МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное учреждение науки

Институт белка Российской академии наук

Программа подготовки научно-педагогических кадров в аспирантуре

06.06.01 Биологические науки

Направленность (профиль) – Молекулярная биология

Мещеряков Георгий Андреевич

# Оценка аллельного дисбаланса по данным высокопроизводительного секвенирования для поиска регуляторных однонуклеотидных вариантов

*Аннотация научно-квалификационной работы*

Научный руководитель:

к. ф-м.н. и д.б.н. И.В. Кулаковский

Выпускник:

Г.А. Мещеряков

Пущино

2025

Индивидуальные геномы особей одного вида, включая людей, значительно различаются по своим нуклеотидным последовательностям. Самыми частыми отличиями являются единичные замены в ДНК: однонуклеотидные варианты (ОНВ) т.е. соматические мутации и популяционные однонуклеотидные полиморфизмы (ОНП). Относительно эталонной (также называемой «референсной») последовательности генома человека, традиционно используемой в молекулярно-биологических и биоинформатических исследованиях, индивидуальный геном насчитывает десятки миллионов однонуклеотидных замен. Часть ОНВ находится в участках, кодирующих белки, но большинство располагается в некодирующих областях генома, в том числе, в районах управляющих транскрипцией генов: промоторах и энхансерах. Присутствие ОНВ в регуляторных районах может оказывать прямое влияние на активность транскрипции, например, через изменение эффективности инициации транскрипции, которая, в свою очередь, зависит от ДНК-белковых взаимодействий, и, как следствие, от нуклеотидной последовательности. Аннотация и предсказание функциональной роли отдельных ОНВ важны для решения задач медицинской генетики и персонализированной медицины.

Сегодня наиболее распространенным методом выявления маркерных полиморфизмов, связанных с заболеваниями, является проведение полногеномных и полнотранскриптомных исследований ассоциаций, которые выявляют связи «генотип-фенотип» в большом масштабе. Такие исследования ограничены размером используемой популяционной выборки и не предоставляют функциональной информации о молекулярном механизме, объясняющим вовлечение вариантов в формирование нормального фенотипа или предрасположенности к патологии. Тот факт, что геном человека в большинстве клеток является диплоидным, позволяет использовать альтернативный подход: аллель-специфический анализ, когда изучаются гетерозиготные ОНП, отличающие материнскую и отцовскую хромосомы одного организма.

Аллель-специфичный анализ это функциональное сравнение аллелей, расположенных на гомологичных хромосомах, с точки зрения экспрессии гена или его альтернативной регуляции. Массовое распространение и удешевление высокопроизводительного секвенирования позволяет масштабировать аллель-специфичный анализ отдельных генов до полного генома или транскриптома, в зависимости от типа проведенного «омиксного» высокопроизводительного эксперимента. На практике, для полногеномного аллель-специфичного анализа, как

правило, доступны результаты небольшого числа экспериментальных повторностей, что ограничивает применение конвенциальных статистических моделей, требующих больших выборок. Достаточно часто ограниченное число повторностей обусловлено трудоемкостью или стоимостью проведения функциональных омиксных экспериментов. Таким образом, необходимы вычислительные и статистические модели, учитывающие природу экспериментальных данных и менее требовательные к размеру выборки. Строго говоря, во многих случаях требуется не точная количественная оценку соотношения сигналов с альтернативных аллелей, а качественный ответ на вопрос «является ли разница между аллельными сигналами значимой», например, достигается ли достаточно низкое P-значение. При этом требуется учитывать шумы, техническую вариабельность и другие источники отклонений в распределениях числа прочтений между вариантами и аллелями, например, вызванные ошибками картирования коротких прочтений на референсную геномную последовательностей. К сожалению, существующие подходы ошибочно переоценивают либо недооценивают дисперсию и не способны учитывать информацию о вариациях копийности участков генома, вплоть до хромосомных дупликаций. Также, недостаточно внимания уделяется дифференциальной оценке аллель-специфичности между содержательно отличными группами образцов, например, различными типами клеток.

Цель настоящей работы состоит в разработке вычислительного метода, отвечающего этим вызовам. В работе удалось показать, что распределение чисел прочтений удовлетворяет смеси отрицательных мультиномиальных распределений, и это утверждение легло в основу нового подхода к определению аллель-специфичных вариантов. Предложенный в работе метод был реализован в рамках программного пакета MIXALIME, протестирован в сравнении с существующими подходами и успешно использован для поиска аллель-специфичных ОНВ по данным экспериментов CAGE-Seq для здоровых и больных человеческих сердец. Среди них присутствовали как известные варианты, ассоциированные с развитием патологий, так и сотни новых ранее не описанных регуляторных вариантов.