Для авторизованных студентов работает фильтр предложений, основанный на требованиях работодателя. Студент не видит должностей, для которых требуемые работодателем навыки превышают указанные им в личном кабинете, что позволяет существенно упростить процедуру поиска.

Созданная нами страница направлена на то, чтобы не только оптимизировать работу администратора, но и позволить студентам иметь больше возможностей для выбора своей дальнейшей деятельности.

На данный момент сайт функционирует с урезанными возможностями. Перейдя по этой ссылке http://sfpractice.tilda.ws, можно его протестировать.

Литература

1. Интернет-адрес: http://www.jvetrau.com/uxstrategy-3/.
2. Интернет-адрес: http://tilda.education/articles-best-resources-for-web-designers.

ПРОГРАММНАЯ СБОРКА ГЕНОМА ЧЕЛОВЕКА

Д. П. Артанова

ВВЕДЕНИЕ

Сборка секвенированных геномных последовательностей, продолжает быть одной из центральных проблем биоинформатики. Это обусловлено, главным образом, непрекращающимся развитием технологий секвенирования нового поколения, с помощью которых создаются короткие прочтения ДНК – риды. Большие объемы данных и изменения свойств ридов, таких как длина и среднее количество ошибок, вызывают новые трудности для сборки [3]. Для обработки данных необходимы программные пакеты, использующие быстрые и оптимальные алгоритмы.

Выделяют два принципиально различных подхода к сборке геномных последовательностей: с использованием референсного генома (референса) – уже собранного генома данного организма, или организма, родственного исследуемому, и сборка де ново. У каждого из этих подходов свои возможности и ограничения, но при использовании любого из них сборка генома остается сложной вычислительной задачей.

Целью данной работы является изучение наиболее популярных средств для сборки геномов против референсной последовательности, исследование особенностей их работы в процессе сборки ридов генома человека, полученных в результате секвенирования, и разработка оптимального способа сборки генома.

СБОРКА С ВЫРАВНИВАНИЕМ НА РЕФЕРЕНС

Основная идея в сборке геномных последовательностей с выравниванием на референс состоит в том, чтобы найти в референсной последовательности участки, максимально похожие на экспериментально полученные риды, – картировать риды на референс – тем самым определив исходный порядок их следования [7]. В самой простой формулировке эта задача представляет собой задачу поиска подстроки в строке. В связи с тем, что геном человека состоит из миллиардов нуклеотидов, и при его секвенировании методами нового поколения, такими как Illumina, создаются десятки миллионов прочтений, эта задача имеет большую вычислительную сложность.

Процесс сборки генома с выравниванием на референс можно разделить на несколько этапов: индексирование референсной последовательности, картирование ридов, сортировка файлов с координатами ридов, поиск вариаций, сборка.

Индексирование референсной последовательности размечает ее таким образом, чтобы на этапе картирования инструменту, который его осуществляет, потребовалось меньше времени на поиск вхождения очередного рида. Референсная последовательность обычно хранится в формате FASTA – текстовый формат для нуклеотидных последовательностей, в котором азотистые основания обозначаются с помощью буквенных кодов. Самый популярный формат для хранения ридов, полученных в результате секвенирования, – FASTQ – текстовый формат, в котором для каждого нуклеотида приводится оценка качества его прочтения в форме специального кода. После того, как риды выровнены на референс файл с их координатами имеет формат SAM, который позже преобразуется в BAM – бинарный файл, содержащий ту же самую информацию. На этапе поиска вариаций создается файл формата VCF, в котором хранится информация о нуклеотидах, которые различаются в референсной последовательности и ридах, и о возможных причинах такого различия. Для этапа сборки используется только файл с проиндексированной референсной последовательностью, который изменяется на основе информации, содержащейся в файле с вариациями.

Для выравнивания ридов на референсную последовательность были использованы программные пакеты Bowtie 2 [4] и BWA [5]. Обе программы основаны на алгоритме, использующем преобразование Барроуза-Уилера и FM-индекс [1][2].

Вызов программ происходил из эмуляции терминала в графическом интерфейсе ОС Ubuntu 16.04 LTS. В качестве референсной последовательности была использована сборка генома человека hg38 в формате FASTA и два файла с парными ридами в формате FASTQ.

Результаты выравнивания ридов с помощью Bowtie 2 представлены на рис. 1.

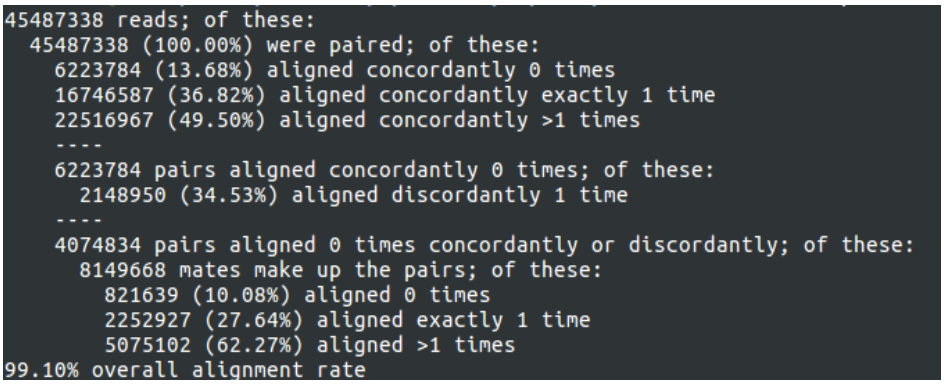

Рис. 1. Результаты картирования ридов с помощью Bowtie 2

В результате выравнивания 36,82 % парных ридов выровнялись совместно один раз, 49,5 % выровнялись совместно несколько раз, причиной чего может быть большое количество повторов, что для генома человека нормально [6]. Из тех ридов, которые не выровнялись совместно (13,68 % от общего количества) 34,53 % выровнялись поодиночке. Процент всех, выровнявшихся ридов 99,1 % – в целом, хороший результат.

BWA не дает развернутой информации о качестве выравнивания ридов после завершения своей работы, как Bowtie 2, и информацию о том, насколько успешно произошло выравнивание можно получить для каждого рида в отдельности из SAM-файла. Выравнивание производилось с помощью алгоритма BWA-MEM (maximal exact matches – максимальные точные совпадения), рекомендуемый создателями программного пакета, как наиболее быстрый и эффективный [5]. Фрагмент промежуточных результатов выравнивания с помощью BWA представлен на рис. 2.

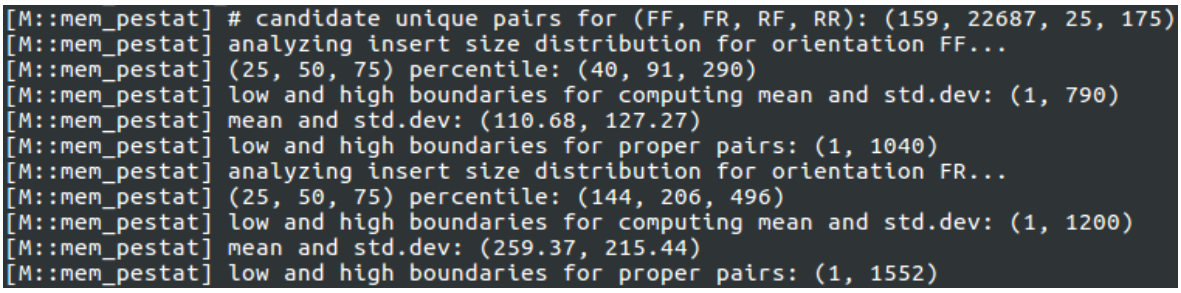

Рис. 2. Промежуточные результаты выравнивания с помощью BWA

BWA анализирует риды блоками по 99010 за один раз. Сначала алгоритм определяет ориентацию парных ридов друг относительно друга. В данном эксперименте алгоритм обнаружил, что в большинстве парных ридов из блока первый рид прочитан в прямом порядке, второй – в обратном. После этого алгоритм рассчитывает характеристики распределения длины ридов.

Оба инструмента для выравнивания выполнили задачу за сравнимое время (около 8 ч), однако при использовании Bowtie 2 есть возможность распараллелить вычисления на несколько ядер процессора, кроме этого Bowtie 2 использует меньше оперативной памяти (около 4 ГБ), чем алгоритм BWA-MEM в BWA (5,1 ГБ).

После выравнивания полученные файлы в формате SAM были конвертированы в формат BAM с помощью программного пакета инструментов SAMtools, отсортированы и проанализированы на наличие вариаций, после чего была произведена сборка секвенированной последовательности с помощью инструмента BCFtools.

ЗАКЛЮЧЕНИЕ

В результате выполнения данной работы экспериментальные риды генома человека были собраны против референсной последовательности с помощью программных пакетов Bowtie 2 и BWA на компьютере с процессором Intel Core i5-4210Ux4 и 8 ГБ памяти в ОС Ubuntu 16.04 LTS. Оба инструмента выполнили задачу за сравнимое время (около 8 ч), однако при использовании Bowtie 2 есть возможность распараллелить вычисления на несколько ядер процессора, кроме этого Bowtie 2 использует меньше оперативной памяти (около 4 ГБ), чем алгоритм BWA-MEM в BWA (5,1 ГБ). Кроме этого Bowtie 2 выводит результаты выравнивания в более удобном формате, чем BWA. В связи с чем для сборки генома на не очень высокопроизводительных машинах рекомендуется Boowtie 2. Наилучший пакет инструментов для обработки результатов картирования – SAMtools. Сборка генома против референса может использоваться для выявления различий в геномах различных организмов одного и того же вида, в том числе геномных мутаций.

Литература

1. *Burrows M., Weeler D.J.* A Block-sorting Lossless Data Compression Algorithm. // Palo Alto : Systems Research Center, 1994.
2. *Ferragina P., Venturini R.* FM-Index Version 2 // FM-index. - University of Pisa, 2000.
3. *Henson J., Tischler G., Ning Z.* Next-generation sequencing and large genome assemblies. // PMC. - Jun 2012.
4. *Langmead B.* Fast gapped-read alignment with Bowtie 2 // NCBI. - Mar 4, 2012.
5. *Li H.,* Durbin R.Fast and accurate short read alignment with Burrows–Wheeler transform // NCBI. - May 18, 2009.
6. Nagarajan Niranjan and Pop Mihai Sequence assembly demystified // Nature Reviews. : Macmillan Publishers Limited, 2013. - Vol. 14.
7. *Tamazian, G.* Chromosomer: a reference-based genome arrangement tool for producing draft chromosome sequences // BioMed Central. – Aug 22, 2016