Применение Hi-C к метагеномике

Бзикадзе Александр Важевич
группа 444

Научный руководитель к.т.н., доц. Литвинов Ю.В.
Научный консультант к.ф.-м.н., доц. Коробейников А.И.

СПбГУ

24 мая 2019 г.

Hi-C протокол

- Метод, позволяющий восстанавливать пространственное строение ДНК
- Результат работы – библиотека парных ридов, т.е. последовательности нуклеотидов определенной длины, составленные из двух различных участков генома

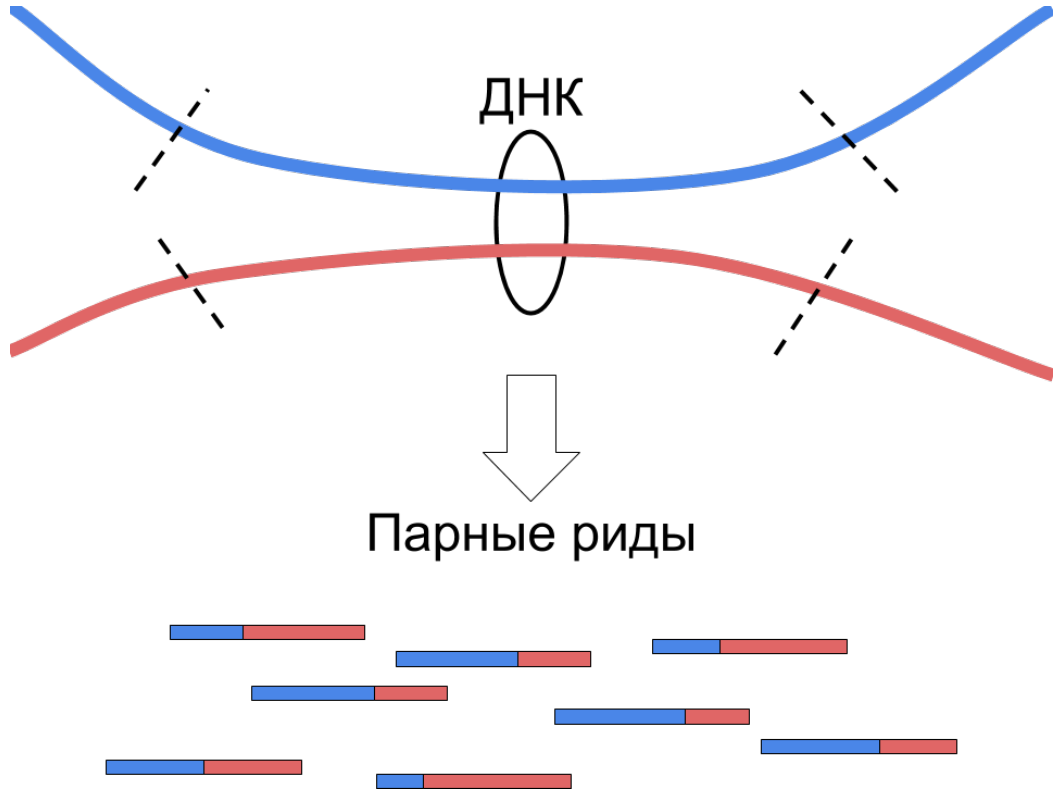


Рис.: Hi-C протокол

Граф Сборки

- Входные риды (к-меры) представляют ребра в мультиграфе
- Исходный геном – один из Эйлеровых путей
- Упрощенный граф называется графом сборки

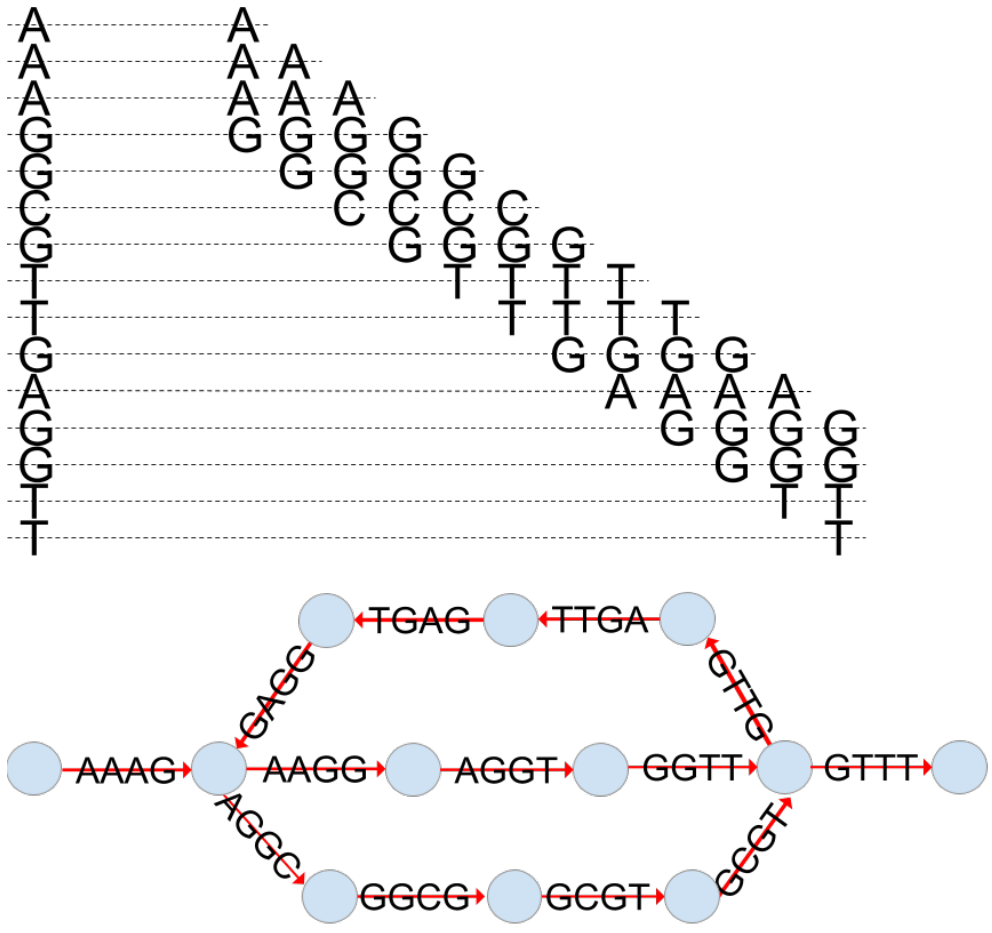


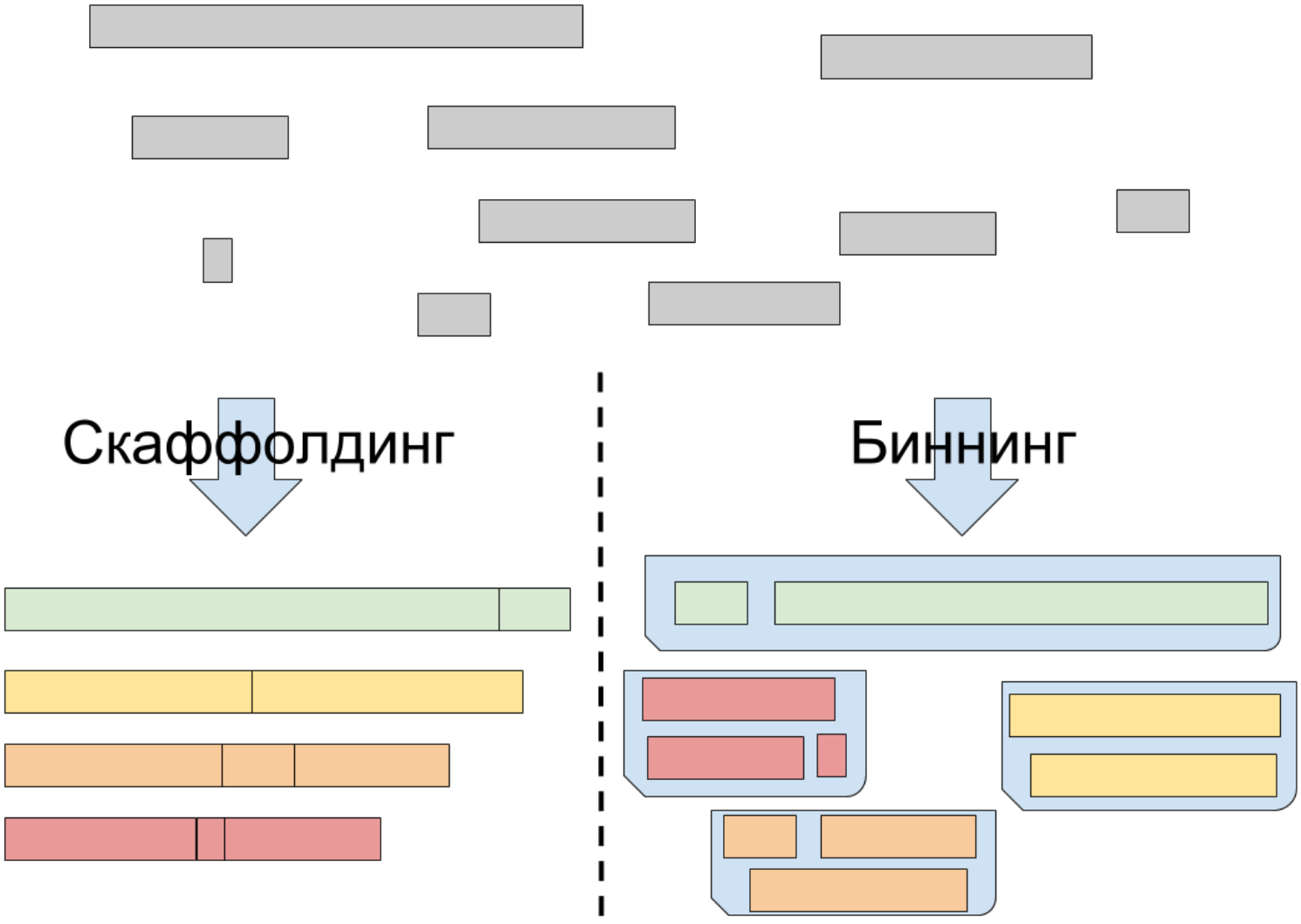
Рис.: Граф Де Брюина (эйлеров)

SPAdes

- Геномный ассемблер
- Разработан в Центре Алгоритмической Биотехнологии, СПбГУ
- Большой инструмент, написанный на C++ и Python

Рис.: SPAdes logo

Задачи метагеномики



Цели работы

Расширение геномного ассемблера SPAdes поддержкой Hi-C данных

- Понимание работы Hi-C протокола
- Исследование свойств данных и выявление типичных проблем
- Исследование возможности применения Hi-C к скаффолдингу
- Прототип решения
- Разработка расширения на основе SPAdes

Скаффолдинг: SALSA2

- Исправление графа сборки с помощью Hi-C, построение графа специального вида
- Генерация скаффолдов поиском максимального взвешенного паросочетания в гибридном графе
- Анализ полученных соединений с помощью Hi-C
- Продолжение, пока большая часть соединений ”корректная”

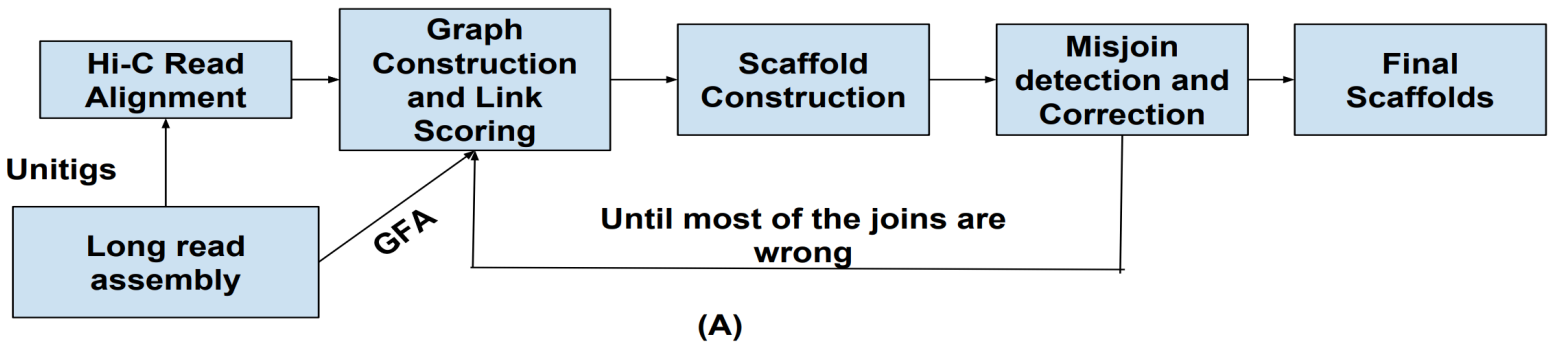


Рис.: Integrating Hi-C links with assembly graphs for chromosome-scale assembly, Jay Ghurye et al.

Прототип скаффолдера: сжатие графа

- Поиск Hi-C связей на коротких ребрах затруднен
- По биологическим причинам Hi-C связей на коротких ребрах может не быть
- Короткие ребра графа сжимаются, порождая сжатый граф
- Короткими считаются ребра короче 500bp

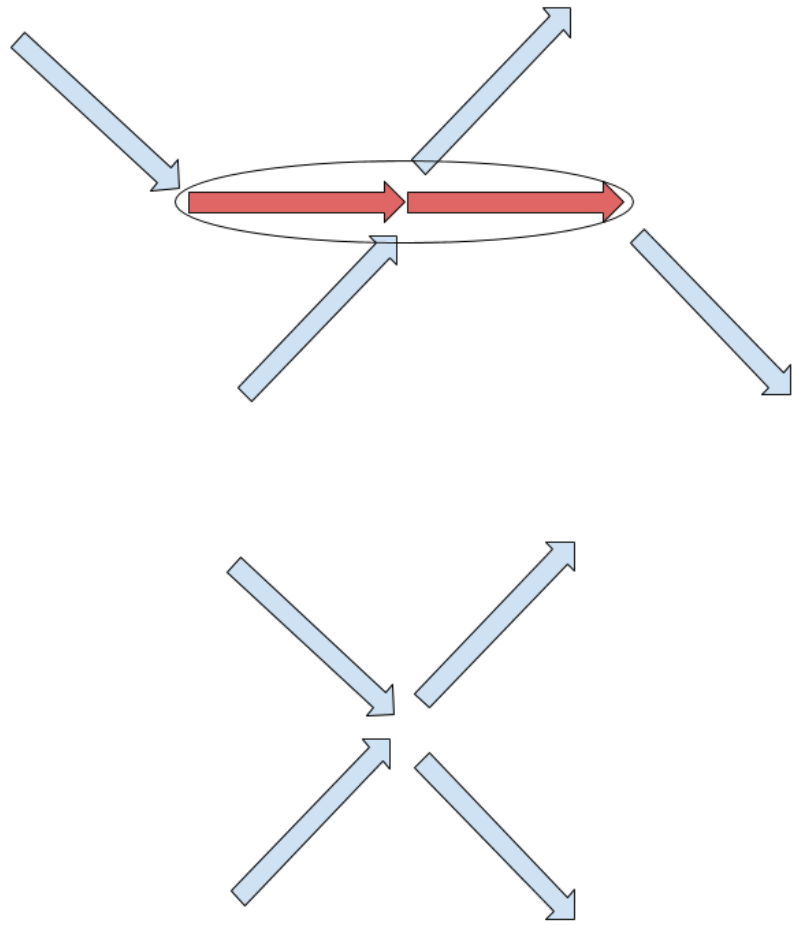


Рис.: Сжатие графа

Прототип скаффолдера

- Возьмем последнее ребро (i1) текущего пути
- Рассмотрим Hi-C связи с последних 4Kbp
- Выбрать (o1) – максимальное подтвержденное Hi-C продолжение i1 среди o1, o2, o3

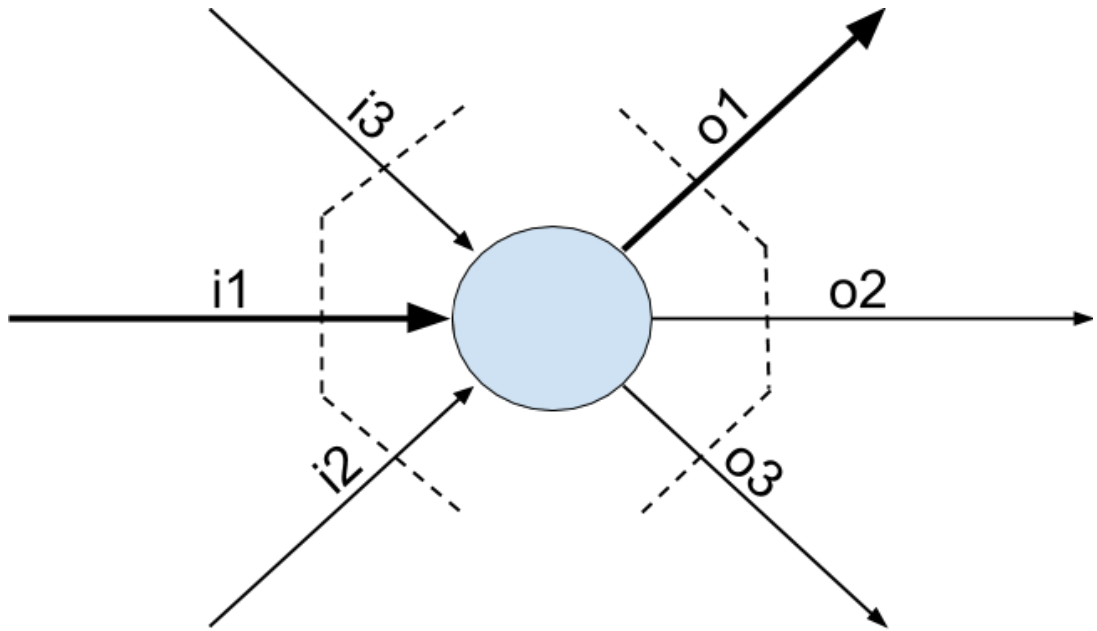


Рис.: Разрешение конфликта

Прототип скаффолдера: продолжение

- Абсолютный критерий (>= 2 Hi-C связи)
- Относительный критерий (в два раза больше связей чем на другие)
- Обратный критерий (i1 – максимум по Hi-C связям для o1 среди i1, i2, i3)

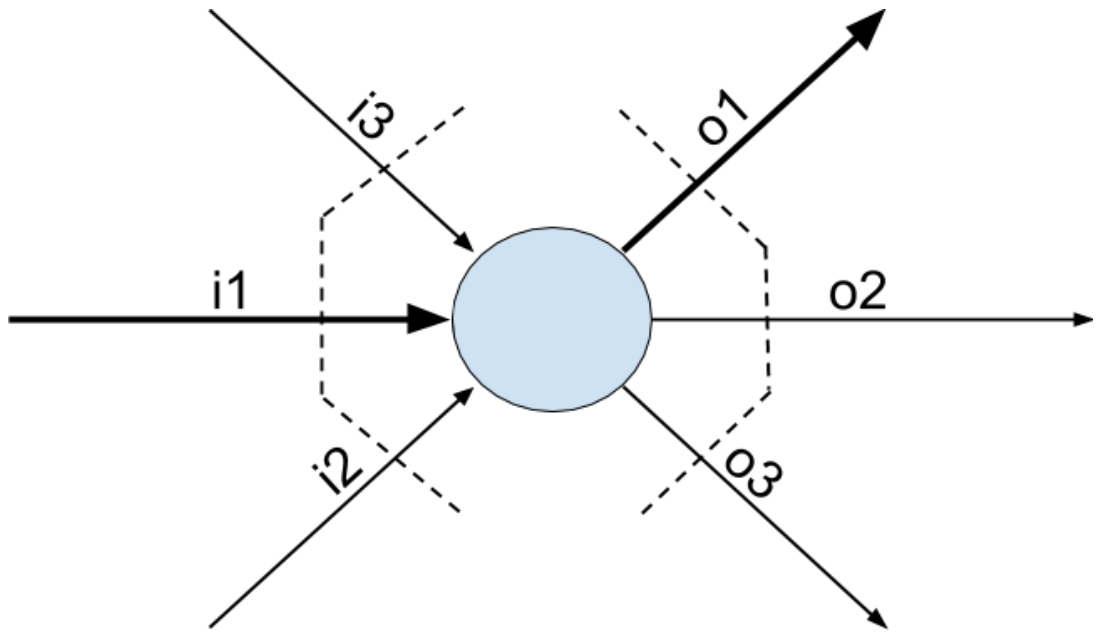


Рис.: Разрешение конфликта

Прототип скаффолдера: результаты

- Симулированные данные:
  - Продолжение выбрано в 50% случаев
  - Продолжение выбрано верно в 97% случаев
- Реальные данные:
  - Продолжение выбрано в 18% случаев
  - Продолжение выбрано верно в 90% случаев

SPAdes

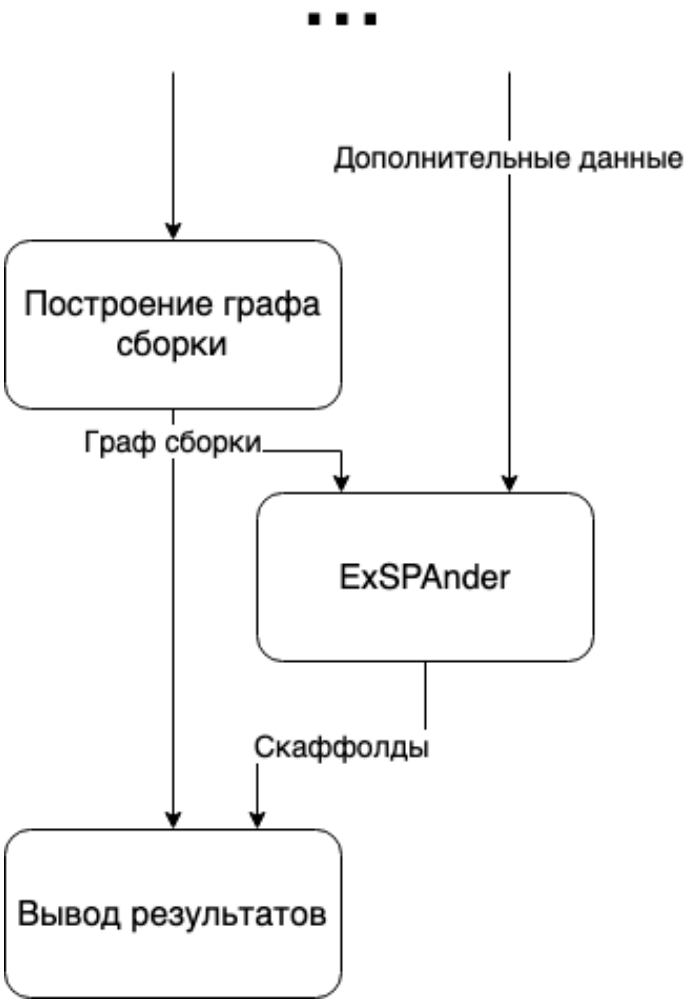


Рис.: Пайплайн SPAdes

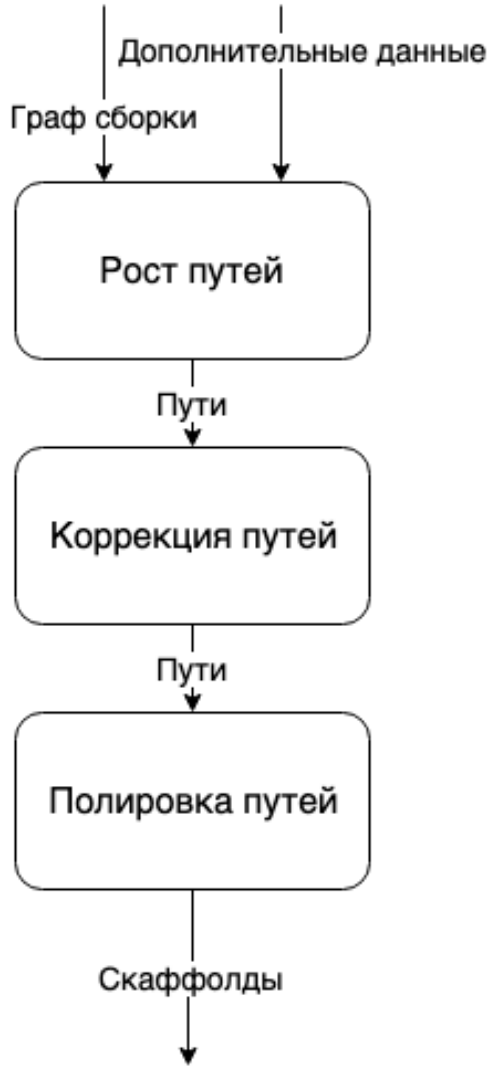


Рис.: ExSPAnder

HiCPathToPathExtensionChooser

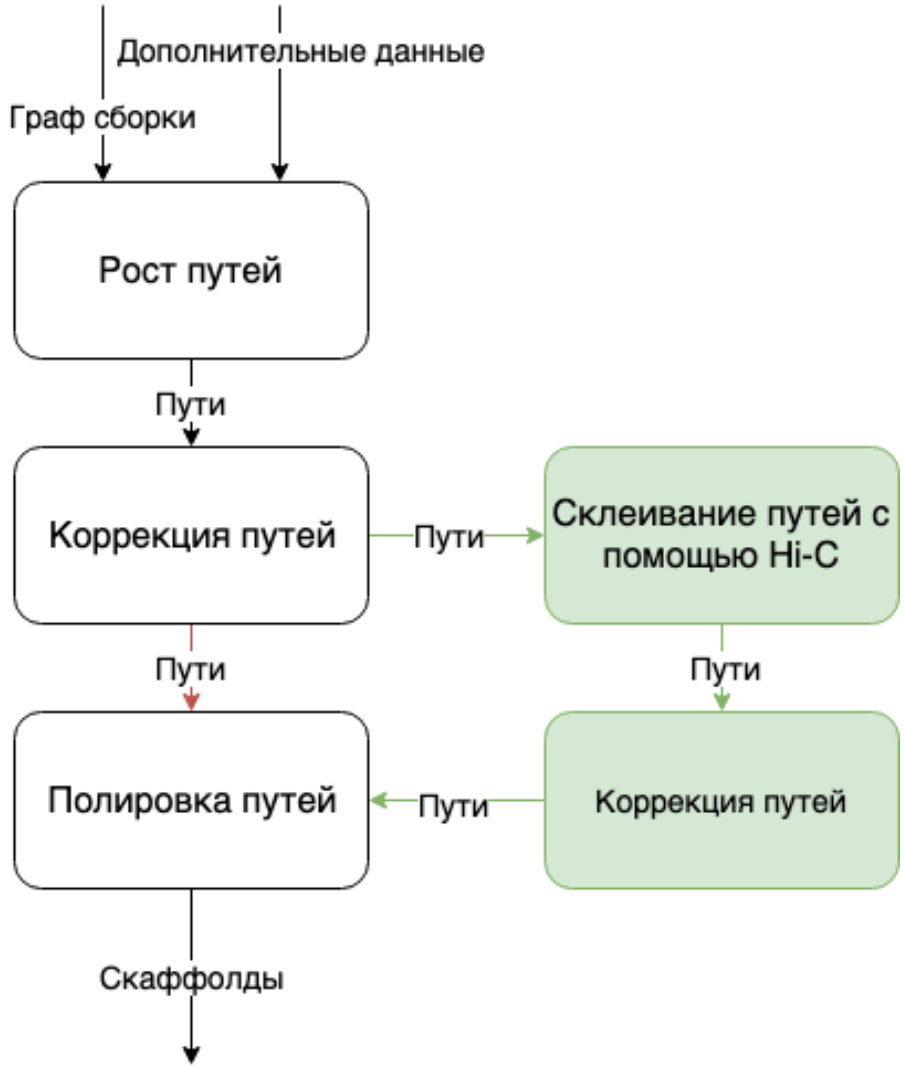


Рис.: Добавление склеивания путей

HiCPathToPathExtensionChooser: алгоритм

- Кандидаты-пути
- Кандидаты фильтруются по большим пересечениям с продолжаемым путем
- По последним 5000bp путей выбирается максимум суммированием Hi-C связей между концами обоих путей
- Запускаются абсолютный и относительный критерии
- Если отношение длин выбранного пути и текущего пути превышает 0.8, запускается обратный критерий

HiCPathToPathExtensionChooser: симулированные данные

- Сборка первого референса улучшилась от двух длинных скаффолдов до одного, но с инверсией в середине длиной 2Kbp при общей длине более 3Mbp
- Сборка второго референса проходит с мизассемблами (ошибками сборки) с особенностями симулиции данных
- Сборка третьего референса улучшилось от 8 длинных скаффолдов до одного
- Четвертый ранее уже собирался в один скаффолд, но на более поздней стадии

HiCPathToPathExtensionChooser: реальные данные

- Образец данных взят из *Epichloe festucae Fl1*, суммарная длина референсного генома – 35Mbp.
- Умеренное положительное влияние на качество сборки
- Добавление ошибок сборки
- Основной тип ошибок – инверсия

HiCPathToPathExtensionChooser: значения метрик на реальных данных

Метрика	*после Hi-C*	*до Hi-C*
Genome fraction	70.784	70.642
Duplication ratio	1.04	1.005
NG50	89507	44857
NGA50	52397	43968
Количество ошибочных скаффолдов	78	1
Количество скаффолдов со значительными ошибками	18	1

Результаты

- Изучено физическое устройство протокола Hi-C
- Проведен анализ свойств Hi-C данных
- Разработан прототип решения задачи скаффолдинга
- Разработано расширение геномного ассемблера SPAdes
- Проведено тестирование расширения
  - На симулированных данных
  - На реальных данных

Приложение: метрики

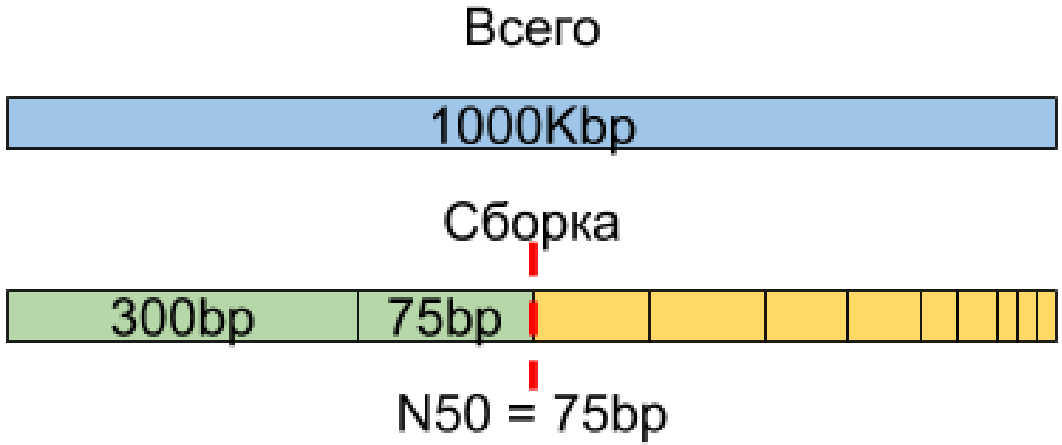


Рис.: Метрика N50

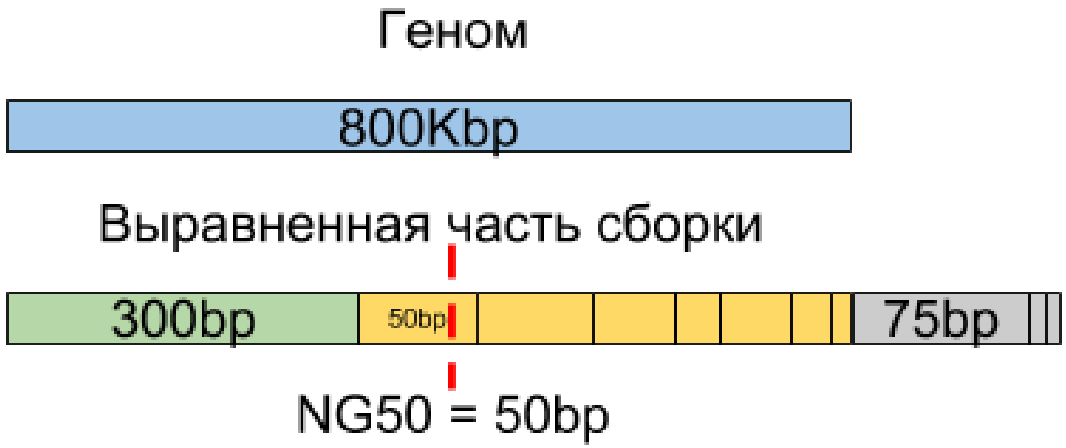


Рис.: Метрика NG50

Приложение: метрики

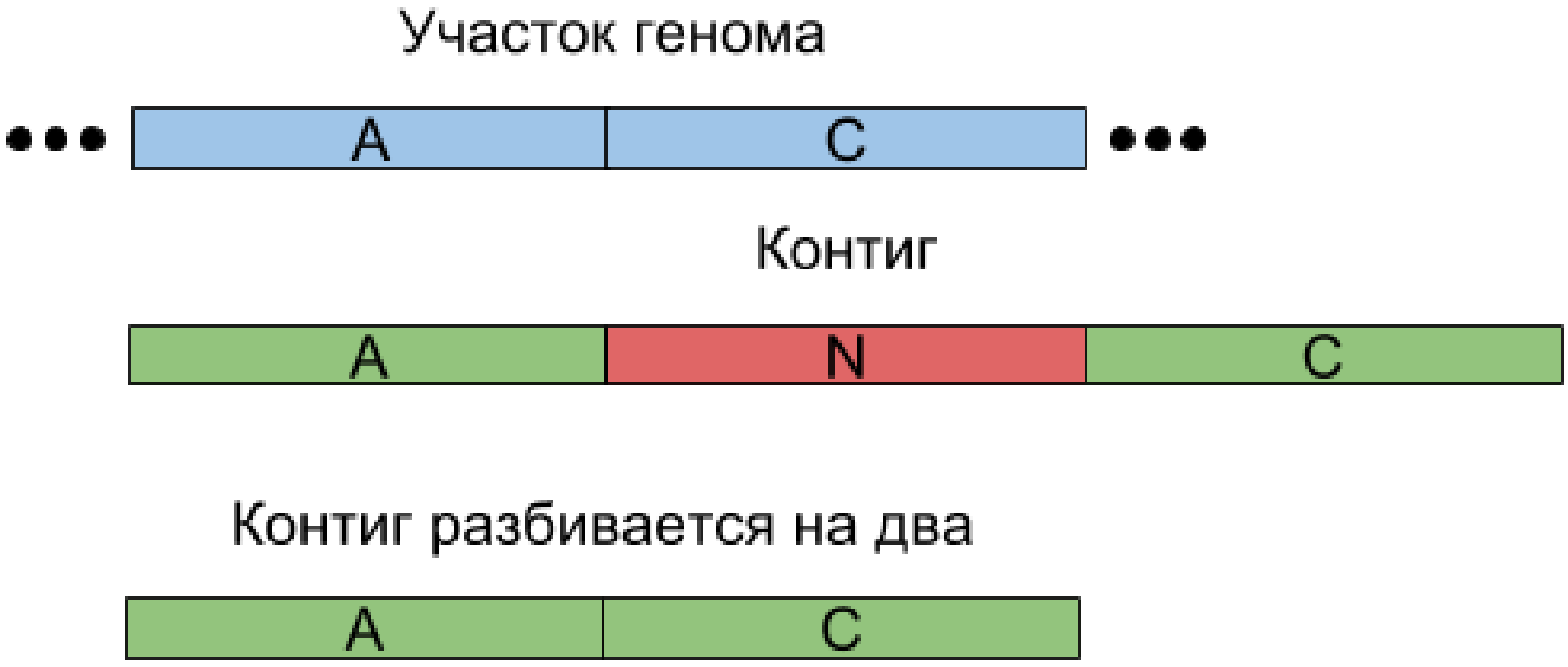


Рис.: Метрика NGA50

Приложение: NG50, симулированные данные

Референсный геном (длина)	с Hi-C	без Hi-C
GCF_001645615 3053012bp	3068998	1847785
GCF_001678755 3176352bp	1930487	494002
GCF_001888565 3312306bp	3268481	634402
GCF_900064425 2938933bp	2923079	2219132

Приложение: NGA50, симулированные данные

Референсный геном (длина)	с Hi-C	без Hi-C
GCF_001645615 3053012bp	1848153	1847785
GCF_001678755 3176352bp	493859	352179
GCF_001888565 3312306bp	3267605	784850
GCF_900064425 2938933bp	2923079	2923128

Приложение: NGA75, симулированные данные

Референсный геном (длина)	с Hi-C	без Hi-C
GCF_001645615 3053012bp	1205270	955853
GCF_001678755 3176352bp	352393	322586
GCF_001888565 3312306bp	3267605	537853
GCF_900064425 2938933bp	2923079	2923128

Приложение: Hi-C протокол

- Соединить ДНК
- Нарезать ДНК рестриктазами
- Заполнить концы, помечая их биотином
- Соединить концы соединенных отрезков
- Очистить и притянуть соединенные отрезки за биотин
- Подготовить соединения для парного секвенирования

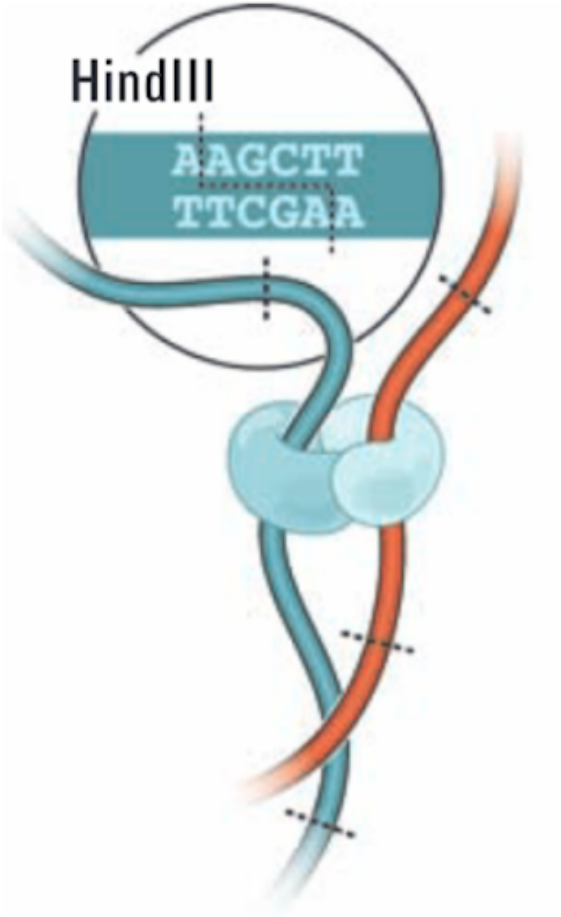


Рис.: Hi-C протокол

Приложение: Hi-C протокол

- Соединить ДНК
- Нарезать ДНК рестриктазами
- Заполнить концы, помечая их биотином
- Соединить концы соединенных отрезков
- Очистить и притянуть соединенные отрезки за биотин
- Подготовить соединения для парного секвенирования

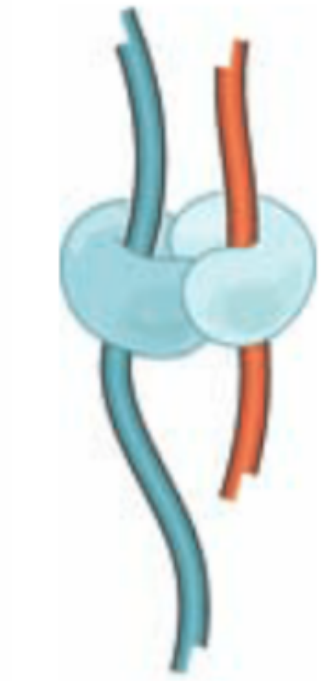


Рис.: Hi-C протокол

Приложение: Hi-C протокол

- Соединить ДНК
- Нарезать ДНК рестриктазами
- Заполнить концы, помечая их биотином
- Соединить концы соединенных отрезков
- Очистить и притянуть соединенные отрезки за биотин
- Подготовить соединения для парного секвенирования

Рис.: Hi-C протокол

Приложение: Hi-C протокол

- Соединить ДНК
- Нарезать ДНК рестриктазами
- Заполнить концы, помечая их биотином
- Соединить концы соединенных отрезков
- Очистить и притянуть соединенные отрезки за биотин
- Подготовить соединения для парного секвенирования

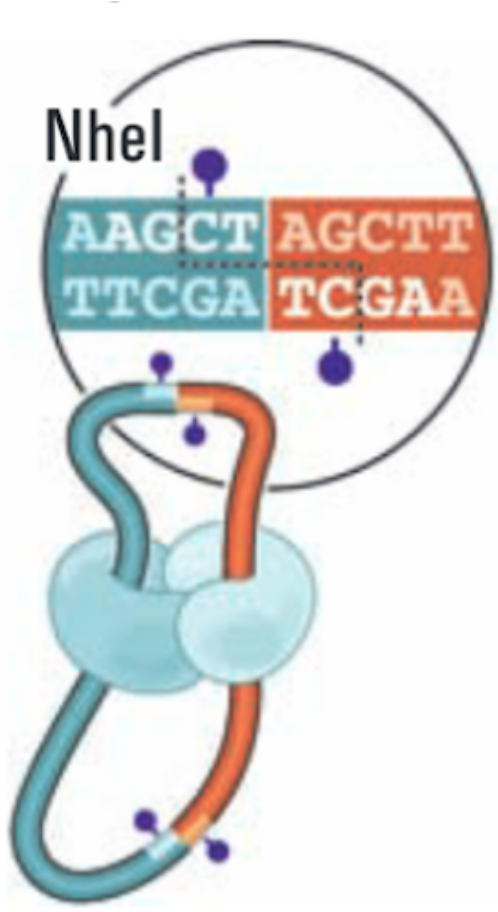


Рис.: Hi-C протокол

Приложение: Hi-C протокол

- Соединить ДНК
- Нарезать ДНК рестриктазами
- Заполнить концы, помечая их биотином
- Соединить концы соединенных отрезков
- Очистить и притянуть соединенные отрезки за биотин
- Подготовить соединения для парного секвенирования

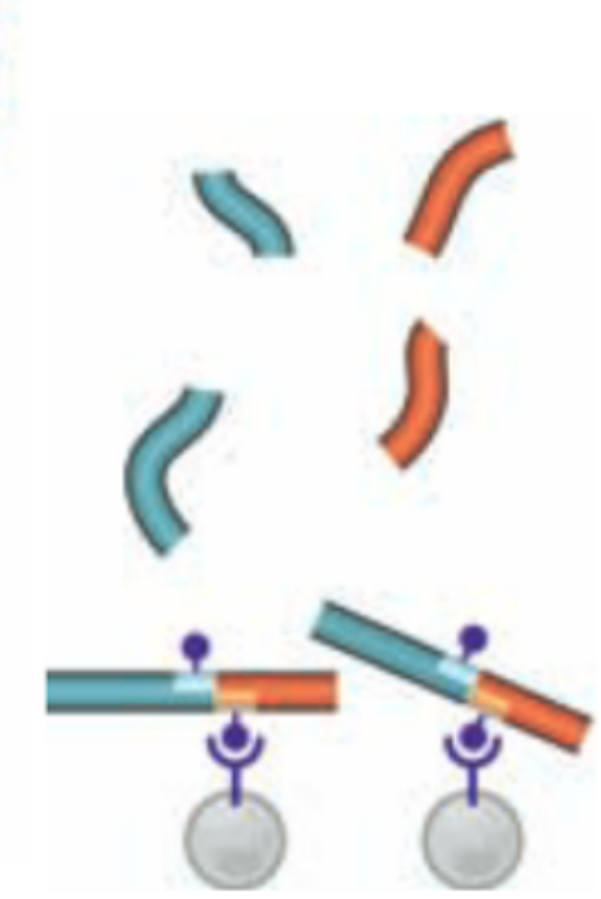


Рис.: Hi-C протокол

Приложение: Hi-C протокол

- Соединить ДНК
- Нарезать ДНК рестриктазами
- Заполнить концы, помечая их биотином
- Соединить концы соединенных отрезков
- Очистить и притянуть соединенные отрезки за биотин
- Подготовить соединения для парного секвенирования

Приложение: анализ Hi-C протокола

- Соединения опираются на пространственную близость участков ДНК
- Результат зависит от выбора рестриктазы – фермента, расщепляющего заданную последовательность нуклеотидов
- Возможны соединения между хромосомами одной клетки
- Редко возникают соединения между ДНК разных клеткок
- Присутствует значительная по количеству доля парных ридов с незначительным расстоянием вставки

Стандартной рекомендацией является использование нескольких Hi-C библиотек с различными рестриктазами

Приложение: анализ данных

- Парные риды с малым расстоянием вставки присутствуют в большом количестве
- Присутствуют соединения между разными организмами

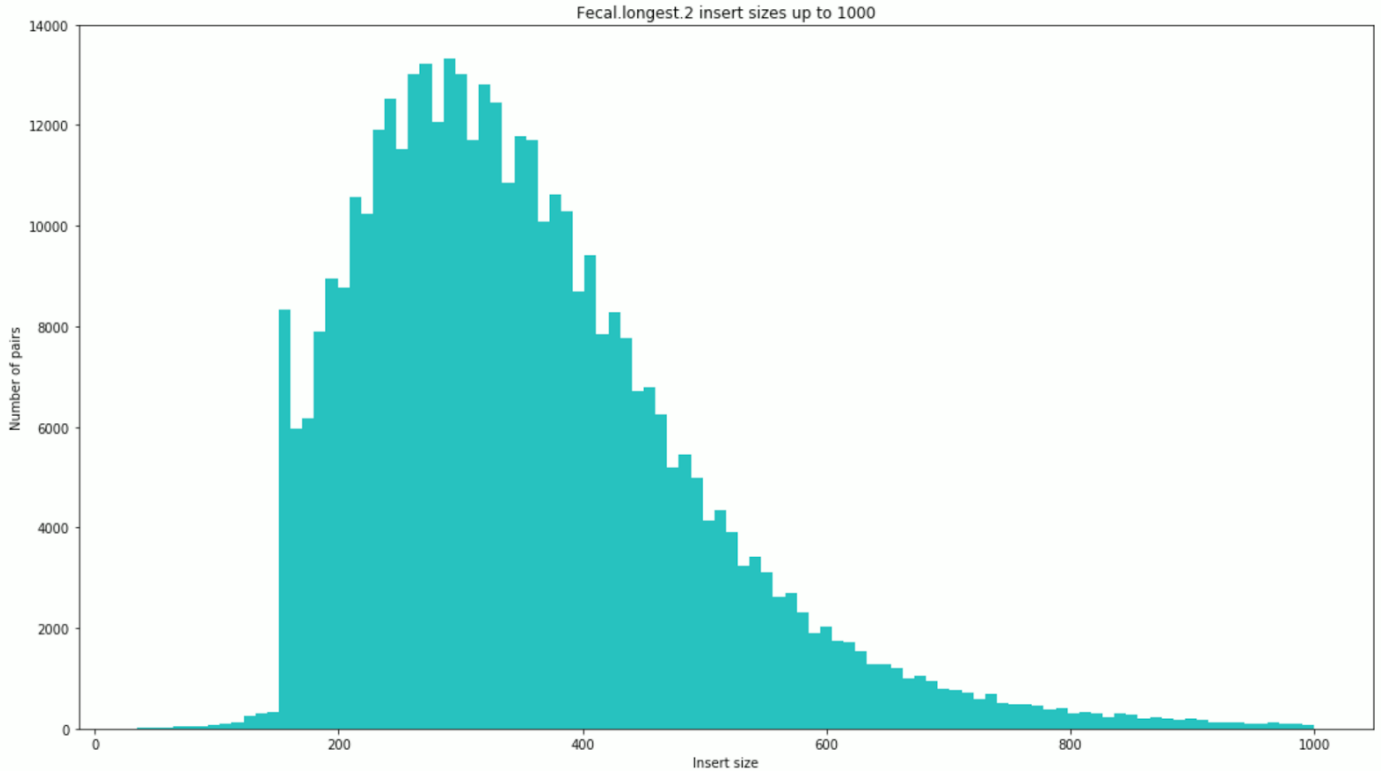


Рис.: 96% пар