Секция «2.2 Биоинформатика»

Создание базы данных корреляций экспериментов ENCODE для изучения эпигеномных взаимодействий

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

База данных ENCODE[4] содержит более 23 000 функциональных геномных экспериментов, включая ChIP-seq, ATAC-seq и DNase-seq. Систематический анализ пространственных корреляций между эпигенетическими метками позволяет выявить функциональные взаимодействия хроматиновых меток и регуляторных элементов в различных клеточных линиях. Ранее нами была разработана концепция интегративного анализа корреляций для ограниченного набора клеточных линий (K562, MCF-7, IMR-90), где были выявлены закономерности взаимодействия меток открытого и закрытого хроматина, а также обнаружены клеточно-специфичные особенности, такие как активность бивалентного домена в эмбриональной линии IMR-90. В настоящей работе представлена полномасштабная реализация подхода с масштабированием на весь массив экспериментов ENCODE.

Для расчёта попарных корреляций использовалось приложение StereoGene[1], позволяющее оценивать пространственное сходство геномных экспериментов с учётом нормализации на контрольные данные и множественного тестирования. Анализ был масштабирован на более чем 20 000 экспериментов, включая анализ гистоновых модификаций, транскрипционных факторов, профилей открытости хроматина и CAGE экспериментов. Для оценки качества данных внедрён анализ воспроизводимости биологических и технических реплик методом IDR (Irreproducible Discovery Rate)[3], который широко применяется в стандартах ENCODE для контроля согласованности результатов между репликами. Разработан публичный веб-портал StereoGeneDB, позволяющий просматривать тепловые карты корреляций для выбранной клеточной линии, сравнивать корреляции между пиками и сырыми данными, оценивать воспроизводимость реплик, а также выполнять сравнение паттернов корреляций между различными типами клеток.

Анализ расширенного набора данных позволил выявить следующие закономерности: при анализе ядерных корреляций в окнах генома фиксированного размера, учитывается пространственный контекст. Кроме того, при подсчете корреляций между ранее посчитанными корреляционными профилями по геномным окнам, определяются пространственно сближенные фрагменты хроматина и регуляторные комплексы белков и РНК. Это позволяет предсказывать топологически ассоциированные домены и функциональные кластеры хроматина. В отдельных типах клеток при этом выявляются изменения регуляторной архитектуры, организации хроматина, состава ассоциированных белков, уровня открытости хроматина и экспрессии генов. Данный механизм был применен для изучения изменений организации хроматина иммунных клеток при активации механизма иммунной памяти, для изучения изменений регуляторного ландшафта нервных клеток при болезни Паркинсона и для изучения траекторий развития регуляторной структуры стволовых клеток человека.

StereoGeneDB представляет собой первый систематический ресурс полногеномных корреляций эпигенетических данных ENCODE с открытым доступом. Портал может быть использован для генерации гипотез о функциональных взаимодействиях хроматиновых

меток, валидации предсказаний регуляторных элементов и интеграции с данными GWAS и eQTL-исследований. Данный ресурс можно использовать для построения функциональной аннотации хроматина по аналогии с алгоритмом разметки ChromHMM[2].

Источники и литература

1) 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

2) Ernst, J., Kellis, M. Chromatin-state discovery and genome annotation with ChromHMM. Nat Protoc 12, 2478–2492 (2017). https://doi.org/10.1038/nprot.2017.124

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

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