Секция «Генетика»

Полногеномные ассоциативные исследования (GWAS) с признаками воспроизводства у свиней на основе алгоритмов машинного обучения

Научный руководитель – Вольчик Вячеслав Витальевич

Бакоев Фаридун Сирожидинович
Студент (бакалавр)
Южный федеральный университет, Ростов-на-Дону, Россия
E-mail: bakoevfaridun@yandex.ru

Бакоев Ф.С.[1,2], Колосова М.А.[1,2]
[1]ФБНУ ФНЦ Всероссийский институт животноводства им. Л.К. Эрнста, г. Подольск, пос. Дубровицы, Россия
[2]Южный федеральный университет, г.Ростов на Дону, Россия

Полногеномный поиск ассоциаций (Genome wide association study, или GWAS) является мощным инструментом исследования генетической архитектуры полигенных признаков, который применяется для выявления генетических факторов, связанных с селекционно-значимыми признаками с.-х. животных [1]. Большинство исследований GWAS основаны на статистике Фишера и Райта, однако анализ больших геномных данных затруднен такими проблемами, как небольшое количество наблюдений и большое количество прогнозирующих переменных (обычно известных как «большое P маленькое N»), высокая размерность или сильно коррелированные структуры данных. Одним из инструментов, позволяющий решить эту проблему является машинное обучение [2].

Целью работы является проведение GWAS с признаками воспроизводства у свиней крупной белой породы с помощью различных алгоритмов машинного обучения. Исследования проводили на свиноматках крупной белой породы (n=360) из селекционно-генетического центра РФ. Для анализа использовали данные по количеству поросят при рождении за первые три опороса. Геномную ДНК экстрагировали из проб (ушной выщип) набором реагентов ДНК-Экстран-2 (ООО «НПФ Синтол», Россия). Генотипирование животных проводили с использованием GeneSeek GGP Porcine HD (San Diego, USA). Функции пакета NAM использовали для вменения (импутации) пропущенных значений с помощью алгоритма машинного обучения Random Forest (RF, случайный лес). На основе функций, заложенных в пакет caret (R) выбрали наиболее значимые SNP, в отношение которых применили нейтральное кодирование (0,1,2). Также удалили все предикторы с нулевой дисперсией (неполиморфные SNP). Данные были разделены в соотношении 70% (обучающие) и 30% (тестовые). Из 55328 SNP после предварительного анализа были выбраны 44268 в качестве предикторов и на их основе составлены модели, основанные на алгоритмах Random Forest, Neural Networks (нейронные сети) и XGboost (экстремальный градиентный спуск) с параметрами, установленными по умолчанию. Наиболее точную оценку (RMSE) имела модель на основе алгоритма Random Forest. С применение диаграммы Венна были выбраны наиболее информативных SNP по всем построенным моделям, которые можно рассматривать в качестве значимых генетических вариантов, ассоциированных с количеством поросят при рождении у свиней.

Исследование выполнено за счет гранта Российского научного фонда (проект №19-76-10012).

Источники и литература

1) 1. Tang Z., Xu J., Yin L. et al. Genome-Wide Association Study Reveals Candidate Genes for Growth Relevant Traits in Pigs. Front Genet. 2019. 5; 10:302.

2) 2. Romagnoni A., Jégou S., Van Steen K. et al. Comparative performances of machine learning methods for classifying Crohn Disease patients using genome-wide genotyping data. Sci Rep. 2019. 9, 10351.