

## Создание пайплайна для автоматизации поиска O-антигена граммотрицательных бактерий

Научный руководитель – Комиссаров Алексей Сергеевич

*Кучур П.Д.<sup>1</sup>, Зилов Д.С.<sup>2</sup>*

1 - Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики, Санкт-Петербург, Россия, *E-mail: tniapp@yandex.ru*; 2 - Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики, Санкт-Петербург, Россия, *E-mail: zilov@scamt-itmo.ru*

Соматический O-антиген является компонентом наружной мембраны грамотрицательных бактерий. По своей природе это полисахарид, имеющий определенную специфичность, которая способствует идентификации бактерий. Разнообразие O-антигенных форм лежит в основе классификации путем выделения O-серогрупп микроорганизмов. Кроме того, O-антиген содержит эпитопы - специфические сайты связывания с антителами организма-хозяина, что обуславливает развитие иммунных реакций в ответ на инфицирование.

Гены, вовлеченные в биосинтез O-антигена, зачастую объединены в кластер. Несмотря на относительную простоту организации этого кластера его выявление в геноме микроорганизмов до сих пор остается сложным, а серотипирование основана на поиске нескольких генов из этого кластера, что значительно снижает точность. А большинство доступных инструментов ограничены только геномом *E. coli*. Наша цель - создание пайплайна, позволяющей автоматизировать поиск, сборку и аннотация кластера генов вовлеченных в синтез O-антигена.

Для создания программы была использована система управления рабочими процессами Snakemake. Мы выбрали этот инструмент так как в него с одной стороны интегрирован менеджер пакетов Conda, который содержит в себе множество биоинформатических инструментов, необходимых в работе, а с другой стороны можно использовать собственные инструменты и скрипты.

На вход разработанному пайплайну подаются сырые данные секвенирования или геномная сборка. Программа автоматически оценивает качество исходных данных, убирает адаптеры и оптические дубликаты. Далее запускается сборка генома с использованием SPAdes и Unicycler. После этого геномная сборка проверяется на наличие контаминаций и плазмид. Очищенная от контаминаций сборка аннотируется с помощью Prokka и eggNOG. На основе полученной сборки создается полная функциональная аннотация генома, из которой вычлняются известные гены, вовлеченные в биосинтез O-антигена. Используя полученную аннотацию и совпадение с базой данных STRING, программа аннотирует опероны, содержащие гены, вовлеченные в биосинтез O-антигена, с учетом возможных ошибок сборки и разрывов контигов.

В настоящий момент мы проводим аннотацию O-антигенов всех грамотрицательных бактериях доступных в базе данных NCBI. Полученная в результате информация будет использована для улучшения существующего пайплайна.

Созданный пайплайн позволяет существенно экономить время при классификации грамотрицательных бактерий с идентификацией их серогрупп, а также при выявлении компонентов наружной мембраны, связанных как с вирулентностью микроорганизмов, так и с симбиотическим взаимодействием между бактериями и растениями.