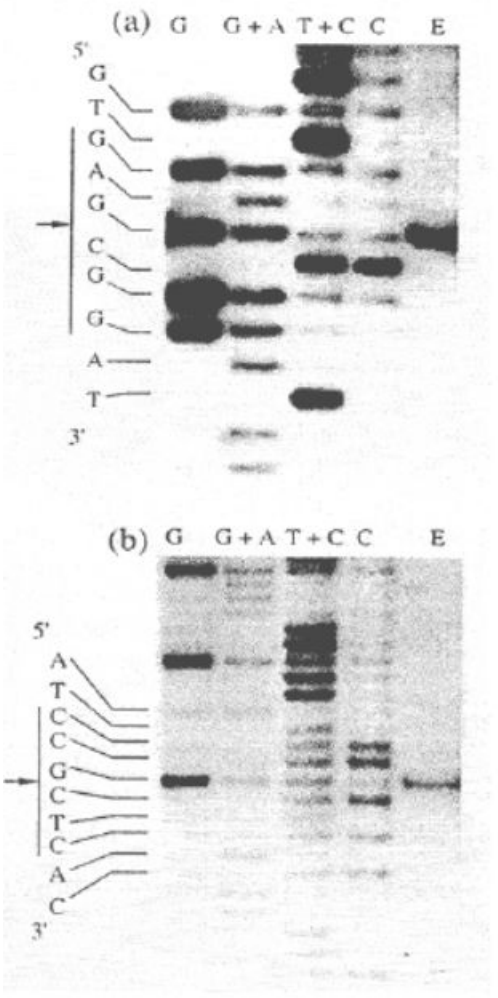
Применение	Решаемая задача
Полногеномное секвенирование *de novo*	Реконструкция работы клетки и организма на молекулярном уровне, эволюционная геномика
Полногеномное повторное секвенирование	Поиск генетических вариаций
Метагеномное секвенирование	Исследование биоценоза, поиск новых видов живых систем
Секвенирование транскриптомов	Исследование генной экспрессии, аннотация генома
Секвенирование малых РНК	Исследование генной экспрессии
Таргетное секвенирование	Поиск генетических вариаций
Секвенирование обработанной бисульфидом ДНК	Исследование профиля ДНК
Секвенирование иммунопреципитированного хроматина (ChIP)	Полногеномное картирование ДНК-белковых взаимодействий
Секвенирование единичных клеток	Исследование генной экспрессии, секвенирование некультивируемых бактерий

Основные термины

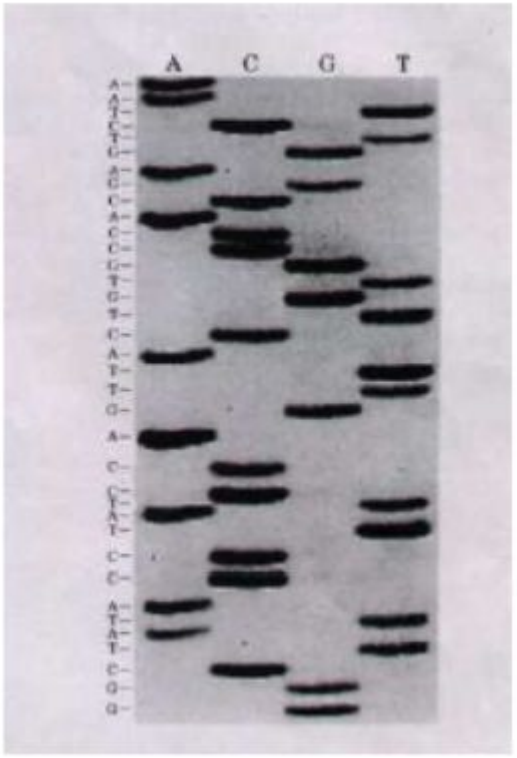
- **Геномные библиотеки** - это коллекция геномной ДНК полученная от одного организма и подготовленная для секвенирования
- **Sequence Read (сиквенсное прочтение, рид)** - нуклеотидная последовательность определённая секвенатором
- **Производительность секвенатора -** набор сиквенсных прочтений, полученных во время одного запуска секвенатора. Выражается в количестве прочитанных нуклеотидов: 1000, 100 тыс., миллионы, миллиарды нуклеотидов
- **Уровень ошибок** – доля неправильных нуклеотидов определенная при секвенировании

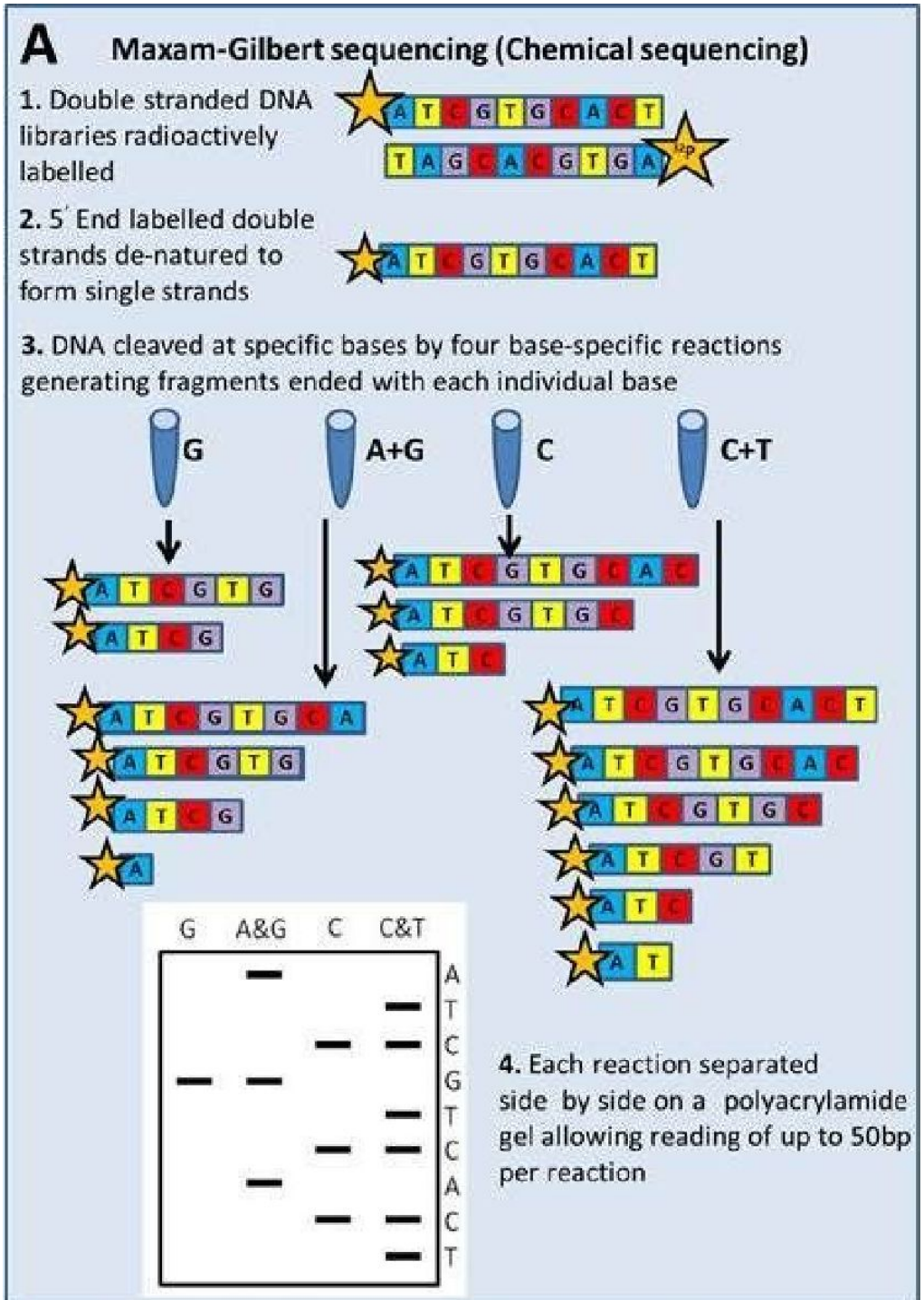
Первые методы секвенирования

- **Maxam-Gilbert (1976-1977)**



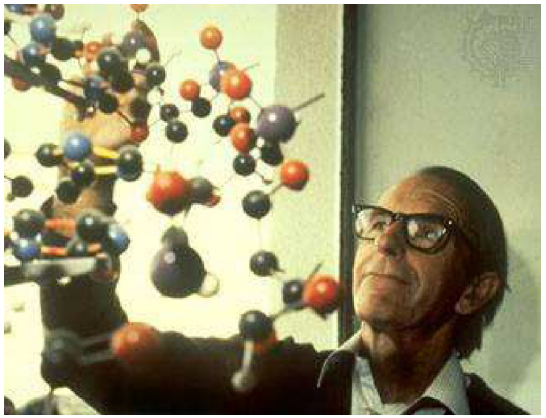
- **Sanger (1977)**





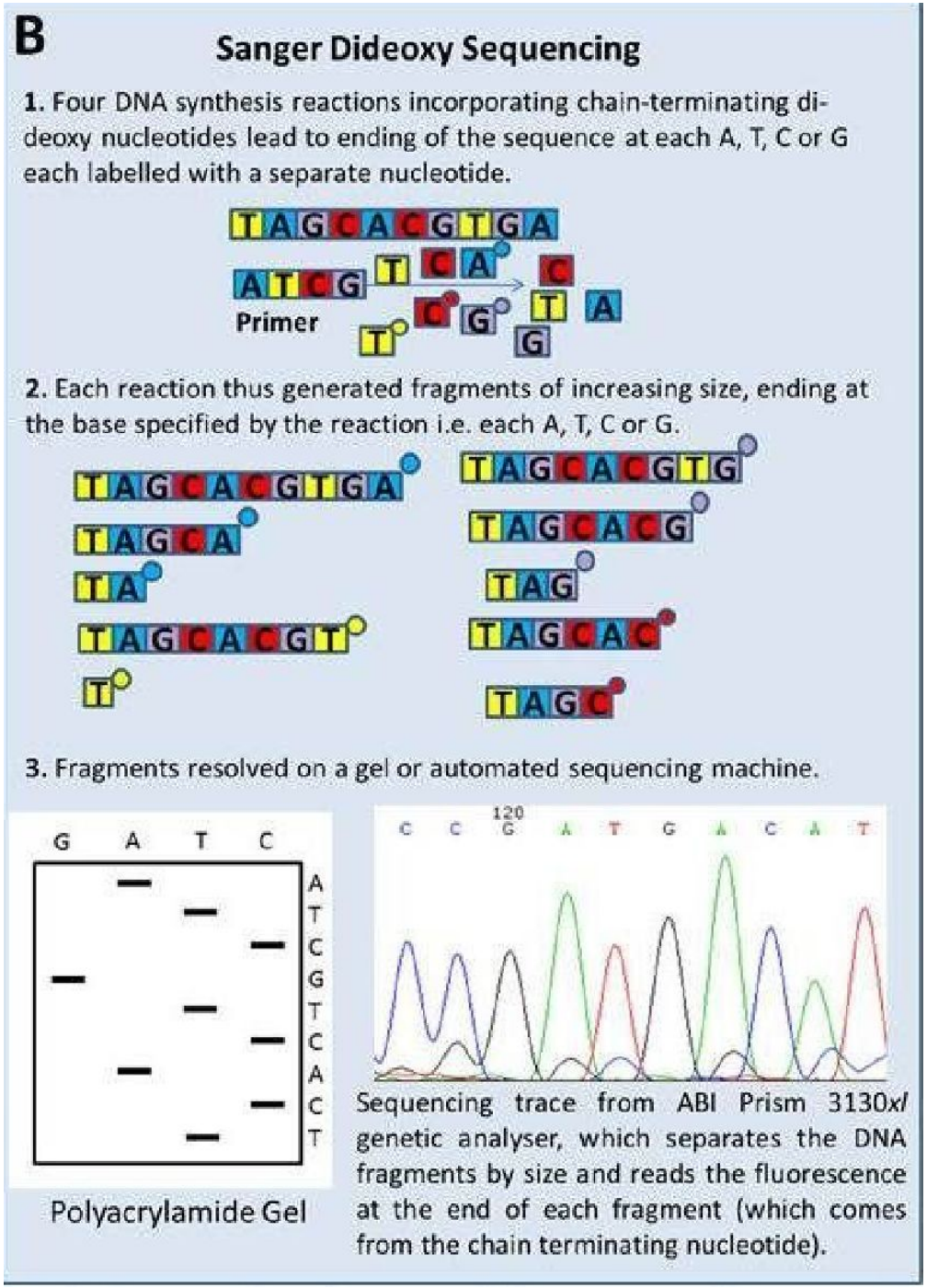
Нуклеотид-специфическая деградация ДНК при обработке различными веществами

Секвенирование по Сенгеру (Золотой стандарт)

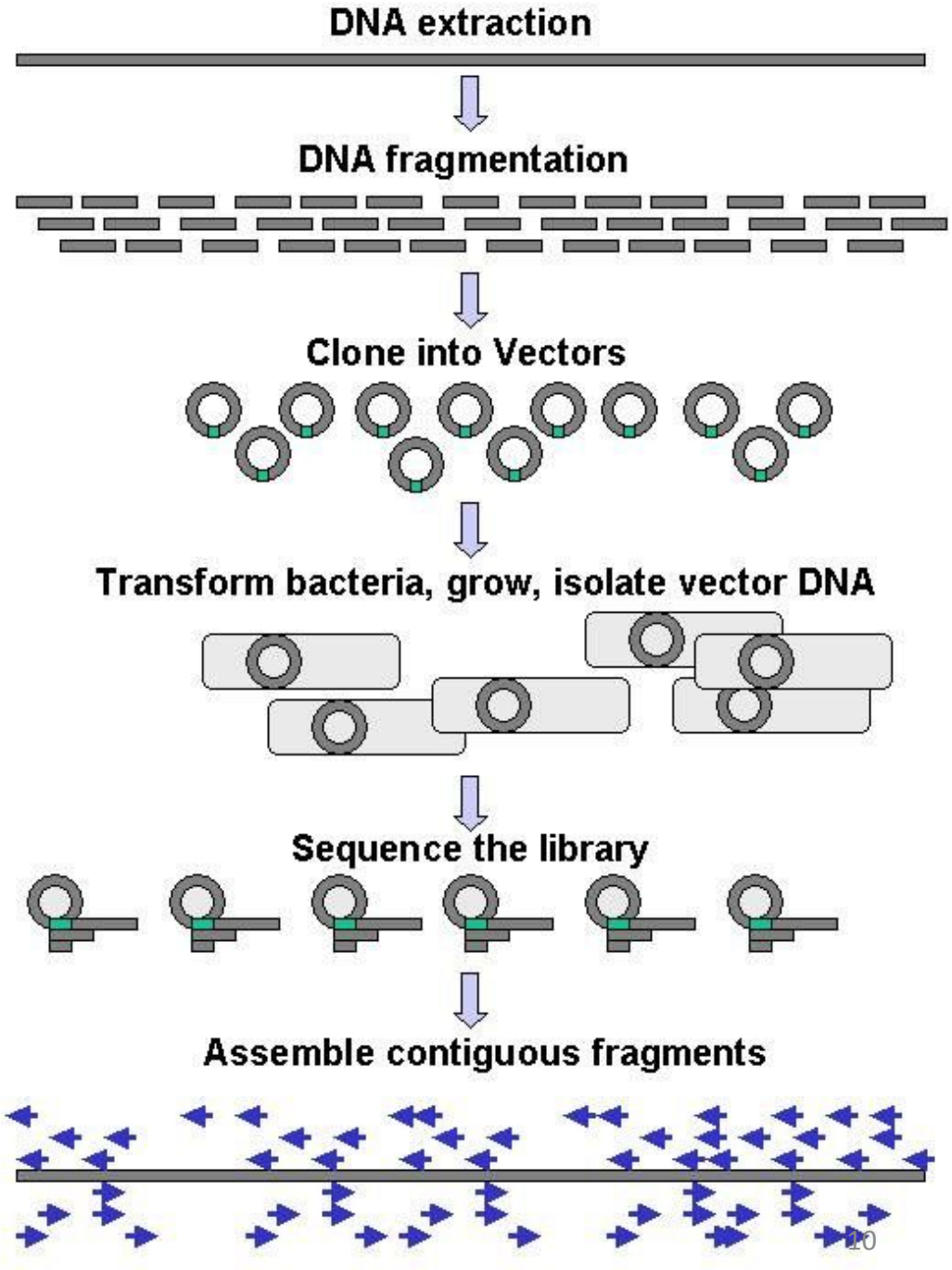


Длинна секвенирования:
300-1000 bp
Ошибки: 0.1-1%

Phi X 174 (ФХ174) бактериофаг был первым секвенированным ДНК геномом (5386 нуклеотидов) в 1977 году

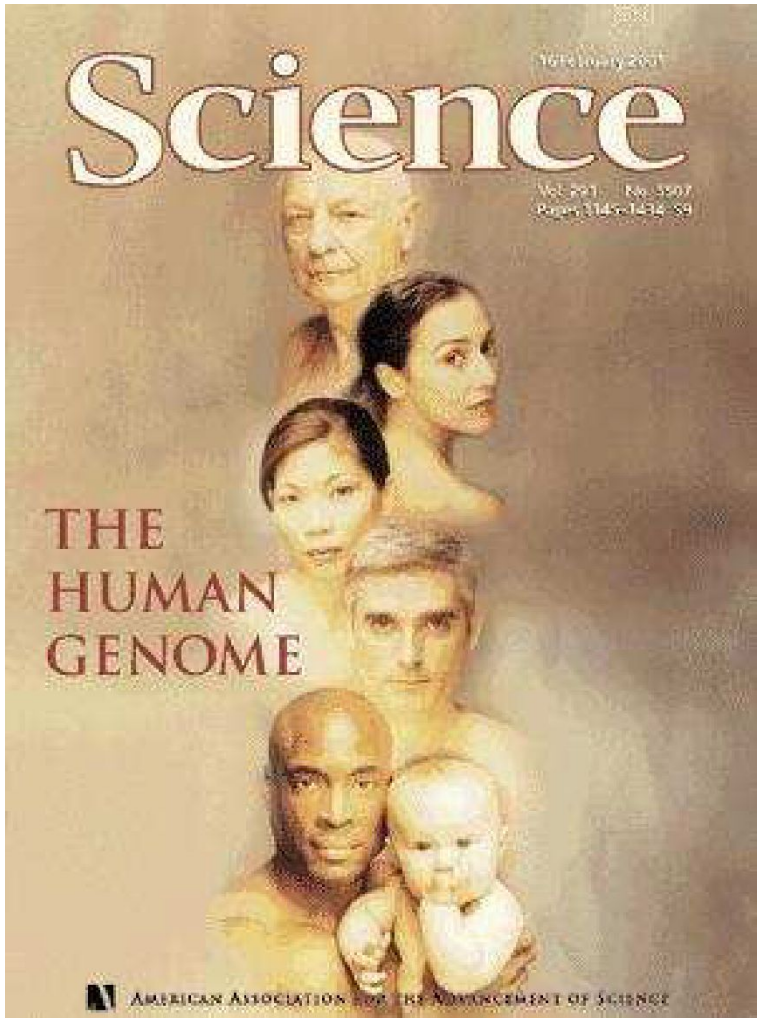


Полногеномное секвенирование с использованием метода Сенгера



Проект геном человека

- Размер генома – 3.2 Гб
- Длительность – 10 лет
- 1990 – 2000
- Цена – 3 млрд. $
- Метод - секвенирование по Сенгеру



Секвенирование по Сенгеру

Плюсы:

- Относительно низкий уровень ошибок
- Удобное и дешевое секвенирование небольших фрагментов генома (16S РНК, Hsp65, и т.д.)

Минусы:

- Высокая стоимость полногеномного секвенирования
- Трудоемкость
- Низкая производительность


Welcome
illumina
HiSeqX™

New Generation Sequencing

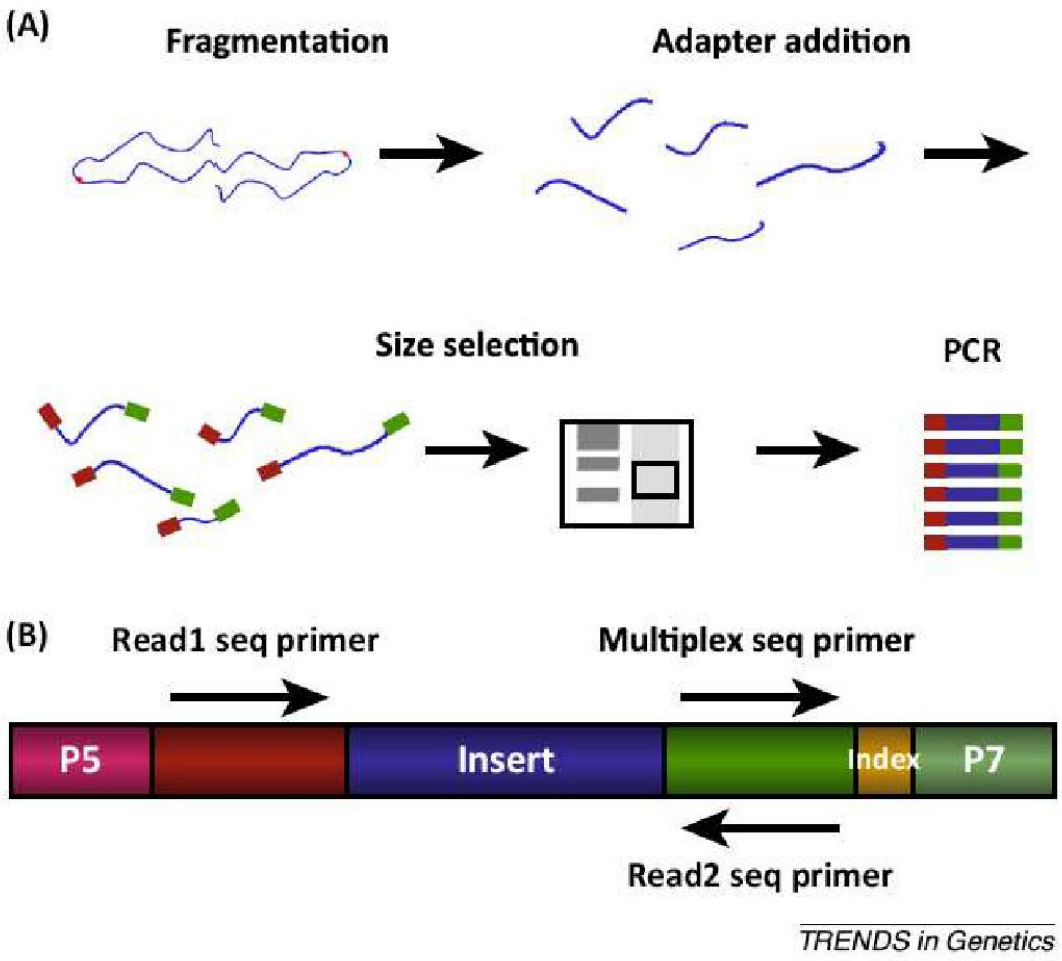
Плюсы:

- Простая подготовка ДНК библиотек (пробоподготовка)
- Высокая производительность
- Низкая стоимость секвенирования

Минусы:

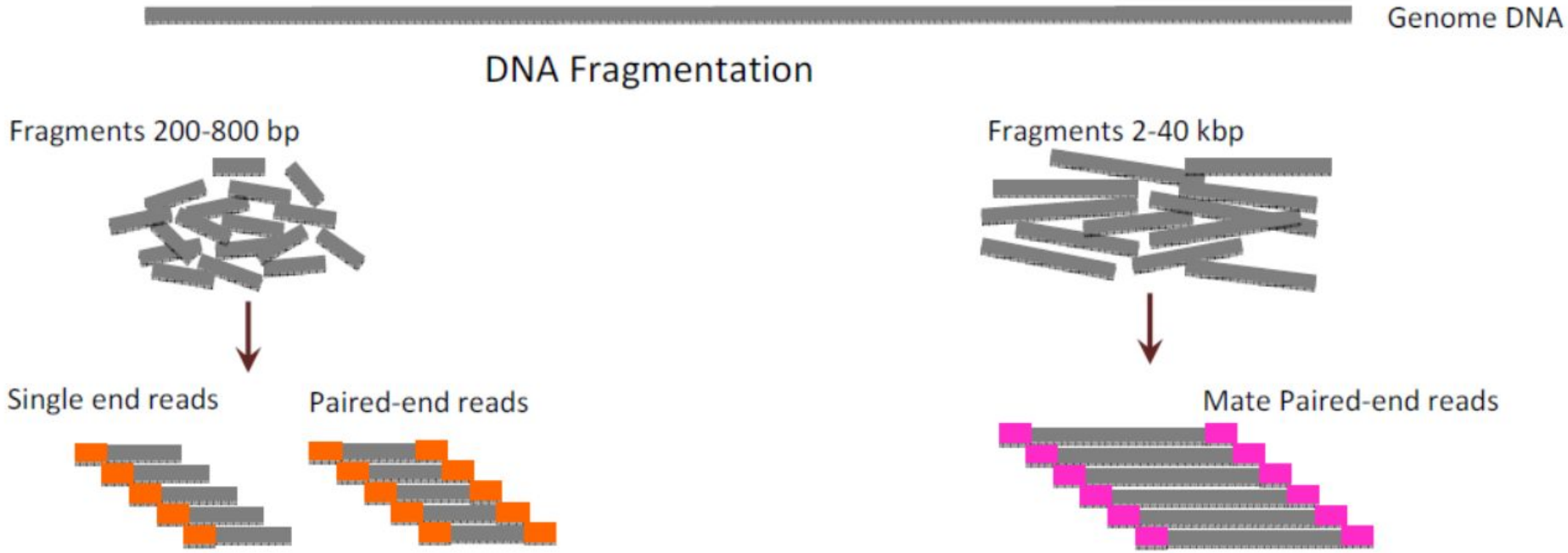
- Короткие риды
- Относительно высокий уровень ошибок

Основные принципы подготовки ДНК библиотек

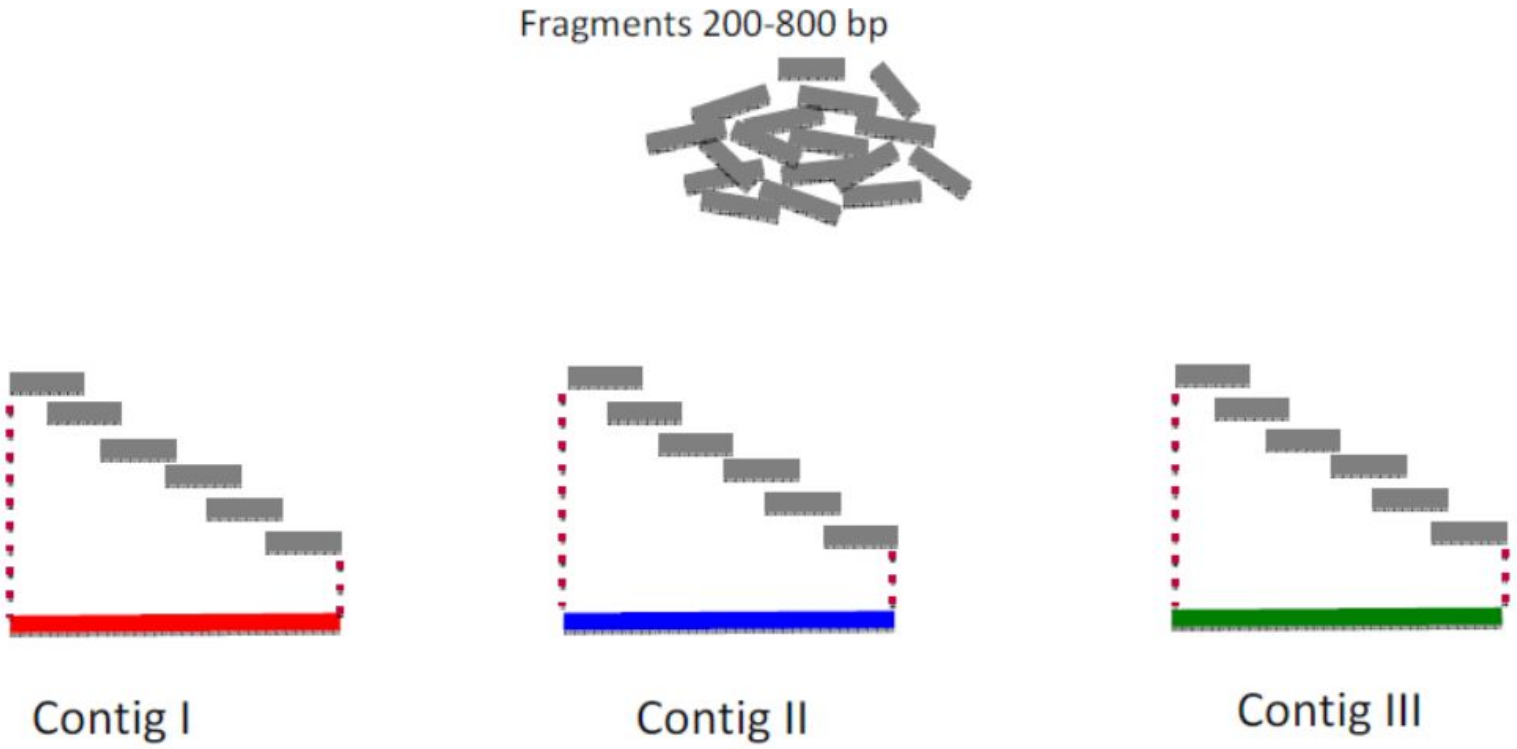


1. Фрагментация ДНК
2. Отбор размера
3. Лигирование адаптора
4. Амплификация библиотеки

Стратегия полногеномного секвенирования использует NGS платформы

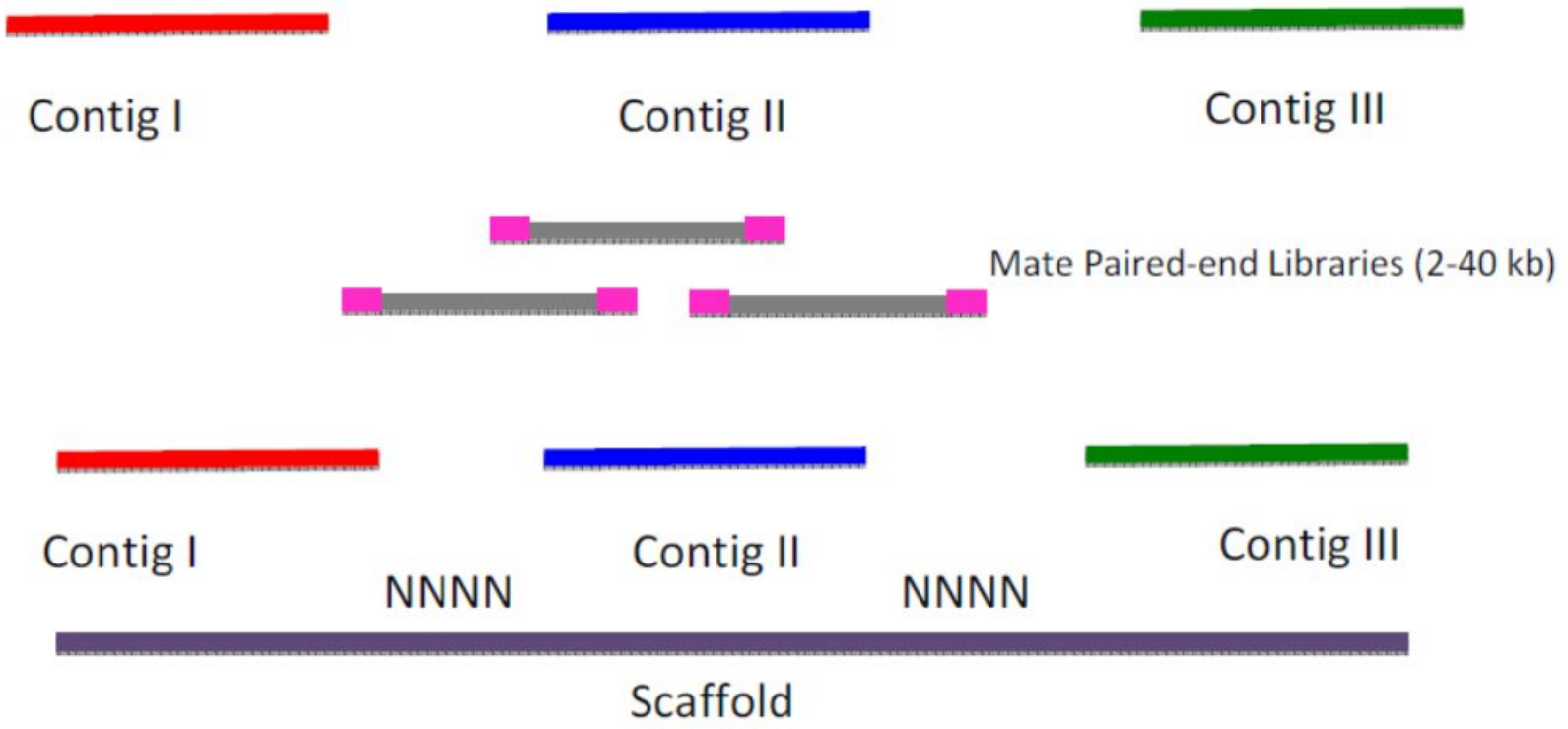


Контиг (Contig) - группа перекрывающихся прочтений, представляющие участок генома.



Contig is a group of overlapping clones representing regions of the genome; the contiguous sequence of DNA created by assembling these overlapping chromosome fragments.
Definition from: NCI Thesaurus via Unified Medical Language System at the National Library of Medicine

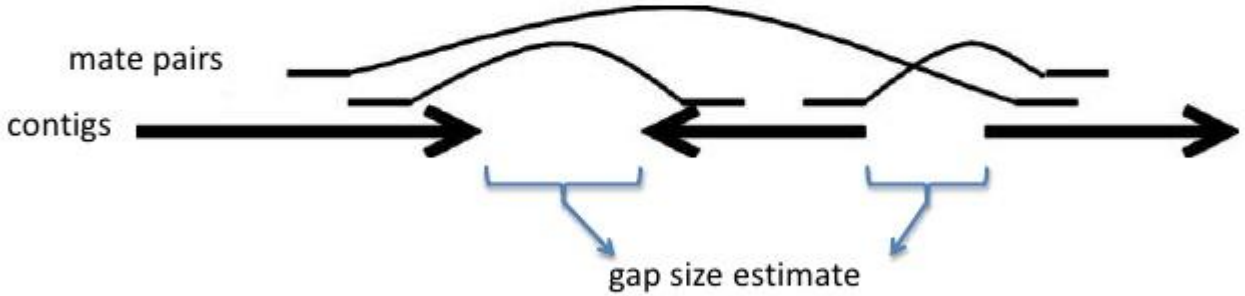
Scaffold (Скафолд) – реконструированная часть генома, полученная в результате анализа библиотек большого размера и правильного взаимного расположения контигов



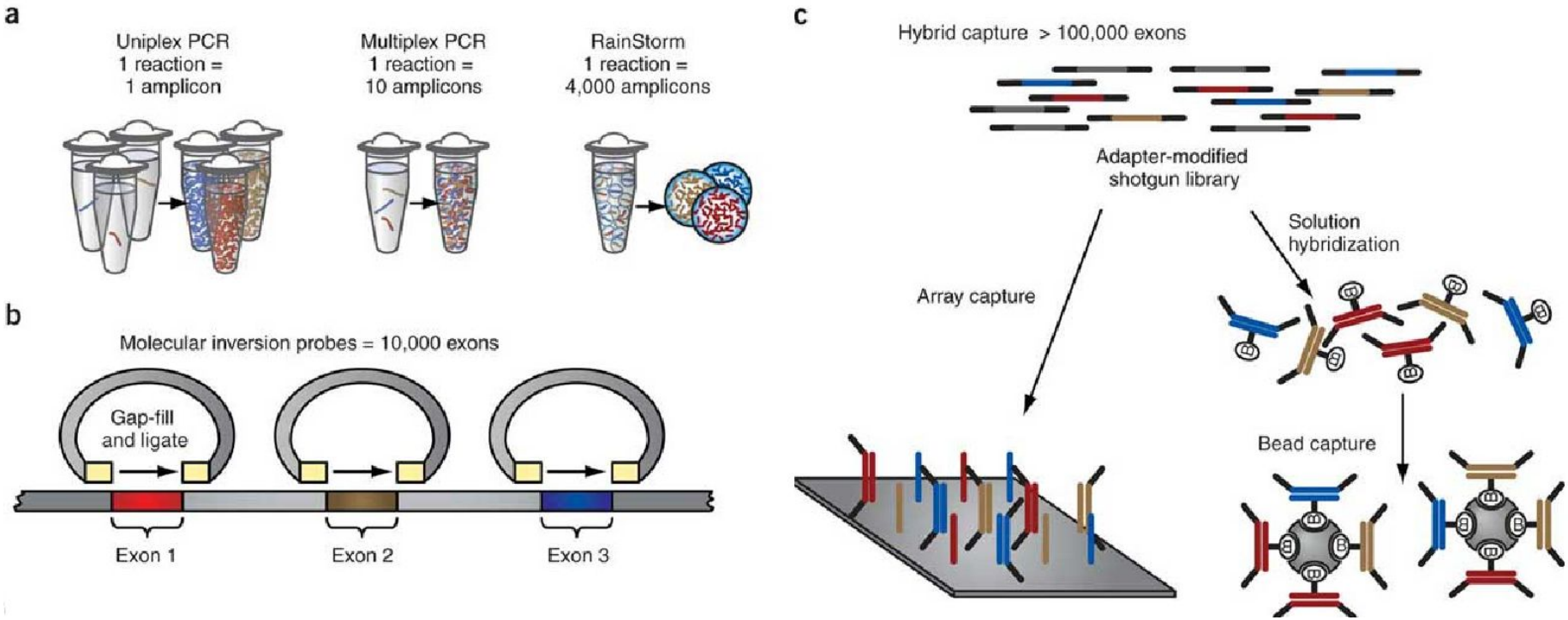
Scaffold is a portion of the genome sequence reconstructed from end-sequenced whole-genome shotgun clones. Scaffolds are composed of contigs and gaps.
Definition from: http://genome.jgi-psf.org/help/scaffolds.html

Scaffolds

Ordered, oriented contigs



Таргетное секвенирование



Nature Methods 7, 111 - 118 (2010)

Индексирование (Баркодинг)

Можно за один запуск секвенатора прочитать несколько геномов или геномных участков

Индексы – короткие олигонуклеотиды с различной последовательностью, которые фланкируют ДНК библиотеки и секвенируются вместе с целевой ДНК. На основе известной индексной последовательности можно дифференцировать несколько образцов секвенированных в одно время.

Примеры индексов

Dual indexing

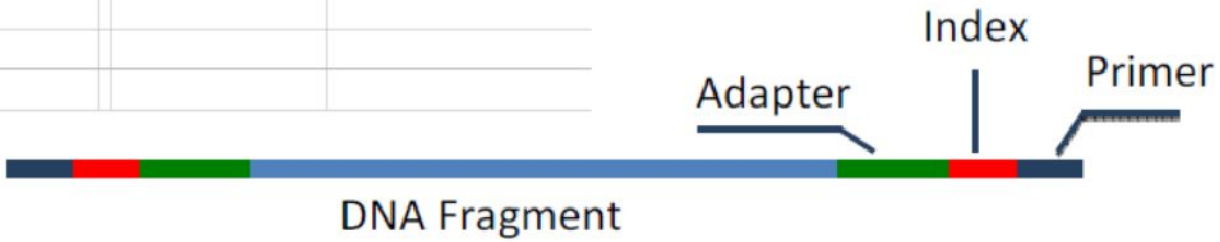
Illumina Nextera

Index 1 (i7)	Sequence	Index 2 (i5)	Sequence
N701	TAAGGCGA	S501	TAGATCGC
N702	CGTACTAG	S502	CTCTCTAT
N703	AGGCAGAA	S503	TATCCTCT
N704	TCCTGAGC	S504	AGAGTAGA
N705	GGACTCCT	S505	GTAAGGAG
N706	TAGGCATG	S506	ACTGCATA
N707	CTCTCTAC	S507	AAGGAGTA
N708	CAGAGAGG	S508	CTAAGCCT
N709	GCTACGCT		
N710	CGAGGCTG		
N711	AAGAGGCA		
N712	GTAGAGGA		

Single indexes

Illumina TruSeq

Adapter	Sequence	Adapter	Sequence
AD002	CGATGT(A)	AD013	AGTCAA(C)
AD004	TGACCA(A)	AD014	AGTTCC(G)
AD005	ACAGTG(A)	AD015	ATGTCA(G)
AD006	GCCAAT(A)	AD016	CCGTCC(C)
AD007	CAGATC(A)	AD018	GTCCGC(A)
AD012	CTTGTA(A)	AD019	GTGAAA(C)



Платформы

- **Roche 454**
  - *pyrosequencing , introduced in 2005*
- **SOLID**
  - *sequencing by ligation, Introduced in 2006*
- **Illumina**
  - *sequencing by synthesis, introduced in 2006*
- **Ion Torrent, PGM/Proton**
  - *ion semiconductor sequencing, released in February 2010*
- **Pacific Biosiences**

  single molecule real-time (SMRT) sequencing, commercially released in early 2011
- **Oxford Nanopore Technologies**
  - *nanopore sequencing, released in 2014*

Roche 454

Pyrosequencing

Reads length up to 1000 b.p.
Output ~ 700 Mb
Run time – 23 h

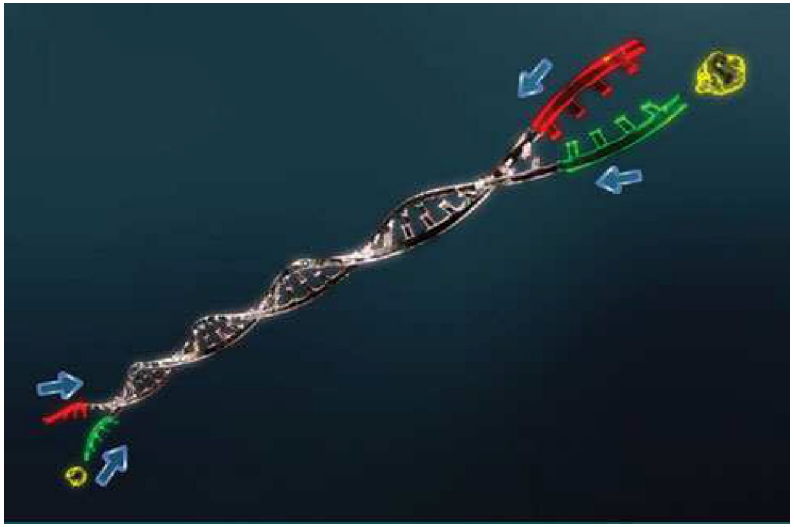
GS FLX+ System

Reads length ~ 450 b.p.
Output ~ 35 Mb
Run time – 10 h

GS Junior System

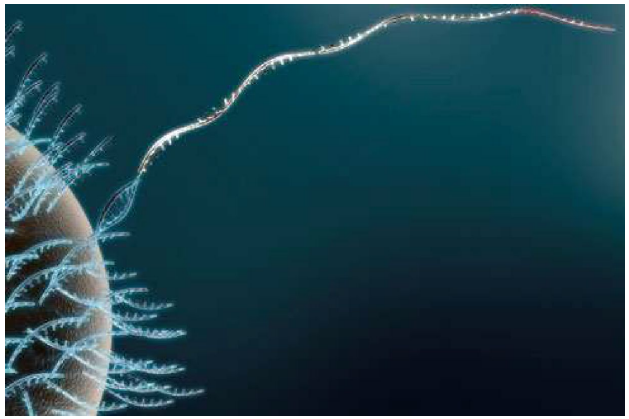
454 Sequencing Technology

Фрагментация ДНК

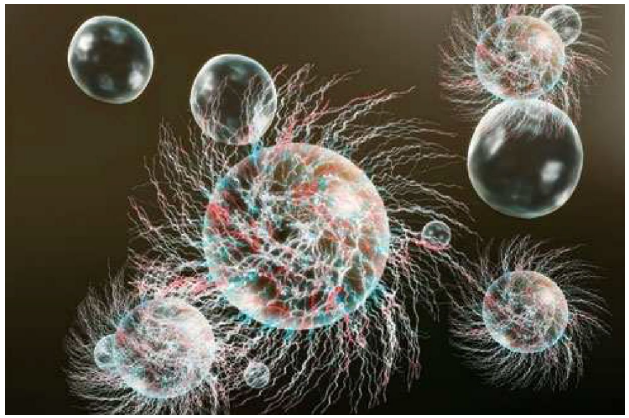


Подготовка библиотеки

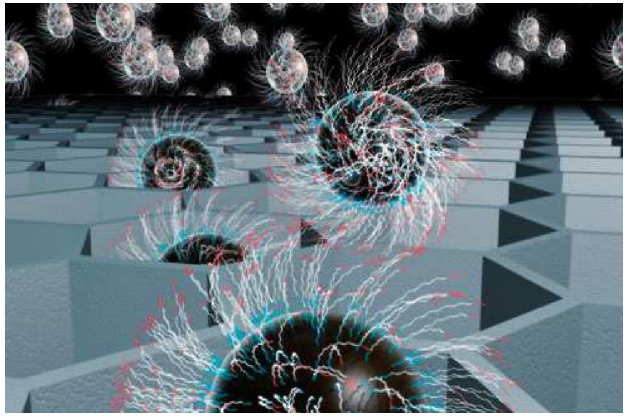
Пришивание адапторов к молекулам ДНК с двух концов.



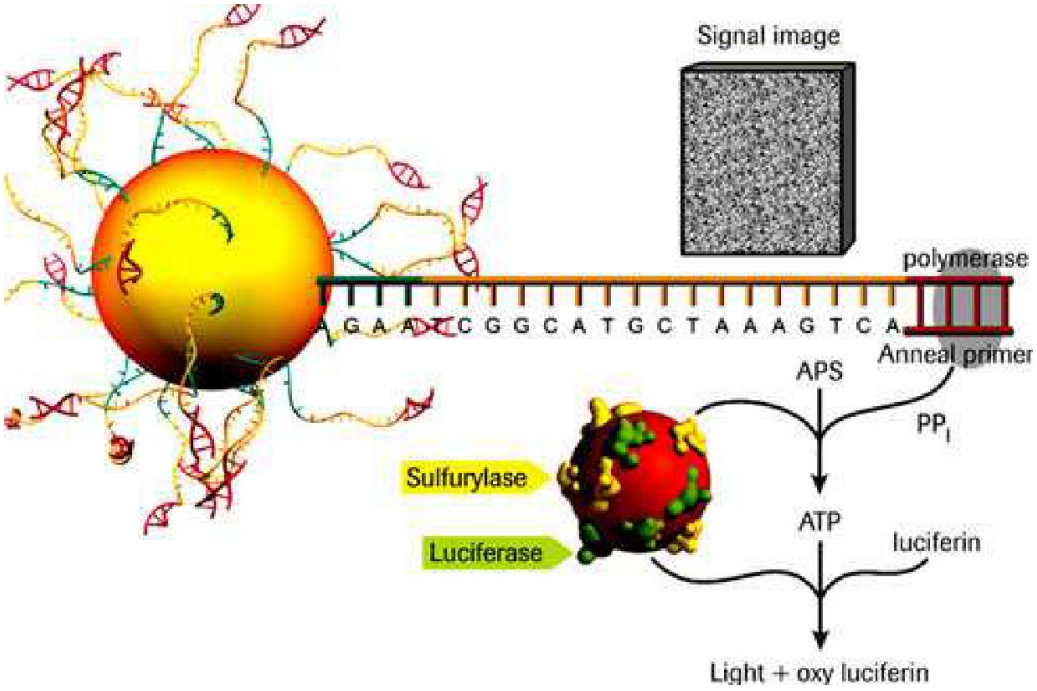
Один фрагмент = одна бусина (bead)
Библиотека фрагментов ДНК прикрепляется к бусинам после денатурации ДНК. Каждая бусина имеет уникальный фрагмент библиотеки. Шарики эмульгируют с реагентами амплификации в смеси вода-в-масле.



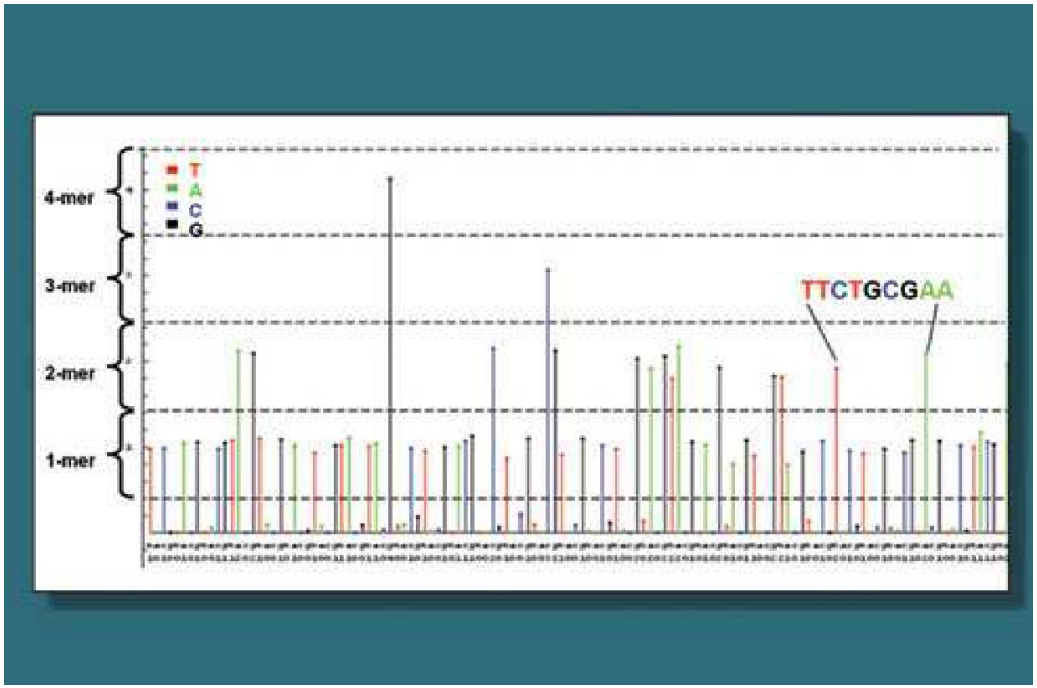
emPCR: Эмульсионная ПЦР-амплификация
Амплификация ведется в эмульсии параллельно, чтобы создать миллионы клонных копий каждого фрагмента библиотеки на каждом шарике. Если фрагмент не присоединяется к шарику, то он остается гладким.



Секвенирование: один шарик = один рид
Шарик помещается в лунку, где дизайн поверхности допускает только один шарик. Отдельные нуклеотиды протекают последовательно через лунки. Каждое включение нуклеотида, комплементарное к матричной нити, приводит к хемилюминесцентному световому сигналу, записанному камерой.



Секвенирование начинается с присоединения праймера, потом присоединение комплементарного нуклеотида приводит к высвобождению пирофосфата, который взаимодействуя с сулфирилазой и люциферазой приводит к образованию светового сигнала, детектируемого камерой.



По интенсивности сигнала определяется какое количество нуклеотидов присоединяется. При этом зная какие нуклеотиды подаются в текущее время определяют последовательность ДНК.

Ion torrent

Ion semiconductor sequencing

Ion PGM – Personal Genome Machine

Reads length – 200 and 400 bp
Output up to 2 Gb
Run time – 2-7 h

Ion Proton

Reads length – 200 bp
Output up to 10 Gb
Run time – 2-4 h

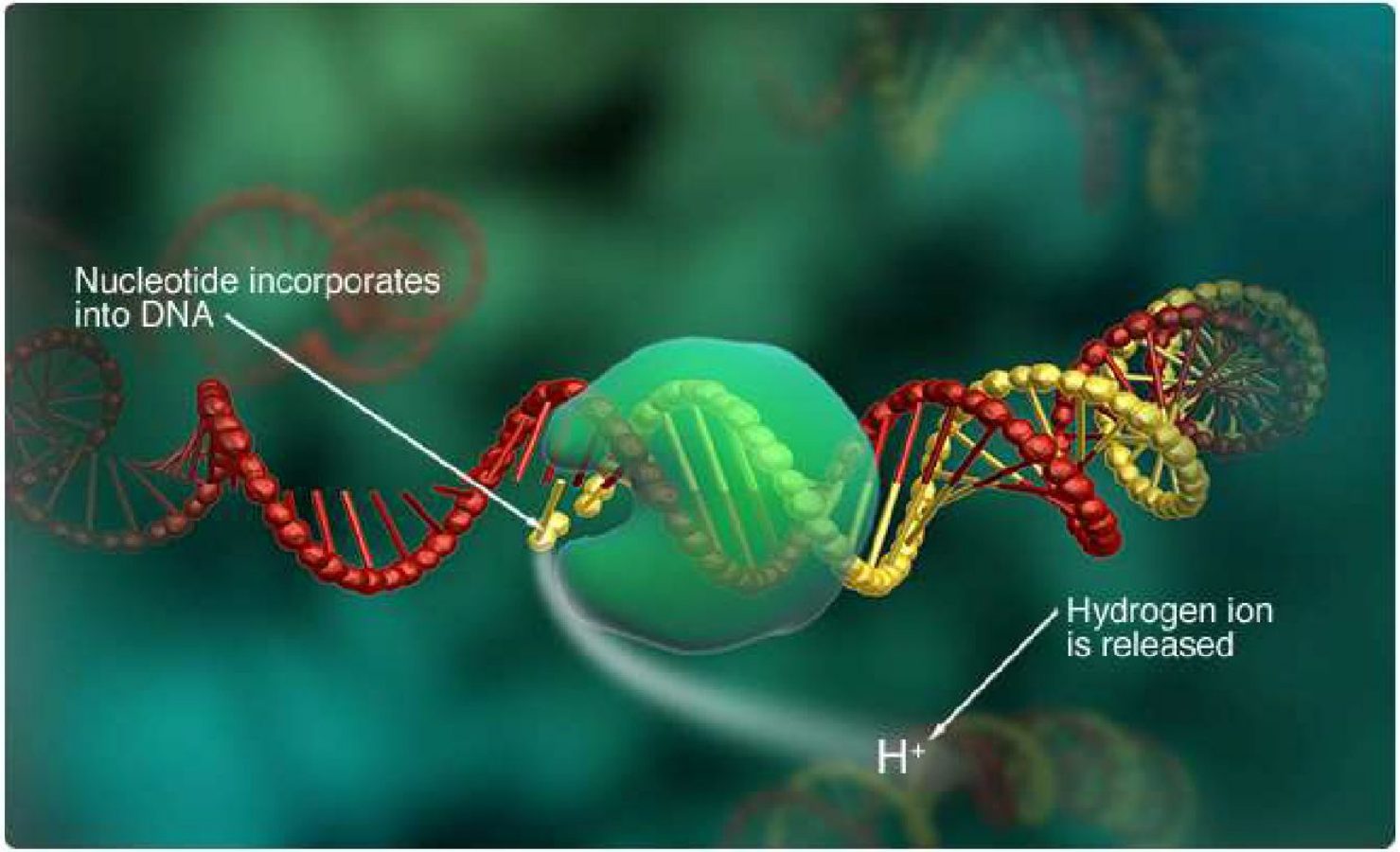
Ion Torrent

Подготовка библиотеки похожа на Roche 454

- фрагментация ДНК
- Прикрепление адаптера
- Эмульсионная ПЦР

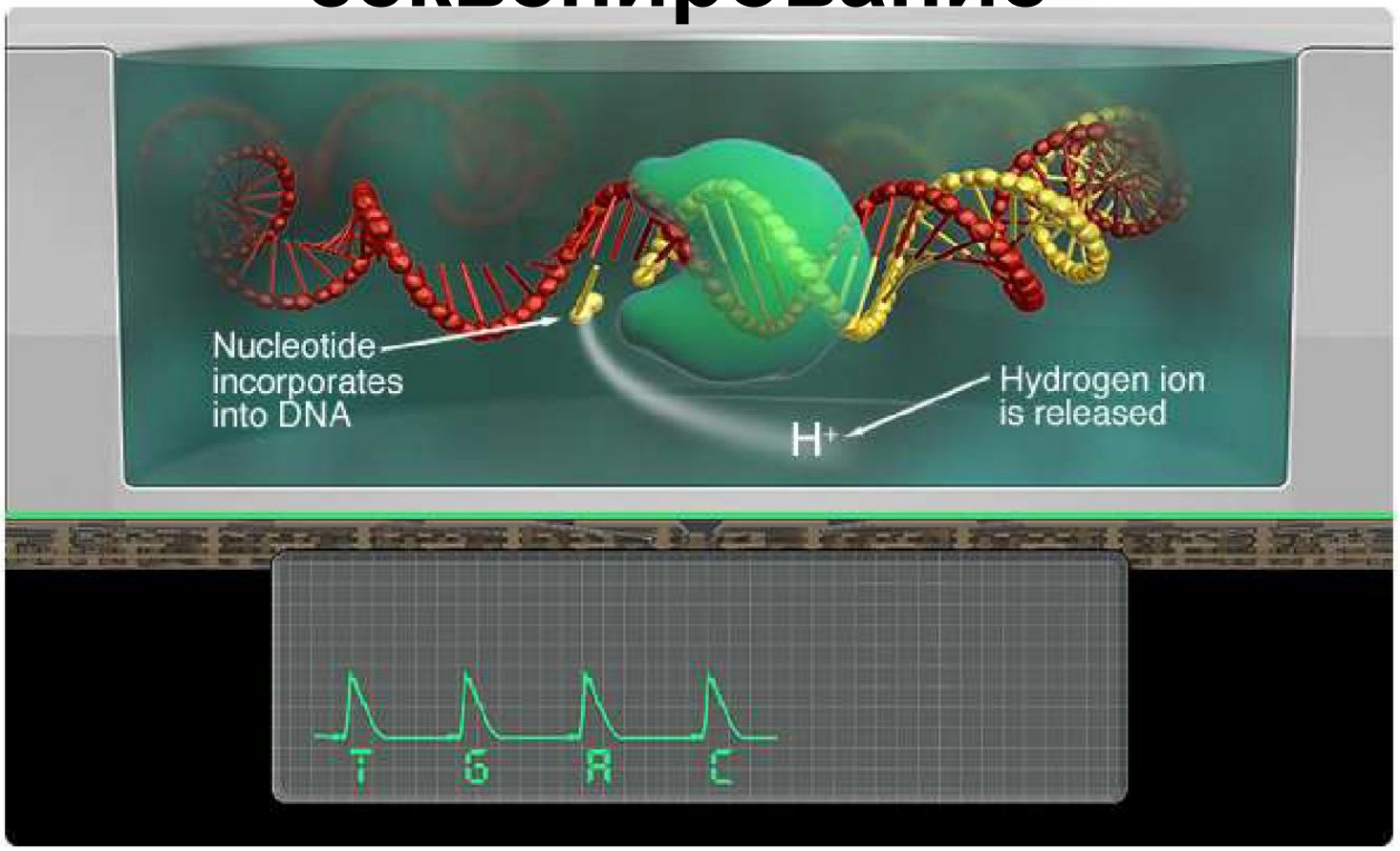
Технология секвенирования отличается

Ion Torrent полупроводниковое секвенирование



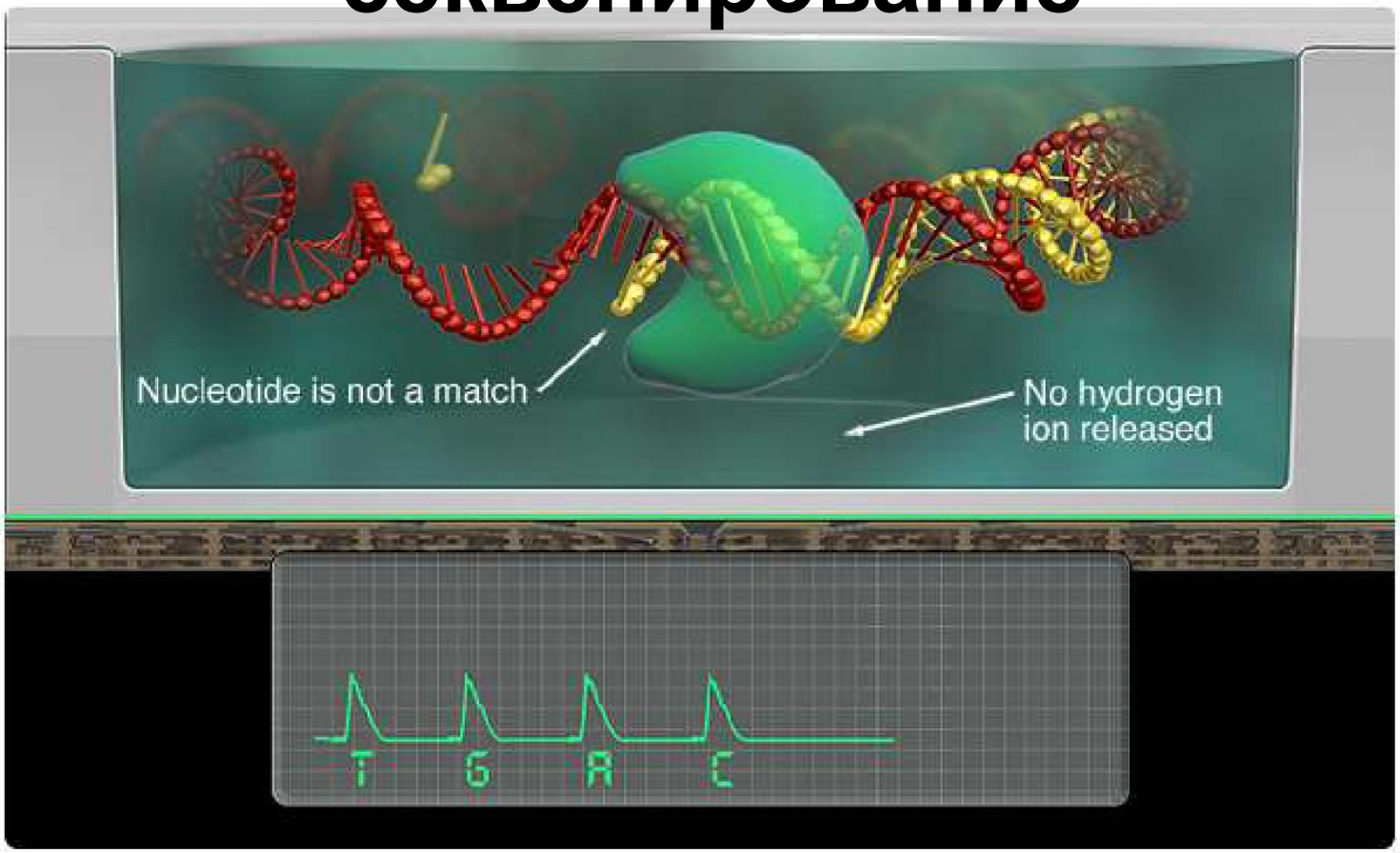
Во время секвенирования, последовательно подаются нуклеотиды, при встраивании которых выделяются ионы водорода.

Ion Torrent полупроводниковое секвенирование

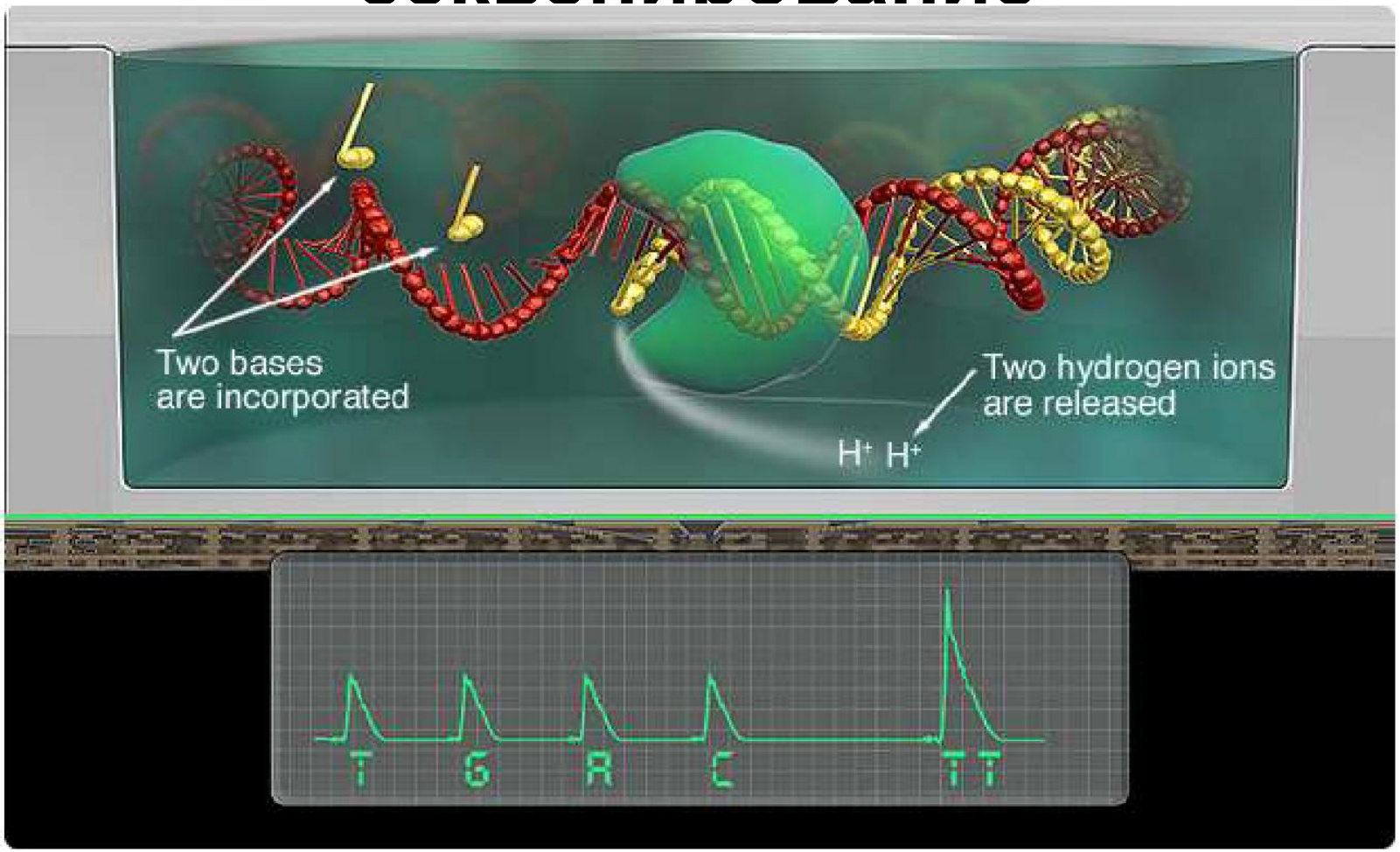


Выделение ионов водорода приводит к изменению кислотности среды, что детектируются высокочувствительным pH-метром

Ion Torrent полупроводниковое секвенирование

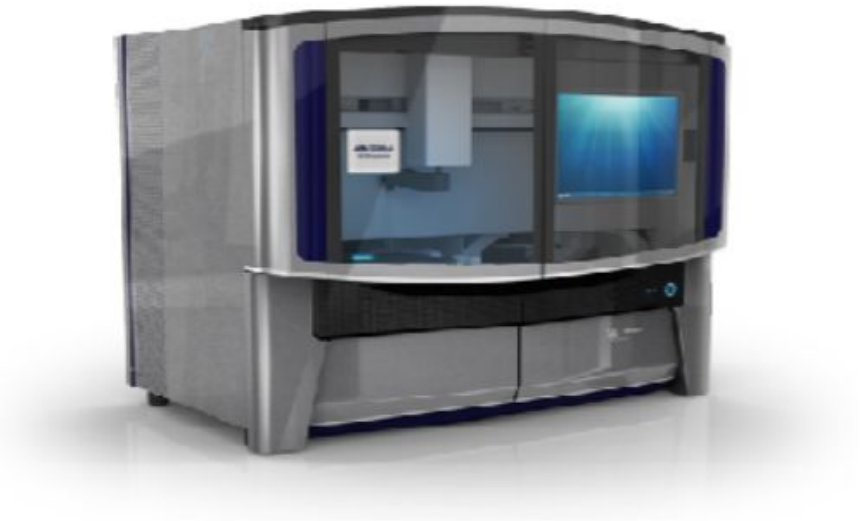


Ion Torrent полупроводниковое секвенирование



SOLiD

Sequencing by Oligonucleotide Ligation and Detection



Commercially available since 2006

Reads length ~ 50 b.p.

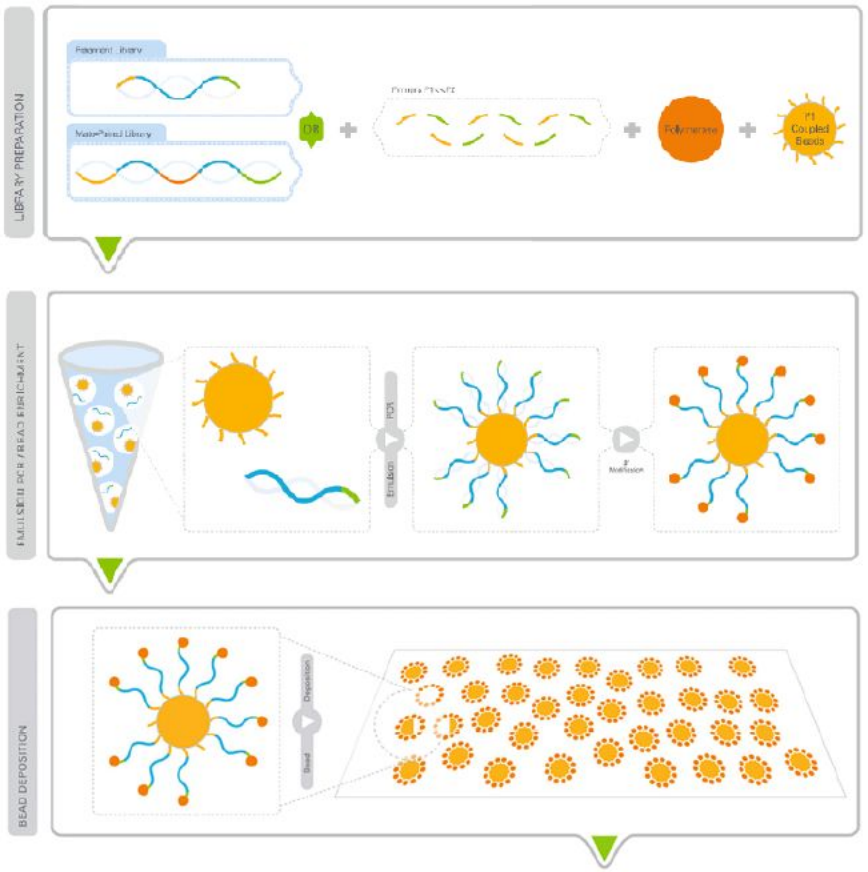
Output ~ 320 Gb

SOLiD

Подготовка библиотеки похожа на Roche 454

- фрагментация ДНК
- Прикрепление адаптера
- Эмульсионная ПЦР
- Технология секвенирования отличается - секвенирование путем лигирования олигонуклеотидов

SOLiD

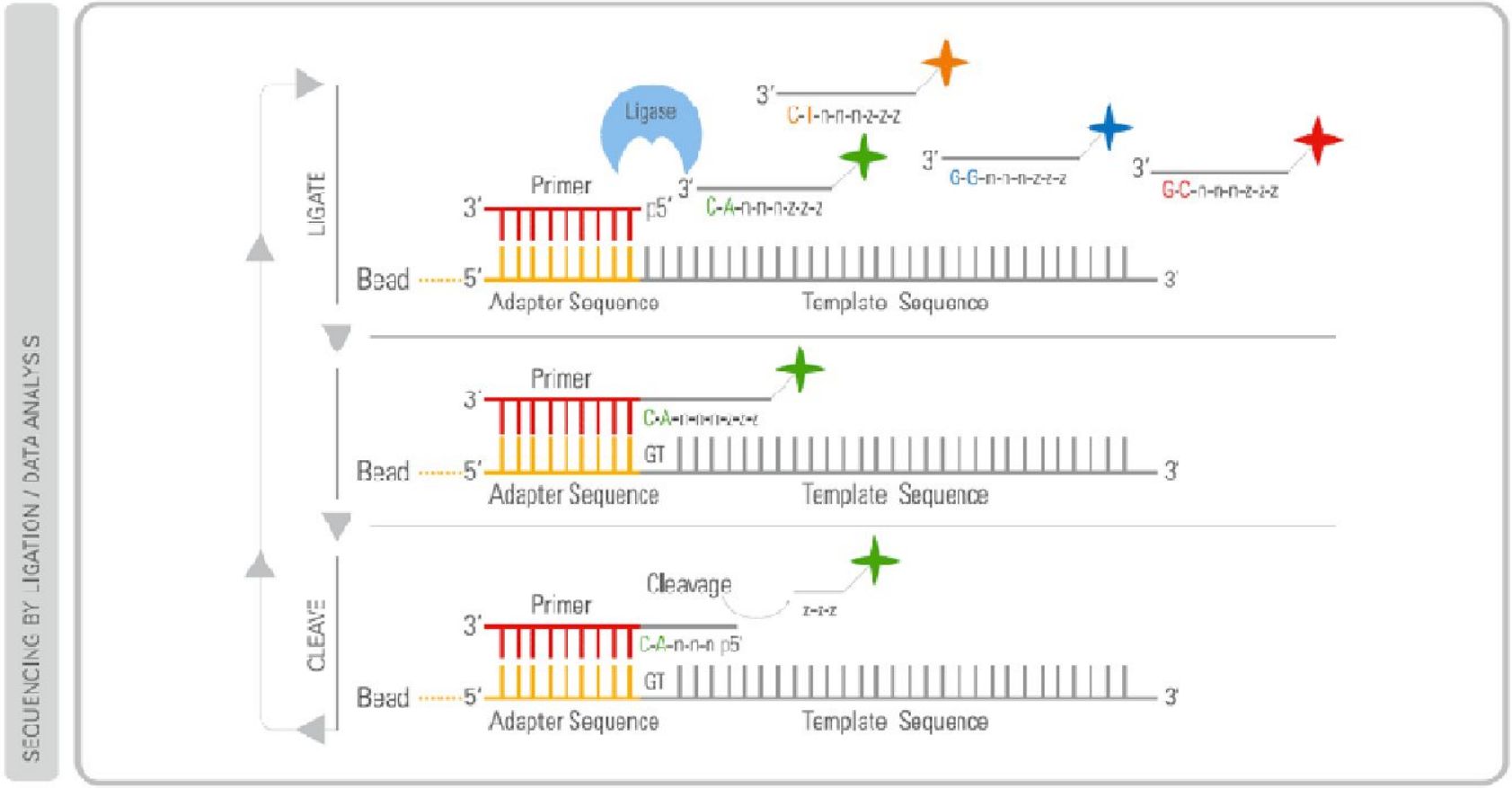


Library Preparation:
-DNA Fragmentation
-Adapters (P1, P2) ligation

Emulsion PCR/Bead Enrichment

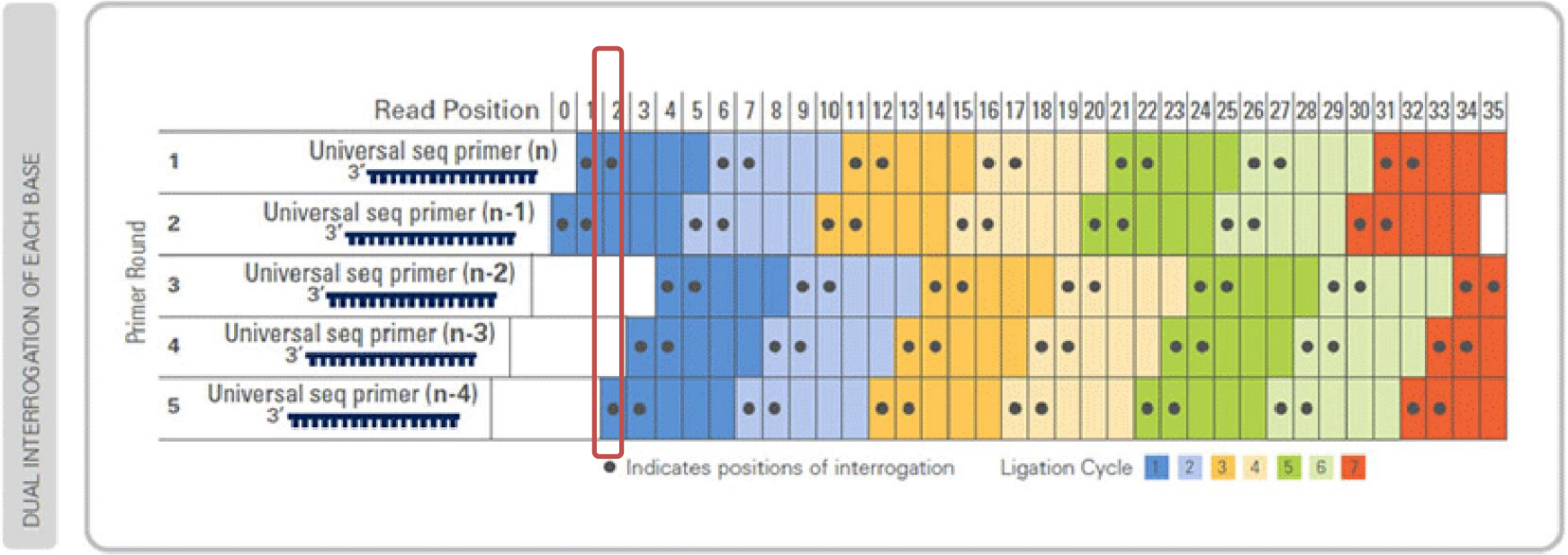
Bead Deposition

SOLiD



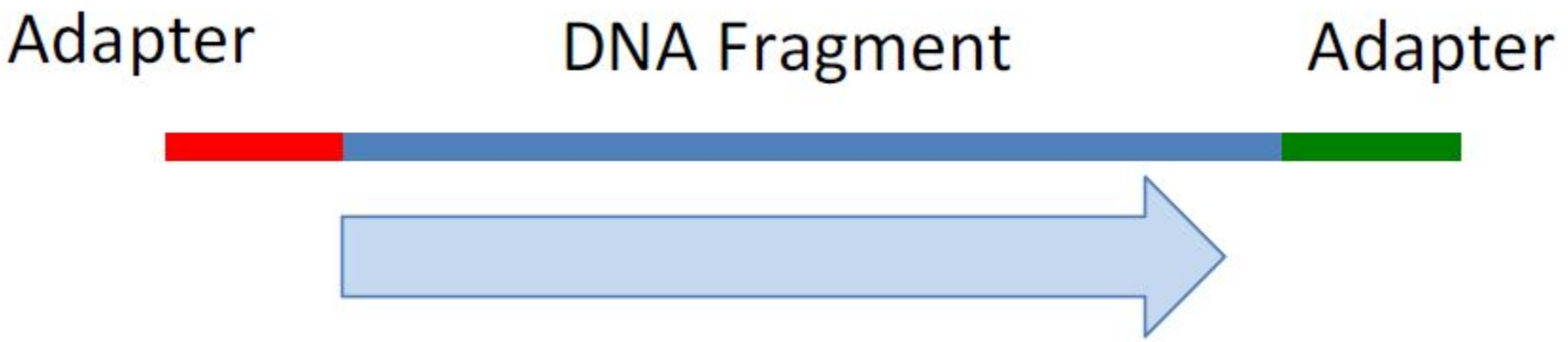
Происходит последовательное взаимодействие олигонуклеотида, состоящего из специфичного динуклеотида, пяти неспецифичных нуклеотидов и флуорафора, что приводит к специфическому связыванию динуклеотида (лигирование) и отщеплению флуорафора и детекция флуоресцентного сигнала.

SOLiD



Для борьбы с неспецифичными нуклеотидами используют новые праймеры, которые короче на 1,2,3,4 нуклеотида (всего 5 раундов секвенирования). Это увеличивает точность секвенирования, т.к. каждый нуклеотид прочитывается дважды, но длина ридов небольшая.

Все описанные технологии обеспечивают односторонние прочтения ДНК



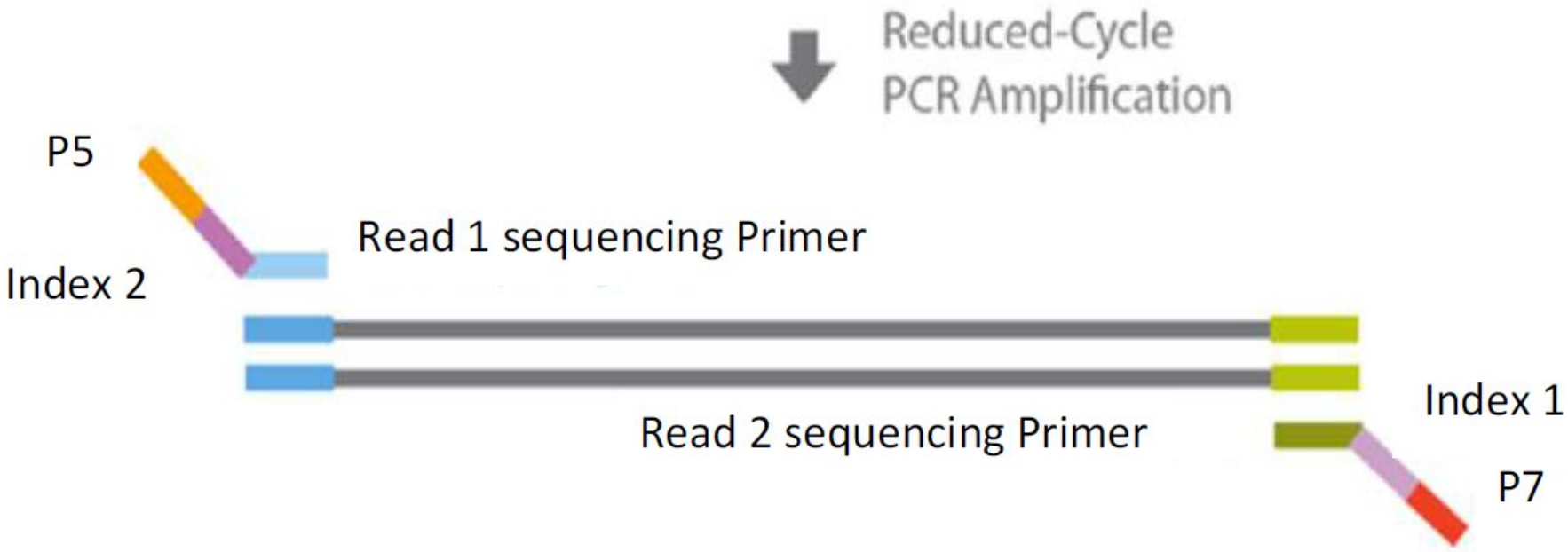
Solexa/Illumina

MiSeq
Up to 15 Gb of output with 25 M sequencing reads and 2x300 bp read lengths

NextSeq500
Up to 120 Gb of output with 400 M sequencing reads and 2x150 bp read lengths

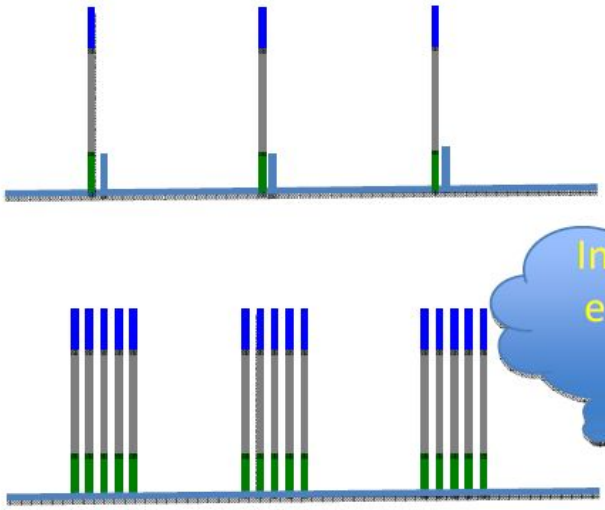
HiSeq 2500
Up to 1000 Gb of output with 4000 M sequencing reads and 2x125 bp read lengths

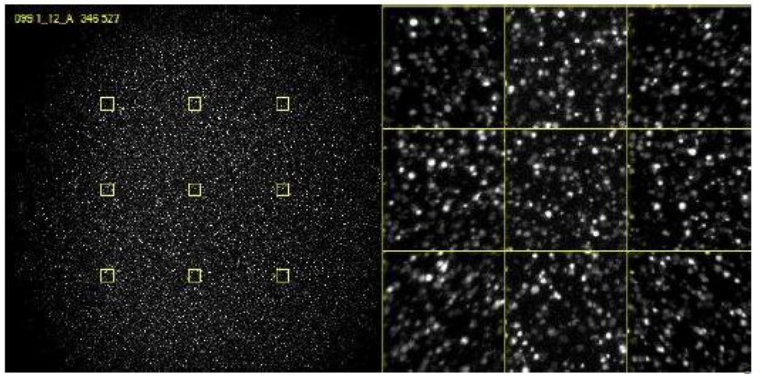
Подготовка библиотеки ДНК



Illumina

- **Гибридизация ДНК-библио**
- **Генерация кластеров (ПЦР)**
- **Секвенирование синтезом**

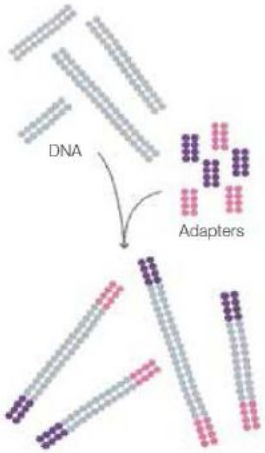




http://www.youtube.com/watch?v=HMyCqWhwB8E

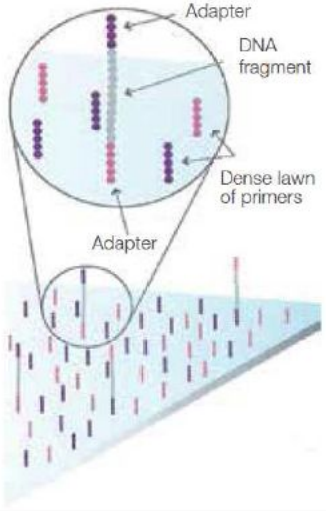
Illumina

a



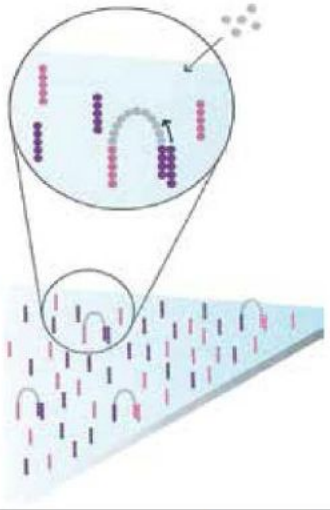
Randomly fragment genomic DNA and ligate adapters to both ends of the fragments.

b



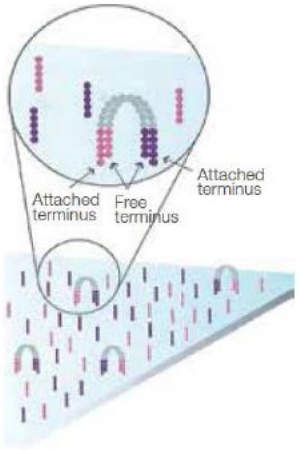
Bind single-stranded fragments randomly to the inside surface of the flow cell channels.

c

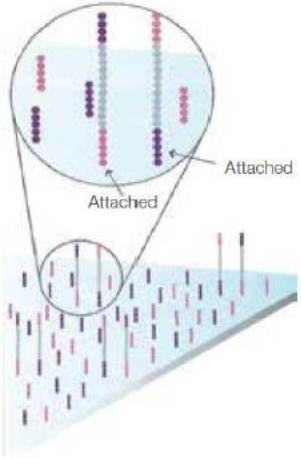


Add unlabeled nucleotides and enzyme to initiate solid-phase bridge amplification.

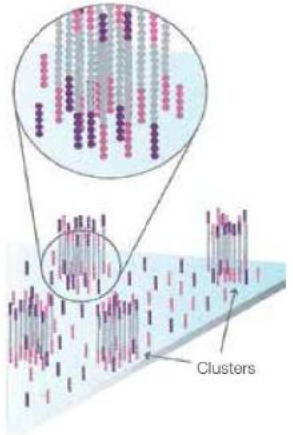
d



e

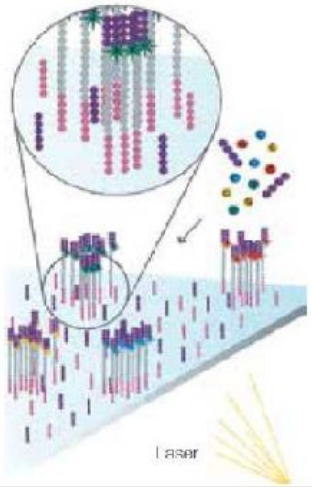


f



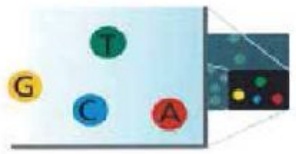
Illumina

g



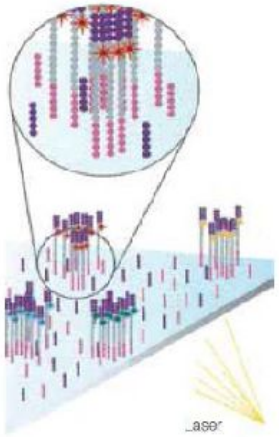
The first sequencing cycle begins by adding four labeled reversible terminators, primers, and DNA polymerase.

h



After laser excitation, the emitted fluorescence from each cluster is captured and the first base is identified.

i

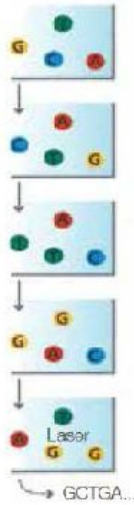


The next cycle repeats the incorporation of four labeled reversible terminators, primers, and DNA polymerase.

j

After laser excitation, the image is captured as before, and the identity of the second base is recorded.

k

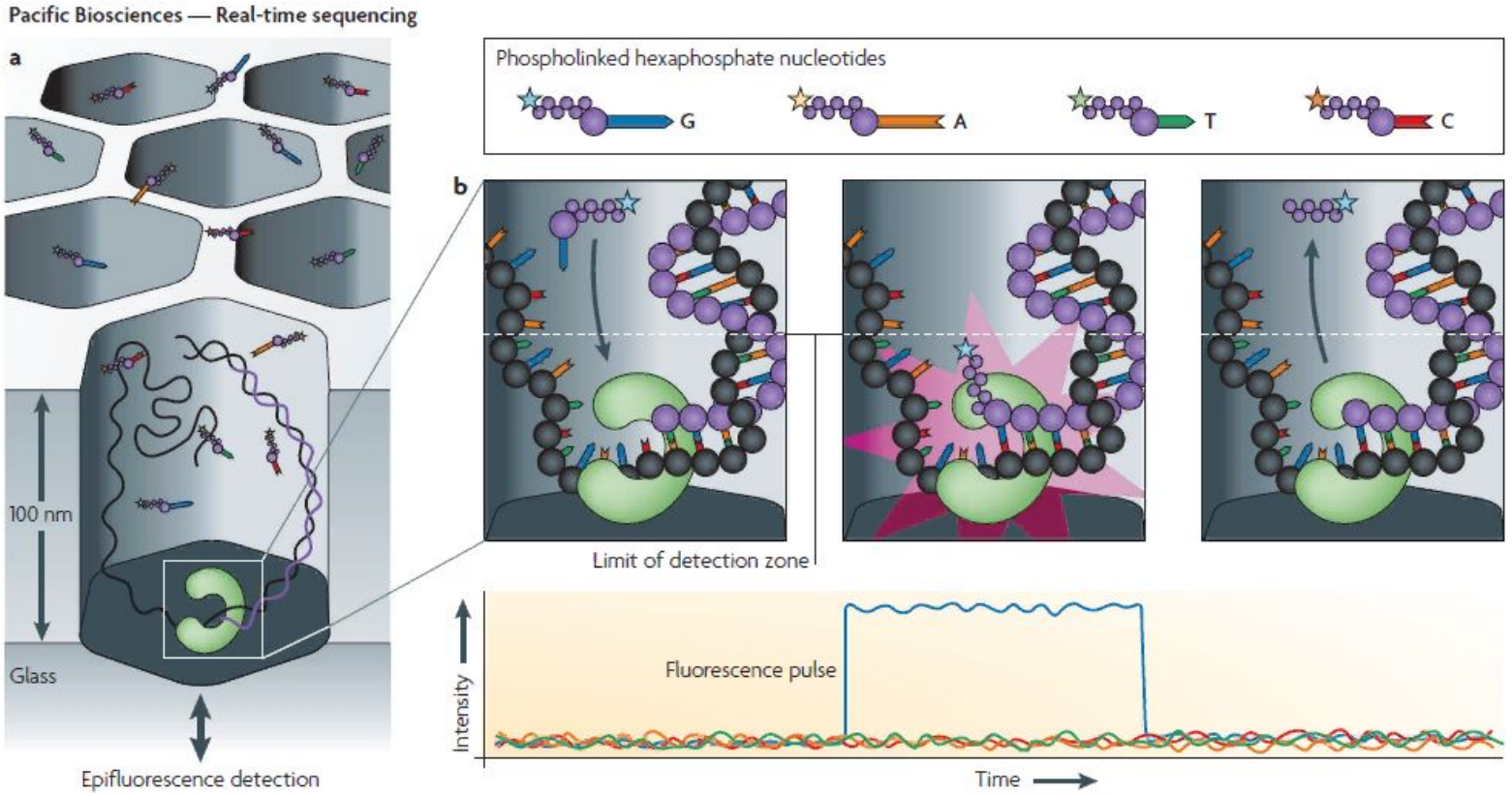


The sequencing cycles are repeated to determine the sequence of bases i a fragment, one base at a time.

Pacific Biosciences

single molecule real-time (SMRT) sequencing
Одномолекулярное секвенирование в реальном времени

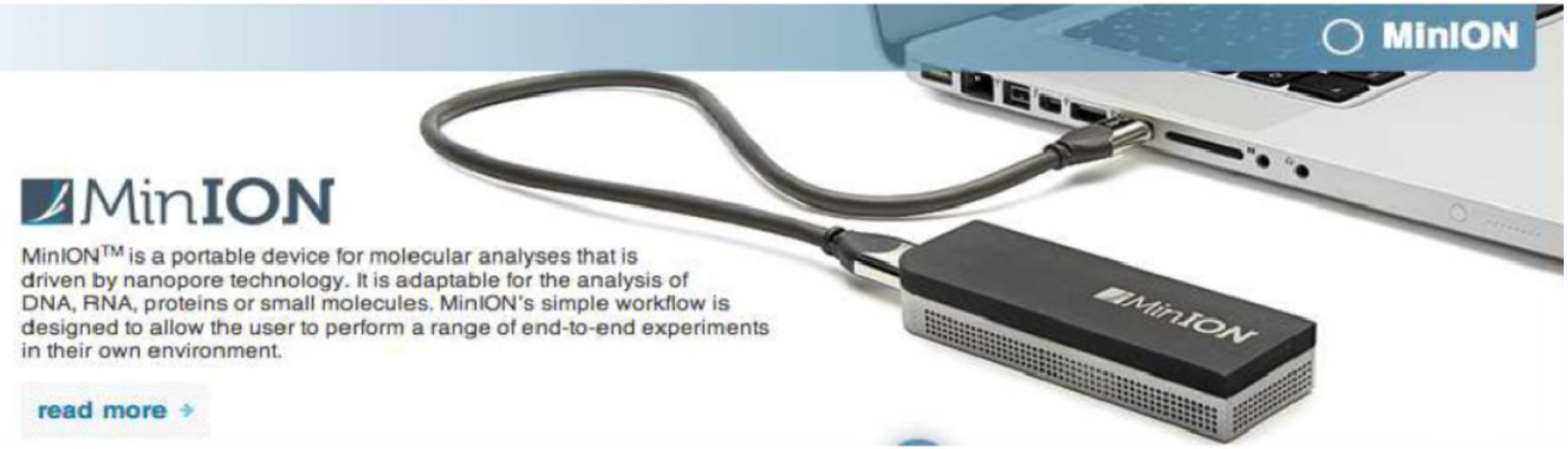
- Секвенировании без амплификации
- Очень длинные риды
  - Производит чтения со средней длиной от 10 000 до 15 000 пар оснований, причем самые длинные риды могут быть более 30 000 пар оснований


Pacific Biosciences — Real-time sequencing
a
100 nm
Glass
Epifluorescence detection
Phospholinked hexaphosphate nucleotides
G
A
T
C
b
Limit of detection zone
Intensity
Fluorescence pulse
Time

Nanopore

Oxford Nanopore Technologies

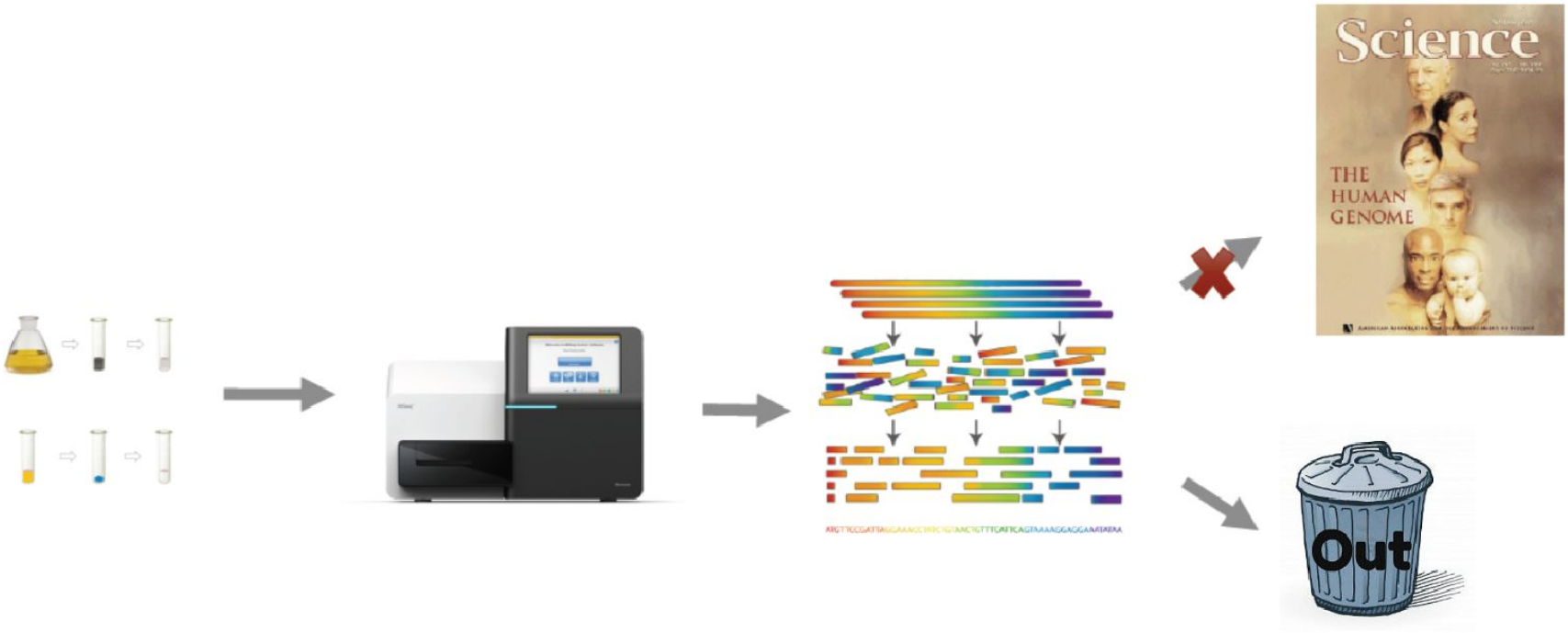
nanopore sequencing



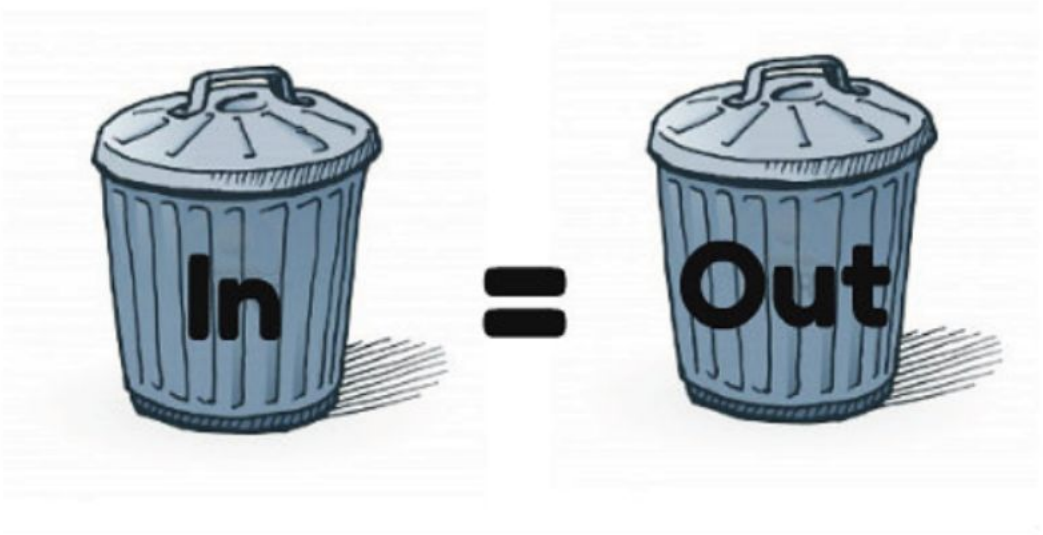
Сравнение платформ NGS

Instrument	Max Read length (nucleotides)	Output (Gb)	Runtime	Error rate (%)	Error type	Reagent cost/Gb ($)
Roche 454 FLX+	1000	0.70	20 hours	1	Indel	6200
Roche 454 Junior	450	0.035	10 hours	1	Indel	19540
IonTorrent Proton	200	10	4-6 hours	~1	Indel	82
IonTorent PGM	400	2	2-7 hours	~1	Indel	460
Illumina HiSeq 2500 v4	125	500	10 hours – 6 days	~0.1	substitution	30
Illumina Miseq v3	300	15	55 hours	~0.1	substitution	109
PacBio RS II	30000	0.187	2 hours	~13	Indel	1111
AB 3730 (capillary)	650	-	2 hours	0.1-1	substitution	2307692

Контроль качества данных



Garbage in – Garbage out



- Your analysis is only as good as your data

Алгоритм контроля качества

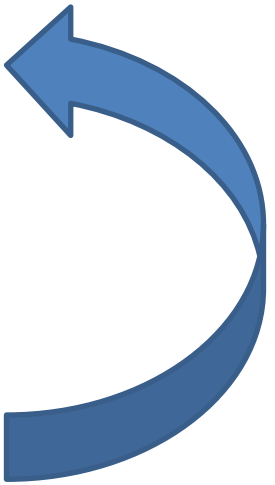
Проверка качества

Определение проблемы

Решение проблемы

Проверка качества

Последующий анализ



Зачем чистить данные?

- Риды низкого качества
- Контаминация (примесь ДНК другого организма)
- Служебные последовательности (адаптеры, индексы)
- Артефакты создания библиотек (некоторые последовательности встречаются чаще, а не равномерно)
- Различный формат данных
- Человеческий фактор

FASTA и FASTQ форматы

FASTA

```
>gi|31563518|ref|NP_852610.1| microtubule-associated proteins 1A/1B light chain 3A isoform b [Homo sapiens]
MKMRFFSSPCGKAAVDPADRCKEVQQIRDQHPSKIPVIIERYKGEKQLPVLDKTKFLVPDHVNMSELVKI
IRRRLQLNPTQAFFLLVNQHSMVSVSTPIADIYEQEKDEDGFLYMVYASQETFGF
```

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

1. Линия начинающаяся с @ содержит идентификатор последовательности
2. Последовательность
3. Линия начинающаяся с + заполняется факультативно
4. Линия с величинами качества прочтения, кодируемые в ASCII формате

Шкала качества Фред (Phred)

Оценки качества нуклеотида Q определяются как величина, которая логарифмически зависит от вероятностей ошибки P

$$Q = -10 \log_{10} P$$

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 10,000	99.99%
50	1 in 100,000	99.999%
60	1 in 1,000,000	99.9999%

Таблица ASCII символов

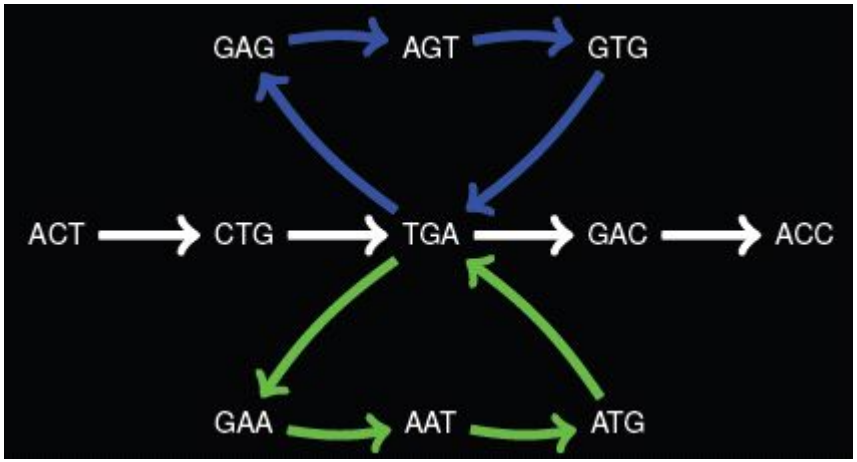
Dec	Hex	Char	Dec	Hex	Char	Dec	Hex	Char	Dec	Hex	Char
0	00	Null	32	20	Space	64	40	@	96	60	`
1	01	Start of heading	33	21	!	65	41	A	97	61	a
2	02	Start of text	34	22	"	66	42	B	98	62	b
3	03	End of text	35	23	#	67	43	C	99	63	c
4	04	End of transmit	36	24	$	68	44	D	100	64	d
5	05	Enquiry	37	25	%	69	45	E	101	65	e
6	06	Acknowledge	38	26	&	70	46	F	102	66	f
7	07	Audible bell	39	27	'	71	47	G	103	67	g
8	08	Backspace	40	28	(	72	48	H	104	68	h
9	09	Horizontal tab	41	29	)	73	49	I	105	69	i
10	0A	Line feed	42	2A	*	74	4A	J	106	6A	j
11	0B	Vertical tab	43	2B	+	75	4B	K	107	6B	k
12	0C	Form feed	44	2C	,	76	4C	L	108	6C	l
13	0D	Carriage return	45	2D	-	77	4D	M	109	6D	m
14	0E	Shift out	46	2E	.	78	4E	N	110	6E	n
15	0F	Shift in	47	2F	/	79	4F	O	111	6F	o
16	10	Data link escape	48	30	0	80	50	P	112	70	p
17	11	Device control 1	49	31	1	81	51	Q	113	71	q
18	12	Device control 2	50	32	2	82	52	R	114	72	r
19	13	Device control 3	51	33	3	83	53	S	115	73	s
20	14	Device control 4	52	34	4	84	54	T	116	74	t
21	15	Neg. acknowledge	53	35	5	85	55	U	117	75	u
22	16	Synchronous idle	54	36	6	86	56	V	118	76	v
23	17	End trans. block	55	37	7	87	57	W	119	77	w
24	18	Cancel	56	38	8	88	58	X	120	78	x
25	19	End of medium	57	39	9	89	59	Y	121	79	y
26	1A	Substitution	58	3A	:	90	5A	Z	122	7A	z
27	1B	Escape	59	3B	;	91	5B	[	123	7B	{
28	1C	File separator	60	3C	<	92	5C	\	124	7C	|
29	1D	Group separator	61	3D	=	93	5D	]	125	7D	}
30	1E	Record separator	62	3E	>	94	5E	^	126	7E	~
31	1F	Unit separator	63	3F	?	95	5F	_	127	7F	□

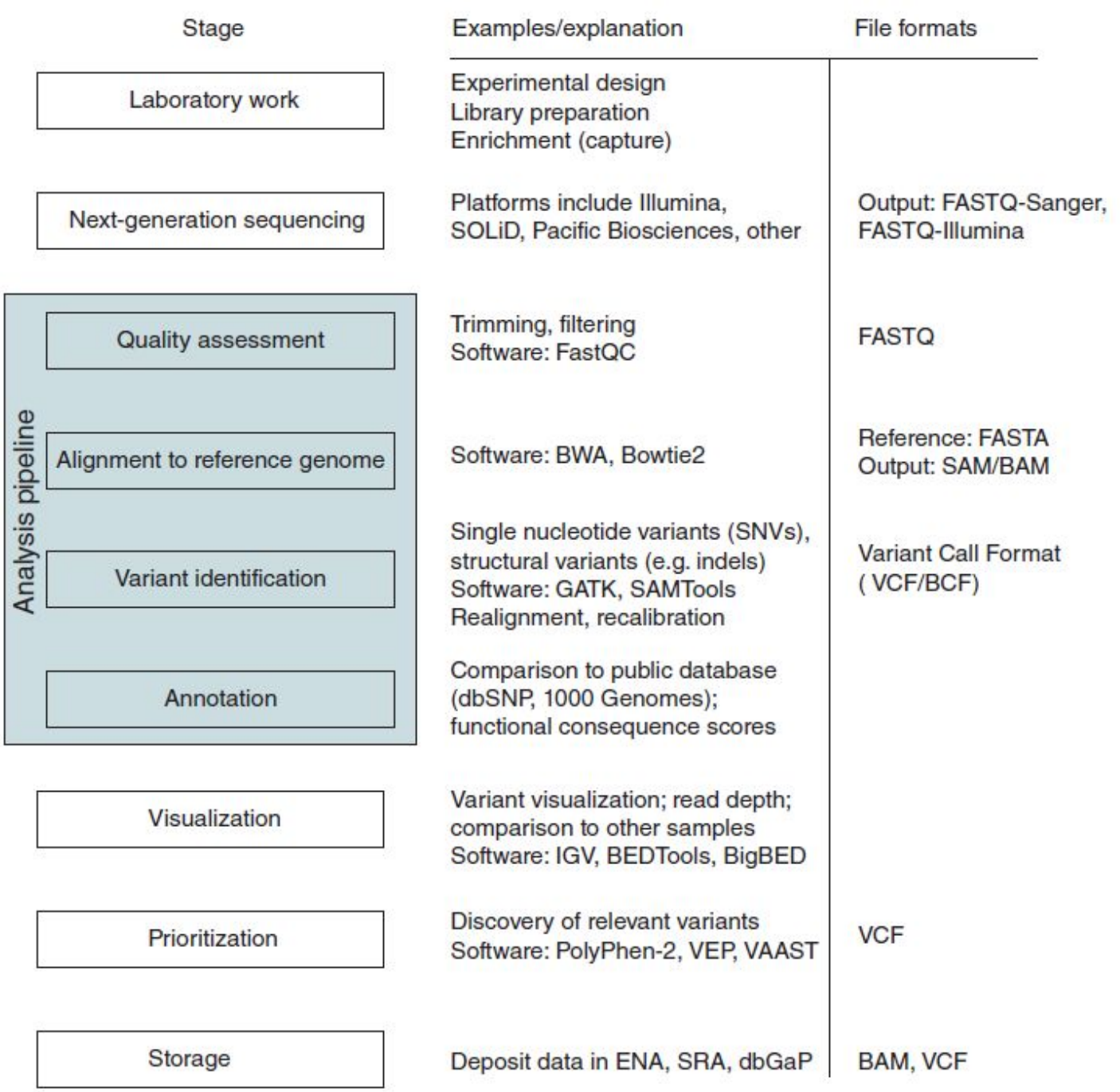
Разные Phred шкалы

```
 SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
 ..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX.....................
 ..............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII.....................
 ................................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ.....................
 LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL.....................................................
 !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
 |                          |    |        |                              |                     |
33                          59   64       73                             104                   126
 0........................26...31.......40
                           -5....0........9.............................40
                                0........9.............................40
                                   3.....9.............................40
 0.2......................26...31.........41

S - Sanger        Phred+33,  raw reads typically (0, 40)
X - Solexa        Solexa+64, raw reads typically (-5, 40)
I - Illumina 1.3+ Phred+64,  raw reads typically (0, 40)
J - Illumina 1.5+ Phred+64,  raw reads typically (3, 40)
    with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
    (Note: See discussion above).
L - Illumina 1.8+ Phred+33,  raw reads typically (0, 41)
```

Сборка генома




Stage
Examples/explanation
File formats
Laboratory work
Experimental design
Library preparation
Enrichment (capture)
Next-generation sequencing
Platforms include Illumina, SOLiD, Pacific Biosciences, other
Output: FASTQ-Sanger, FASTQ-Illumina
Analysis pipeline
Quality assessment
Trimming, filtering
Software: FastQC
FASTQ
Alignment to reference genome
Software: BWA, Bowtie2
Reference: FASTA
Output: SAM/BAM
Variant identification
Single nucleotide variants (SNVs), structural variants (e.g. indels)
Software: GATK, SAMTools
Realignment, recalibration
Variant Call Format (VCF/BCF)
Annotation
Comparison to public database (dbSNP, 1000 Genomes); functional consequence scores
Visualization
Variant visualization; read depth; comparison to other samples
Software: IGV, BEDTools, BigBED
Prioritization
Discovery of relevant variants
Software: PolyPhen-2, VEP, VAAST
VCF
Storage
Deposit data in ENA, SRA, dbGaP
BAM, VCF

FastQC – инструмент для контроля качества данных

- На вход – исходные данные с секвенатора
- HTML отчет
- Графический интерфейс и версия с командной строкой

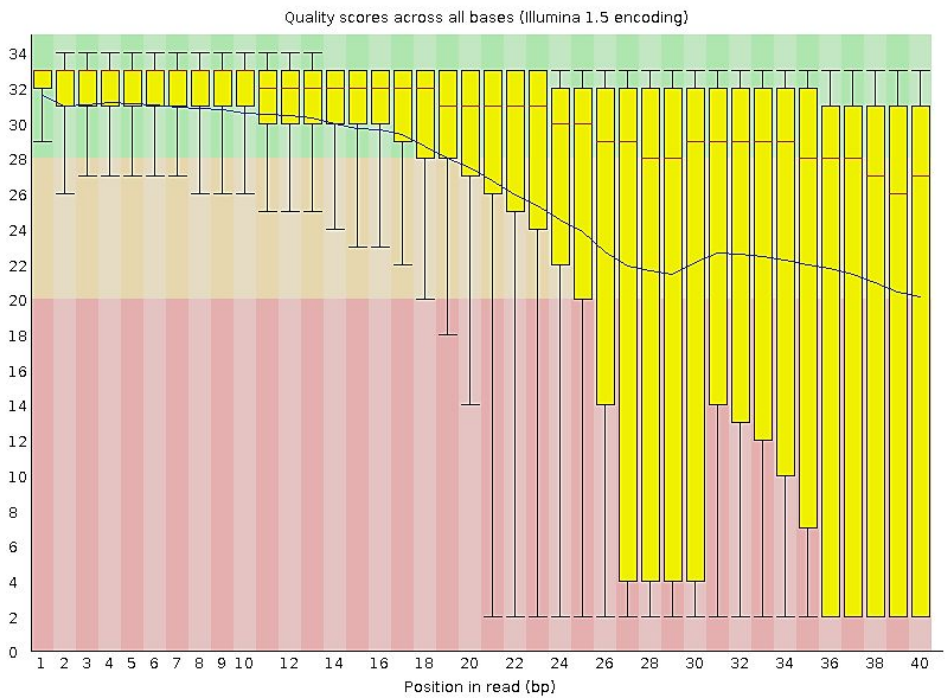
www.bioinformatics.babraham.ac.uk/projects/fastqc

FastQC

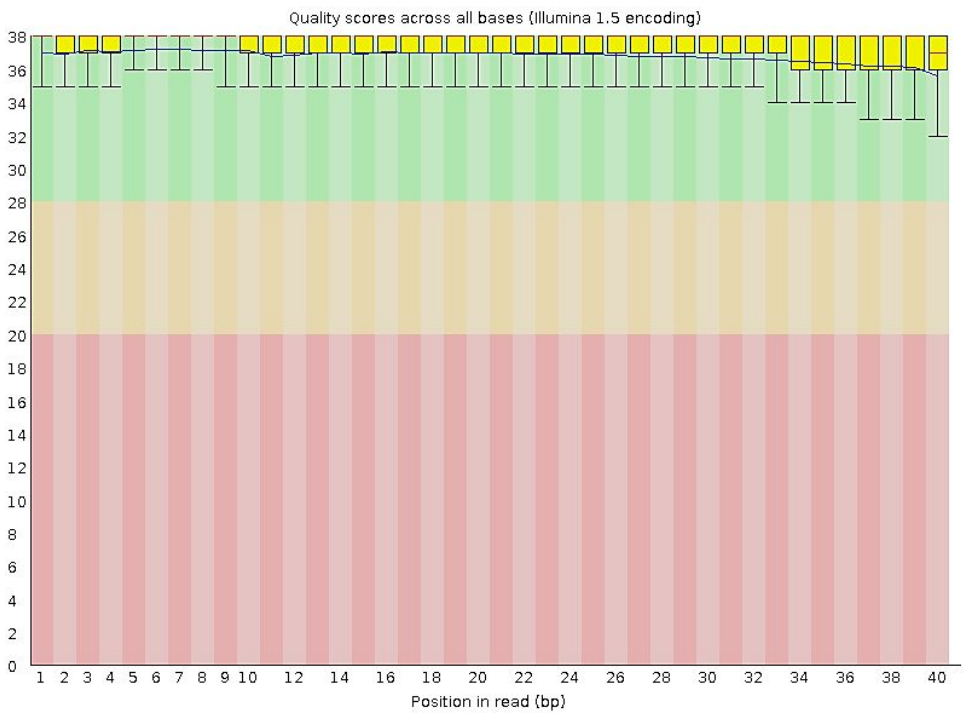
Basic Statistics

Measure	Value
Filename	good_sequence_short.txt
File type	Conventional base calls
Encoding	Illumina 1.5
Total Sequences	250000
Sequences flagged as poor quality	0
Sequence length	40
%GC	45

FastQC: распределение качества по остаткам



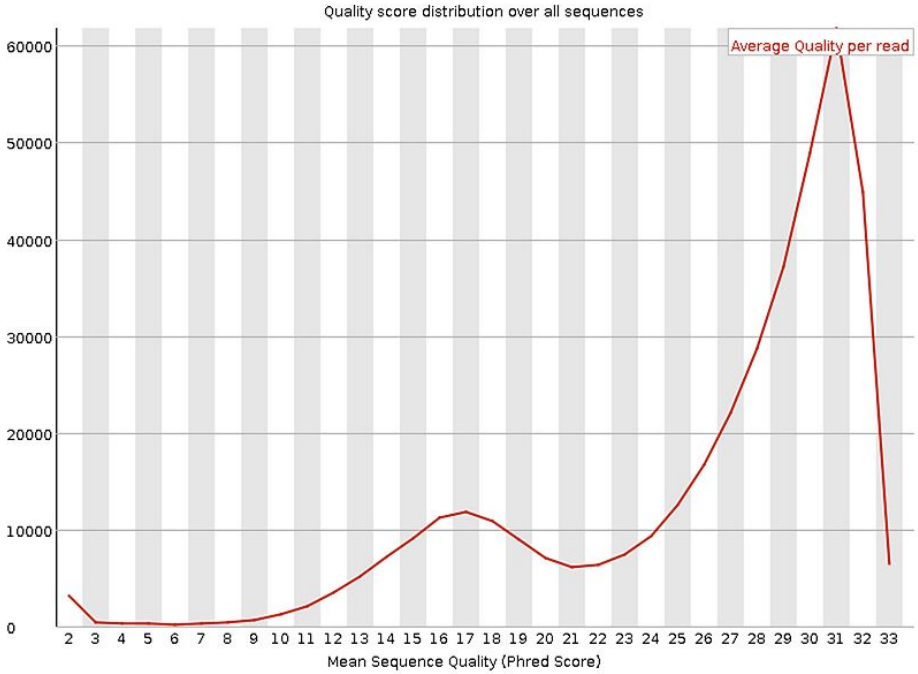
Плохое

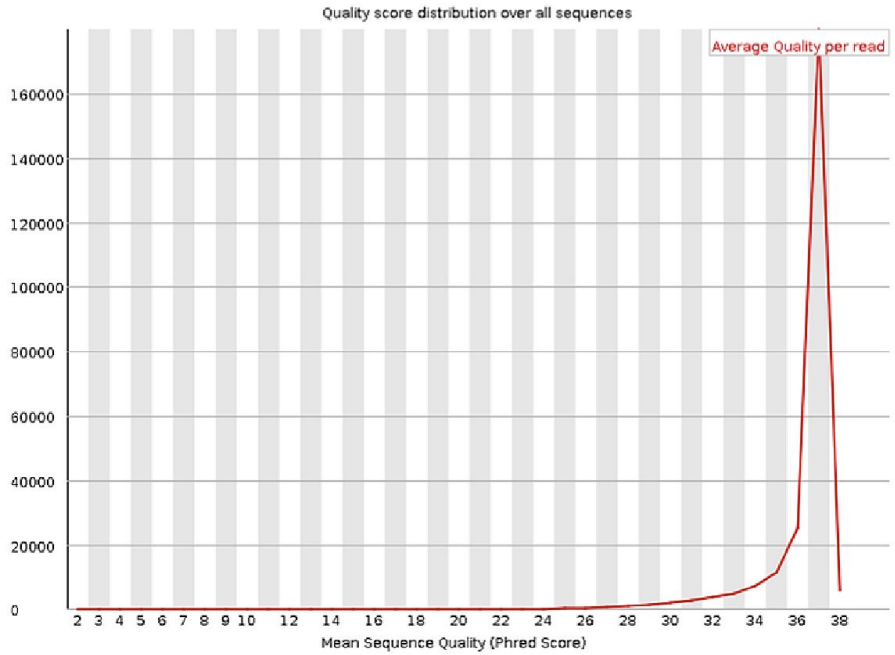


Хорошее

У Illumina качество ридов обычно уменьшается к 3' концу

FastQC: распределение качества по ридам



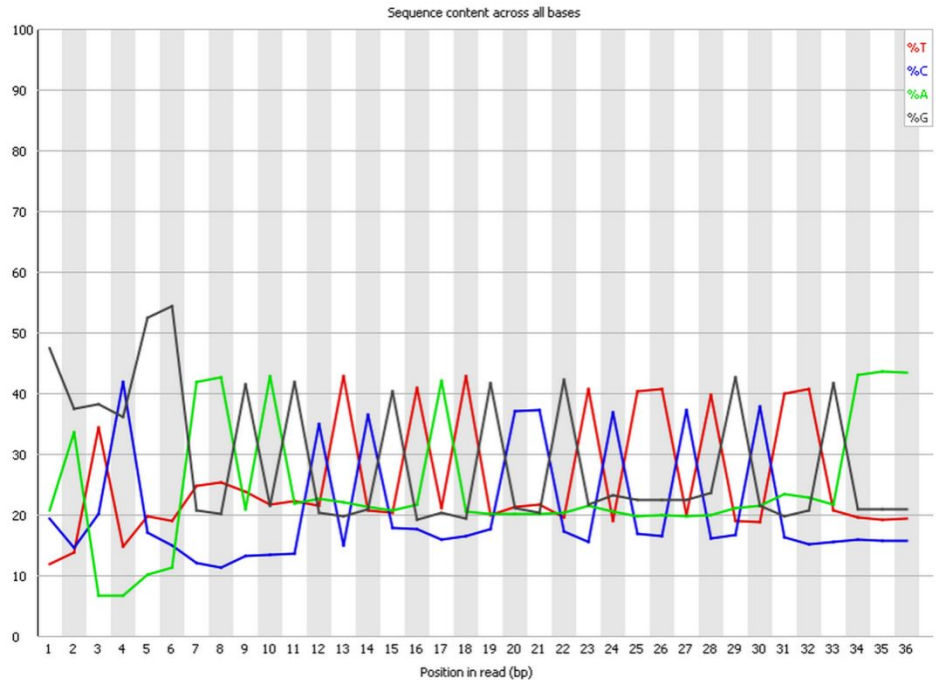


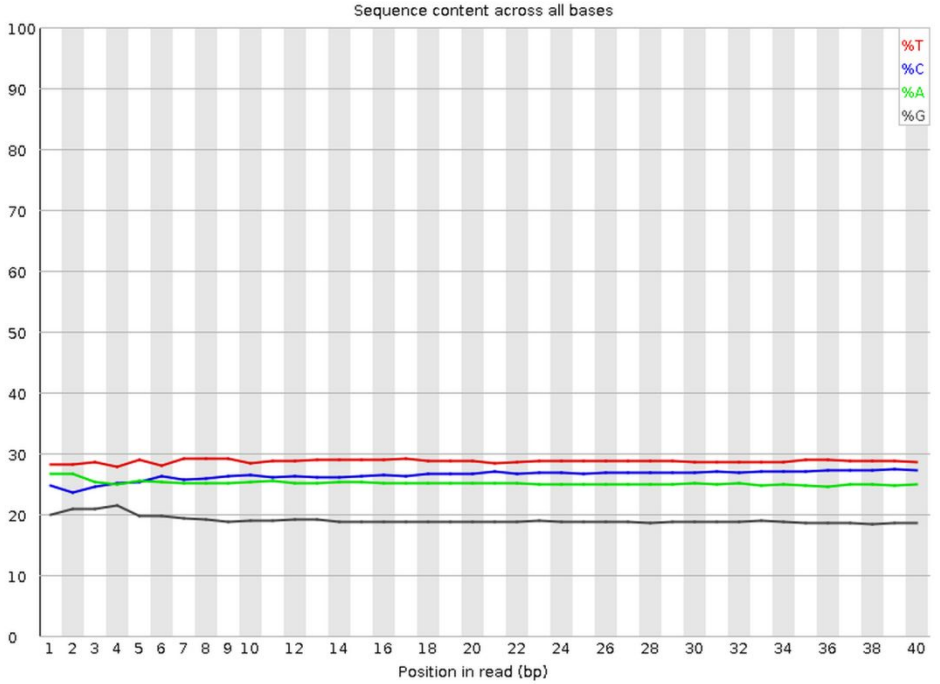
Плохое

Хорошее

Мы можем выделить группы ридов с низким и высоким качеством

FastQC: распределение качества по составу остатков



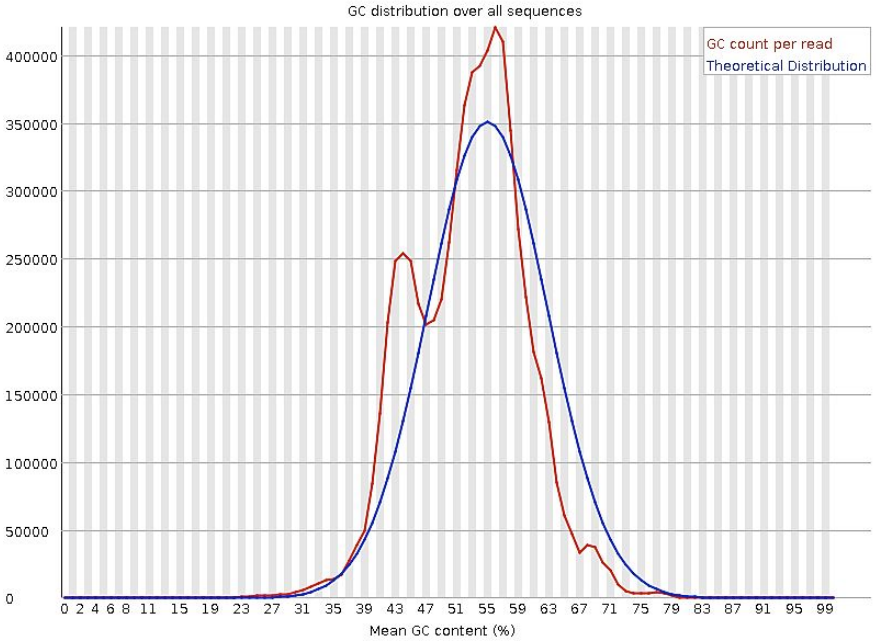


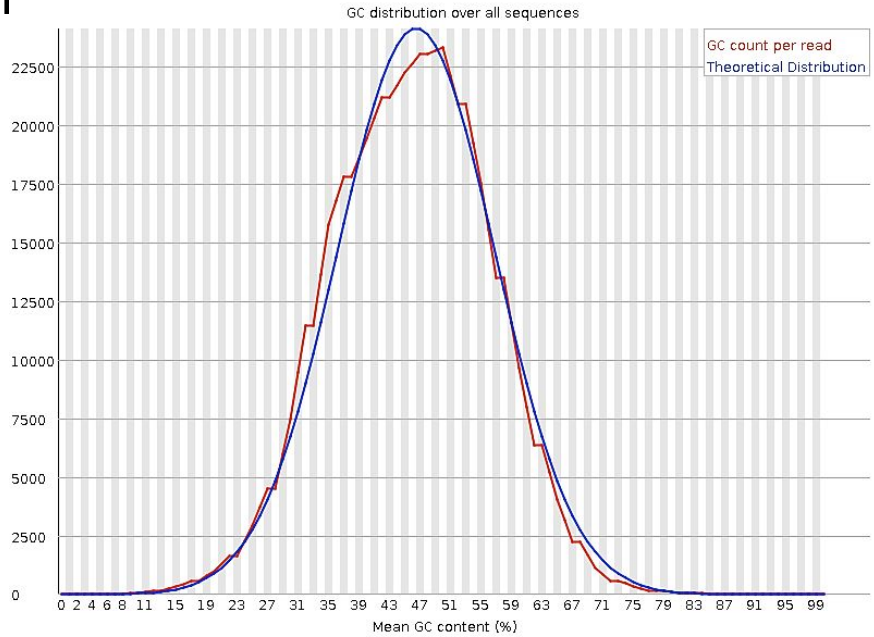
Плохое

Хорошее

Мы можем определить адаптеры или сдвиг

FastQC: распределение ридов по GC составу



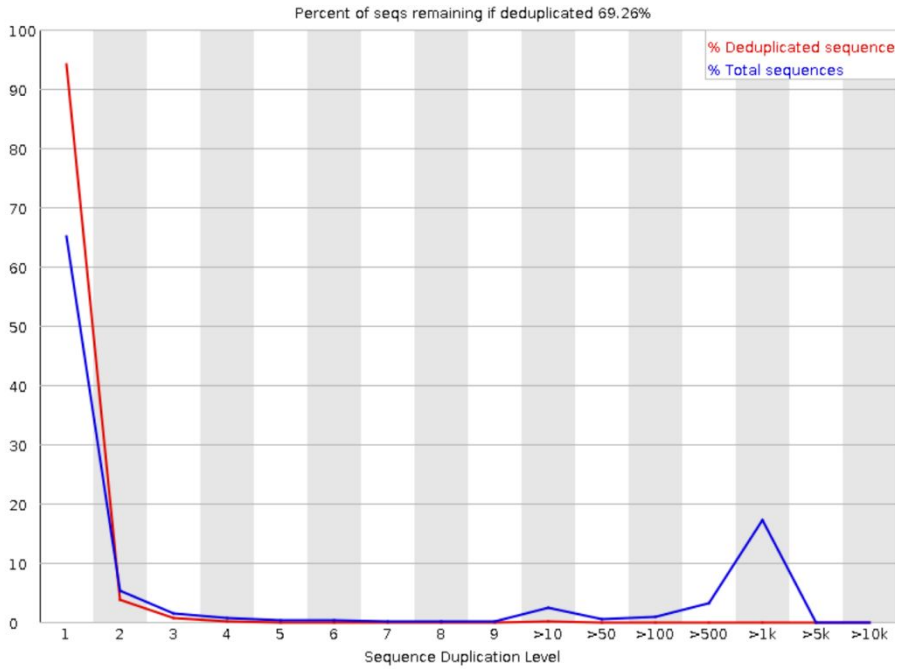


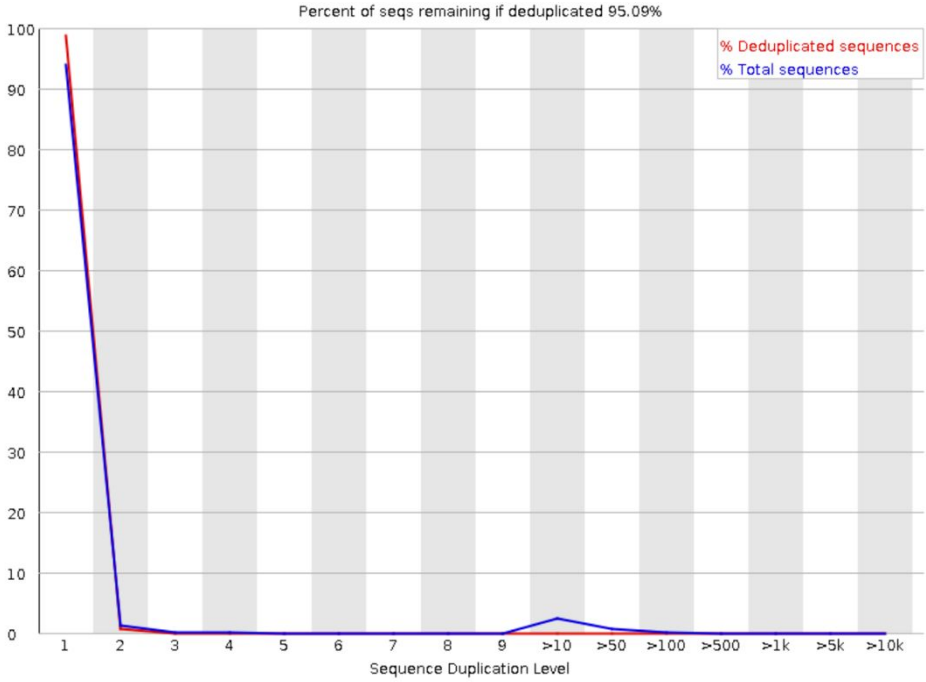
Плохо

Хорошо

GC пики могут свидетельствовать о контаминации

FastQC: уровни дупликаций последовательностей





Плохо

Хорошо

Высокий уровень дупликации свидетельствует об оверамплификации некоторых последовательностей при PCR

FastQC: Overrepresented sequences

Overrepresented sequences

Sequence	Count	Percentage	Possible Source
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCT	8122	8.122	Illumina Paired End PCR Primer 2 (100% over 40bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGATCGGAAG	5086	5.086	Illumina Paired End PCR Primer 2 (97% over 36bp)
AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTAC	1085	1.085	Illumina Single End PCR Primer 1 (100% over 40bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGGAAG	508	0.508	Illumina Paired End PCR Primer 2 (97% over 36bp)
AATTATACGGCGACCACCGAGATCTACACTCTTTCCCTAC	242	0.242	Illumina Single End PCR Primer 1 (97% over 40bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAAGATCGGAA	235	0.23500000000000001	Illumina Paired End Adapter 2 (96% over 31bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCGAGATCGGAAGA	228	0.22799999999999998	Illumina Paired End Adapter 2 (96% over 28bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGGACG	205	0.20500000000000002	Illumina Paired End PCR Primer 2 (97% over 36bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGGATCGGAA	183	0.183	Illumina Paired End Adapter 2 (100% over 32bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGGTCGGAAG	183	0.183	Illumina Paired End Adapter 2 (100% over 32bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGAACT	164	0.164	Illumina Paired End PCR Primer 2 (97% over 40bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGGTCT	129	0.129	Illumina Paired End PCR Primer 2 (97% over 40bp)
AATTATACTTCTACCACCTATATCTACACTCTTTCCCTAC	123	0.123	No Hit
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGGACT	122	0.122	Illumina Paired End PCR Primer 2 (97% over 36bp)
CGGTTCAGCAGGAATGCCGAGATCGGAAGAGCGGTTCAGC	113	0.11299999999999999	Illumina Paired End PCR Primer 2 (96% over 25bp)

Плохо

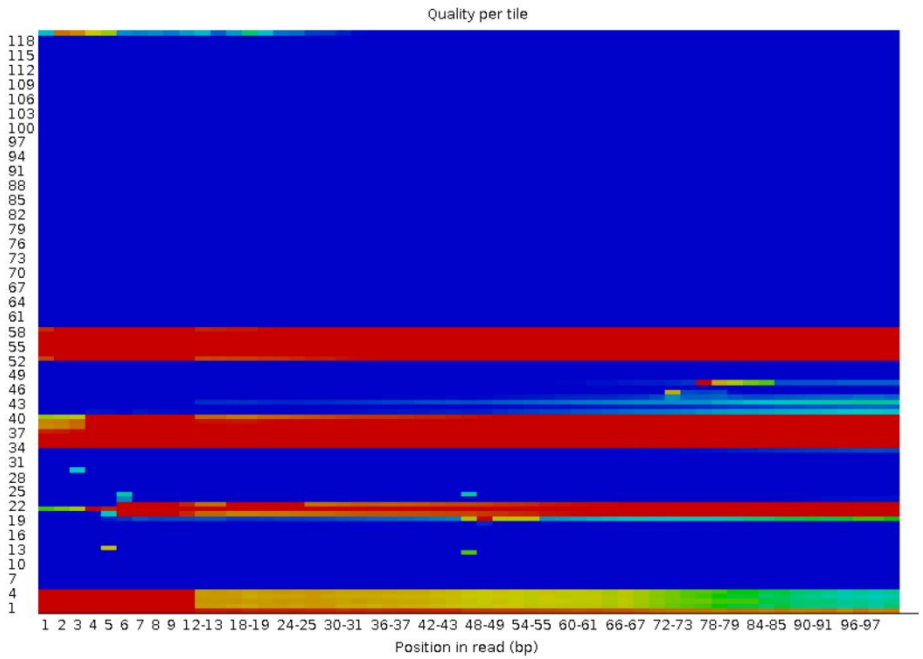
Overrepresented sequences

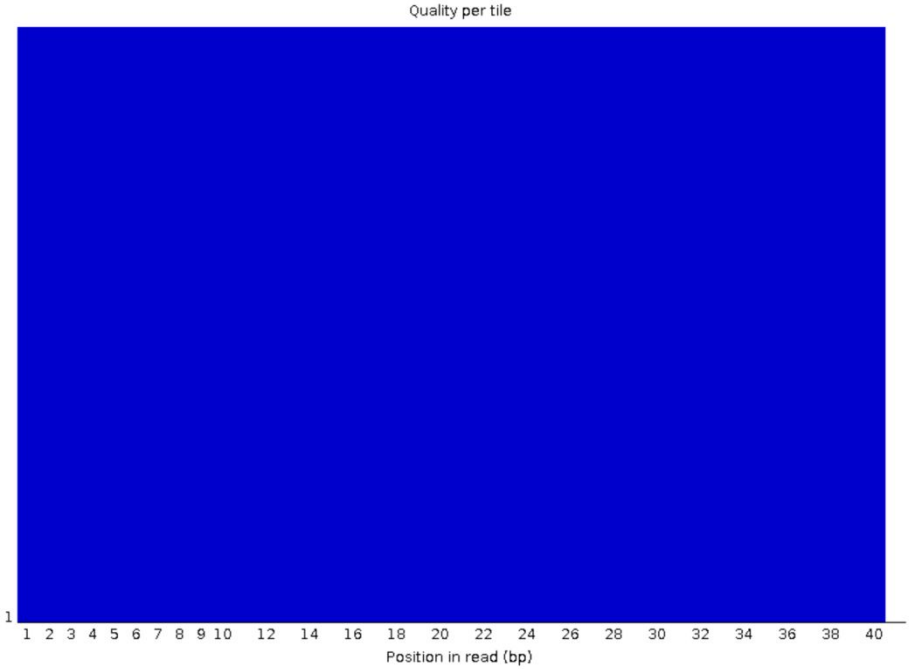
No overrepresented sequences

Хорошо

Перепредставленные последовательности могут показывать источник контаминации

FastQC: Качество ячеек





Плохо

Хорошо

У Illumina можно определить проблемы с ячейками

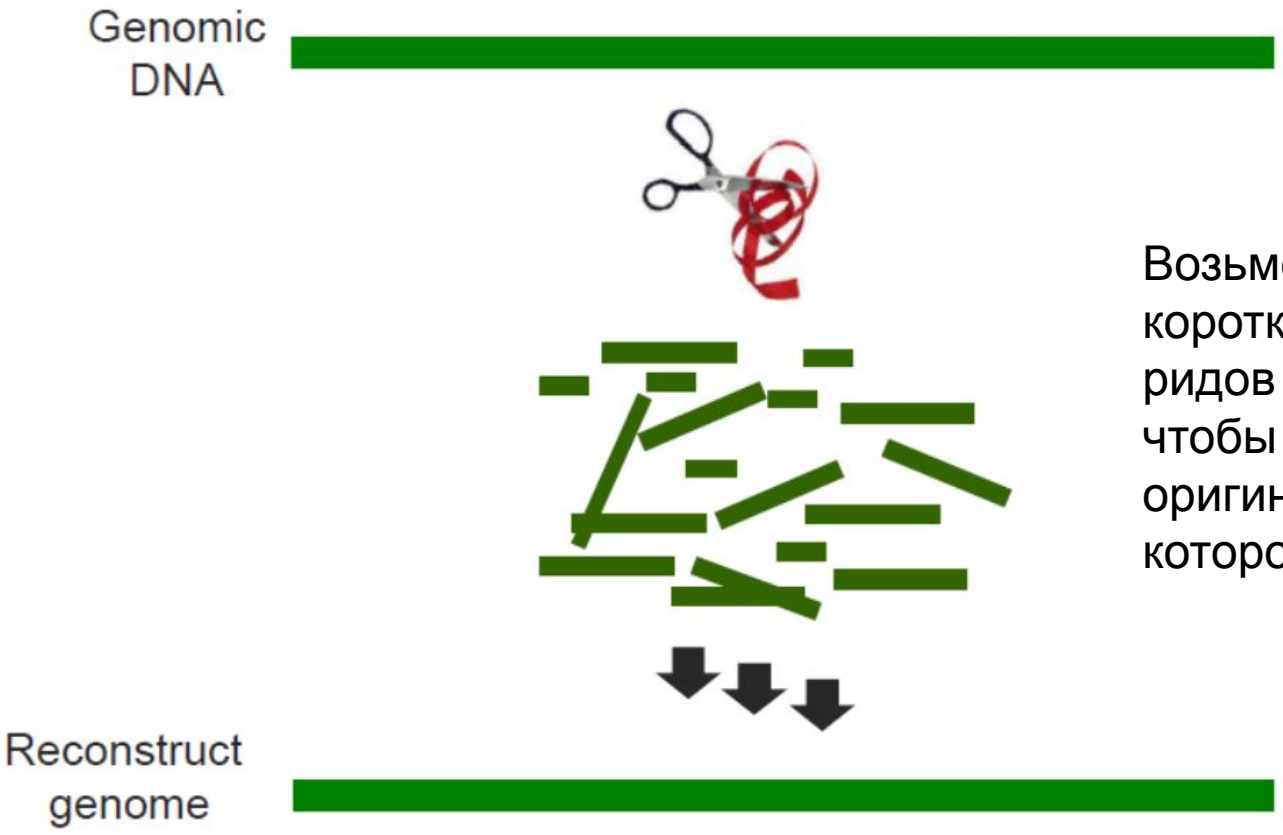
Шаги препроцессинга

- **Фильтрация данных по качеству**
  - Удаление ридов, качество которых ниже определенного порога;
  - Обрезание части ридов, где качество плохое
- **Удаление контаминации**
  - Биологическая контаминация: определение и удаление ридов
  - Контаминация адапторами: вырезание адапторов и удаление поврежденных ридов

У нас есть очищенные данные. Что дальше?

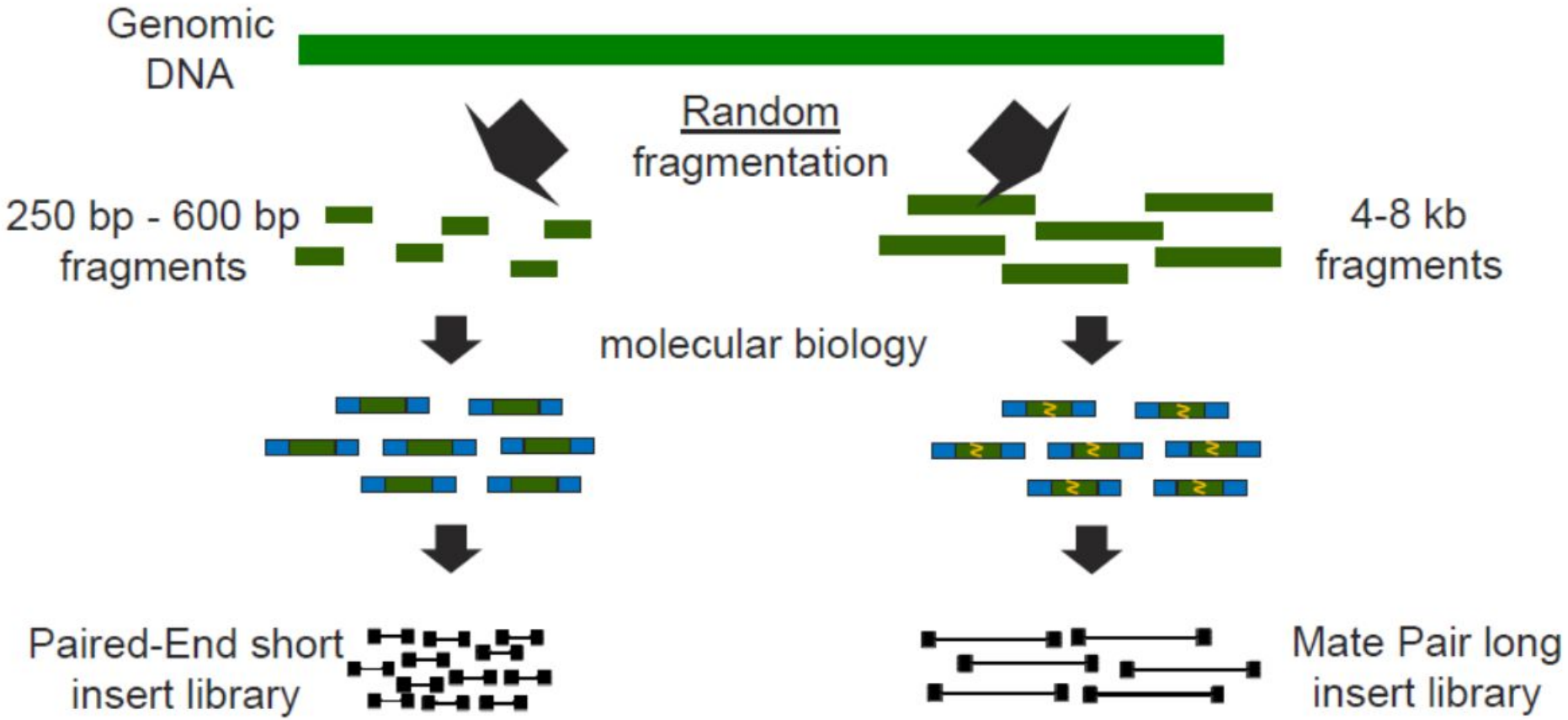
- Сборка de novo
- Сборка по референсному геному
- Выравнивание с референсным геномом

Сборка de novo

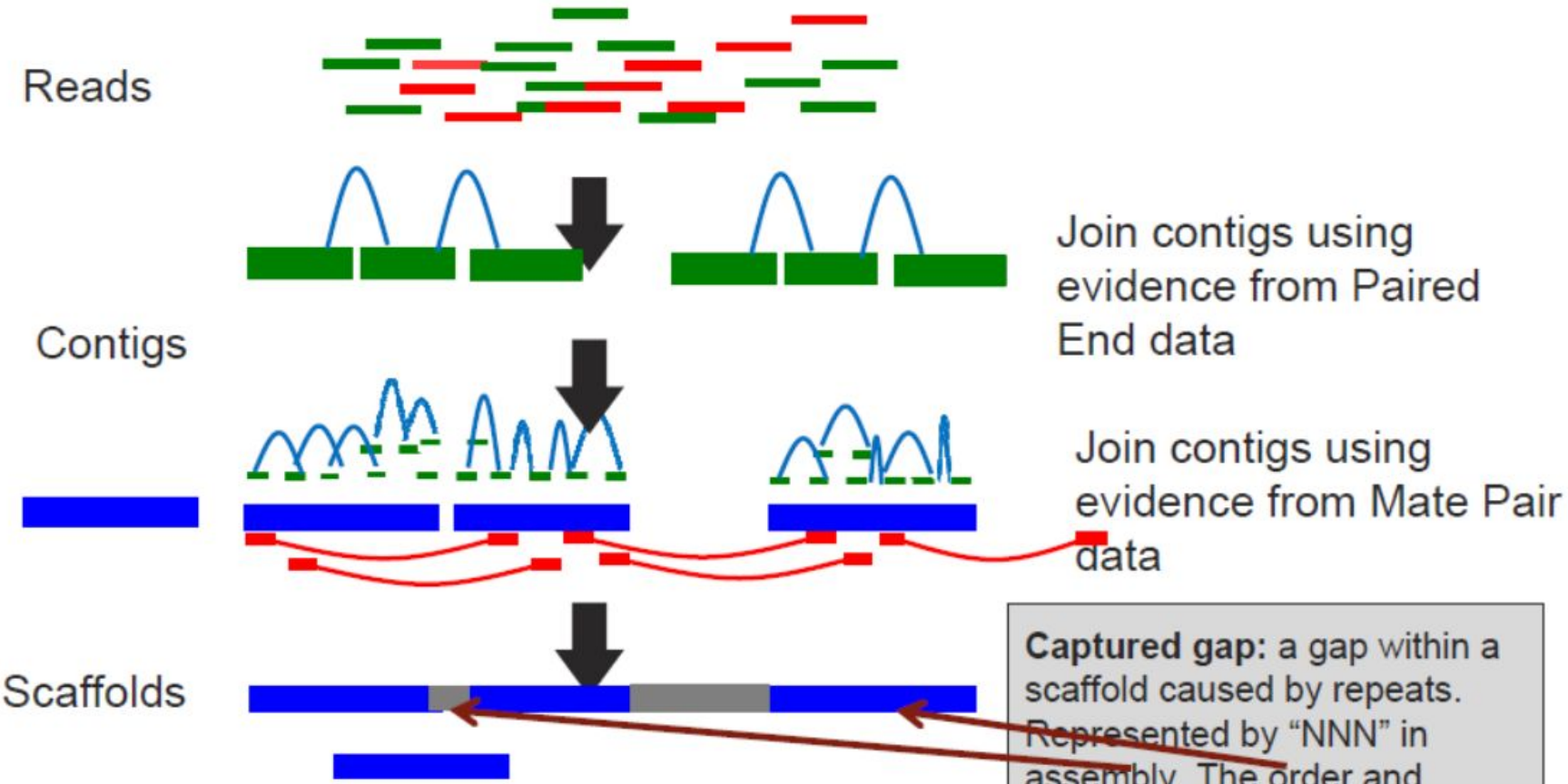


Возьмем большое количество коротких секвенированных ридов и поместим их вместе, чтобы воссоздать полный оригинальный геном из которого они были получены

Секвенирование геномов с использованием коротких ридов



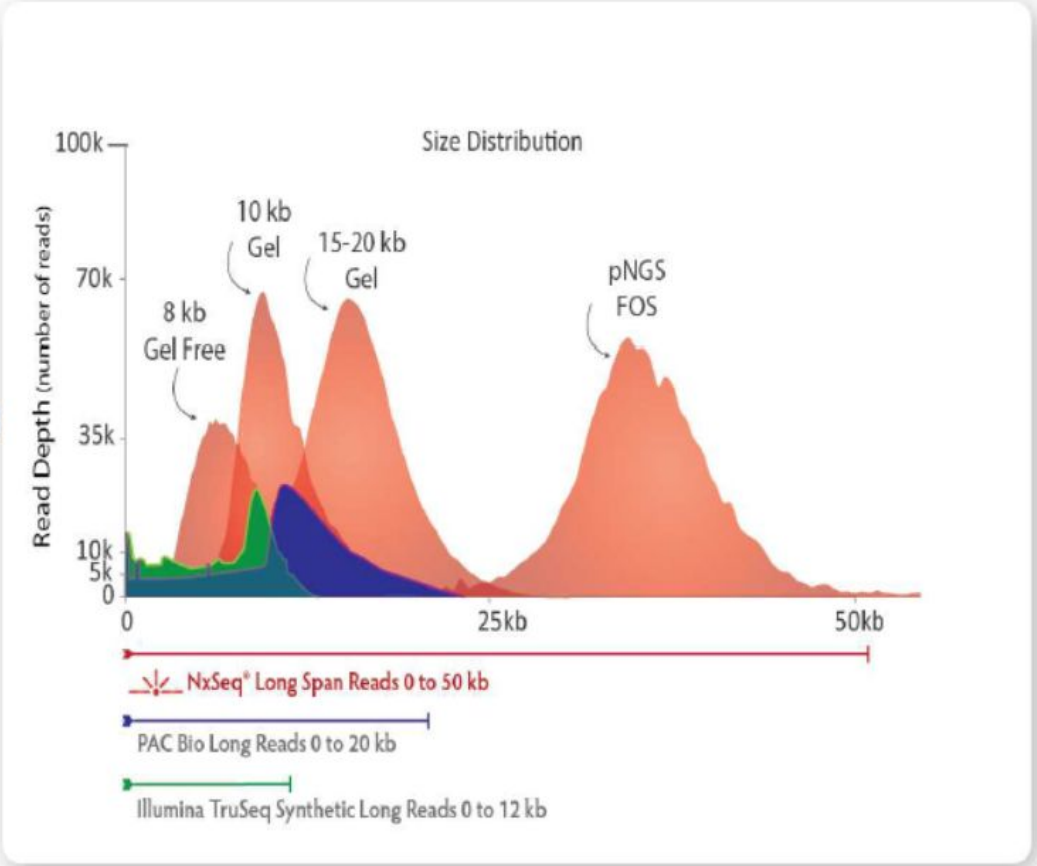
План сборки


Reads
Contigs
Scaffolds
Join contigs using evidence from Paired End data
Join contigs using evidence from Mate Pair data
Captured gap: a gap within a scaffold caused by repeats. Represented by "NNN" in assembly. The order and orientation of the contigs spanning the gap is known

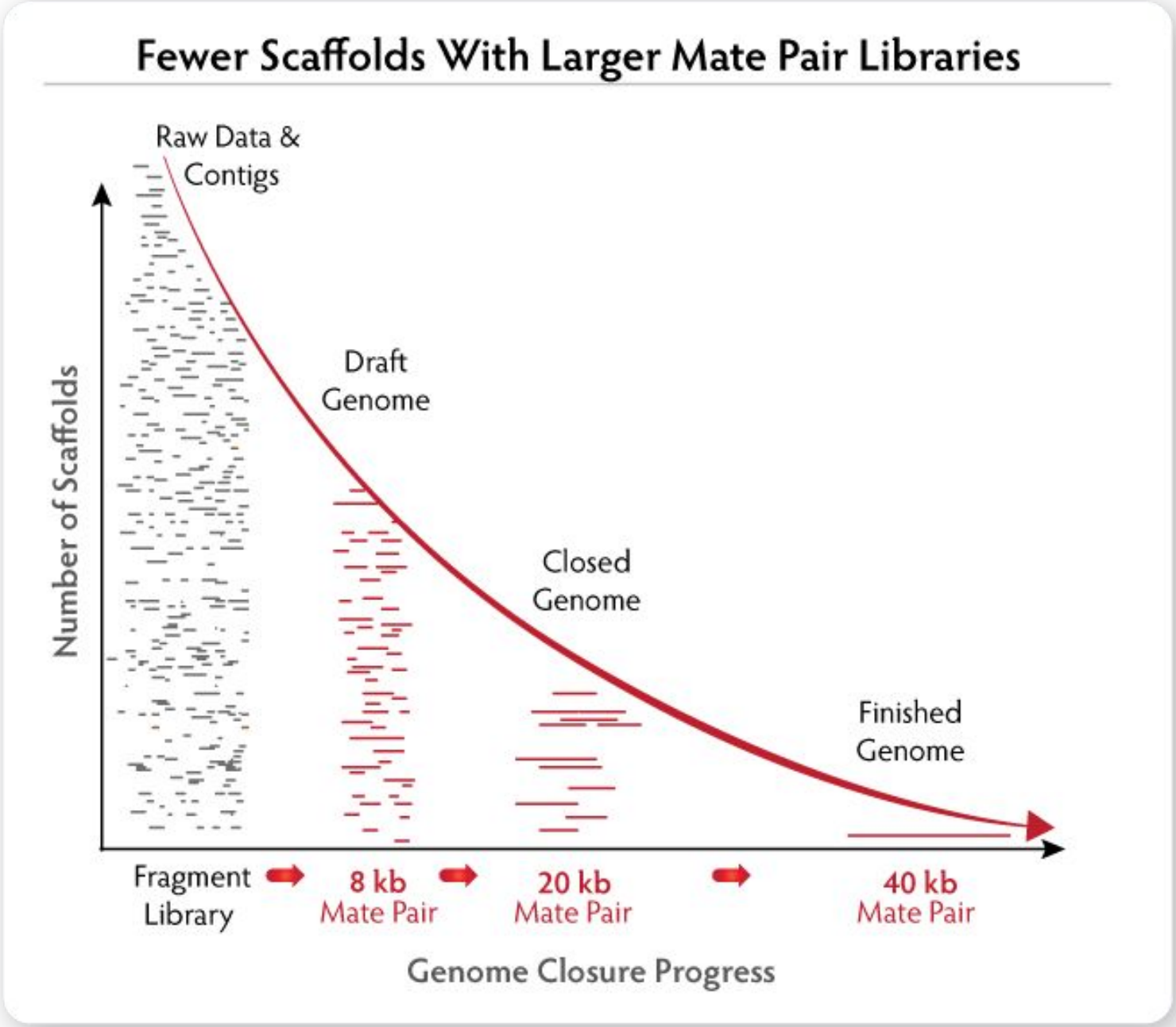
Разноразмерные библиотеки ДНК

Sanger era

-3 kb (IS elements)
-8 kb (ribosomal operons)
-25 kb (duplications)

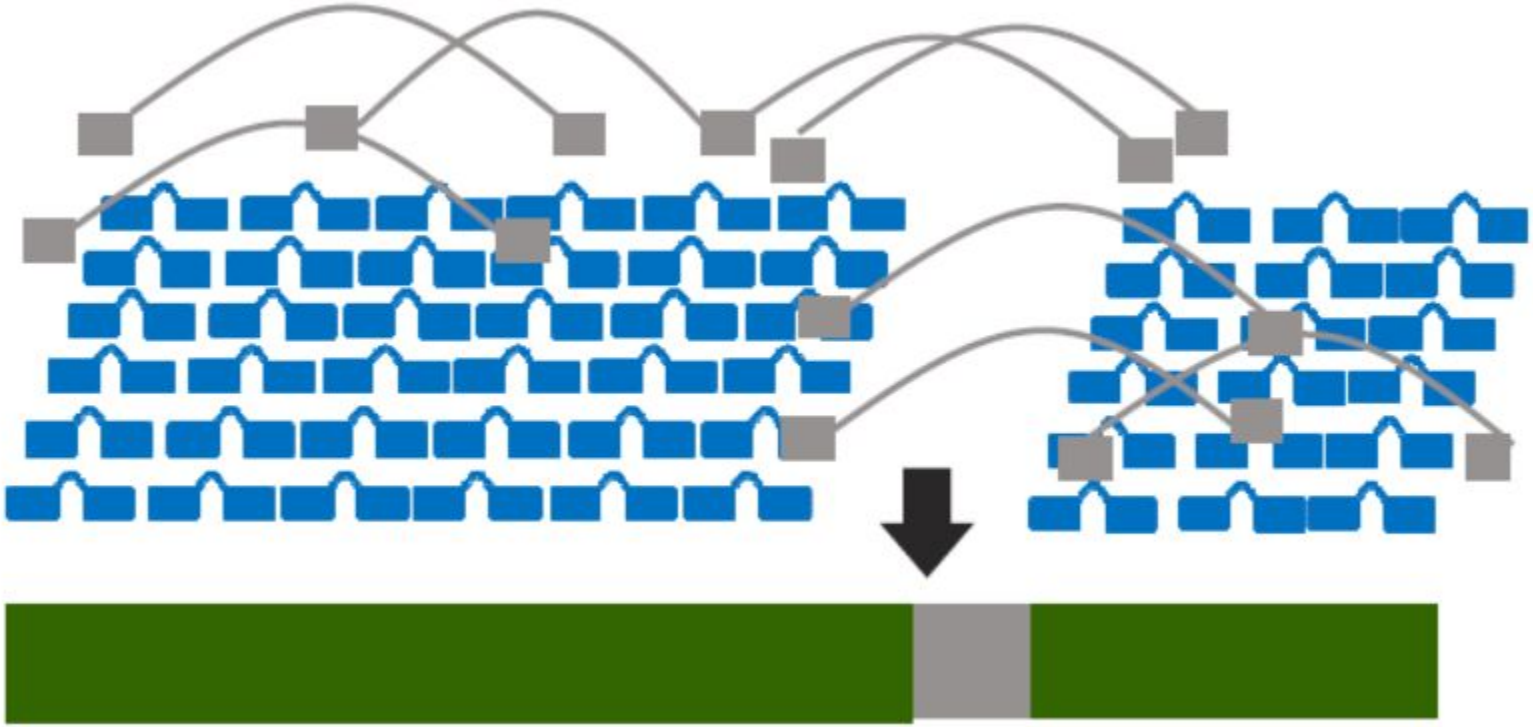


http://lucigen.com/landingpage/matepair/ir



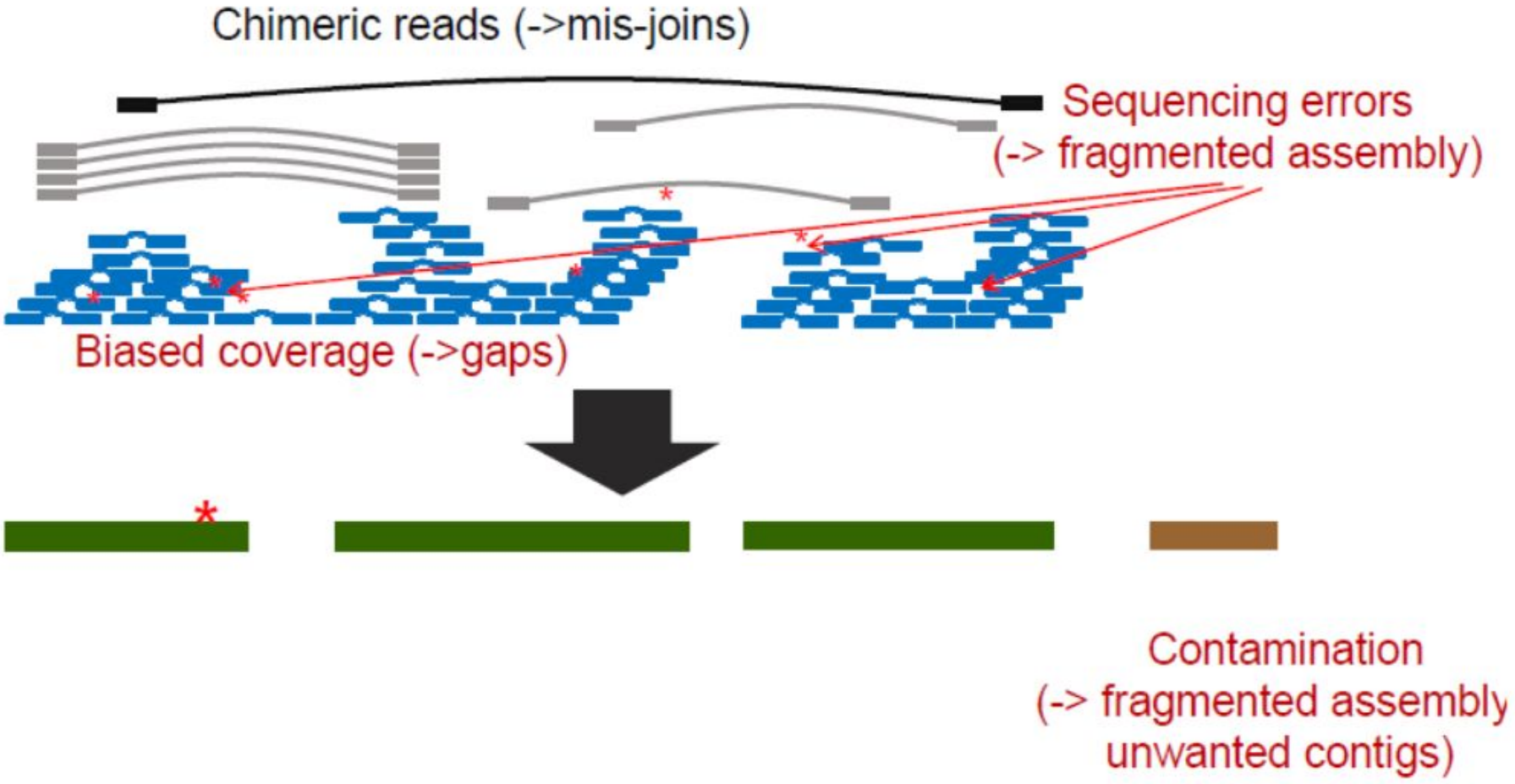
http://lucigen.com/landingpage/matepair/

Сборка генома в идеальном



Однородное покрытие ридами, нет ошибок и контаминации

Сборка генома в реальности

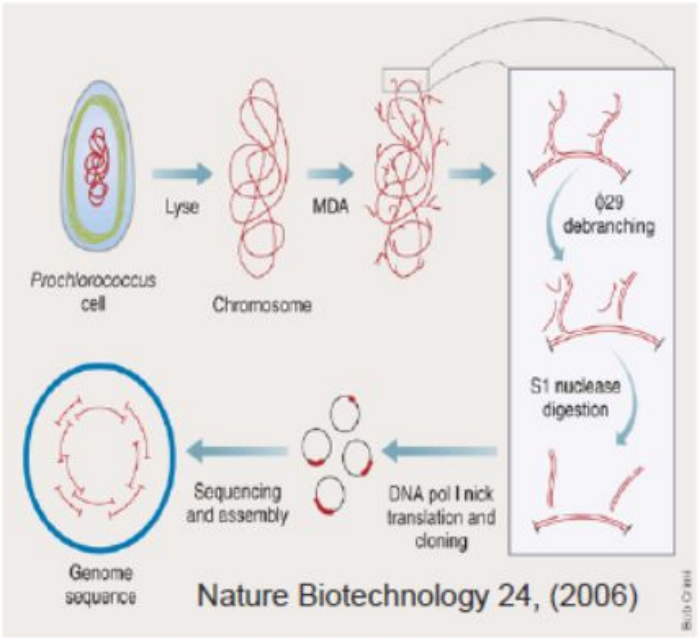


Metagenomes

- Typically size of metagenomic sequencing project is very large
- Different organisms have different coverage. Non-uniform sequence coverage results in significant under- and over-representation of certain community members
- Low coverage for the majority of organisms in highly complex communities leads to poor (if any) assemblies
- Chimerical contigs produced by co-assembly of sequencing reads originating from different species.
- Genome rearrangements and the presence of mobile genetic elements (phages, transposons) in closely related organisms further complicate assembly.
- No assemblers developed for metagenomic data sets

Single cell

- Uneven coverage
- High level of errors
- High level of chimerical data



Выбор правильной программы - сборщика геномов (ассемблер)

- На сколько большой геном?
- Существуют ли известные особенности этого генома (например, наличие большого числа повторов, GC состав)?
- Какое количество данных ожидается?
- Какого типа данные у вас есть?
- Какое качество данных и необходим ли их препроцессинг перед сборкой генома?

Сборщики геномов

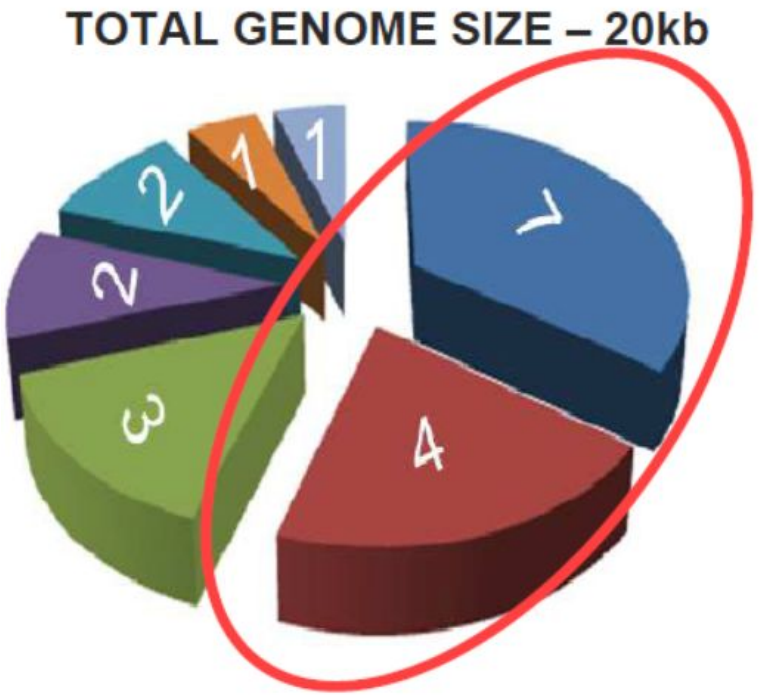
	Sequencing Platform	Error Correction	Assembly Approach	Genome Size	Libraries	Input Datasets	Reads	Reference
Celera	Sanger, Illumina, 454, IonTorrent, PacBio CCS	corrects reads, trims reads, removes poor quality reads and duplicates	OLC	prokaryotic/ mammalian	unpaired, paired-end	microbes, mammals	reads >75 bp	http://sourceforge.net/apps/mediawiki/wgs-assembler/index.php?title=Main_Page
AllPaths-LG	Illumina, PacBio	corrects reads, removes poor quality reads	de Bruijn graph	prokaryotic/ mammalian	unpaired, paired-end, mate pairs	microbes (regular isolates), mammals	short	http://www.broadinstitute.org/software/allpaths-lg/blog
SPAdes	Illumina, IonTorrent, PacBio, hybrid data sets	corrects reads, removes poor quality reads	de Bruijn graph	prokaryotic/ fungal	unpaired, paired-end, mate pairs	Microbes (single-cell, regular isolates), fungi	medium + long	http://bioinf.spbau.ru/spades
SSAKE	Illumina, IonTorrent,	n/a	gready	prokaryotic/ mammalian	unpaired, paired-end	microbes (regular isolates), mammals	short	http://www.bcgsc.ca/bioinfo/software/ssake

Оценка качества сборки генома

- Количество контигов
- Общая длинна всех контигов
- Длинна наибольшего контига
- Количество неправильно собранных контигов
- Количество идентифицированных генов
- GC состав %
- N50

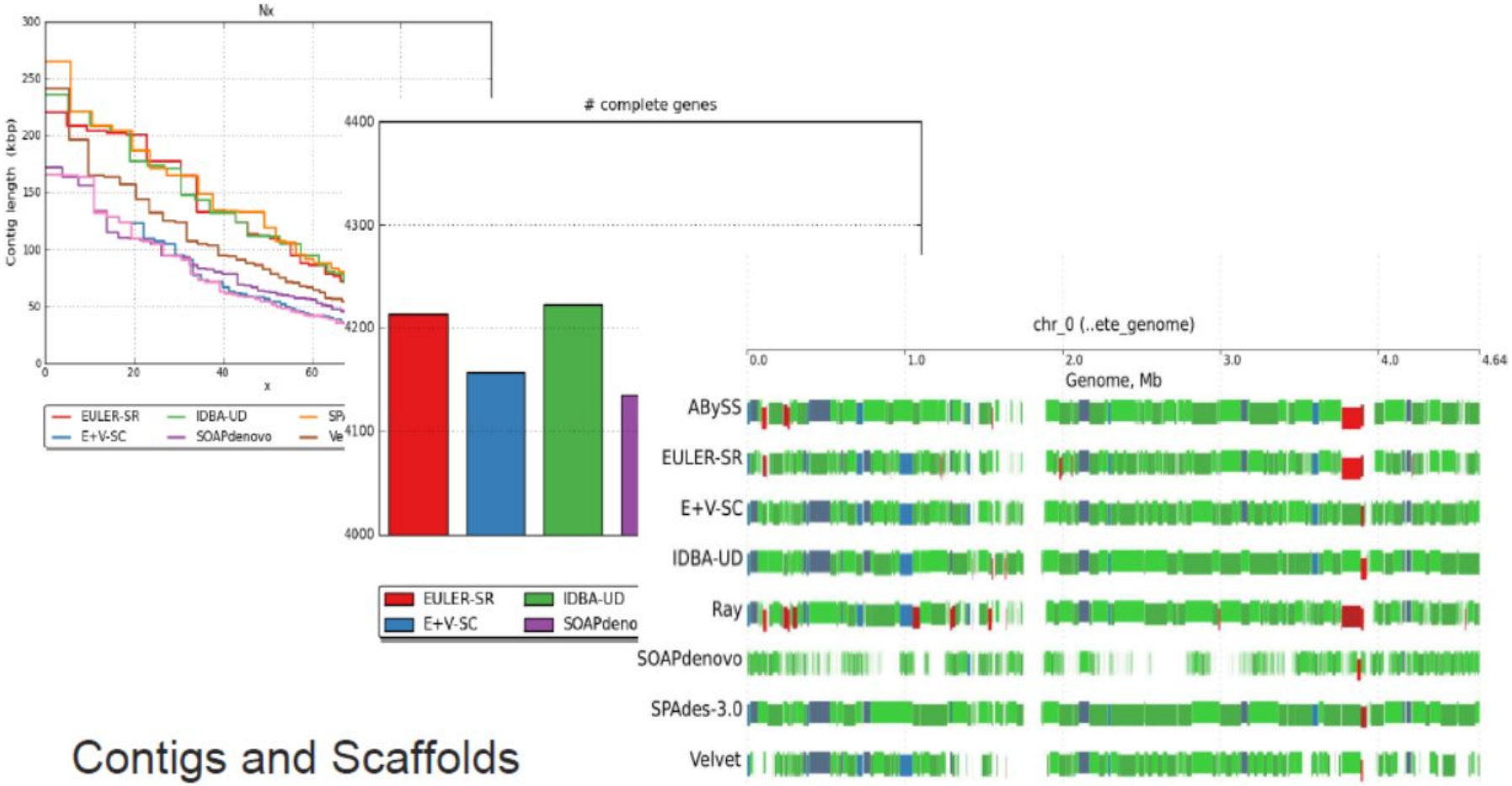
N50

Размер контига, который представляет из себя наиболее длинный контиг, такой, начиная с которого, все остальные контиги составляют не менее 50% длинны генома.



=> N50 = 4

QUAST - QUality ASsesment Tool for Genome Assemblies



Contigs and Scaffolds

http://quast.bioinf.spbau.ru/

GENOMICS

Genome Project Standards in a New Era of Sequencing

More detailed sequence standards that keep up with revolutionary sequencing technologies will aid the research community in evaluating data.

P. S. G. Chain,[1,2,3*†§] D. V. Grafham,[4†§] R. S. Fulton,[5†] M. G. FitzGerald,[5†] J. Hostetler,[7†] D. Muzny,[8†] J. Ali,[9] B. Birren,[6] D. C. Bruce,[1,10] C. Buhay,[8] J. R. Cole,[3] Y. Ding,[8] S. Dugan,[8] D. Field,[11] G. M. Garrity,[3] R. Gibbs,[8] T. Graves,[5] C. S. Han,[1,10] S. H. Harrison,[3*] S. Highlander,[8] P. Hugenholtz,[1] H. M. Khouri,[12] C. D. Kodira,[6*] E. Kolker,[13,14] N. C. Kyrpides,[1] D. Lang,[12] A. Lapidus,[1] S. A. Malfatti,[12] V. Markowitz,[15] T. Metha,[6] K. E. Nelson,[7] J. Parkhill,[4] J. Schmutz,[17] S. Sozhamannan,[18] P. Sterk,[11] R. L. Strausberg,[7] G. Tiedje,[3] G. Weinstock,[5] A. Wollam,[5] Genomic Standards Consortium Jumpstart Consortium,[‡] J. C. Detter[16†‡]

For over a decade, genome sequences have adhered to only two standards that are relied on for purposes of sequence analysis by interested third parties (*1*, *2*). However, ongoing developments in revolutionary sequencing technologies have resulted in a redefinition of traditional whole-genome sequencing that requires reevaluation of such standards. With commercially available 454 pyrosequencing (followed by Illumina, SOLiD, and now Helicos), there has been an explosion of genomes sequenced under the moniker "draft"; however, these can be very poor quality genomes (due to inherent errors in the sequencing technologies, and the inabil-

draft genon
gating thro
ber and typ
trace repos
this availab
contigs or g
database. T
ity of such
some havoc
has contrib
nential leap
greatlyredu
time- and c
versus the
and finishin

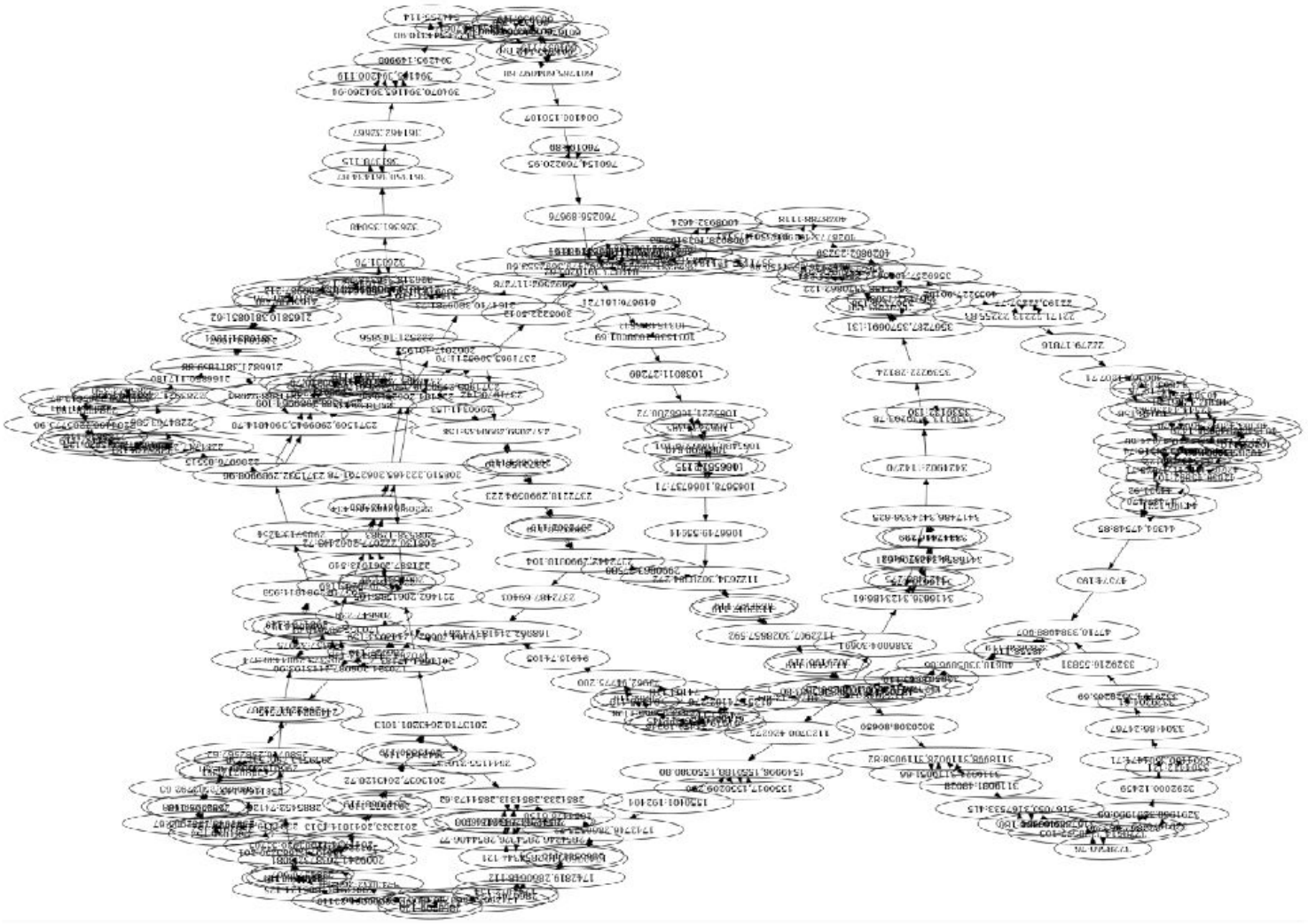
the figure, page 236); hence, there is an urgent need to distinguish good from poor data sets.

The sequencing institutes and consortia whom we represent believe that a new set of

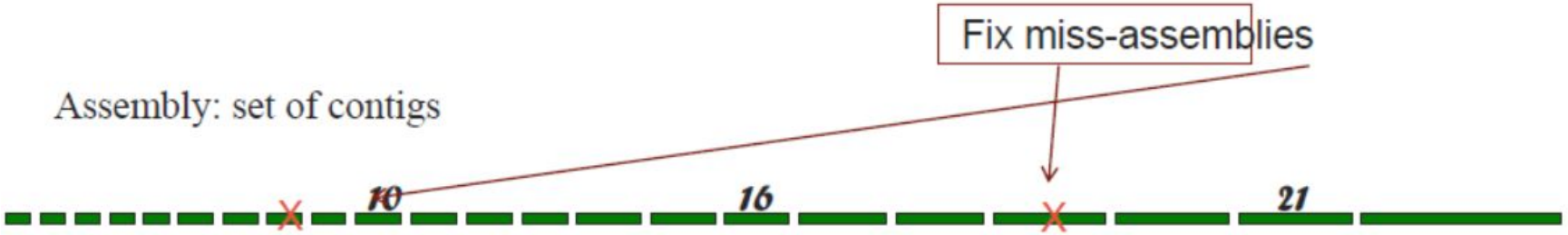
- *Standard Drafts*
- *High-Quality draft*
- *Improved High-Quality draft*
- *Annotation directed Improved Draft*
- *Noncontiguous Finished*
- ***Finished:*** refers to the current gold standard and represents genome sequences with less than 1 error per 100 000 base pairs and where each replicon is assembled into a single contiguous sequence. The Finished product is appropriate for all types of detailed analyses and acts as a high-quality reference genome for comparative purposes

Science 9 October 2009: Vol. 326 no. 5950 pp. 1 237

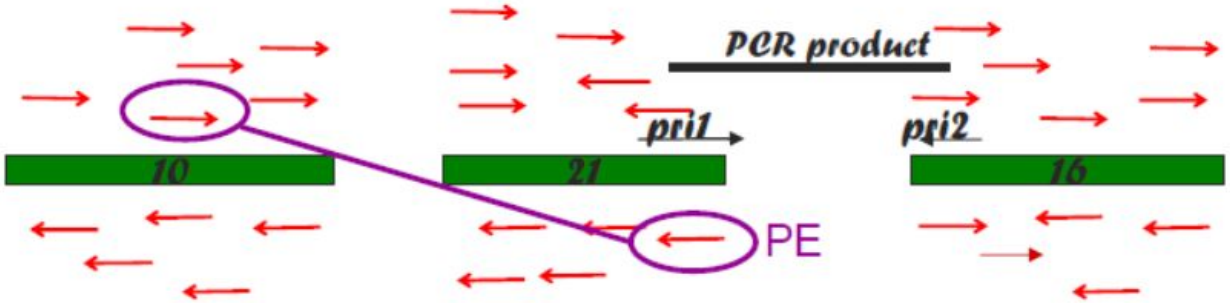
Реальные графы де Брюйна



Улучшение сборки генома



Ordered sets of contigs (scaffolds)

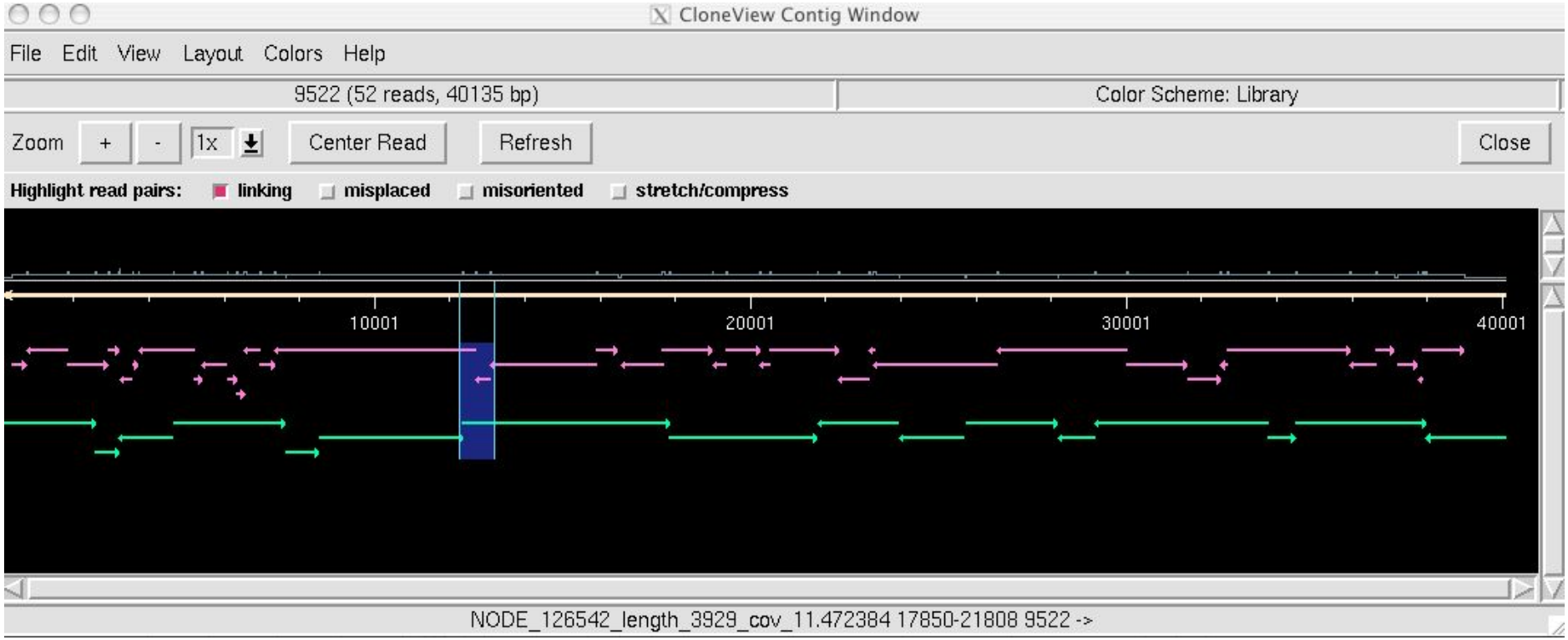


Clone walk
(Sanger lib)

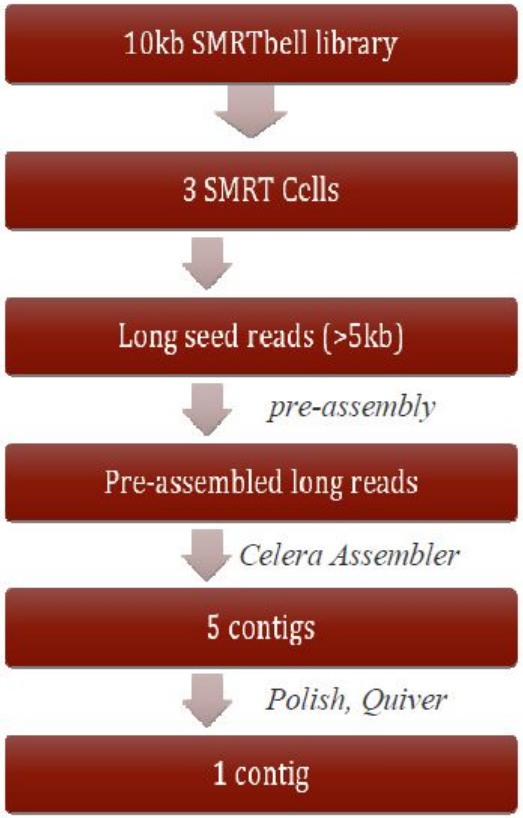
PCR - sequence

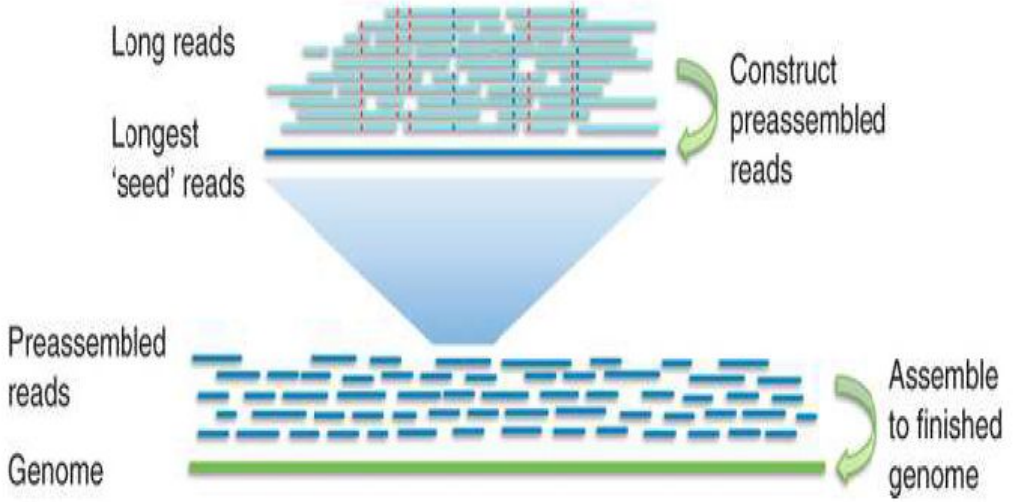
New technologies: no clones to walk off even if you can scaffold contigs

Гибридная сборка



Сборка на основе данных PacBio





Nat Methods. 2013 Jun;10(6):

Получение финишного генома

Historic timeline of sequencing of bacteria and archaea:

2002-2006	2006-2008	2008-2011	2011	2011-2013	2013-
Sanger	Sanger/454	454/Illumina	Illumina	Illumina/PacBio	PacBio
$50k	$35k	$10k	$1.5-3k	$5k	<$2k
49 contigs	22 contigs	44 contigs	69 contigs	6 contigs	1 contig

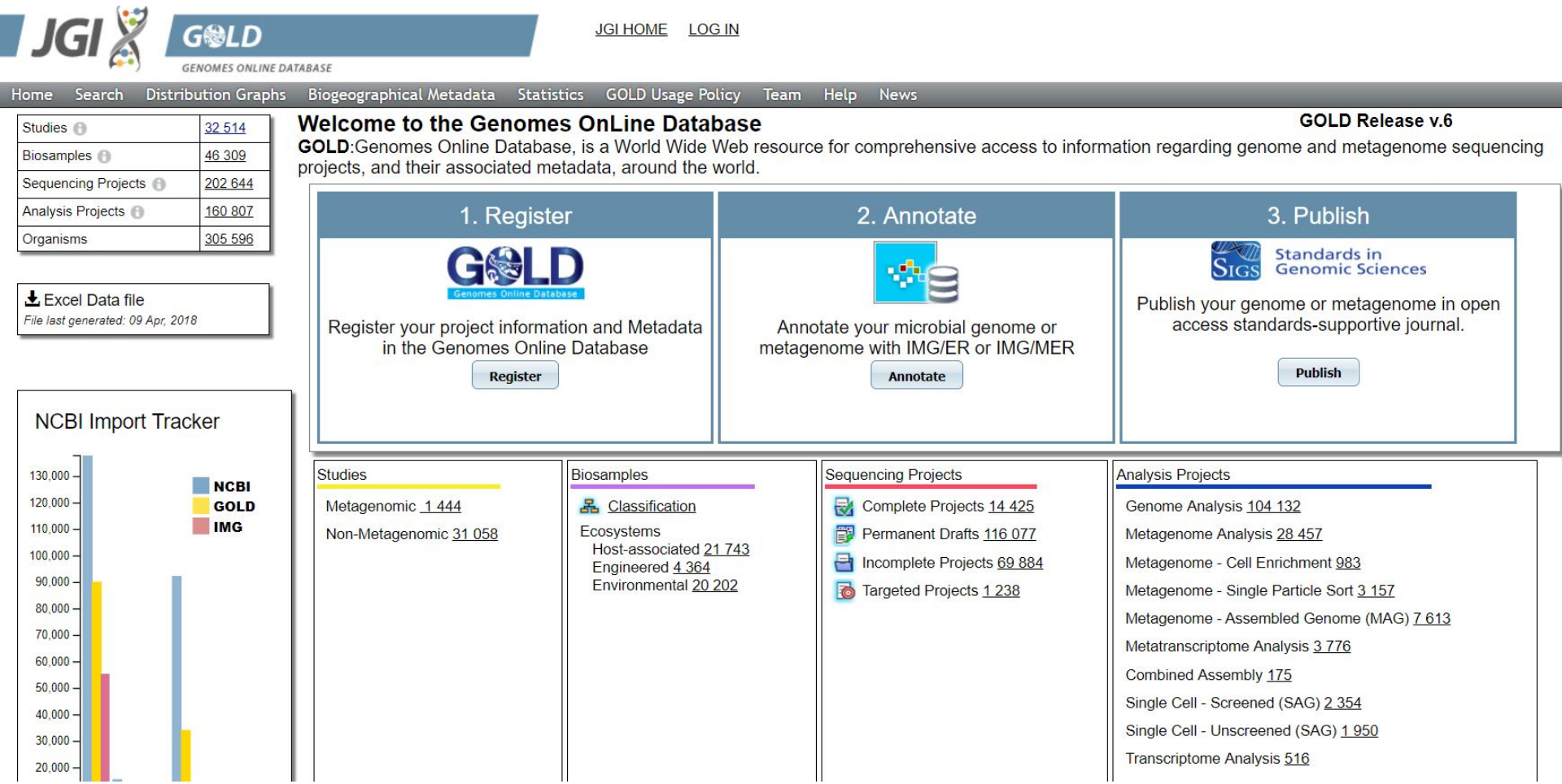
Manual Finishing ($45-85k/ genome) | No Finishing

Nat Methods. 2013 Jun;10(6):

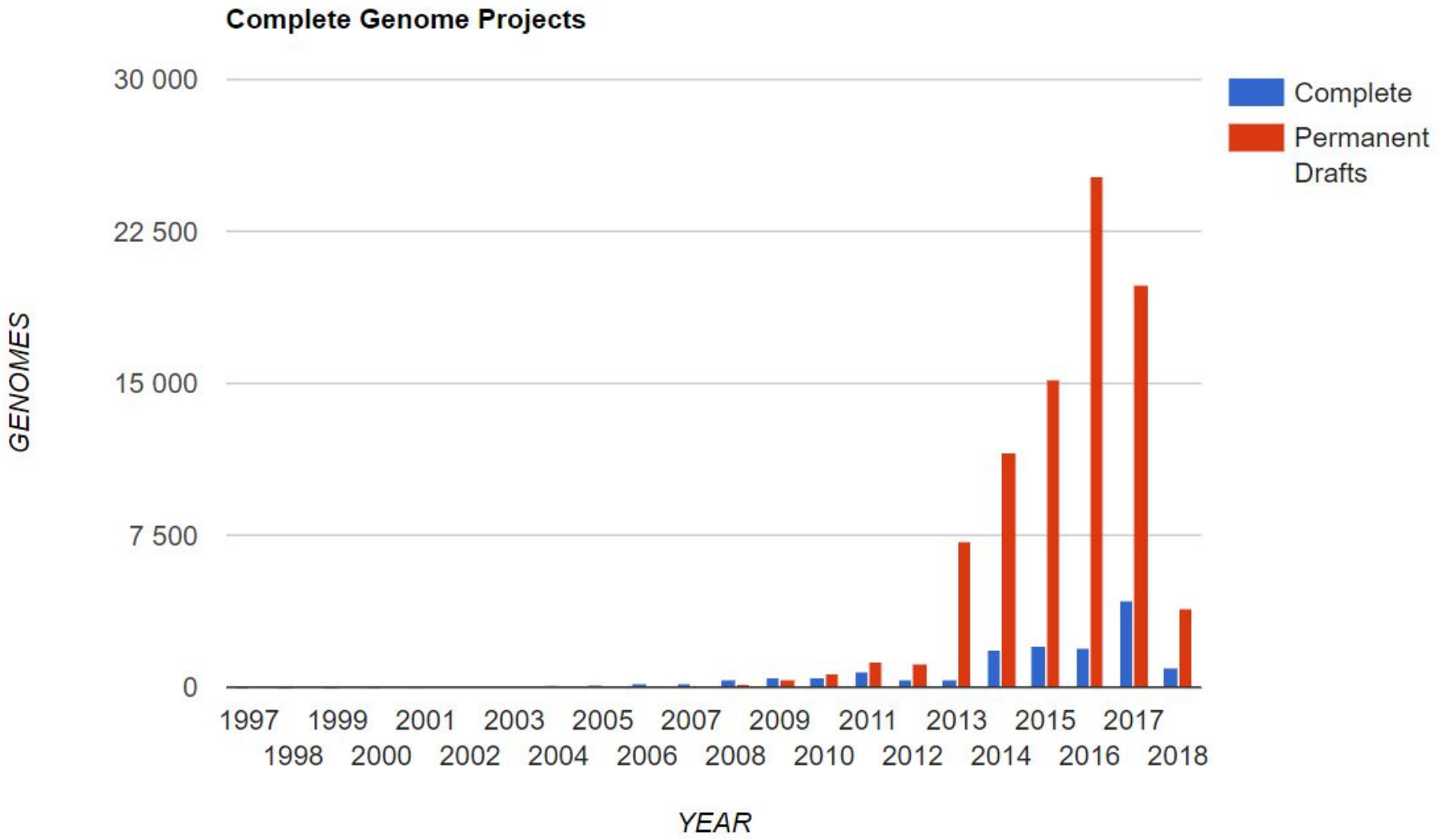
Зачем нужны финишные геномы?

- Функциональные геномные исследования требуют высококачественной, полной последовательности генома в качестве отправной точки
- Сравнительная геномика имеет смысл только в терминах полных последовательностей генома
- Исследования бактериальных геномов требует по крайней мере одной полной эталонной последовательности генома
- Финишные геномы помогают в идентификации источника вспышки инфекций и филогенетическом анализе
- Полный геном - это постоянный научный ресурс
- Полный геном человека является наилучшим источником для улучшения лечения пациентов (переход к персонализированной медицине)

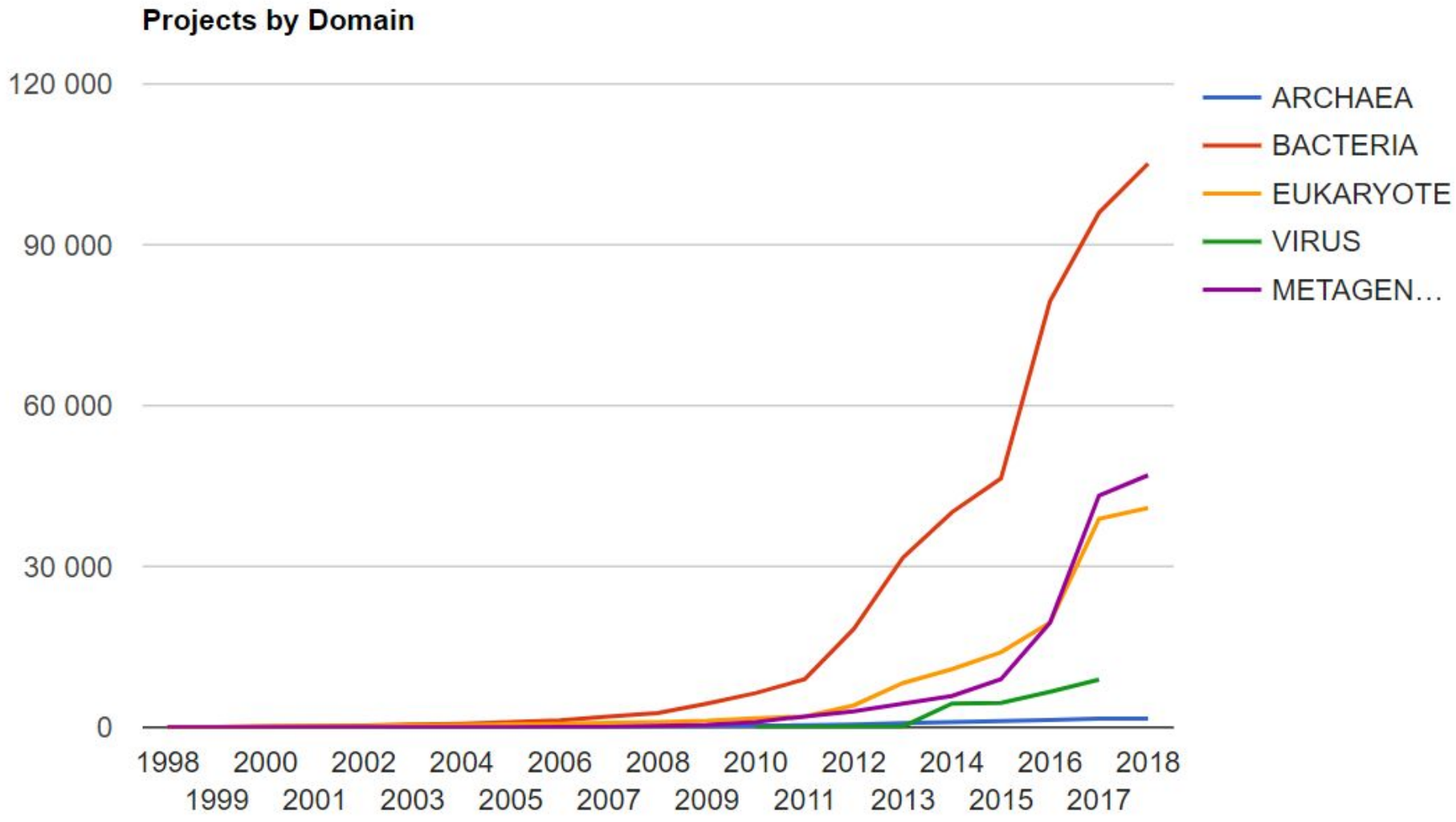
GOLD: Genomes OnLine Database



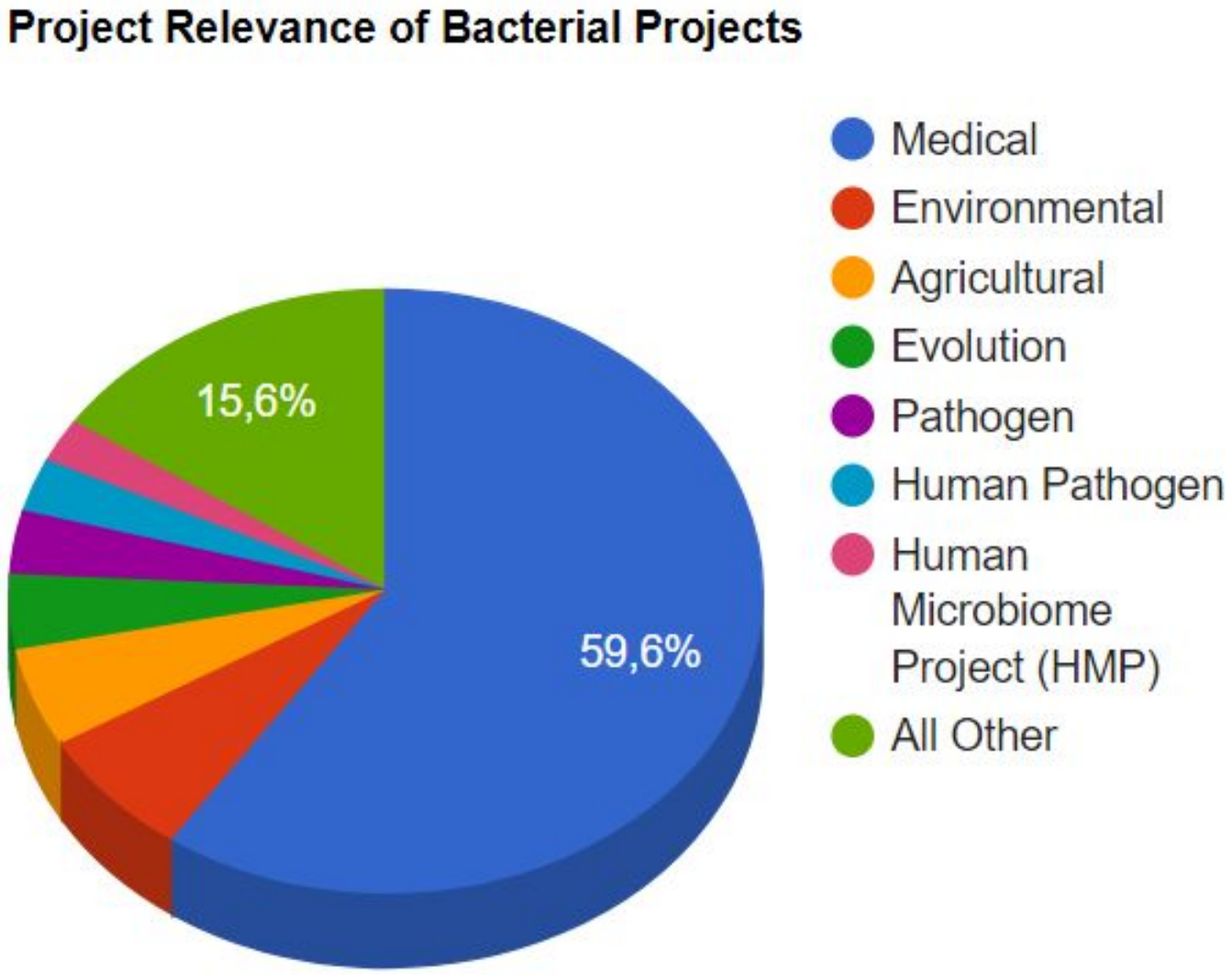
Статистика GOLD



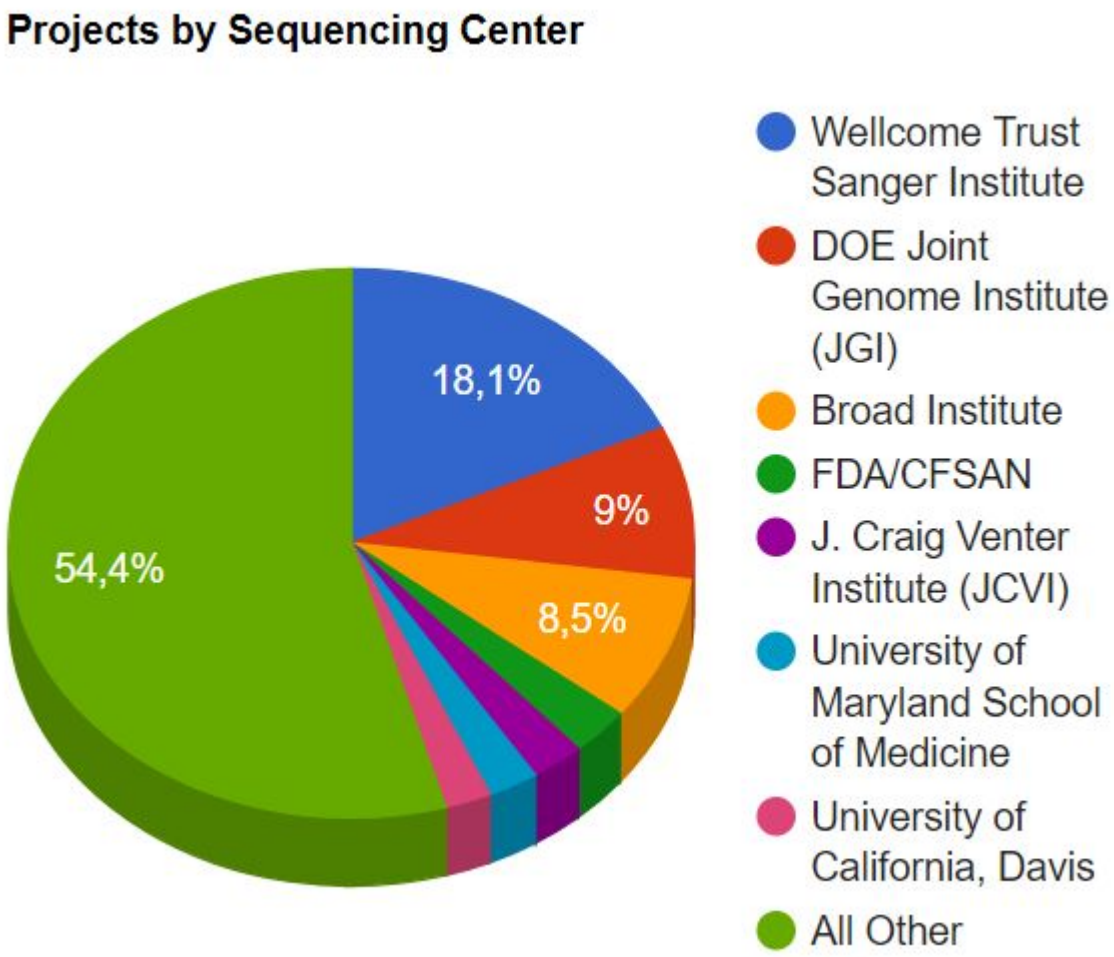
Статистика GOLD



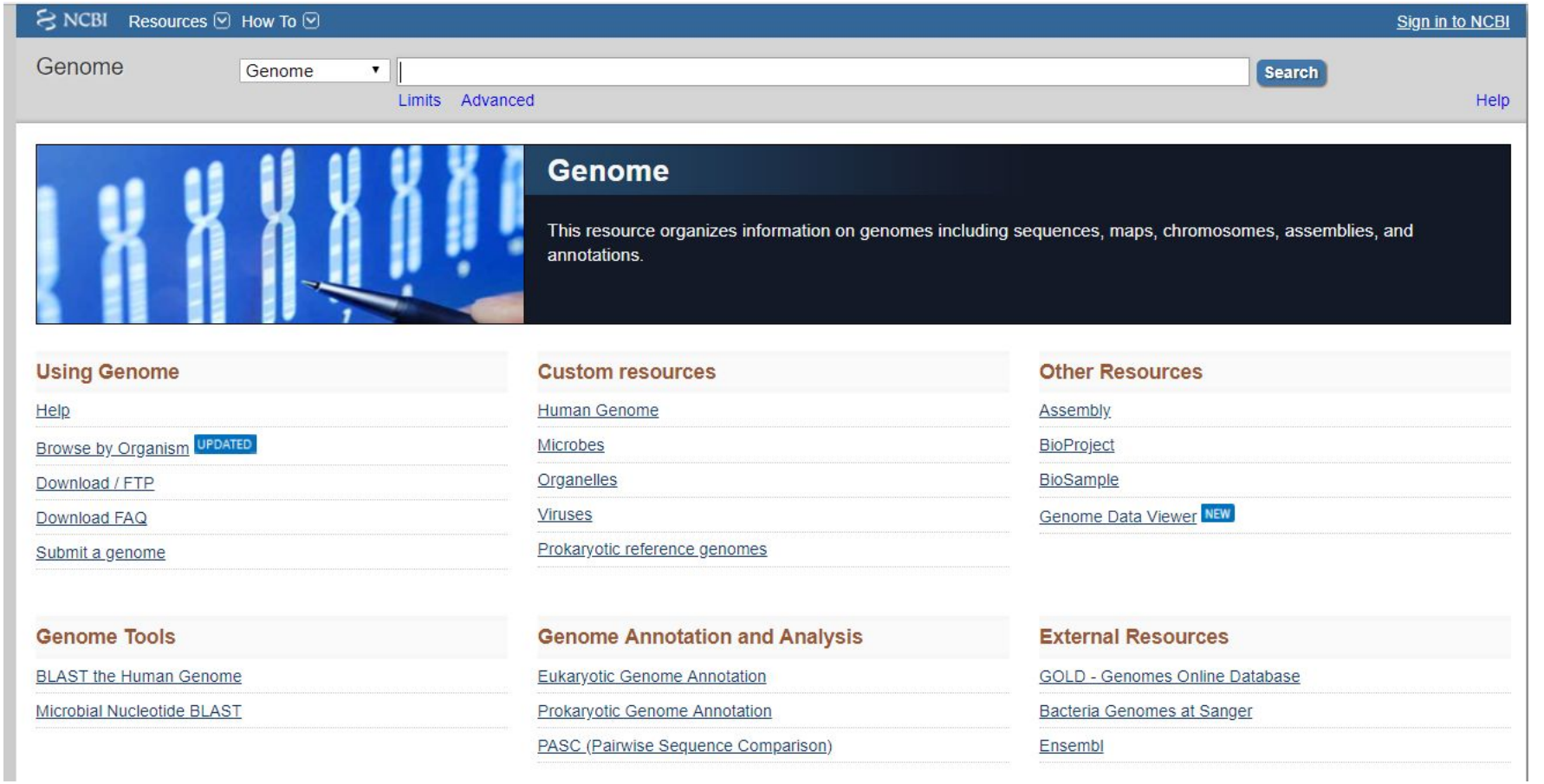
Статистика GOLD



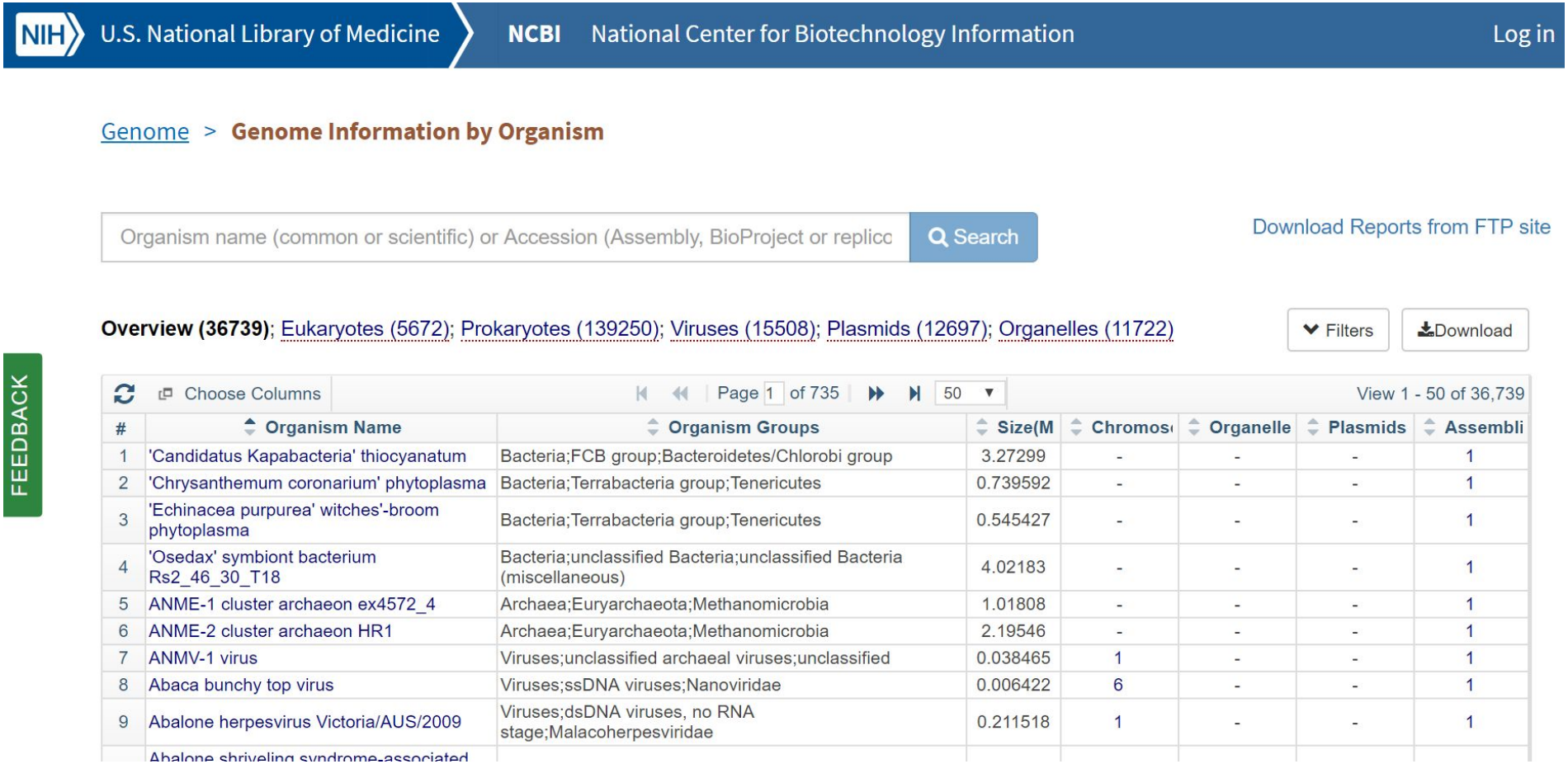
Статистика GOLD



NCBI Genome

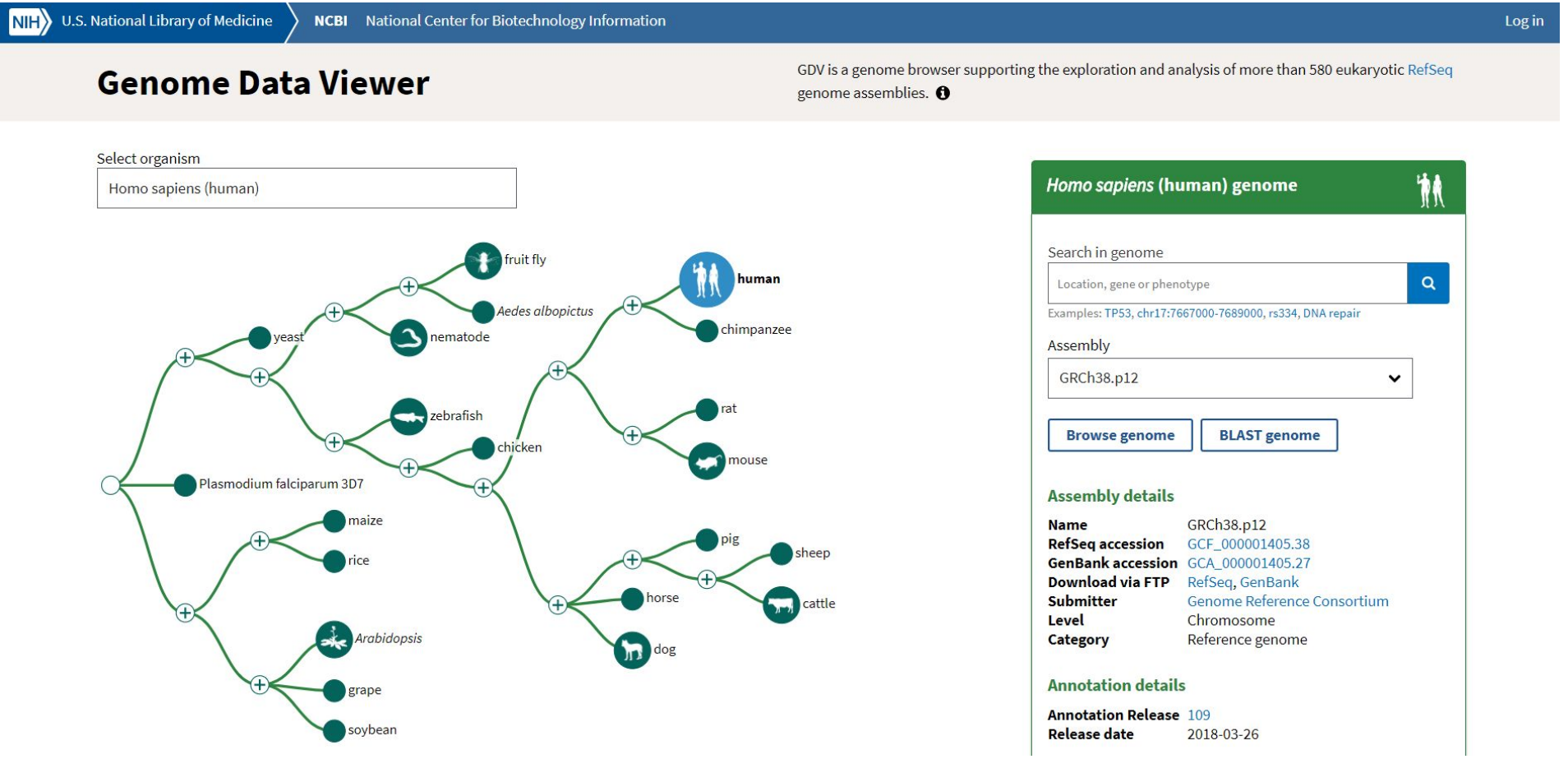

NCBI
Resources
How To
Sign in to NCBI
Genome
Genome
Search
Limits
Advanced
Help
Genome
This resource organizes information on genomes including sequences, maps, chromosomes, assemblies, and annotations.
Using Genome
Help
Browse by Organism
UPDATED
Download / FTP
Download FAQ
Submit a genome
Custom resources
Human Genome
Microbes
Organelles
Viruses
Prokaryotic reference genomes
Other Resources
Assembly
BioProject
BioSample
Genome Data Viewer
NEW
Genome Tools
BLAST the Human Genome
Microbial Nucleotide BLAST
Genome Annotation and Analysis
Eukaryotic Genome Annotation
Prokaryotic Genome Annotation
PASC (Pairwise Sequence Comparison)
External Resources
GOLD - Genomes Online Database
Bacteria Genomes at Sanger
Ensembl

NCBI Genome

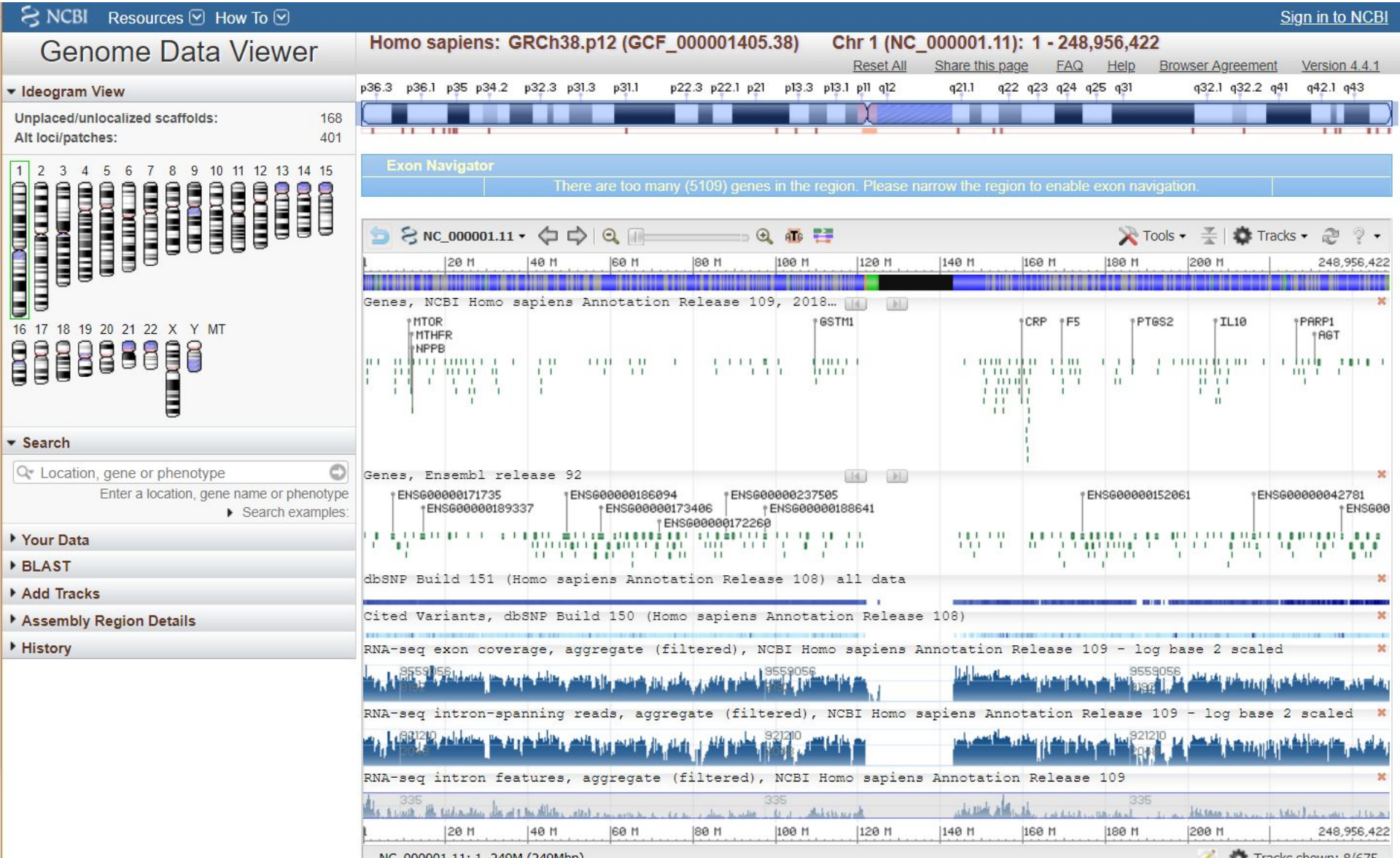


#	Organism Name	Organism Groups	Size(M	Chromos	Organelle	Plasmids	Assembli
1	'Candidatus Kapabacteria' thiocyanatum	Bacteria;FCB group;Bacteroidetes/Chlorobi group	3.27299	-	-	-	1
2	'Chrysanthemum coronarium' phytoplasma	Bacteria;Terrabacteria group;Tenericutes	0.739592	-	-	-	1
3	'Echinacea purpurea' witches'-broom phytoplasma	Bacteria;Terrabacteria group;Tenericutes	0.545427	-	-	-	1
4	'Osedax' symbiont bacterium Rs2_46_30_T18	Bacteria;unclassified Bacteria;unclassified Bacteria (miscellaneous)	4.02183	-	-	-	1
5	ANME-1 cluster archaeon ex4572_4	Archaea;Euryarchaeota;Methanomicrobia	1.01808	-	-	-	1
6	ANME-2 cluster archaeon HR1	Archaea;Euryarchaeota;Methanomicrobia	2.19546	-	-	-	1
7	ANMV-1 virus	Viruses;unclassified archaeal viruses;unclassified	0.038465	1	-	-	1
8	Abaca bunchy top virus	Viruses;ssDNA viruses;Nanoviridae	0.006422	6	-	-	1
9	Abalone herpesvirus Victoria/AUS/2009	Viruses;dsDNA viruses, no RNA stage;Malacoherpesviridae	0.211518	1	-	-	1
	Abalone shriveling syndrome-associated						

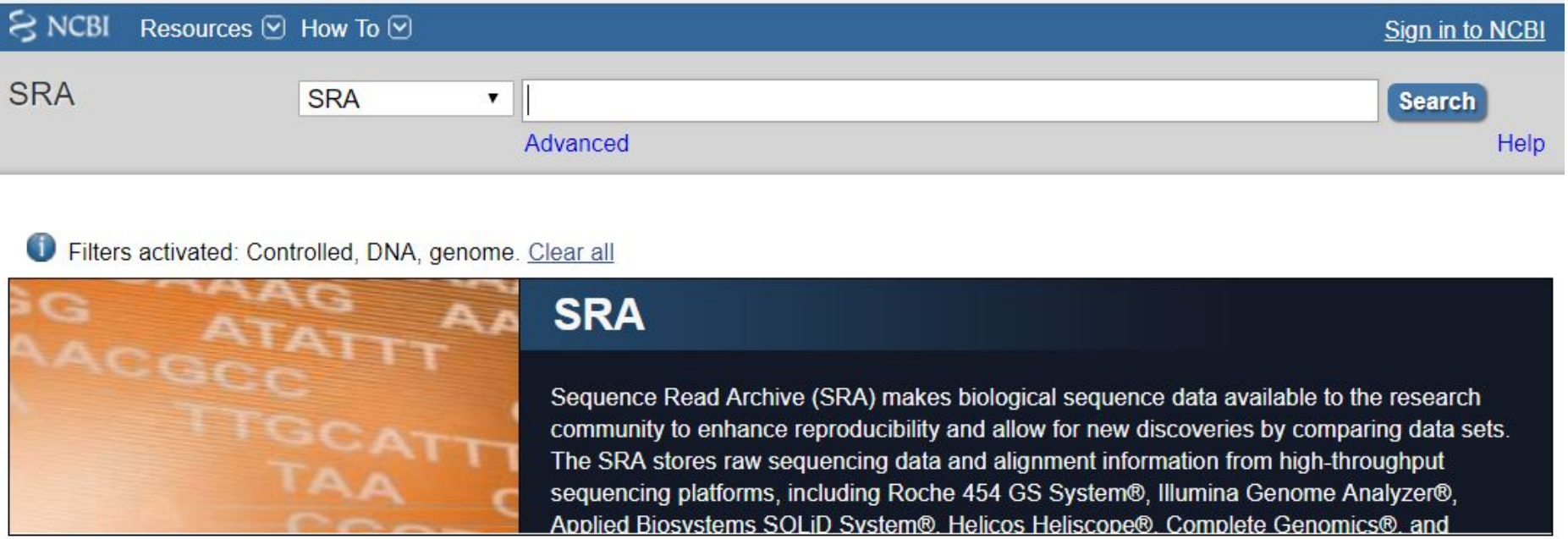
NCBI Genome


NIH U.S. National Library of Medicine
NCBI National Center for Biotechnology Information
Log in
Genome Data Viewer
GDV is a genome browser supporting the exploration and analysis of more than 580 eukaryotic RefSeq genome assemblies.
Select organism
Homo sapiens (human)
fruit fly
Aedes albopictus
nematode
yeast
zebrafish
chicken
human
chimpanzee
rat
mouse
Plasmodium falciparum 3D7
maize
rice
Arabidopsis
grape
soybean
pig
sheep
cattle
horse
dog
Homo sapiens (human) genome
Search in genome
Location, gene or phenotype
Examples: TP53, chr17:7667000-7689000, rs334, DNA repair
Assembly
GRCh38.p12
Browse genome
BLAST genome
Assembly details
Name GRCh38.p12
RefSeq accession GCF_000001405.38
GenBank accession GCA_000001405.27
Download via FTP RefSeq, GenBank
Submitter Genome Reference Consortium
Level Chromosome
Category Reference genome
Annotation details
Annotation Release 109
Release date 2018-03-26

NCBI Genome


NCBI Resources How To
Sign in to NCBI
Genome Data Viewer
Homo sapiens: GRCh38.p12 (GCF_000001405.38) Chr 1 (NC_000001.11): 1 - 248,956,422
Reset All Share this page FAQ Help Browser Agreement Version 4.4.1
Ideogram View
Unplaced/unlocalized scaffolds: 168
Alt loci/patches: 401
Exon Navigator
There are too many (5109) genes in the region. Please narrow the region to enable exon navigation.
NC_000001.11
Tools
Tracks
Genes, NCBI Homo sapiens Annotation Release 109, 2018…
Genes, Ensembl release 92
dbSNP Build 151 (Homo sapiens Annotation Release 108) all data
Cited Variants, dbSNP Build 150 (Homo sapiens Annotation Release 108)
RNA-seq exon coverage, aggregate (filtered), NCBI Homo sapiens Annotation Release 109 - log base 2 scaled
RNA-seq intron-spanning reads, aggregate (filtered), NCBI Homo sapiens Annotation Release 109 - log base 2 scaled
RNA-seq intron features, aggregate (filtered), NCBI Homo sapiens Annotation Release 109
Search
Location, gene or phenotype
Enter a location, gene name or phenotype
Search examples:
Your Data
BLAST
Add Tracks
Assembly Region Details
History

NCBI SRA database



NCBI SRA database

NCBI Site map | All databases | Search

Sequence Read Archive

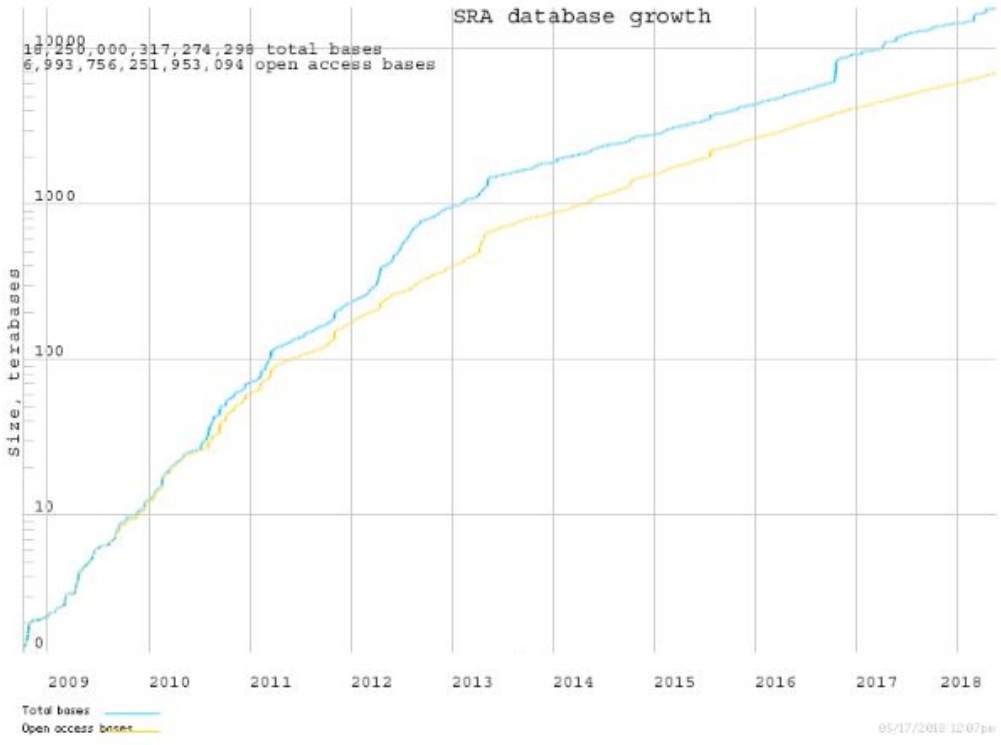
Main | Browse | Search | Download | Submit | Software | Trace Archive | Trace Assembly | Trace BLAST

Overview

The Sequence Read Archive (SRA) stores raw sequence data from "next-generation" sequencing technologies including Illumina, 454, IonTorrent, Complete Genomics, PacBio and OxfordNanopores. In addition to raw sequence data, SRA now stores alignment information in the form of read placements on a reference sequence.

SRA is NIH's primary archive of high-throughput sequencing data and is part of the international partnership of archives (INSDC) at the NCBI, the European Bioinformatics Institute and the DNA Database of Japan. Data submitted to any of the three organizations are shared among them.

Please check SRA Overview for more information.



Save in CSV format

Submitting to SRA

Making data available to the research community enhances reproducibility and allows for new discovery by comparing data sets.

- Submission Quick Start
- Frequently Asked Questions and Troubleshooting
- Log in to Submission Portal (for submitting sequence data)
- Log in to SRA (for updating and troubleshooting submissions)

Using SRA Data with SRA Toolkit

Use SRA data to validate experimental results, increase sample sizes, determine variance and open up new avenues of research.

- SRA Download Guide
- SRA Toolkit Usage Guide
- Software Download
- Get sources code on GitHub (for developers using SRA)