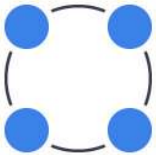
Estimate Library Complexity Optimization



ИНСТИТУТ
БИОИНФОРМАТИКИ

Студент: Свичкарев Анатолий

Научный руководитель:
Льянов Заал, EPAM Systems

Picard // аналог SamTools

Broad Institute

Cambridge, MA

broadinstitute / picard

```
java -jar picard.jar -h
```

Estimate Library Complexity

– утилита Picard для оценки числа уникальных ДНК прочтений в изучаемой библиотеке парных ридов.

Прочтения – дубликаты, если:

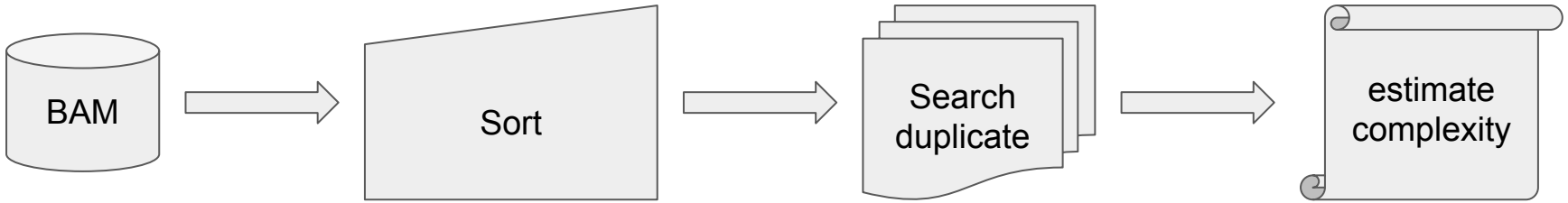
- они совпадают без пропусков
- общий уровень несовпадений < порога (0.03 по умолчанию).

```
java -jar picard.jar EstimateLibraryComplexity \
     I=input.bam \
     O=est_lib_complex_metrics.txt
```

Задача

Исследовать возможность оптимизации метрики EstimateLibraryComplexity из Picard Tools и реализовать найденные подходы.

Существующий алгоритм ELC



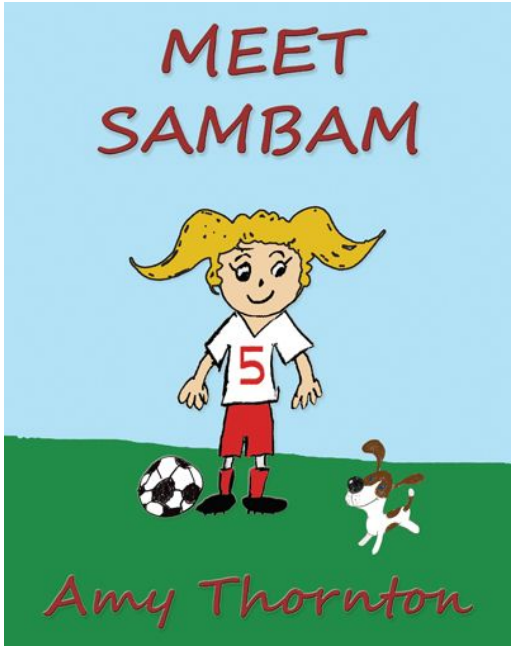
≈10-100 Gb

Слиянием с использованием ПЗУ

В группах, $O(n^2)$

Используя информацию о дубликатах

Текущие результаты



1. Прошёл инструктаж в предметной области:
   выравнивание ридов на референс,
   структура SAM/BAM форматов;
2. Разобрал алгоритм ELC
   и его реализацию в Picard на Java;
3. Обсуждение этапов оптимизации;
4. Реализация одного из пунктов плана оптимизации
   (снижение затрат на определение принадлежности к группе).

Цель

Оптимизировать производительность работы метрики ELC

План

1. Разработка способа замера производительности
2. Выявление стадий оптимизации и их описание
3. Повышение производительности программы:
   3.1. Распараллелить поиск дубликатов в группах
   3.2. Сортировка ридов с фоновой записью
4. Сравнительный анализ численных экспериментов