УДК 575

ПОИСК ГЕНЕТИЧЕСКИХ АССОЦИАЦИЙ НА ОСНОВЕ ОТКРЫТЫХ И АНКЕТНЫХ ДАННЫХ

А.С. Ракитько[1,2], И.И. Низамутдинов[1], А.У. Елмуратов[1,4], Я.В. Попов[1], Н.А. Слепов[1,2], В.В. Ильинский[1,3,4,5]

[1]ООО «Генотек ИТ»; [2]Кафедра теории вероятностей, механико-математический факультет, МГУ им. М.В. Ломоносова; [3]Российский национальный исследовательский медицинский университет им. Н.И. Пирогова; [4]НИИ биомедицинской химии им. В.Н. Ореховича; [5]Институт общей генетики им. Н.И. Вавилова, Москва, Россия

Задача выявления генетических маркеров, которые значимо ассоциированы с некоторым фенотипом, является одной из основных в современной генетике. Анализ баз данных показывает, что к 2018 году было проведено уже около 3000 GWAS (genome-wide association study) исследований. Как правило, научные группы при проведении GWAS исследований отталкиваются от конкретного фенотипа: собирают выборку больных ("cases") и здоровых ("controls"), осуществляют генотипирование или секвенирование, статистически анализируют полученные данные. В дальнейшем, данная выборка может участвовать в мета-анализах, однако, использование этих данных для проведения новых GWAS исследований затруднительно, поскольку образцы в коллекции описаны в контексте только одного конкретного фенотипа (заболевания). С точки зрения повторного использования генетических данных более перспективным выглядит проект UK 100 000 Genomes [1], в котором результатам полногеномного секвенирования сопоставляются электронные медицинские карты. Это позволяет генерировать обучающие выборки для широкого спектра относительно частых мультифакторных заболеваний. Однако, возможно, ещё более интересным источником данных для ассоциативных исследований являются клиенты direct-to-consumer (D2C) компаний. На текущий момент, уже более 10 миллионов людей имеют результаты микрочипового генотипирования [3]. Дополнение генетических данных ответами на опросники создает беспрецедентный массив данных для GWAS исследований. Одним из примеров платформ, агрегирующих генетические и фенотипические данные клиентов D2C компаний является OpenSNP [2]. Количество пользователей платформы, самостоятельно предоставивших доступ к своим данным, превышает 6 000 людей, которые в совокупности ответили на более 50 000 анкетных вопросов. Мы использовали эти данные для обучения и тестирования модели по предсказанию цвета глаз и волос. Полученные результаты оказались сопоставимы по точности исследованиям, в которых собиралась собственная коллекция образцов для изучения указанных выше фенотипов (AUC > 0.9). Кроме того, были проанализированы анкеты 3 000 клиентов медико-генетического центра Genotek с последующим проведением GWAS исследований. Подобные "proof of concept" работы демонстрируют перспективность использования анкетных и генетических данных, находящихся в открытом доступе.

Ключевые слова: открытые данные, генетические тесты, GWAS.

Литература

1. The 100,000 Genomes Project Protocol v3, Genomics England. doi:10.6084/m9.figshare.4530893.v2. 2017.
2. Greshake B. et al. openSNP–a crowdsourced web resource for personal genomics //PLoS One. – 2014. – Т. 9. – №. 3. – С. e89204.
3. Khan R., Mittelman D. Consumer genomics will change your life, whether you get tested or not //Genome biology. – 2018. – Т. 19. – №. 1. – С. 120.