

Методы анализа дифференциальной экспрессии генов на основе RNA-SEQ с нормализацией TMM

Научный руководитель – Майка Алексей Николаевич

Лобанов Владислав Константинович

Студент (бакалавр)

Воронежский государственный университет инженерных технологий, Воронежская область, Россия

E-mail: vladl9474@gmail.com

В условиях стремительного развития молекулярной биологии точный анализ экспрессии генов становится критически важным для диагностики заболеваний и разработки новых терапевтических стратегий. Использование RNA-Seq обеспечивает высокую точность и чувствительность при количественной оценке уровней экспрессии генов в различных биологических состояниях [1].

Нормализация данных является ключевым этапом анализа, устраняющим технические артефакты и обеспечивающим корректное сравнение образцов. Среди методов нормализации выделяются: TMM (edgeR), корректирующий различия между образцами и устойчивый к выбросам; RPKM/FPKM, учитывающие длину транскрипта; TPM, подходящий для внутригруппового сравнения; DESeq2, использующий медианные соотношения с автоматической коррекцией дисперсий; edgeR, применяющий метод TMM с байесовской оценкой дисперсий. После тщательного анализа для данного исследования выбраны TMM и DESeq2 как наиболее эффективные методы, обеспечивающие высокую чувствительность и статистическую мощность.

Для выявления генов с изменённой экспрессией применяются t-тест с поправкой Уэлча, который сравнивает средние значения экспрессии между группами, учитывая неравенство дисперсий а также алгоритмы DESeq2. DESeq2 основан на обобщённых линейных моделях и использовании негативного биномиального распределения, что делает его особенно эффективным при малом числе образцов.

Результаты анализа визуализируются через Volcano Plot (рис. 1), отражающий fold change и $-\log_{10}$ p-значение, и MA-plot, показывающий соотношение изменения экспрессии к среднему уровню [4]. Volcano Plot позволяет быстро выделить биологически значимые гены, а MA-plot помогает выявить тренды в данных, оценить влияние нормализации и обнаружить гены с высокой изменчивостью экспрессии.

Дополнительную ценность исследованию придаёт интеграция методов машинного обучения. Применение алгоритмов случайного леса, метода опорных векторов и методов понижения размерности [4] позволяет выявлять скрытые паттерны в больших наборах данных и выделять прогностические биомаркеры [2].

Практическая значимость исследования подтверждается применением описанных методов на масштабном наборе данных GSE289146 (305 образцов: 210 от пациентов с психиатрическими расстройствами и 95 от здоровых доноров). Разработанный комплексный подход демонстрирует эффективность в изучении молекулярных основ психических заболеваний, открывая новые перспективы в их диагностике и прогнозировании.

Источники и литература

- 1) "Bioinformatics Programming Using Python" (Mitchell L. Model)
- 2) "Machine learning for RNA sequencing-based intrinsic subtyping of breast cancer" (Scientific Reports, 2020)

- 3) "t-Distributed Stochastic Neighbor Embedding (t-SNE)" (van der Maaten & Hinton, 2008)
- 4) Scanpy Tutorial (<https://scanpy-tutorials.readthedocs.io>)

Иллюстрации

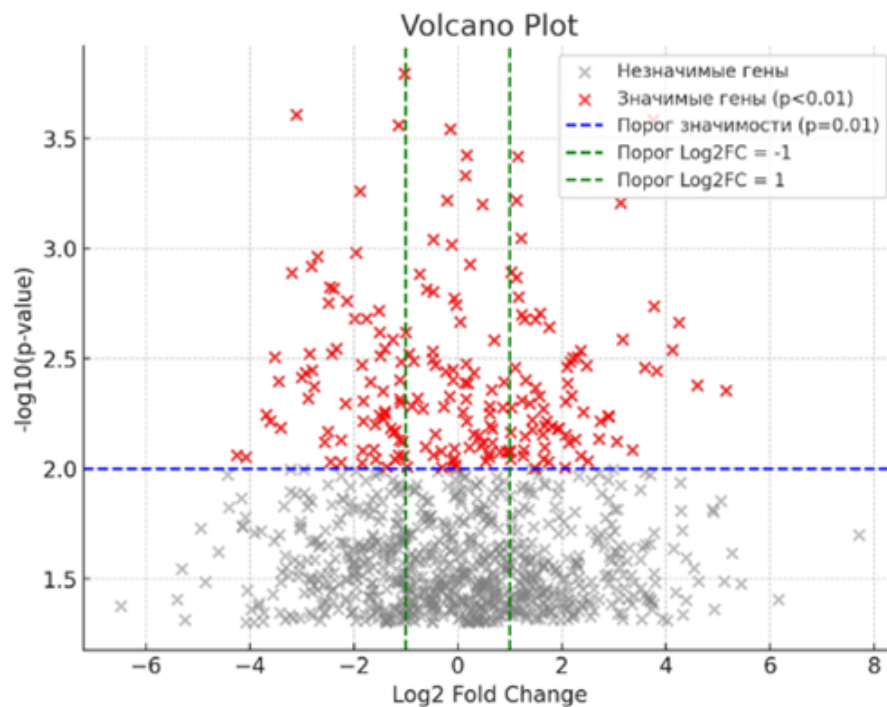


Рис. : 1 – пример Volcano Plot