DOI 10.18699/GenBio2022-14

ВЫЧИСЛИТЕЛЬНЫЙ КОНВЕЙЕР ДЛЯ АНАЛИЗА РЕЗУЛЬТАТОВ ЭКСПЕРИМЕНТОВ ПО ГЕНОТИПИРОВАНИЮ ПУТЕМ СЕКВЕНИРОВАНИЯ (GBS)

Пронозин Артем Юрьевич, Афонников Дмитрий Аркадьевич, Салина Елена Артемовна.
Kurchatov Genomic Center of the Institute of Cytology and Genetics of Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
pronozinartem95@gmail.com

Метод Genotyping-by-sequencing (GBS) применяется для идентификации генетической изменчивости и более быстрого генотипирования образцов, а также является более экономически эффективным методом в сравнении с полногеномным анализом. GBS часто применяют в исследованиях геномов сельскохозяйственных растений. При том, что результат генотипирования как правило дает около сотни тысяч геномных вариаций, а современные эксперименты требуют генотипировать сотни образцов, возникает необходимость использования биоинформатических методов, ориентированных на сверхбольшие объемы данных. В работе предложен вычислительный конвейер для биоинформатической обработки результатов экспериментов GBS. С его помощью проведен анализ Российских сортов ячменя из коллекции ГенАгро ИЦиГ СО РАН.

В работе использовались данные эксперимента GBS для сортов и линий ячменя из коллекции ГенАгро ИЦиГ СО РАН (195 образцов Российских сортов и 24 зарубежных), всего получено 219 библиотек. Конвейер биоинформатического анализа включал несколько этапов: 1. Предварительная обработка данных. 2. Поиск SNP. 3. Анализ вариаций на основе сравнения полиморфизмов. В качестве референсного генома использована последовательность *Hordeum vulgare* (IBSC_v2).

Предварительный анализ библиотек показал, что средняя глубина прочтения образцов – 9.7. В среднем на образец приходится 250 тысяч маркеров SNP и 13 тысяч инделов. Кластеризация методом главных компонент (PCA) проведенная на основе SNP, выявила 2 кластера, которые включают 64 и 134 генотипов, 21 образец не относится ни к одному из указанных кластеров. Анализ распространения генотипов по климатическим зонам России показал, что в 1 кластере преобладают образцы из северных районов России во 2 кластере из южных. Некоторые образцы, не принадлежащие кластерам, представлены многорядными голозерными видами, тогда как образцы, относящиеся к двум указанным кластерам в основном двурядные. Филогенетическое дерево подтвердило структуру популяции, которая была получена методом PCA.

Разработан вычислительный конвейер для биоинформатического анализа результатов эксперимента GBS. В результате анализа 219 библиотек ячменя, в геномах образцов выявлено 13.668.021 SNP маркеров и 845.413 инделов в сравнении с референсным геномом ячменя. Полученные 2 кластера (метод PCA) демонстрируют четкое разграничение генотипов ячменя по климатическим зонам России.

Ключевые слова: GBS, генотипирования, ячмень.

COMPUTATIONAL PIPELINE FOR ANALYZING THE RESULTS OF GENOTYPING BY SEQUENCING (GBS) EXPERIMENTS

Pronozin Artem Yuriovich, Afonnikov Dmitry Arkadyevich, Salina Elena Artemovna.
Kurchatov Genomic Center of the Institute of Cytology and Genetics of Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
pronozinartem95@gmail.com

Genotyping-by-sequencing (GBS) method is used for identification of genetic variability and faster genotyping of samples, and is a more cost-effective method compared to whole genome analysis. GBS is often used in studies of agricultural plant genomes. While the result of genotyping usually yields about hundreds of thousands of genomic variations, and modern experiments require genotyping hundreds of samples, there is a need to use bioinformatic methods focused on extremely large volumes of data. In this work we proposed a computational pipeline for bioinformatic processing of the results of GBS experiments. It was used to analyze Russian barley varieties from the GenAgro collection of ICG SB RAS.

GBS experiment data for barley varieties and lines from the GenAgro collection of ICG SB RAS (195 samples of Russian varieties and 24 foreign ones) were used, in total 219 libraries were obtained. The bioinformatic analysis pipeline included several stages: 1. Data pre-processing. 2. SNP search. 3. Analysis of variation based on polymorphism comparison. The *Hordeum vulgare* sequence (IBSC_v2) was used as a reference genome.

Preliminary analysis of the libraries showed that the average reading depth of the samples is 9.7. On average, there are 250 thousand SNP markers and 13 thousand indels per sample. Principal component analysis (PCA) clustering based on SNPs revealed 2 clusters including 64 and 134 genotypes, 21 samples did not belong to any of these clusters. Analysis of the distribution of genotypes by climatic zones of Russia showed that in the 1st cluster samples from the northern regions of Russia prevail in the 2nd cluster from the southern regions of Russia. Some specimens not belonging to the clusters are represented by multiserial holosere species, while the specimens belonging to the two mentioned clusters are mainly biserial. The phylogenetic tree confirmed the population structure, which was obtained by the PCA method.

A computational pipeline for bioinformatic analysis of the GBS experiment results was developed. As a result of the analysis of 219 barley libraries, 13,668,021 SNP markers and 845,413 indels were identified in the genomes of the samples compared with the reference barley genome. The obtained 2 clusters (PCA method) demonstrate a clear distinction of barley genotypes by climatic zones of Russia.

Keywords: GBS, genotyping, barley.