doi 10.18699/bgrs2024-06-17

Разработка биоинформатического конвейера для обработки данных, полученных методом GBS

Пронозин А.[1, 2]*, Афонников Д.[1, 2]

[1] *Институт цитологии и генетики СО РАН, Новосибирск, Россия*

[2] *Курчатовский геномный центр ИЦиГ СО РАН, Новосибирск, Россия*

** pronozinartem95@gmail.com*

Ключевые слова: GBS; ячмень; биоинформатический конвейер

Мотивация и цели: Развитие технологий секвенирования нового поколения открыло новые возможности для генотипирования различных организмов, включая растения. Метод генотипирования путем секвенирования (GBS) применяется для идентификации генетической изменчивости и более быстрого генотипирования образцов, а также является более экономически эффективным методом в сравнении с полногеномным секвенированием. GBS продемонстрировал свою надежность и гибкость для ряда видов и популяций растений. Этот метод был применен для генетического картирования, выявления молекулярных маркеров, геномной селекции, в исследовании генетического разнообразия, идентификации сортов, а также в исследованиях в области биологии-охраны природы и эволюционной экологии.

GBS существенно сокращает как стоимость, так и время, необходимое для секвенирования исследуемых образцов. Это привело к потребности в разработке качественного биоинформатического анализа для постоянно расширяющегося количества секвенированных данных.

В настоящей работе нами был разработан биоинформатический конвейер GBS-DP для анализа данных, полученных методом GBS. Конвейер применим для любых видов организмов. Конвейер полностью автоматизирован и позволяет обрабатывать большие объемы данных (более 400 образцов).

Методы и алгоритмы: Конвейер состоит из трех основных этапов: предобработка данных, поиск полиморфизмов, анализ генетического разнообразия. Реализация конвейера на платформе Snakemake позволила полностью автоматизировать процесс расчета и установки необходимых программных пакетов. Для тестового применения конвейера GBS-DP в настоящей работе был использован проект PRJEB39633 из базы данных European Nucleotide Archive (ENA) [1], который содержит библиотеки GBS для популяции ячменя.

Результаты: Конвейер предоставляет результаты оценки базовых характеристик секвенированных библиотек: длина прочтения для каждой библиотеки, средняя глубина прочтения, количество прочтений, приходящихся на одну библиотеку, среднее покрытие каждой библиотеки и всех библиотек в целом.

Также конвейер предоставляет результаты поиска полиморфизмов между исследуемыми генотипами. Для 272 исследуемых образцов выявлено 447409 SNP. Общее количество индел 46557. Параметр неравновесия по сцеплению (LD) (r2) был выбран равным 0.5. После применения фильтра LD осталось 45402 полиморфных и независимых SNP.

Конвейер предоставляет распределение выявленных SNP по хромосомам, анализ главных компонент генотипов на основе выявленных SNP и построение филогенетического дерева. Результаты анализа главных компонент на основе 45402 SNP показывают, что внутри исследованной популяции на диаграмме рассеяния в пространстве двух первых компонент четко выделяется несколько кластеров.

Выводы: В настоящей работе нами был предложен биоинформатический конвейер GBS-DP, который позволяет обрабатывать данные широкомасштабного секвенирования, проведенного методом GBS. Результаты демонстрируют достаточно высокую скорость работы конвейера как для больших данных (более 400 библиотек), так и для малых (30 библиотек). Конвейер предоставляет также анализ выявленных полиморфизмов.

Финансирование: Работа выполнена за счет финансирования Курчатовского геномного центра ФИЦ ИЦиГ СО РАН, соглашение с Министерством образования и науки РФ № 075-15-2019-1662. Вычисления проводились с использованием ресурсов ЦКП «Биоинформатика».

Development of a bioinformatics pipeline for GBS data processing

Pronozin A.[1, 2]*, Afonnikov D.[1, 2]

[1] *Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia*

[2] *Kurchatov Genomic Center of the Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia*

* *pronozinartem95@gmail.com*

Key words: GBS; barley; bioinformatics pipeline

Motivation and Aim: The development of next-generation sequencing technologies has opened up new opportunities for genotyping various organisms, including plants. Genotyping by sequencing (GBS) is used to identify genetic variability and to genotype samples more rapidly, and is more cost-effective than whole-genome sequencing. GBS has demonstrated its reliability and flexibility for a number of plant species and populations. The method has been applied to genetic mapping, molecular marker detection, genomic selection, in genetic diversity studies, variety identification, and in conservation biology and evolutionary ecology studies.

GBS significantly reduces both the cost and the time required for sequencing the samples under study. This has led to the need to develop high-quality bioinformatics analysis for the ever-expanding amount of sequenced data.

In the present work, we have developed the GBS-DP bioinformatics pipeline for analyzing GBS-derived data. The pipeline is applicable to any species of organisms. The pipeline is fully automated and allows processing large amounts of data (more than 400 samples).

Methods and Algorithms: The pipeline consists of three main stages: data preprocessing, search for polymorphisms, and genetic diversity analysis. The implementation of the pipeline on the Snakemake platform allowed us to fully automate the process of calculation and installation of the necessary software packages. For the test application of the GBS-DP pipeline in this work, we used the PRJEB39633 project from the

European Nucleotide Archive (ENA) database [1], which contains GBS libraries for the barley population.

Results: The pipeline provides results of evaluation of basic characteristics of sequenced libraries: read length for each library, average read depth, number of reads per library, average coverage of each library and all libraries in general.

The pipeline also provides the results of the search for polymorphisms between the genotypes analyzed. For the 272 samples analyzed, 447,409 SNPs were identified. The total number of indels is 46,557. The linkage disequilibrium (LD) parameter (r2) was chosen to be 0.5. After applying the LD filter, 45,402 polymorphic and independent SNPs remained.

The pipeline provides the distribution of the identified SNPs across chromosomes, principal component analysis of genotypes based on the identified SNPs, and construction of a phylogenetic tree. The results of principal component analysis based on 45,402 SNPs show that within the studied population, several clusters are clearly distinguished in the scatter diagram in the space of the first two components.

Conclusion: In the present work, we proposed the GBS-DP bioinformatics pipeline, which allows us to process large-scale sequencing data performed by the GBS method. The results demonstrate quite high speed of the pipeline for both large data (more than 400 libraries) and small data (30 libraries). The pipeline also provides analysis of detected polymorphisms.

Funding: This work was funded by the Kurchatov Genomic Center of the Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences, agreement with the Ministry of Education and Science of the Russian Federation No. 075-15-2019-1662. Calculations were performed using the resources of the Bioinformatics Center.

Список литературы/References

1. Leinonen R. et al. The European nucleotide archive. *Nucleic Acids Res.* 2010;39(Suppl.1):D28-D31