ПРОГРАММНАЯ РЕАЛИЗАЦИЯ СБОРКИ ТРАНСКРИПТОМА ЧЕЛОВЕКА

Артанова Д. П., Яцков Н. Н., Гринев В. В.
Кафедра системного анализа и компьютерного моделирования, кафедра генетики, Белорусский государственный университет
Минск, Республика Беларусь
E-mail: {yatskou, grinev_vv}@bsu.by, darya.artanova@gmail.com

В данной работе изучена и выполнена сборка транскриптома против референса. Предложены рекомендации для оптимальной программной реализации сборки транскриптома против референсной последовательности.

Введение

Оптимальное применение существующих алгоритмов и программных средств для сборки секвенированных последовательностей ДНК и РНК является одной из важнейших задач биоинформатики [1]. Выделяют два принципиально различных подхода к сборке: с использованием референсной последовательности – уже собранного транскриптома или генома данного организма, или организма, родственного исследуемому, и сборка de novo.

Целью данной работы является изучение наиболее популярных программных средств для сборки данных секвенирования против референсной последовательности, исследование особенностей их работы на примерах смоделированных наборов данных и в процессе сборки прочтений транскриптома человека, полученных в результате секвенирования, а также разработка программной реализации оптимального способа сборки в условиях ограниченных вычислительных ресурсов.

I. Моделирование прочтений

Моделирование данных секвенирования представляет собой программную генерацию прочтений существующей (также смоделированной или полученной экспериментально) последовательности нуклеотидов. Необходимость в моделировании прочтений возникает, когда нужно проверить правильность работы тех или иных инструментов для обработки прочтений, а сделать это на экспериментально полученных прочтениях невозможно, так как их характеристики обычно неизвестны.

В процессе моделирования данных секвенирования можно выделить 2 этапа:

- фрагментация;
- секвенирование.

На этапе фрагментации последовательность, с которой будут моделироваться прочтения, разбивается на участки, длины которых распределены по определенному закону. На этапе секвенирования с фрагментов последовательности считываются прочтения.

Для моделирования прочтений использовался программный пакет Polyester, который создает прочтения, максимально похожие на получаемые с помощью технологии секвенирования Illumina, для чего использует программную имитацию этой технологии, генерируя и комплементарные последовательности.

II. Сборка с выравниванием на референс

Основная идея в сборке последовательностей с выравниванием на референс состоит в том, чтобы найти в референсной последовательности участки, максимально похожие на экспериментально полученные прочтения, – картировать прочтения – тем самым определив исходный порядок их следования [2]. В связи с тем, что геном человека состоит из 3 253 848 404 пар нуклеотидов, эта задача имеет большую вычислительную сложность.

Процесс сборки с выравниванием на референс можно разделить на следующие этапы:

- индексирование референсной последовательности;
- картирование прочтений;
- поиск вариантов и сборка.

Сборка транскриптома производилась на компьютере с процессором Intel Core i5-4210U @1.70 ГГц x4 и 8 ГБ оперативной памяти.

Для картирования экспериментальных прочтений в формате FASTQ на референсную последовательность использовались программные пакеты Bowtie 2 [3] и BWA [4]. В их основе лежит алгоритм, использующий преобразование Барроуза–Уилера и FM-индекс [5,6]. В качестве референсной последовательности рассмотрен геном человека hg38 в формате FASTA. Для обработки результатов выравнивания использовались программные пакеты SAMtools и BCFtools.

III. Результаты

В пакете Polyester смоделированы наборы прочтений с различающимися характеристиками, такими как длина и вероятность ошибочного нуклеотида. На смоделированных данных протестирована работа инструментов картирования BWA и Bowtie 2.

На наборе данных с прочтениями, длиной около 15 пар оснований, с задачей картирования не справился ни BWA, ни Bowtie 2; на наборе данных с прочтениями, длиной больше 250 пар оснований, удалось выровнять на референс абсолютно все прочтения с помощью обоих программных пакетов. BWA и Bowtie 2 одинаково хорошо справились с обработкой данных, содержащих большое количество ошибок: на наборе данных, содержащем 5% ошибочных нуклеотидов, 99,7% прочтений были успешно картированы, при использовании как BWA, так и Bowtie 2. При этом эти программные пакеты не рекомендуются для обработки данных, содержащих более, чем 2% ошибок, ввиду того, что нецелесообразно использование метода секвенирования, создающего настолько неточные наборы данных.

Из полученных результатов картирования можно сделать вывод, что BWA и Bowtie 2 справились с задачей картирования на одном уровне, причем существует явная зависимость качества выравнивания прочтений от их длины и процентного содержания в них ошибочных нуклеотидов.

Экспериментальный набор содержал 101 948 412 парных прочтения, и в результате работы BWA 80 354 002 из них (78,8%) выровнялись на референс попарно, в результате работы Bowtie 2 попарно выровнялись 78 527 108 (78%) прочтений; около 1% прочтений не выровнялось вообще, что является приемлемым результатом картирования транскриптома человека; остальные прочтения выровнялись на референсную последовательность по отдельности в ходе работы каждого из инструментов. Такой результат картирования прочтений транскриптома человека можно считать успешным.

При сборке в условиях ограниченных вычислительных ресурсов большую роль играют встроенные средства оптимизации вычислений программных пакетов, определяющих и время работы, и потребление ресурсов. В таблице 1 представлены оценки характеристик вычислительных ресурсов, используемых в процессе сборки (через дробь указаны значения для смоделированных и экспериментальных прочтений).

Таблица 1 – Сравнительный анализ потребляемых вычислительных ресурсов

Этап	Программа	Память, ГБ	Параллельные вычисления	Время, мин
Картирование	Bowtie 2	<1/4,1	да	11/570
Картирование	BWA	<1/5,3	нет	11/590
Сборка	SAMtools	3	нет	3/30
Сборка	BCFtools	3	нет	20/180

Кроме проанализированных в таблице вычислительных ресурсов, необходимо учитывать емкость накопителя, используемого для хранения информации. Суммарный размер всех файлов, необходимых для сборки и создаваемыми в ее процессе, составляет около 80 ГБ.

Заключение

В данной работе реализована имитационная модель секвенирования транскриптома и исследованы программные пакеты сборки транскриптома против референсной последовательности; предложен эффективный программный способ сборки транскриптома, включающий индексирование референсной последовательности и выравнивание на нее прочтений с помощью преобразования Барроуза–Уилера. С использованием предложенного программного способа сборки выполнен анализ смоделированных данных и сборка экспериментальных прочтений транскриптома человека и проведен анализ необходимых для этого вычислительных ресурсов. По результатам этого анализа можно сделать следующие выводы:

- минимальное необходимое количество оперативной памяти для сборки транскриптома человека – 6 ГБ;
- примерное необходимое количество времени для сборки полного транскриптома человека – 23 ч;
- время сборки напрямую зависит от вычислительной мощности процессора и от количества процессорных единиц, так как в некоторых программных пакетах существует возможность распараллеливать вычисления;
- при использовании предложенного программного способа, сборка транскриптома человека с выравниванием на референс выполнима в условиях ограниченных вычислительных ресурсов, например, на ПК.

1. Henson, J., Tischler, G., Nin, Z. Next-generation sequencing and large genome assemblies. // PMC. / J. Henson, G. Tischler // – Cambridge, 2012.
2. Tamazian, G. Chromosomer: a reference-based genome arrangement tool for producing draft chromosome sequences / G. Tamazian // – BioMed Central, 2012.
3. Langmead, B. Fast gapped-read alignment with Bowtie 2 /B. Langmead // – NCBI, 2012
4. Li, H., Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform / H. Li, R. Durbin // – NCBI, 2009
5. Burrows, M., Weeler D. J. A Block-sorting Lossless Data Compression Algorithm / M. Burrows, D. J. Weeler // – Palo Alto : Systems Research Center, 1994.
6. Ferragina, P., Venturini, R. FM-Index Version 2 /P. Ferragina, R. Venturini // – University of Pisa, 2000