

MDSearch: алгоритм выбора минимального информативного набора SNP для сравнительного анализа широкого спектра видов

А.С. Ермолаев*, М.А. Самарина, Д.С. Ульянов, П.Ю. Крупин, Г.И. Карлов, М.Г. Дивашук

Всероссийский научно-исследовательский институт сельскохозяйственной биотехнологии, Москва, 127550, Россия

* alerm@al-erm.ru

Цель: Набор SNP-маркеров часто используется для анализа наборов сортов широкого спектра сельскохозяйственных видов. Во многих случаях, не требуется использовать все найденные маркеры для однозначной дискриминации всех образцов в наборе. Исходя из этого, может быть определен минимальный информативный набор (МИН) маркеров, который однозначно отличит все пары образцов в наборе. Ранее создавались программы, способных находить МИН маркеров для набора образцов [1], однако в настоящее время эти программы недоступны для загрузки. В то же время поиск МИН маркеров для набора образцов является важной задачей в селекции и семеноводстве [2], что делает актуальным создание алгоритмов для решения данной задачи.

Материалы и методы: MDSearch написан на Python3 и находится в открытом доступе на Github (<https://github.com/alermol/MDSearch>). Пользователь предоставляет на вход файл в формате VCF, содержащий генотипы целевых образцов. Файл может содержать пропущенные генотипы, что значительно расширяет возможности выбора МИН, позволяя использовать его для данных, полученных с помощью технологий, дающих неравномерное покрытие (Genotyping-by-sequencing). Поддерживаются VCF любой пloidности, что важно для сельскохозяйственных видов, среди которых множество видов с полиплоидными геномами. При запуске программы пользователь может определить минимальное количество SNP, которые будут отличаться у пары образцов, что позволяет контролировать степень устойчивости созданного МИН к ошибкам генотипирования. По умолчанию MDSearch подбирает один набор, содержащий МИН, однако пользователь может определить как количество маркеров в итоговом наборе, так и количество наборов, которые подбирает программа. В случае необходимости расширить МИН до требуемого количества маркеров, выбираются маркеры с наибольшим полиморфизмом на основании параметра PIC (Polymorphic Index Content).

Алгоритм MDSearch включает в себя два основных шага для поиска МИН. На первом этапе происходит прямой жадный отбор маркеров, имеющих наибольшую частоту минорного аллеля (MAF). Отбор происходит до тех пор, пока полученный набор не станет способен однозначно дискриминировать набор образцов как минимум по одному маркеру или по количеству, заданному пользователем при запуске. Когда такой набор найден, происходит минимизация набора путем обратного отбора. Из набора убирается один случайный маркер; если полученный набор теряет дискриминационную способность, маркер возвращается в набор, и убирается другой маркер. Этап минимизации может выполняться параллельно, что повышает скорость нахождения МИН. В каждом потоке происходит независимый процесс минимизации. На финальном этапе полученный МИН может расширяться наиболее полиморфными маркерами до размера, определенного пользователем при запуске программы. Полученный МИН маркеров возвращается пользователю в виде VCF и может быть использован в дальнейшем анализе.

Результаты: Нами была создана программа для поиска МИН маркеров MDSearch. Данная программа позволяет работать с вариантами, полученными из широкого диапазона видов с различной структурой генома, и позволяет быстро выбрать один или несколько наборов маркеров для их однозначной дискриминации.

Выводы: MDSearch позволяет эффективно решать задачу поиска МИН маркеров для заданного набора образцов и может быть полезна в селекционном и семеноводческом процессах для защиты авторских прав и контроля чистоты семенного материала. Результаты апробации инструмента MDSearch опубликованы [3].

Финансирование: Исследование выполнено при поддержке Министерства науки и высшего образования Российской Федерации, государственное задание FGUM-2023-0005.

Список литературы

- 1 Gale K., Jiang H., Westcott M. An optimization method for the identification of minimal sets of discriminating gene markers: application to cultivar identification in wheat. *J Bioinform Comput Biol.* 2005;3(2):269-279
- 2 Henning J.A., Coggins J., Peterson M. Simple SNP-based minimal marker genotyping for *Humulus lupulus* L. identification and variety validation. *BMC Res Notes.* 2015;8(1):542
- 3 Ermolaev A. et al. A Tool for the Design of the Minimal Fingerprinting SNP Set: Use Case for Barley. *Agronomy.* 2024;14(8):1802