Секция «Биоинформатика»

Применение методов кластеризации для поиска групп SNPs, ассоциированных с ишемическим инсультом

Сапожников Никита Алексеевич
Аспирант
Институт молекулярной генетики РАН, Москва, Россия
E-mail: nikita.sapozhnikov1@gmail.com

Актуальность. На сегодняшний день инсульты являются одной из наиболее частых патологий. Ежегодно регистрируется около 15 млн. случаев инсульта, из которых треть заканчивается смертельным исходом, а 70-80% выживших становятся инвалидами [1]. На долю ишемического инсульта приходится 80-85% всех инсультов.
Цель исследования. Исследовать возможность и особенности применения методов кластеризации для поиска генетических локусов (групп однонуклеотидных полиморфизмов, SNPs), связанных с риском развития ишемического инсульта. Материалы. Исходные данные были загружены из международной базы данных генотипов и фенотипов dbGaP [2]. После предварительной обработки данных с целью контроля их качества, суммарное число SNPs составило 883908, а индивидов - 5580, из которых 652 составляли контрольную группу. Мужчины и женщины были представлены в равных пропорциях. Исходные генотипические данные с использованием программы Plink были трансформированы в матрицы неравновесия по сцеплению (LD), включающие попарные значения LD (r2) между всеми SNPs в составе аутосом.

Методы. Кластеризация решает задачу разбиения исходного набора объектов на группы таким образом, чтобы объекты в одной группе были максимально схожи, а в разных группах максимально различны. В нашем исследовании объектами являлись SNPs, а мерой их близости - 1-r2. Для кластеризации SNPs применили алгоритмы DBSCAN и HDBSCAN. В полученных каждым методом группах SNPs восстановили гаплотипы и, используя критерий Хи-квадрат, сравнили частоты их встречаемости в контрольной и тестовой выборках при помощи программы Plink. Статистическую значимость различий устанавливали по p-value после применения поправки на множественность тестирования. SNPs значимых гаплотипов общих для обоих алгоритмов аннотировали с помощью программы snpEff. Полученные списки генов-кандидатов проверили на сверхпредставленность в базах данных канонических путей и генных онтологий MSigDB.

Результаты. 16 идентифицированных генов оказались сверхпредставленными в подсемействах A и B гамма-протокадеринов, что позволяет предположить значимость процессов адгезии клеточных мембран и межклеточных взаимодействий в развитии инсульта.
Особенность работы в том, что DBSCAN и HDBSCAN образовали мозаичные кластеры SNPs, когда SNPs из одного кластера не обязательно являются соседними на геноме. Стоит также отметить, что предложенная кластеризация снижает размерность данных, что актуально для ассоциативных исследований с применением методов машинного обучения.

Источники и литература

1) Gattringer T. [и др.]. Predicting Early Mortality of Acute Ischemic Stroke // Stroke. 2019. № 2 (50). С. 349–356.

2) Meschia J. F. [и др.]. NINDS Stroke Genetics Network (SiGN) Experience with the Causative Classification System // International Journal of Stroke. 2013. № 4 (8). С. E9–E9.