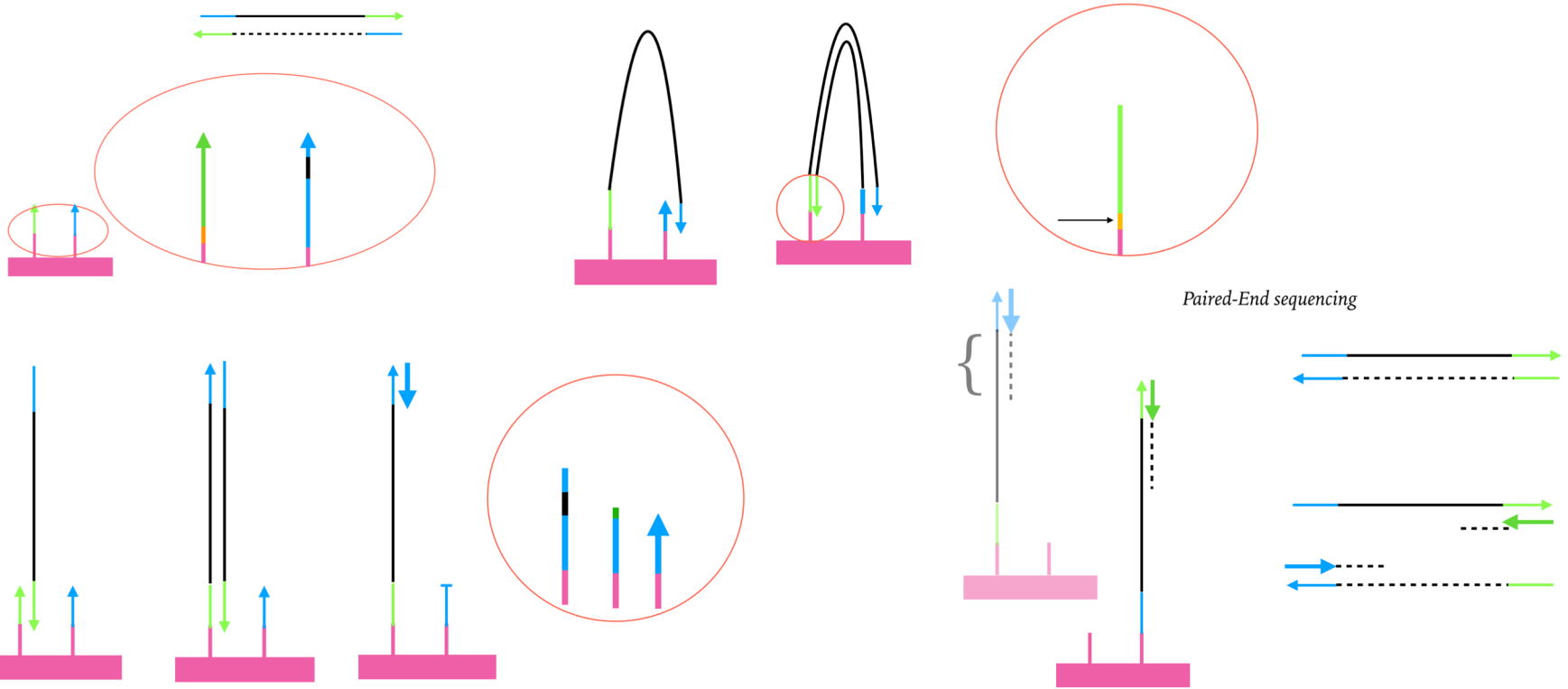
Картирование

Герасимов Евгений
jalgard@gmail.com

Парные чтения (Illumina)



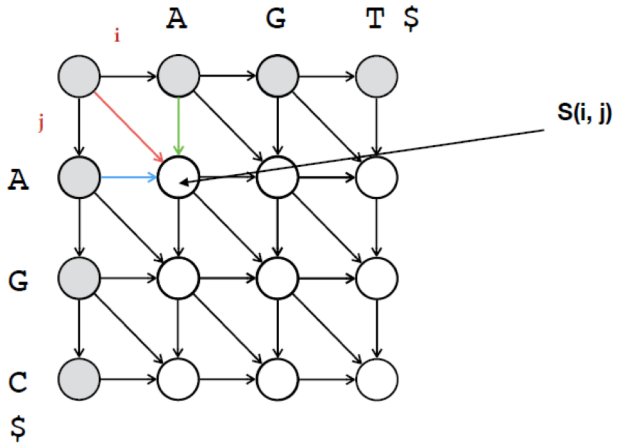
Выравнивания

```
 Y   P   S   Y   L   H       A   D   D       S   A   V   V   Q   A   A   S   A
CTATCCGTCCTATCTGC-ATGCCGATGAC--TCCGCCGTCGTTCAAGCGGCTTCGGCGGC
CTATCCGTCGT-TCCGCTATACGGAAGCCGATCCG--GTCGT----GAGGC-TCGCCGTC
 Y   P   S   F       R   Y   T   E   A   D   P       V   V           R   L       A   V
```

	C	T	A	T	C	T	G	C	A	T	G
G	↘										
T		↘	→								
T				↘							
C					↘						
C						↘					
G							↘				
C								↘			
T									↓		
A										↘	
T											↘
A											

$$S(i,j) = \max \begin{cases} S_{(i-1,j-1)} + M(i,j) \\ S_{(i,j-1)} + G(i) \\ S_{(i-1,j)} + G(j) \end{cases}$$

G = -1
M(i == j) = +2
M(i != j) = -1



Задача картирования

Референсный геном

AGTAGTTTTCTCAAGTTATTATTTTTGTAAGTGTATGTTAATATCTCTAGCAAATTTAAGAAGGAATTCC
ATTTAATTCTAAATATGCAAAAATGAAATTTTTTACAAACATAGAATAATATAAATTCATAAACCTTAAA
AAACAAATTCATTTGATATTTGATAAATAAATTAAGTAAGGCAAATTTAATCTTGATTCGTAAGGTATGT
CATATTTGTGCGTAGTCTAATCCCAATCCCAATATTTGTAAGCGGTTTCACTAAATATGTTATTTTTTTA
TTTGTATATAATATGGCATTATGAAAATGTAAACTTATCTTTTATTATAAATTTAAAAACAAATTCTTCT
ATTTCAAAATCCCAAAAACAACTTAAAATTTGTAAATATGTTTTACCTATCCAAATTGTTATTTTCATTT
GTACTTGTACAAACTTTTTTAACAAATTTAAAAAACAAAATTCCTTTCTTTCCACTTTAGAATTTAAAAG
AAATTAAAACGCACCAAATAAATAAATATAAATCTTTTTAAATGTTTAGATAGTTTTATCCAAATTGTTA
TTTTTGTAAATAAGCATAAATCCTTTAGCCTTAAAAAATCTAATAATAAATTAAACACCCCTGATGAAGA
...

AAATCCTTTAGCCTT **AGGGGTGTTTAATTT**

Задача картирования

Референсный геном

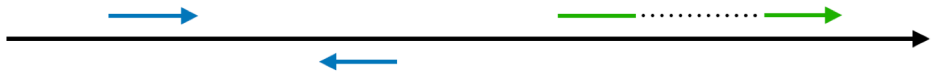
GTACTTGTACAAACTTTTTTAACAAATTTAAAAAACAAAATTCCTTTCTTTCCACTTTAGAATTTAAAAG
AAATTAAAACGCACCAAATAAATAAATATAAATCTTTTTAAATGTTTAGATAGTTTTATCCAAATTGTTA
TTTTTGTAAATAAGCATAAATCCTTTAGCCTTAAAAAATCTAATAATAAATTAAACACCCCTGATGAAGA
AACAAACACAAACATTCCATAAATAATTACCCCTTTAAGCGAAATTATACGGTGTAAAAAACAAAGTCCC
CATTTCATAATTCAAAAAACAAGATTTTTACAAATGTAAATCCTTATGAAAATGAAACAAAAACAAAATA
ATAAAGAGAATCTAATTATGGTATTTGTAAACAATTTCTCACAAATCCTAAAAAACAAAAATAAAGTTAT
TTAACTTTGTAAATTACAAATTAAAAAGATAAATCTTTTATGCTAATATCATATACACAATAACCATTAC
TTTATGTAAATTAAATCTTTAAGTAGTTTCCCTAAATTATTATTTTTGTAAGTATGTATAAATGTTTCTA
...

AAATCCTTTAGCCTT

AGGGGTGTTTAATTT

Терминология

- Reference
- Mapping / alignment
- Forward / reverse strand
- Paired-end / Mate-pair
- Primary / secondary alignment
- Chimeric / junction / split alignment
- Unique / multimap / unmapped
- Concordant / discordant
- Insert size



Особенности задачи

- Геном (текст) достаточно большой (~3Gb геном человека)
- Геном известен заранее и не изменяется
- Чтений (паттернов) очень много, но они все короткие
- Чтения очень похожи (часто идентичны) геному
- Каждое чтение можно обрабатывать независимо

- Алгоритм может требовать много ресурсов (оперативной памяти)
- Геном можно подготовить заранее
- Алгоритм может требовать много ресурсов (процессорного времени)
- Можно использовать ***алгоритмы поиска точного совпадения***, а не выравнивания
- Задача будет хорошо параллелиться

Препроцессинг генома

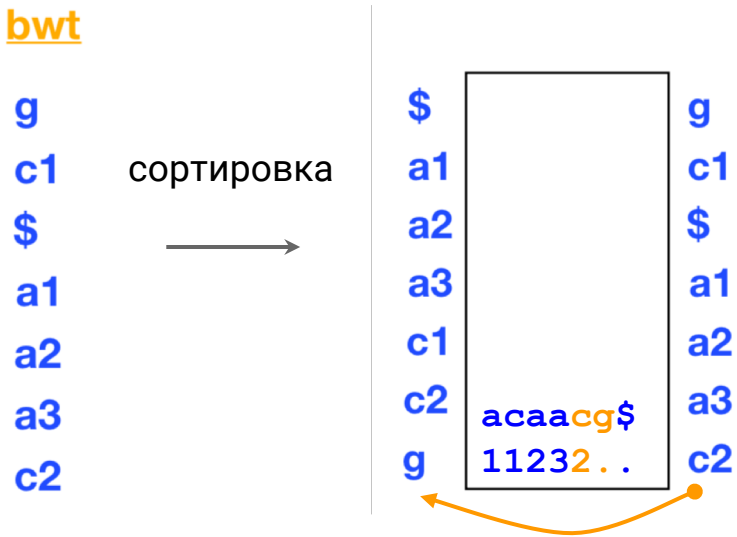
Преобразование Барроуза — Уилера

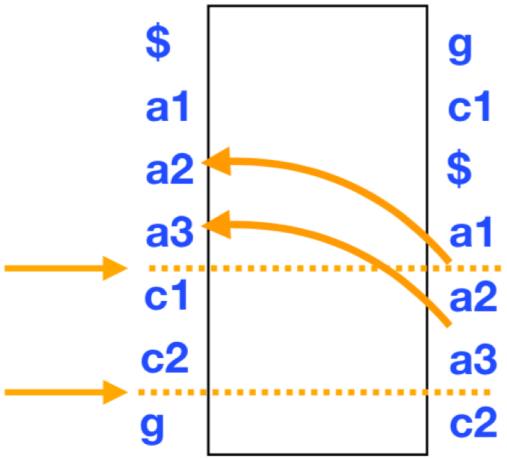
Полезные свойства BWT

1. Сжатие (без потери) acaacg$

   gc$3ac_

2. Быстрый поиск





собственно поиск

Дальнейшее развитие темы: FM-index

$$LF(i) = C[\, L[i] \,] + Occ(\, L[i],\ i \,)$$

bwt

g
c1
$
a1
a2
a3
c2

C:

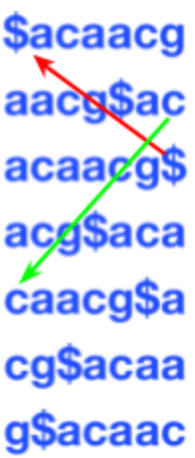
$	0
a	1
c	4
g	6

Occ:

	1	2	3	4	5	6	7
$	0	0	1	1	1	1	1
a	0	0	0	1	2	3	3
c	0	1	1	1	1	1	2
g	1	1	1	1	1	1	1

$$LF(3) = C[\, L[3] \,] + Occ(\, L[3],\ 3 \,) = C[\, \text{‘\$’} \,] + Occ(\text{‘\$’}, 3) = 0 + 1 = 1$$

$$LF(2) = C[\, L[2] \,] + Occ(\, L[2],\ 2 \,) = C[\, \text{‘c’} \,] + Occ(\text{‘c’}, 2) = 4 + 1 = 5$$



Реальные инструменты, использующие BWT / FM-index

bowtie / **bowtie2**:

- использует FM-index для поиска seeds, seed может иметь 0, 1 или 2 замены
- каждый найденный ‘seed’-локус продляется с использованием Smith-Waterman
- имеется режим end-to-end (‘глобального’) и local (‘локального’ выравнивания) работы с ридом
- есть некоторые технические ограничения на количество seeds, которое находится для рида

BWA:

- использует FM-index, чтобы найти области максимального точного совпадения (mem)
- использует Smith-Waterman для продления выравнивания
- может быть использован для ридов очень большой длины (PacBio)
- каждый рид может иметь несколько локальных выравниваний (разные части длинного рида)

Преимущества и недостатки

\+

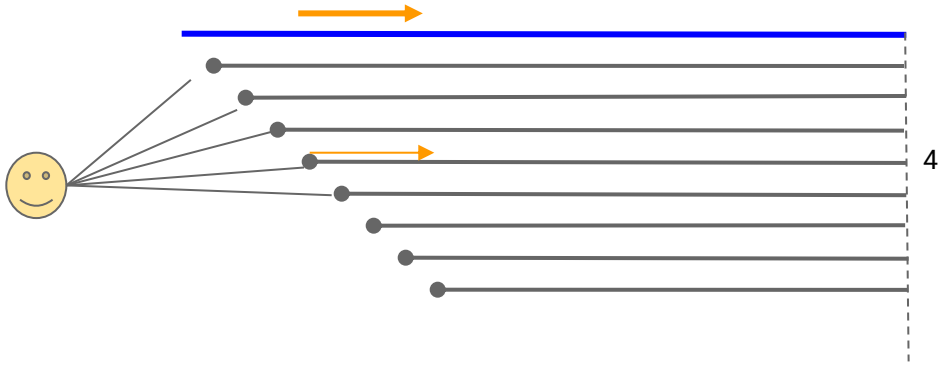
- геном может храниться в сжатом виде
- найти seeds / mems довольно “просто”
- хорошее сочетание скорости и качества выравнивания (прохождения пространства поиска)
- лучшее соотношение скорости к размеру индекса в памяти
- отлично адаптируются к длинным ридам

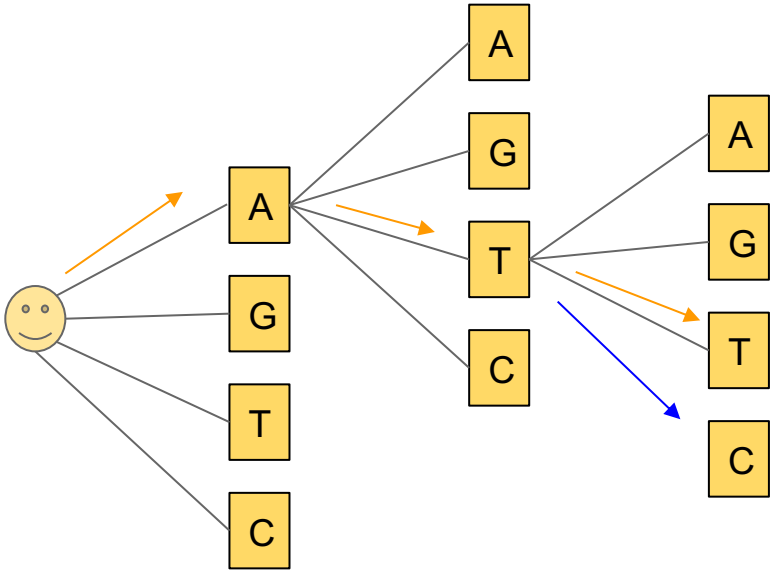
\-

- не эффективно искать seed, если есть несовпадения (mismatches), с коротким seed сильно теряется скорость работы
- время поиска seed по-прежнему ~log(геном)
- нет четких оснований, по которым выбирается оптимальная длина seed

Другие подходы: суффиксы и префиксы

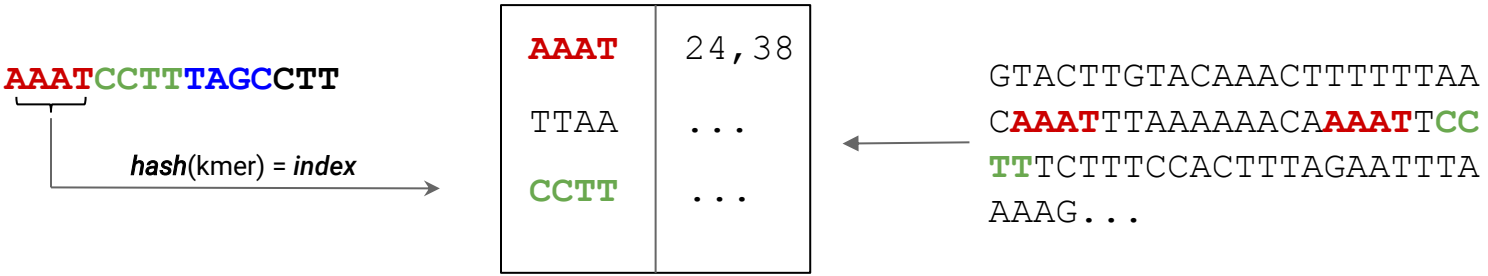
Суффиксное дерево





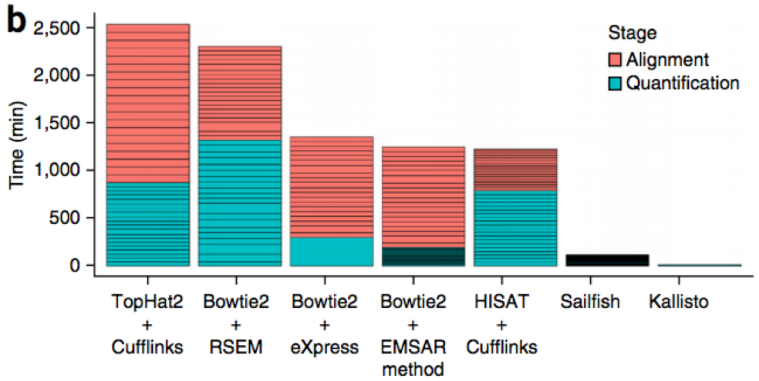
Другие подходы: хэширование

Хэш-таблицы



“Alignment-free” методы:

- используют частоты kmer-ов для кластеризации ридов между заданным набором образцов
- быстры (очень)
- используются в метагеномике и транскриптомике



doi:10.1038/nbt.3519

Некоторые замечания

По разным данным FM-index может превосходить по скорости отдельные хэш-картировщики = многое зависит от конкретной реализации

Сравнение особенностей алгоритмов

\+

- хэш-основанные алгоритмы - самые быстрые и самые (или одни из самых) эффективных в плане требований по памяти
- алгоритмы на суффиксных деревьях тоже очень быстрые (сравнимо с хэш), но позволяют получать гораздо более оптимальное выравнивание (как?)

\-

- хэш-основанные подходы плохо работают с ридами, если в них есть ошибки и замены, еще хуже - индели
- хэш-основанные алгоритмы, строго говоря, не дают выравнивания
- алгоритмы на суффиксных деревьях очень требовательны к памяти

Слишком короткие чтения

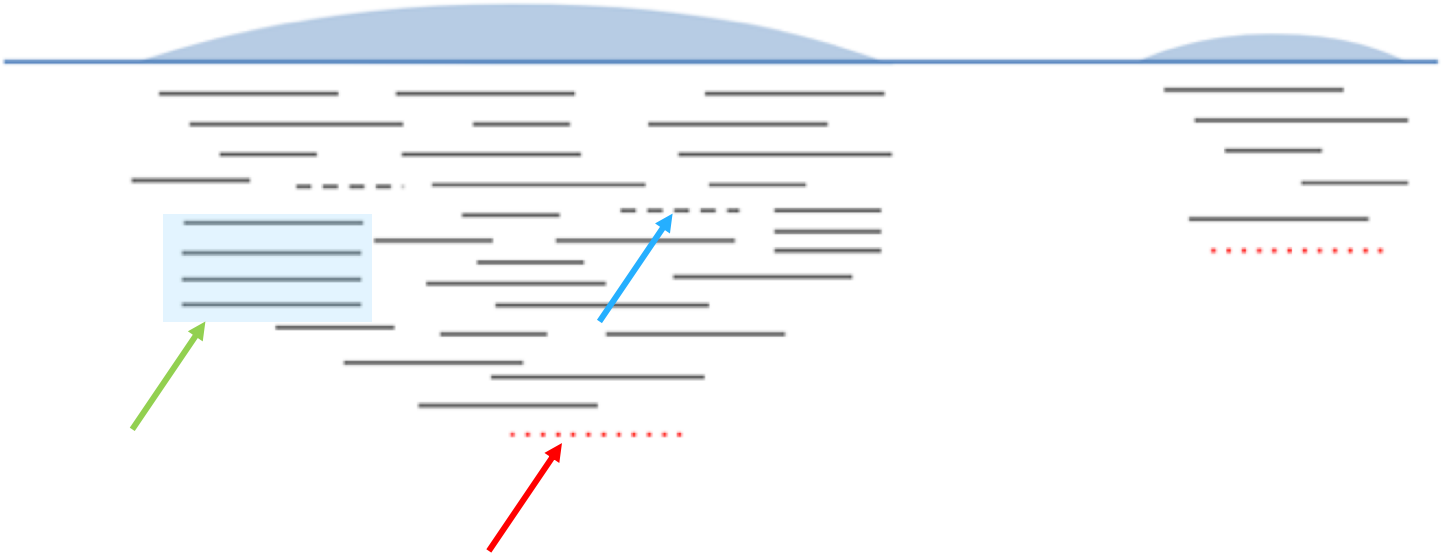
В коротком тексте слишком мало информации

Технические проблемы:

- недостаточно kmer-ов можно извлекать из рида, недостаточное количество хороших seeds можно подобрать
- выравнивание всегда будет иметь невысокий score, легче найти альтернативное выравнивание с таким же score

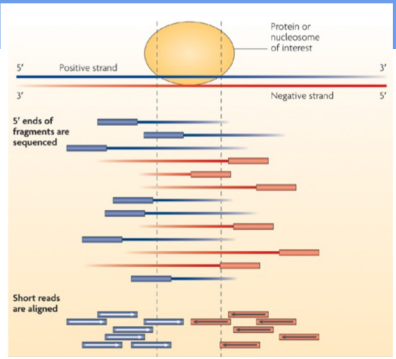
Короткие чтения, как правило, картируются в несколько мест (не уникально)

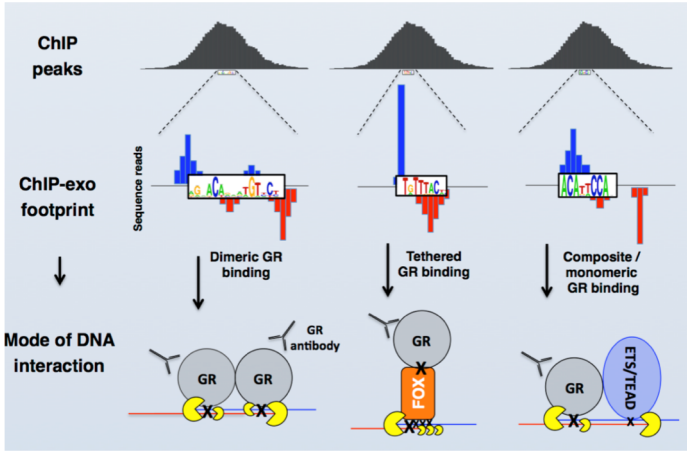
Профиль покрытия



О чем говорит профиль покрытия?

- Структурные вариации разного масштаба
  - SNP
  - Indels
  - CNV
  - Fusion
- Экспрессия генов
- Взаимодействия молекул
  - ChIP-seq
  - HiC
  - Mnase-seq
  - CAGE





SNP-calling

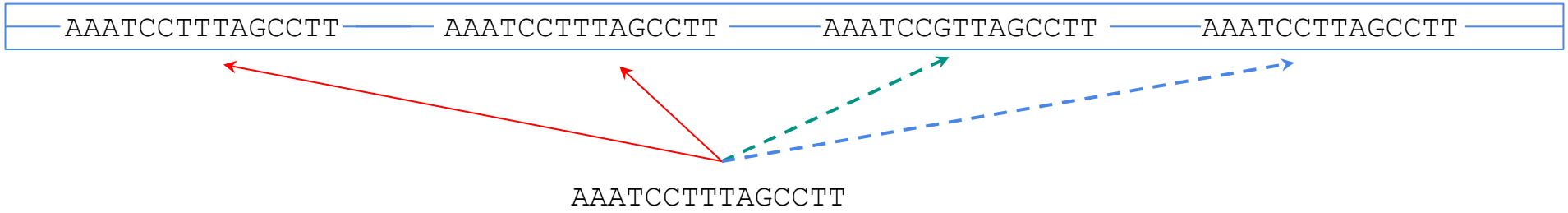
SNP нельзя найти, если разрешить только картирование с полным совпадением!

Некоторые конкретные следствия:

- SNP-calling будет зависеть от выбора картировщика
- SNP-calling будет зависеть от настроек картировщика
- SNP-caller должен уметь фильтровать потенциальные баги картирования
- SNP не может находиться в seed / kmer
- Сближенные SNP сложнее “ловить”

Уникальность картирования

разные программы по-разному подходят к вопросу об уникальности картирования



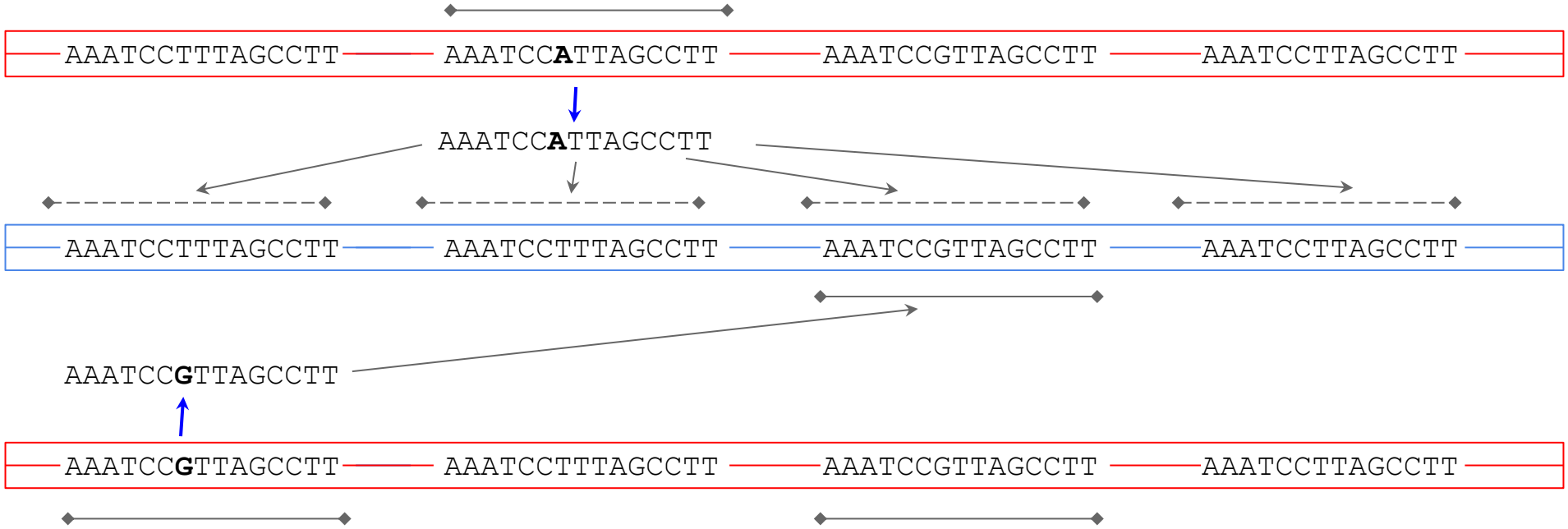
специальные тэги в sam-файле:

bwa:

XT:A:R

Нельзя просто смотреть на MAPQ!

Эффект “притяжения” к референсу



Сложности с интронами

split alignment и chimeric alignment

```
AAACAAATTCATTTGATATTTGTATAAATAAATTAAGTAAGGTCAAATTTAATCTTGATTCGTAAGGTATGT
          ATTTGATATTT--------------------GTCAAA (ПРАВИЛЬНО)
          ATTTGATATTTGT------------------CAAA
```