Секция «Биоинформатика»

База данных корреляций экспериментов ENCODE

Научный руководитель – Миронов Андрей Александрович

Абзалимов Амир Ришатович

Студент (специалист)

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

E-mail: amilim.general@gmail.com

С накоплением большого объема экспериментальных данных, полученных методами ChIP-seq, ATAC-seq, DNase-seq и другими, возникает необходимость в создании структурированной базы данных, которая позволит рассчитывать попарные корреляции и делать выводы о функциональной значимости и взаимодействиях хроматиновых меток. Проведение корреляционного анализа для всех экспериментов, имеющихся в базе данных ENCODE [1], поможет раскрыть полную картину эпигеномных взаимодействий, поэтому перед нами была поставлена задача построения базы данных корреляций экспериментов ENCODE. Также была поставлена задача разработки веб-интерфейса для визуализации данных и исследования полученных результатов.

Созданный веб-сервис позволяет изучать тепловую карту корреляций для различных хроматиновых меток внутри каждой клеточной линии, сравнивать попарно корреляции выравниваний и пиков, анализировать корреляции данных, нормированных на входные данные, оценивать уровень воспроизводимости реплик экспериментов, а также отображать значения различных статистических параметров и вспомогательной мета-информации о каждом эксперименте.

Анализ геномного распределения и интерпретация структуры взаимодействий осуществлялись с применением программного обеспечения StereoGene [2], обеспечивающего оценку попарных корреляций между хроматин-ассоциированными метками из различных экспериментов базы данных ENCODE. Расчет попарной корреляции, нормирование на входные данные, построение ковариационной матрицы для снижения шума и учет множества статистических параметров, реализованные в StereoGene, позволяют построить детальную карту предполагаемых взаимодействий, а также верифицировать существующие гипотезы о функциональной значимости хроматиновых меток.

В результате анализа экспериментальных данных Histone ChIP-seq для клеточных линий K562, MCF-7 и IMR-90 были получены результаты, которые согласуются с уже имеющимися данными о функциях различных хроматиновых меток, однако в ряде случаев выявлены расхождения.

На данный момент разработан веб-интерфейс на языке программирования JavaScript с использованием библиотек React и Plotly и построена база данных корреляций экспериментов TF ChIP-seq базы ENCODE. Спектр используемых методов был расширен: помимо тепловых карт корреляций, применяется анализ воспроизводимости технических и биологических повторов экспериментов с использованием метода IDR [3], что позволяет оценить корректность данных.

Источники и литература

1 ENCODE portal (Sloan et al. 2016) (https://www.encodeproject.org/)

2 Elena D Stavrovskaya, Tejasvi Niranjan, Elana J Fertig, Sarah J Wheelan, Alexander V Favorov, Andrey A Mironov, StereoGene: rapid estimation of genome-wide correlation of

continuous or interval feature data, Bioinformatics, Volume 33, Issue 20, October 2017, Pages 3158–3165, https://doi.org/10.1093/bioinformatics/btx379

3 Li, Qunhua, et al. “MEASURING REPRODUCIBILITY OF HIGH-THROUGHPUT EXPERIMENTS.” The Annals of Applied Statistics, vol. 5, no. 3, 2011, pp. 1752–79. JSTOR, DOI:10.1214/11-AOAS466!