Секция «2.2 Биоинформатика»

# Оценка влияния однонуклеотидных замен в регуляторных элементах на экспрессию в рамках конкурса “CAGI7 LentiMPRA Challenge”

**Научный руководитель – Пензар Дмитрий Дмитриевич**

***Василенко Алексей Анатольевич***

*Студент (магистр)*

Московский государственный университет имени М.В.Ломоносова, Факультет биоинженерии и биоинформатики, Москва, Россия

*E-mail: genadijgenadev@gmail.com*

В настоящий момент эффекты однонуклеотидных вариантов (SNV) в кодирующих областях возможно оценить с относительно высокой точностью, однако предсказание влияния SNV в регуляторных элементах на экспрессию генов, что широко востребованно в современной медицинской геномике при персонифицированном подходе к лечению пациентов, и в лабораторной практике, для создания синтетических последовательностей с заданными свойствами, по-прежнему остается сложной задачей. Таким образом, разработка инструментов, позволяющих качественно предсказывать эффекты SNV в регуляторных областях, является актуальным направлением.

В рамках конкурса CAGI7 LentiMPRA Challenge перед участниками была поставлена задача разработать предсказательную модель для учёта влияния SNV в регуляторных областях на экспрессию генов на примере клеточной линии HepG2. Командам был предоставлен доступ к данным по экспрессии для $\sim 20.000$ вариантов в регуляторных областях, полученным в ходе MPRA-эксперимента.

На основе этих данных нами была изучена предсказательная способность трёх инструментов, основанных на разных подходах: ChromBPNet (CNN) [2], deltaSVM [1] и мотивные признаки транскрипционных факторов. ChromBPNet и deltaSVM обучаются на треках ATAC- или DNase-Seq и предсказывают доступность хроматина, однако в рамках парадигмы Zero-Shot Learning полученные предсказания можно экстраполировать и на изменение экспрессии. В качестве мотивных признаков были взяты изменения p-value PERFECTOS-APE [3] матриц PFM различных транскрипционных факторов (HOCOMOCO, JASPAR и Factorbook).

Наилучшее качество было получено при использовании ансамбля моделей ChromBPNet, обученных на данных эксперимента DNase-Seq для HepG2 – PearsonR 0,43 (0,66 на данных со значимым изменением экспрессии), что сопоставимо с результатами Borzoi. Модели, обученные на данных по близким клеточным линиям или клеткам печени, показали более скромные результаты. Полученное решение вошло в состав итогового ансамбля, занявшего первое место в конкурсе. Предсказания deltaSVM оказались значительно менее точными – PearsonR 0,29 (0,59 на данных со значимым изменением экспрессии). На основе одних мотивных признаков не удалось получить качественное предсказание, так как лишь небольшая часть из них имеет хорошую корреляцию с экспериментальными данными. Однако учёт таких мотивных признаков в ансамбле с другими предсказательными моделями приводит к приводит к ограниченному росту корреляции.

По результатам конкурса была проведена аннотация использованных регуляторных элементов и выполнен сравнительный анализ эффективности рассмотренных нейросетевых архитектур и других моделей при оценке регуляторного потенциала в зависимости от типа исследуемой последовательности.

## Источники и литература

1) Lee, D., Gorkin, D., Baker, M., et al. (2015). A method to predict the impact of regulatory variants from DNA sequence. Nature Genetics, 47(8), pp. 955–961. https://doi.org/10.1038/ng.3331.

2) Pampari, A., et al. (2025). ChromBPNet: bias factorized, base-resolution deep learning models of chromatin accessibility reveal cis-regulatory sequence syntax, transcription factor footprints and regulatory variants. bioRxiv [Preprint]. https://doi.org/10.1101/2024.12.25.630221.

3) Vorontsov, I.E., Kulakovskiy, I.V., Khimulya, G., Nikolaeva, D.D. and Makeev, V.J. (2015). PERFECTOS-APE - Predicting Regulatory Functional Effect of SNPs by Approximate P-value Estimation. In: Proceedings of the International Conference on Bioinformatics Models, Methods and Algorithms (BIOSTEC 2015). Lisbon, Portugal, 12-15 January 2015. SciTePress, pp. 102–108. https://doi.org/10.5220/0005189301020108.

## Иллюстрации

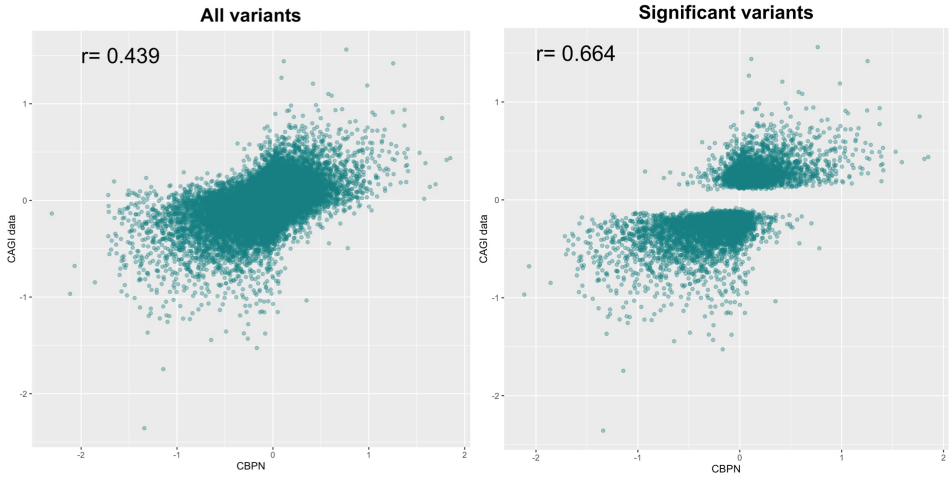


Рис. : ChromBPNet и данные MPRA.

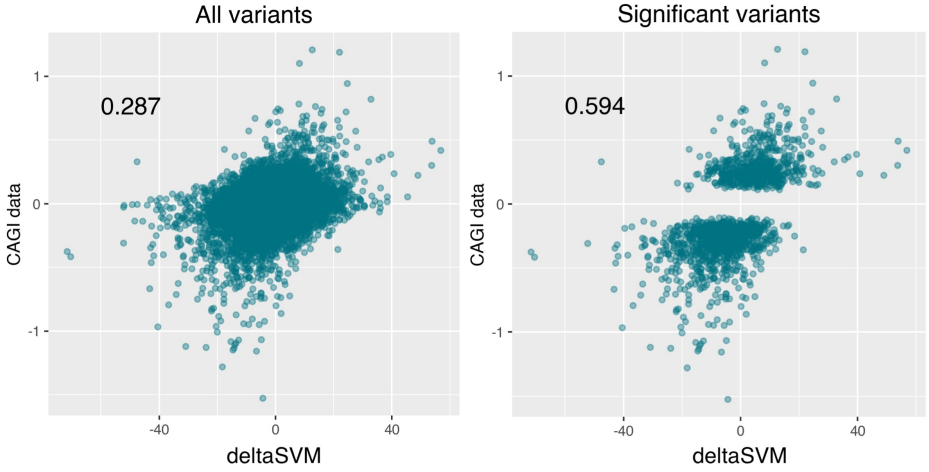


Рис. : deltaSVM и данные MPRA.

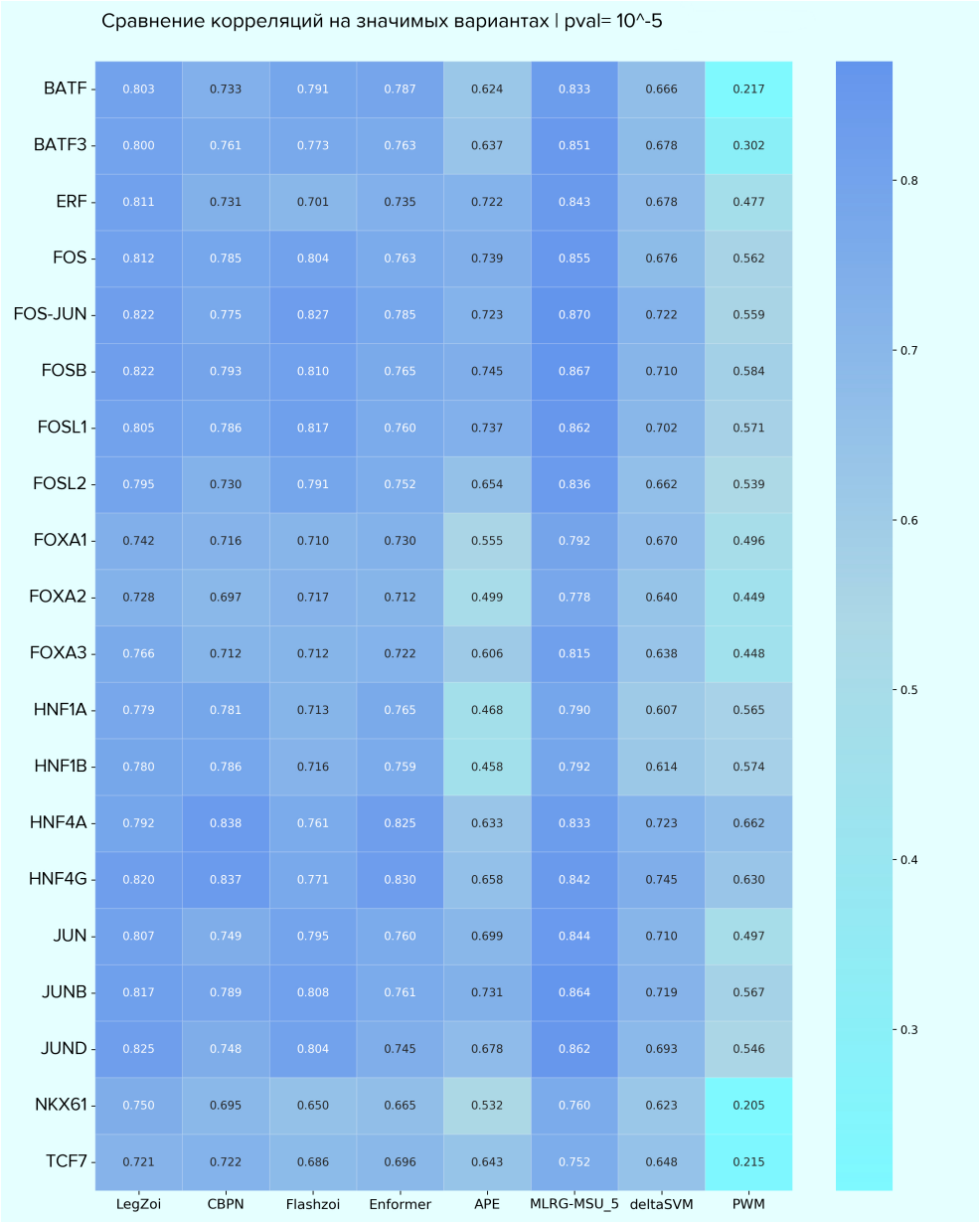


Рис. : сравнение корреляций на значимых вариантах