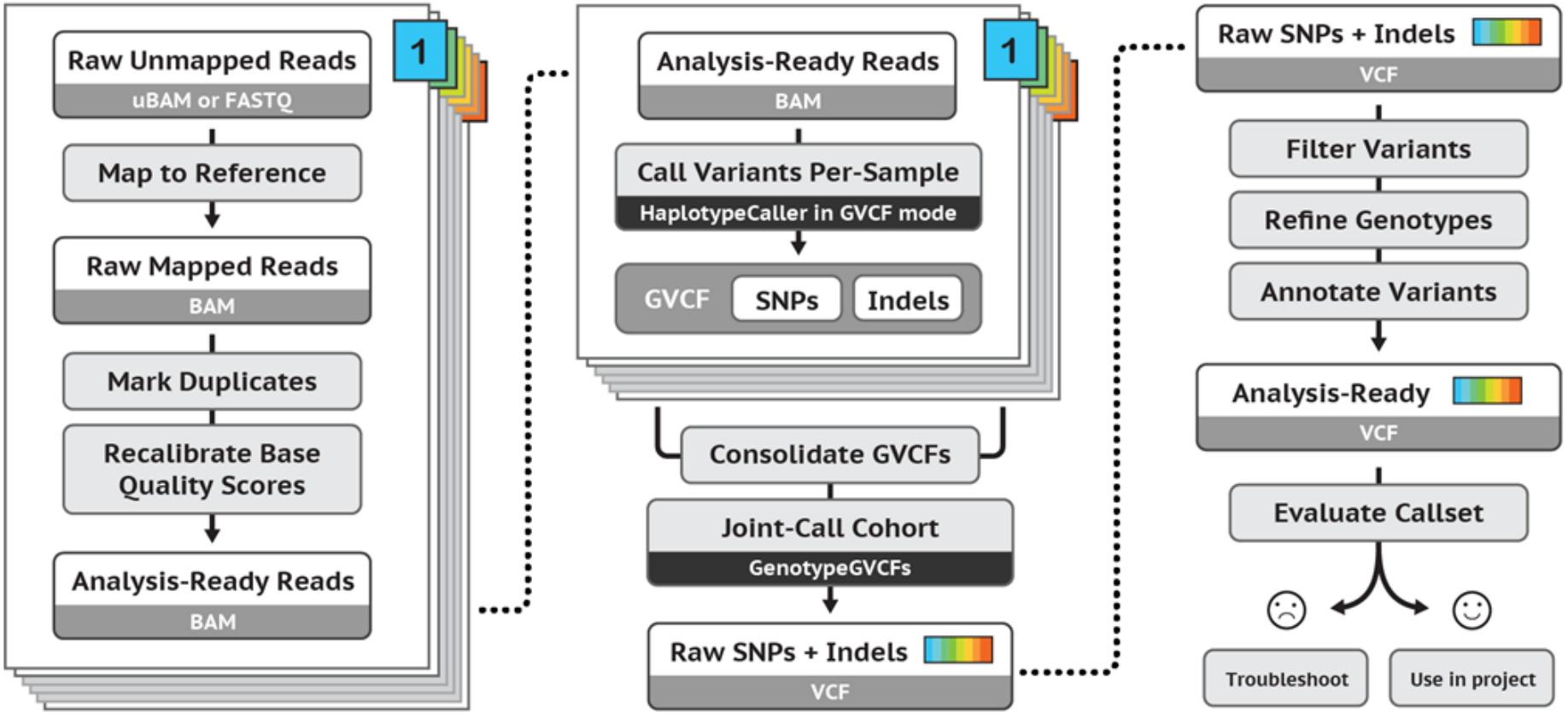
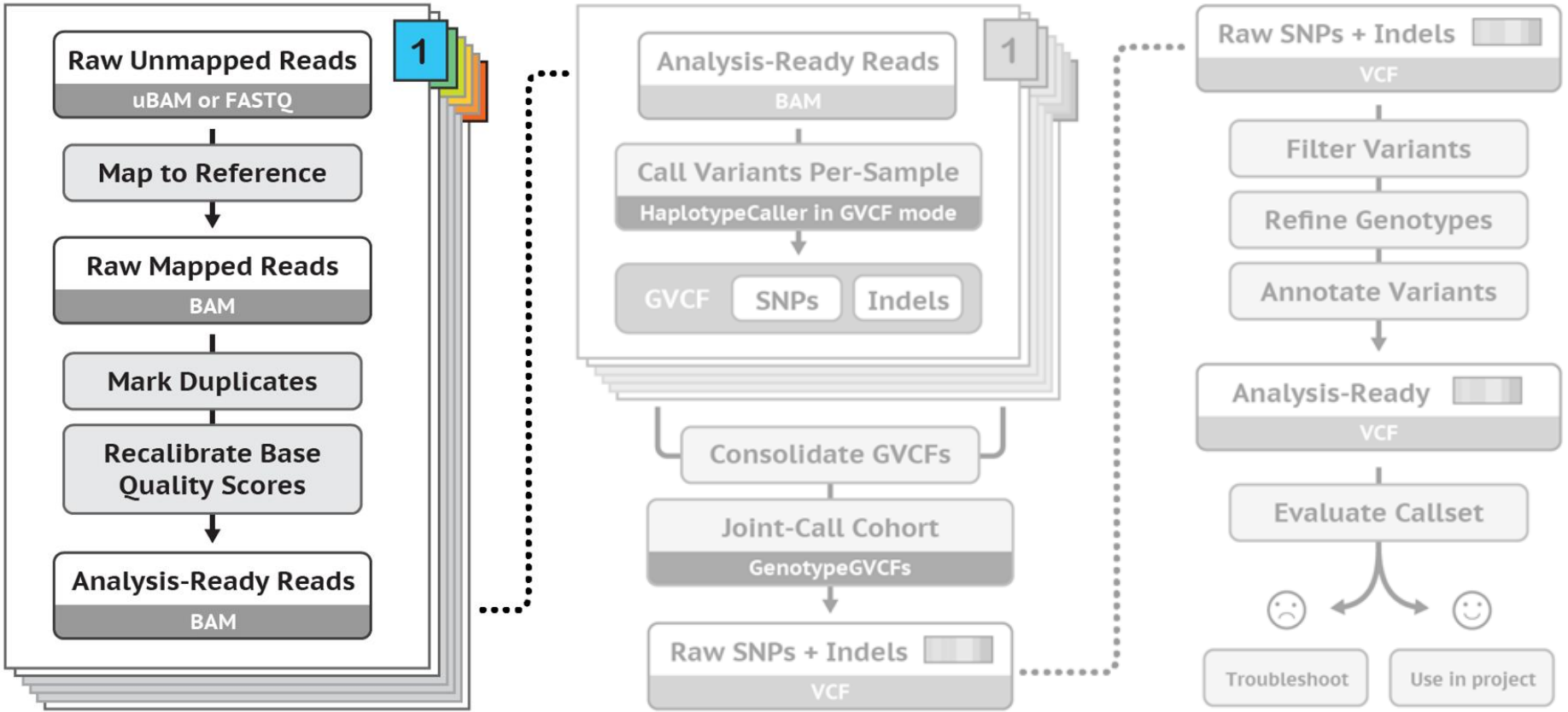
Стандартный клинический пайплайн

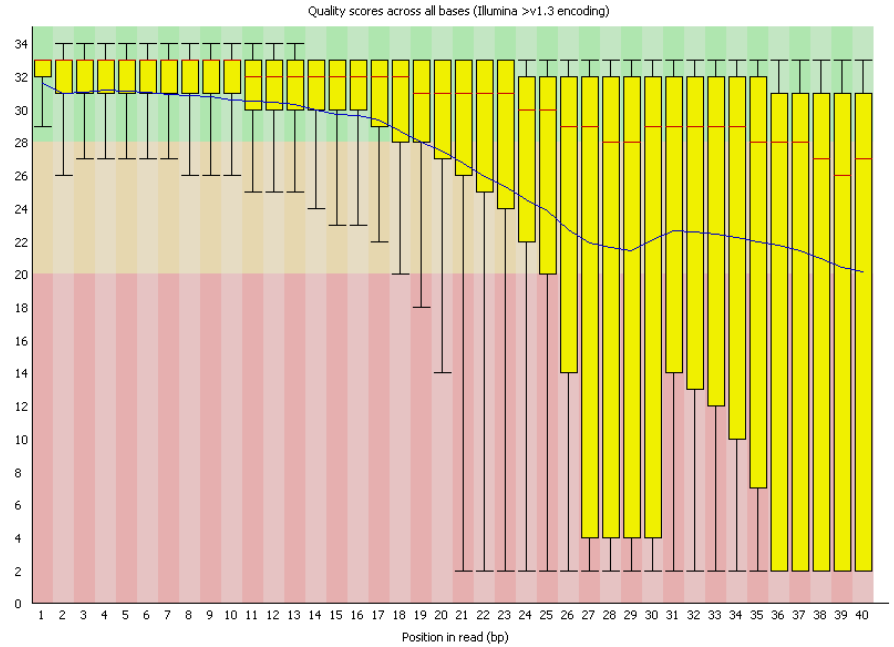
Прикладная биоинформатика: NGS-технологии и Omics-анализ


Raw Unmapped Reads
uBAM or FASTQ
1
Map to Reference
Raw Mapped Reads
BAM
Mark Duplicates
Recalibrate Base Quality Scores
Analysis-Ready Reads
BAM
Analysis-Ready Reads
BAM
1
Call Variants Per-Sample
HaplotypeCaller in GVCF mode
GVCF
SNPs
Indels
Consolidate GVCFs
Joint-Call Cohort
GenotypeGVCFs
Raw SNPs + Indels
VCF
Raw SNPs + Indels
VCF
Filter Variants
Refine Genotypes
Annotate Variants
Analysis-Ready
VCF
Evaluate Callset
Troubleshoot
Use in project


Raw Unmapped Reads
uBAM or FASTQ
Map to Reference
Raw Mapped Reads
BAM
Mark Duplicates
Recalibrate Base Quality Scores
Analysis-Ready Reads
BAM
1
Analysis-Ready Reads
BAM
Call Variants Per-Sample
HaplotypeCaller in GVCF mode
GVCF
SNPs
Indels
1
Consolidate GVCFs
Joint-Call Cohort
GenotypeGVCFs
Raw SNPs + Indels
VCF
Raw SNPs + Indels
VCF
Filter Variants
Refine Genotypes
Annotate Variants
Analysis-Ready
VCF
Evaluate Callset
Troubleshoot
Use in project

Контроль качества

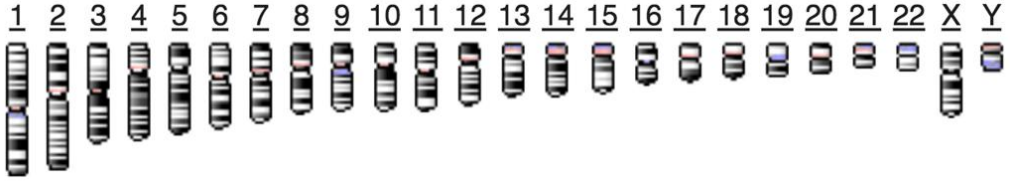
- FASTQC
- Зачастую требуется проводить тримминг (качество, адаптеры).



Формат BAM. uBAM

- Для корректной работы пайплайна необходимо включить в исходные FASTQ-файлы метаданные: идентификатор библиотеки, координаты секвенирования в ячейке (лейн).
- Для этого вместо FASTQ-файлов в GATK пайплайне используется формат uBAM, который является модификацией формата BAM.
- О формате BAM речь пойдёт позднее, сейчас важно лишь понимать, что uBAM – это по сути FASTQ с метаданными.

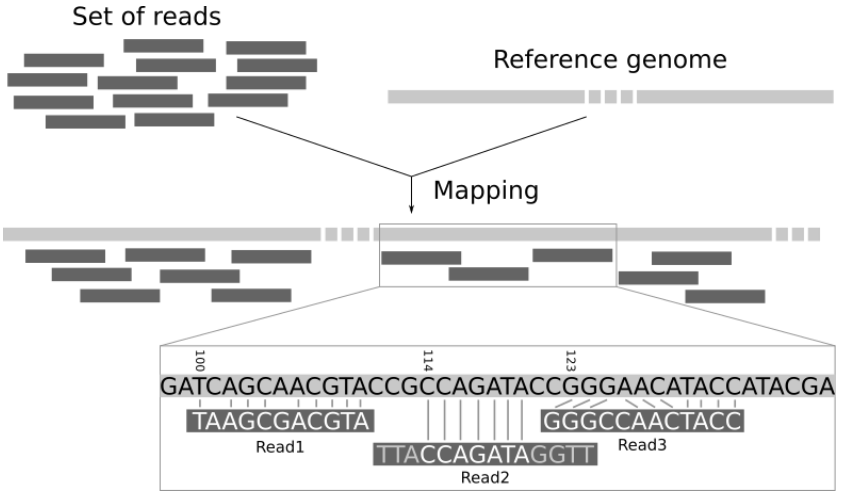
Сборка генома человека




```
>gi|568336023|gb|CM000663.2| Homo sapiens chromosome 1, GRCh38 reference primary assembly
CCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTA
ACCCTAACCCTAACCCTAACCCTAACCCAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCC
TAACCCTAACCCTAACCCTAACCCTAACCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACC
CTAACCCTAACCCCTAACCCTAACCCTAAACCCTAAACCCTAACCCTAACCCTAACCCTAACCCTAACCC
CAACCCCAACCCCAACCCCAACCCCAACCCCAACCCTAACCCCTAACCCTAACCCTAACCCTACCCTAAC
CCTAACCCTAACCCTAACCCTAACCCTAACCCCTAACCCCTAACCCTAACCCTAACCCTAACCCTAACCC
TAACCCTAACCCCTAACCCTAACCCTAACCCTAACCCTCGCGGTACCCTCAGCCGGCCCGCCCGCCCGGG
TCTGACCTGAGGAGAACTGTGCTCCGCCTTCAGAGTACCACCGAAATCTGTGCAGAGGACAACGCAGCTC
CGCCCTCGCGGTGCTCTCCGGGTCTGTGCTGAGGAGAACGCAACTCCGCCGTTGCAAAGGCGCGCCGCGC
CGGCGCAGGCGCAGAGAGGCGCGCCGCGCCGGCGCAGGCGCAGAGAGGCGCGCCGCGCCGGCGCAGGCGC
AGAGAGGCGCGCCGCGCCGGCGCAGGCGCAGAGAGGCGCGCCGCGCCGGCGCAGGCGCAGAGAGGCGCGC
CGCGCCGGCGCAGGCGCAGACACATGCTAGCGCGTCGGGGTGGAGGCGTGGCGCAGGCGCAGAGAGGCGC
GCCGCGCCGGCGCAGGCGCAGAGACACATGCTACCGCGTCCAGGGGTGGAGGCGTGGCGCAGGCGCAGAG
AGGCGCACCGCGCCGGCGCAGGCGCAGAGACACATGCTAGCGCGTCCAGGGGTGGAGGCGTGGCGCAGGC
GCAGAGACGCAAGCCTACGGGCGGGGGTTGGGGGGGCGTGTGTTGCAGGAGCAAAGTCGCACGGCGCCGG
GCTGGGGCGGGGGGAGGGTGGCGCCGTGCACGCGCAGAAACTCACGTCACGGTGGCGCGGCGCAGAGACG
GGTAGAACCTCAGTAATCCGAAAAGCCGGGATCGACCGCCCCTTGCTTGCAGCCGGGCACTACAGGACCC
GCTTGCTCACGGTGCTGTGCCAGGGCGCCCCCTGCTGGCGACTAGGGCAACTGCAGGGCTCTCTTGCTTA
GAGTGGTGGCCAGCGCCCCCTGCTGGCGCCGGGGCACTGCAGGGCCCTCTTGCTTACTGTATAGTGGTGG
```

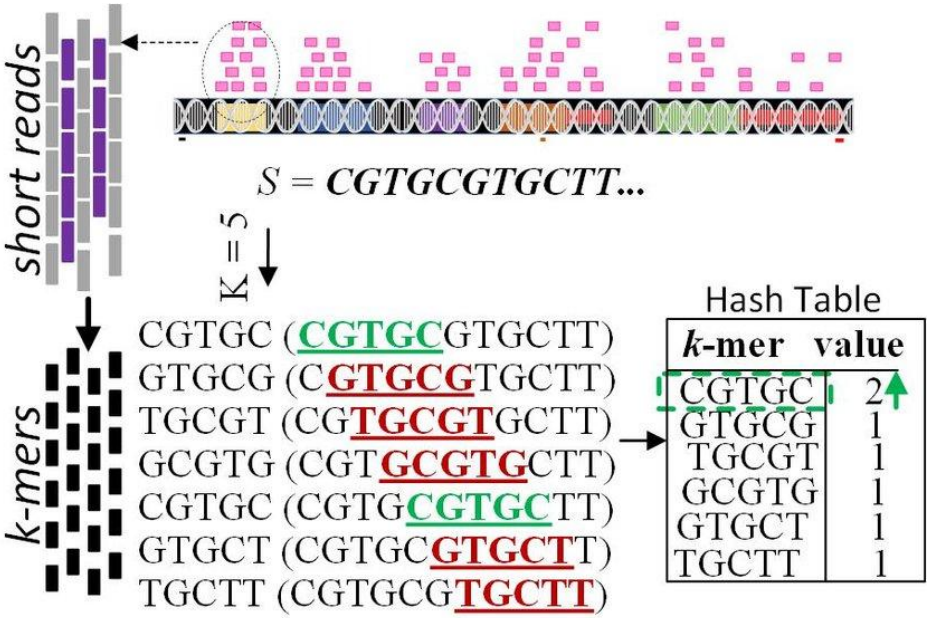
Выравнивание на референс

- Процесс выравнивания на ридов на последовательность референсного генома – mapping.
- Сложность выравнивания состоит в том, что необходимо допускать возможность обрезания концов ридов, инделов и нуклеотидных замен.
- Неверное выравнивание => неверные результаты => неверные выводы!



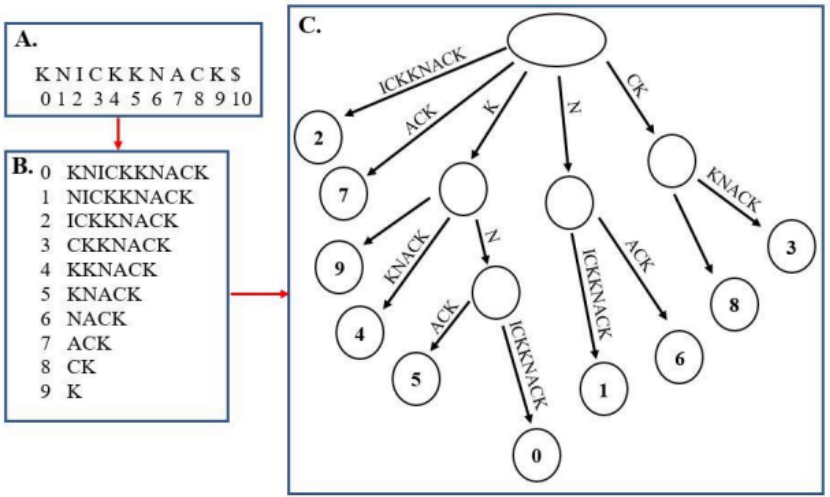
Историческая основа выравнивания – алгоритм BLAST

- Разделение референсной последовательности на k-mer’ы и формирование таблицы хэшей (hash-table), в которой ищутся k-mer’ы ридов. Получив множество вариантов различных выравниваний, выбираем наиболее правдоподобное, с наименьшим числом несостыковок.
- Однако, BLAST подразумевает большое время вычисления.



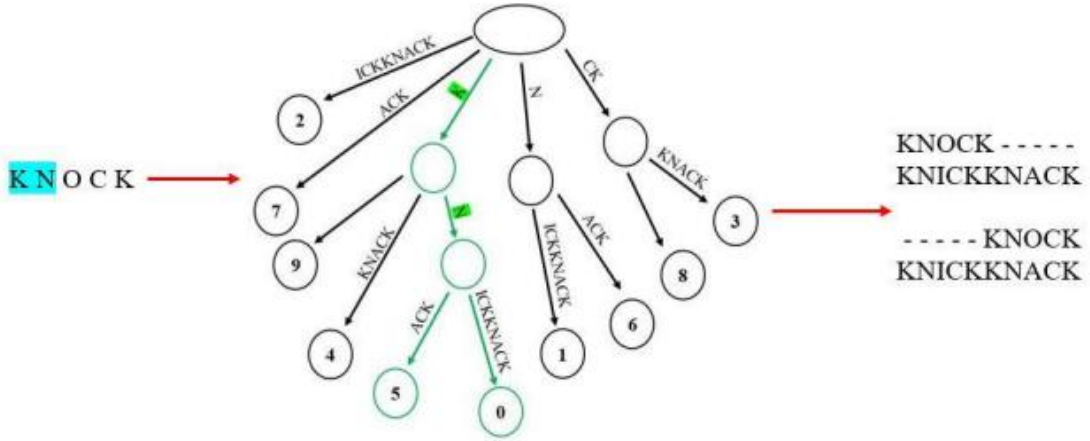
Ускорение поиска k-mer’ов – MUMmer

- MUMmer решил проблему длительного поиска k-mer’ов в референсной последовательности, преобразовав её в суффиксное дерево.
- Это структура данных для хранения массива строк, являющихся суффиксами некоторой исходной строки.



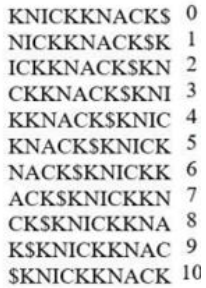
Поиск по суффиксному дереву

- Поиск по суффиксному дереву осуществляется по k-mer’ам (на рисунке k=2), которые берутся из ридов (на рисунке – KNOCK).
- Исходя из того, что k-mer’у KN соответствуют два возможных варианта исходной строки (KNICKKNACK и KNACK), формируется два выравнивания.
- Однако, суффиксное дерево занимает большое количество памяти.



От суффиксного дерева к FM-индексу

- В конец референсной последовательности добавляется “$”.
- Каждый “rotation” референсной последовательности – смещение начала на один символ вправо и добавление недостающих символов в конец.
- Вместо “rotation” (аналог суффиксов) хранятся индексы символов после “$”.



Строка BWT

- “Rotation” сортируются по лексикографическому порядку.
- Одновременно формируется последовательность их номеров (0-10 на рисунке: 10, 7, 8, 3, 2…).
- Последний “столбец” полученной матрицы (выделен зелёным) является центральной сущностью преобразования Барроуза-Уилера (Burrows-Wheeler transform, BWT) – A (соответствующая ей последовательность индексов – B).

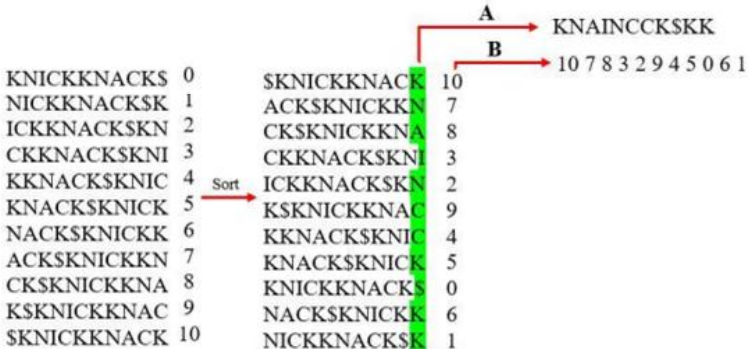
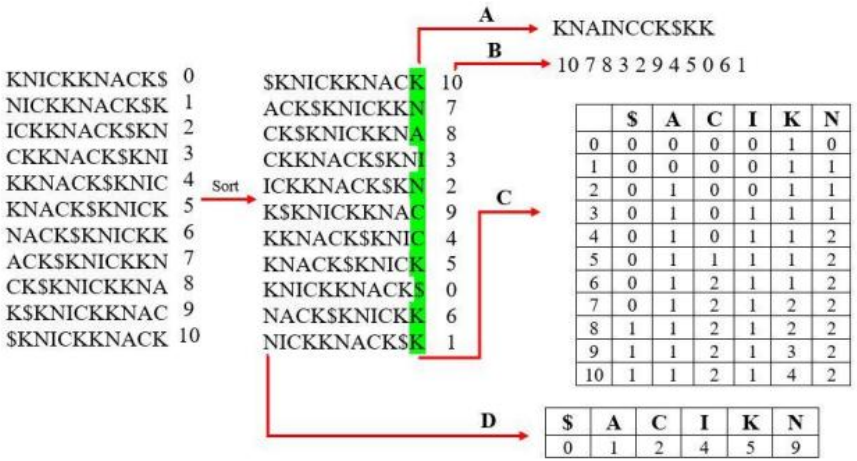


Таблица рангов и таблица поиска

- Формируется таблица рангов: строки – “rotation” референсной последовательности, столбцы – уникальные символы.
- Таблица заполняется при прохождении вдоль строки BWT (“KNAINCCK$KK”): каждый новый символ строки увеличивает значение в соответствующем столбце последующей строки.
- Индексы первых встреч с символами первого столбца также сохраняются в отдельной таблице поиска (D).

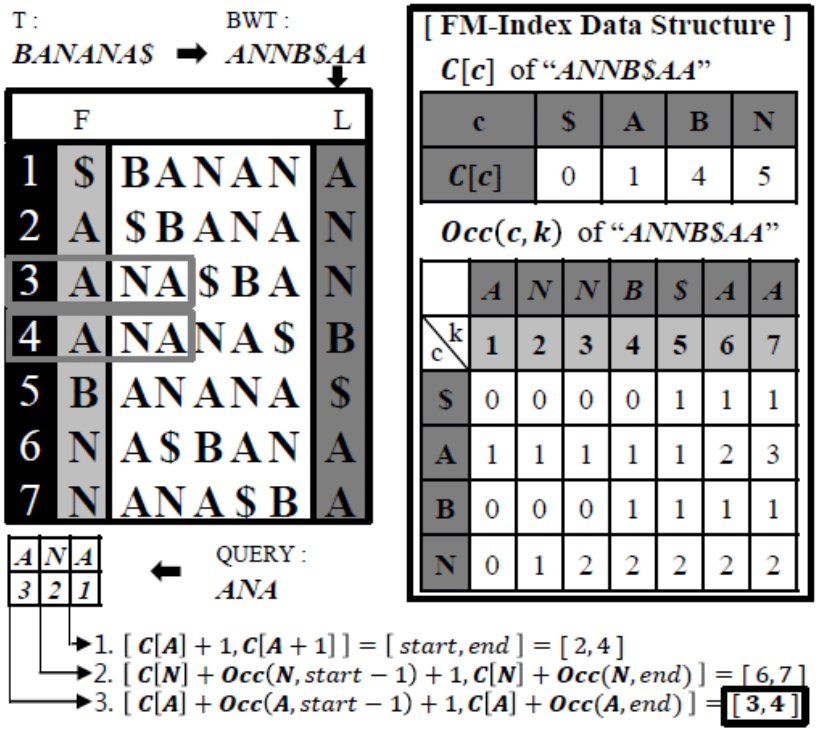


	$	A	C	I	K	N
0	0	0	0	0	1	0
1	0	0	0	0	1	1
2	0	1	0	0	1	1
3	0	1	0	1	1	1
4	0	1	0	1	1	2
5	0	1	1	1	1	2
6	0	1	2	1	1	2
7	0	1	2	1	2	2
8	1	1	2	1	2	2
9	1	1	2	1	3	2
10	1	1	2	1	4	2

$	A	C	I	K	N
0	1	2	4	5	9

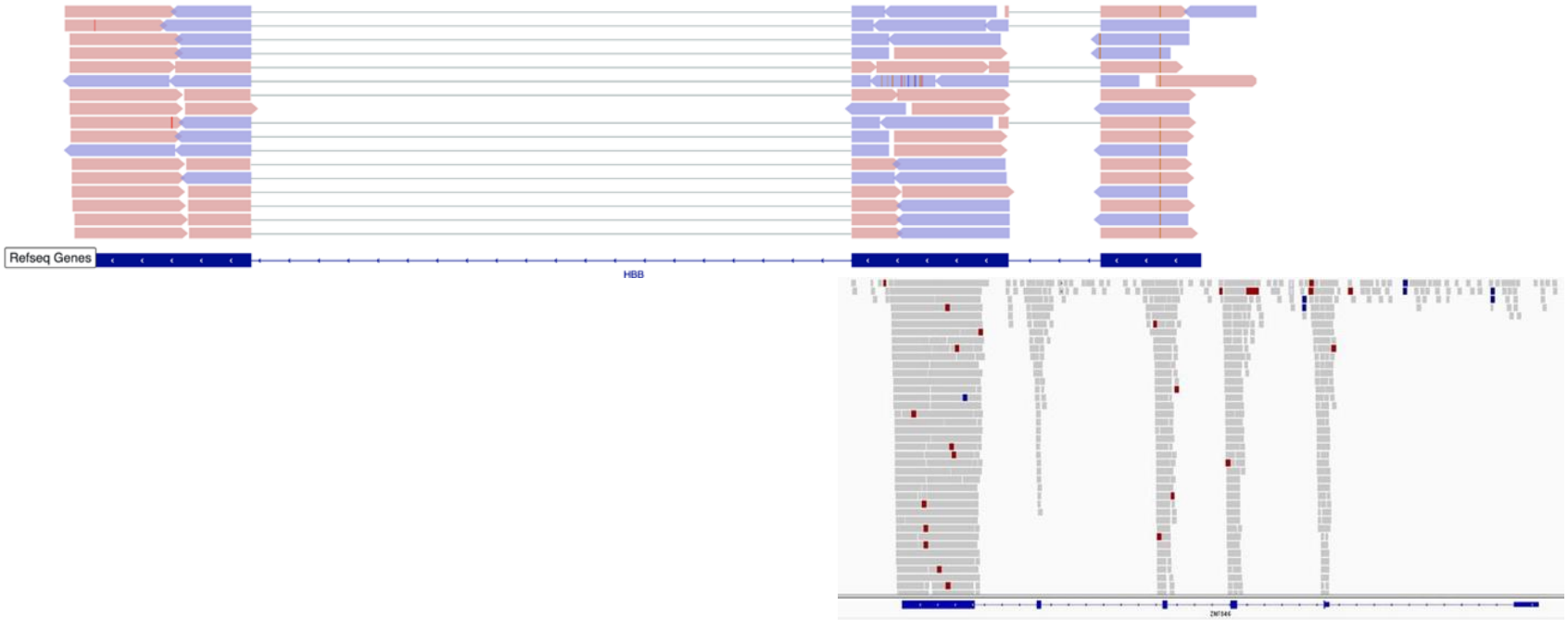
Поиск в алгоритме FM-index

- Риды делятся на k-mer’ы, которые берутся в обратном порядке.
- Осуществляется итеративный поиск символов k-mer’ов по формуле на рисунке.
- “ANA” – рид. Ищем с конца.
- [2,4] – индексы в F, которым соответствуют последовательности, начинающиеся с “A”.
- [6,7] – то же для “NA”.
- [3,4] – то же для “ANA”.



ДНК или РНК выравнивание

- Масса аспектов выравнивания связана с видом материала – ДНК или РНК.

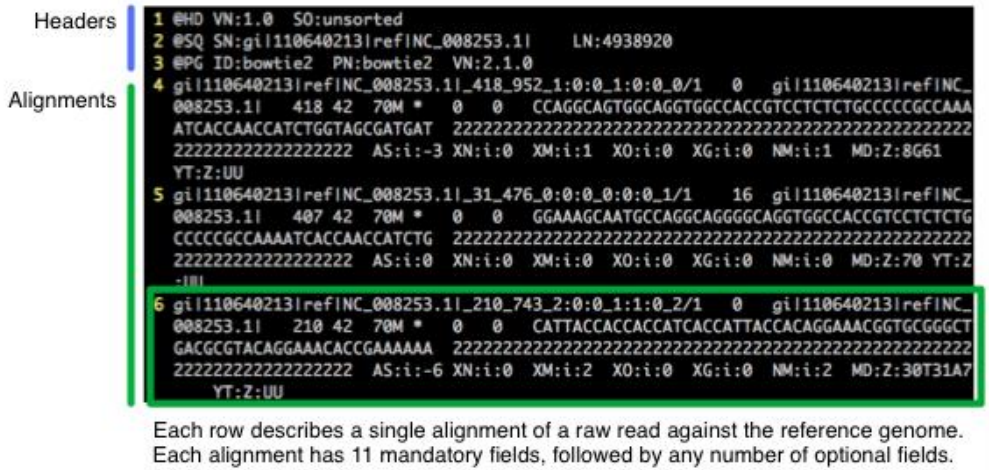


Программы для выравнивания ДНК

- Bowtie2 – два режима работы: end-to-end (риды используются целиком) и local (используется центральная часть рида, если концы низкокачественные).
- BWA-MEM – выбирает лучшее выравнивание из всех возможных end-to-end или local для каждого рида. Рекомендуется в GATK Best Practices.
- HISAT2 – изначально HISAT создавался для выравнивания РНК, HISAT2 же способен работать и с геномными данными.
- MUMmer4 – изначально MUMmer разрабатывался для выравнивания геномов различных организмов.

SAM-файл

- Результат работы программ для выравнивания – SAM-файл.
- Формат сжатого, бинарного SAM файла – BAM.

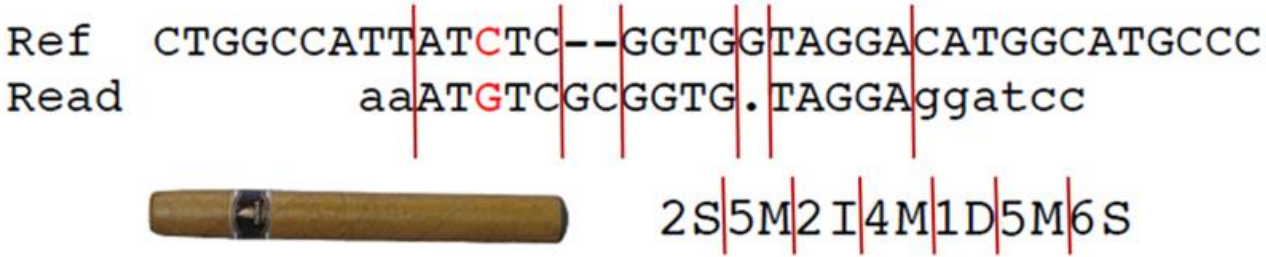


Each row describes a single alignment of a raw read against the reference genome. Each alignment has 11 mandatory fields, followed by any number of optional fields.

- Поля выравниваний в SAM формате содержат следующую информацию: **QNAME** (наименование рида), **FLAG**, **RNAME** (наименование референсной последовательности), **POS** (позиция самого левого нуклеотида референса, на который выравнивается этот рид), **MAPQ** (качество выравнивания – не прочитывания!), **CIGAR**, **RNEXT** (наименование следующего или спаренного с текущим ридом рида), **PNEXT** (его позиция), **TLEN** (длина выравнивания пары ридов), **SEQ** (последовательность рида), **QUAL** (качество прочтения).

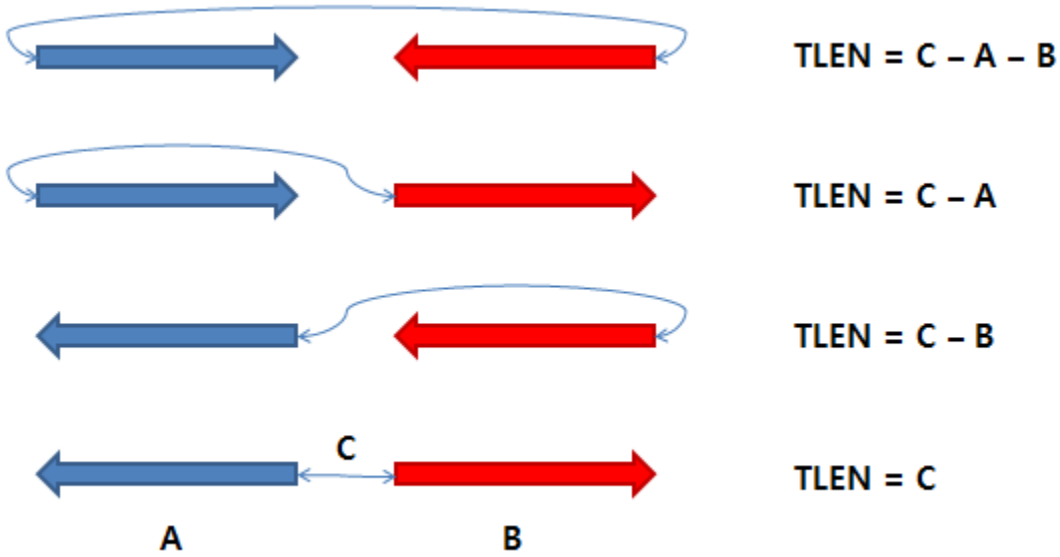
CIGAR string

- Строка CIGAR описывает, как фрагменты рида соответствуют/не соответствуют последовательности референса.
- M – match, I – insertion, D – deletion, S – soft clipping (в риде остались эти символы), H – hard clipping (символы удалены из последовательности рида в bam-файле)...



TLEN

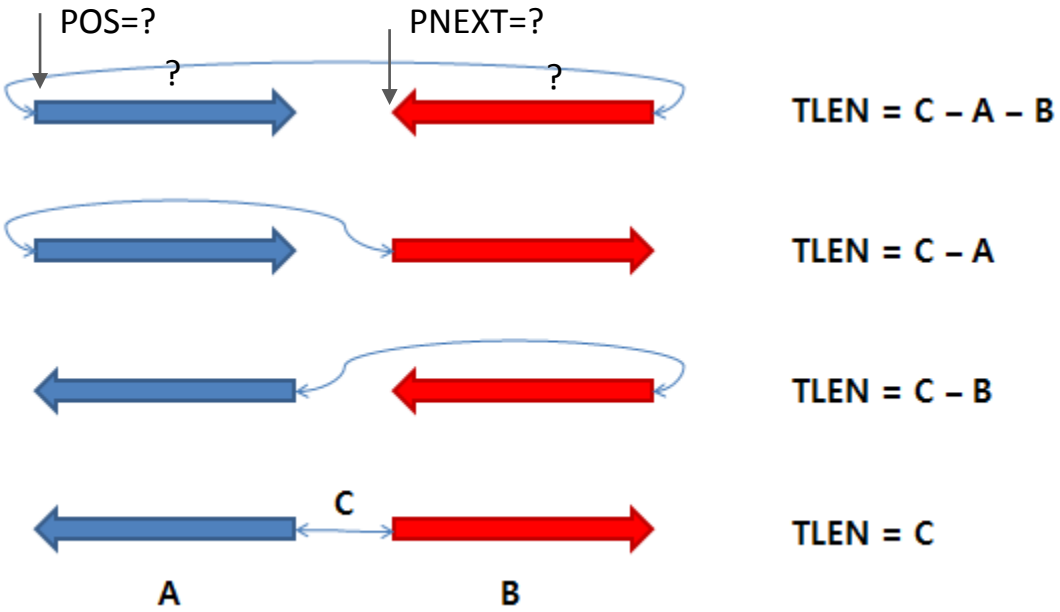
- Как рассчитывается длина выравнивания для пары ридов сразу?



TLEN

- Как рассчитывается длина выравнивания для пары ридов сразу (на примере пары ридов с ID = r001)?

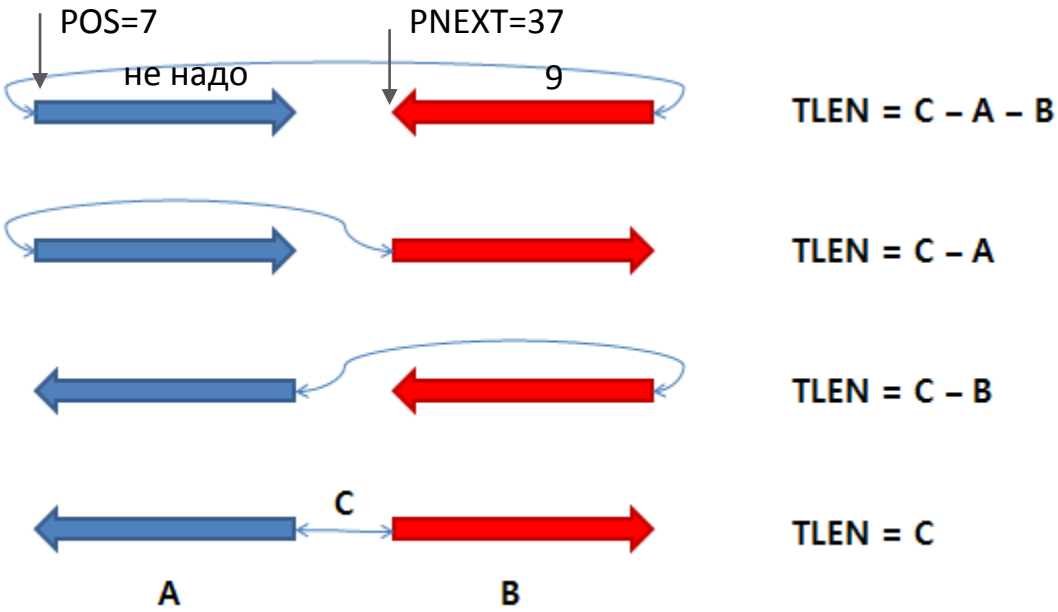
```
r001    99 ref  7 30 8M2I4M1D3M = 37
r002     0 ref  9 30 3S6M1P1I4M *  0
r003     0 ref  9 30 5S6M       *  0
r004     0 ref 16 30 6M14N5M    *  0
r003  2064 ref 29 17 6H5M       *  0
r001   147 ref 37 30 9M         =  7
```



TLEN

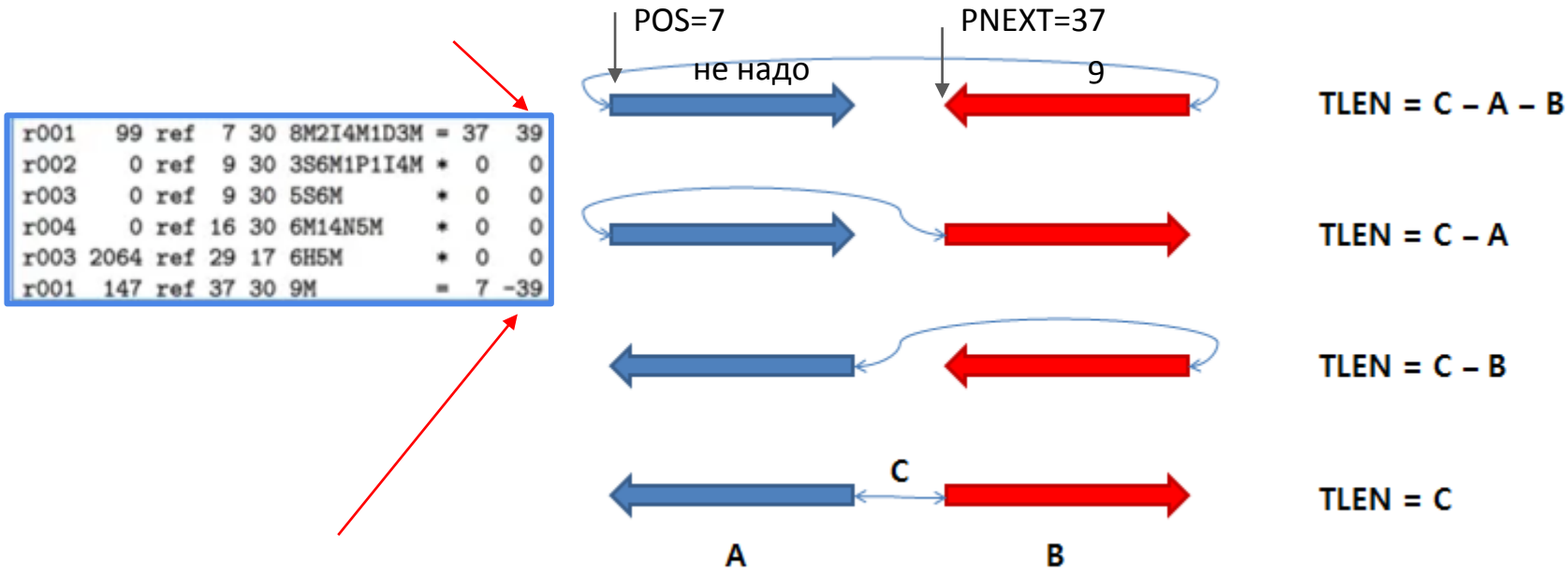
- Как рассчитывается длина выравнивания для пары ридов сразу (на примере пары ридов с ID = r001)?

```
r001    99 ref  7 30 8M2I4M1D3M = 37
r002     0 ref  9 30 3S6M1P1I4M *  0
r003     0 ref  9 30 5S6M       *  0
r004     0 ref 16 30 6M14N5M    *  0
r003  2064 ref 29 17 6H5M       *  0
r001   147 ref 37 30 9M         =  7
```



TLEN

- Как рассчитывается длина выравнивания для пары ридов сразу (на примере пары ридов с ID = r001)?



10-ричная	16-ричная	Описание
1	0x1	Спаренный рид
2	0x2	Оба риды в паре корректно выровнены
4	0x4	Рид не выровнен
8	0x8	NEXT (т.е. парный) рид не выровнен
16	0x10	SEQ – обратнокомплементарная последовательность
32	0x20	SEQ в NEXT – обратнокомплементарная последовательность
64	0x40	Первый рид из пары (NEXT – второй)
128	0x80	Второй рид из пары (NEXT – первый)
256	0x100	Вторичное выравнивание
512	0x200	Рид не проходит контроль качества прибора
1024	0x400	Дупликат
2048	0x800	Добавочное (supplementary) выравнивание

FLAG

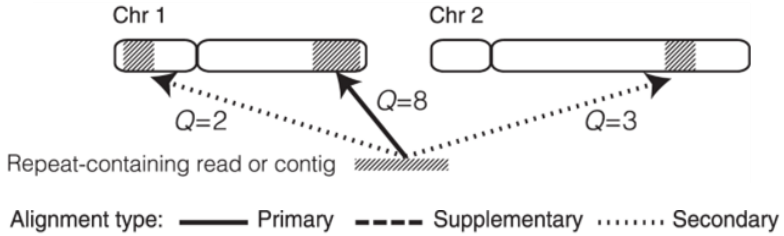
Если несколько флагов характеризуют данное выравнивание, их значения складываются – степени двойки используются для того, чтобы сумма трактовалась однозначно.
0 – исключение всего остального: неспаренный, корректно выровненный на прямую последовательность рид.

Химерное выравнивание

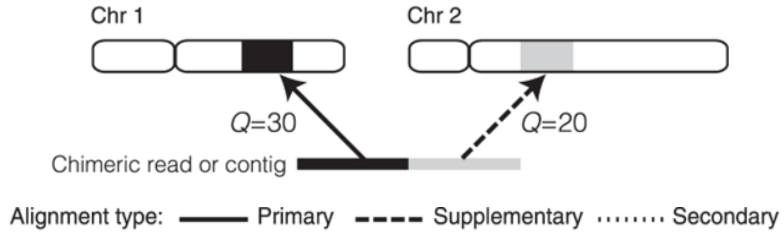
- Если рид целиком не может обеспечить формирование линейного выравнивания (которое не предусматривает изменения направления, а только замены и инделы), то, возможно, различные его участки могут сформировать несколько линейных выравниваний.
- Выравнивание сегментов рида сразу на несколько участков референсной последовательности называется химерным выравниванием. Химерное выравнивание занимает несколько строк с одинаковыми QNAME в SAM-файле.
- Химерное выравнивание содержит одно репрезентативное (определяется по-разному) и одно или несколько добавочных (supplementary).

Добавочные (2048) и вторичные (256) выравнивания

- Вторичное выравнивание появляется в том случае, если один и тот же рид одинаково хорошо выравнивается на несколько участков референса.



- Добавочное выравнивание – составная часть химерного выравнивания, выравнивание сегмента на референс, не являющееся репрезентативным.



samtools

- Перевод из SAM в BAM (экономит место):

```
samtools view -bS file.sam > file.bam
```

- Чтение BAM:

```
samtools view file.bam
```

- Отбор строк с определёнными флагами:

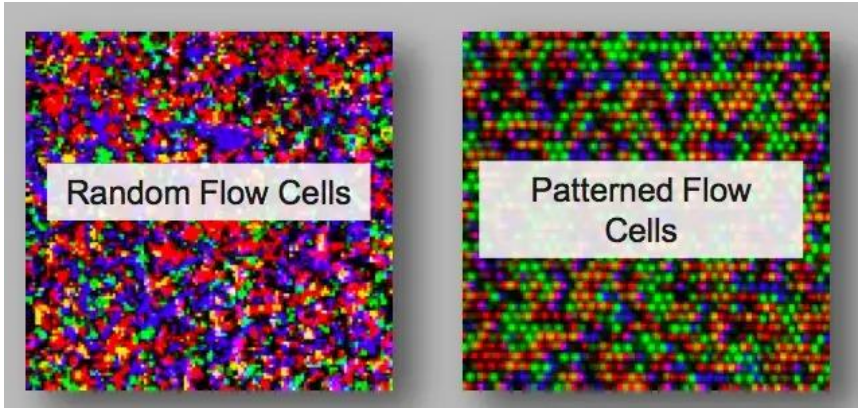
```
samtools view -F 0x4 file.bam
```

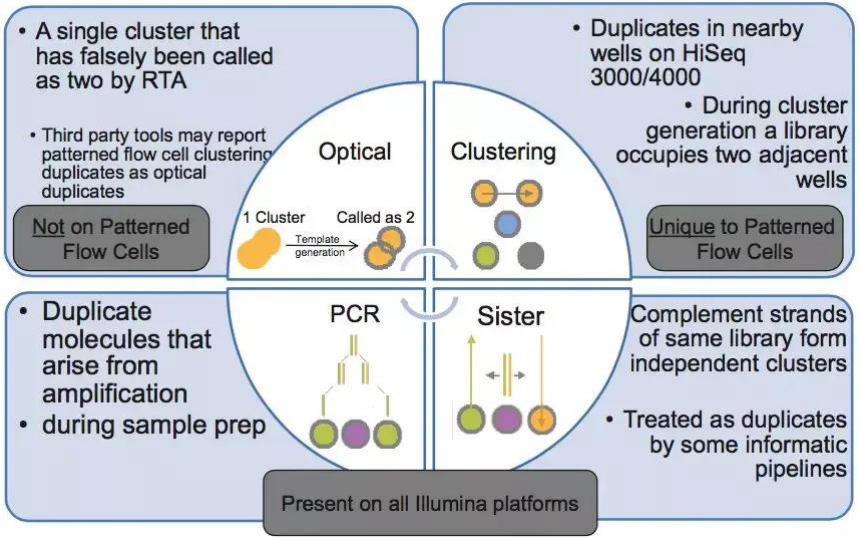
- Отбор строк без определённых флагов:

```
samtools view -f 0x4 file.bam
```

Дупликаты в данных

- Сестринские (от комплементарных цепей) – являются истинными биологическими дупликатами;
- ПЦР-дупликаты;
- Оптические дупликаты (иногда отдельно выделяют кластерные, см. рисунок – когда продукты одной молекулы занимают более одной ячейки – ExAmp).
- В упорядоченных ячейках (patterned flow cells) две проблемы: высокая концентрация => поликлональные ячейки; и низкая => ExAmp дупликаты.

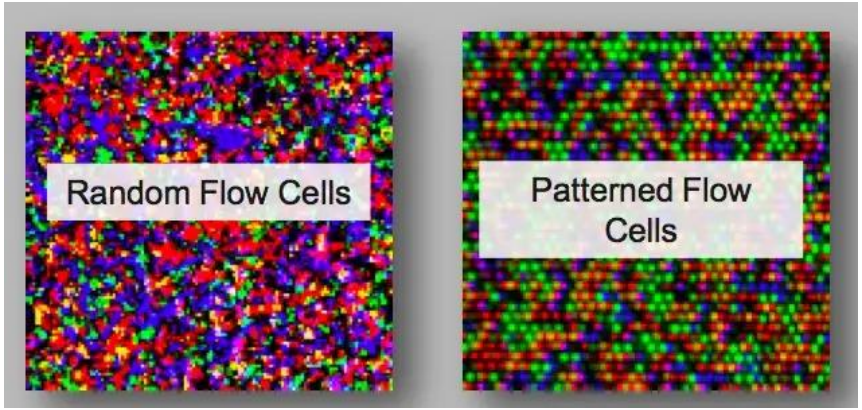




RTA – real time analysis

MarkDuplicates

- Все “ложные” дупликаты можно разделить на оптические и ПЦР-дупликаты.
- MarkDuplicates являются частью пакета Picard, а также включён в GATK.
- Алгоритм сравнивает 5’-концы ридов, и определяет группы дупликатов по их схожести. Наиболее качественный по сумме качеств нуклеотидов рид среди дупликатов считается первичным, остальные помечаются как дупликаты.



- В первую очередь оптические и ПЦР-дупликаты различают по расстоянию между кластерами дупликатов (OPTICAL DUPLICATE PIXEL DISTANCE):
  - Случайно распределённые ячейки – расстояние 100;
  - Упорядоченные – 2500.

Отчёт MarkDuplicates

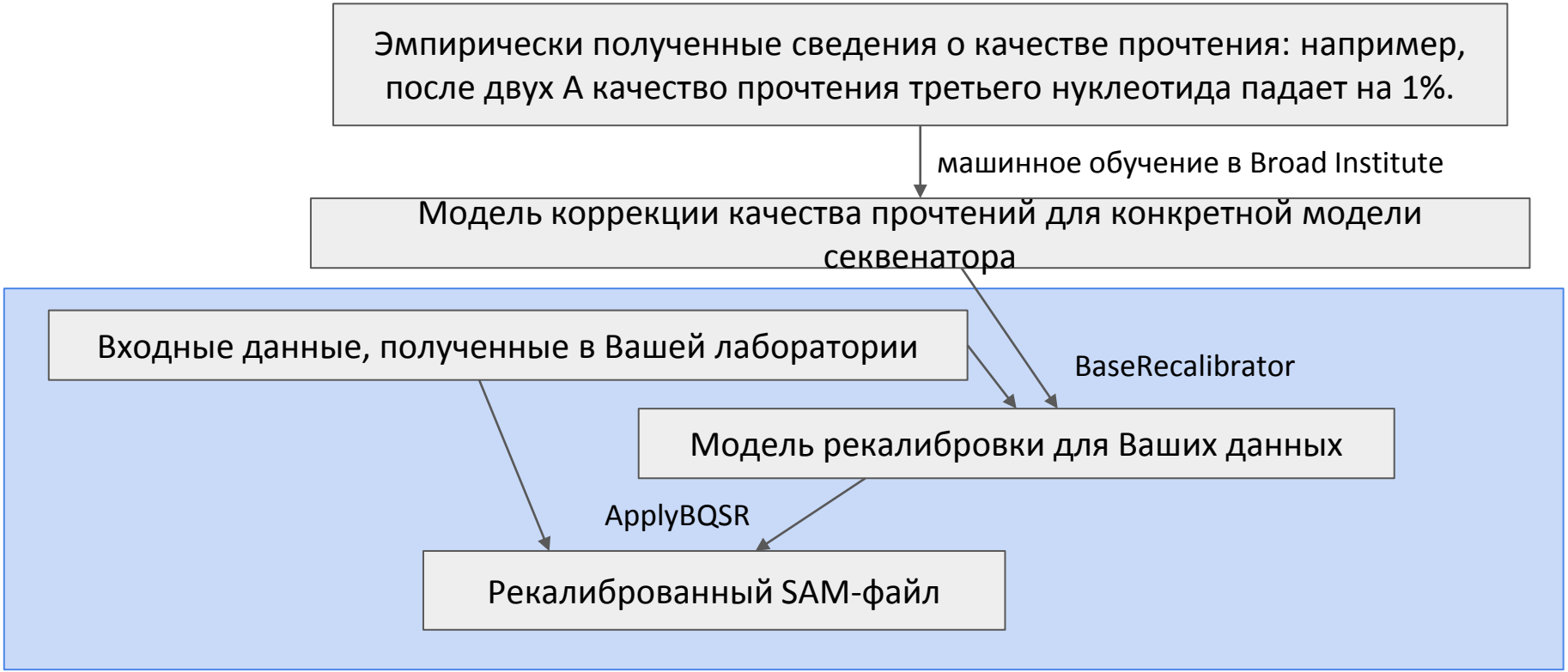
- В отчёте MarkDuplicates (metrics) указывается название библиотеки, её размер, число неспаренных и спаренных ридов, доля дупликатов в библиотеке и т.д.

1	2
LIBRARY	Unknown Library
UNPAIRED_READS_EXAMINED	0
READ_PAIRS_EXAMINED	135813
SECONDARY_OR_SUPPLEMENTARY_RDS	0
UNMAPPED_READS	0
UNPAIRED_READ_DUPLICATES	0
READ_PAIR_DUPLICATES	3584
READ_PAIR_OPTICAL_DUPLICATES	8
PERCENT_DUPLICATION	0.026389
ESTIMATED_LIBRARY_SIZE	2533249

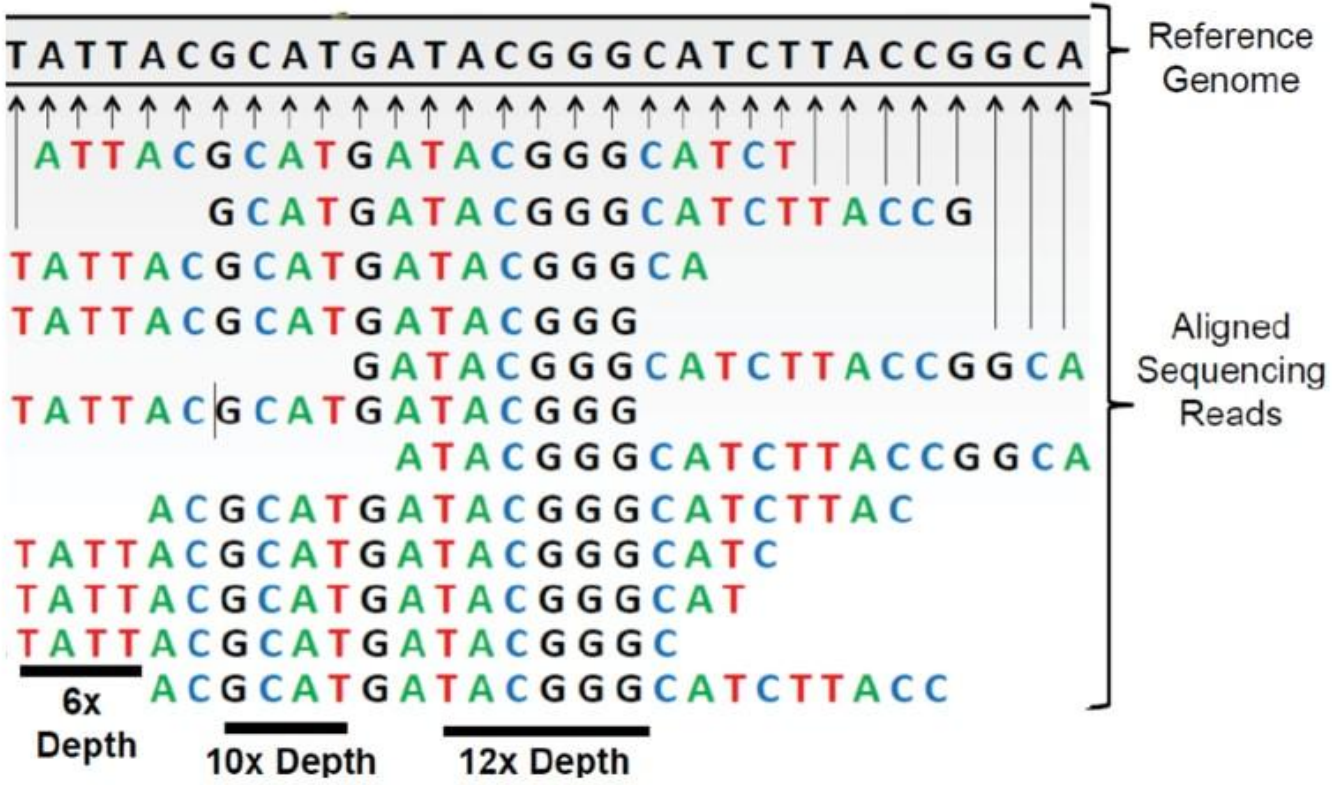
Рекалибровка качества

- Последний этап перед calling’ом – рекалибровка качества.
- Различные аспекты работы программ для обнаружения вариантов во многом зависят от показателей качества прочтения нуклеотидов.
- Однако, выдаваемые прибором показатели качества часто бывают скомпрометированы систематическими ошибками, приводящими к недооценке или переоценке качества прочтения в целом.
- Для переоценки качества существуют инструменты GATK, BaseRecalibrator и ApplyBQSR, BQSR расшифровывается как Base Quality Score Recalibration.

Как работает BQSR



SAM/BAM


TATTACGCATGATACGGGCATCTTACCGGCA
Reference Genome
ATTACGCATGATACGGGCATCT
GCATGATACGGGCATCTTACCG
TATTACGCATGATACGGGCA
TATTACGCATGATACGGGG
GATACGGGCATCTTACCGGCA
TATTACGCATGATACGGG
ATACGGGCATCTTACCGGCA
ACGCATGATACGGGCATCTTAC
TATTACGCATGATACGGGCATC
TATTACGCATGATACGGGCAT
TATTACGCATGATACGGGC
ACGCATGATACGGGCATCTTACC
Aligned Sequencing Reads
6x Depth
10x Depth
12x Depth