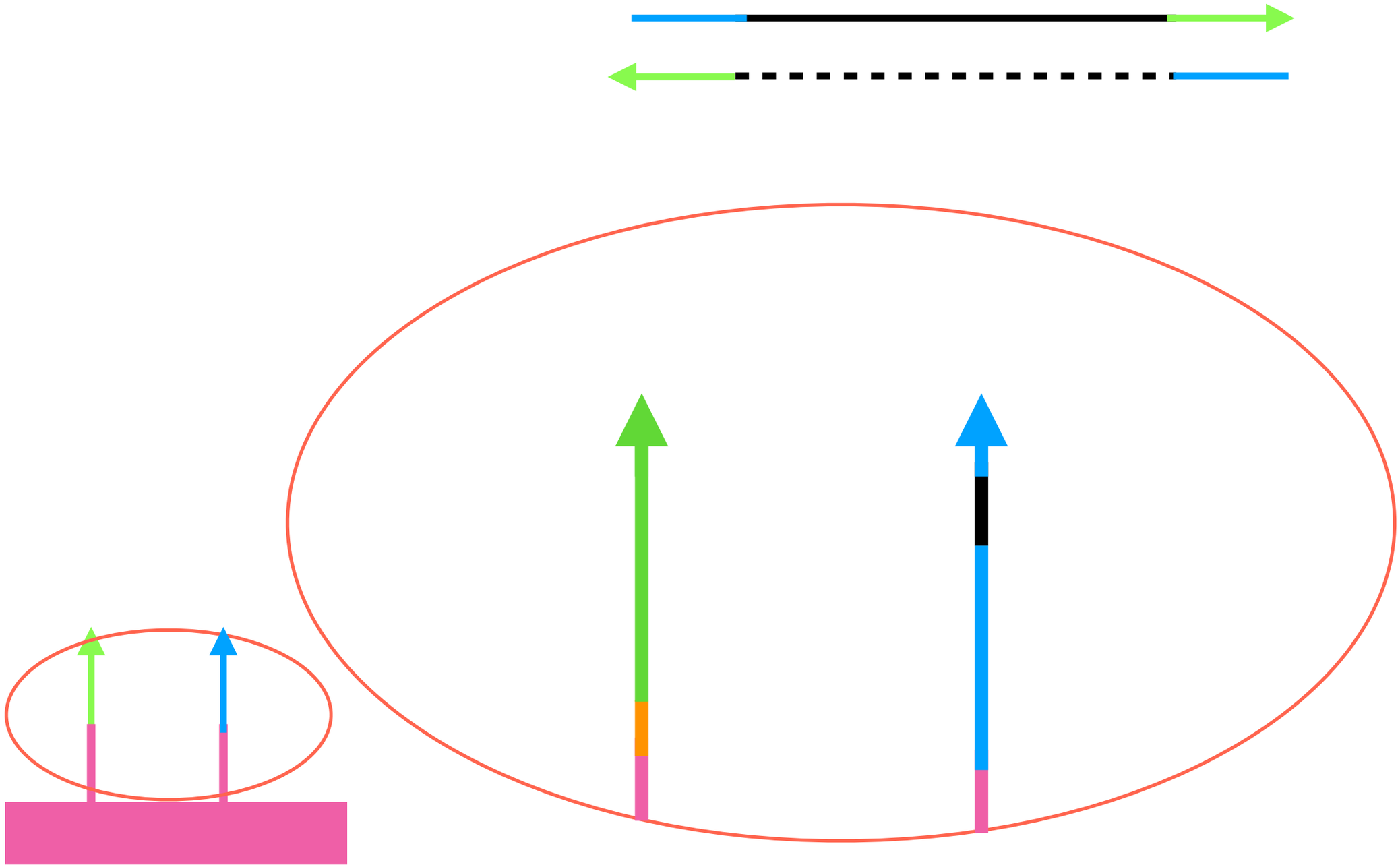
Биоинформатика

2018

Лекция 6
Герасимов Евгений Сергеевич
jalgard@yandex.ru

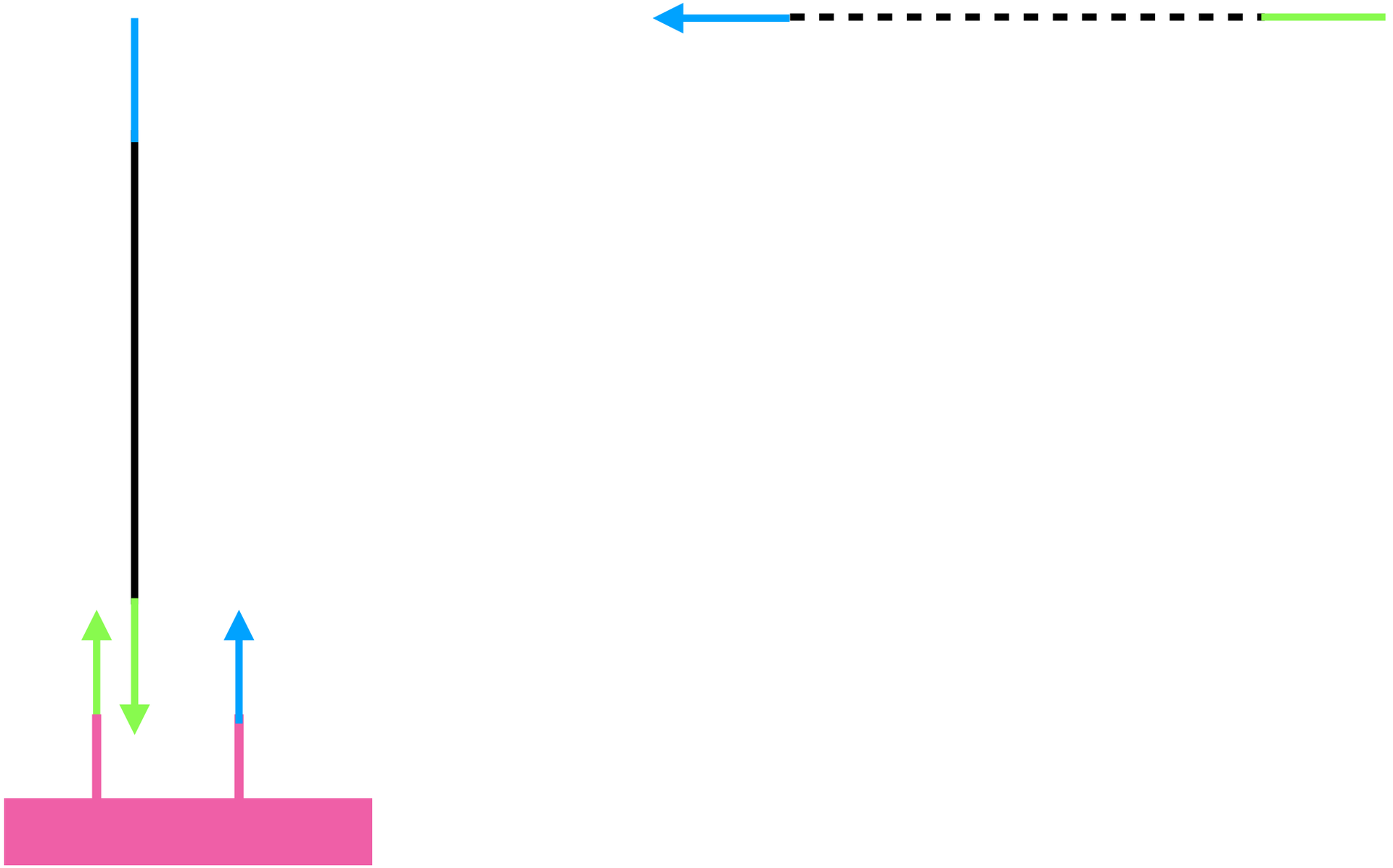
Парно-концевые риды

Paired-End sequencing



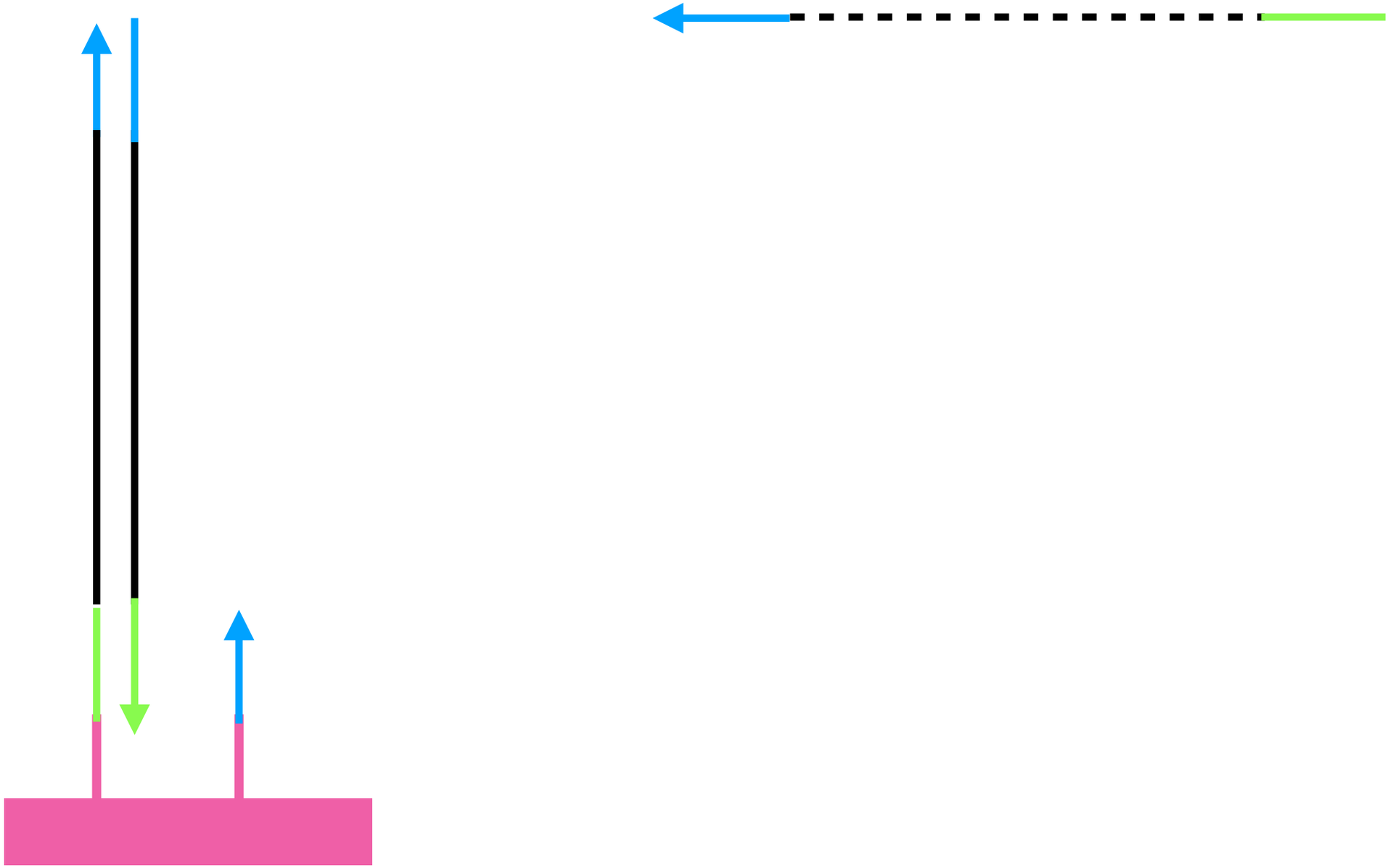
Парно-концевые риды

Paired-End sequencing



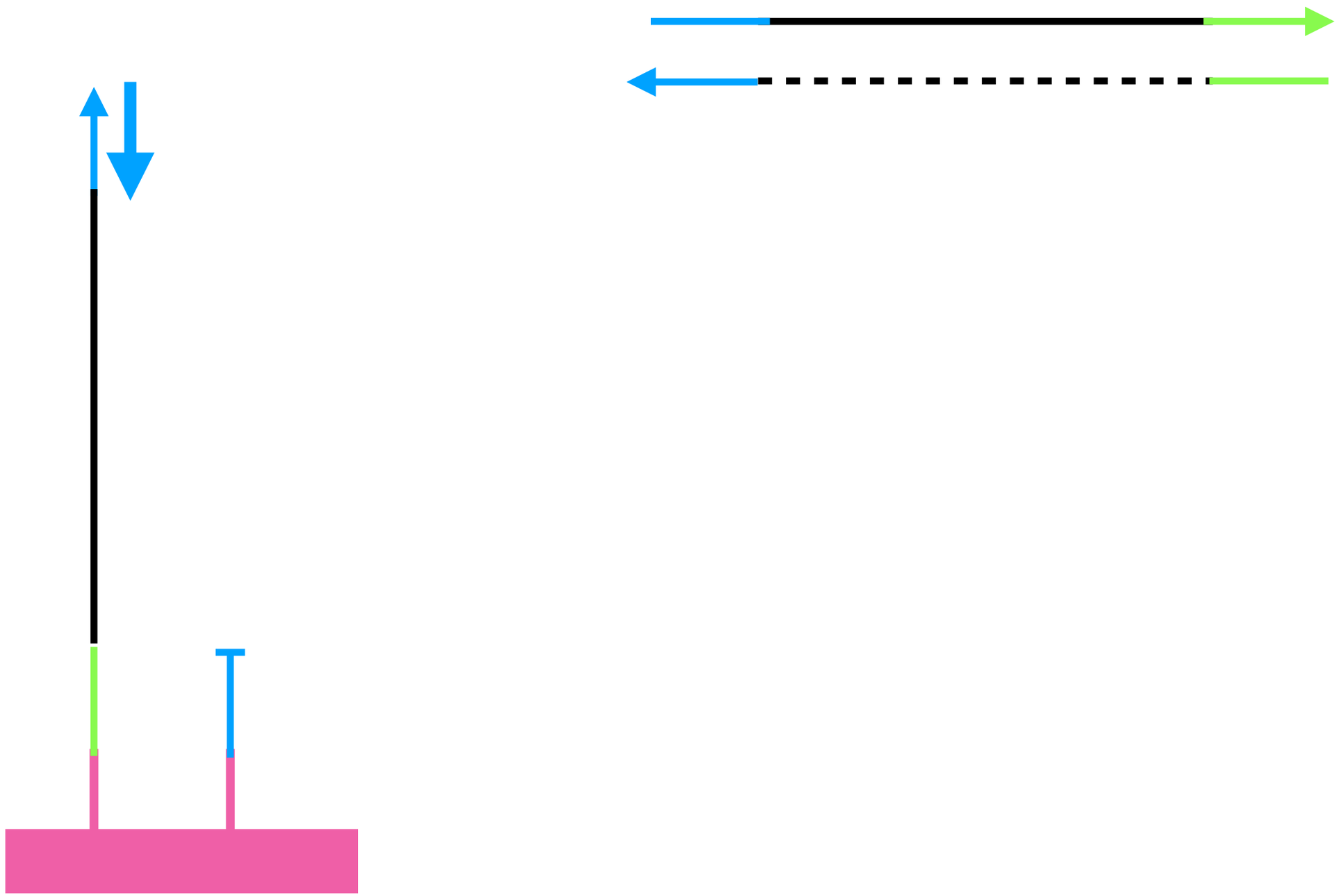
Парно-концевые риды

Paired-End sequencing



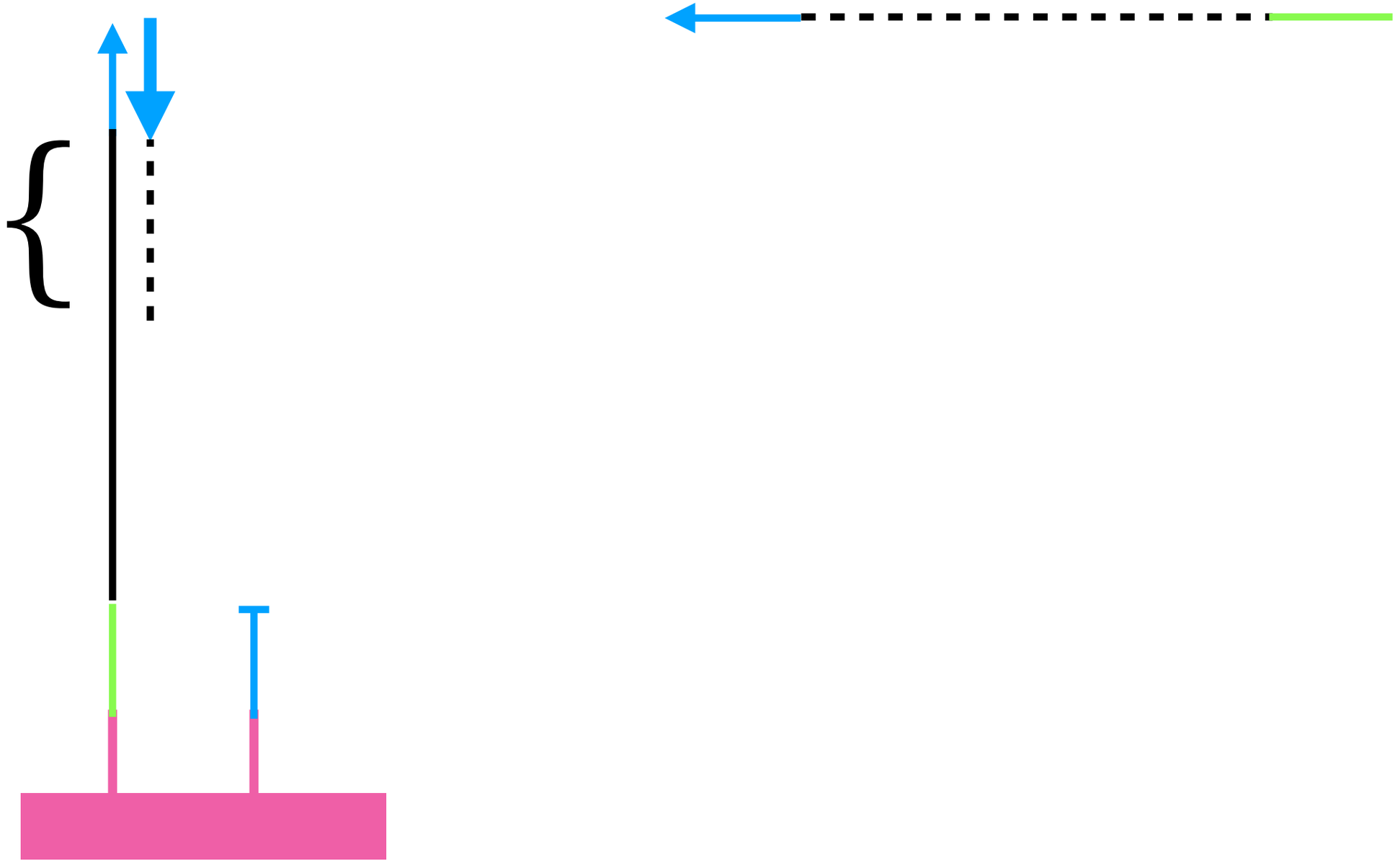
Парно-концевые риды

Paired-End sequencing



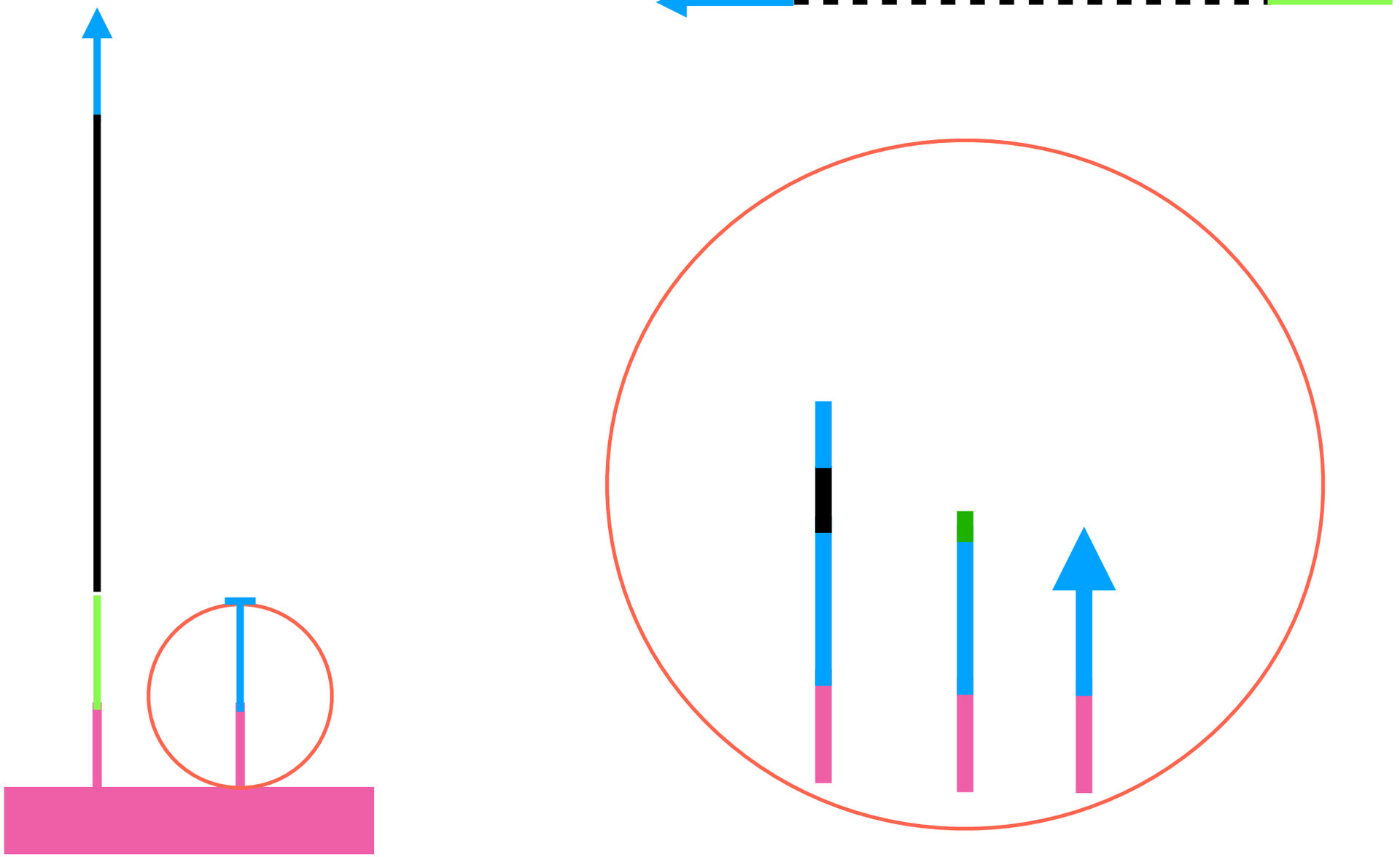
Парно-концевые риды

Paired-End sequencing



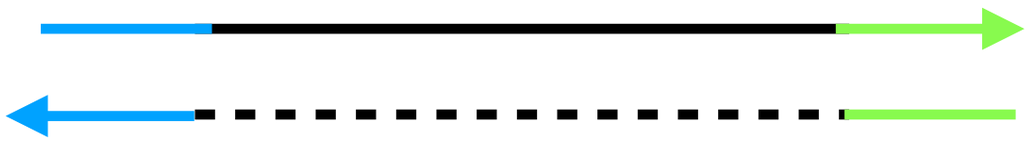
Парно-концевые риды

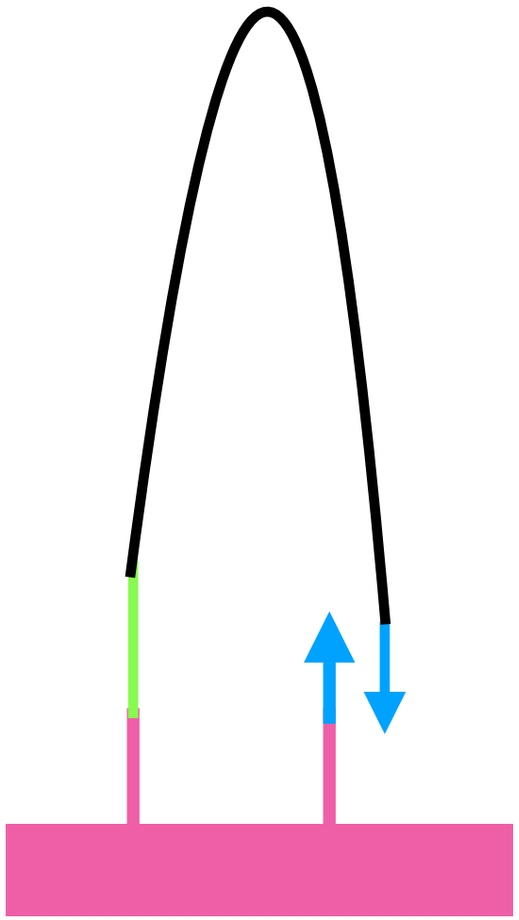
Paired-End sequencing



Парно-концевые риды

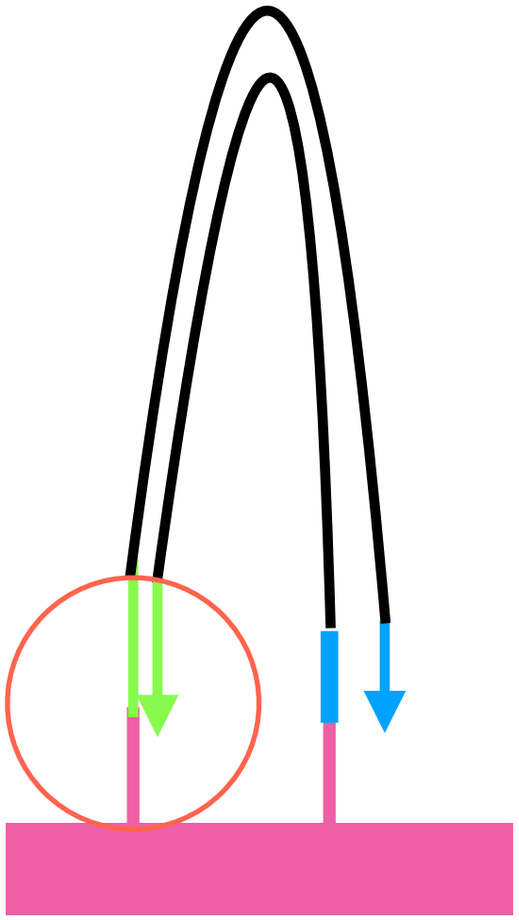
Paired-End sequencing

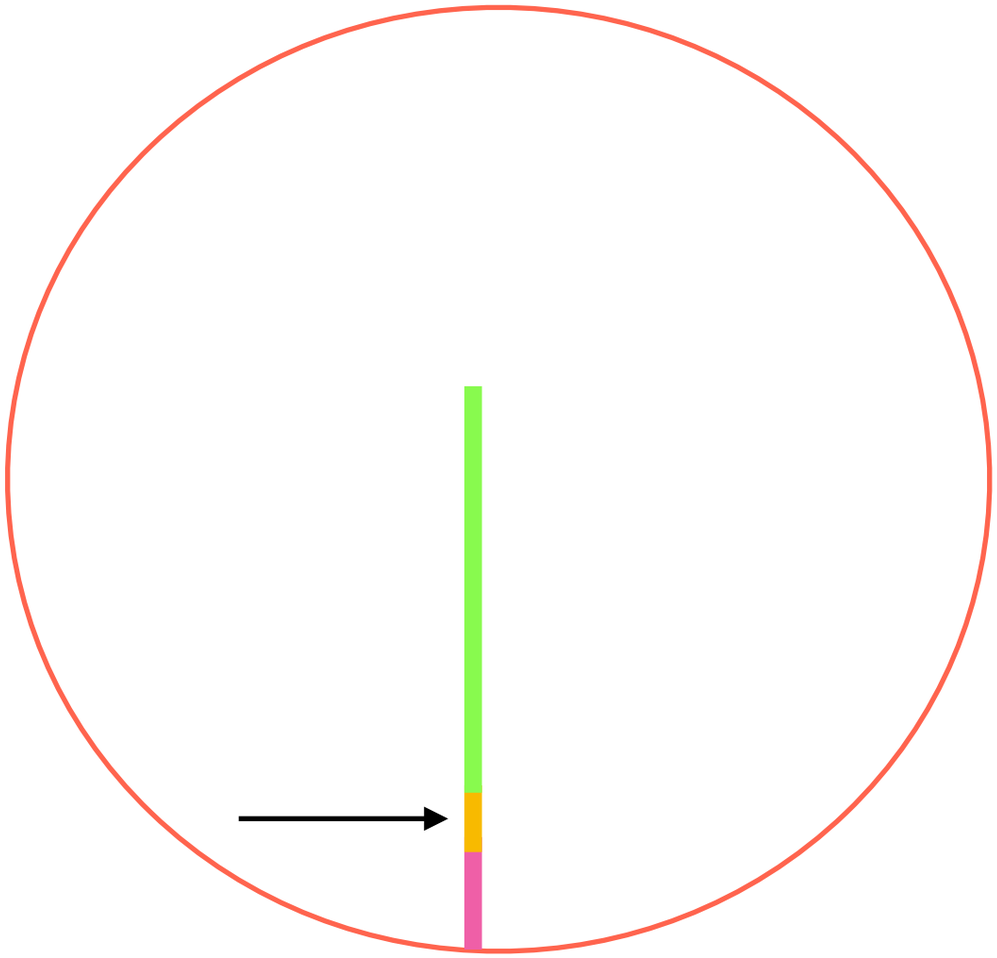




Парно-концевые риды

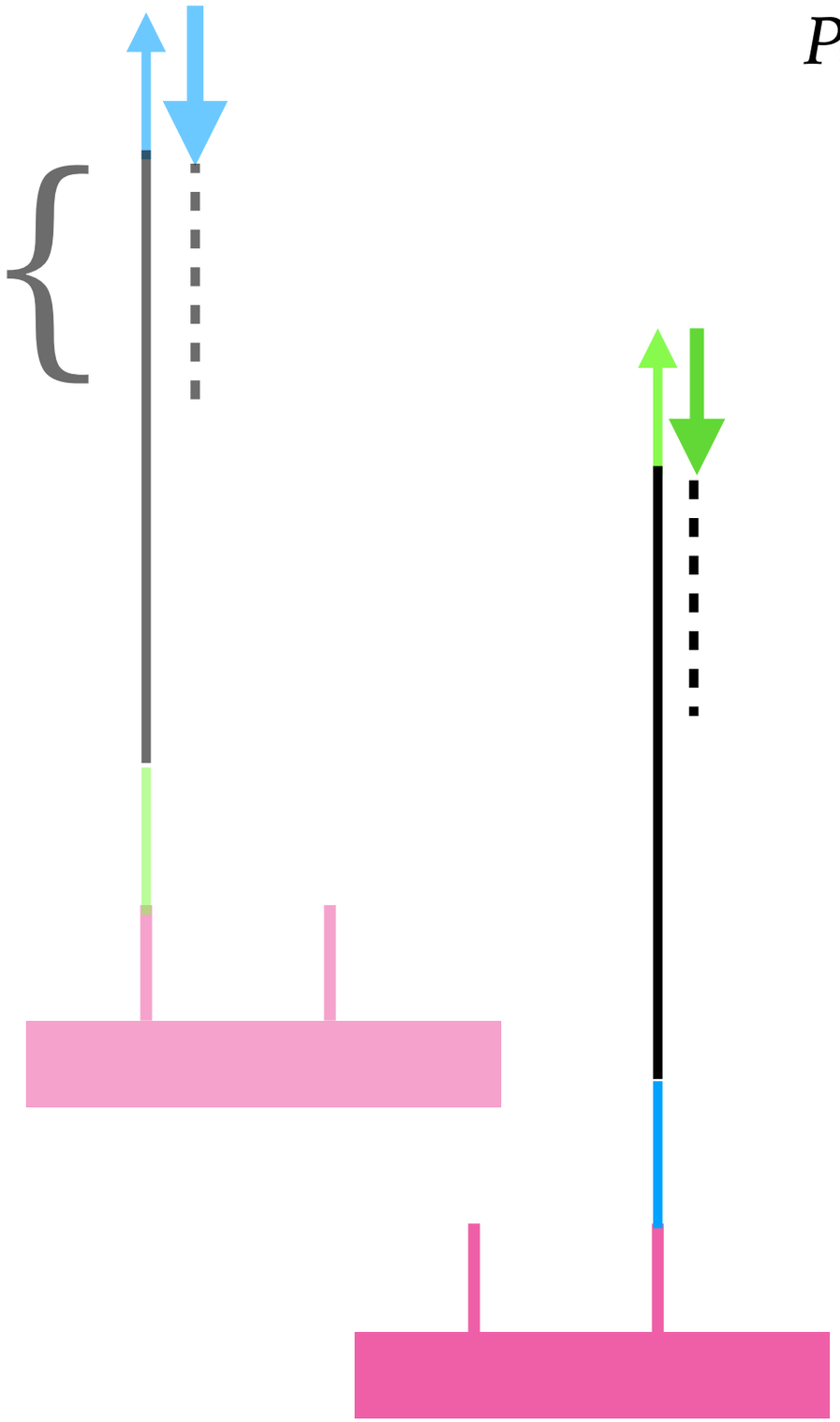
Paired-End sequencing

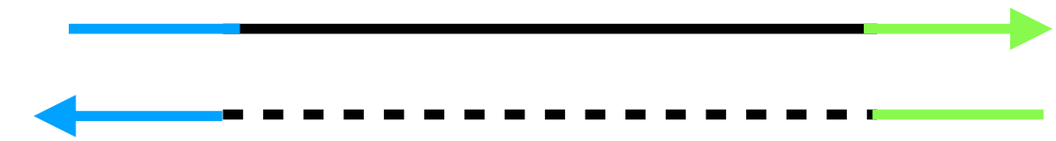




Парно-концевые риды

Paired-End sequencing

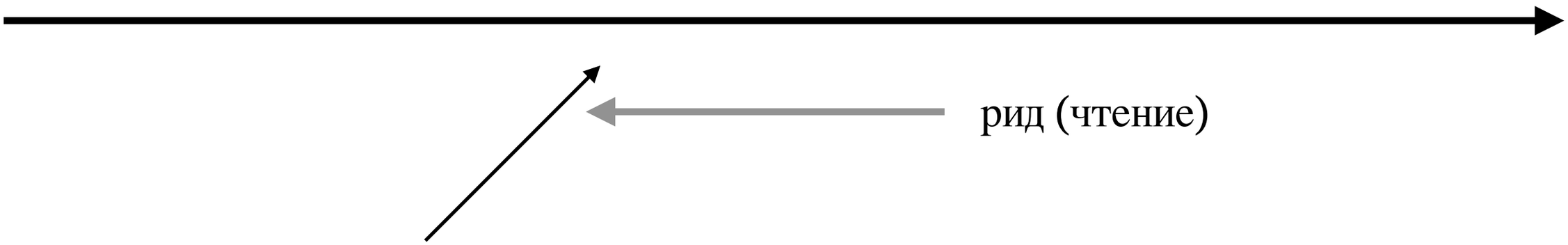




Термины

связанные с картированием чтений

референс (геном, сборка, скаффолд, контиг)



выравнивание / картирование

Основные понятия:

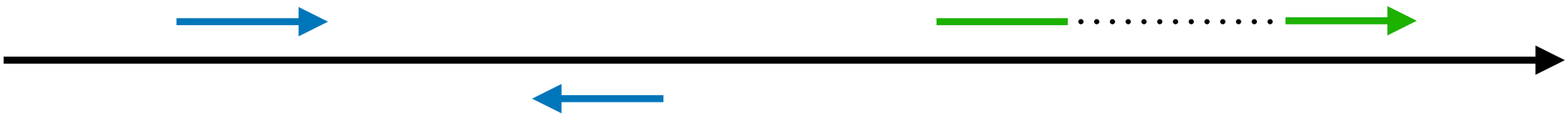
multimapper / unique mapper / unmapped

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

Техники, основанные на NGS

Что мы можем понять, используя данные NGS

RNA-seq

ChIP-seq

Hi-C

MNase-seq

CLIP-seq

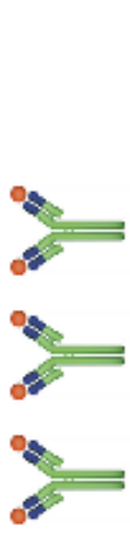
NET-seq

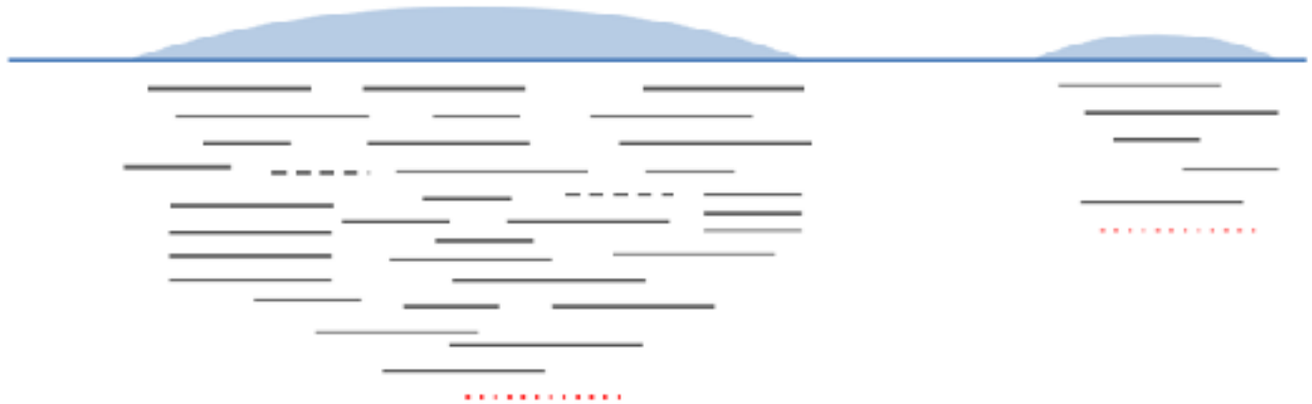
Ribo-seq

CAGE-seq

ATAC-seq

Exome-seq





Single-cell: да или нет?

Подход позволяет получить информацию о:

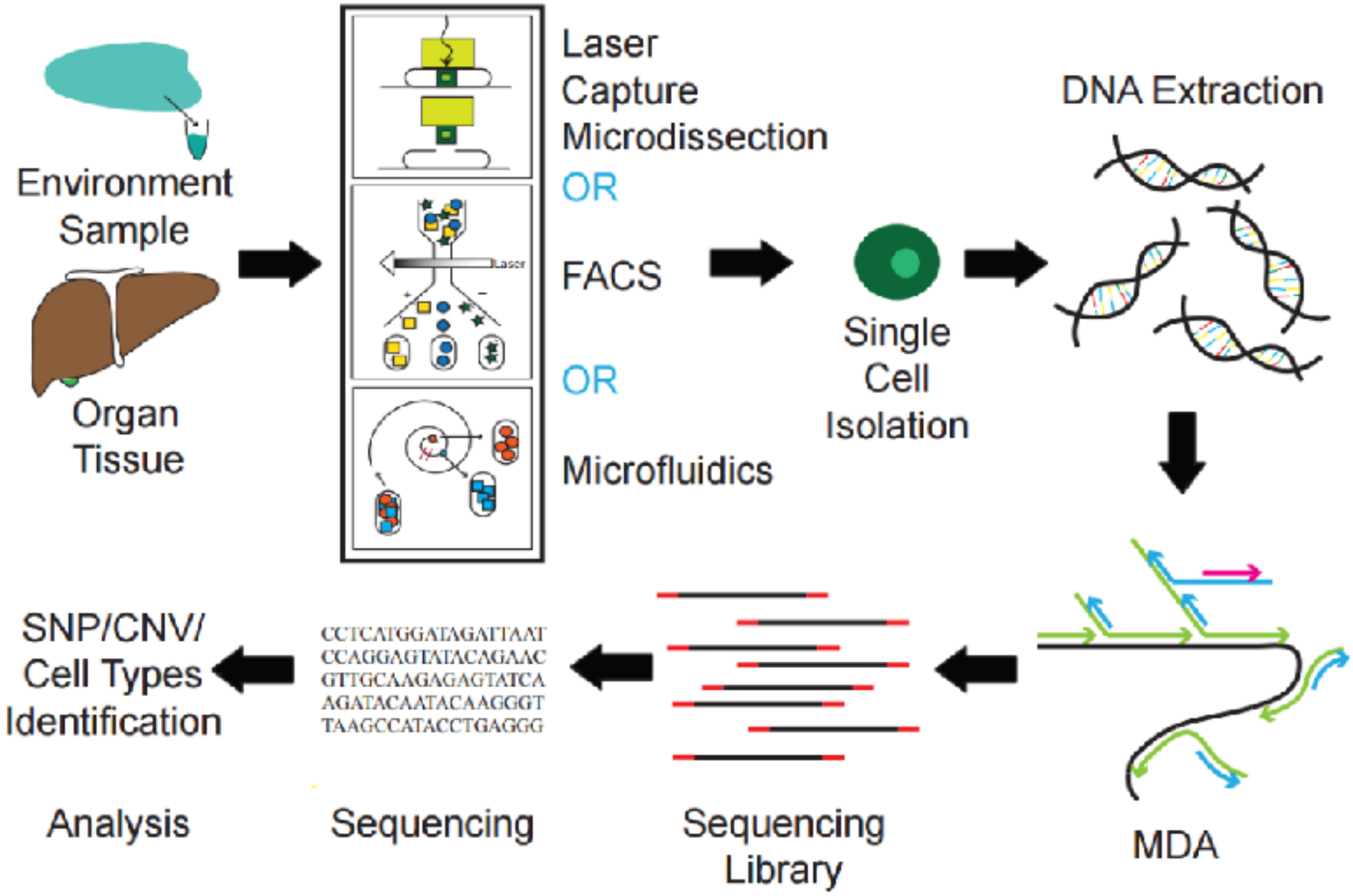
- Некультивируемых микроорганизмах (часто единственный возможный путь изучения)
- Редких типах клеток
- Гетерогенных образцах
- Полиморфизме соматических тканей
- Изменениях в клеточных линиях
- Развитии заболеваний

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})

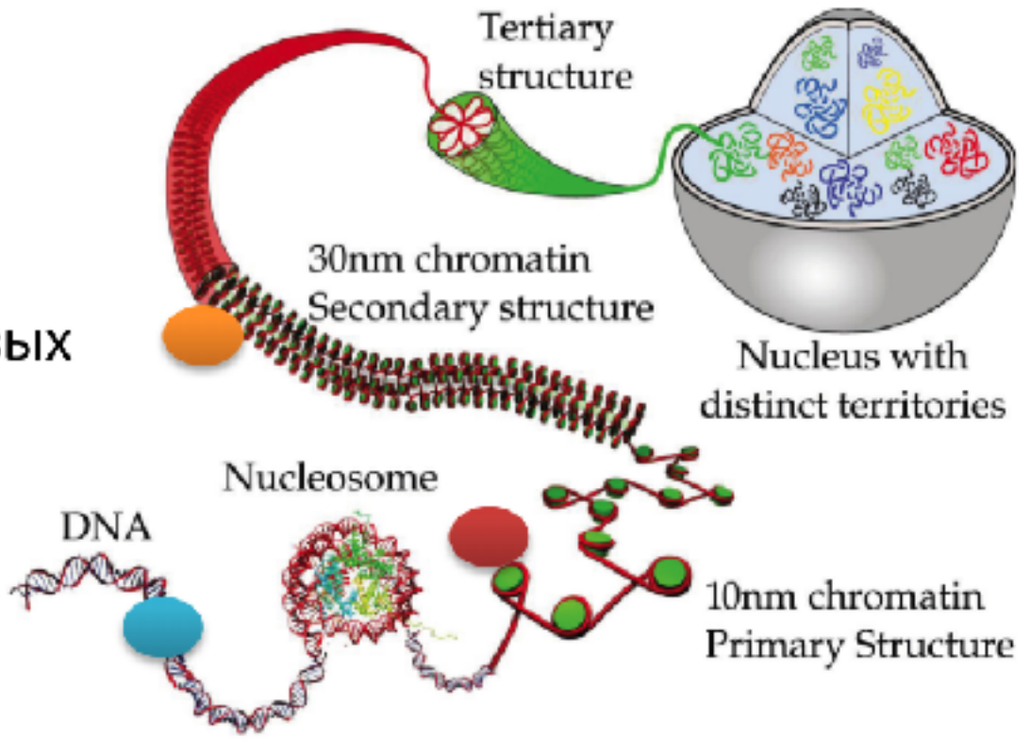
В реакции используется ДНК-полимераза бактериофага phi29 и случайные праймеры (random primers). В ходе изотермической реакции при 30°C происходит синтез многих копий исходной ДНК с **вытеснением** цепей. Средняя длина продукта – 12 т.п.н. (до 100 т.п.н.)

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

DNA-DNA / DNA-Protein Interactions

HiC / ChipSeq

Изучение ДНК-белковых взаимодействий



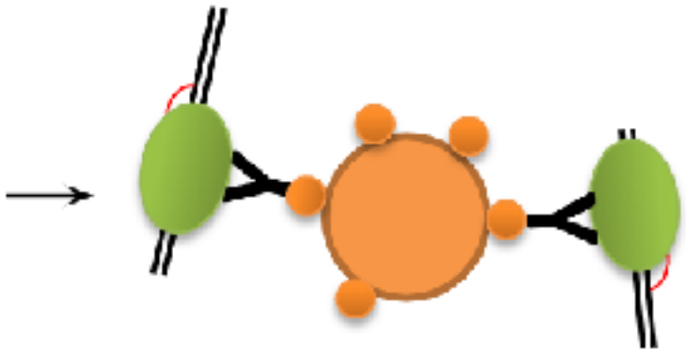
Изучение пространственной организации

ChipSeq

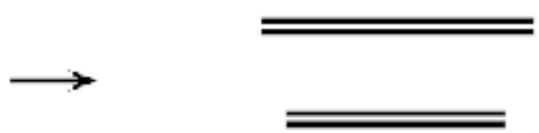
"Wet-lab"

Сшивка формальдегидом

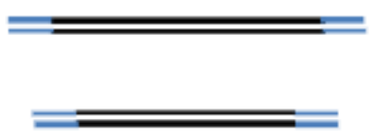
Фрагментация ДНК



Иммунопреципитация сшитых ДНК-белковых комплексов



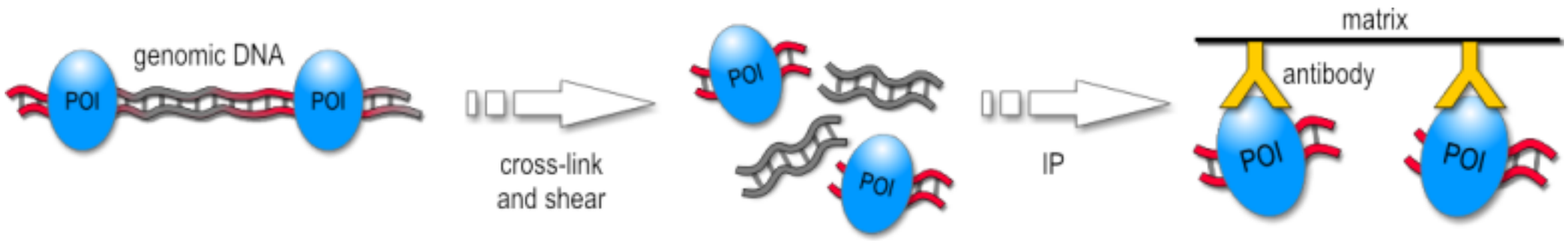
Освобождение ДНК (прогреванием)



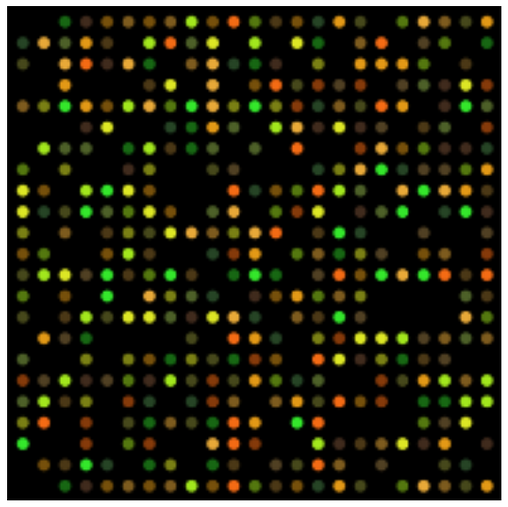
Подготовка библиотеки и секвенирование

ChipSeq

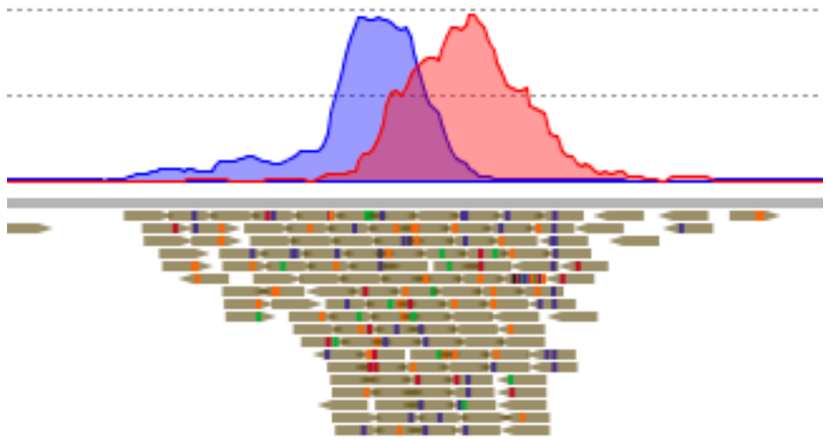
Mapping of DNA-protein contacts



Chip-on-Chip

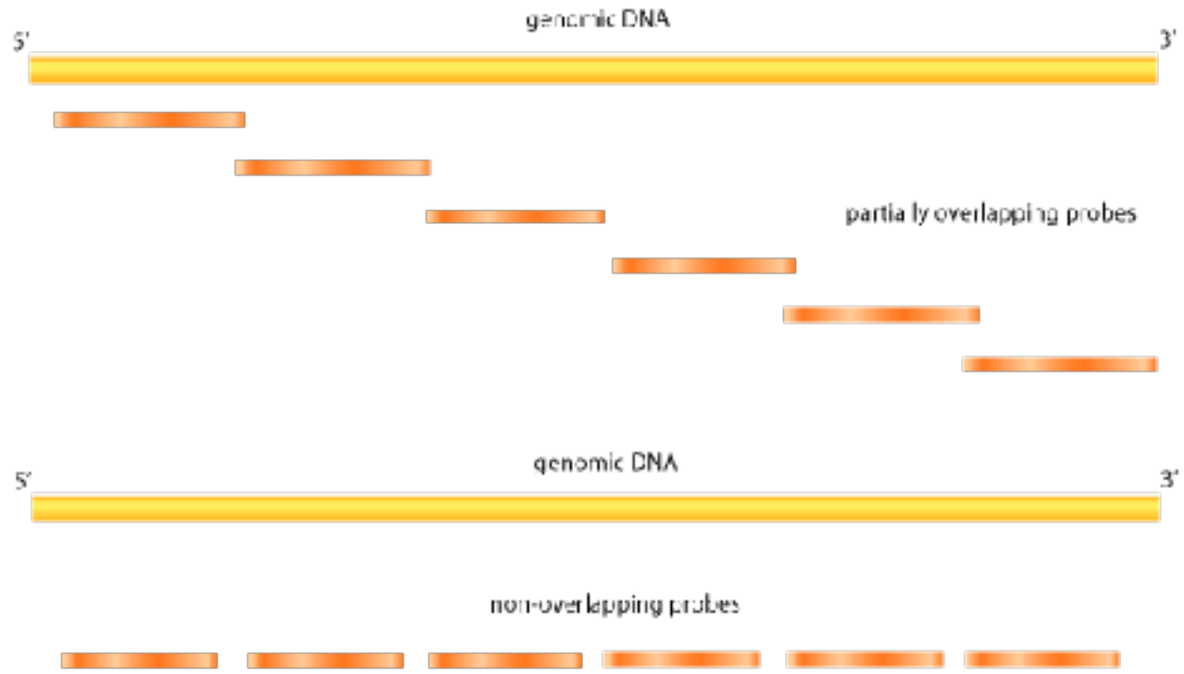


Chip-Seq



Chip-on-Chip

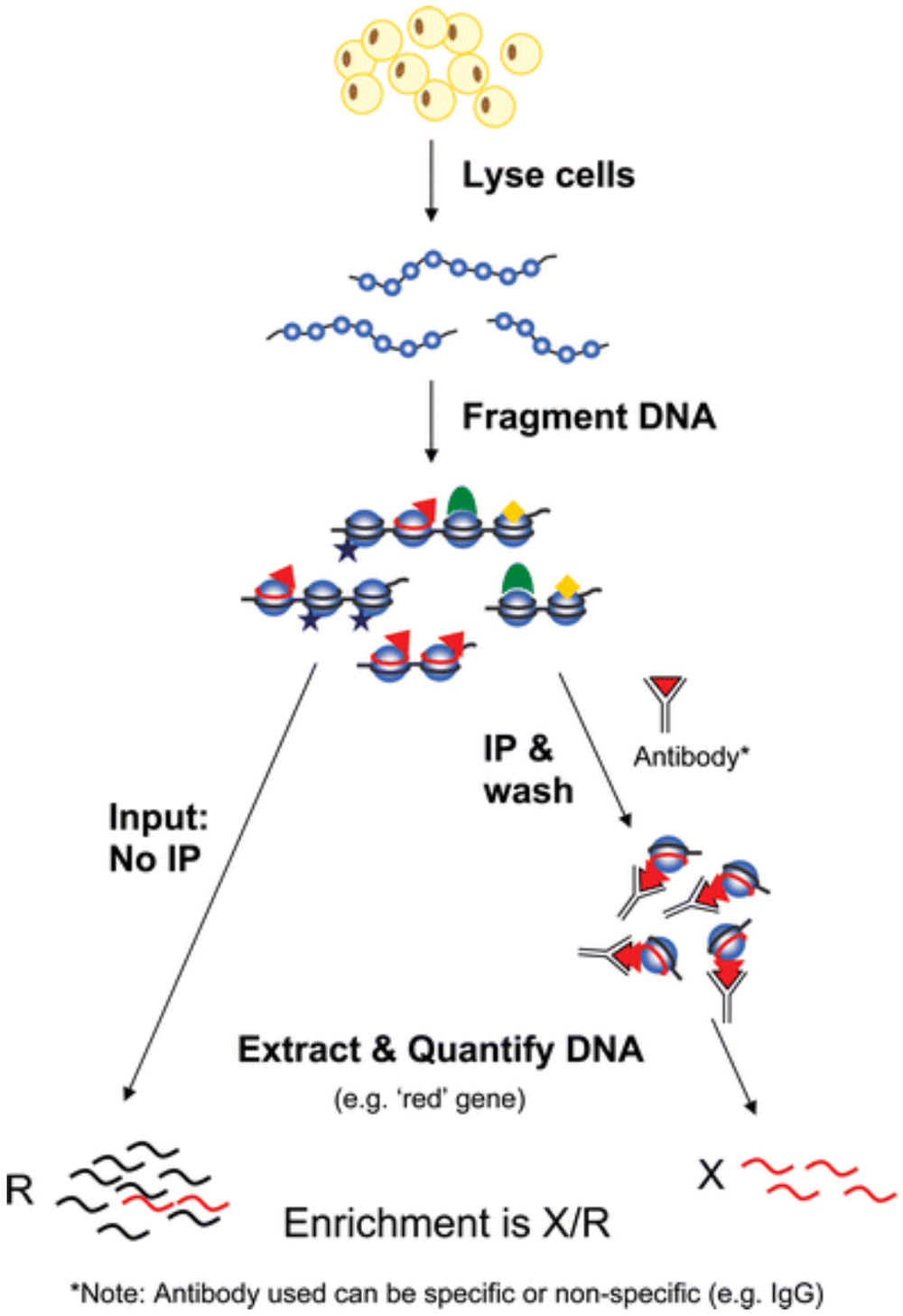
Enrichment of target fragments detection



Affymetrix, NimbleGene, Agilent

ChipSeq

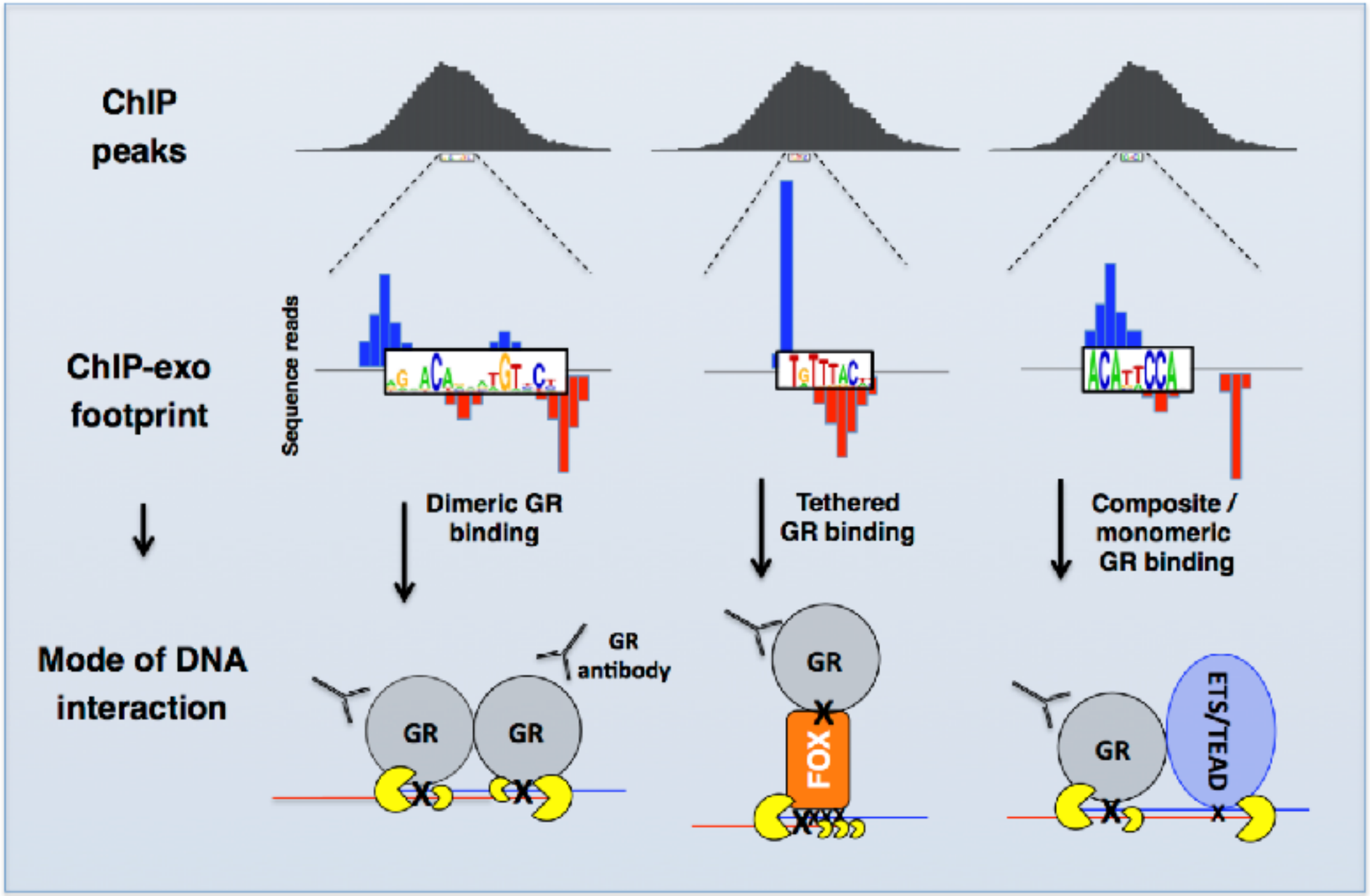
Enrichment of target fragments detection



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

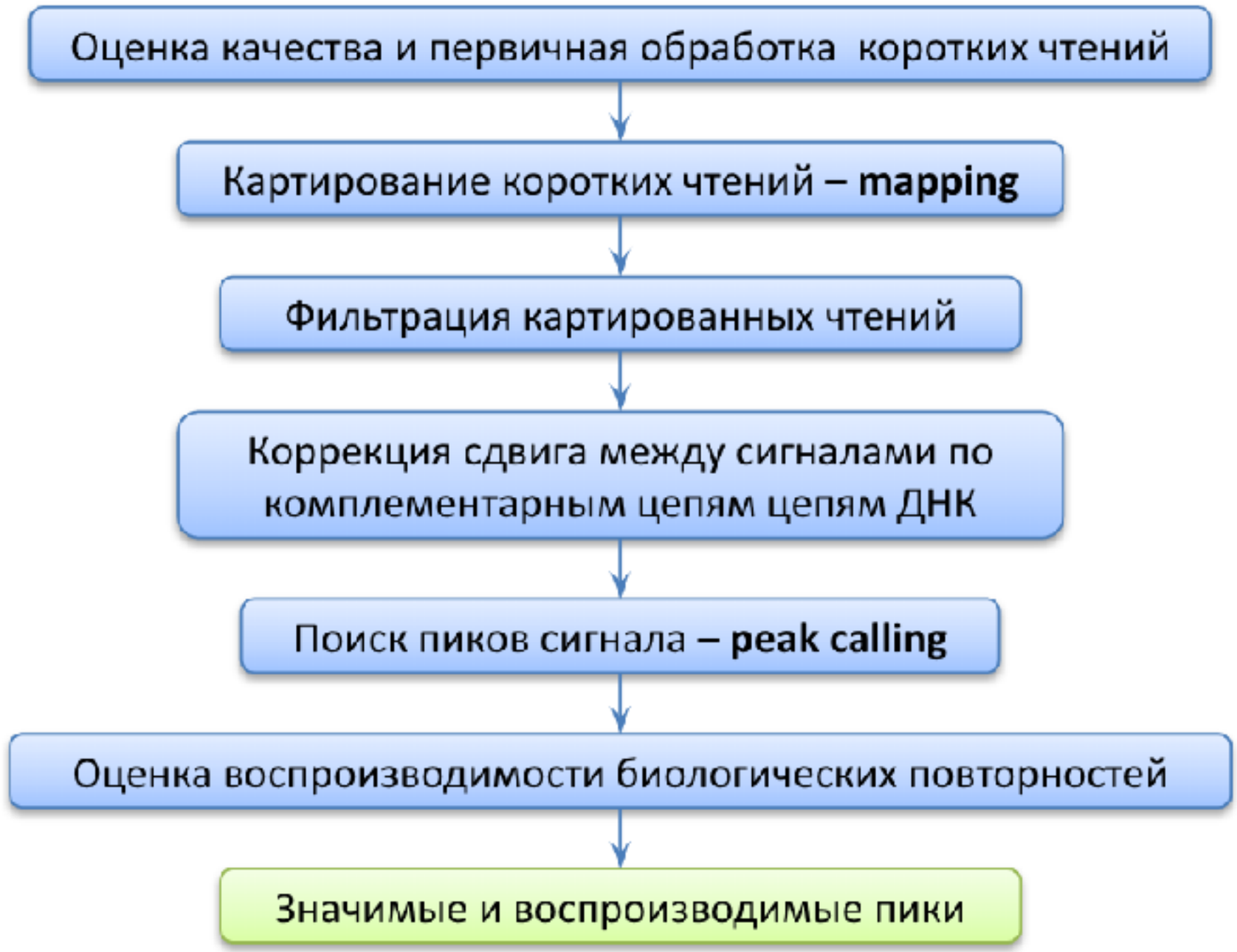
ChipExo

Определение типа взаимодействия



ChipSeq

Pipeline



ChipSeq

Processing

Remove multimappers

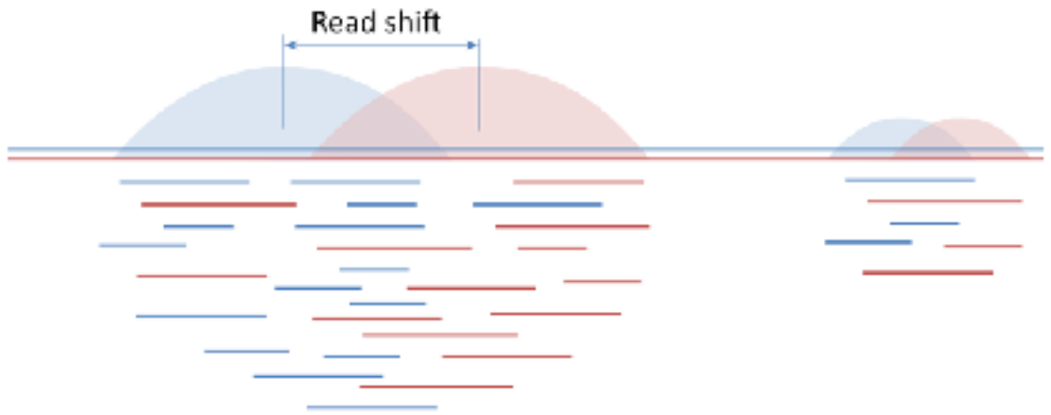
Remove duplicates

Remove low-quality reads

ChipSeq

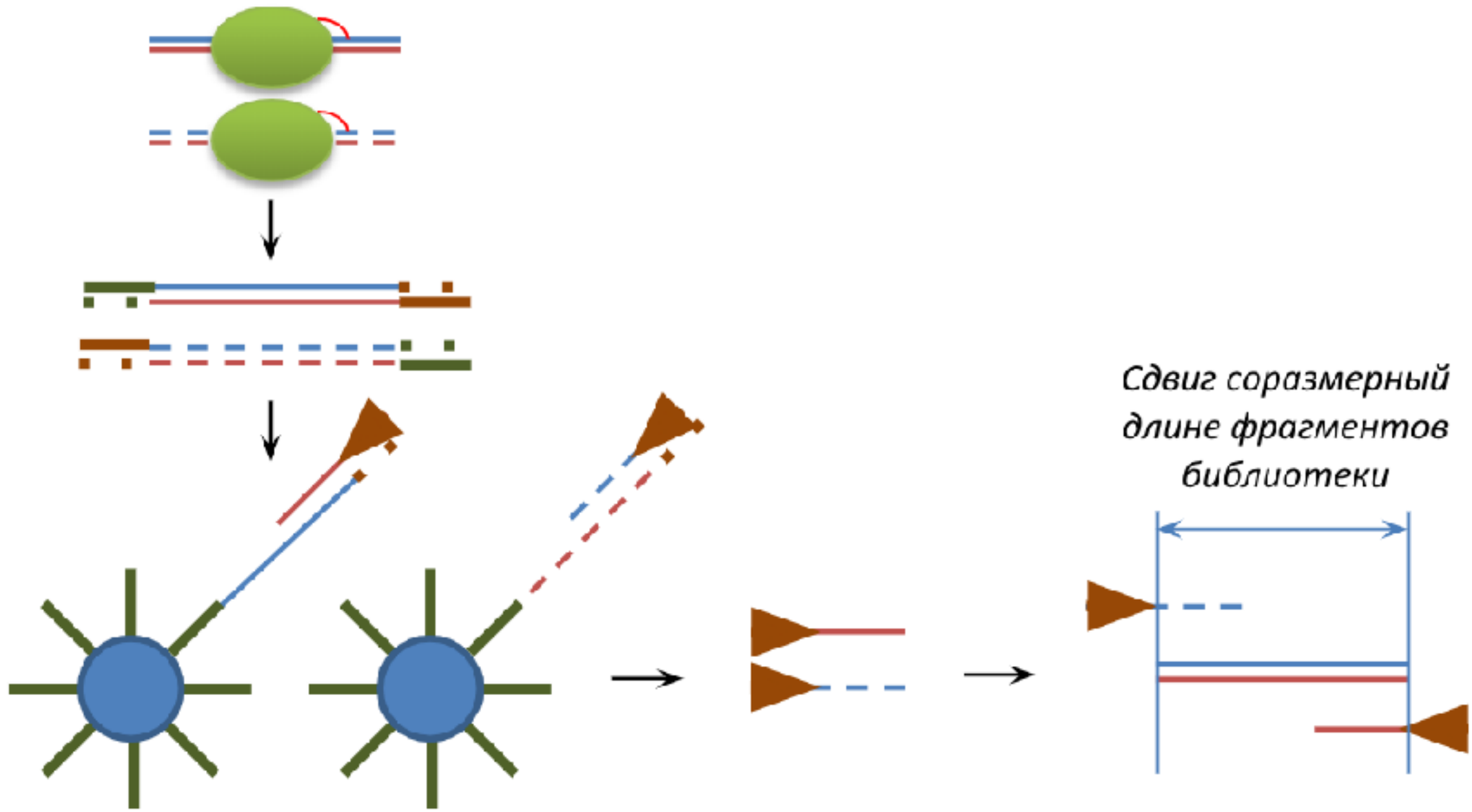
Tag shift correction

Detect peak



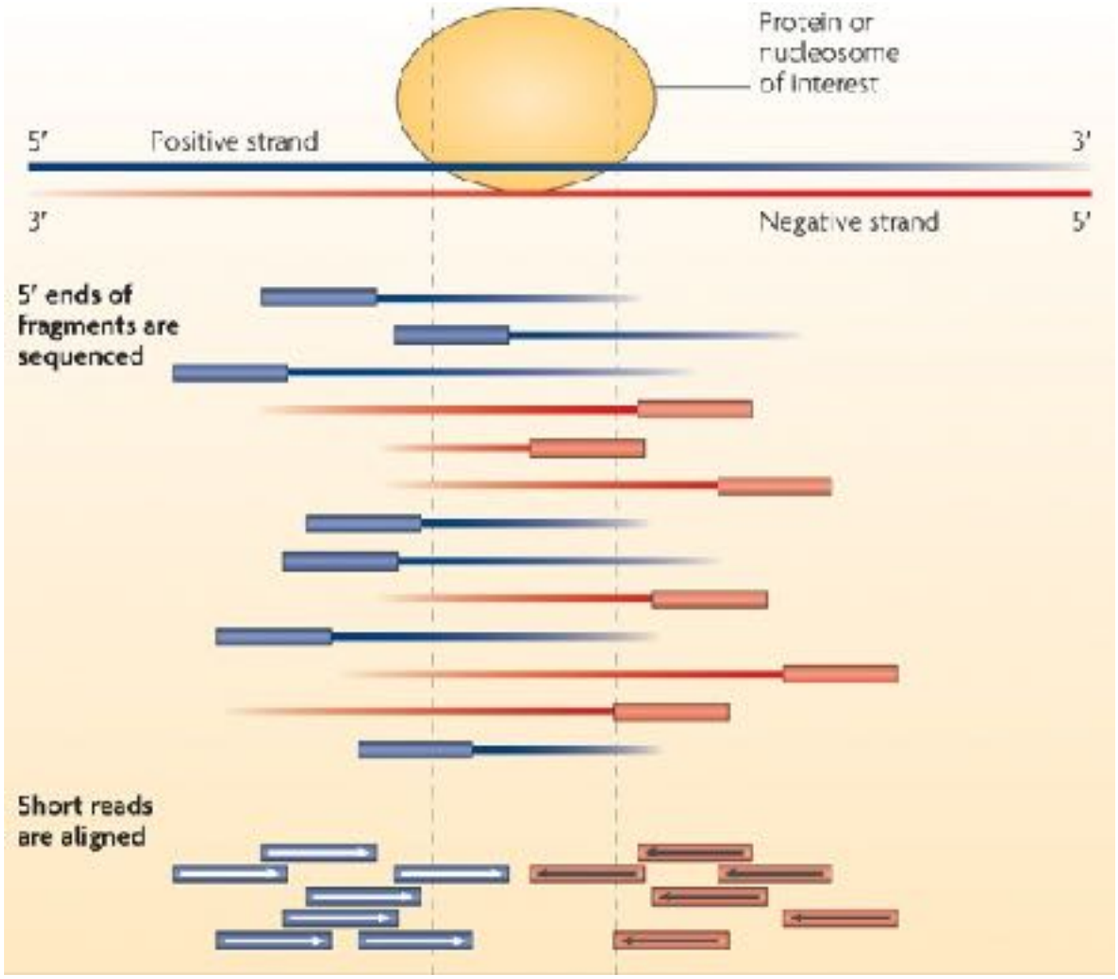
ChipSeq

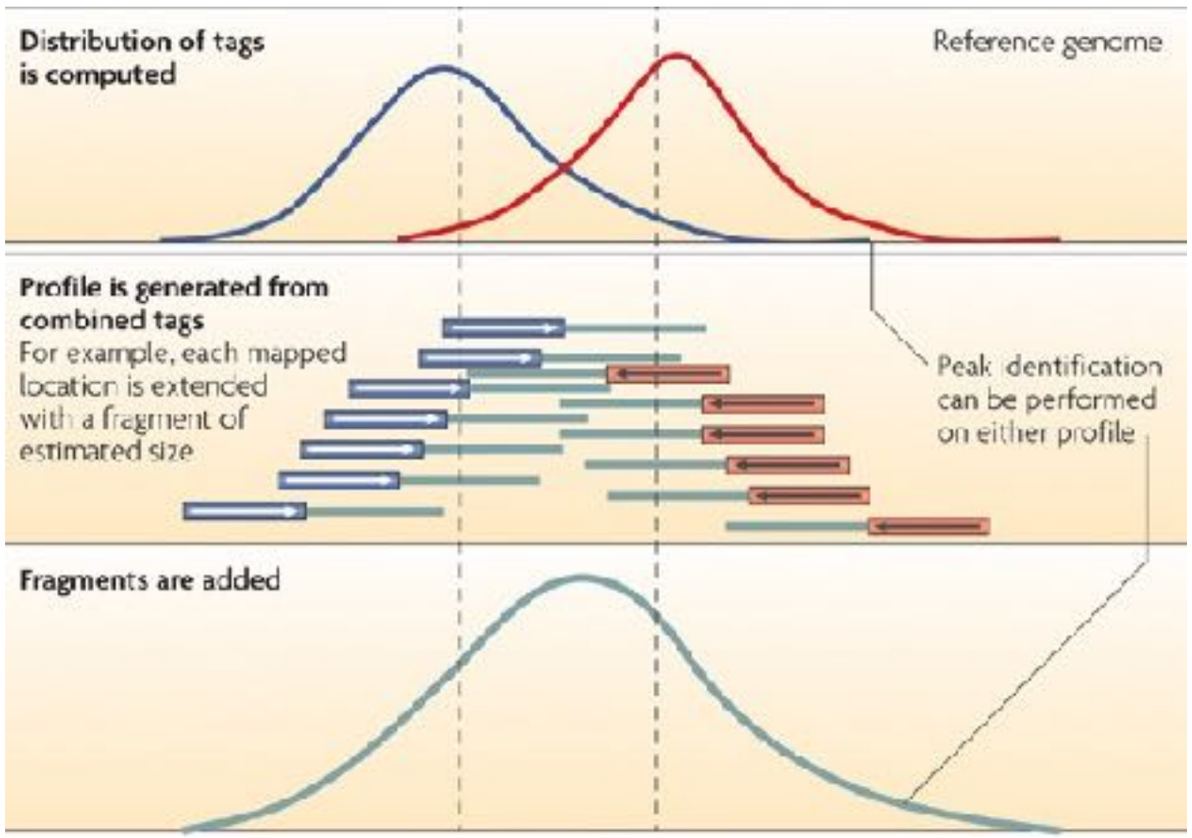
Why tags are shifted?



ChipSeq

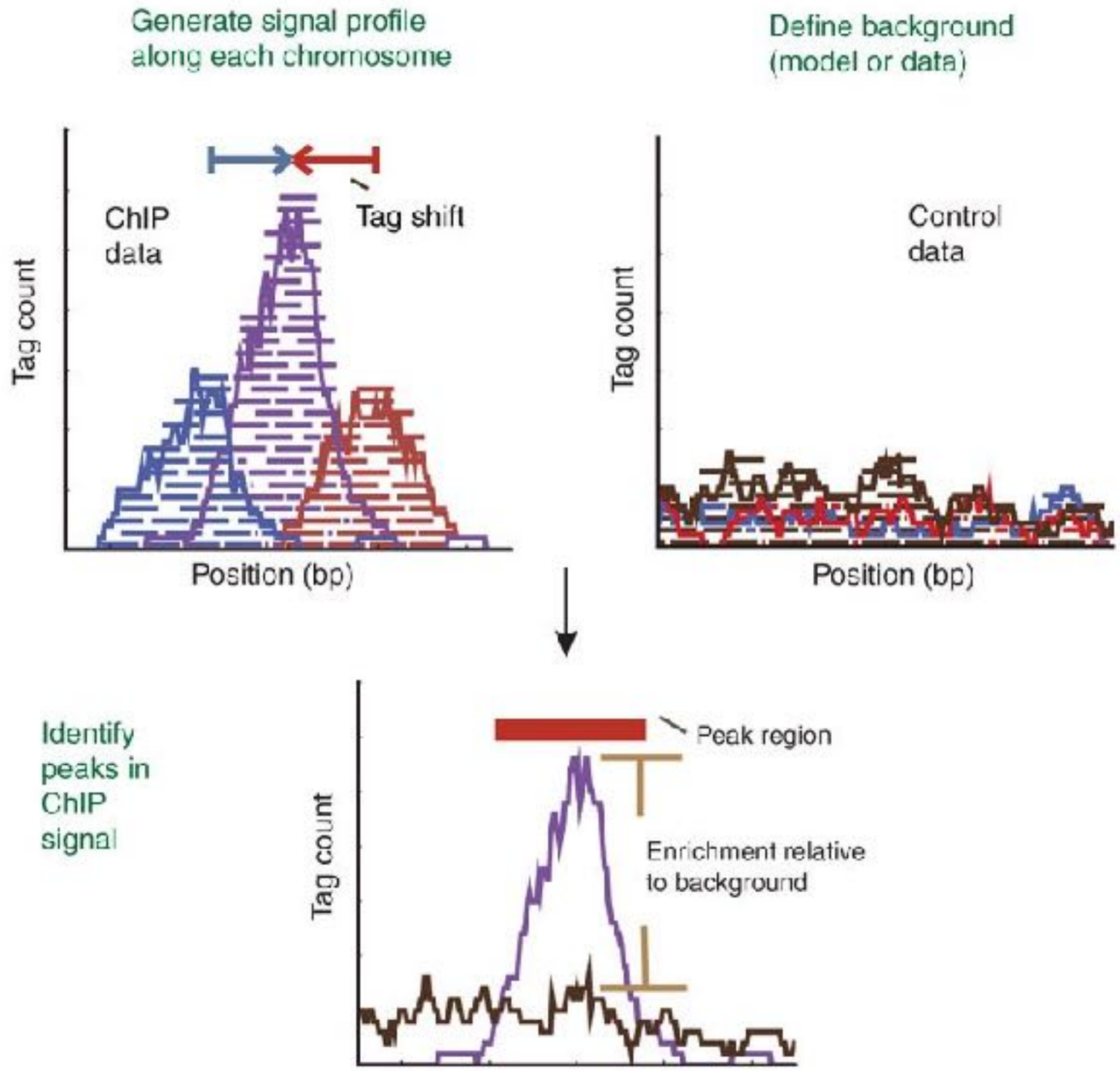
Tag shift corrected





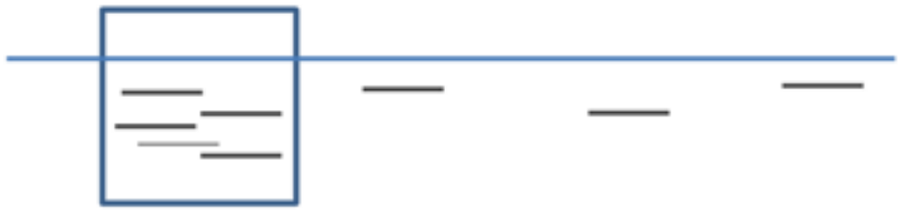
ChipSeq

Pipeline



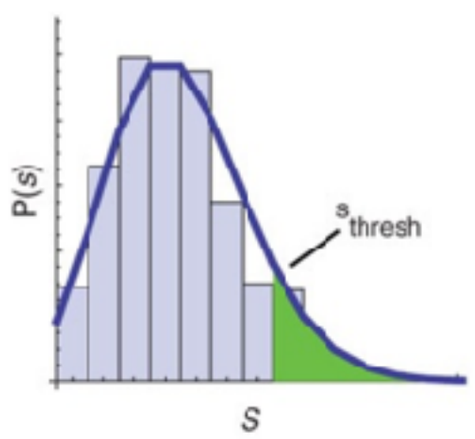
ChipSeq

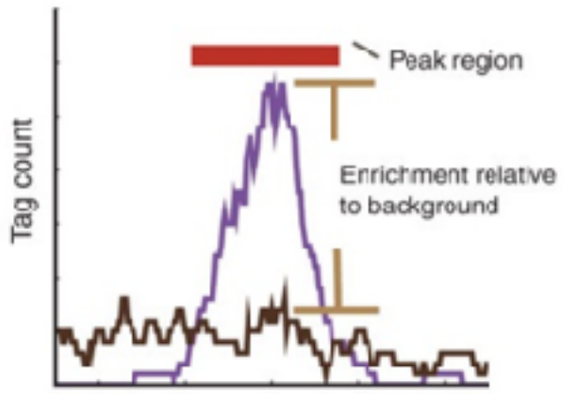
Enrichment detection



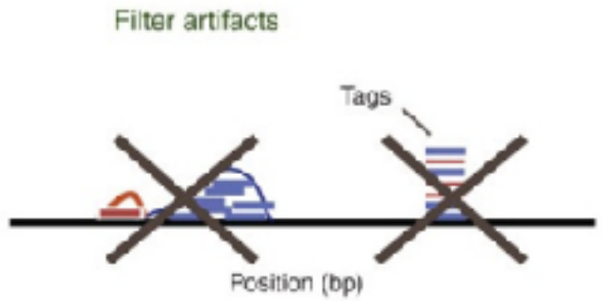
Какова вероятность обнаружить такое количество чтений в окне шириной *d* нт?

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





Pepke et al. (2009) Nature Methods. 6



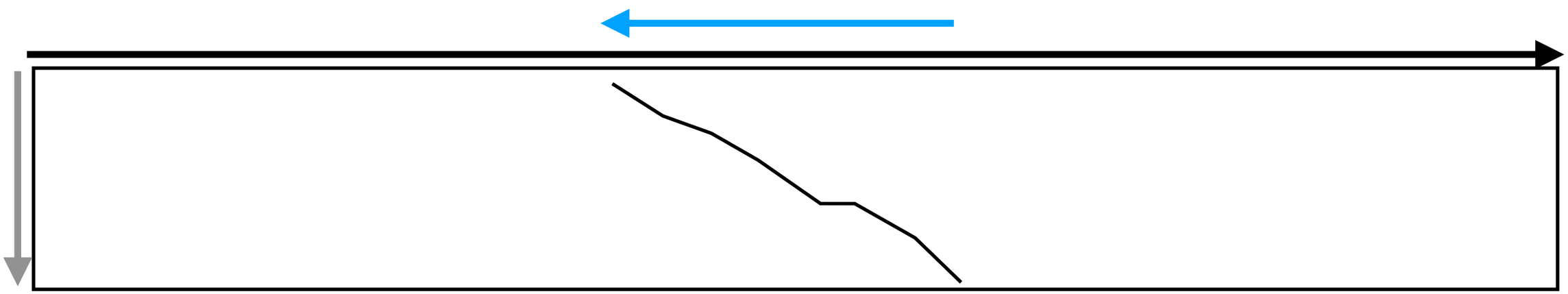
2 основных типа артефактных сигналов:

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

Картирование

Как быстро и качественно получить сигнал?

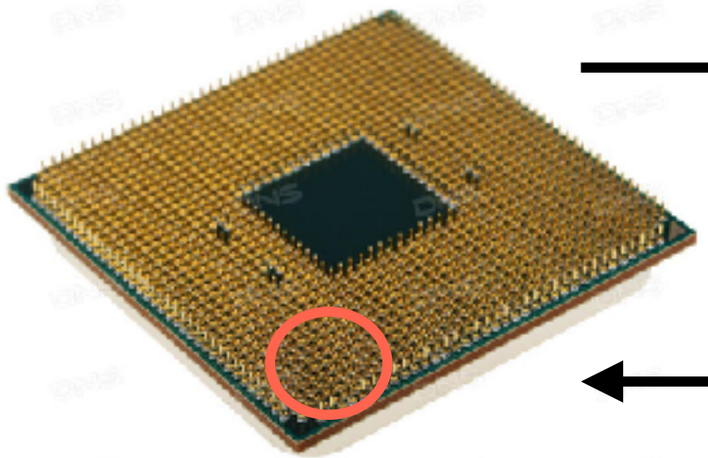
Mapping



Займет $\sim O(kN \times M)$

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

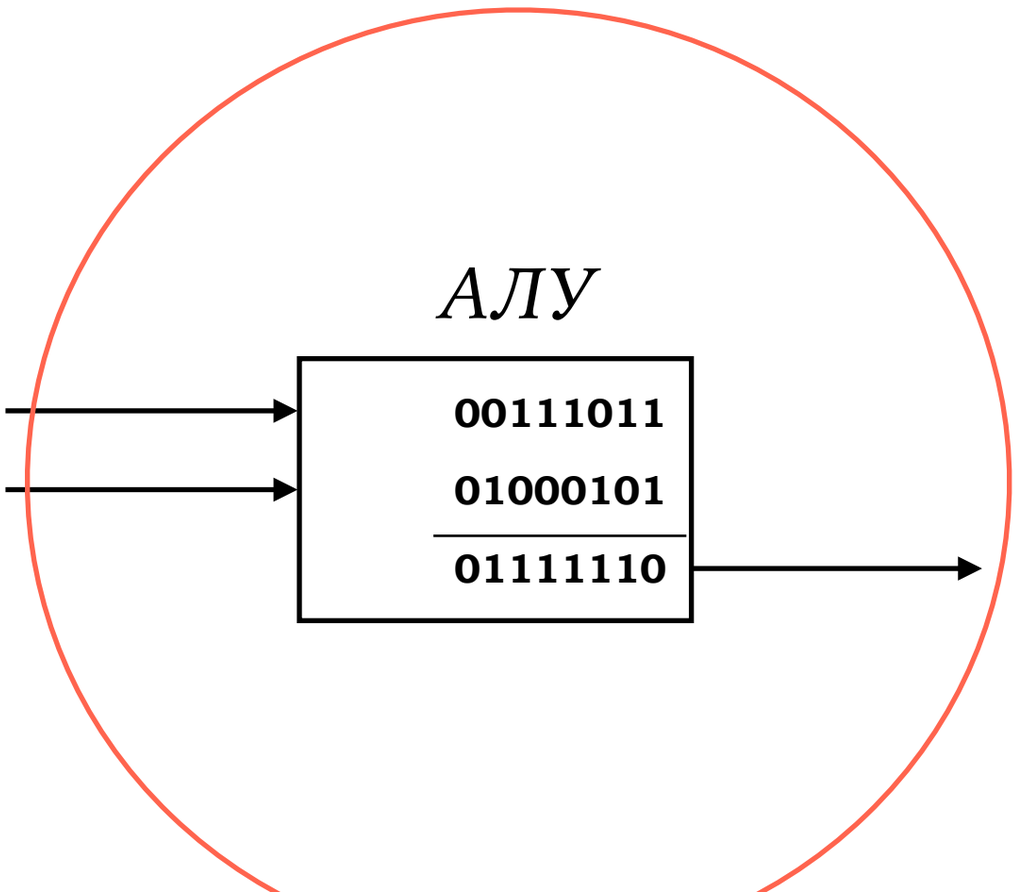
Немного о big-O



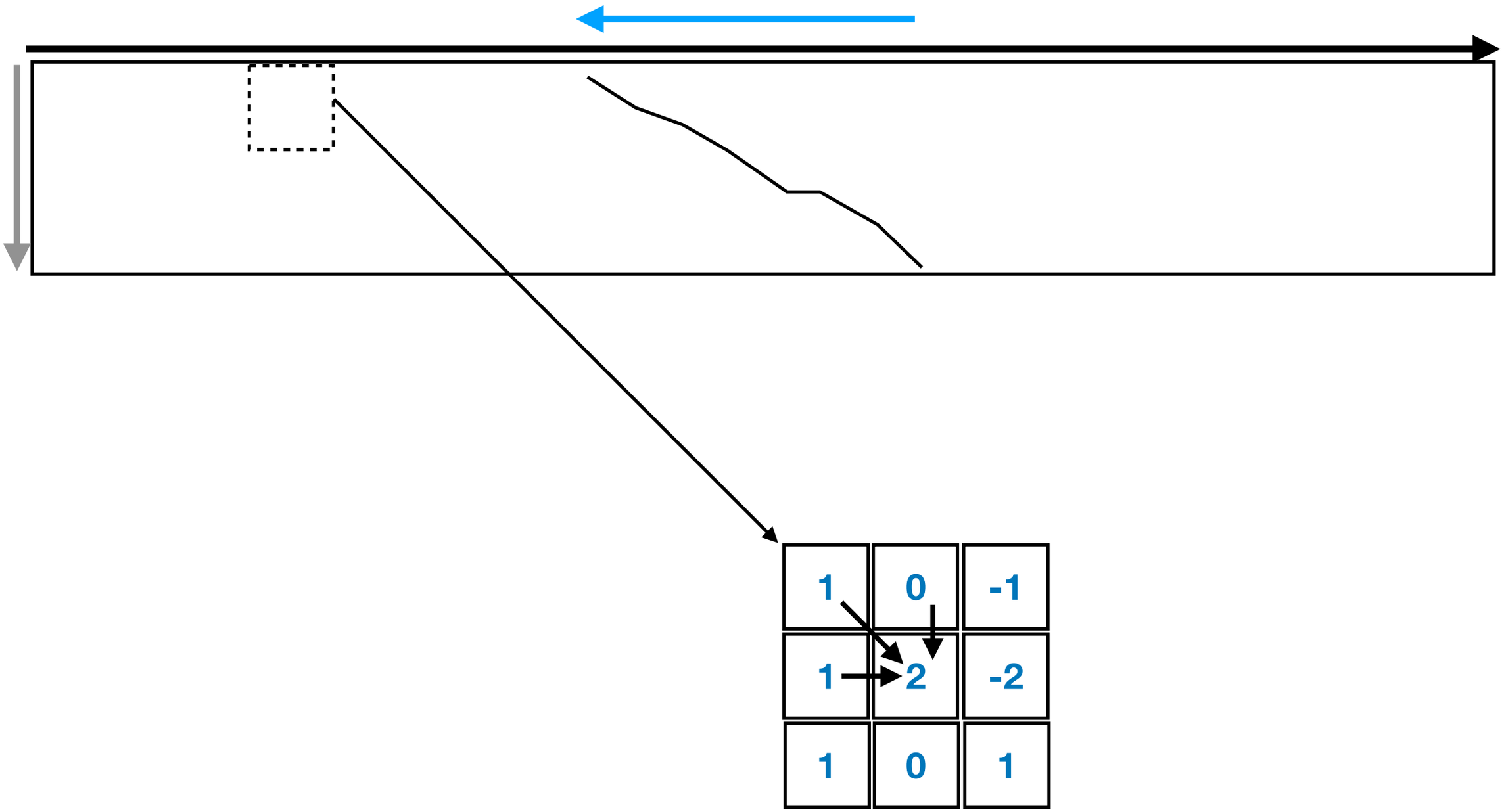
адрес

данные

56266211	00001000
45623467	00111011
12375478	11110000
67224784	11010100

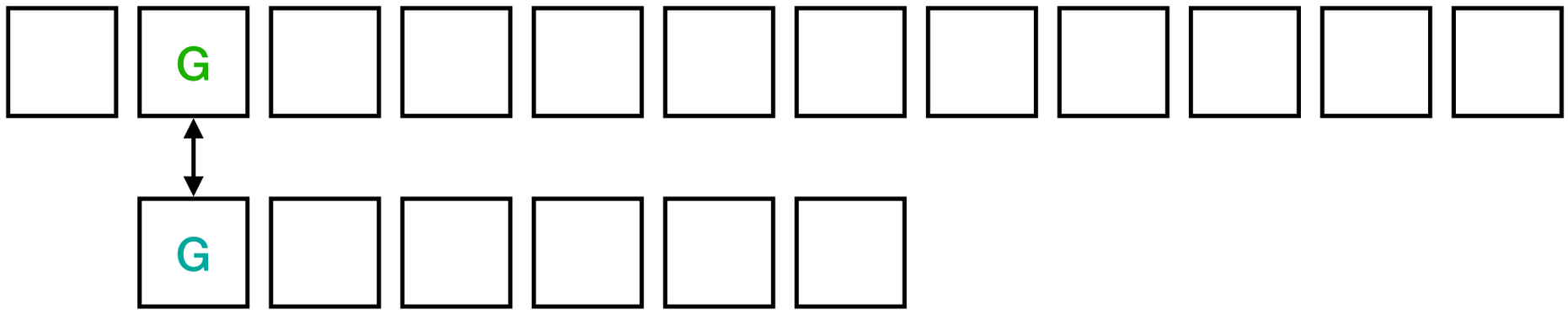


Mapping

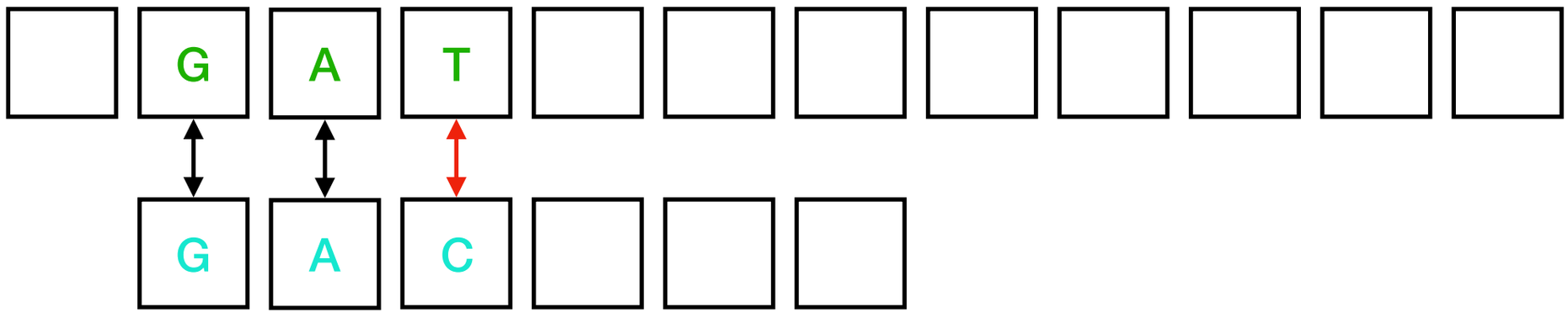


Mapping

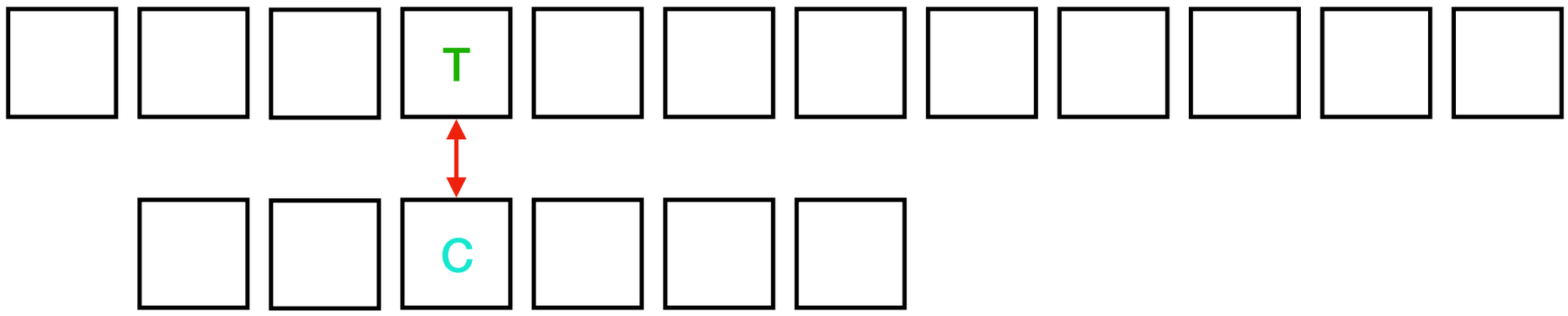
Попробуем exact pattern match?



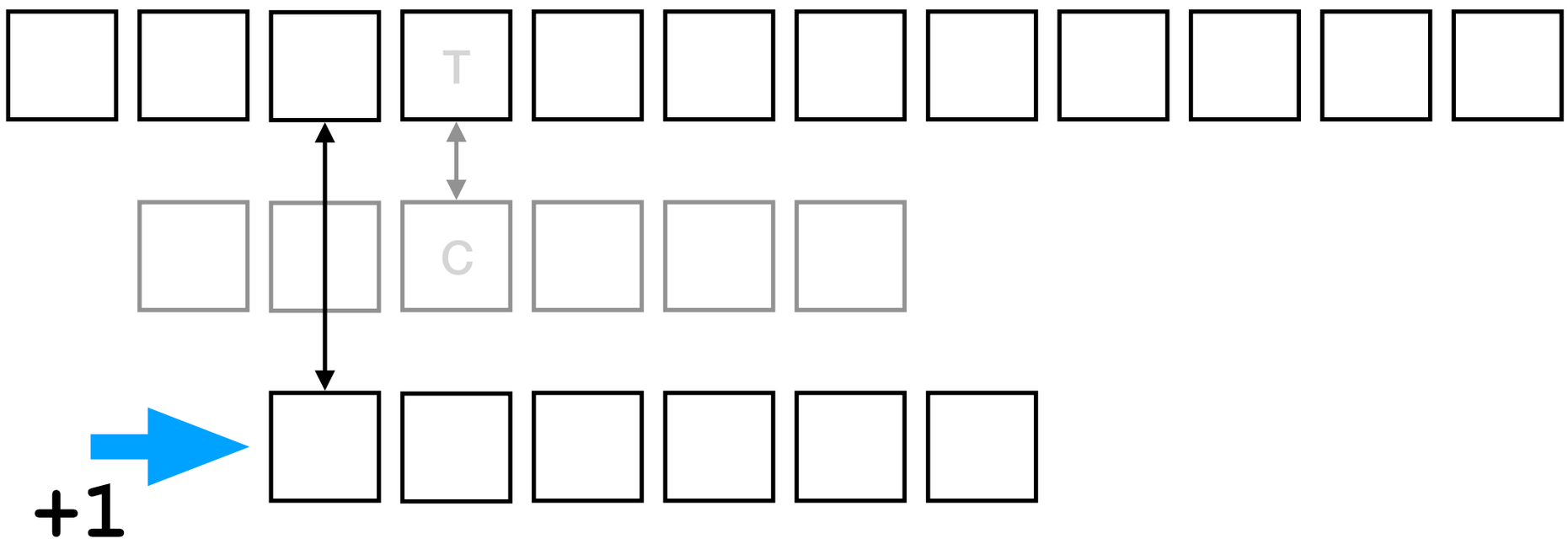
Mapping



Mapping



Mapping

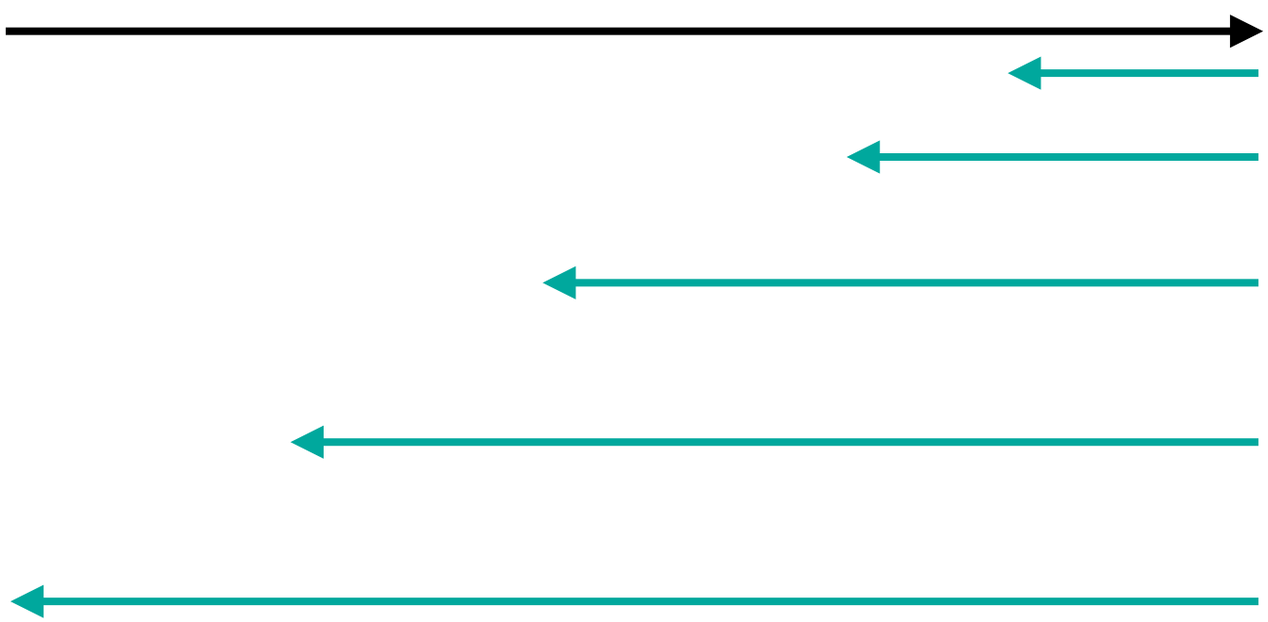


"Наивный" алгоритм опять требует O(N×M) времени

(но уже с лучшей константой)

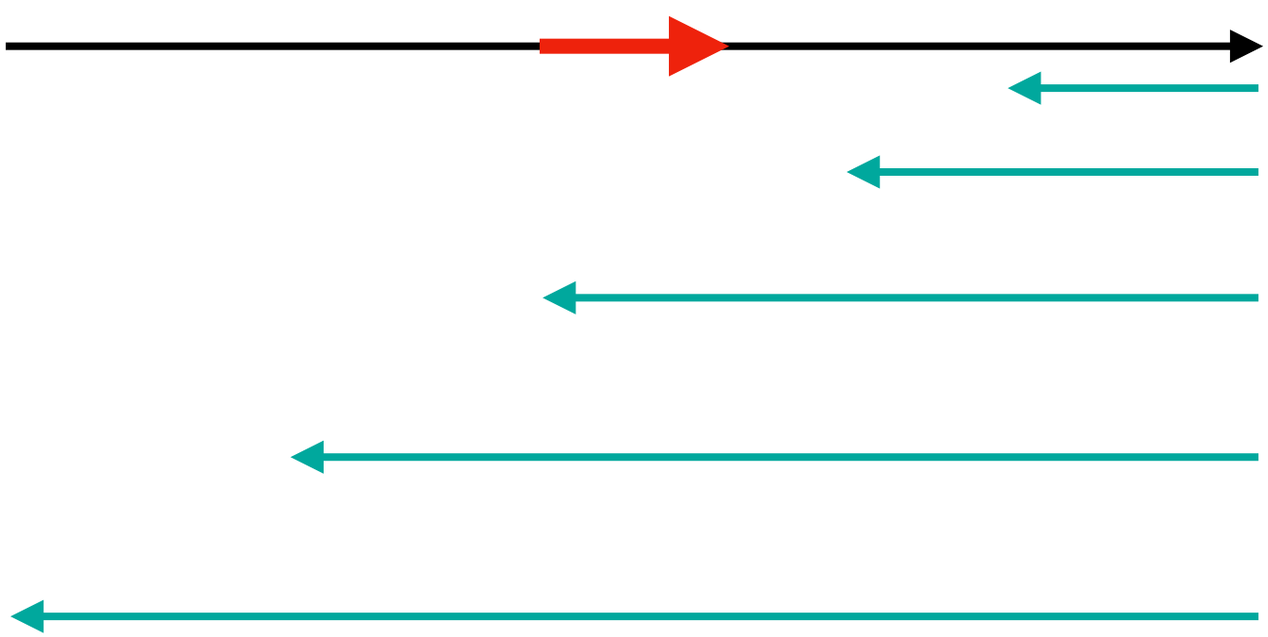
Mapping

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



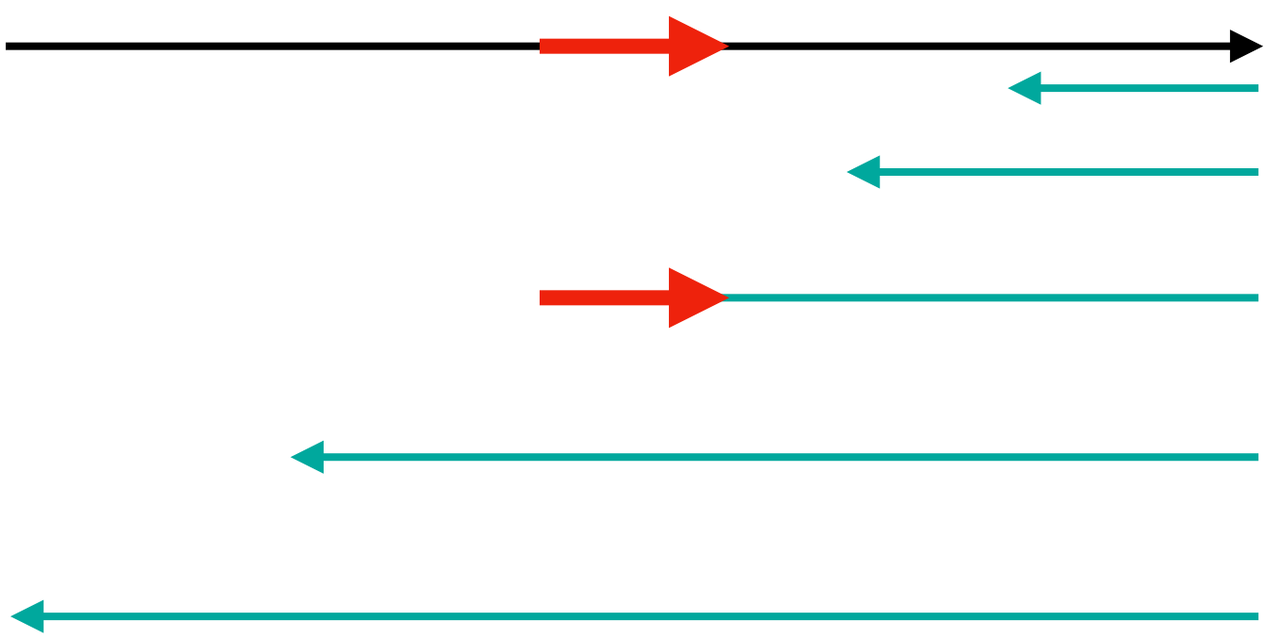
Mapping

Менее тривиальное решение



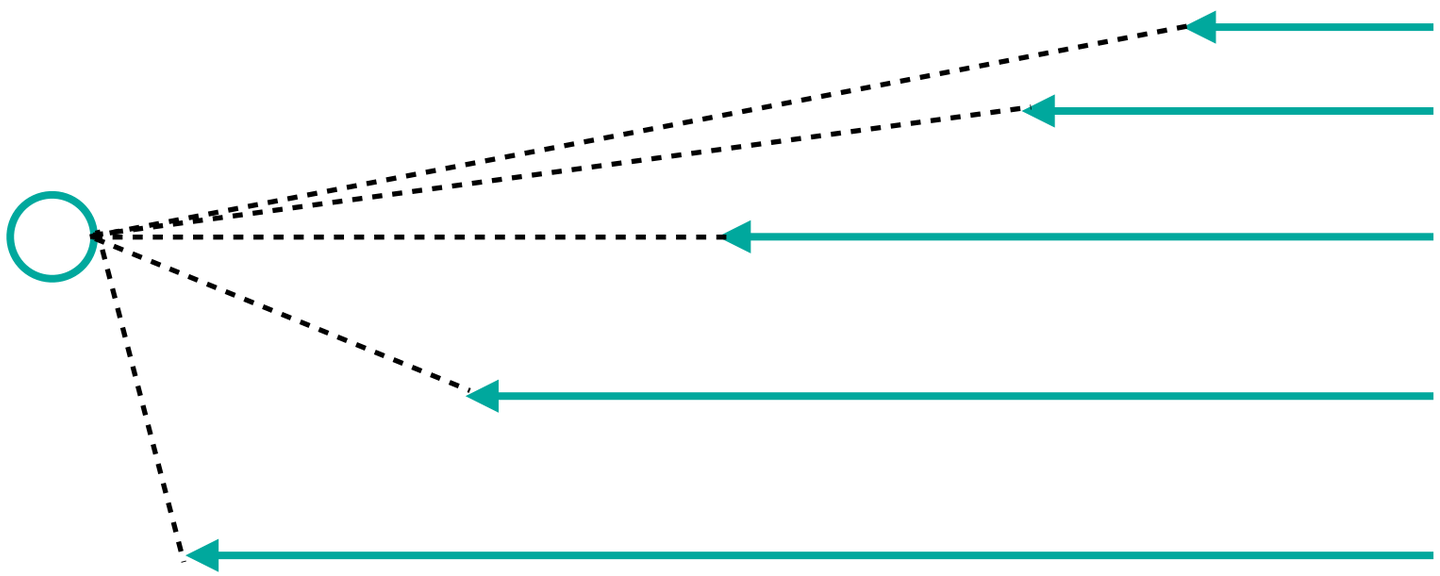
Mapping

Подстрока исходной строки
есть префикс какого-то
суффикса исходной строки



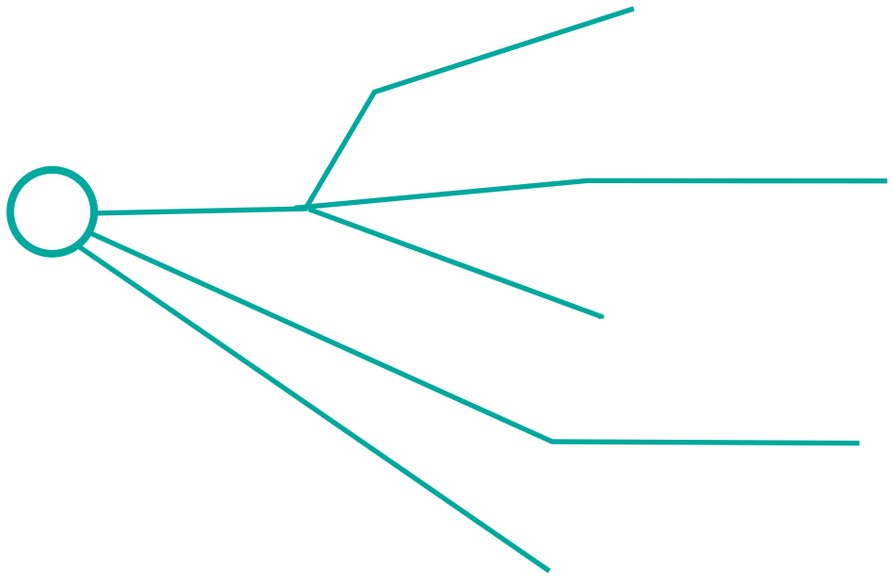
Mapping

Подстрока исходной строки
есть префикс какого-то
суффикса исходной строки



Mapping

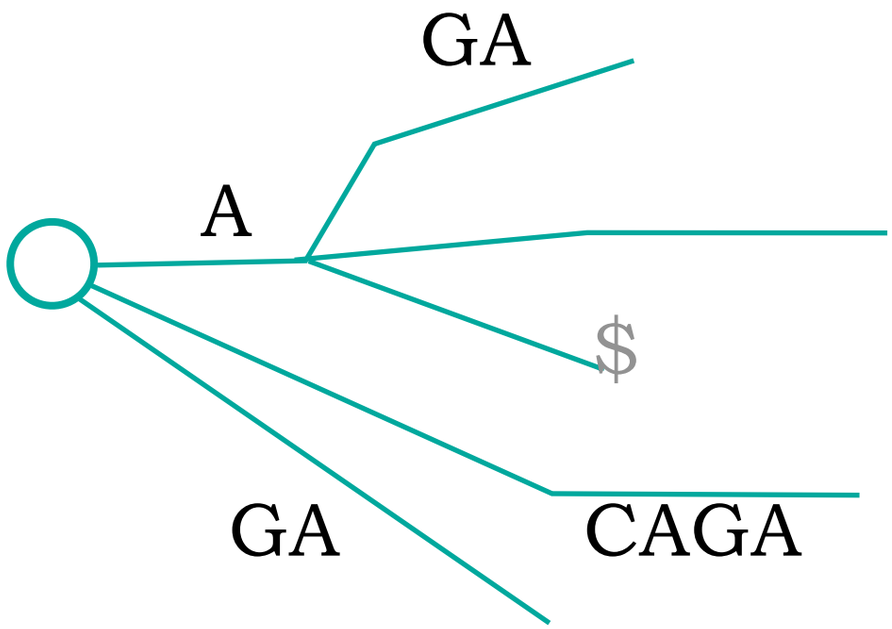
Подстрока исходной строки
есть префикс какого-то
суффикса исходной строки



ACAGA

Mapping

Подстрока исходной строки
есть префикс какого-то
суффикса исходной строки



ACAGA$

Поиск по суффиксному дереву потребует
максимум O(M) времени, какой бы
ни был длины текст

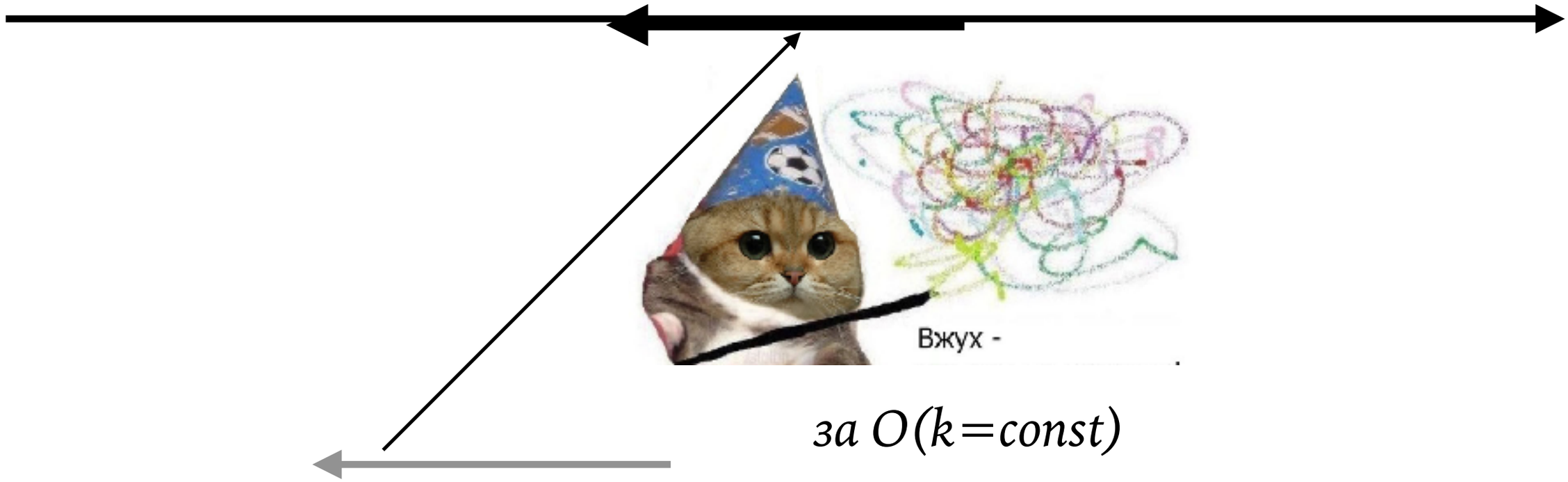
Mapping

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!

Mapping

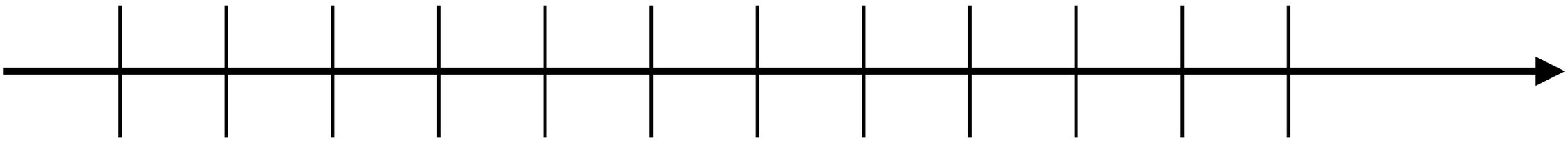
А что если так?



за $O(k=const)$

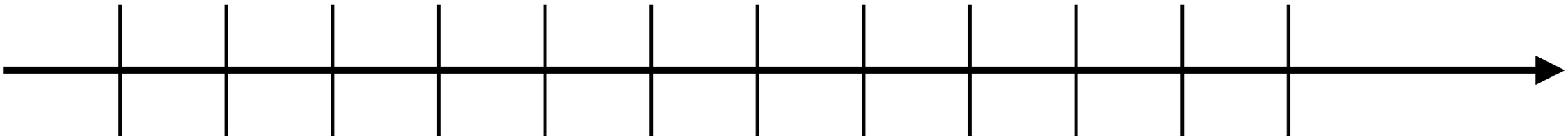
Mapping

Разберем геном на kmer-ы

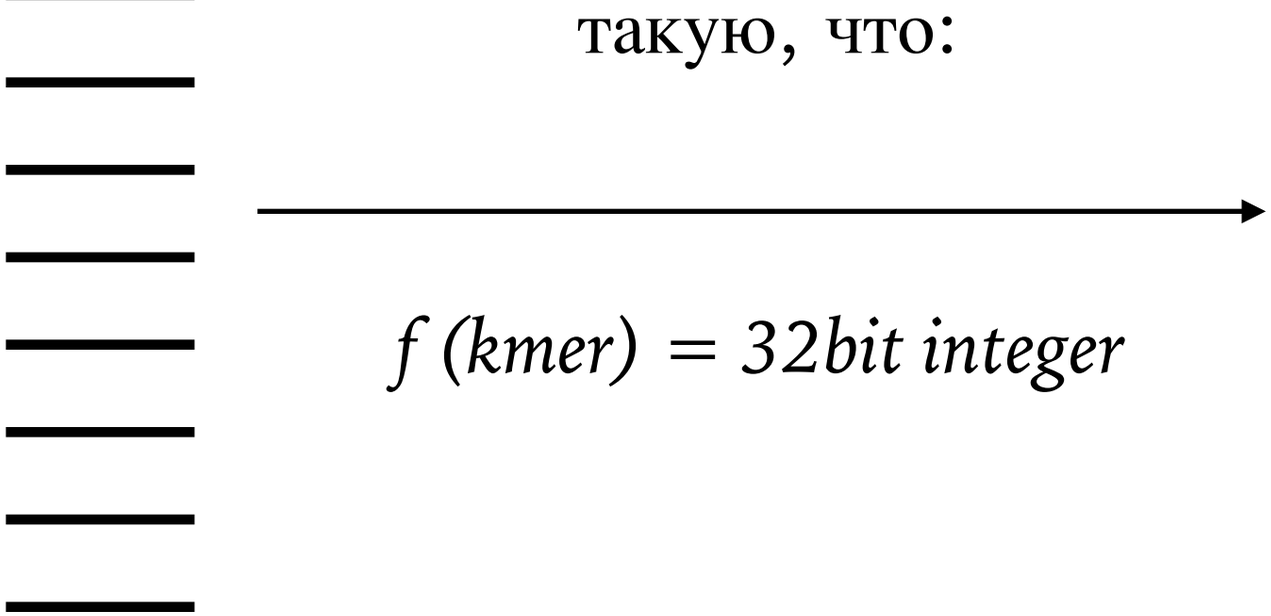


Mapping

Разберем геном на kmer-ы



зададим функцию отображения,
такую, что:



Mapping

56266211

45623467

12375478

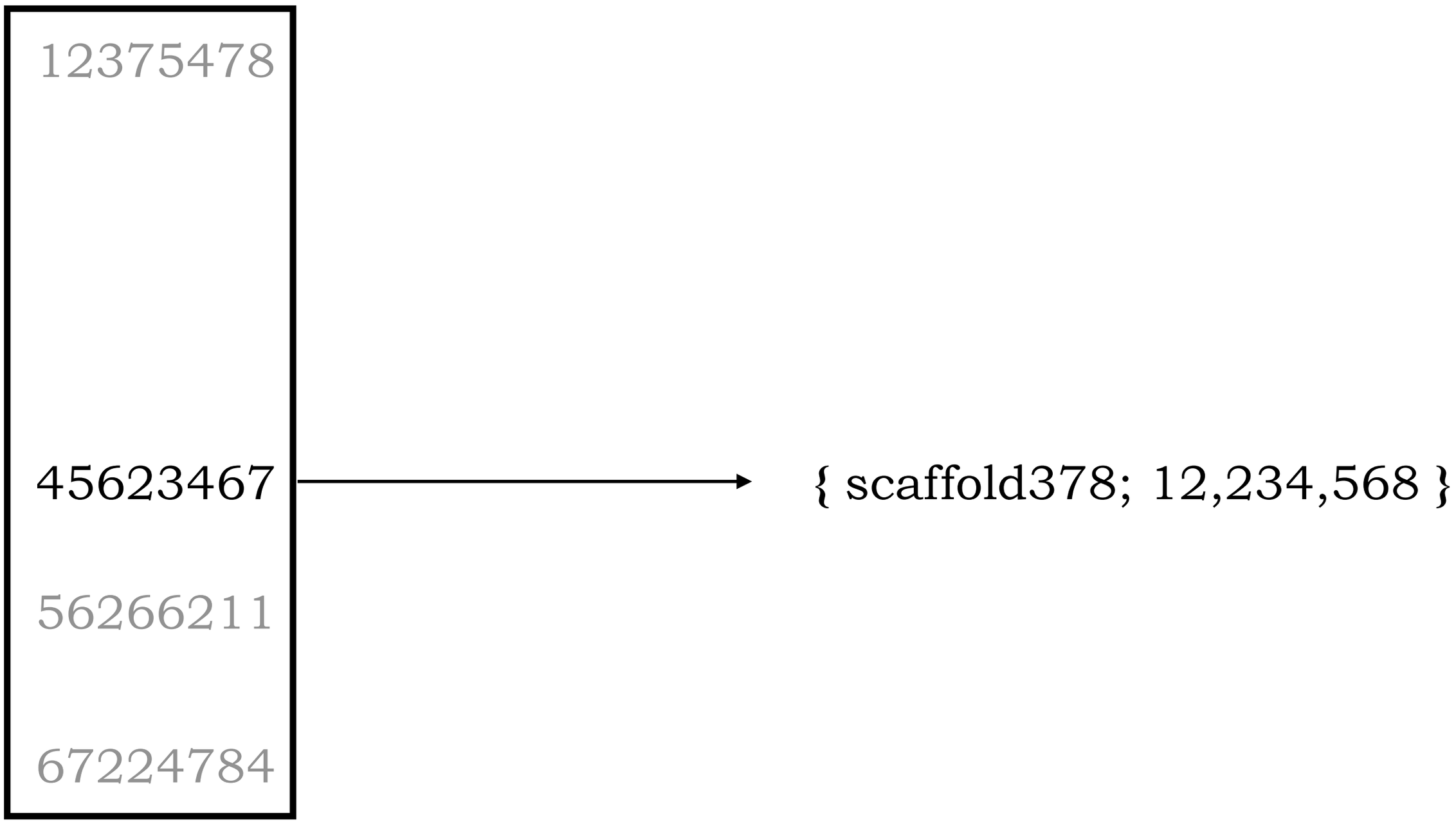
67224784

А теперь вычислить f(read)

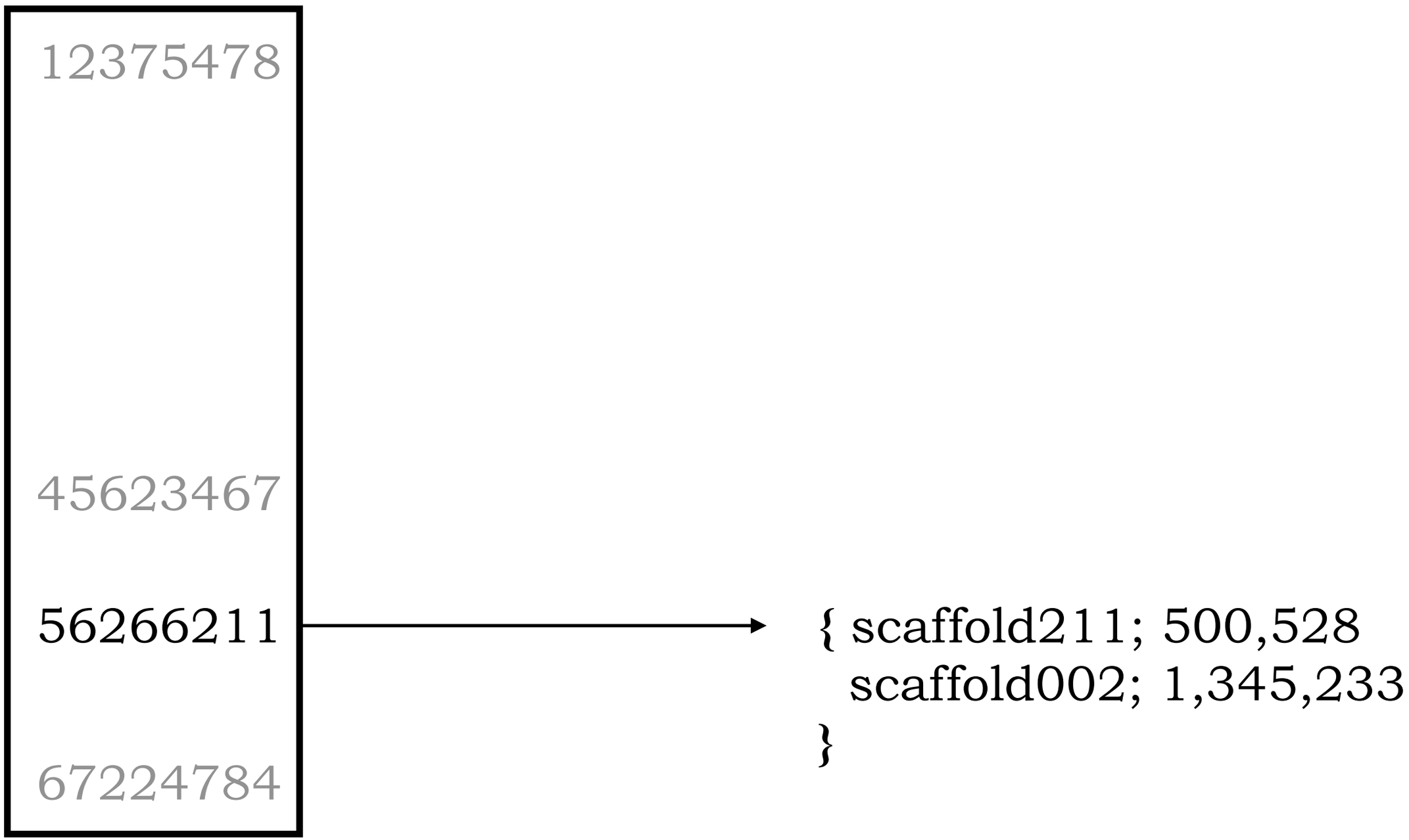
56266211

$$f\,(kmer) = 32bit\ integer$$

Mapping


12375478
45623467
56266211
67224784
{ scaffold378; 12,234,568 }

Mapping


12375478
45623467
56266211
67224784
{ scaffold211; 500,528
scaffold002; 1,345,233
}