БОЛЬШИЕ ВЫЗОВЫ

ВСЕРОССИЙСКИЙ КОНКУРС
НАУЧНО-ТЕХНОЛОГИЧЕСКИХ ПРОЕКТОВ

МГК 2021

Региональный трек
Всероссийского конкурса
научно-технологических проектов

«БОЛЬШИЕ ВЫЗОВЫ»

направление

Генетика, персонализированная и прогностическая медицина

название работы

Выравнивание результатов секвенирования на геном вируса Ebola

участник(и)

Солдатенкова Ольга Дмитриевна

#большиевызовы
#мгк

mgk.olimpiada.ru

г. Москва
2021

Актуальность и научная новизна

Рис. 1. ВОЗ: Число умерших от лихорадки Эбола превысило 5000 человек.
URL: https://www.dw.com/ru /a-18059522

Актуальность: В связи с обострением эпидемиологической ситуацией в мире становится крайне актуальной задача быстрой идентификации природы патогенов. Современным быстрым и точным методом диагностики является секвенирование с выравниванием на референсный геном.

Научная новизна: В работе проводится выравнивание результатов секвенирования РНК, содержащейся в плазме крови человека на геном вируса Ebola. Используются различные профессиональные биоинформатические инструменты и проводится сравнительный анализ результатов их работы. Написана программа на языке Python, позволяющая в компактной форме собирать необходимую статистику по результатам выравнивания для парных прочтений.

Зависимость стоимости секвенирования генома человека от времени

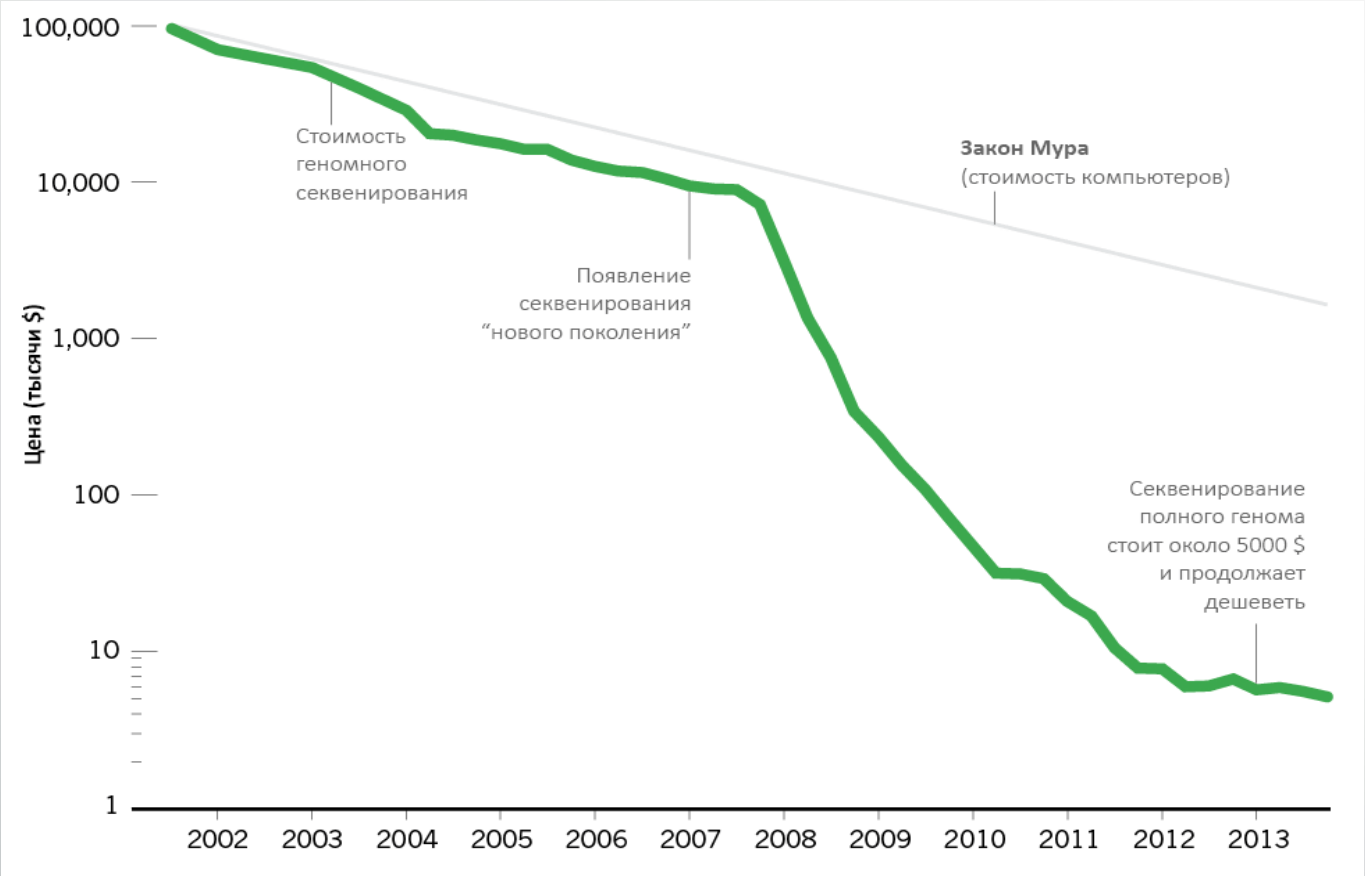

Рис. 2. Стоимость секвенирования генома человека. URL: https://biomolecula.ru/articles/tekhnologiia-1000-za-genom#source-2

Секвенирование - определение последовательности нуклеотидов в ДНК.

Прибор, на котором проводится секвенирование – **секвенатор**.

В 2001 году секвенирование генома человека стоило около 100 млн.$ и выполнялось в основном методом Сэнгера. Сегодня секвенирование полного генома человека стоит около 1000$.

Стало возможным использовать секвенирование для диагностики заболеваний.

Секвенаторы второго поколения

Рис 3. Секвенатор Illumina. URL: https://www.medicalexpo.ru/prod/illumina-inc/product-83632-604911.html (дата обращения: 06.12.2020).

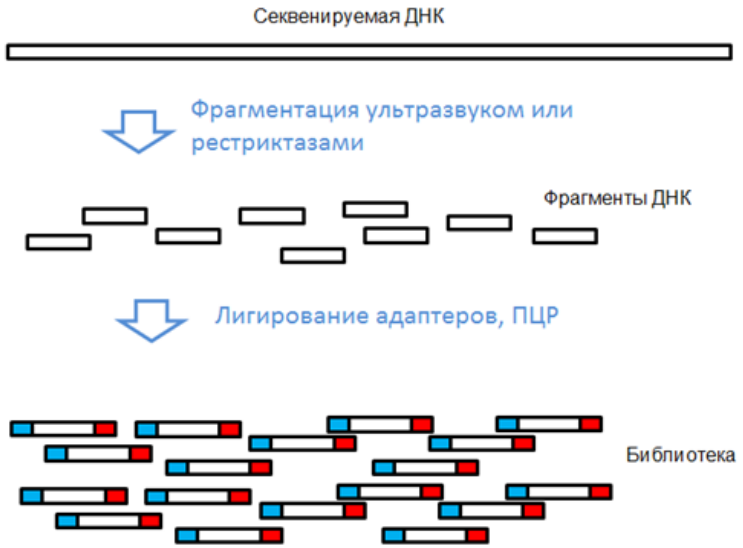

Рис. 4. Получение библиотеки для секвенирования

Вслед за появлением метода секвенирования Сэнгера, появилось множество более продуктивных методов. Их общее название: методы секвенирования нового поколения, секвенирование нового поколения [1]. Как правило, в них используется метод коротких прочтений, называемых **ридами**. Это метод, при котором, общая секвенируемая ДНК разрезается на множество небольших фрагментов. Далее проводится амплификация при помощи полимеразной цепной реакции (ПЦР) и собственно секвенирование.

Выравнивание результатов секвенирования на референсный геном

После секвенирования данные представлены в виде **очень большого** количества ридов - 13.4 миллиона в данной работе. Большинство ридов принадлежит геномам других организмов. Только около 0.0006% ридов принадлежат геному Ebola. В связи с этим, возникает задача по выравниванию ридов на референсный геном.

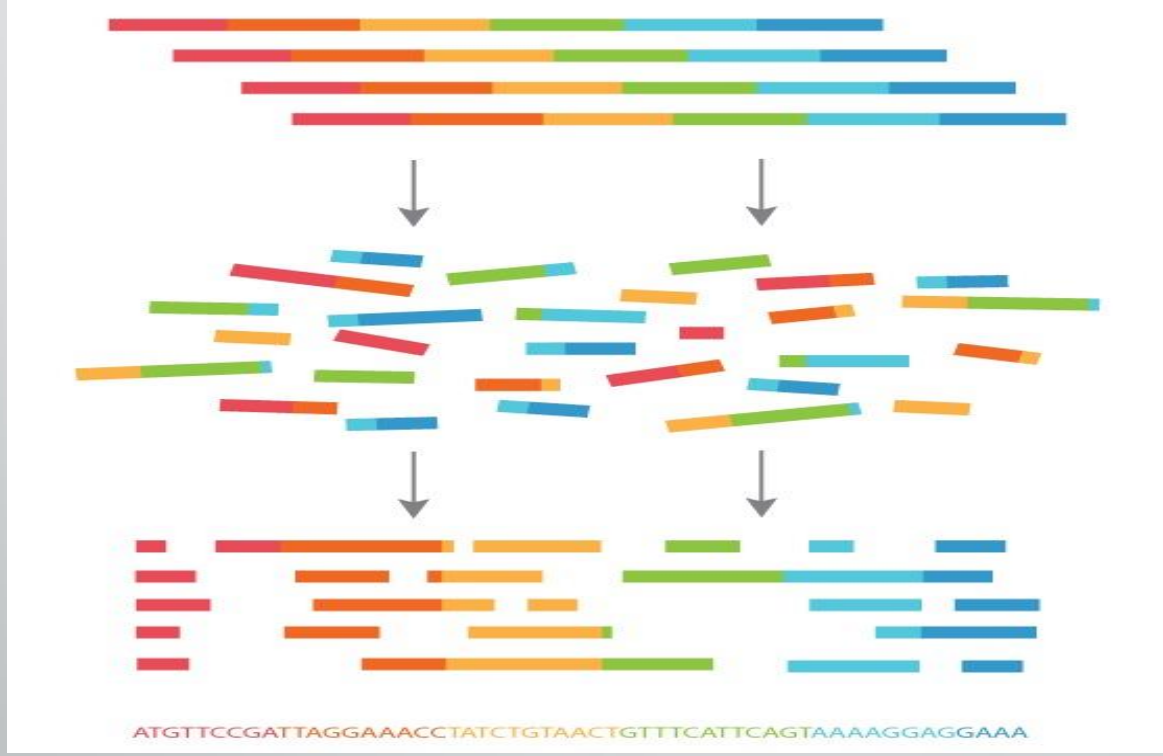

Рис. 5. Выравнивание ридов на геном. URL: https://old.abmgood.com/marketing/knowledge_base/next_generation_sequencing_data_analysis.php

Первая система, созданная для поиска выравнивания - BLAST, она стала важнейшим инструментом исследования для молекулярных биологов и биоинформатиков [2].

Были разработаны специальные форматы для хранения и обработки результатов выравнивания – SAM (Sequence Alignment/Map) и BAM [3], [4].

Характеристики выравнивания результатов секвенирования на геном вируса Ebola различными программами

Проводились одинаковые по содержанию, но разные по форме действия для 4 программ выравнивания: BWA, Bowtie2, Minimap2 и Hisat2. Результаты анализа выравнивания представлены в таблице:

Название программы	Выравненные риды	Выравненные нуклеотиды	Максимальное качество	Среднее качество	Время работы (сек)
BWA	601	86598	60	59.95	291
Bowtie2	590	85127	42	31.59	234
Minimap2	595	85730	60	54.83	68
Hisat2	373	53604	60	60	76

Все четыре исследованные программы имеют в основе алгоритм Борроуза-Вилера, часто со значительными изменениями. Также была попытка использовать программы BLAST, SOAP2 и SnapGene.

Таблица 1. Характеристики выравнивания результатов секвенирования на геном вируса Ebola различными программами.

Сравнение скорости работы программ выравнивания относительно BWA

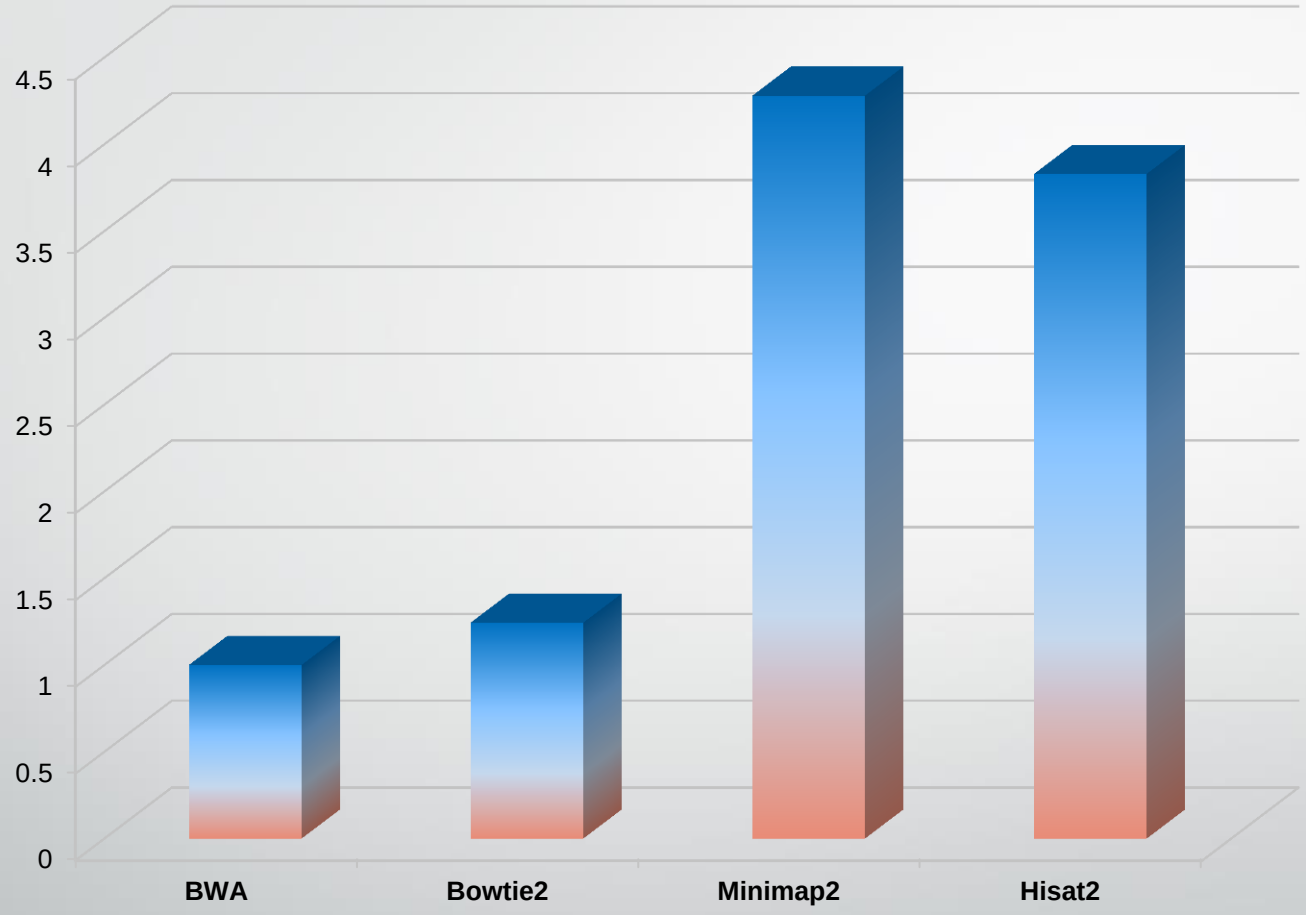

Рис. 6. Сравнение скорости работы программ выравнивания относительно BWA

Можно заметить, что Minimap2 значительно выделяется по этому показателю и превосходит BWA в 4.3 раза. Немного отстает от Minimap2 программа Hisat2. Bowtie2 же значительно медленнее, чем Minimap2.

Сравнение числа выравненных нуклеотидов

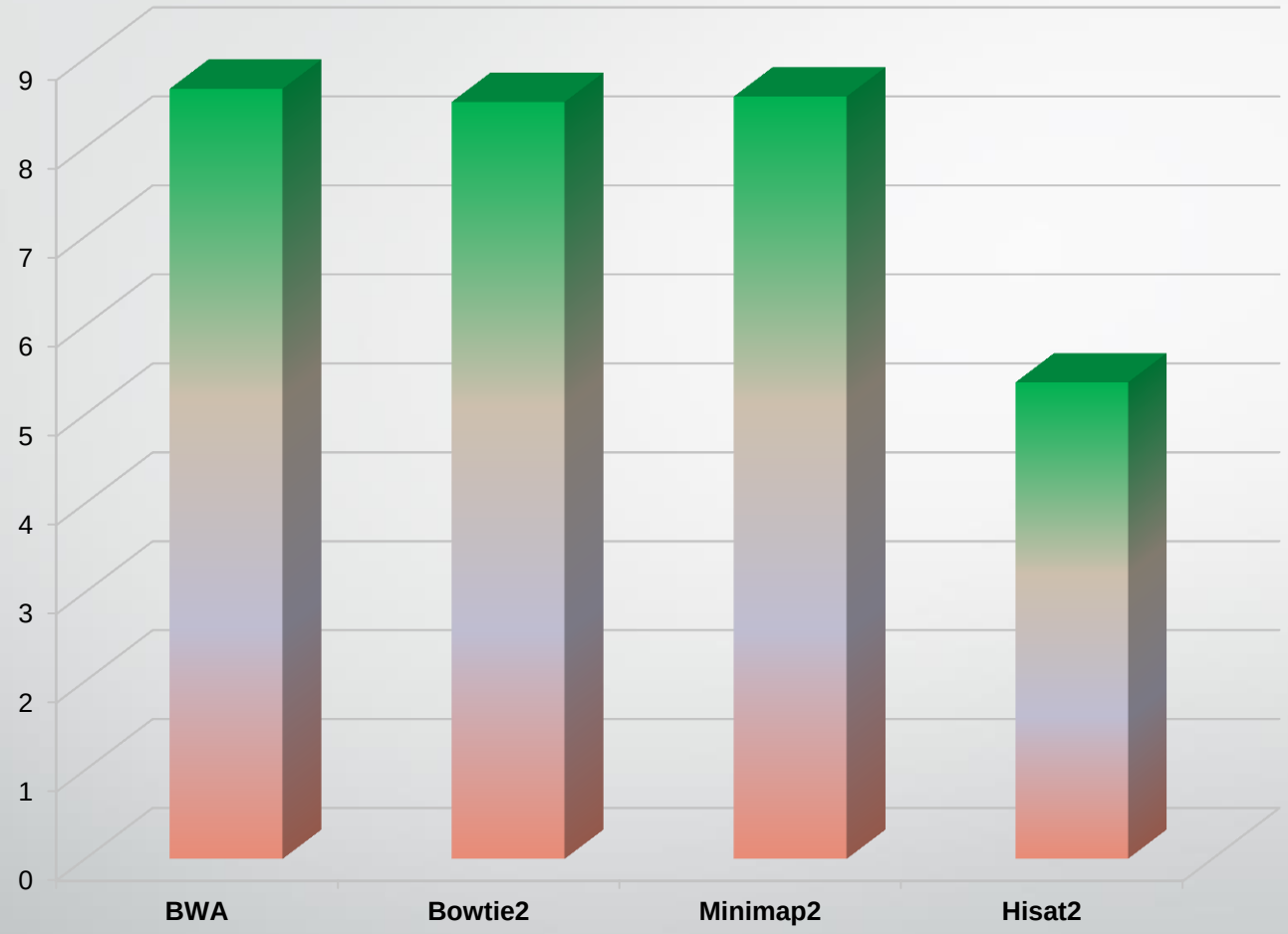

Рис. 6. Сравнение числа выравненных нуклеотидов

BWA, Bowtie2 и Minimap2 выровняли примерно одинаковое количество нуклеотидов, которое с точностью до 3% соответствует таксономическому анализу, проведенному BLAST. Отличие Hisat2 объясняется настройками программы, выставленными «по умолчанию».

Выводы

Из исследованных программ выравнивания наибольшую производительность продемонстрировала **Minimap2** при сохранении очень хорошего качества выравнивания и количества выровненных нуклеотидов. **Hisat2** сильно уступает **Minimap2** по количеству выравненных нуклеотидов и, кроме того, выдает несортированный файл формата **SAM**, который не может быть обработан **SAMtools**. **BWA** и **Bowtie2** сильно уступают **Minimap2** по скорости работы, хотя не уступают по количеству выравненных нуклеотидов.

Исследуемые программы затратили на выравнивание 6.7 миллионов прочтений на геном размером 18871 bp от 1до 5 мин. Учитывая то, что с использованием новых технологий скорость секвенирования постоянно растет, метод диагностики заболеваний при помощи секвенирования с последующим выравниванием на геном становится все более перспективным. Заметим, что благодаря эффективности используемых в программах выравнивания алгоритмов, для решения поставленной в данной работе задачи не потребовалось больших вычислительных ресурсов.

Список литературы

1. Metzker M. L. Sequencing technologies—the next generation //Nature reviews genetics. – 2010. – Т. 11. – №. 1. – С. 31-46.

2. Li H. et al. The sequence alignment/map format and SAMtools //Bioinformatics. – 2009. – Т. 25. – №. 16. – С. 2078-2079.

3. SAM/BAM and related specifications. URL: http://samtools.github.io/hts-specs/ (дата обращения: 22.12.2020).

4. Altschul S. F. et al. Basic local alignment search tool //Journal of molecular biology. – 1990. – Т. 215. – №. 3. – С. 403-410.

Спасибо за внимание