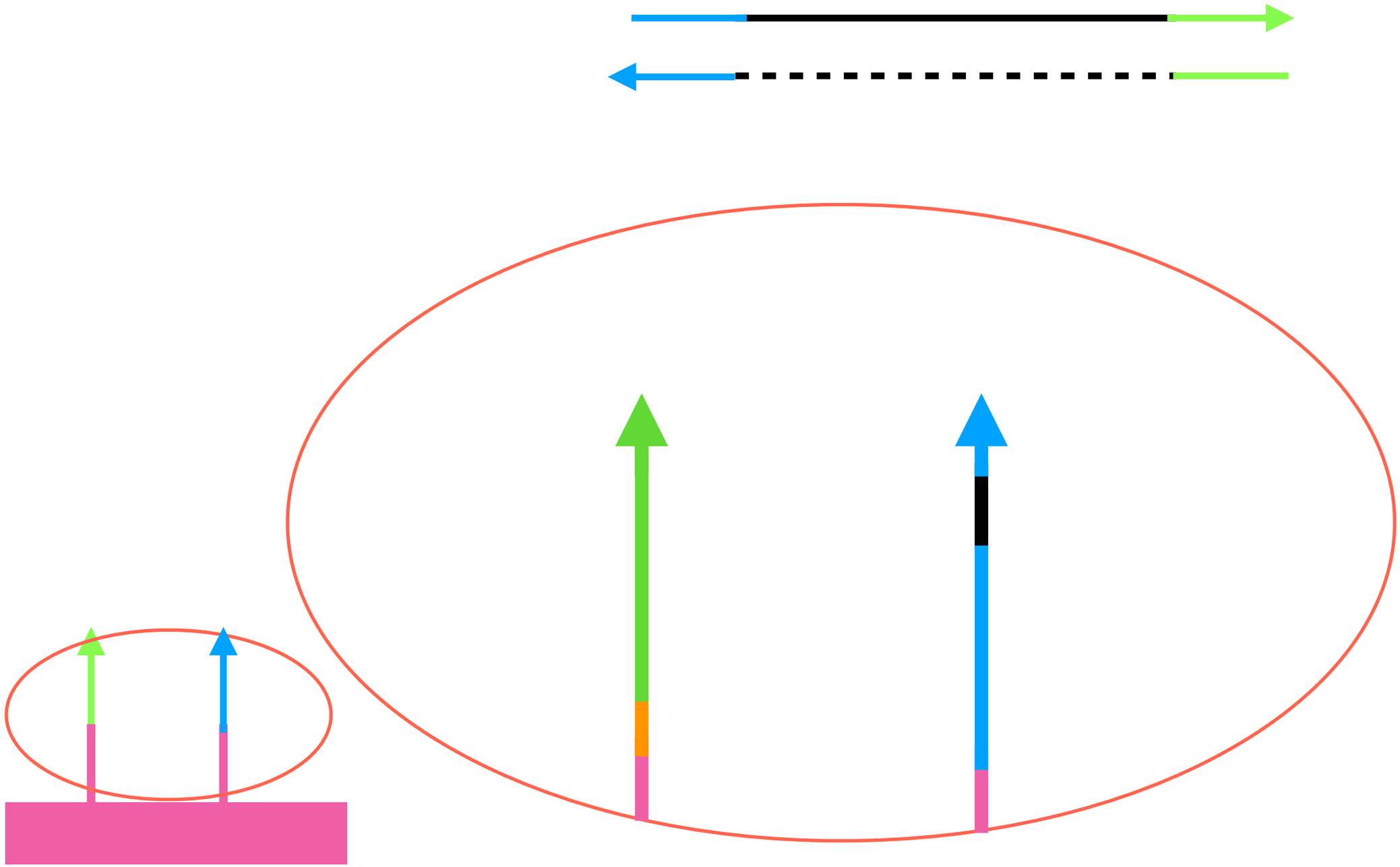
Биоинформатика

2018

Лекция 6
Герасимов Евгений Сергеевич
jalgard@yandex.ru

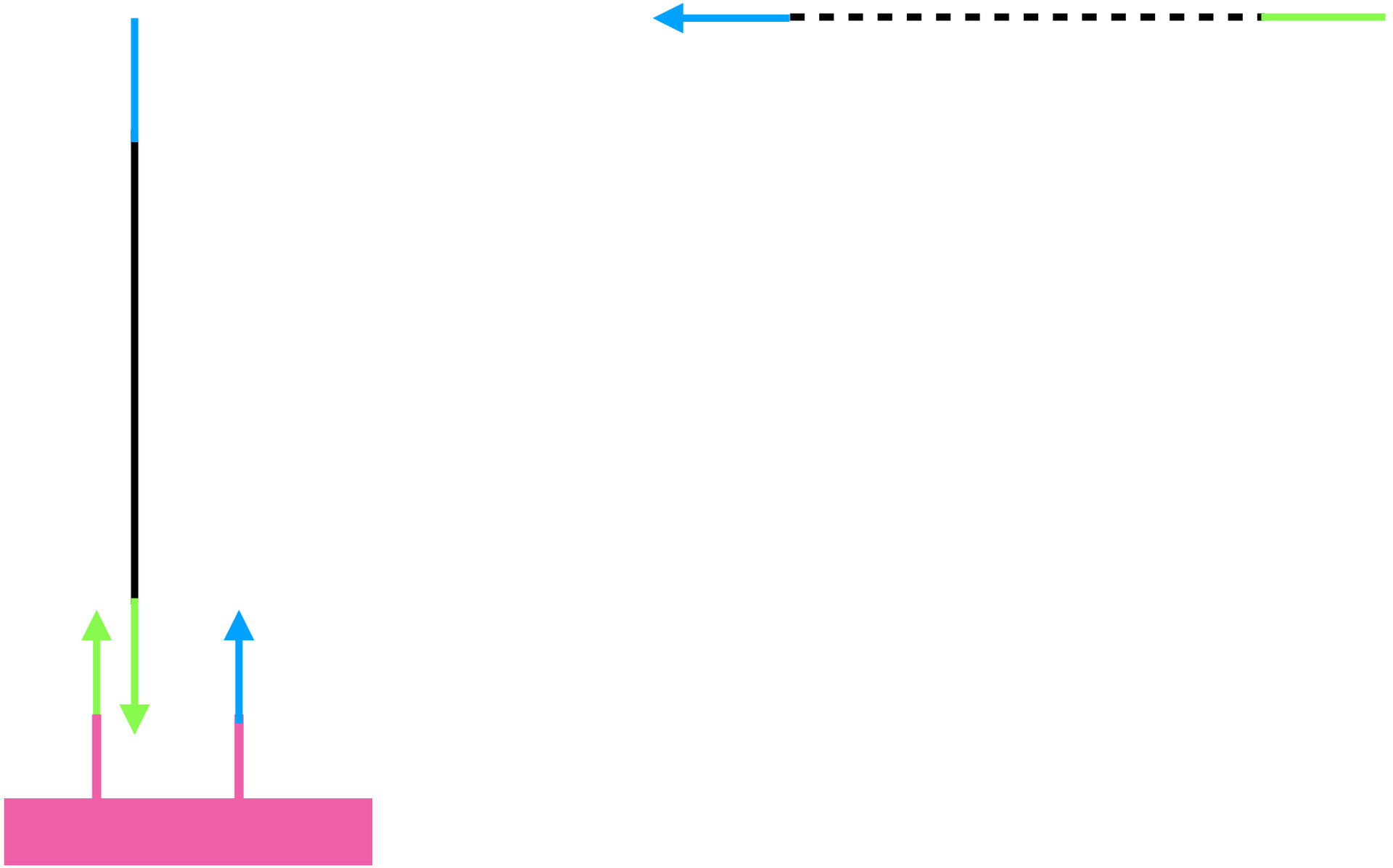
Парно-концевые риды

Paired-End sequencing



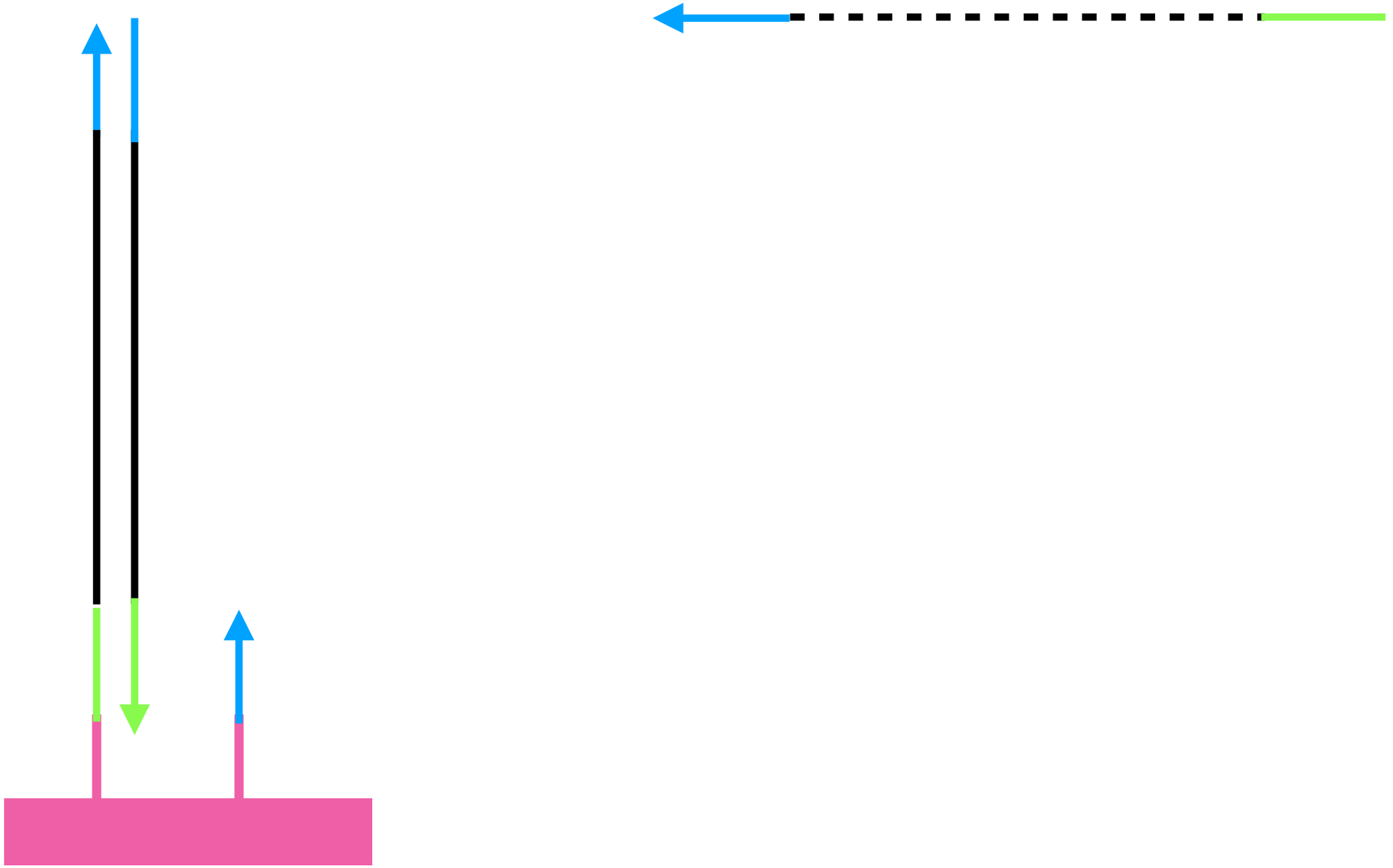
Парно-концевые риды

Paired-End sequencing



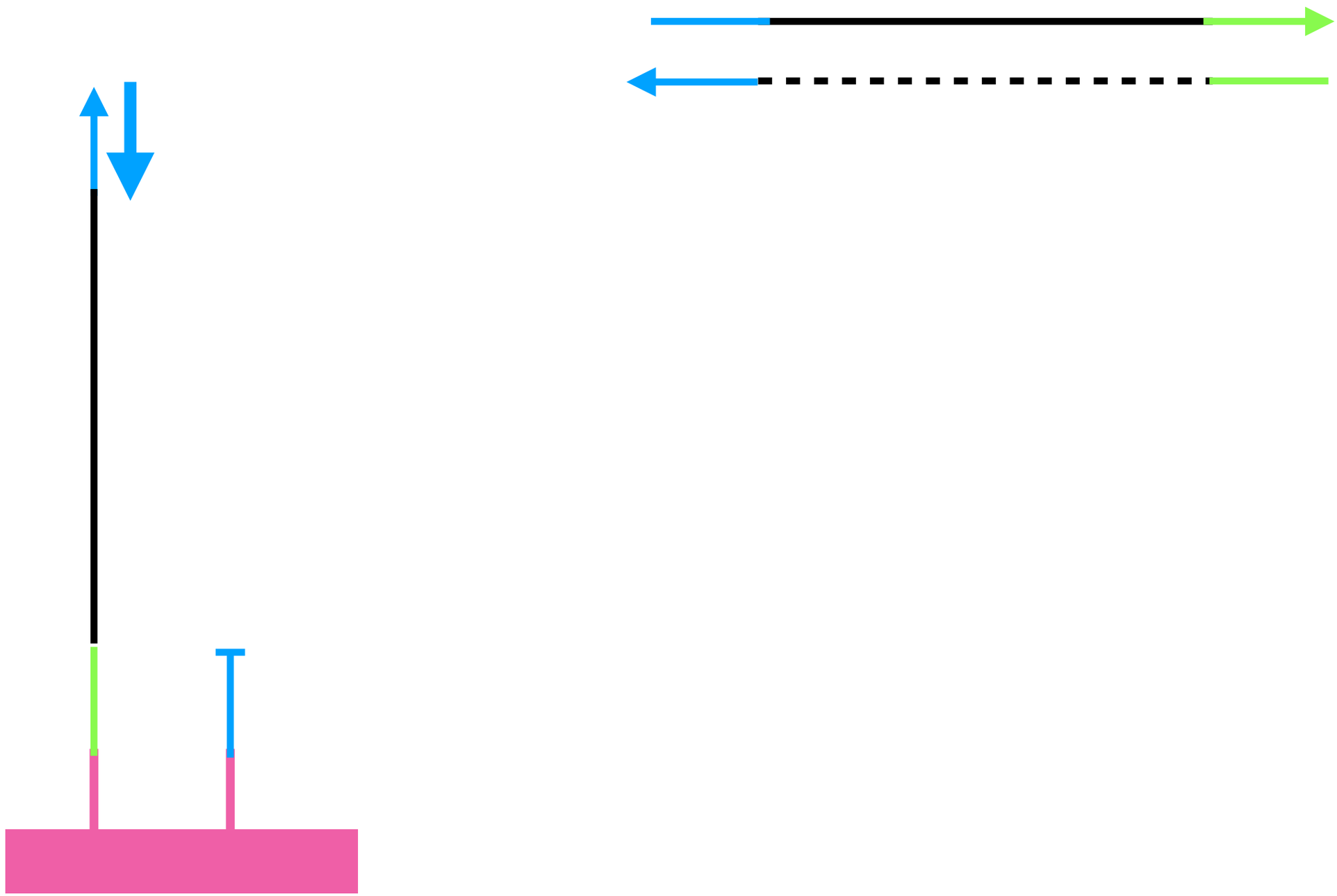
Парно-концевые риды

Paired-End sequencing



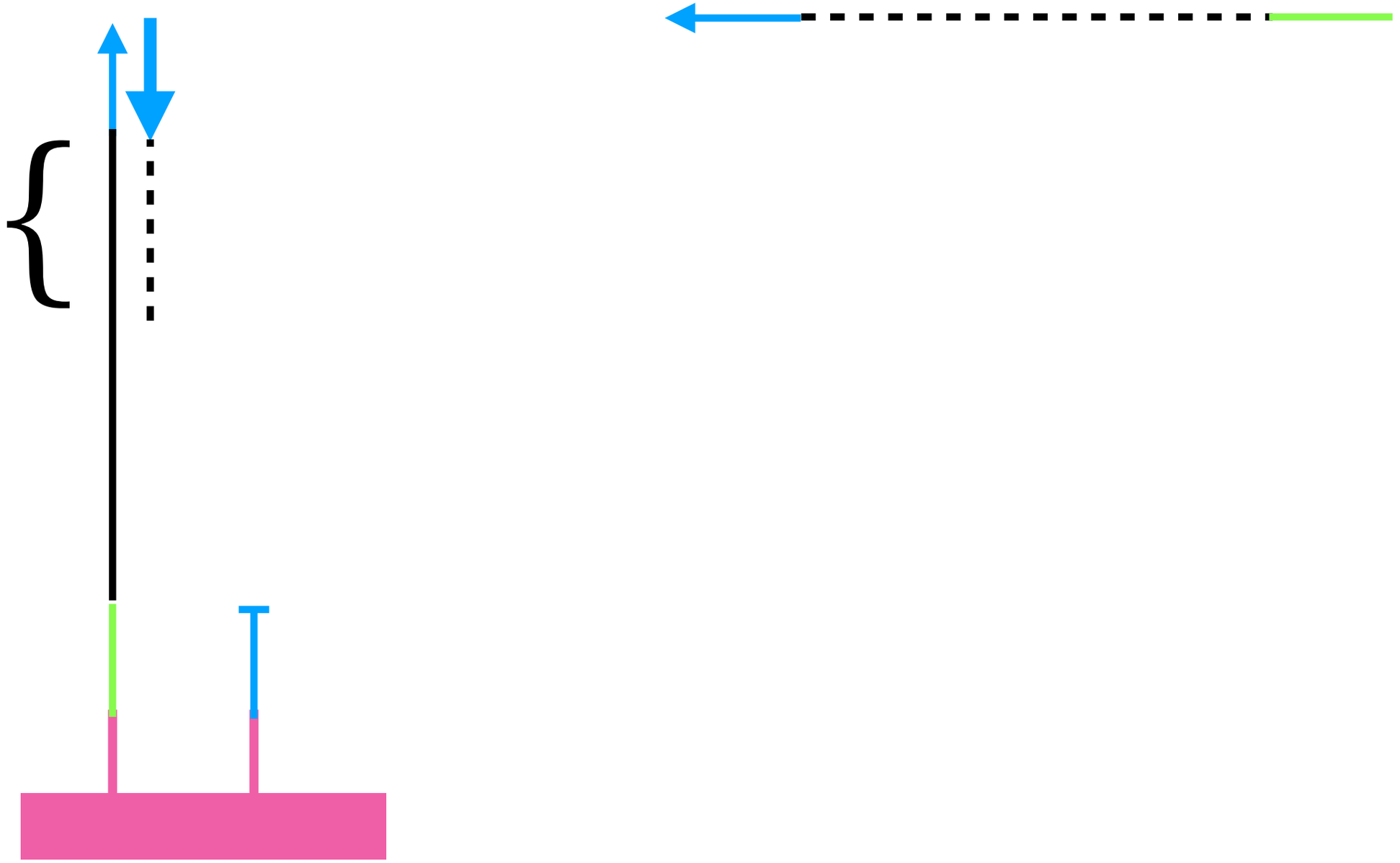
Парно-концевые риды

Paired-End sequencing



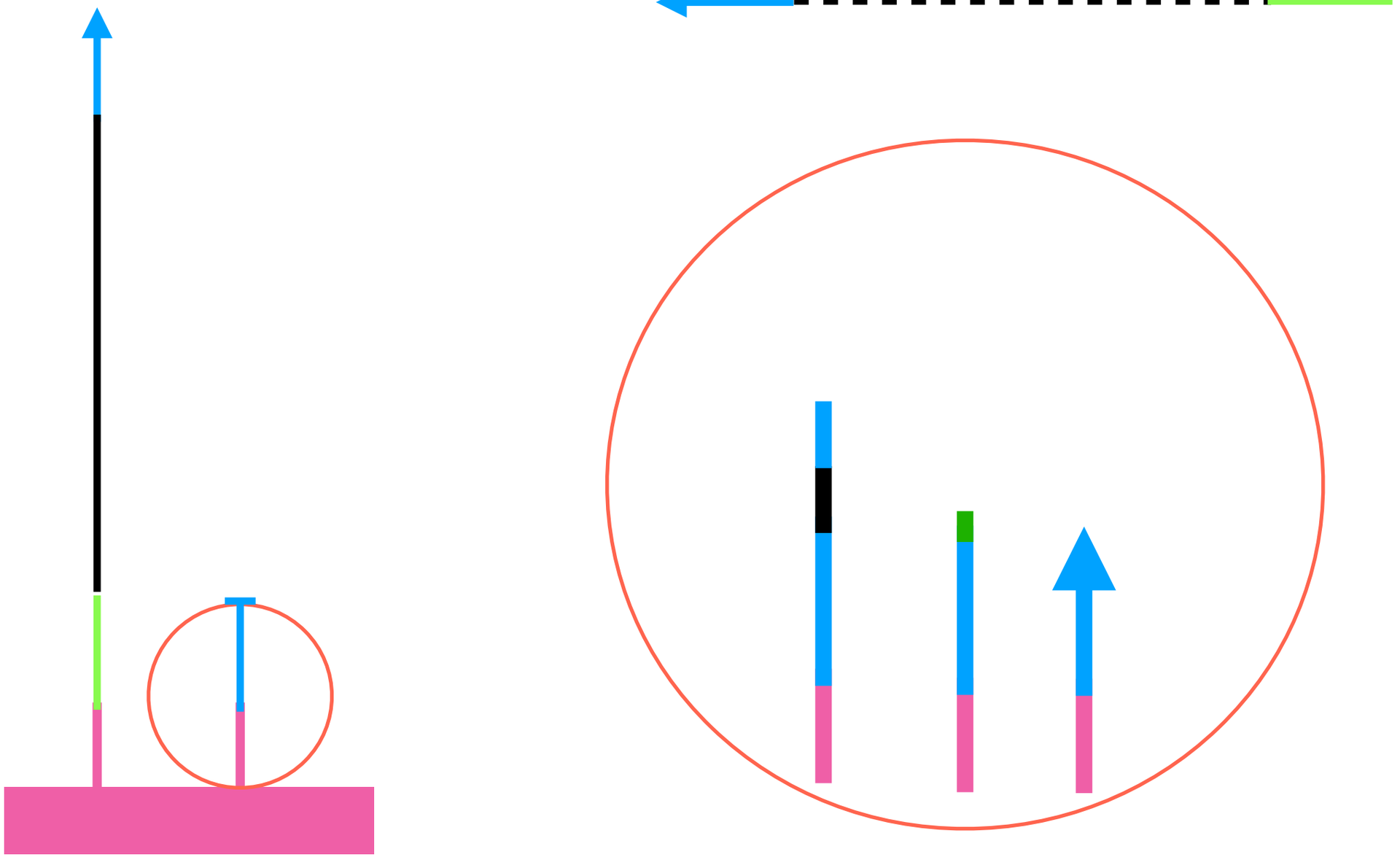
Парно-концевые риды

Paired-End sequencing



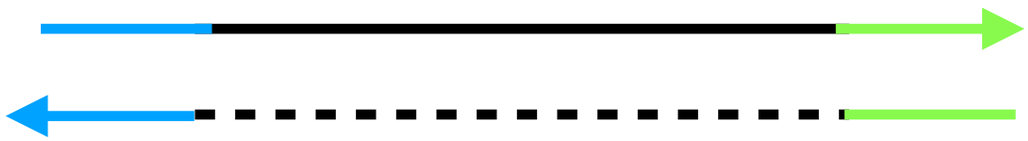
Парно-концевые риды

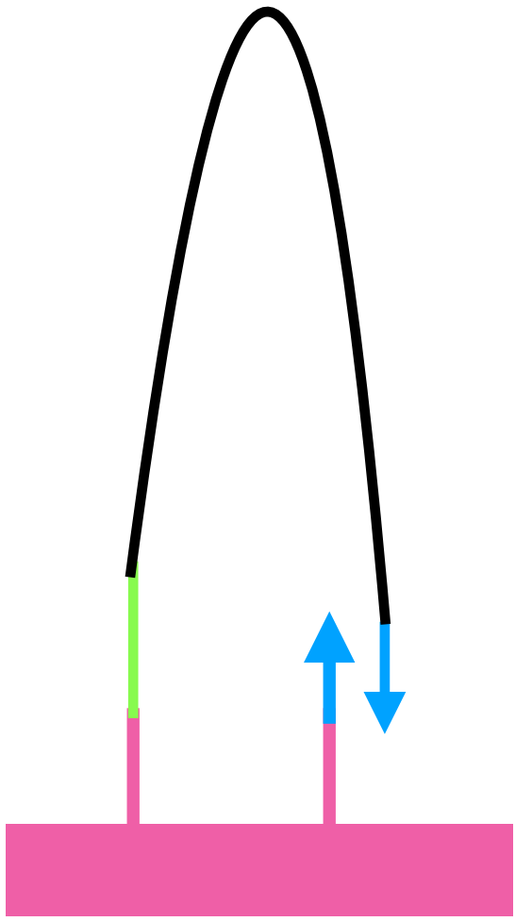
Paired-End sequencing



Парно-концевые риды

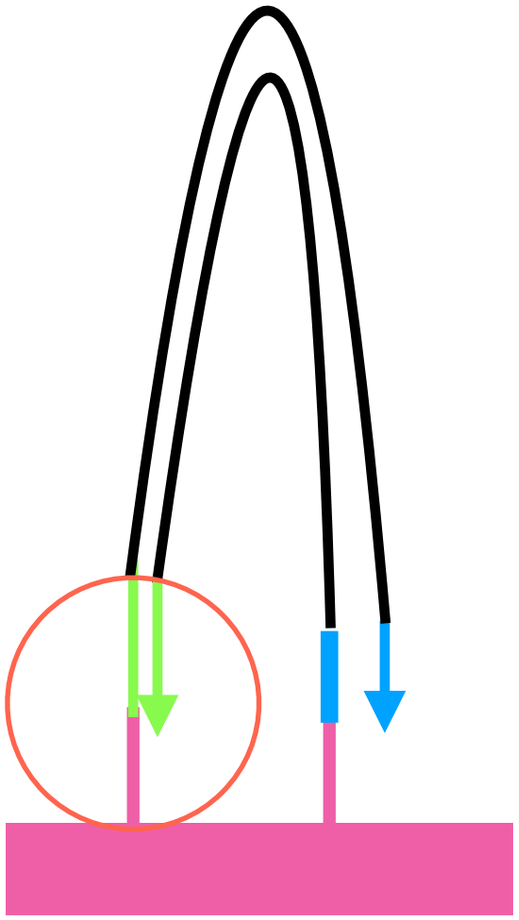
Paired-End sequencing

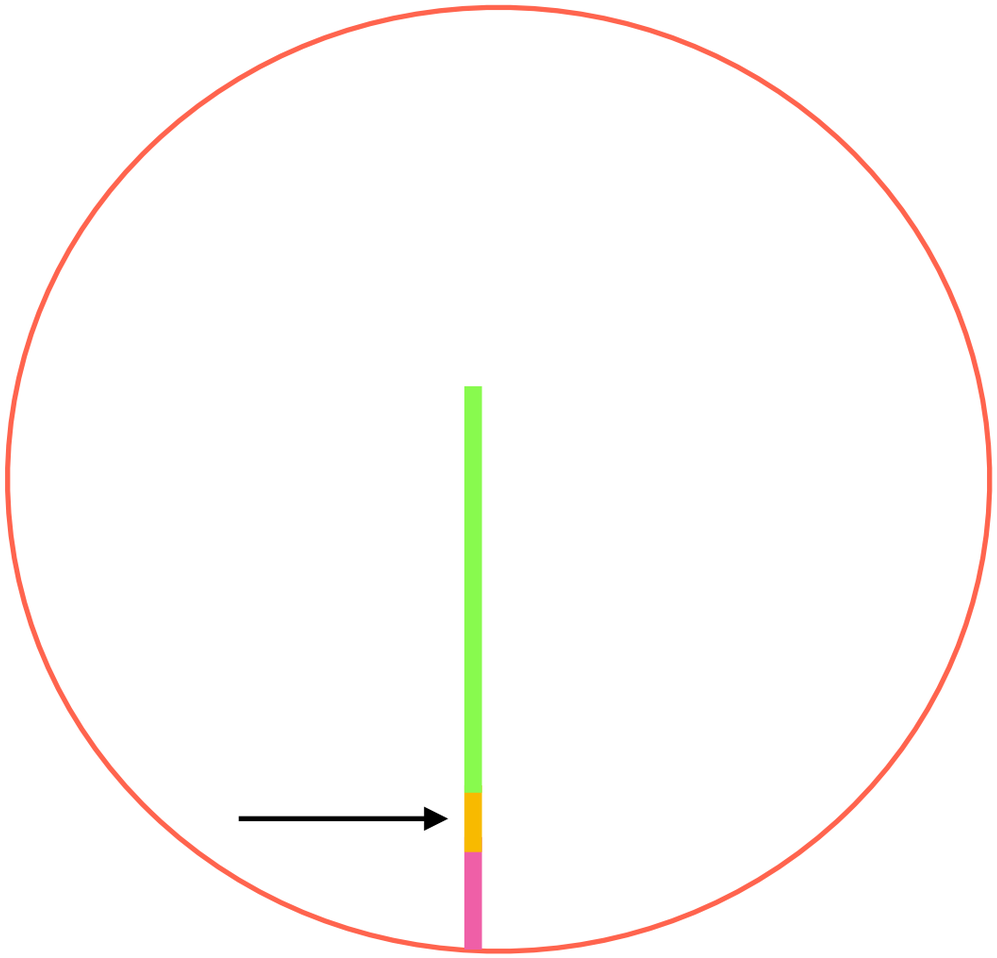




Парно-концевые риды

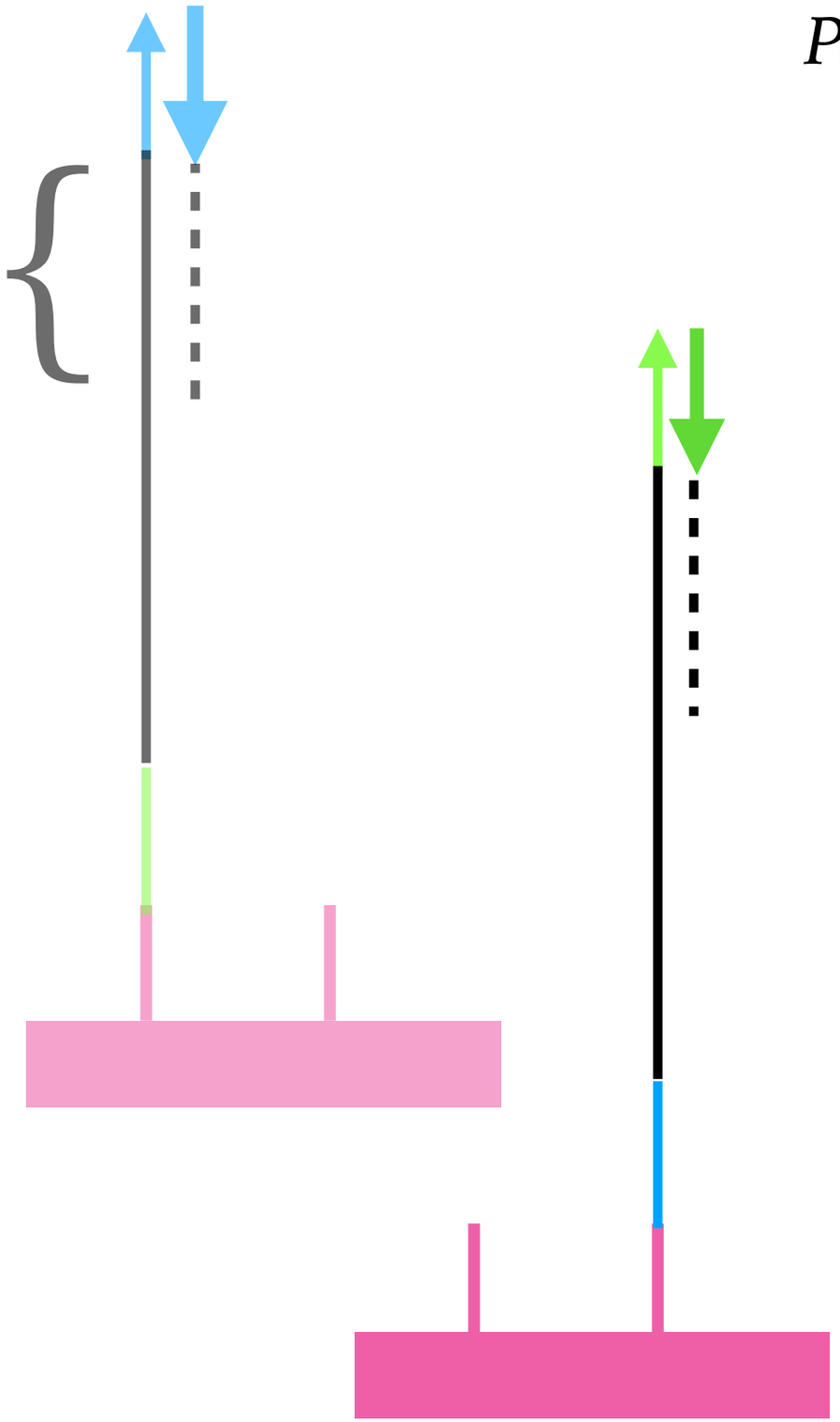
Paired-End sequencing

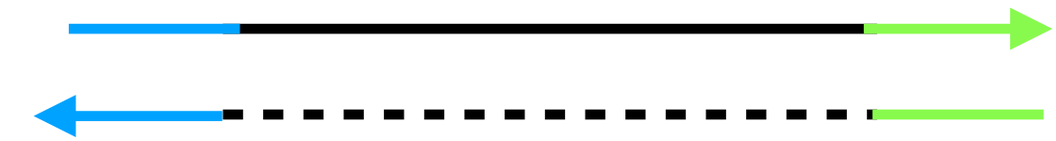




Парно-концевые риды

Paired-End sequencing





Техники, основанные на NGS

Что мы можем понять, используя данные NGS

RNA-seq

ChIP-seq

Hi-C

MNase-seq

CLIP-seq

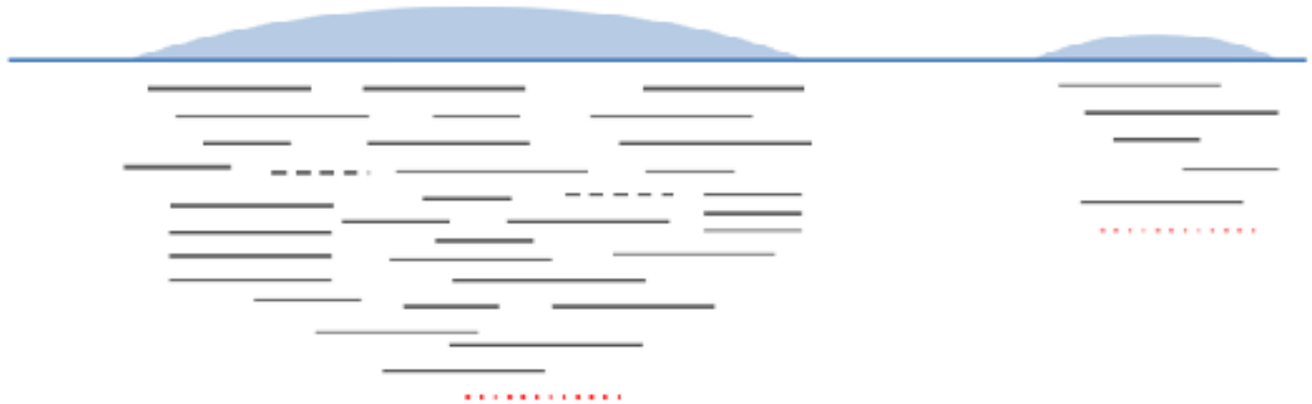
NET-seq

Ribo-seq

CAGE-seq

ATAC-seq

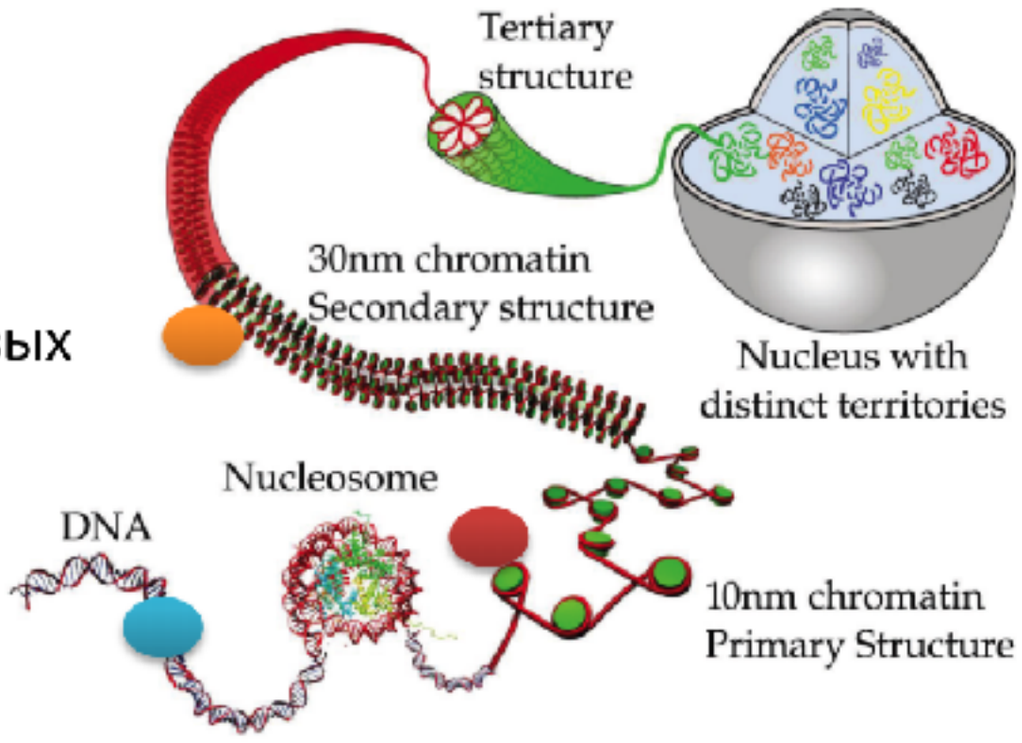
Exome-seq



DNA-DNA / DNA-Protein Interactions

HiC / ChipSeq

Изучение ДНК-белковых взаимодействий



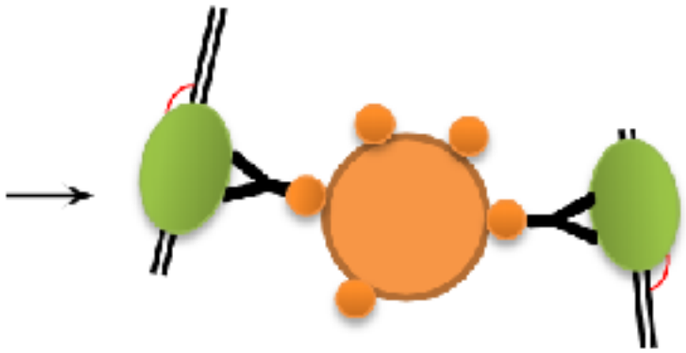
Изучение пространственной организации

ChipSeq

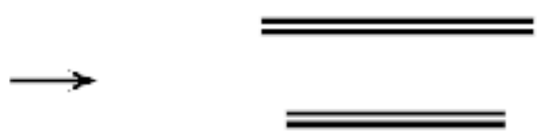
"Wet-lab"

Сшивка формальдегидом

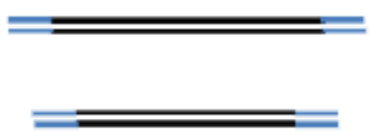
Фрагментация ДНК



Иммунопреципитация сшитых ДНК-белковых комплексов



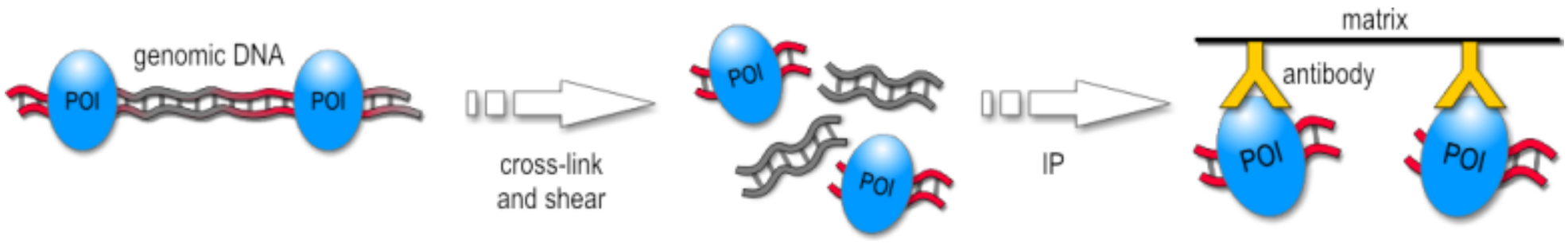
Освобождение ДНК (прогреванием)



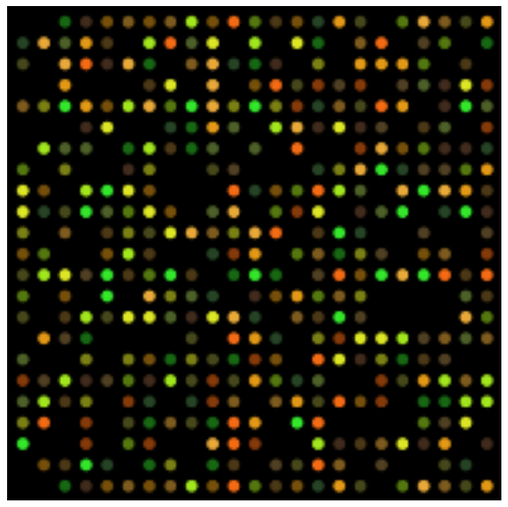
Подготовка библиотеки и секвенирование

ChipSeq

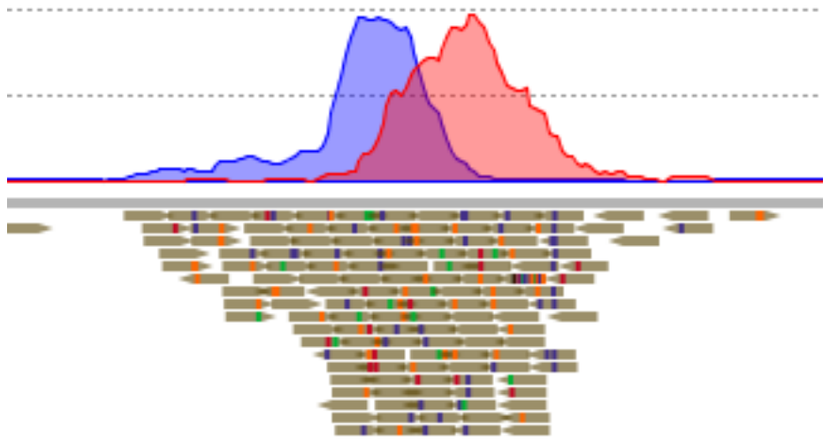
Mapping of DNA-protein contacts



Chip-on-Chip

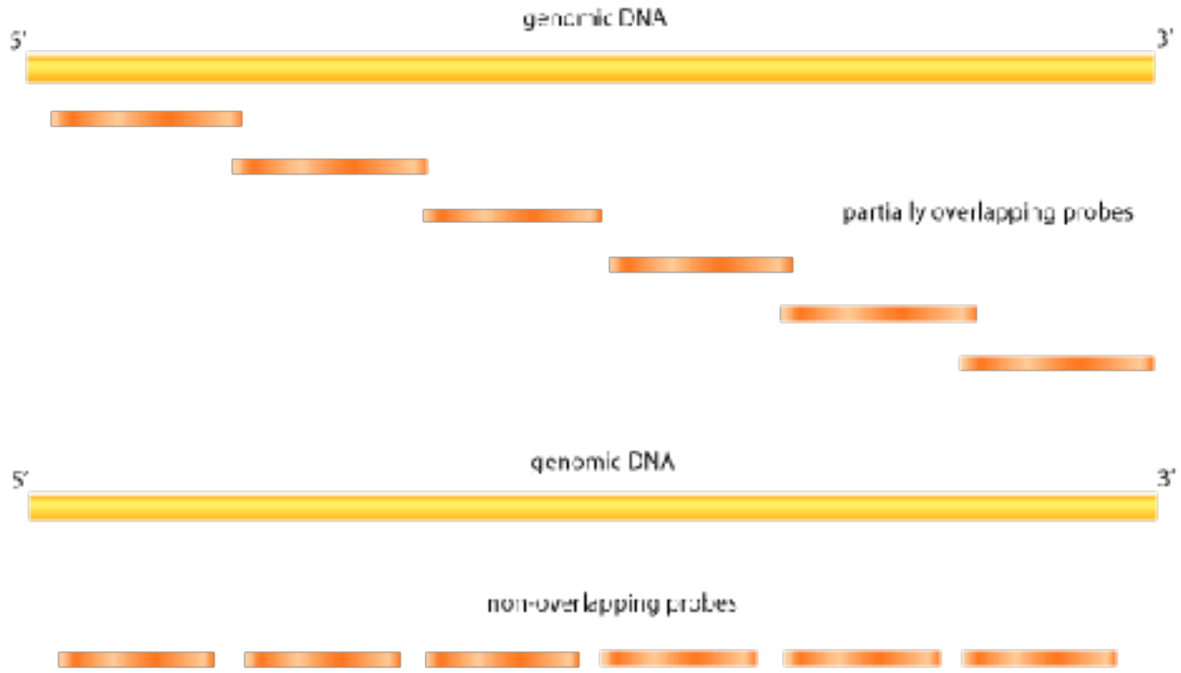


Chip-Seq



Chip-on-Chip

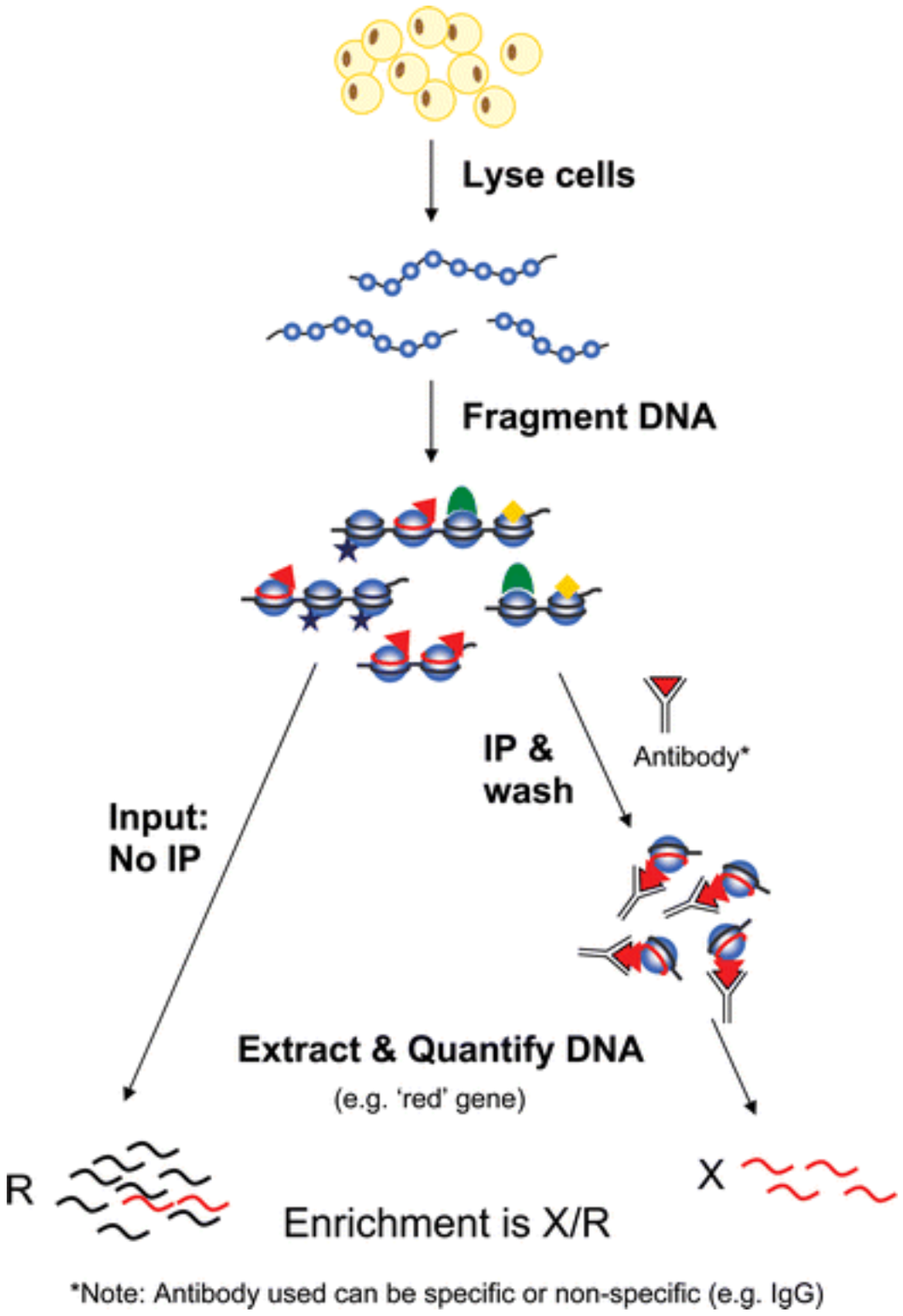
Enrichment of target fragments detection



Affymetrix, NimbleGene, Agilent

ChipSeq

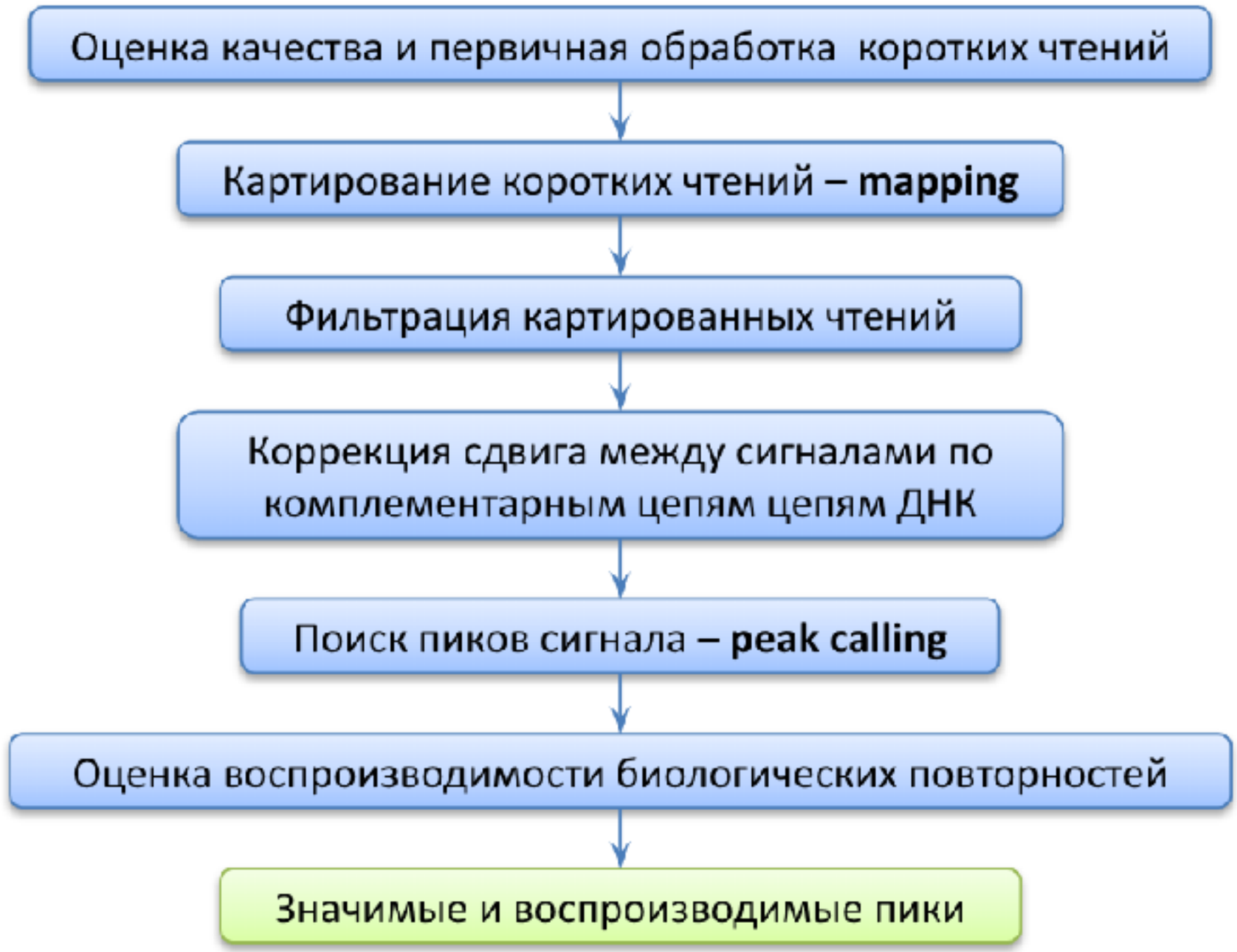
Enrichment of target fragments detection



*Note: Antibody used can be specific or non-specific (e.g. IgG)

ChipSeq

Pipeline



ChipSeq

Processing

Remove multimappers

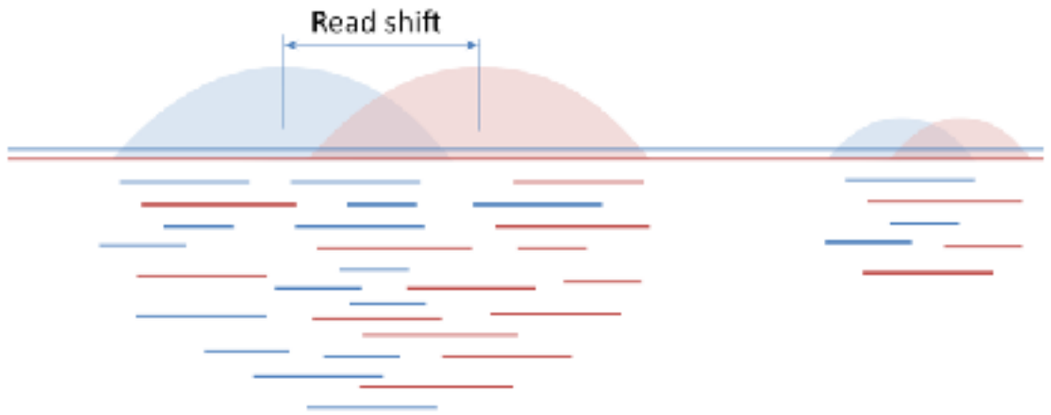
Remove duplicates

Remove low-quality reads

ChipSeq

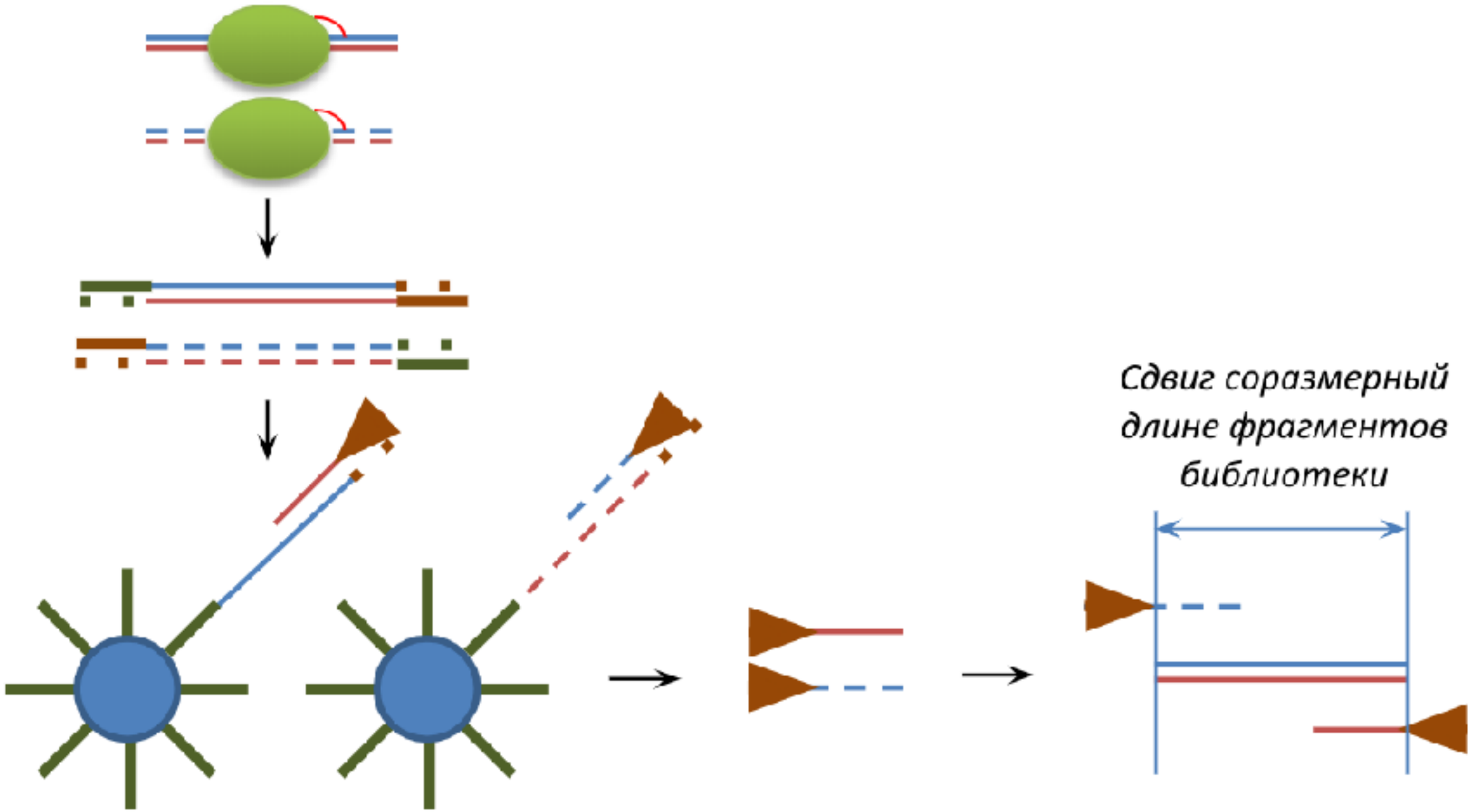
Tag shift correction

Detect peak



ChipSeq

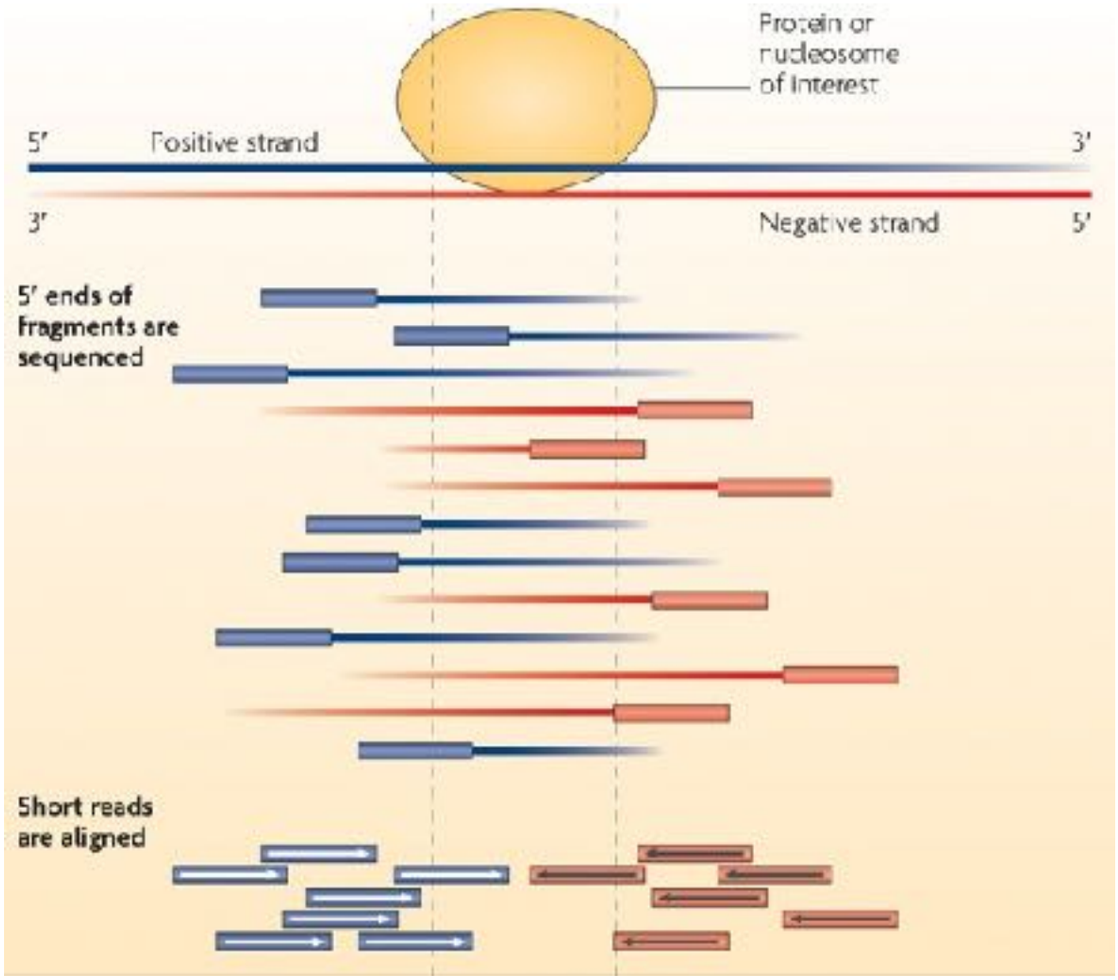
Why tags are shifted?

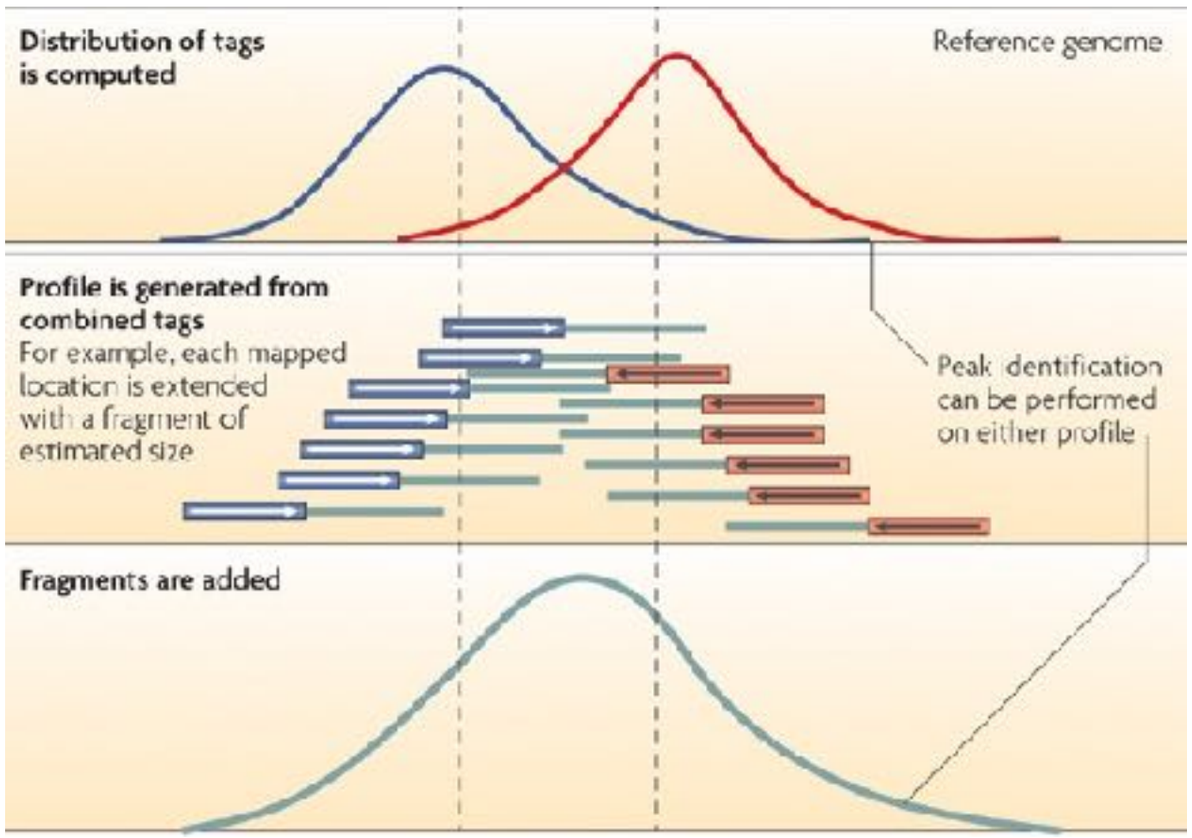


Сдвиг соразмерный длине фрагментов библиотеки

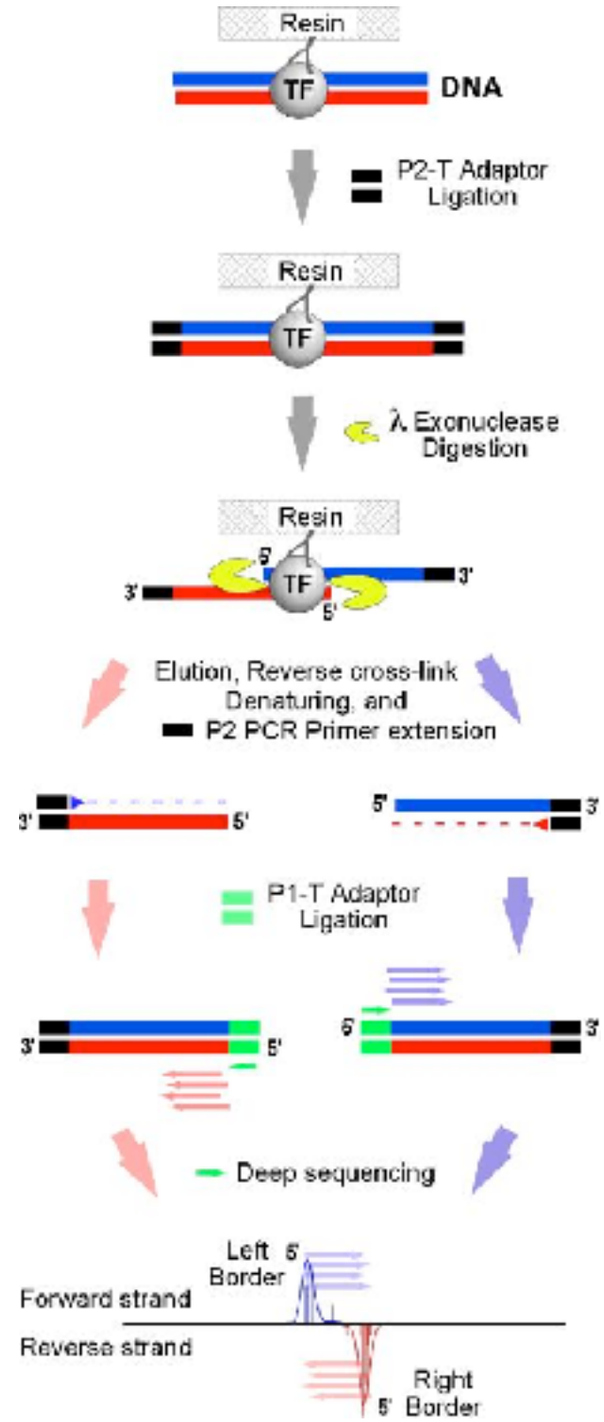
ChipSeq

Tag shift corrected



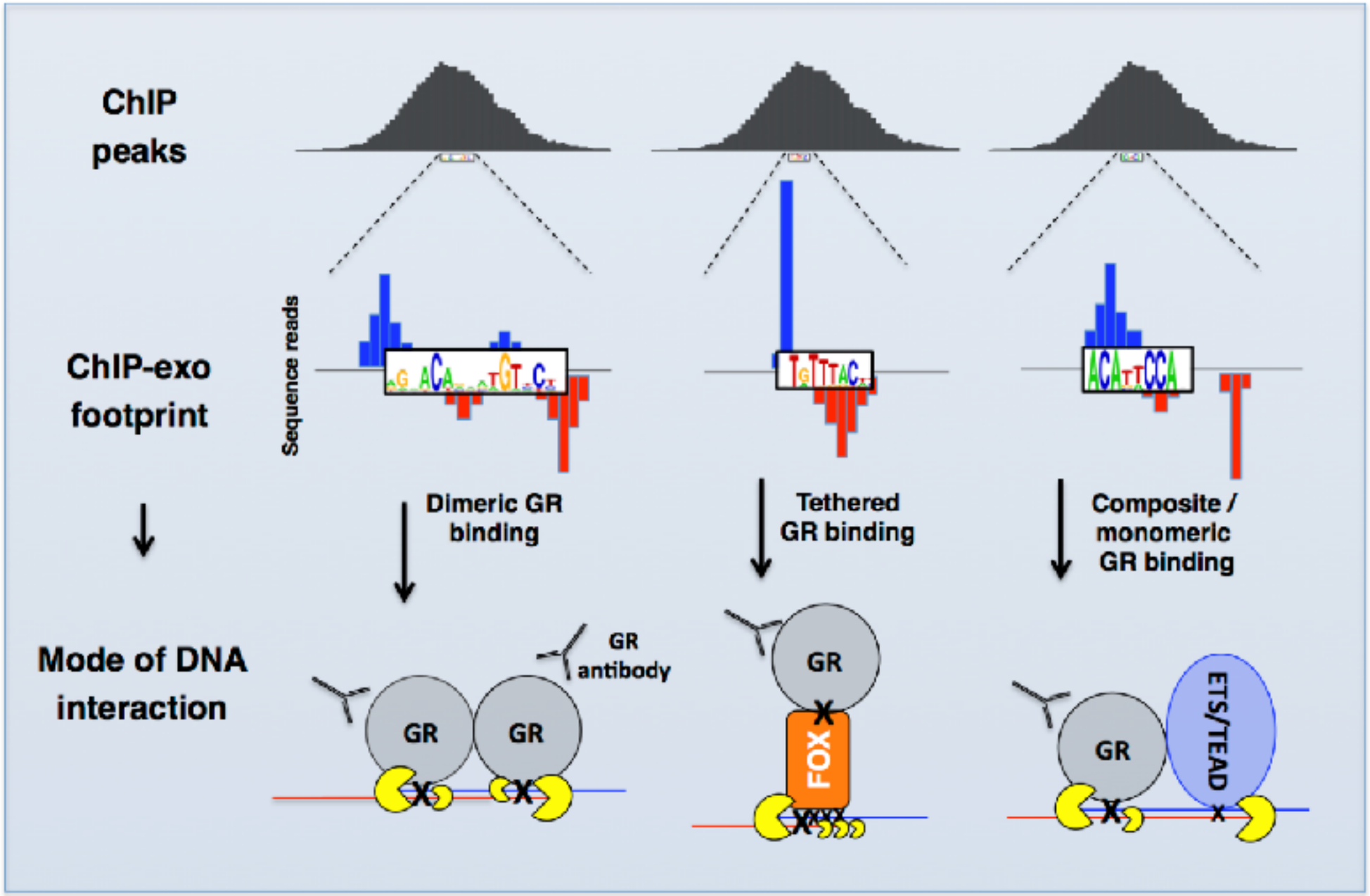


ChipSeq-exo


Resin
TF
DNA
P2-T Adaptor Ligation
Resin
TF
λ Exonuclease Digestion
Resin
5'
TF
3'
3'
5'
Elution, Reverse cross-link Denaturing, and
P2 PCR Primer extension
3'
5'
5'
3'
P1-T Adaptor Ligation
3'
5'
5'
3'
Deep sequencing
Left Border
5'
Forward strand
Reverse strand
Right Border
5'

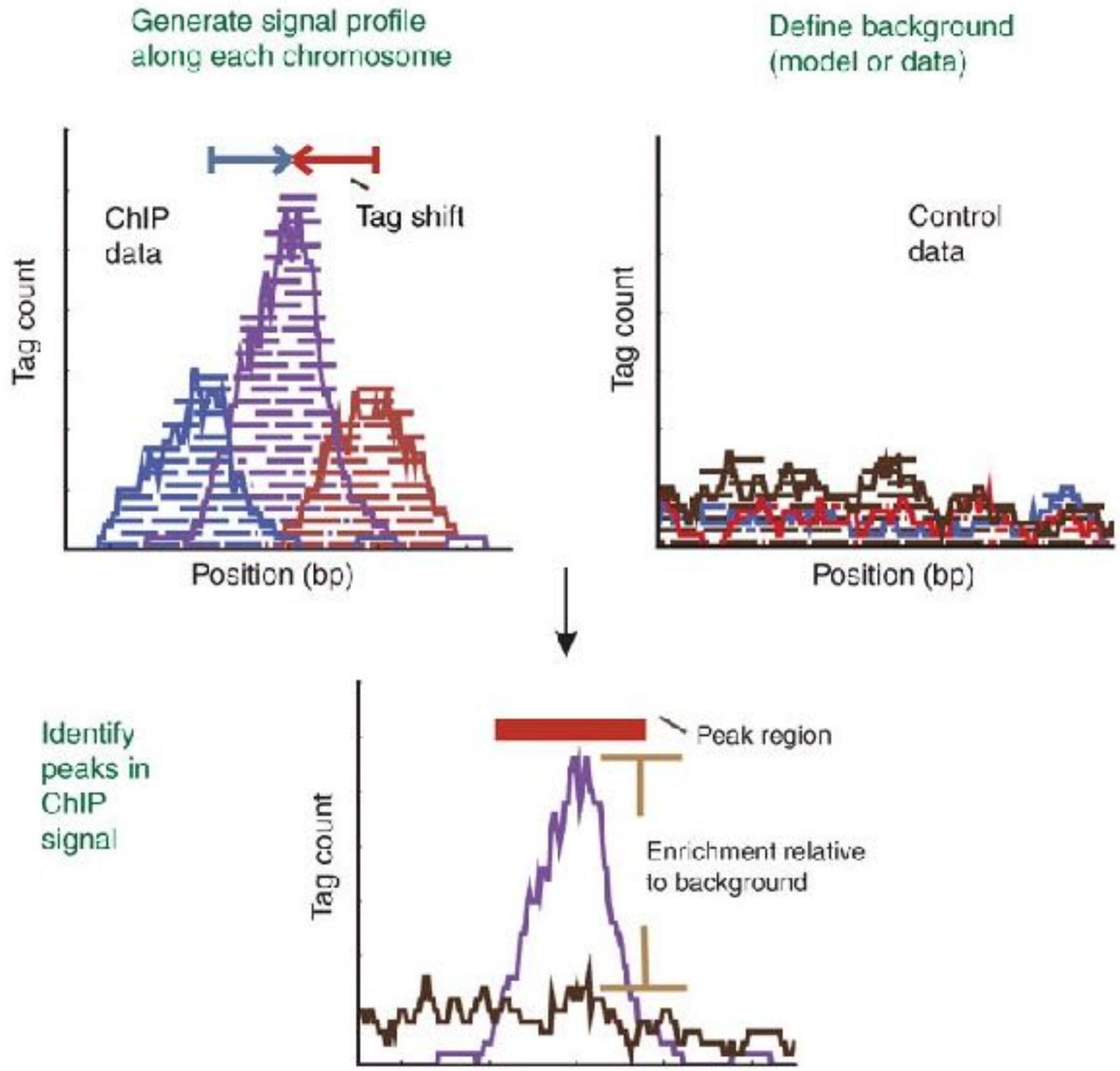
ChipExo

Определение типа взаимодействия



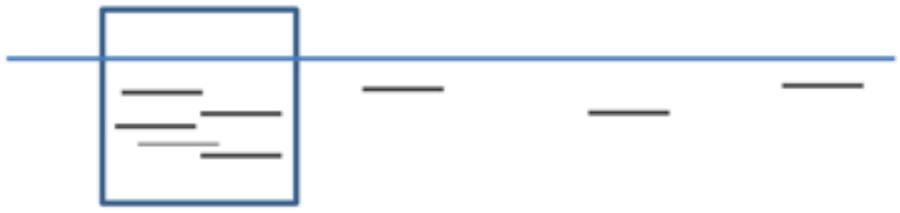
ChipSeq

Pipeline



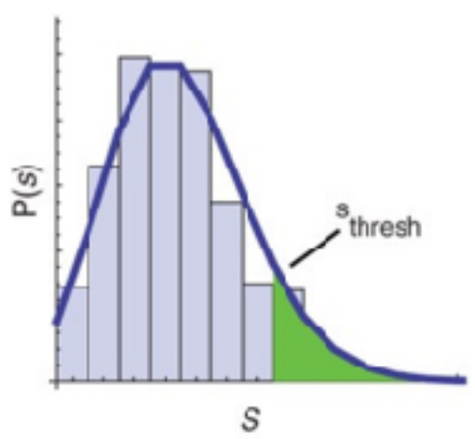
ChipSeq

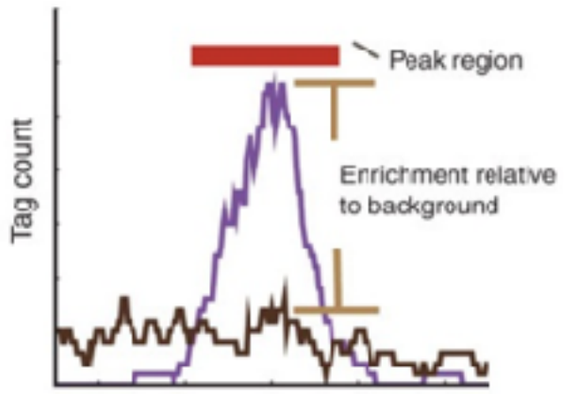
Enrichment detection



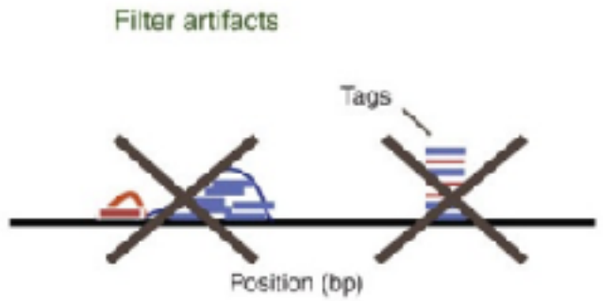
Какова вероятность обнаружить такое количество чтений в окне шириной *d* нт?

p-value, q-value, FDR (false discovery rate)





Pepke et al. (2009) Nature Methods. 6



2 основных типа артефактных сигналов:

- пики со значительным различием в покрытии по комплементарным цепям
- пики с одним чтением или очень небольшим количеством чтений

Single-cell: да или нет?

Подход позволяет получить информацию о:

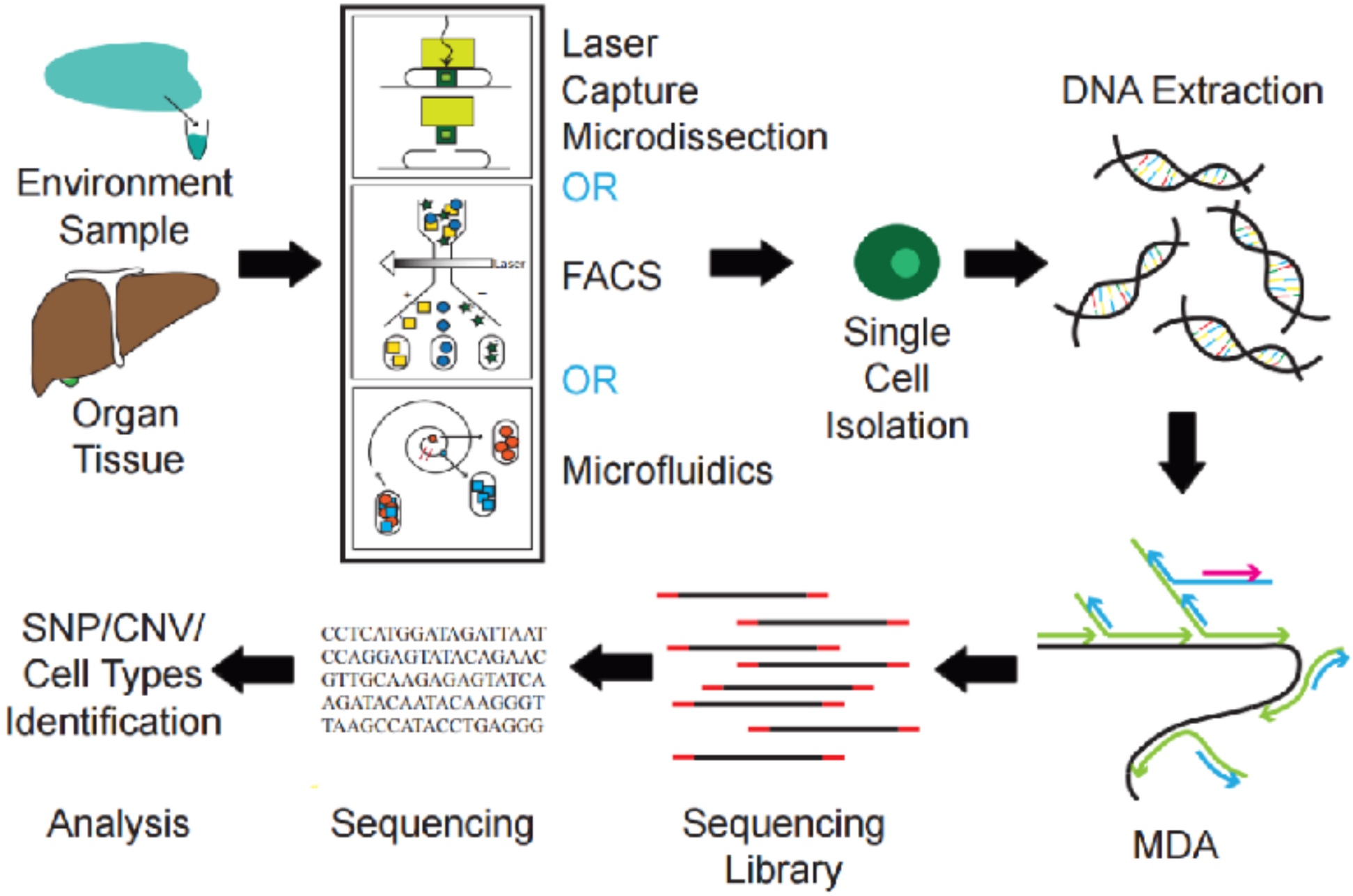
- Некультивируемых микроорганизмах (часто единственный возможный путь изучения)
- Редких типах клеток
- Гетерогенных образцах
- Полиморфизме соматических тканей
- Изменениях в клеточных линиях
- Развитии заболеваний

Отрицательные стороны подхода – более насущными становятся проблемы:

✦Деградации и потери материала

✦Контаминации

Single-cell protocol



Single-cell protocol

Multiple Displacement Amplification

Цель – увеличить количество ДНК, получаемое от одной клетки, до необходимого для секвенирования: от фемтограмм (10^{-12}) к микрограммам (10^{-6})

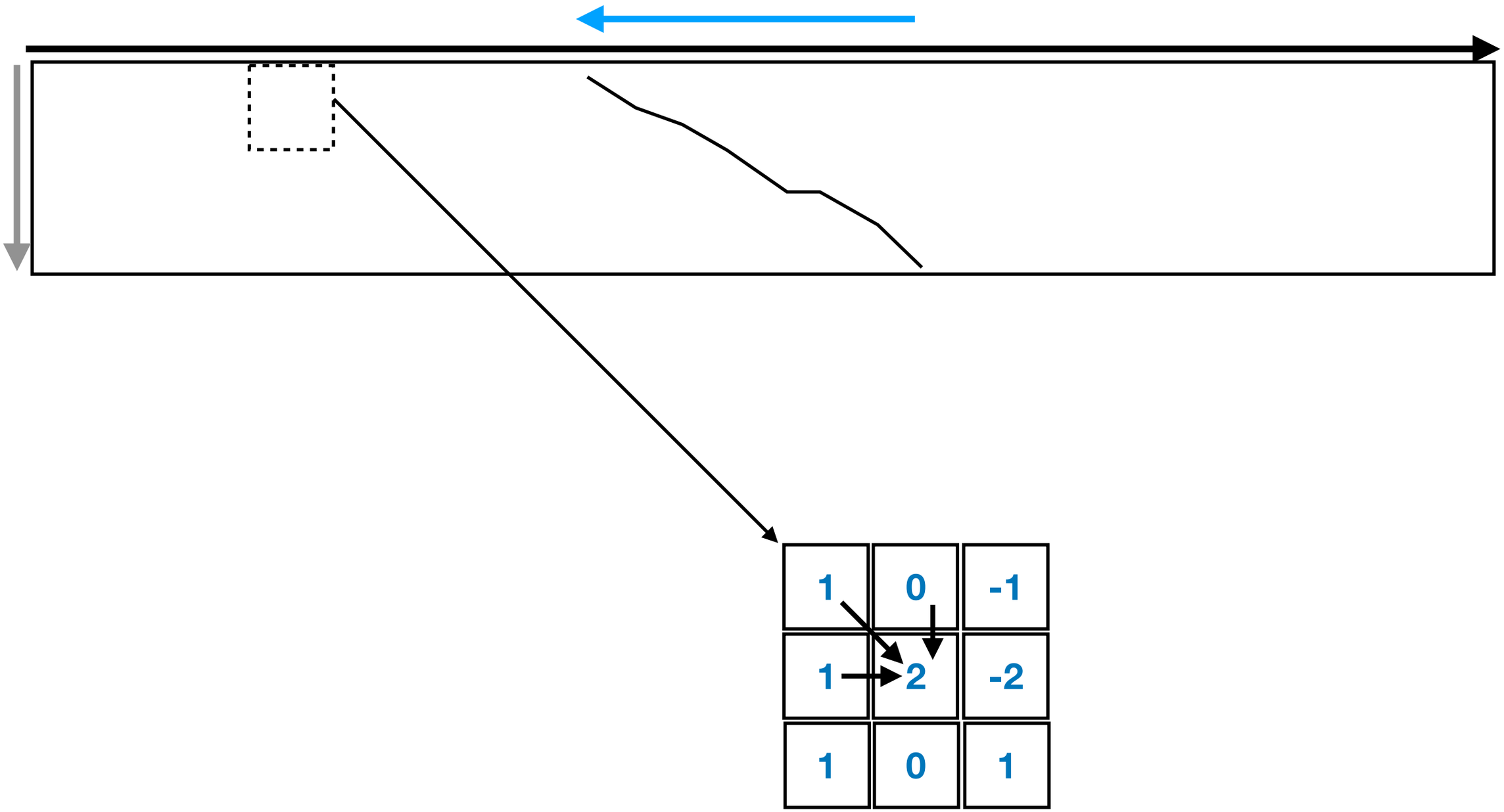
В реакции используется ДНК-полимераза бактериофага phi29 и случайные праймеры (random primers). В ходе изотермической реакции при 30°C происходит синтез многих копий исходной ДНК с **вытеснением** цепей. Средняя длина продукта – 12 т.п.н. (до 100 т.п.н.)

Слабая сторона: неравномерная амплификация – некоторые участки могут быть перепредставлены, некоторые утеряны

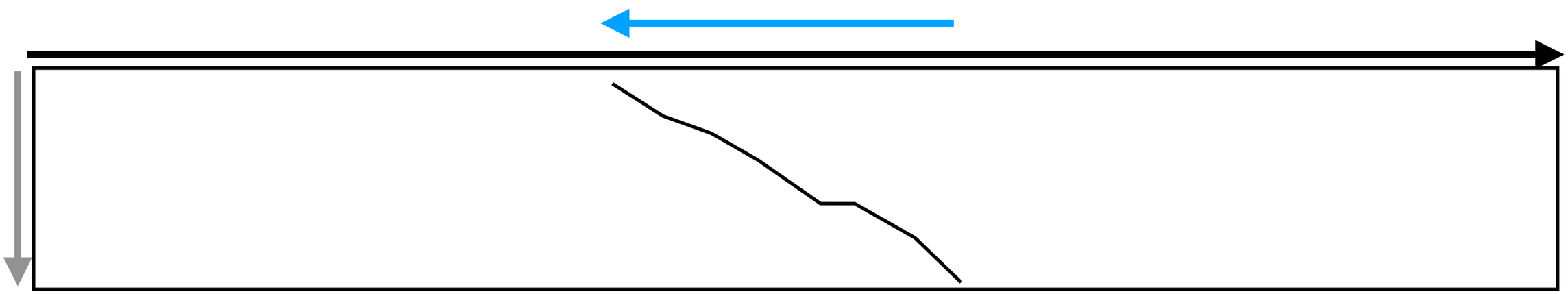
Картирование

Как быстро и качественно получить сигнал?

Картирование



Картирование



Займет $\sim O(kN \times M)$

К тому же, с довольно плохой константой k(=6)

Картирование

seed-extend парадигма

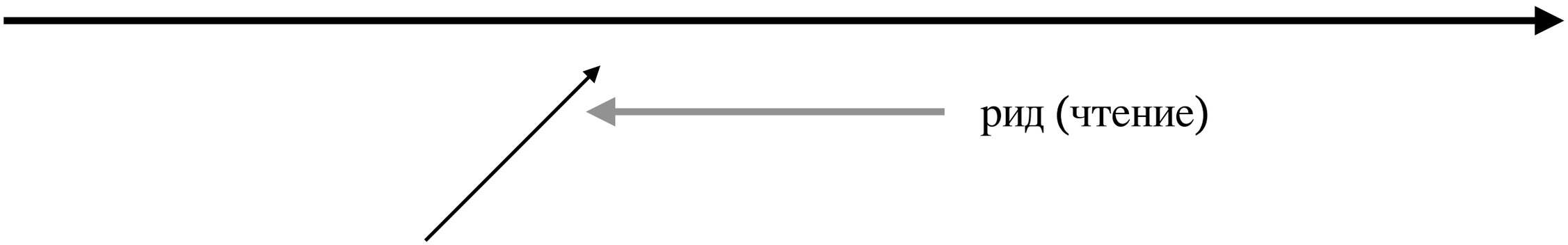
1. локализовать места возможного выравнивания рида с геномом быстрым методом

2. сделать честное выравнивание только в ограниченной области генома

Термины

связанные с картированием чтений

референс (геном, сборка, скаффолд, контиг)



выравнивание / картирование

Основные понятия:

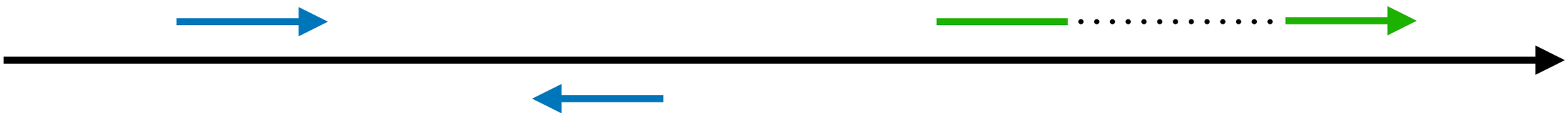
multimapper / unique mapper / unmapped

парные чтения : concordant mapping

exact match / split-alignment / chimeric alignment

Термины

связанные с картированием чтений



Основные понятия:

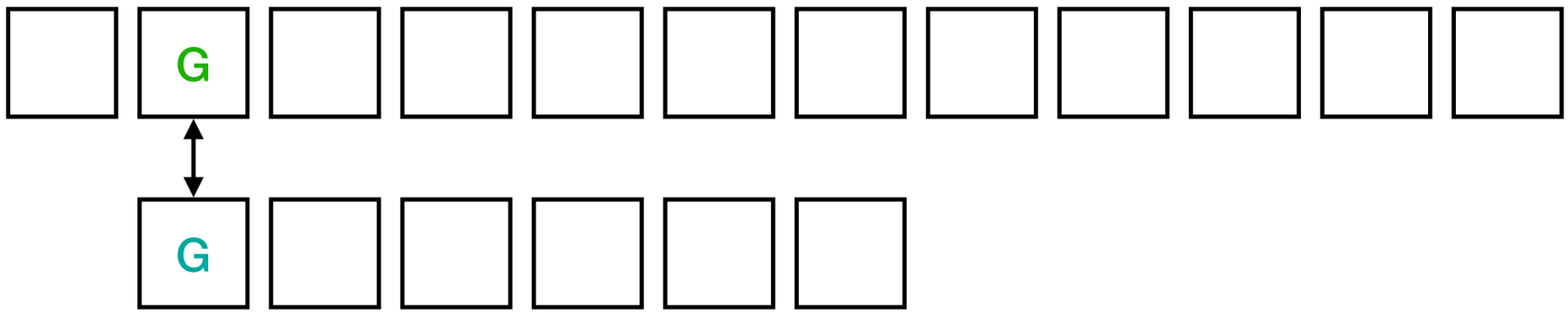
multimapper / unique mapper / unmapped

парные чтения : concordant mapping

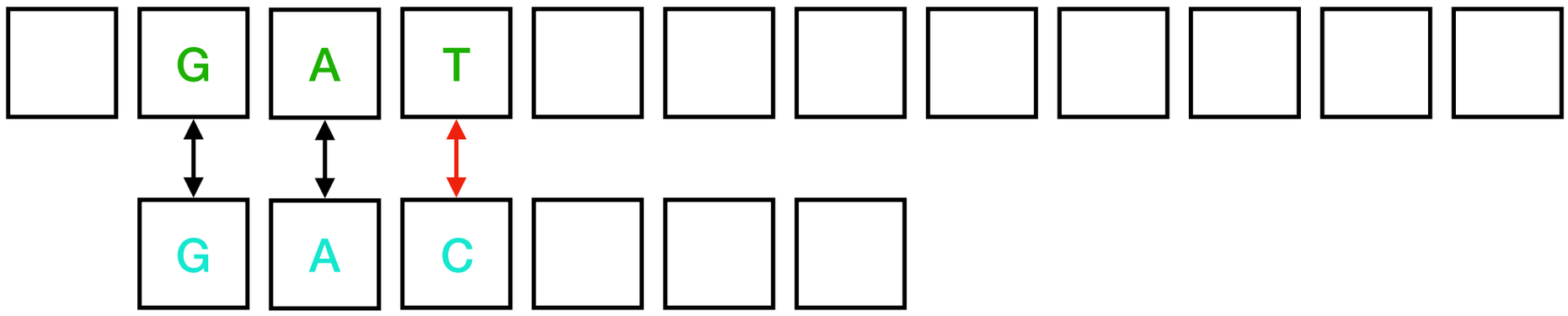
exact match / split-alignment / chimeric alignment

Картирование - наивный метод

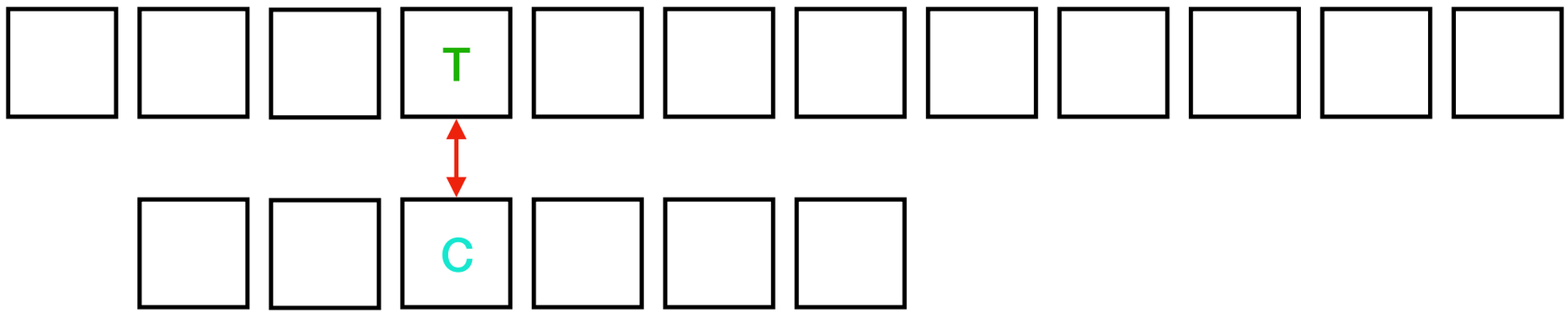
Попробуем exact pattern match?



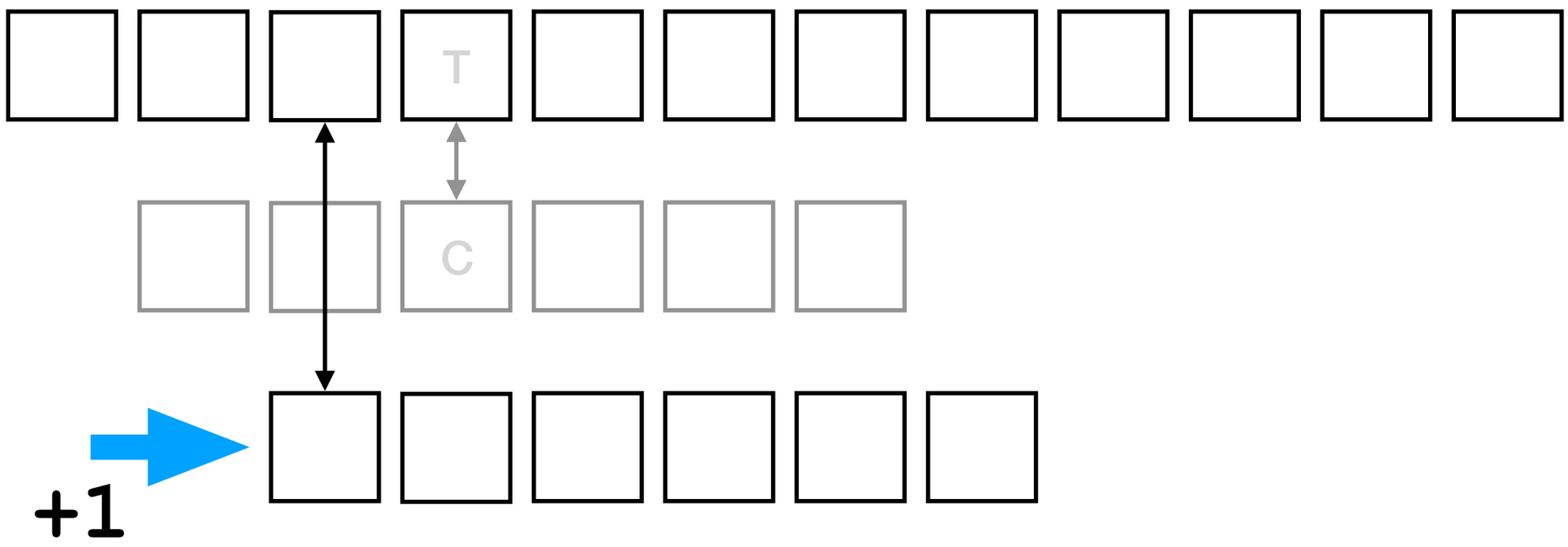
Картирование - наивный метод



Картирование - наивный метод



Картирование - наивный метод

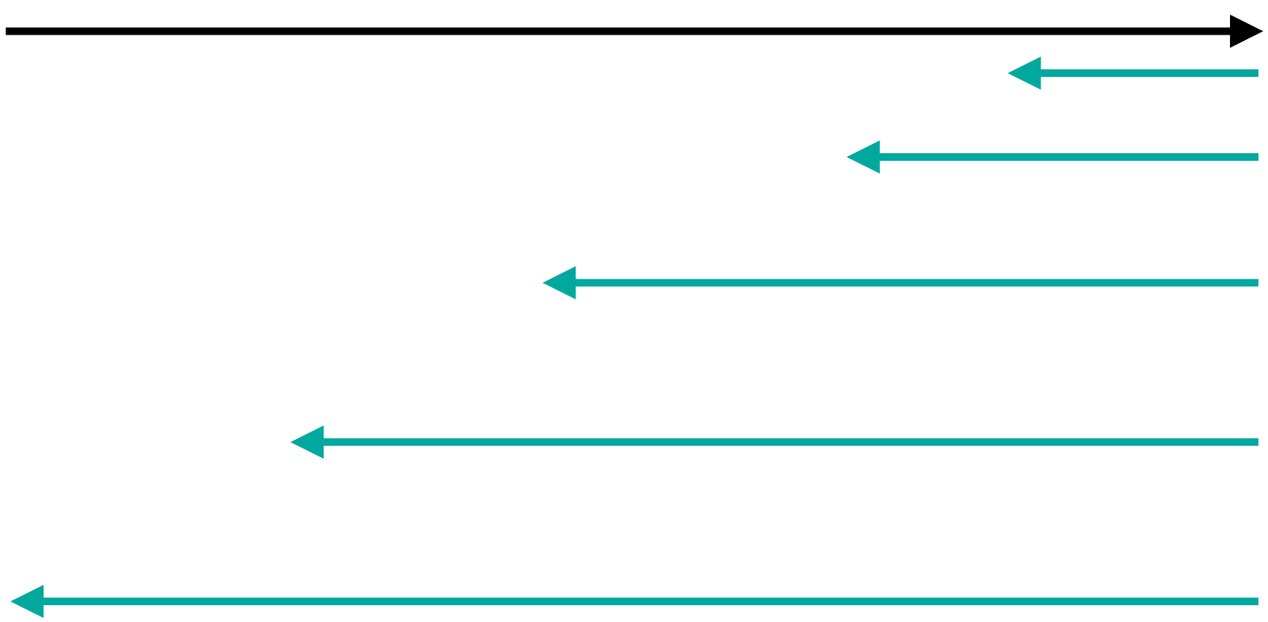


"Наивный" алгоритм опять требует $O(N \times M)$ времени

(но уже с лучшей константой)

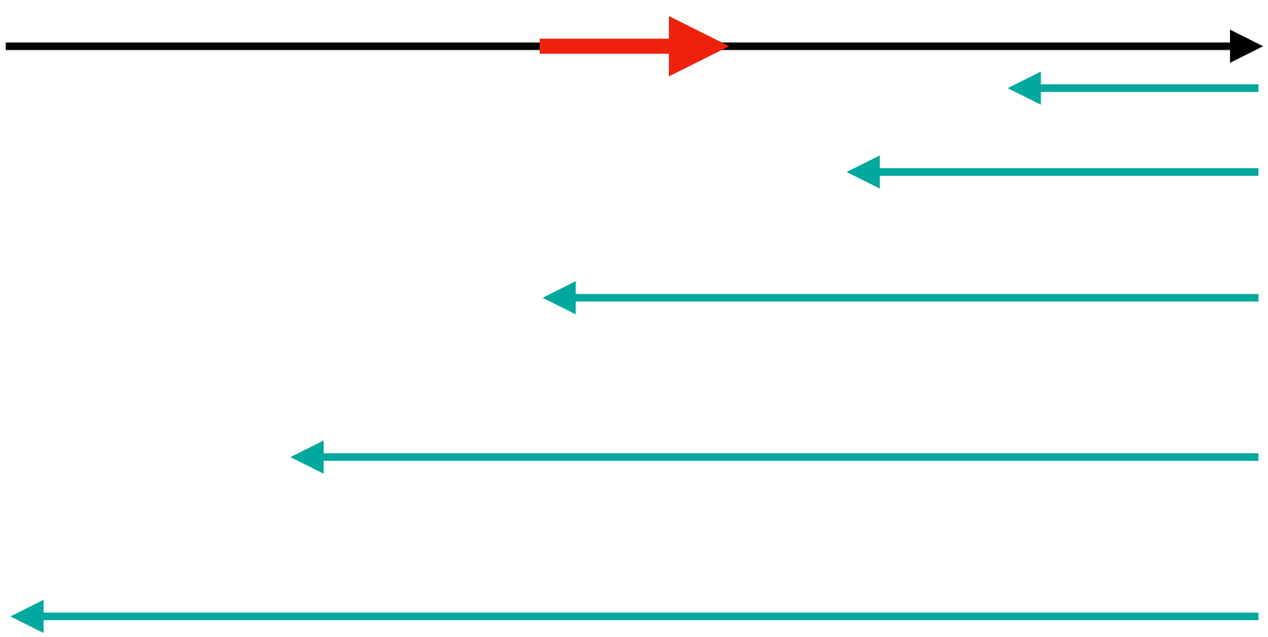
Mapping

Не решить задачу в 6 раз быстрее -
всё равно не решить задачу



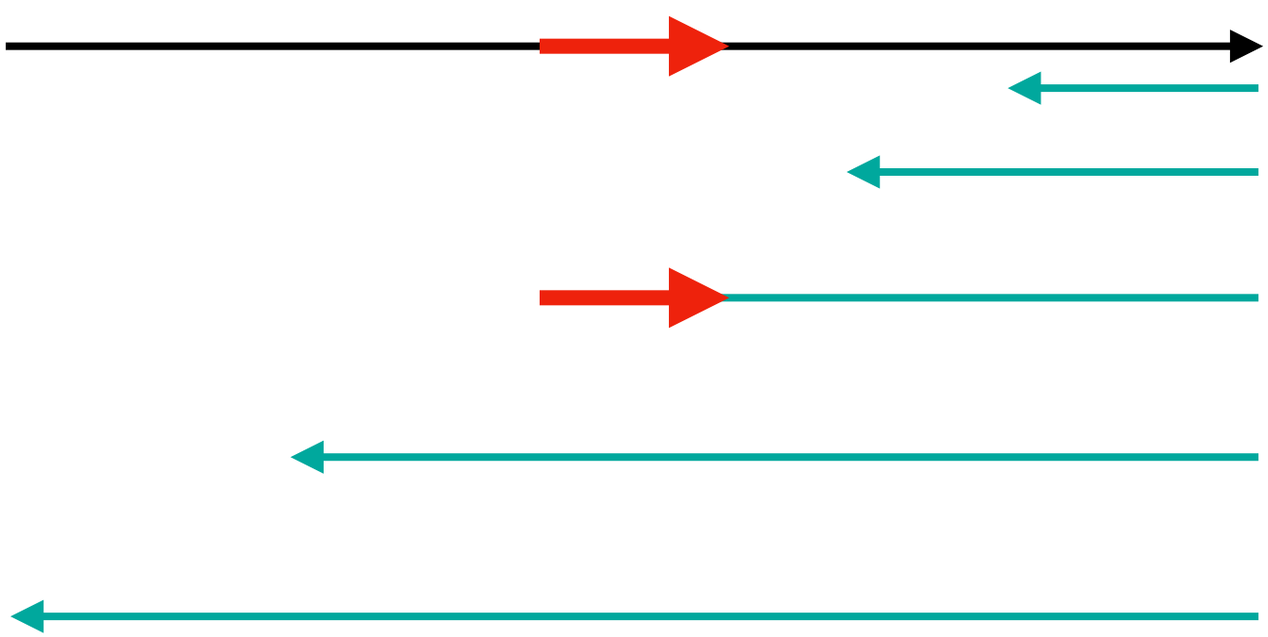
Mapping

Менее тривиальное решение



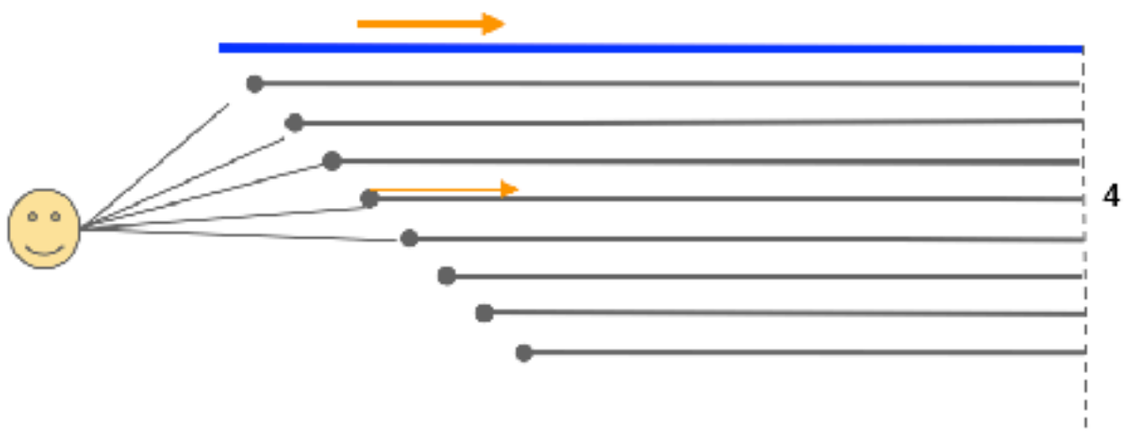
Mapping

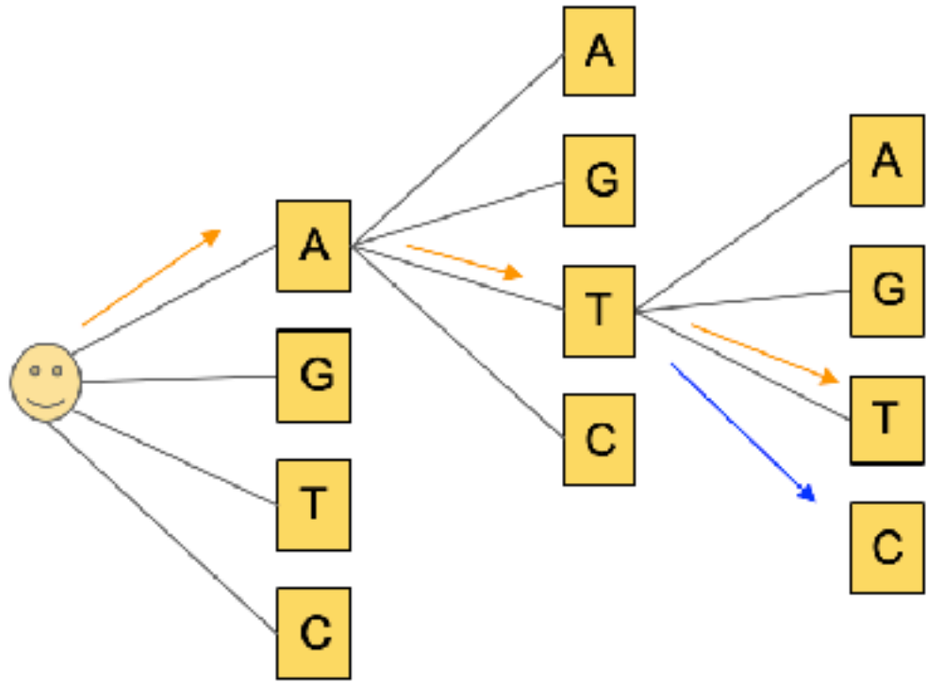
Подстрока исходной строки
есть префикс какого-то
суффикса исходной строки



Mapping

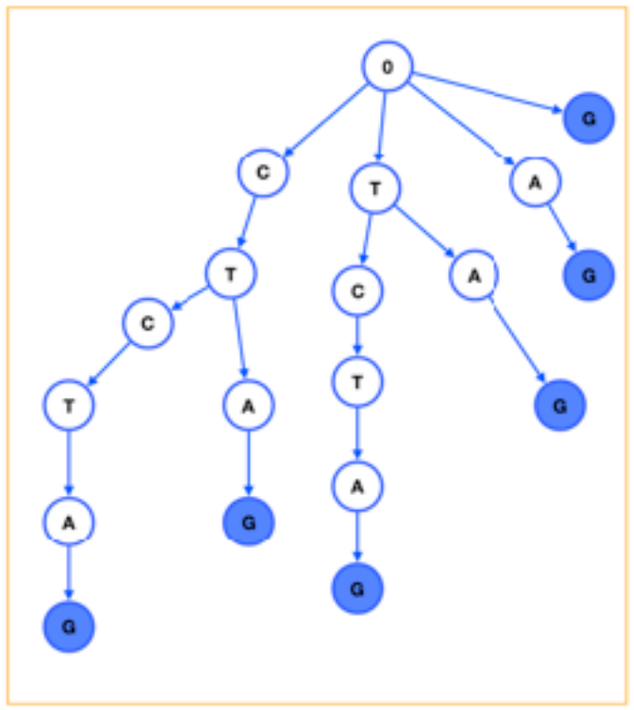
Суффиксное дерево



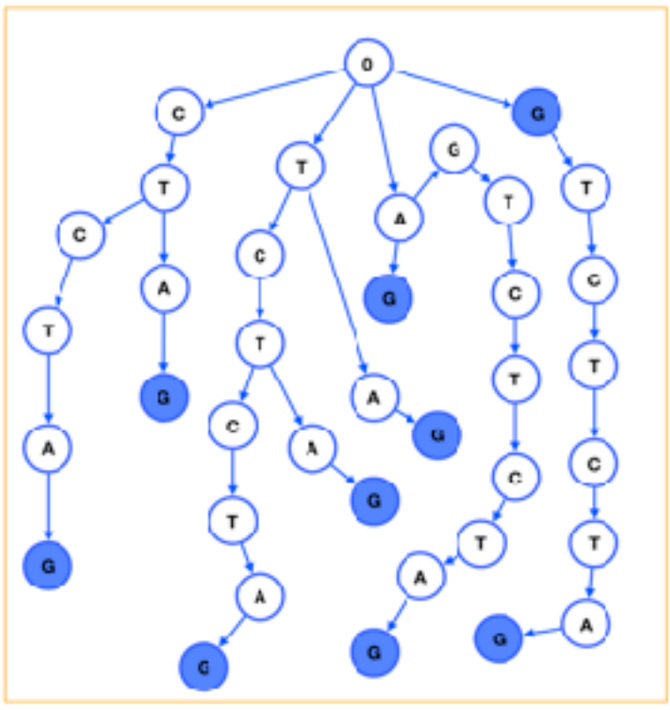


Суффикс строки

AGTCTCTAG	
G	9
AG	
TAG	
CTAG	
TCTAG	
CTCTAG	
TCTCTAG	
GTCTCTAG	
AGTCTCTAG	1



CTCTAG$



Mapping

Получается ли, что скорость работы
алгоритма перестала зависеть от
длины генома, в котором мы ищем?

Ура! Теперь мы можем найти рид
в геноме человека так же быстро,
как в геноме вируса!

BWT, FM-index

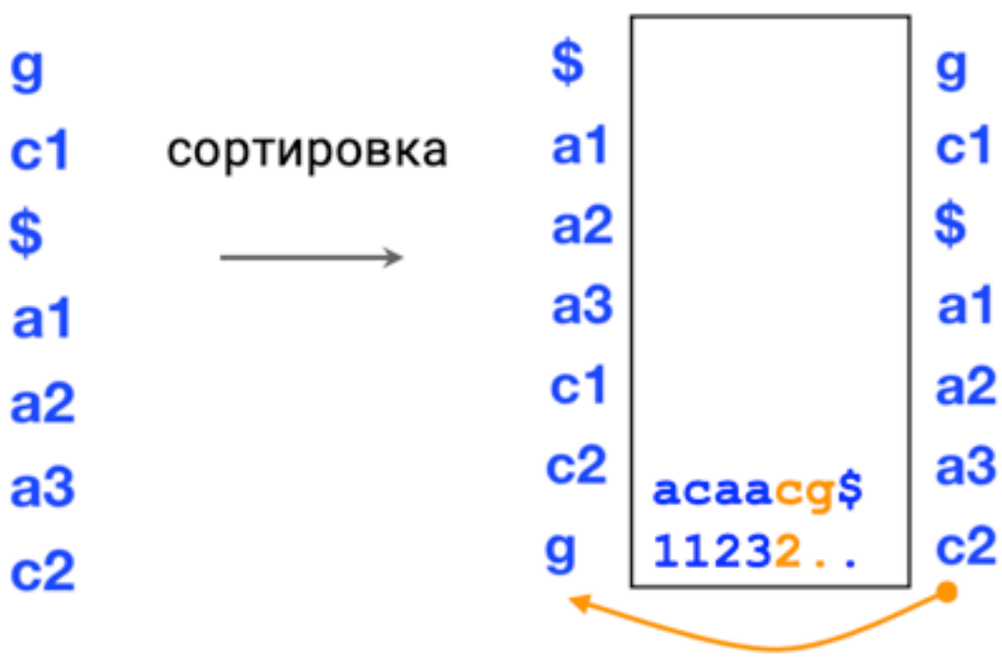
Преобразование Барроуза — Уилера

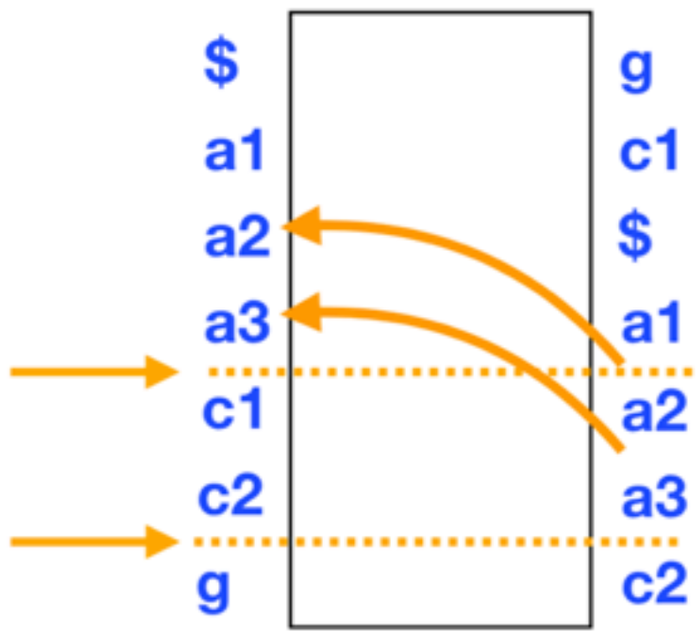
BWT, FM-index

1. Сжатие (без потери) acaacg$

   gc$3ac_

2. Быстрый поиск





собственно поиск

BWT, 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:

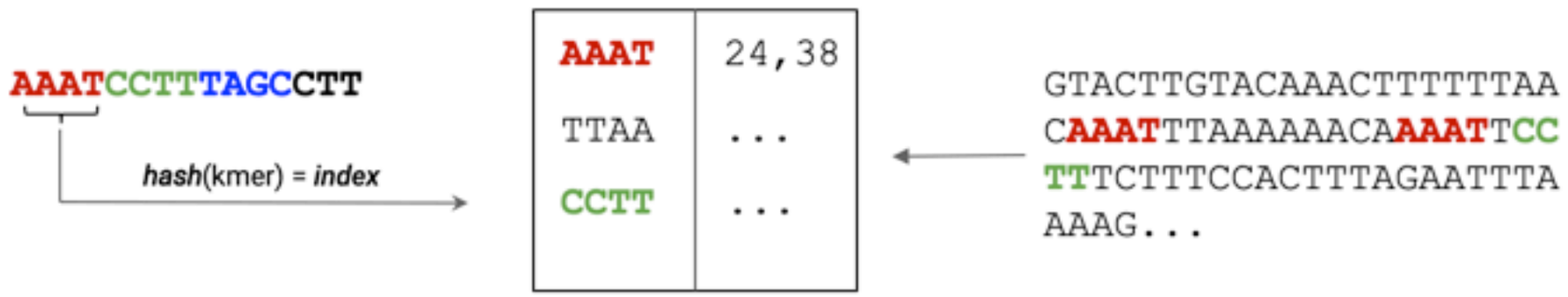
	1	2	3	4	5	6	7
$	0	0	1	1	1	1	1
a	0	0	0	1	2	3	3
c	0	1	1	1	1	1	2
g	1	1	1	1	1	1	1

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

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

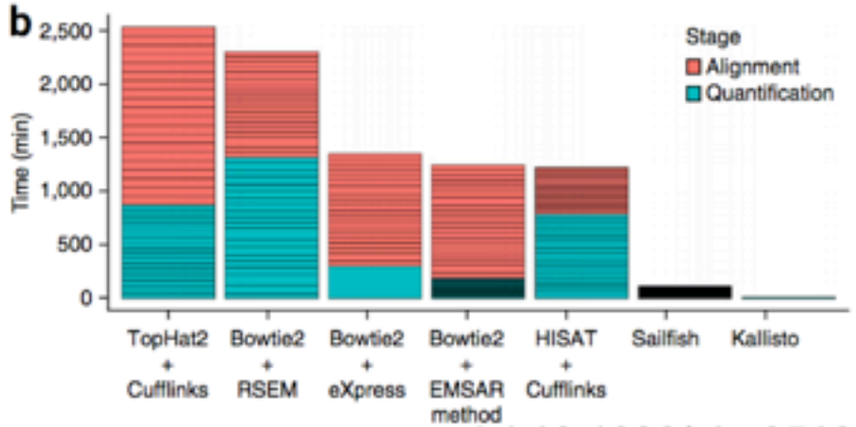
Mapping

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