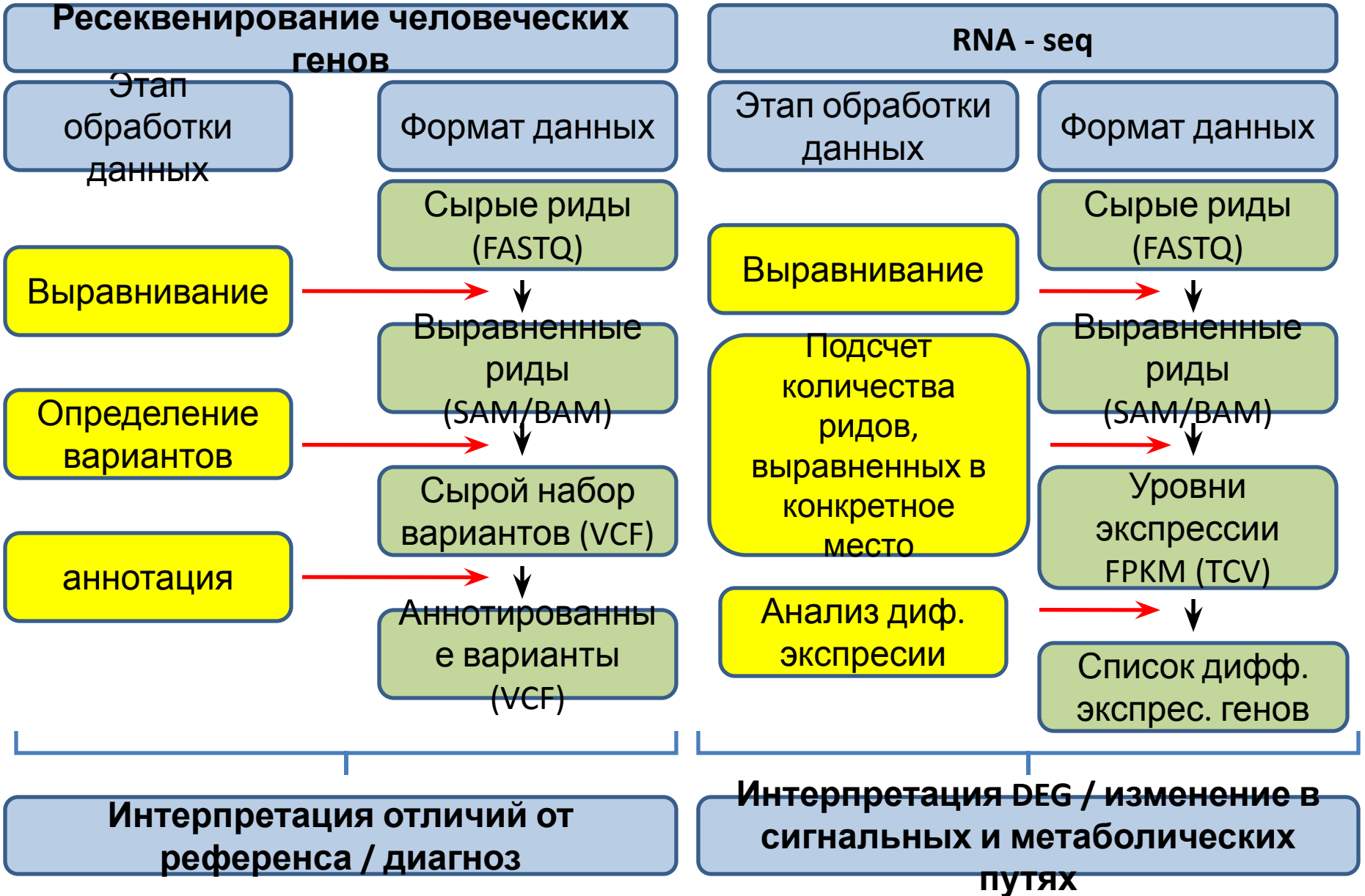
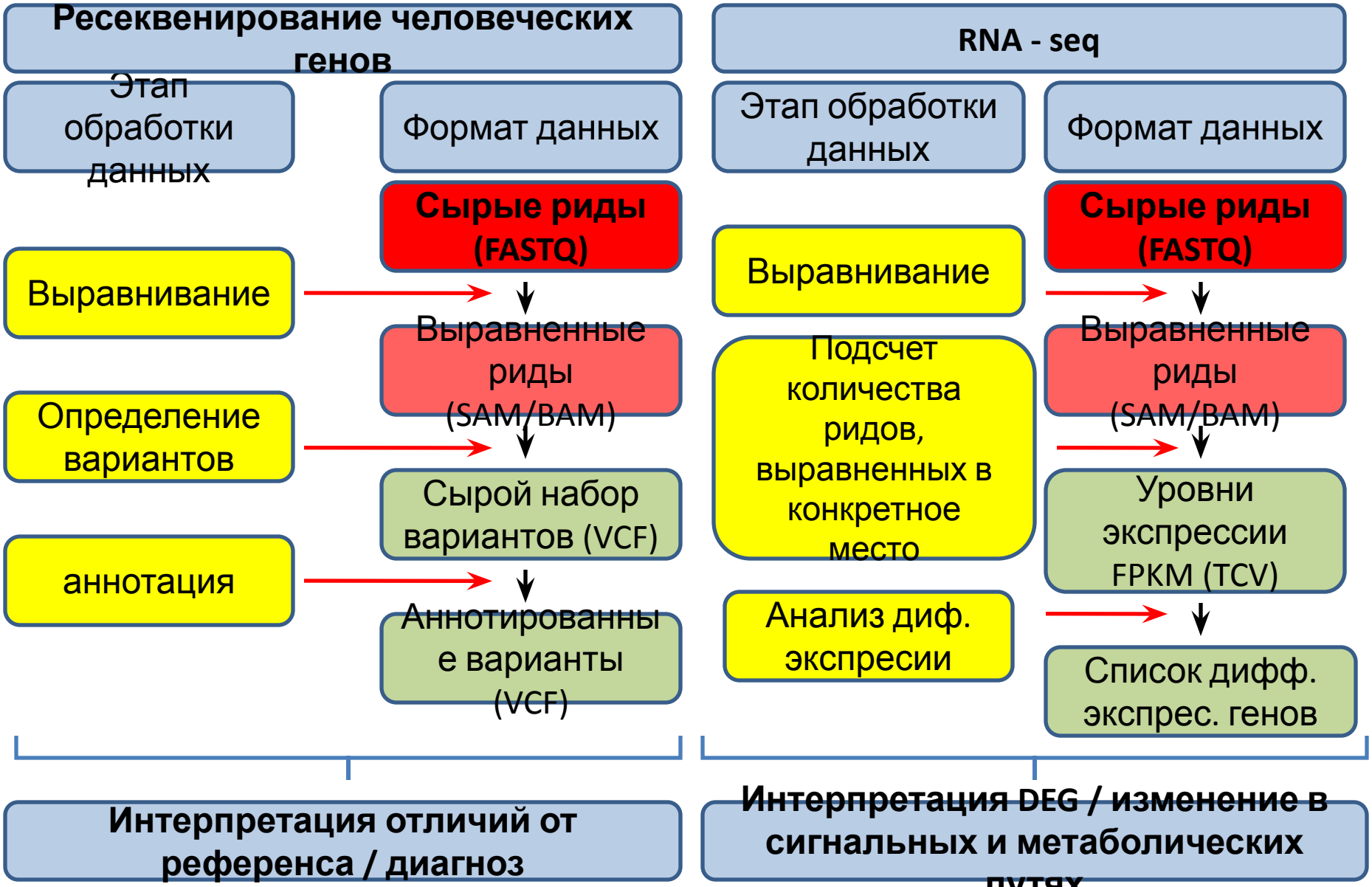
Оценка качества прочтений NGS / FastQC

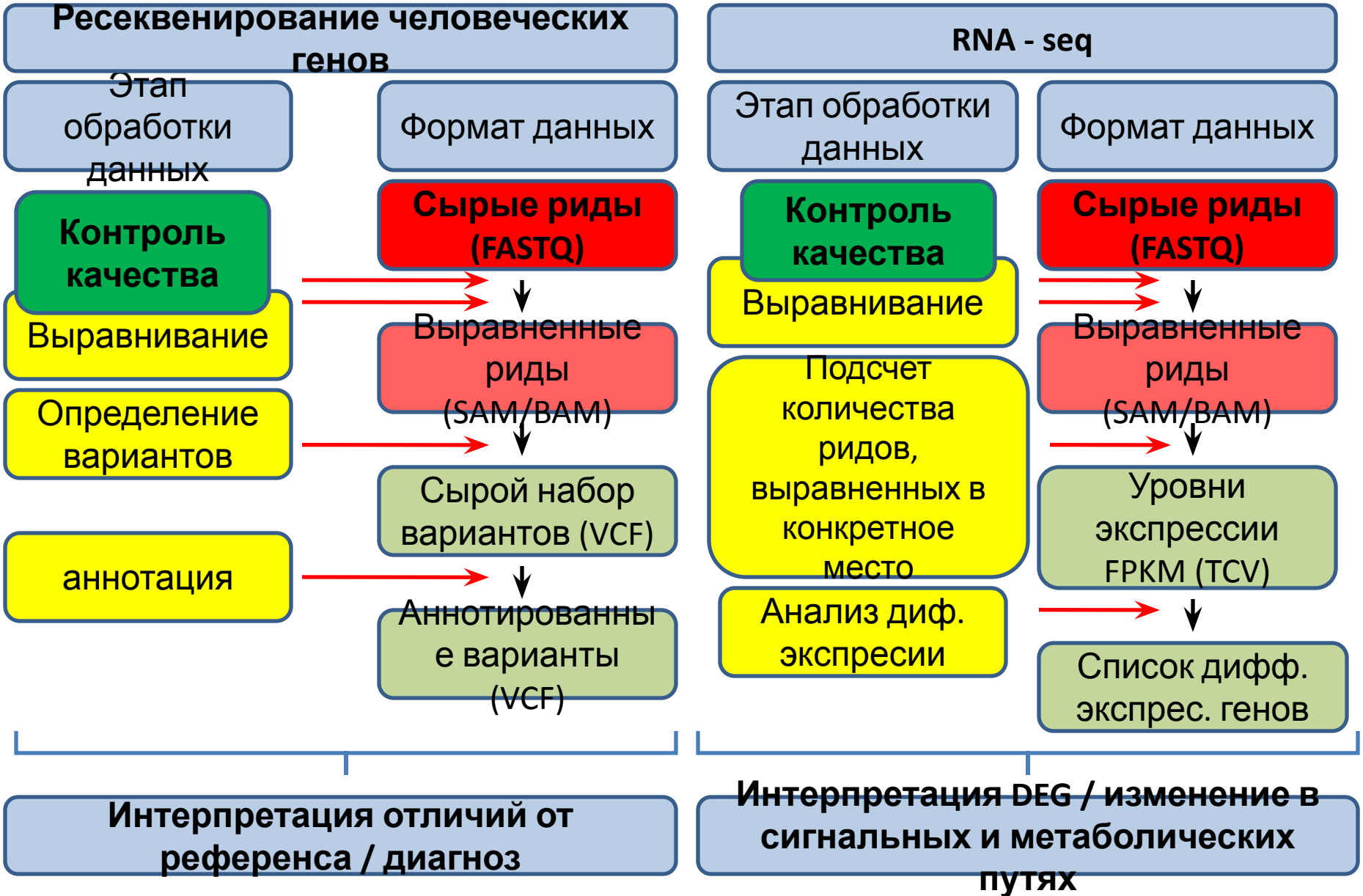
Примеры экспериментов в основе которых лежит NGS



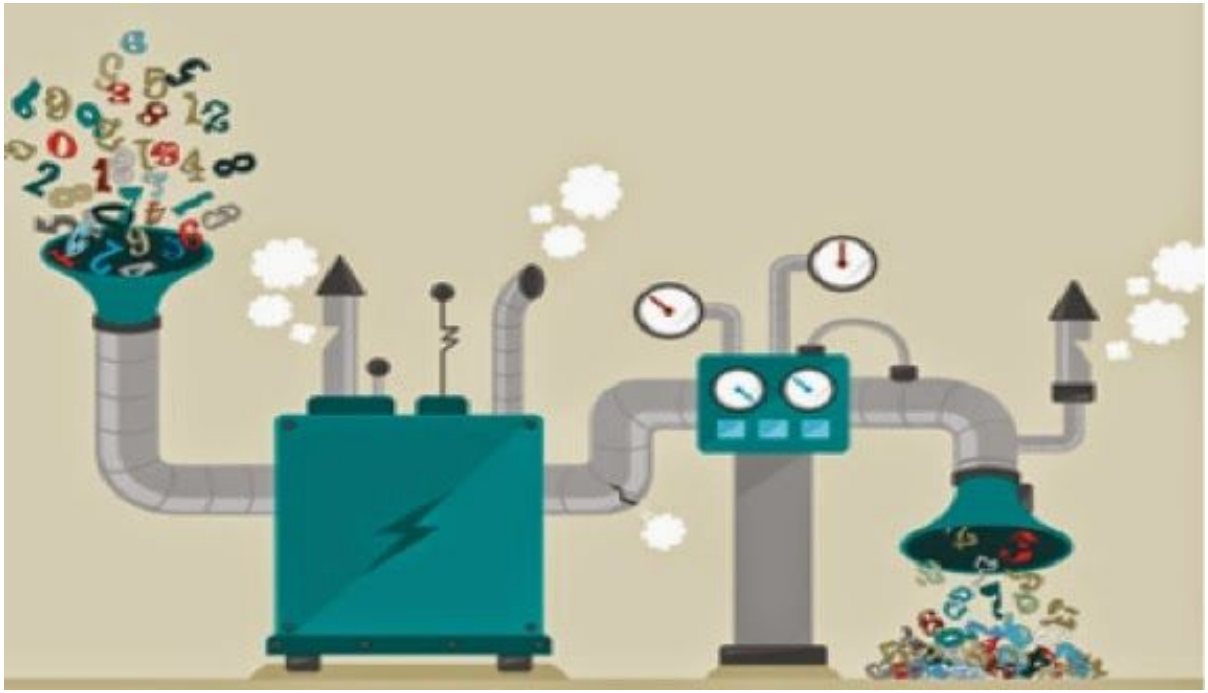
Сырые данные на выходе у секвенатора



Контроль качества обязательный этап

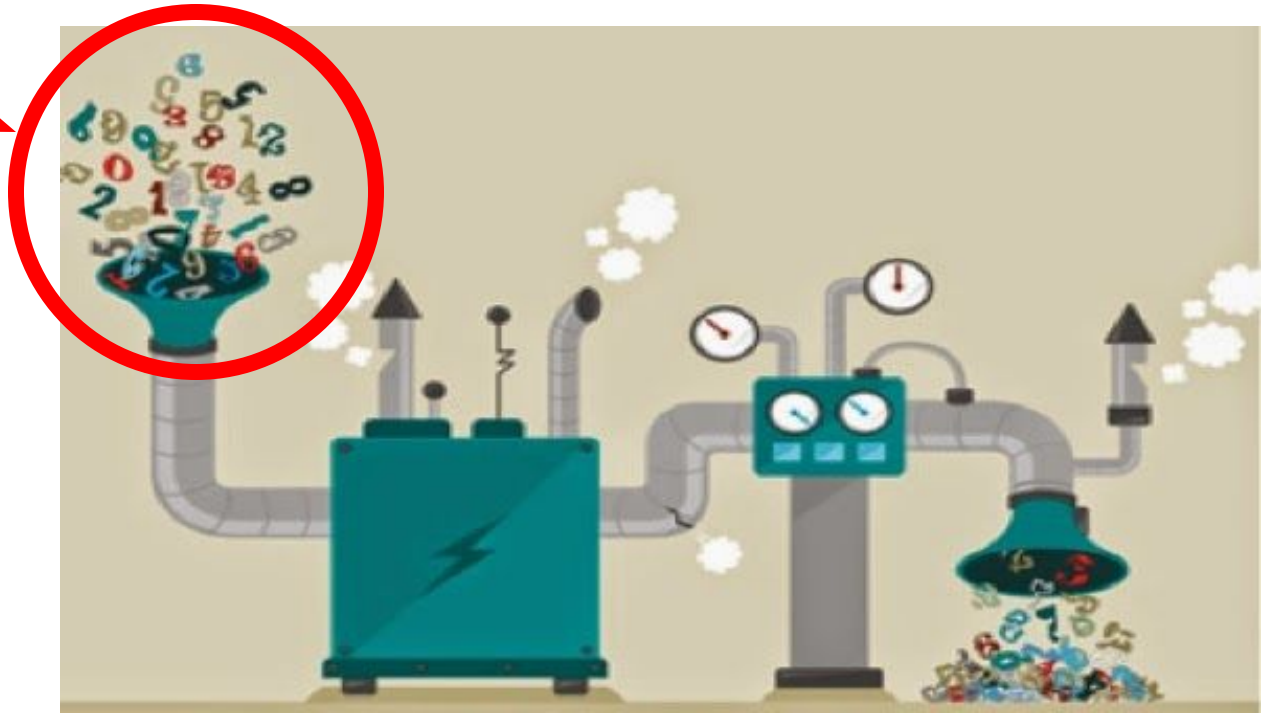


А на кой черт оно собственно надо?



А на кой черт оно собственно надо?

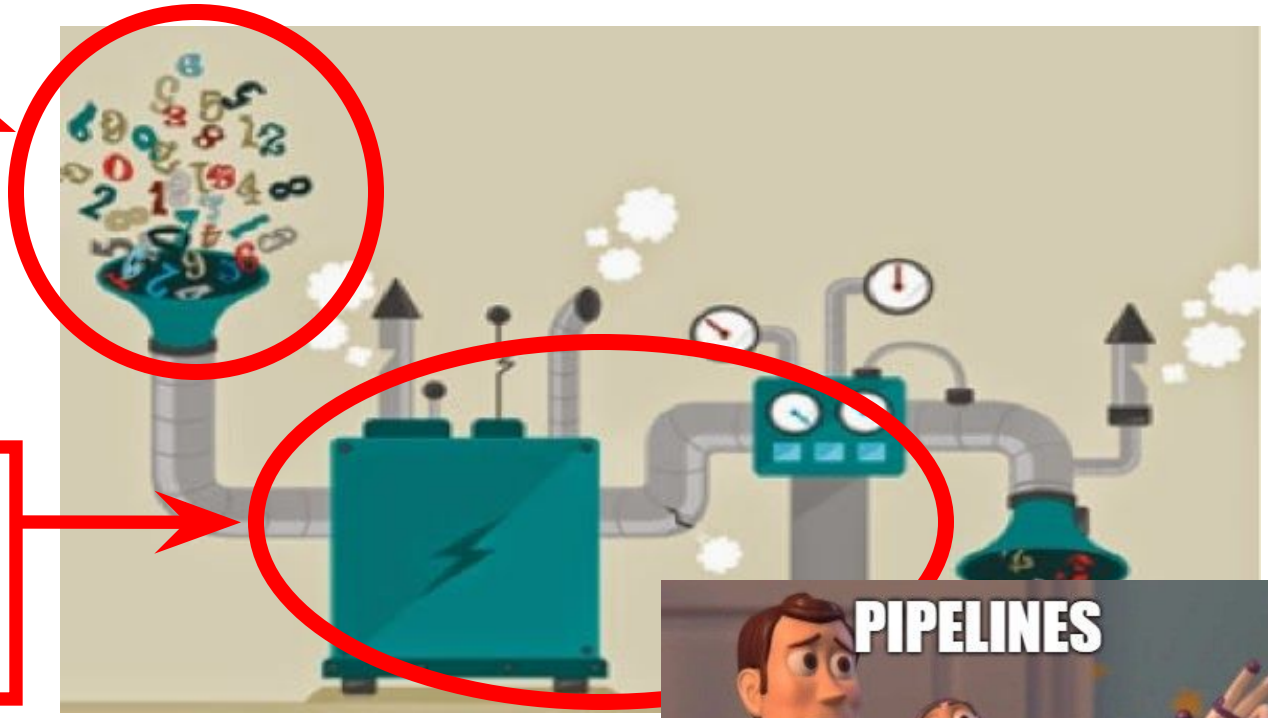
Сырые данные, полученные в ходе работы секвенатора. Их вы подает на вход вашего pipeline. Часто вы начинаете с **FASTQ** файлов.



А на кой черт оно собственно надо?

Сырые данные, полученные в ходе работы секвенатора. Их вы подает на вход вашего pipeline. Часто вы начинаете с **FASTQ** файлов.

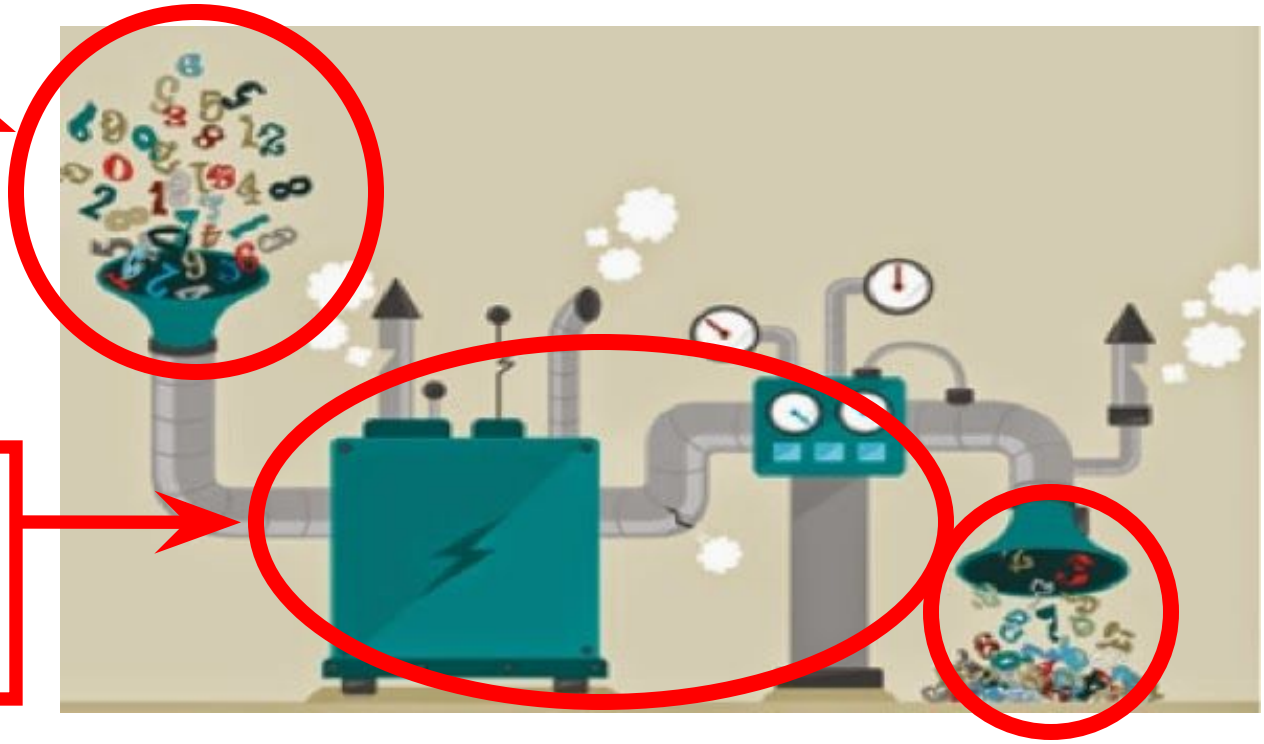
Ваш **pipeline** – последовательность инструментов, которыми вы обрабатываете данные



А на кой черт оно собственно надо?

Сырые данные, полученные в ходе работы секвенатора. Их вы подает на вход вашего pipeline. Часто вы начинаете с **FASTQ** файлов.

Ваш **pipeline** – последовательность инструментов, которыми вы обрабатываете данные

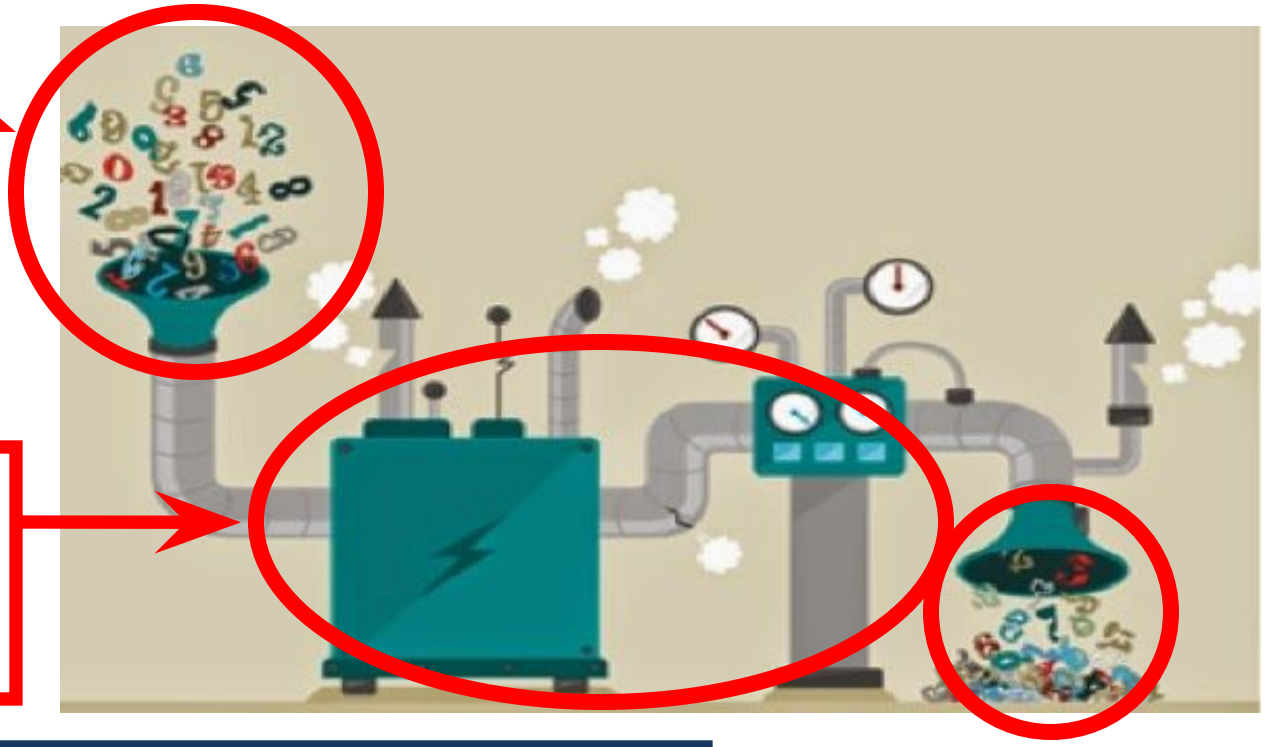


Результат анализа

А на кой черт оно собственно надо?

Сырые данные, полученные в ходе работы секвенатора. Их вы подает на вход вашего pipeline. Часто вы начинаете с **FASTQ** файлов.

Ваш **pipeline** – последовательность инструментов, которыми вы обрабатываете данные



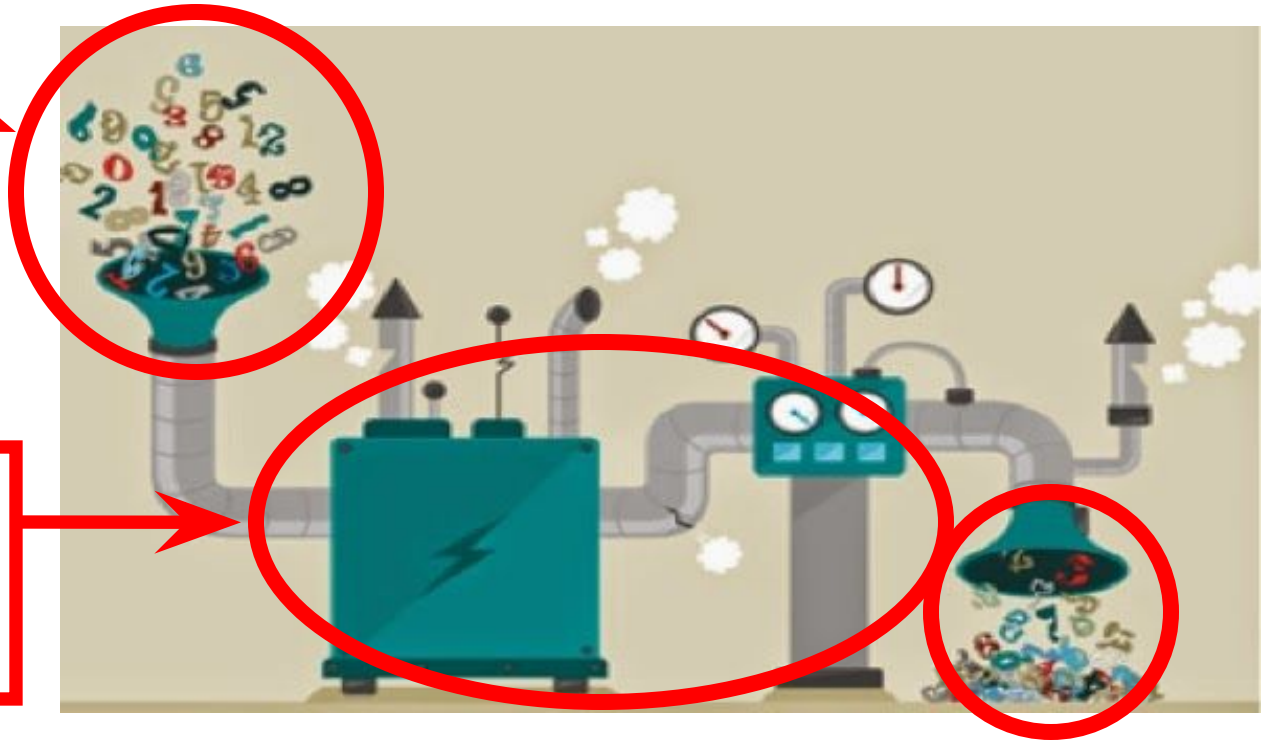
Результат анализа

Основное правило: **GIGO** - garbage **in**, garbage **out**

А на кой черт оно собственно надо?

Сырые данные, полученные в ходе работы секвенатора. Их вы подает на вход вашего pipeline. Часто вы начинаете с **FASTQ** файлов.

Ваш **pipeline** – последовательность инструментов, которыми вы обрабатываете данные



Результат анализа

Основной вывод: обязателен контроль качества входных данных

FASTQ формат

@cluster_2:UMI_ATTCCG

TTTCCGGGGCACATAATCTTCAGCCGGGCGC

+

9C;=;=<9**@4868**>9:67AA<9>65<=>591

FASTQ формат

Идентификатор последовательности с необязательным описанием. Начинается с символа @

@cluster_2:UMI_ATTCCG

TTTCCGGGGCACATAATCTTCAGCCGGGCGC

+

9C;=;=<9**@4868**>9:67AA<9>65<=>591

FASTQ формат

Идентификатор последовательности с необязательным описанием. Начинается с символа @

Последовательность “прочтенных” нуклеотидов

@cluster_2:UMI_ATTCCG

TTTCCGGGGCACATAATCTTCAGCCGGGCGC

+

9C;=;=<9**@4868**>9:67AA<9>65<=>591

FASTQ формат

Идентификатор последовательности с необязательным описанием. Начинается с символа @

Последовательность “прочтенных” нуклеотидов

Служебная строка

@cluster_2:UMI_ATTCCG

TTTCCGGGGCACATAATCTTCAGCCGGGCGC

+

9C;=;=<9**@4868**>9:67AA<9>65<=>591

FASTQ формат

@cluster_2:UMI_ATTCCG

TTTCCGGGGCACATAATCTTCAGCCGGGCGC

+

9C;=;=<9**@4868**>9:67AA<9>65<=>591

Идентификатор последовательности с необязательным описанием. Начинается с символа @

Последовательность “прочтенных” нуклеотидов

Служебная строка

Строка, содержащая значения качества (Q score) для нуклеотидов из второй строки.

FASTQ формат

@cluster_2:UMI_ATTCCG

TTTCCGGGGCACATAATCTTCAGCCGGGCGC

+

9C;=;=<9@4868>9:67AA<9>65<=>591

Идентификатор последовательности с необязательным описанием. Начинается с символа @

Последовательность “прочтенных” нуклеотидов

Служебная строка

Строка, содержащая значения качества (Q score) для нуклеотидов из второй строки.

Q score – показатель, зависящий от вероятности неправильного прочтения данного нуклеотида. Существует несколько вариантов определения Q score в зависимости от платформы, на которой осуществлялось секвенирование.

$Q_{sanger} = -10\log_{10}(p)$

$Q_{solexa} = -10\log_{10}(p/(1-p))$

Где p – вероятность, что соответствующий нуклеотид определен неверно.

Q score кодируется символами ASKII

Таблица 2

Символы с кодами 32–127

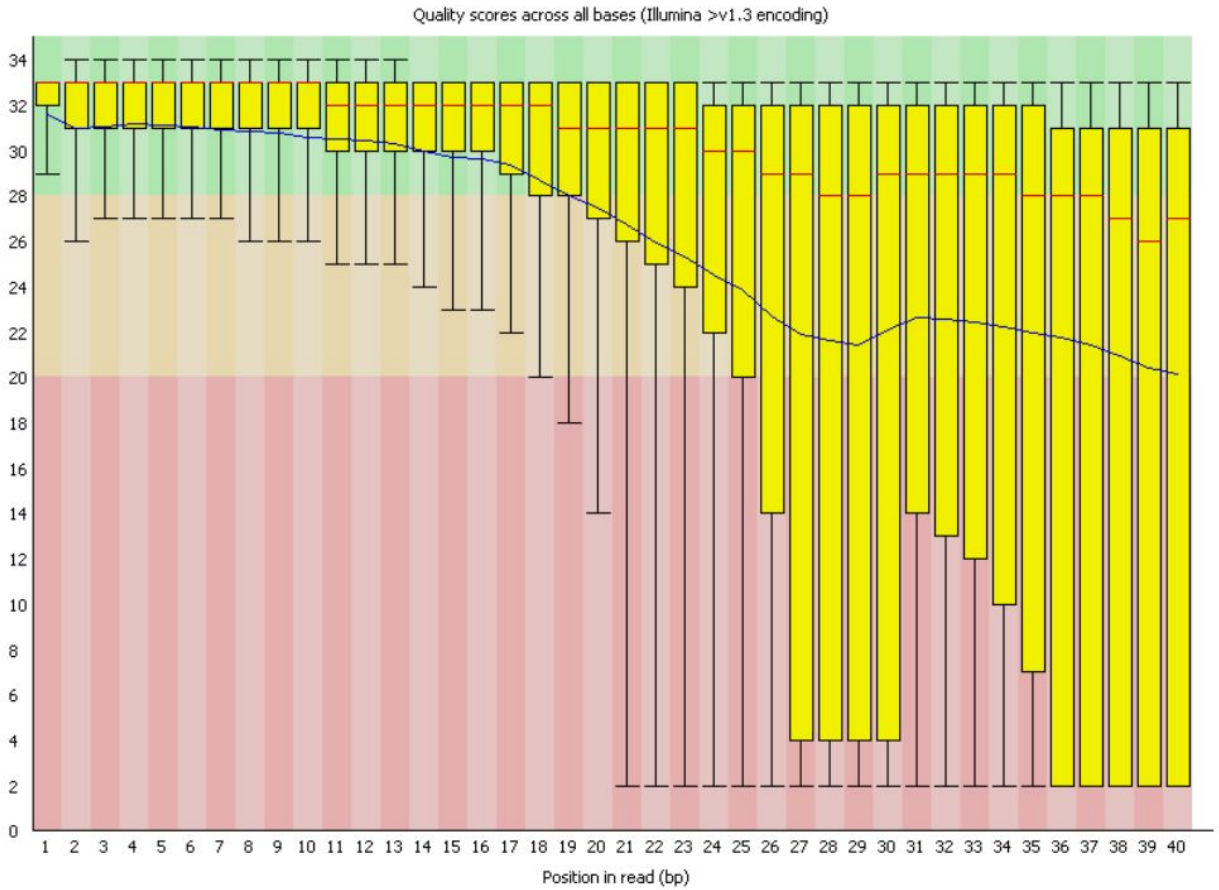
Код	Символ	Код	Символ	Код	Символ	Код	Символ
32	пробел	56	8	80	P	104	H
33	!	57	9	81	Q	105	I
34	"	58	:	82	R	106	J
35	#	59	;	83	S	107	K
36	$	60	<	84	T	108	L
37	%	61	=	85	U	109	m
38	&	62	>	86	V	110	n
39	'	63	?	87	W	111	o
40	(	64	@	88	X	112	p
41	)	65	A	89	Y	113	q
42	*	66	B	90	Z	114	r
43	+	67	C	91	[	115	s
44	,	68	D	92	\	116	t
45	-	69	E	93	]	117	u
46	.	70	F	94	^	118	v
47	/	71	G	95	_	119	w
48	0	72	H	96	`	120	x
49	1	73	I	97	A	121	y
50	2	74	J	98	b	122	z
51	3	75	K	99	c	123	{
52	4	76	L	100	d	124	|
53	5	77	M	101	e	125	}
54	6	78	N	102	f	126	~
55	7	79	O	103	g	127	del

FastQC

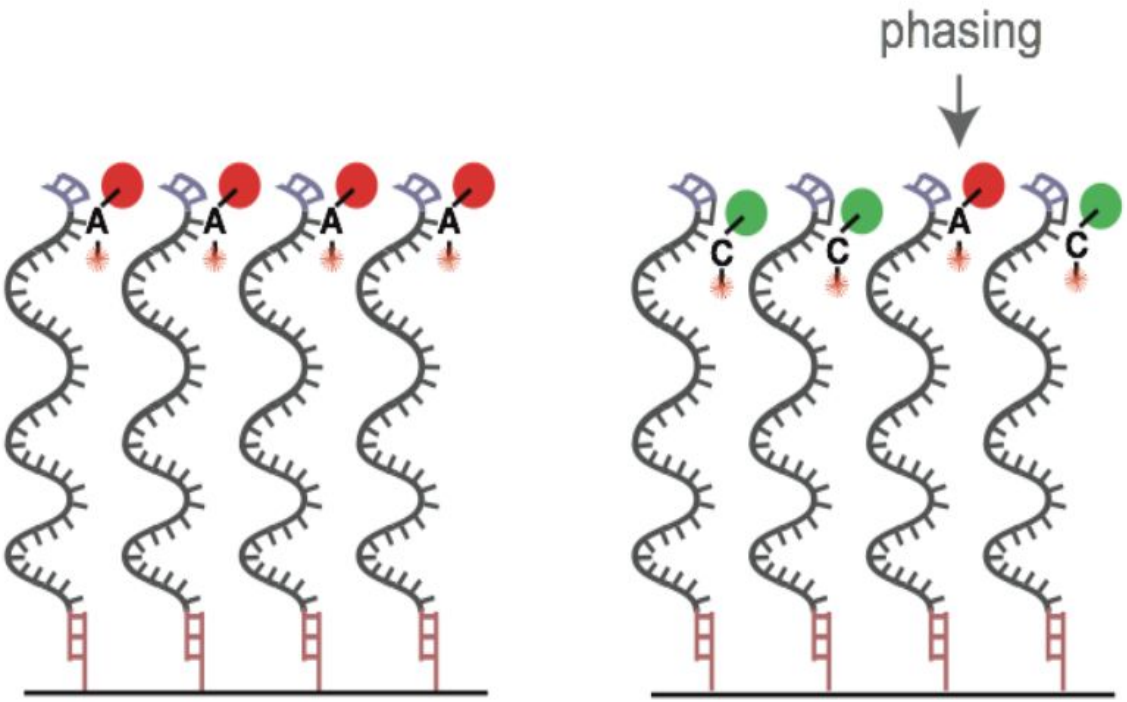
- FastQC – инструмент, позволяющий проводить контроль качества сырых ридов.
- В настоящее время по сути стал стандартом для этой цели

Quality score по основаниям в ридах

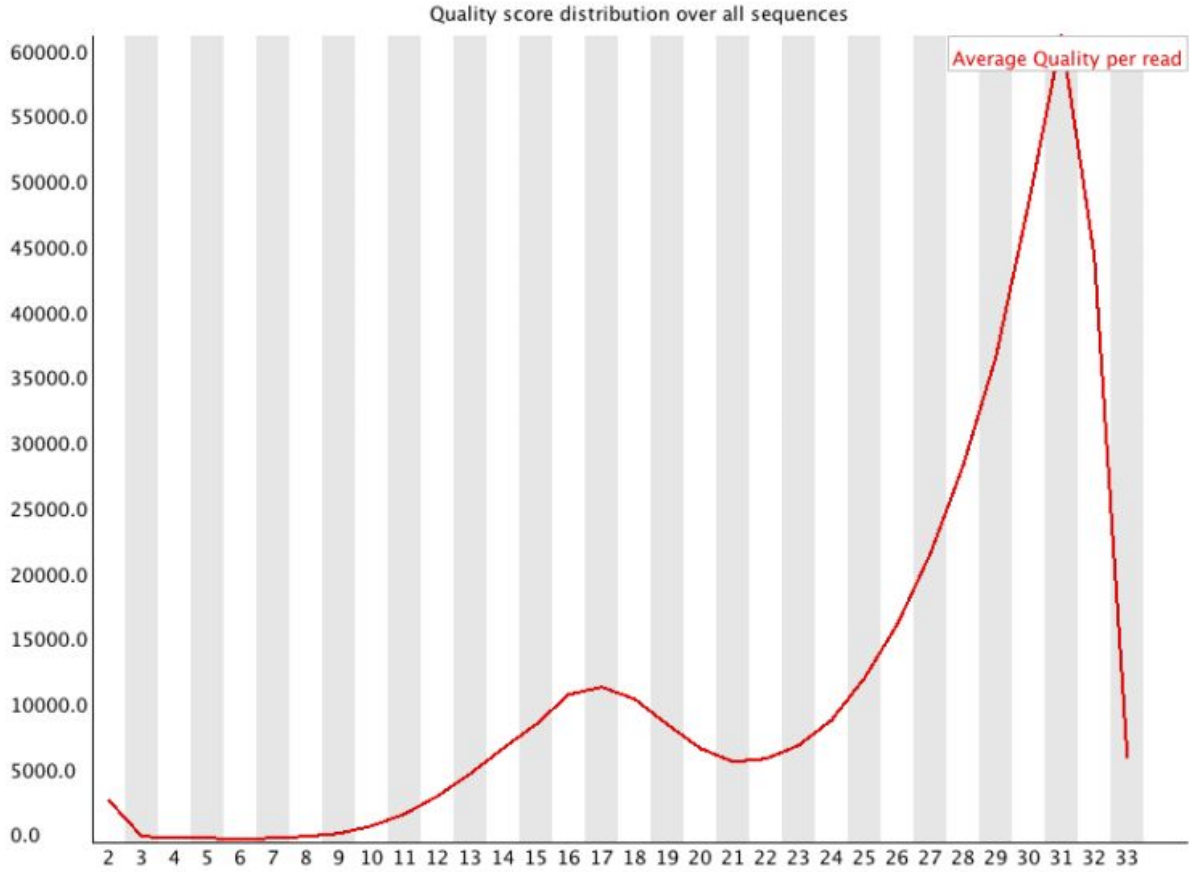
1) Красная линия – медианное значе ние Qscore в данной позиции рида
2) Синяя линия – среднее значение
3) График ящик с усами: желтый прямоугольник – межквартильное расстояние
4) «Усы» - ограничивают часть выборки между 10% и 90% значений



Ухудшение качества прочтения к концу ридов

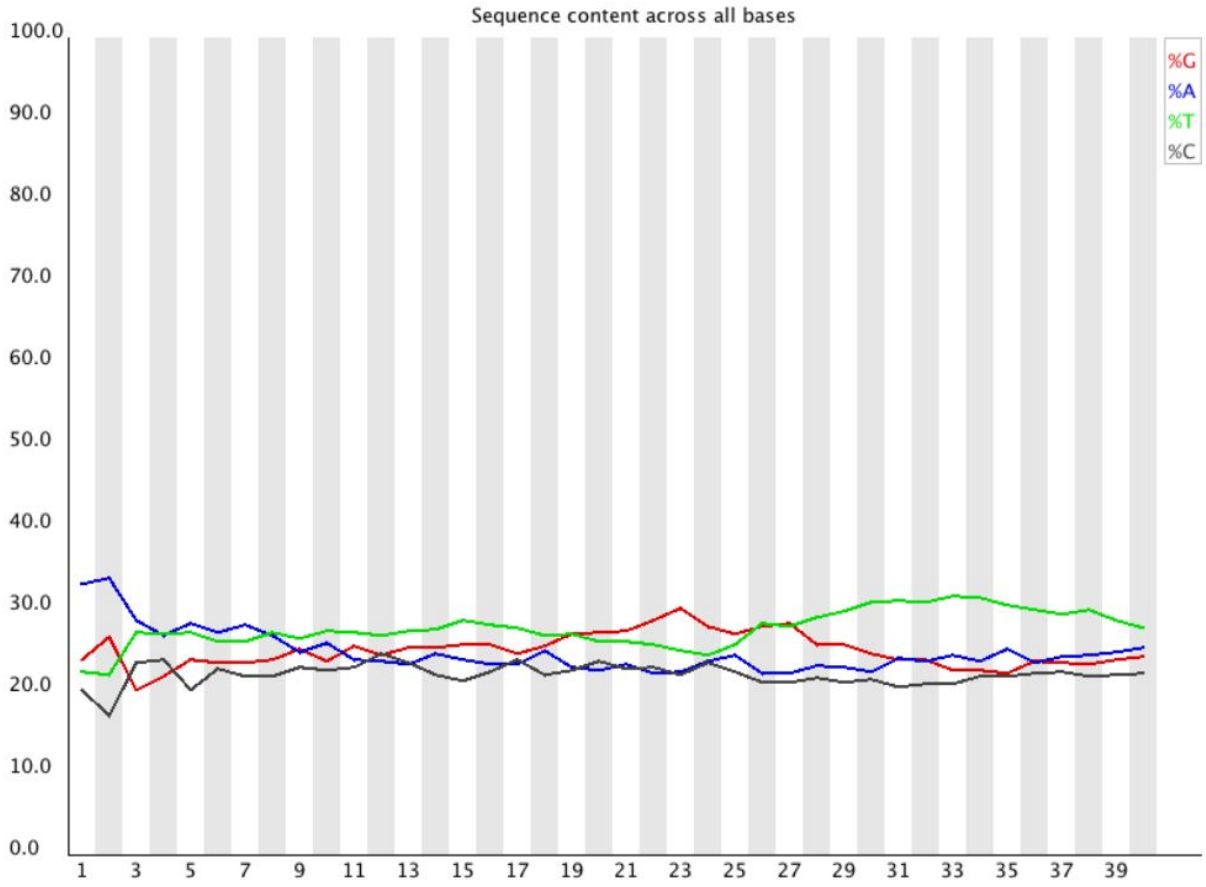


Quality score целых последовательностей



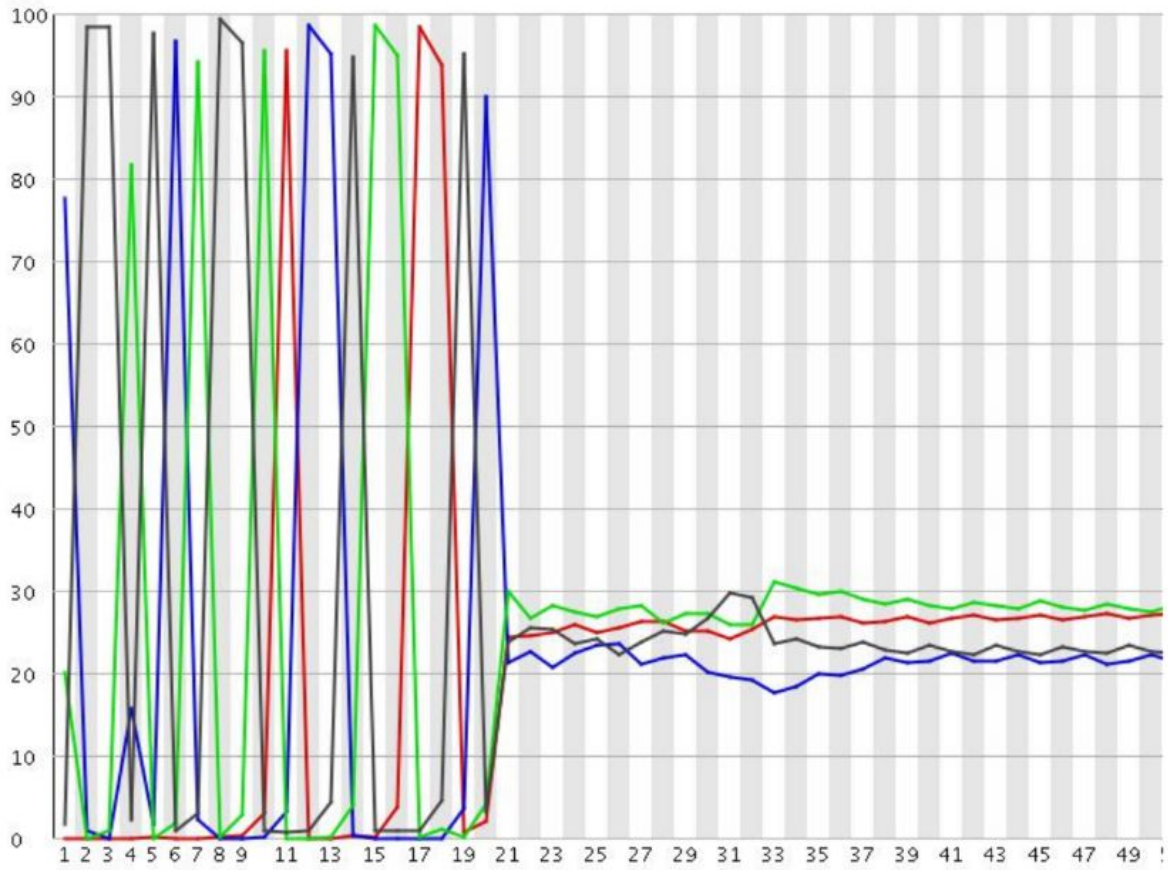
- Этот график позволяет увидеть часть ваших последовательностей, имеющих более низкое среднее качество, чем большинство ридов. Их должно быть не много

Содержание нуклеотидов по позициям в ридах



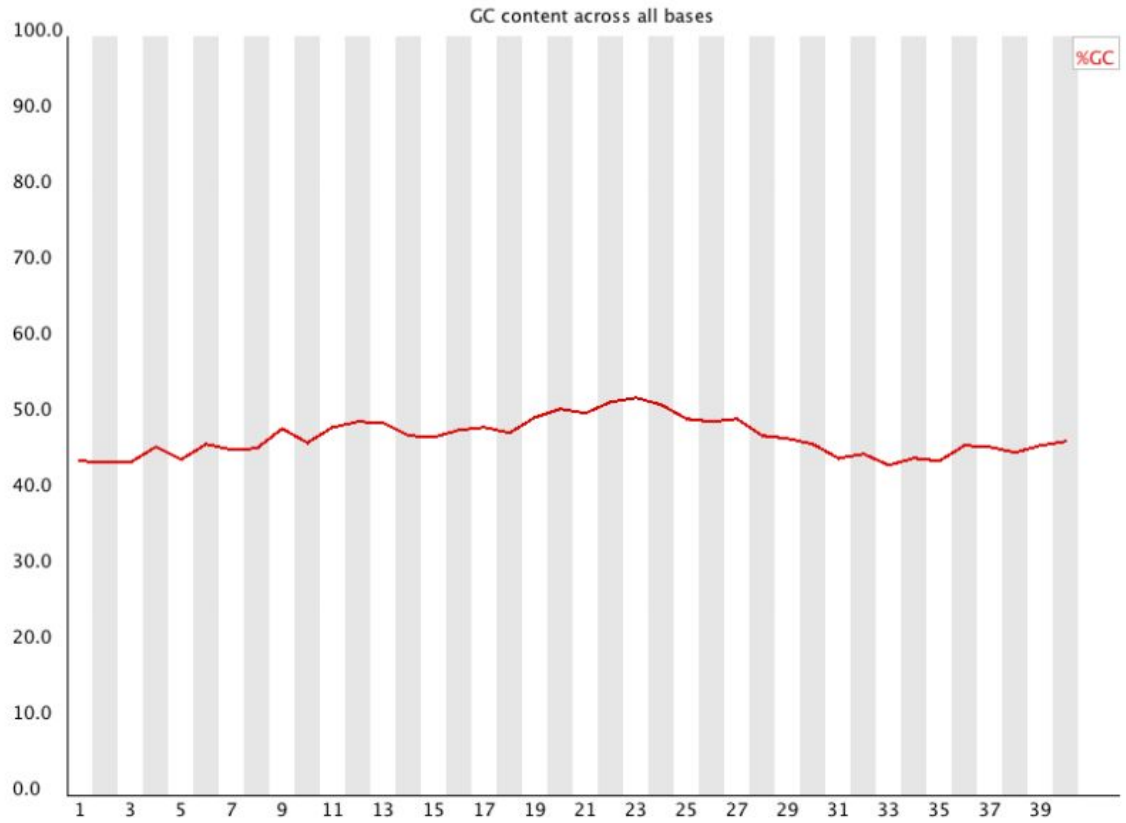
- График показывает пропорцию по нуклеотидам в конкретной позиции ридов. В полностью рандомизированной библиотеке вы ожидаете увидеть незначительные отличия по содержанию конкретного нуклеотида в зависимости от позиции. В общем случае оно должно быть примерно равно доле этого нуклеотида во всей ДНК данного организма.

Содержание нуклеотидов по позициям в



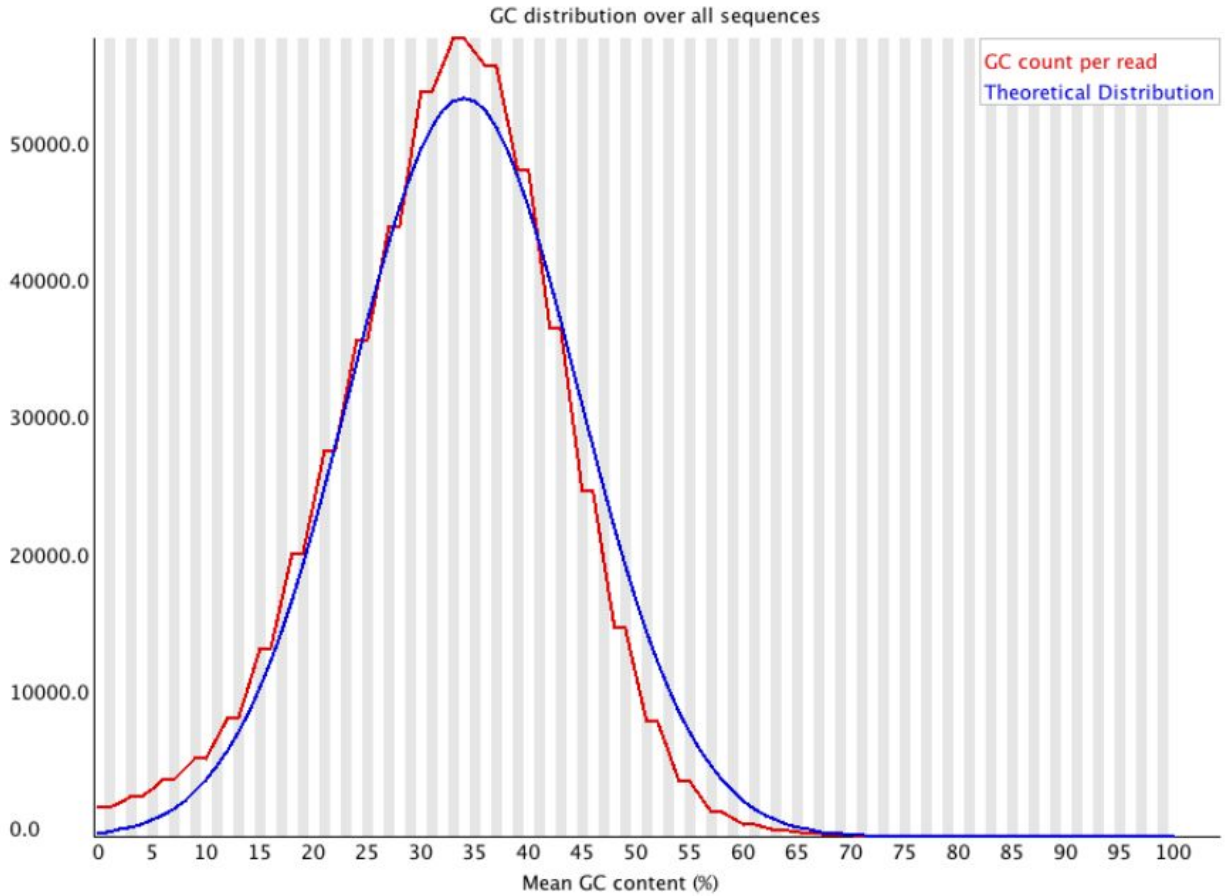
- График показывает пропорцию по нуклеотидам в конкретной позиции ридов. В полностью случайной библиотеке вы ожидаете увидеть незначительные отличия по содержанию конкретного нуклеотида в зависимости от позиции. В общем случае оно должно быть примерно равно доле этого нуклеотида во всей ДНК данного организма.

Содержание GC по позициям в ридах



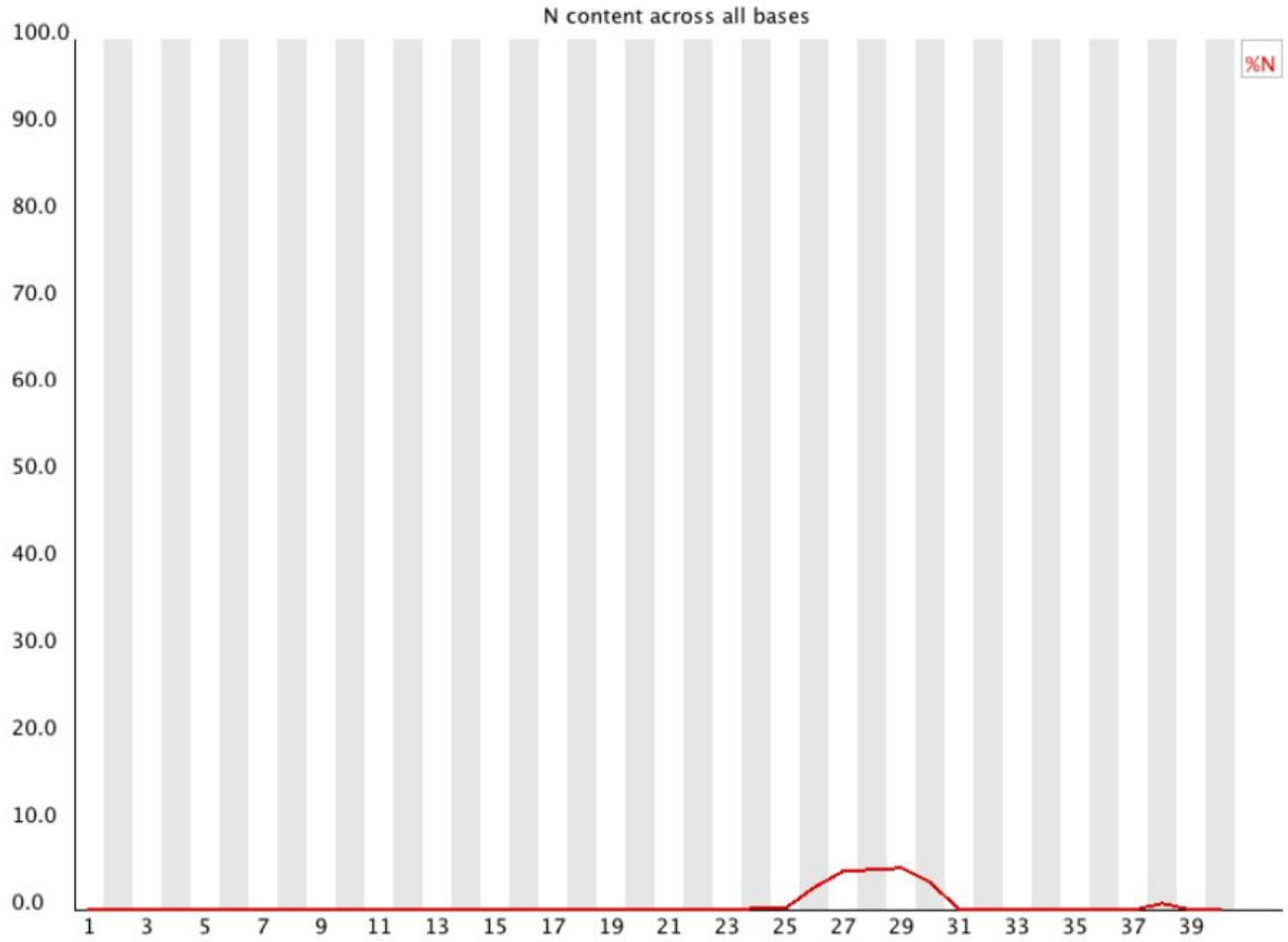
- В случаной библиотеке вы ожидает увидеть незначительную разницу по содержанию GC в зависимости от позиции. Общее содержание GC должно отражать содержание GC в геноме исследуемого организма. Пики на графике могут отражать наличие в вашей библиотеке чрезмерной представленности определенной последовательности.

Содержание GC в целых последовательностях



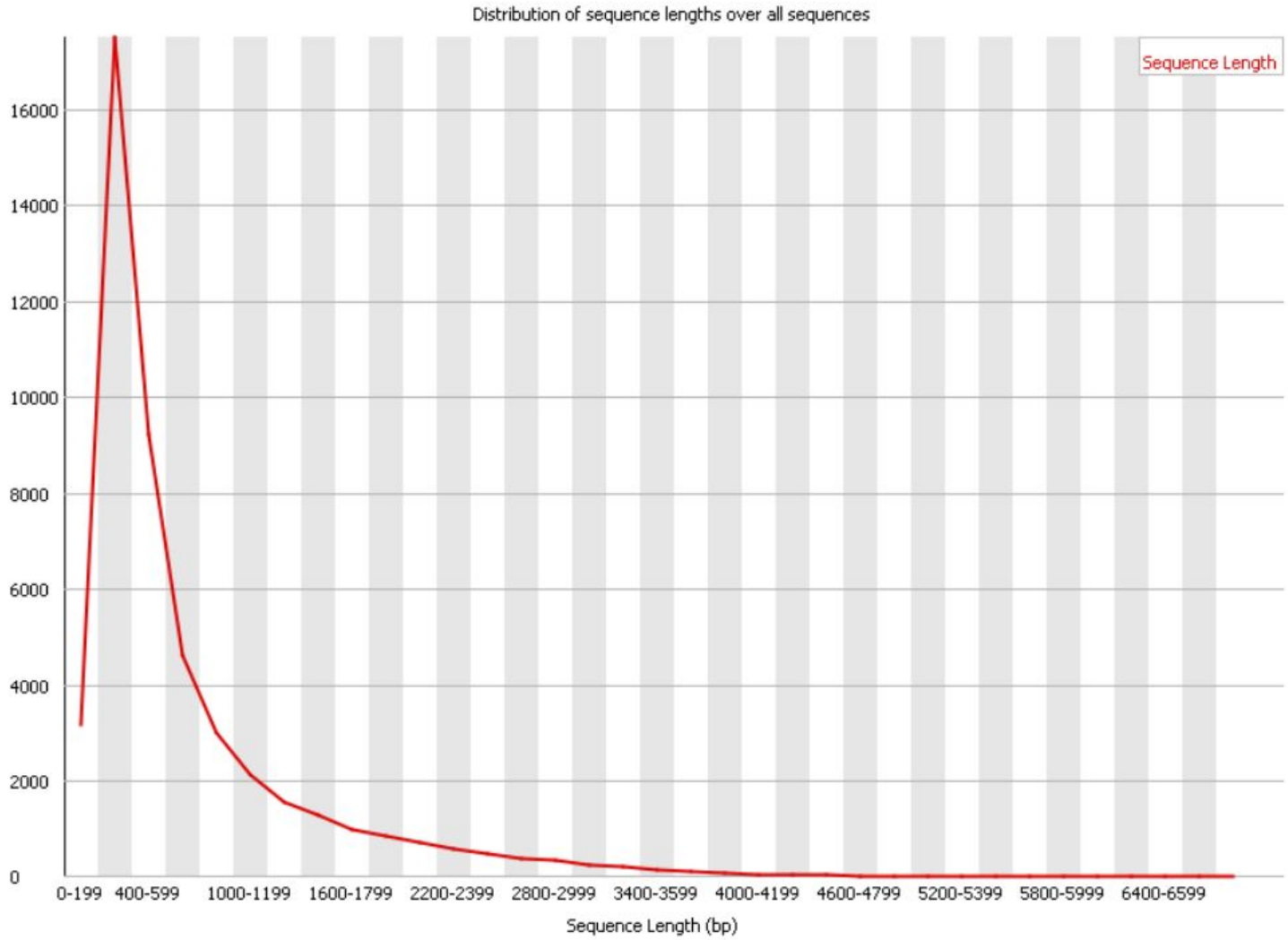
- Вы ожидаете увидеть похожее на нормальное распределение с одним пиком. Наличие второго пика может указывать на загрязнение библиотеки ДНК второго организма.

Содержание N по позиции в ридах

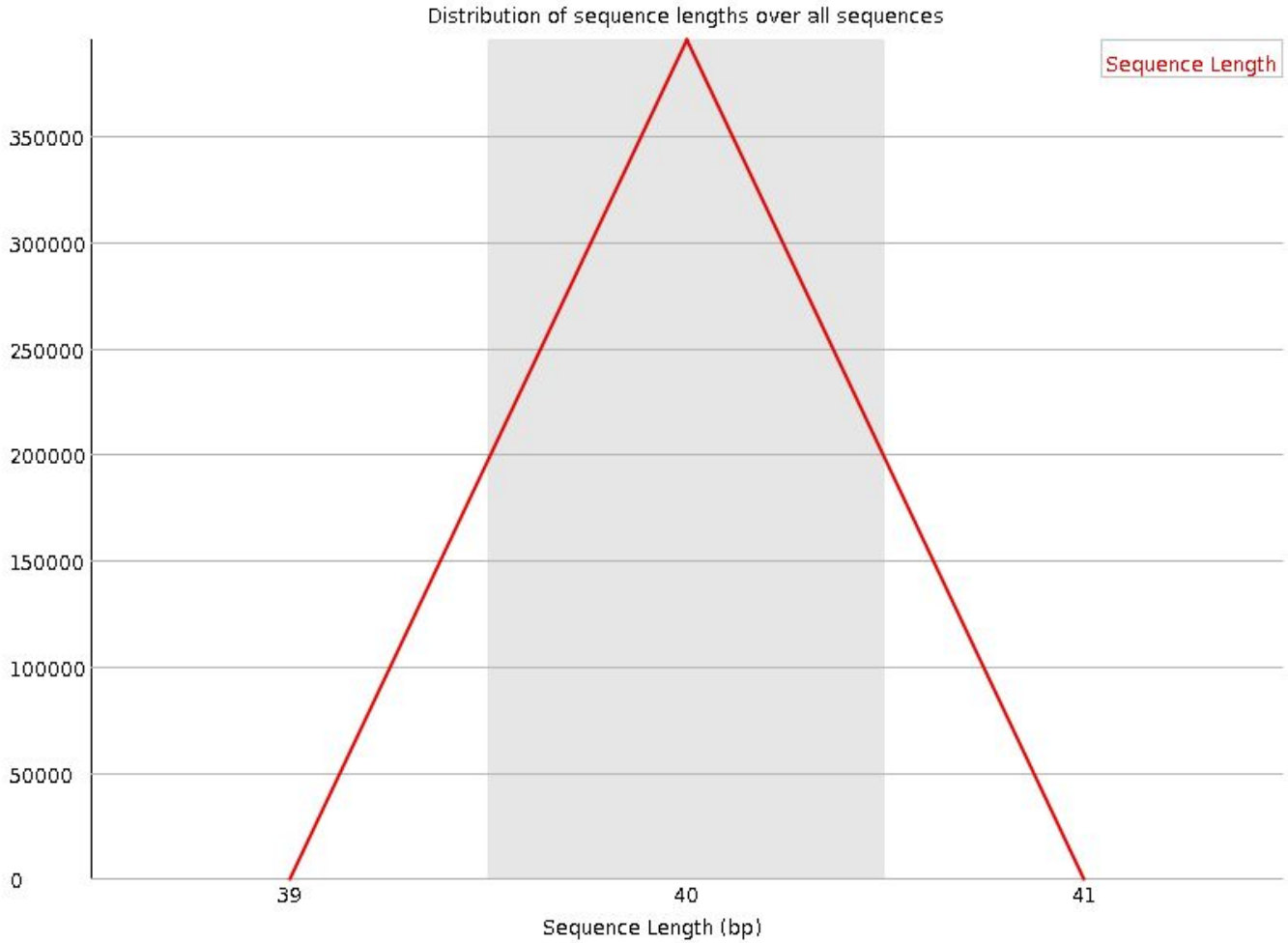


- Наличие небольшого количество N (неопределенных нуклеотидов) в ридах, полученных секвенатором достаточно распространенное явление. FastQC выдает предупреждение, если содержание N больше 5%. Если содержание N более 20% эксперимент считается неудачным.

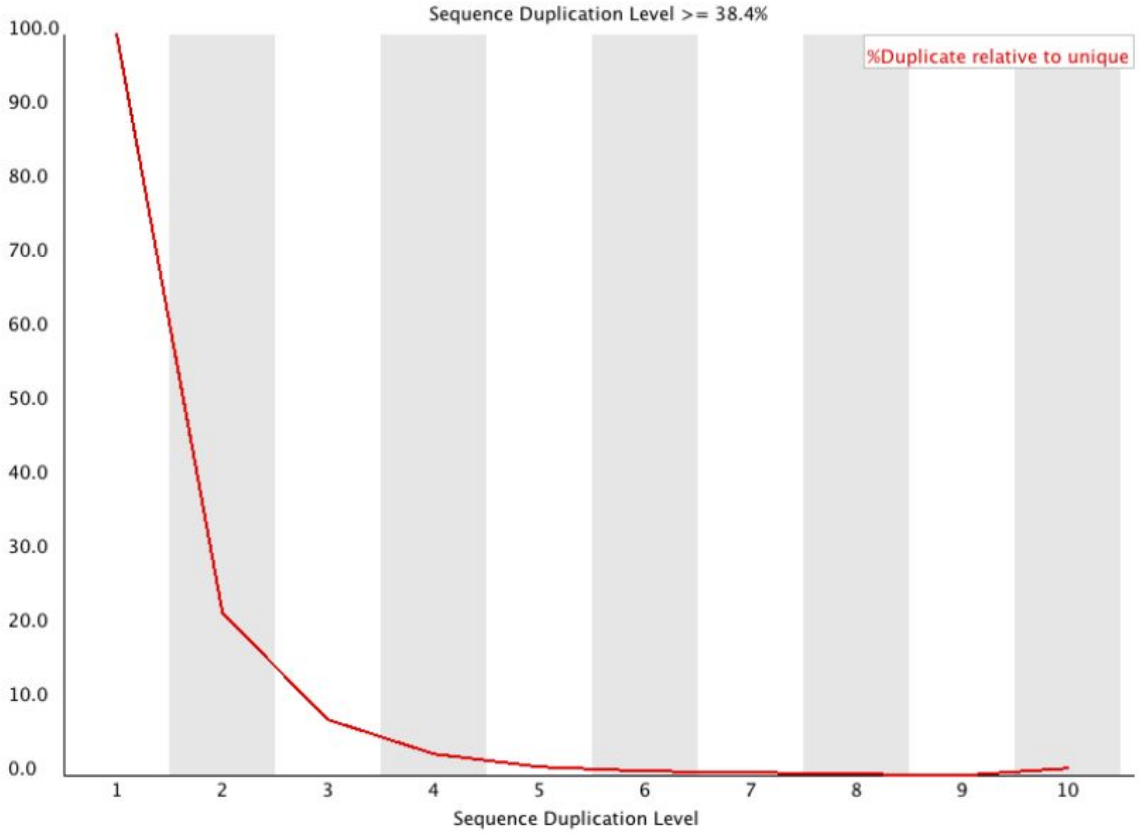
Распределение длин прочтений



Распределение длин прочтений



Дуплицированные последовательности



- В полностью рандомизированной библиотеке большинство сиквенсов встречаются в ридах только 1 раз. Небольшой количество дупликаций может свидетельствовать об очень высоком покрытии таргетного сиквенса. Очень большой уровень дуплекации скорее всего связан с обогащением библиотеки определенным сиквенсом .

Сверхпредставленные последовательности

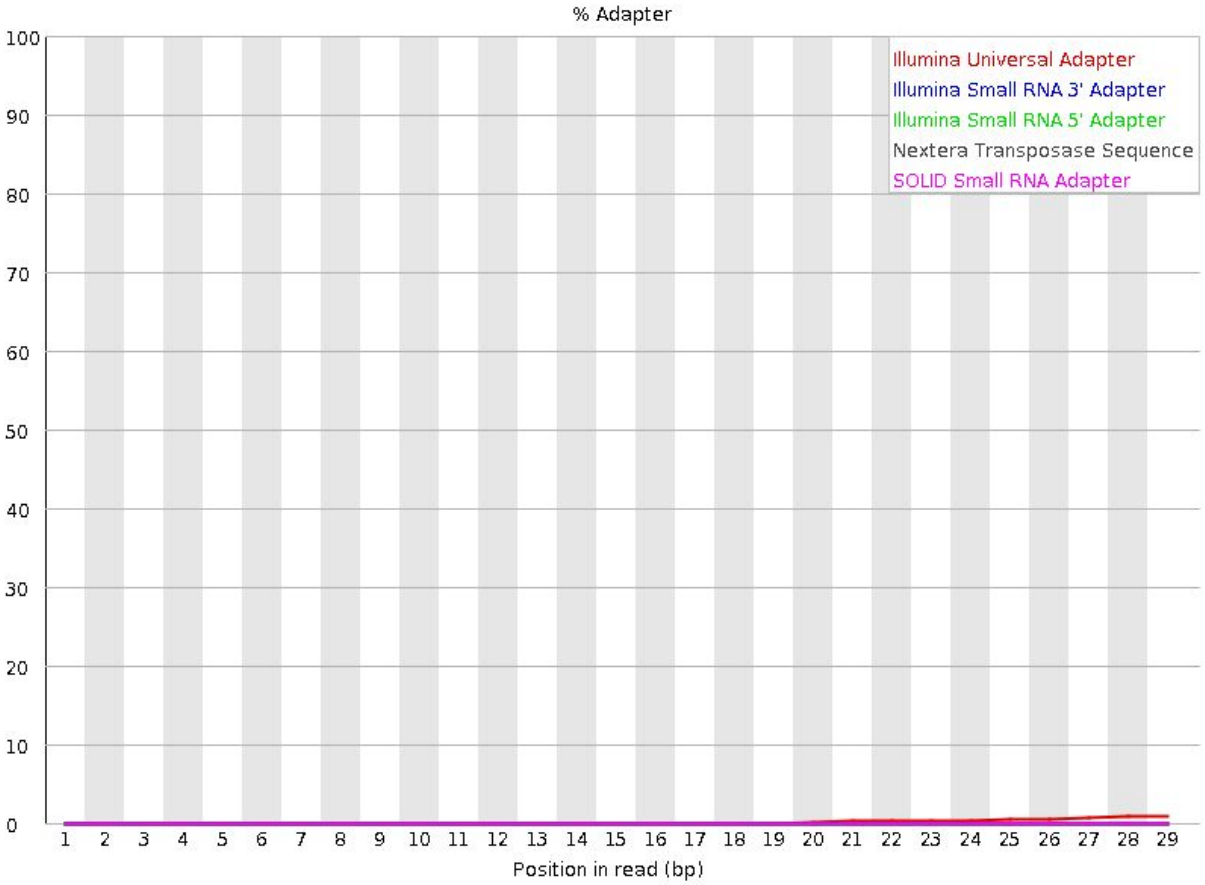
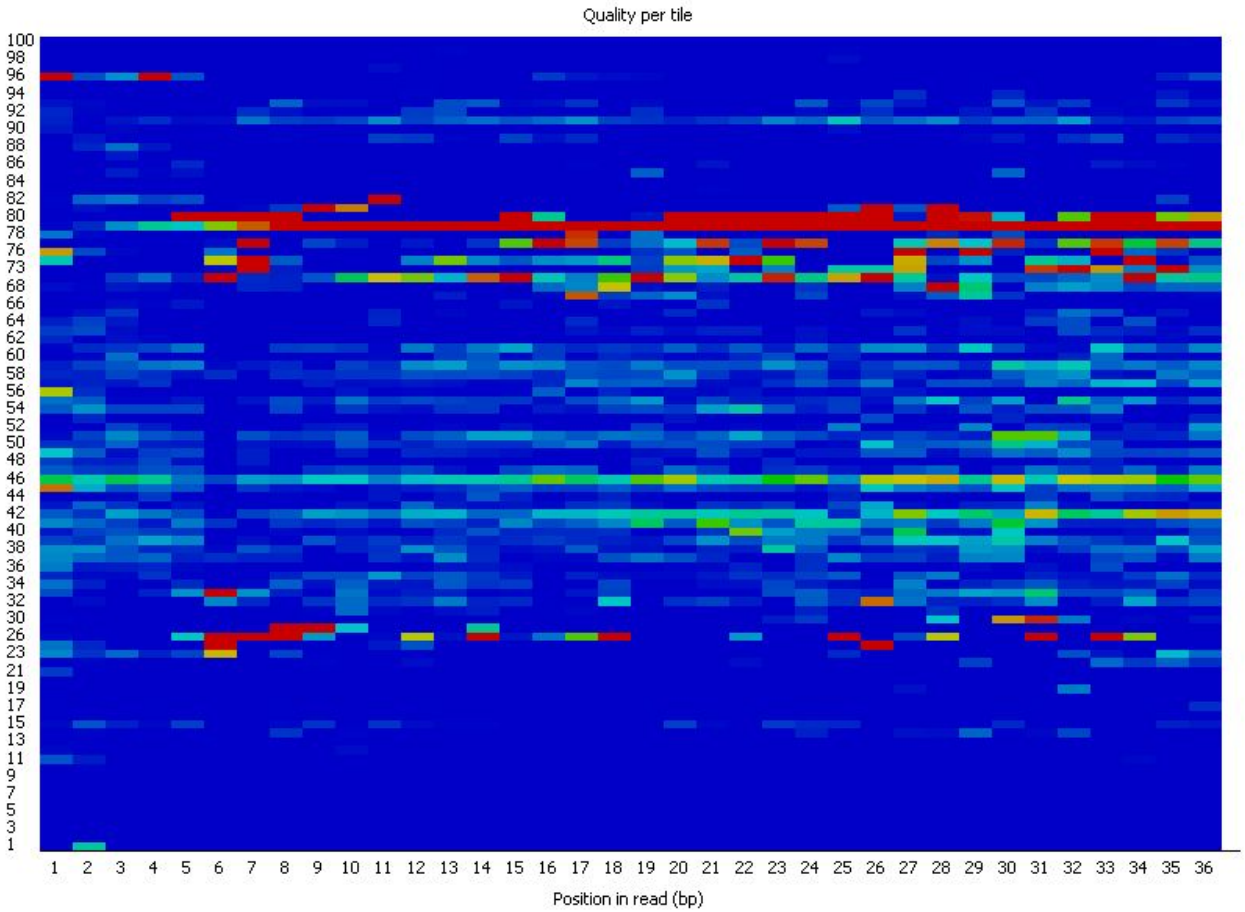
- Обычно библиотека для NGS содержит разнообразный набор последовательностей, без единственной последовательности, составляющая существенную часть всего набора. Обнаружение существенно перепредставленной последовательности может означать, что такая последовательность высоко биологически значима или что при подготовке библиотеки произошла контаминация.
- В этом модуле представляются все последовательности, составляющие более 0,1% от общего количества.
- Для каждой такой последовательности программа произведет поиск совпадений с распространенными контаминирующими агентами и выведет лучшие сопадения. Совпадений не обязательно указывают на конкретный источник контаминации, но может указать правильное направление.
- Стоит помнить, что последовательности адаптеров очень похож.

Сверхпредставленные последовательности

Overrepresented sequences

Sequence	Count	Percentage	Possible Source
AGAGTTTTATCGCTTCCATGACGCAGAAGTTAACACTTTC	2065	0.5224039181558763	No Hit
GATTGGCGTATCCAACCTGCAGAGTTTTATCGCTTCCATG	2047	0.5178502762542754	No Hit
ATTGGCGTATCCAACCTGCAGAGTTTTATCGCTTCCATGA	2014	0.5095019327680071	No Hit
CGATAAAAATGATTGGCGTATCCAACCTGCAGAGTTTTAT	1913	0.4839509420979134	No Hit
GTATCCAACCTGCAGAGTTTTATCGCTTCCATGACGCAGA	1879	0.47534961850600066	No Hit
CCTGCAGAGTTTTATCGCTTCCATGACGCAGAAGTTAACA	613	0.15507680476007366	No Hit
CGGTTCAGCAGGAATGCCGAGATCGGAAGAGCGGTTCAGC	599	0.15153508328105078	Illumina Paired End PCR Primer 2 (96% over 25bp)
TCTGCAGGTTGGATACGCCAATCATTTTTATCGAAGCGCG	585	0.1479933618020279	No Hit
CGCTTAAAGCTACCAGTTATATGGCTGGGGGGTTTTTTTT	552	0.13964501831575965	No Hit
CTCTGCAGGTTGGATACGCCAATCATTTTTATCGAAGCGC	532	0.1345854162028698	No Hit
CTGCGTCATGGAAGCGATAAAACTCTGCAGGTTGGATACG	515	0.13028475440691342	No Hit
CTGCAGGTTGGATACGCCAATCATTTTTATCGAAGCGCGC	505	0.12775495335046852	No Hit
GCTTAAAGCTACCAGTTATATGGCTGGGGGGTTTTTTTTG	411	0.10397482341988626	No Hit

Содержание адаптеров



Per Tile Sequence Quality