МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ
УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ «НОВОСИБИРСКИЙ
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ГОСУДАРСТВЕННЫЙ
УНИВЕРСИТЕТ» (НОВОСИБИРСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ,
НГУ)

Механико-математический факультет
Кафедра дискретной математики и информатики
Направление подготовки «Математика и компьютерные науки»,
бакалавриат

**ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА
БАКАЛАВРА**

Емелина Антона Германовича

Применение вычислительных технологий RNA-Seq в биологии

«К защите допущена»
Заведующий кафедрой
д.ф.-м.н.
Гончаров С. С. / _____
«____» _____20 ____г.

Научный руководитель
д.б.н.,
Орлов Ю.Л. / _____
«____» _____20 ____г.

Дата защиты: «____» _____20 ____г.

Новосибирск
2023

Реферат

Название работы: Применение вычислительных технологий RNA-Seq в биологии
Количество страниц:
Количество рисунков:
Количество таблиц:
Количество использованных источников:
Ключевые слова:
Краткое описание

Содержание

1. Введение

Транскрипция – один из основных процессов, происходящих в клетке. В ходе этого процесса генерируется цепочка РНК, которая представляет собой последовательность из четырёх азотистых оснований (нуклеотидов), закодированных в буквенном виде.

RNA-Seq (РНК секвенирование) — новый революционный метод, позволяющий восстановить последовательности РНК и их количество[1]. Метод производит анализ транскриптома (совокупности всех молекул РНК, образующихся в ходе транскрипции), результатом которого являются нуклеотидные последовательности разной длины (от 50 до 200 пар нуклеотидов), называемые ридами. Метод помогает определить степень активности (экспрессию) генов, закодированных в ДНК. Количественные характеристики экспресии одного гена могут являться важным инструментом отслеживания влияния гена на функции других генов в целом организме.

Экспрессия гена показывает степень его проявления. Чем сильнее экспрессируется ген, тем больше белка синтезируется с закодированной в ДНК генетической информации. Исследование экспрессии генов осуществляется с помощью специализированной библиотеки программ Cufflinks, в которую входят используемые нами Cuffnorm и Cuffdiff. Cuffnorm вычисляет экспрессию в двух или более образцах, а Cuffdiff проверяет статистическую значимость каждого наблюдаемого изменения экспрессии между ними.

В работе использованы ДНК мыши и риды мышей с позитивным и негативным опытом, сформированным в агонистических социальных взаимодействиях: риды контрольной группы, групп агрессивных и подавляемых мышей.

Объектом исследования являются физиологические особенности особей мышей, предметом - применение вычислительных технологий к выборке и анализ полученных данных.

Цель исследовани: определить различия в экспрессии генов между агрессивными и депрессивными мышами на основе данных RNA-Seq. При достижении цели возникают задачи детализации методов, применяемых при работе с данными RNA-Seq, изучения и использования программы STAR, программы для выявления статистической значимости Cufflinks, и задача применения методов анализа к полученным данным с помощью Python и R. Также дополнительно присутствует задача выявления дифференциально экспрессируемых генов и последующей кластеризации для определения отличающихся особей.

Выдвигается предположение о том, что полученная информация после применения биологической интерпретации может говорить о корреляции между агрессивным поведением и выработкой гормона счастья дофамина.

Полная схема нашей работы следующая: подготовка данных с помощью программы STAR, статистический анализ с помощью Cufflinks и языков программирования Python и R.

Теоретическая значимость работы состоит в изучении работы статистических инструментов анализа данных, полученных в процессе RNA-Seq. Практической значимостью являются рекомендации связанные с особенностями выборок особей, приводящими к трудностям и неточностям вычислений.

Основные положения исследования были изложены в форме доклада на 61 научно-практической конференции МНСК-2023 и опубликованы в форме тезисов в материалах конференции.

2. Используемая терминология

Определение 1. *Транскрипт — молекула РНК, образующаяся в результате транскрипции*

Определение 2. *Нуклеотид — одно из четырёх азотистых оснований, закодированное в буквенном виде (A, T, G, C)*

Определение 3. *Рид — последовательность нуклеотидов длиной от 50 до 200 нуклеотидов, синтезируемых в процессе работы RNA-Seq*

3. Работа в программе STAR

Секвенирование РНК — метод определения первичной структуры молекул РНК, представляющий собой высокочувствительный и точный инструмент для изучения транскриптома. STAR — программа для работы с ридами, полученными после RNA-Seq[1]. С её помощью нам удалось провести два этапа подготовки данных перед статистическим анализом.

Выравнивание последовательностей — биоинформатический метод, основанный на размещении двух или более последовательностей РНК друг под другом таким образом, чтобы легко увидеть сходные участки в этих последовательностях. Сходство первичных структур двух молекул может отражать их функциональные, структурные или эволюционные взаимосвязи. Выровненные последовательности оснований нуклеотидов или аминокислот обычно представляются в виде строк матрицы. Добавляются разрывы между основаниями таким образом, чтобы одинаковые или похожие элементы были расположены в следующих друг за другом столбцах матрицы.

Индексация генома заключается в преобразовании исходной последовательности ДНК, представляющей из себя строку длиной примерно в 5 млрд. символов, которым соответсвует нуклеотид, в хеш-таблицу для облегчения дальнейшей работы с ней.

Картирование ридов заключается в поиске и сопоставлении совпадающих участков ридов и генома. Полученная геномная карта является основным источником данных, необходимых для последующего анализа.

Сначала необходимо провести индексацию генома, затем картирование выборок на полученный референсный геном.

3.1. Индексация генома: хеш-таблица

В приложениях, использующихся в биоинформатике, геномы обычно представляются не как линейная последовательность нуклеотидов, а как индексированные структуры данных, которые могут облегчить различные анализы, такие как быстрое выравнивание ридов.

Одним из основных методов индексирования предварительной обработки геномов для эффективного выравнивания является хэш-таблица. Хэш-таблица представляет последовательность генома в виде нескольких списков геномных позиций, по одному списку для каждого возможного k-мера (последовательности) некоторой предварительно выбранной длины k[2].

Хеш-таблица — это структура данных, реализующая интерфейс ассоциативного массива, а именно, она позволяет хранить пары (ключ, значение) и выполнять три операции: операцию добавления новой пары, операцию удаления и операцию поиска пары по ключу[3].

Выполнение операции в хеш-таблице начинается с вычисления хеш-функции от ключа. Получающееся хеш-значение $i = hash(key)$ играет роль индекса в массиве H. Затем выполняемая операция (добавление, удаление или поиск) перенаправляется объекту, который хранится в соответствующей ячейке массива $H[i]$.

Хэш-таблицы используются различными программами для выявления совпадений коротких последовательностей (или сид-фрагментов) между ридом и геномом. Затем эти сид-фрагменты можно комбинировать или расширять для получения более полного выравнивания. Хэш-таблицы особенно полезны для выравнивания ридов, содержащих множественные несовпадения или вставки относительно генома.

В геномике хэш-таблицы обычно реализуются как простая справочная таблица, в которой массив смещений содержит указатели на массив позиций для совокупности возможных k-меров. Эта простая реализация таблицы возможна из-за фиксированного и относительно небольшого значения k.

Определим граф, вершины которого - это все подпоследовательности длины l, рёбрами соединены пары подпоследовательностей, отличающихся одним нуклеотидом[4]. Граф имеет $4l$ вершин и каждая из них имеет $3l$ рёбер. Разделим $4l$ вершин графа на $4l/(3l+1)$ классов эквивалентности таким образом, что каждый класс имеет центральную вершину и $3l$ смежных вершин. Затем используем центральную вершину как хеш-ключ для хранения позиций в геноме для $3l+1$ подпоследовательностей в хеш-таблице (см. рис. 1).

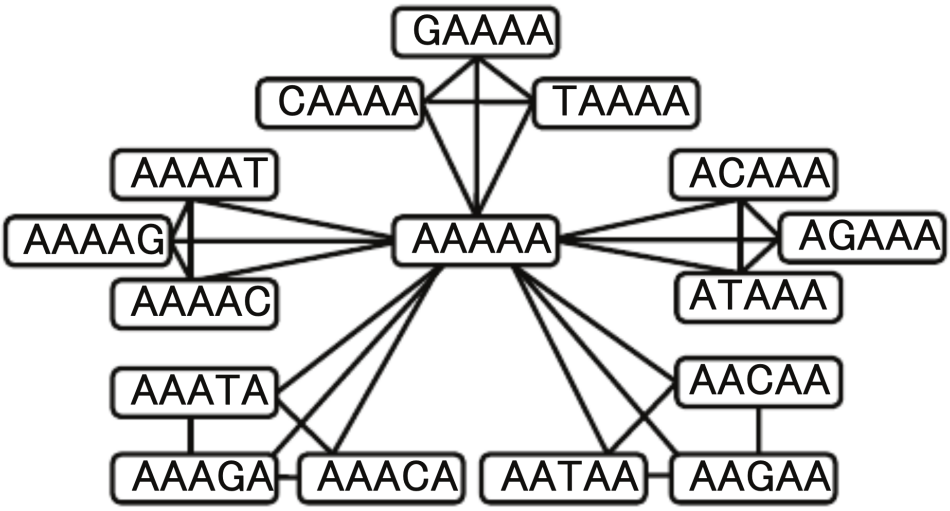

Рис. 1: Визуализация подпоследовательности “AAAAA” в виде 15 смежных подпоследовательностей. Каждая вершина описывает подпоследовательность, и каждое ребро обозначает разницу в один нуклеотид. 15 подпоследовательностей разделены на 5 групп в зависимости от позиции отличающегося нуклеотида.

На вход программы принимается файл формата .fasta с полным ДНК мыши (последовательность нуклеотидов) и вспомогательный .gtf файл аннотации для определения участков генома. Выбирается режим «genomeGenerate» и путь для расположения папки с результатами индексации. На выходе получаем файлы, подробно описывающие референсный геном.

Файлы генома содержат бинарную последовательность генома, массивы суффиксов, текстовые названия/длины хромосом, координаты соединений сплайсинга и информацию о транскриптах/генах. Большинство этих файлов используют внутренний формат STAR

и не предназначены для использования конечным пользователем.

3.2. Картирование выборок: поиск в хеш-таблице

Картирование является процессом наложения ридов на индексированный референсный геном. Программа ищет совпадения участков последовательностей ридов и участков генома. По своей сути процесс представляет собой поиск подстроки в строке. Таким образом, мы можем понять структуру генной сети и определить кодирующие участки — экзоны, и участки, не имеющие значимость, — интроны.

Картирование осуществляется в два этапа: поиск сид-фрагментов и кластеризация/склеивание/подсчёт[5].

Основной идеей фазы поиска сид-фрагментов является последовательный поиск Максимального Картируемого Префикса (Maximal Mappable Prefix)(MMP). MMP является аналогом концепта Максимального Уникального Совпадения (Maximal Exact (Unique) Match), используемого в широкомасштабном инструменте для выравнивания генома Mummer[6]. $MMP(R, i, G)$ определяется как наиболее длинная подстрока $(R_i, R_{i+1}, \ldots, R_{i+\mathrm{MML}-1})$, которая совпадает не менее чем с одной подстрокой из G, где R - последовательность рида, i - расположение рида, G - референсная последовательность генома, MML - максимальная картируемая длина.

На первом этапе алгоритм находит MMP, начиная с первой базы рида. Затем поиск MMP повторяется для неотображенной части рида. Этот подход представляет собой естественный способ нахождения точных местоположений сплайс-соединений в последовательности считывания и имеет преимущество перед произвольным разделением последовательностей считывания, используемым в методах разделенного считывания. Места сращивания обнаруживаются за один проход выравнивания без каких-либо априорных знаний о локусах или свойствах соединений сплайсинга, а также без предварительного непрерывного прохода выравнивания, необходимого для подходов к базе данных соединений. MMP в поиске STAR реализуется через массивы несжатых суффиксов (SA)[7]. Примечательно, что поиск MMP является неотъемлемым результатом поиска стандартной двоичной строки в несжатых SA и не требует дополнительных вычислительных усилий по сравнению с поиском точного совпадения полной длины. Двоичный характер поиска SA приводит к благоприятному логарифмическому масштабированию времени поиска с длиной эталонного генома, что позволяет выполнять быстрый поиск даже в отношении больших геномов. Преимущественно, для каждой MMP поиск SA может найти все отдельные точные геномные совпадения с небольшими вычислительными затратами, что облегчает точное выравнивание ридов, которые сопоставляются с несколькими геномными локусами.

На вход программы принимается путь к папке с результатами индексации генома и файлы формата .fasta, содержащие риды для девяти особей мышей. Используется по три особи для каждой рассматриваемой группы: агрессивные, подавляемые и контрольная группа. На выходе имеем файл формата .bam с информацией о картировании. Формат BAM является бинарным эквивалентом формата SAM (Sequence alignment map), который

разработан специально для больших объёмов данных, полученных в процессе картирования.

4. Обработка данных в Cufflinks

Для дальнейшей работы с картированными данными нам понадобится пакет Cufflinks, чтобы определить экспрессию генов и найти дифференциально экспрессируемые гены.

Экспрессия гена - это степень его проявления в геноме. Она определяется количественный содержанием гена в геноме[8]. Если ген сильно экспрессируется, то его влияние будет сильнее выражено на фоне влияния других генов на организм.

Метод секвенирования РНК становится основным методом определения того, какие гены и на каком уровне экспрессируются в клетке. С помощью РНК секвенирования можно определять различия в экспрессии генов на различных стадиях развития организма или в разных тканях.

Экспрессия транскрипта может быть измерена с помощью подсчёта количества фрагментов, сгенерированных им[11]. Изменение уровня экспрессии измеряется через сравнение количества фрагментов в каждом состоянии. Транскрипт считается значительно дифференциально экспрессированным, если вероятность увидеть изменения в подсчёте достаточно маленькая в соответствии с подходящей статистической моделью.

Cuffnorm вычисляет экспрессию в двух или более образцах в единицах измерения fpkm[10] (фрагменты на килобазу на миллион картированных прочтений), а Cuffdiff проверяет статистическую значимость каждого наблюдаемого изменения экспрессии между ними.

Cuffnorm принимает на вход девять имеющихся выборок ридов мышей и возвращает файл с величинами экспрессии для каждой выборки в единицах измерения fpkm (фрагменты на килобазу на миллион картированных прочтений).

Cuffdiff принимает на вход вспомогательный файл с аннотацией по геному и от двух выборок, по которым будет проводиться сравнение. Нам нужно попарно сравнить три группы мышей: контрольную и подавляемую, контрольную и агрессивную, подавляемую и агрессивную.

Cuffdiff на основании вариации 3 выборок в каждой группе оценивает достоверность отличия величины экспрессии конкретного гена между анализируемыми группами (p_value), с учетом поправки на множественное сравнение (24 тысячи генов) (q_value).

5. Статистический анализ

Рассматриваются данные после сравнительного анализа в Cuffdiff и выбираются статистические значимые гены.

Сравниваются 2 группы с помощью t-теста, если отклонение значимое (более 0,5), то говорится о достоверной разности в экспрессии данного гена. Выбираются гены только со значимым отклонением.

t-критерий для двух независимых выборок (двухвыборочный t-критерий) проверяет гипотезу о равенстве средних в двух выборках (предполагается нормальность распределения переменных, а также равенство дисперсий выборок).

Из результатов работы Cuffnorm берём значения для найденных дифференциально экспрессируемых генов, измеряемые в fpkm.

Получаем матрицу значений fpkm для каждой особи по каждому из генов. Обозначим особей контрольной группы – C, депрессивных – L, агрессивных – W. Это необходимо для последующего анализа.

Цель работы – выявление различий в работе генов стриатума у агрессивных и депрессивных мышей по сравнению с контрольной группой. Для этого мы сосредоточились на наиболее высоко экспрессирующейся генной сети cAMP, связанной с поведением.

Циклический аденозин монофосфат (cAMP) — это каскад реакций возбуждения и сдерживающих реакций в средних шипиковых нейронах стриатума. В него входит 33 гена. Этот каскад ответственен за поведение в стриатуме, поэтому мы выделяем эти гены.

Из полученных генов в матрице оставляем только 33 гена, принадлежащих каскаду, в результате имеем матрицу 9 особей на 33 гена, которую можно подвергнуть статистическому анализу.

5.1. Cuffdiff: разделение смеси распределений

Нужно построить вероятностную модель для предсказания числа фрагментов, которая учитывает биологическую изменчивость[9].

Моделируется изменчивость количества фрагментов для каждого гена. Количество фрагментов для каждой изоформы оценивается в каждой выборке вместе с мерой неопределенности в этой оценке, возникающей из-за неоднозначно картированных ридов.

Используемые переменные:

- T — множество транскриптов
- N — общее число сгенерированных участков
- $\{C_t\}_{t \in T}$ — число участков, сгенерированных с транскрипта t (являются функциями от ρ_t и N)
- ρ_t — экспрессия транскрипта t
- $\rho = \{\rho_t\}_{t \in T}$ — множество величин экспрессии транскриптов

- x_t — число наблюдаемых участков с транскрипта t
- F — множество фрагментов
- $S_f \subseteq 2^T$ — множество транскриптов, на которые картируется фрагмент f
- $\mathbf{S} = \{S_f\}_{f \in F}$ — наблюдаемые транскрипты

Для начала возьмём модель Пуассона[12] для описания числа фрагментов. Но присутствует проблема чрезмерной дисперсии, которую можно решить, преобразовав распределение. Также необходимо учитывать неопределённость в определении транскрипта.

$$C_t \in \Pi(\rho_t N) \tag{1}$$

$\sum_{t \in T} \rho_t = 1$, $\hat{\rho}_t = \frac{x_t}{N}$ — оценка максимального правдоподобия для величины экспрессии.

Число экзонов имеет отрицательное биномиальное распределение $X_t \in NB(r_t, p_t)$ с функцией вероятности

$$\sum_t \mathbb{P}(C_t = x_t | \mathbf{S}) \frac{e^{-x_t} x_t^k}{k!}, \tag{2}$$

где $Y_t = p(C_t|\mathbf{S})$ — условная вероятность, что фрагменты произошли из набора транскриптов $\mathbf{S}$, она имеет гамма распределение $Y_t \in \Gamma(r, \theta)$.

Параметры распределения:

$$r_t = \frac{m_t^2}{\psi_t}, \quad p_t = \frac{m_t}{m_t + \psi_t}, \tag{3}$$

где m_t - начальный момент (матожидание) Y_t, ψ_t - центральный момент (дисперсия) Y_t.

Полученная модель не учитывает неопределённость, поэтому вместо импользования смеси распределения Пуассона и отрицательного биномиального распределения мы стремимся к использованию смеси отрицательных биномиальных распределений, которая моделируется бета отрицательным биномиальным распределением.

Чтобы найти параметры распределения, необходимо найти оценки вариабильности между образцами (в нашем случае один образец) и неопределённости в подсчёте числа фрагментов.

Далее будем считать, что C_t имеют отрицательно биномиальное распределение, а не распределение Пуассона.

Число фрагментов в транскрипте t моделируется с помощью отрицательного биномиального распределения: $x_t \in NB(\mu_t, \sigma_t^2)$.

Оценки дисперсии данного распределения получаются путём подстановки обобщённой линейной модели (ОЛМ) гамма семейства в cross-replicate (матожидание, дисперсия) пары. ОЛМ параметризует NB распределение, что позволяет моделировать число фрагментов. Это реализуется с помощью пакета LOCFIT, откуда берётся функция $V_E(N)$, возвращающая предсказанное значение дисперсии количества фрагментов среди выборок, принимающая на вход значение среднего подсчётов N.

Чтобы оценить число фрагментов для каждой отдельной изоформы, используется функция правдоподобия:

$$L(\rho|\mathbf{S}) = \left(\prod_{g\in G} \beta_g^{X_g}\right)\left(\prod_{g\in G}\prod_{f\in F_g}\sum_{t\in g} \gamma_t \cdot Q(f,t)\right) \tag{4}$$

Модель разделяет экзоны и транскрипты на неперекрывающиеся участки $G = g_1, ..., g_k$, и разделяет каждый параметр экспрессии транскрипта ρ_t на два β_g и γ_t, таким образом $\rho_t = \beta_g\gamma_t$.

Параметр β_g — это вероятность того, что случайный фрагмент из библиотеки попадёт в участок g, параметр γ_t — это вероятность того, что фрагмент из участка g, содержащего множество транскриптов T_g, происходит из транскрипта $t \in T_g$. $F_g \subseteq F$ — множество фрагментов, которые картируются на участок g, и $Q(f,t)$ является константой, которая включает в себя ряд эффектов нормализации для каждого фрагмента, смоделированных алгоритмом.

$X_g = \sum_{f\in F_g} m_f \cdot \frac{1}{n_f}$, где $m_f = \frac{1}{n}$ (n — число участков) — константа масштабирования, n_f — внешний коэффициент размера f. С помощью этого мы можем получить оценку относительной экспрессии ρ_t.

Чтобы учитывать неопределённость в нашей модели, необходимо определить условные вероятности для количества фрагментов. Мы считаем γ_t для каждого транскрипта с помощью EM алгоритма (Expectation-maximization — алгоритм для нахождения оценок максимального правдоподобия параметров вероятностных моделей). Далее вычисляем условное матожидание для числа фрагментов для каждого транскрипта.

В нашей работе рассматривается частный случай с одним локусом и двумя транскриптами:

$$L(\rho|\mathbf{S}) = \prod_{f\in F_g}\Big(\gamma_1 \cdot Q(f,1) + \gamma_2 \cdot Q(f,2)\Big), \tag{5}$$

g — локус, γ_1, γ_2 — вероятности, что экзон произошёл из транскрипта 1 или 2 соответственно.

Каждый фрагмент рассматриваем как множество случайных величин, имеющих распределение Бернулли, одна случайная величина для одного транскрипта из S_f, с вероятностью успеха:

$$p_f^t = \frac{\gamma_t Q(f,t)}{\sum_{i\in S_f}\gamma_i Q(f,i)} \tag{6}$$

Условное матожидание для экзонов, соответствующих транскрипту t в участке g вычисляется следующим образом $\sum_{f\in F_g} p_f^t$ по формуле полного матожидания.

Дисперсия и ковариация для условного распределения фрагментов между транскриптами i и j прослеживаются в матрице ψ, где элементы вычисляются следующим образом:

$$\psi_{i,j} = \sum_{f\in F_g}\psi_{f_i,j} = \begin{cases} p_f^i(1-p_f^i) & \text{if i = j} \\ -p_f^i p_f^j & \text{if i}\neq\text{j} \end{cases} \tag{7}$$

После того как мы получили вариабильность между образцами и неопределённость в числе фрагментов для транскриптов, мы можем объединить их в одно смешанное распределение.

Предполагаем, что экспрессия транскрипта ρ_t моделируется с помощью переменной X_t, имеющей бета отрицательное биномиальное распределение. Бета отрицательное биномиальной распределение может быть рассмотрено как смешанное распределение, состоящее из отрицательных биномиальных. Его параметры r, α, β определяются путём решения трёх уравнений.

Для простоты обозначим $p = \frac{\alpha-1}{\alpha+\beta-1}$, также введём

$$A = X_g\hat{\gamma}_t, \quad B = V_E(X_g)\cdot\hat{\gamma}_t, \quad C = \psi^g_{t,t} \tag{8}$$

Получаем уравнения:

$$\frac{r(1-p)}{p} = A \tag{9}$$

$$\frac{r(1-p)}{p^2} = B \tag{10}$$

$$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} = \frac{A^4}{B^4}\cdot\frac{C}{r^2} \tag{11}$$

(9) и (10) определяют отрицательное биномиальное распределение для значения матожидания и дисперсии параметра p, которые имеют бета распределение и совпадают с избыточной дисперсией, оценённой для данного оценённого значения количества. (10) получено из того факта, что нам бы хотелось, чтобы среднее значение для отрицательного биномиального распределения изменялось в соответствии с неопределённостью в назначении транскрипта, полученной из $\psi^g_{t,t}$.

Из (9) и (10) получаем: $r = \frac{A^2}{B-A}$, $p = \frac{A}{B}$.

Т.к. r неотрицательно, то $B > A$, что эквивалентно предположению о присутствии избыточной диспресии. Если $B = A$, то дисперсия равна матожиданию, и мы можем вернуться к смеси распределения Пуассона и отрицательного биномиального распределения. Аналогично действуем, если нет неопределённости в числе экзонов.

Далее получаем: $\alpha = 1 - \frac{A}{A-B}\cdot\beta$.

Подставляя полученное уравнение в (11) и используя выражение для r, получаем кубическое уравнение, из которого получаем значение для β. Таким образом находим все параметры для бета отрицательного биномиального распределения.

Распределение для количества фрагментов для каждого транскрипта позволяет нам оценивать значимость в видимых изменениях между двумя или более состояниями.

5.2. Тепловая карта

Самоорганизующаяся карта (self-organizing map) - технология машинного обучения, используемая для создания репрезентации данных большой размерности с более низ-

кой размерностью (в основном двухмерной), с сохранением топологических особенностей данных[13]. Например набор данных с p переменными, измеренными в n наблюдениях, может быть представлен в виде кластеров наблюдений с одинаковыми значниями переменных. Эти кластеры в последствие визуализируются в виде двухмерной «карты» таким образом, что центральные кластеры имеют больше общих значений, чем периферийные. Данный метод облегчает визуализацию и анализ данных с большой размерностью.

Самоорганизующиеся карты, как многие нейронные сети, работают в двух режимах: обучающий и картирование. Обучение использует набор данных, чтобы сгенерировать низкоразмерное представление входных данных. Картирование классифицирует дополнительные входные данные, используя сгенерированную карту.

Используемые переменные:

- s - текущая итерация
- λ - общее число итераций
- t - индекс целевого вектора входных данных во входящем датасете $\mathbf{D}$
- *D(t)* - целевой вектор входных данных
- v - индекс узла карты
- $\mathbf{W}_v$ - текущий вектор весов узла v
- u - индекс наилучшего подходящего элемента (НПЭ) в карте
- $\theta(u, v, s)$ - ограничение, связанное с расстоянием от НПЭ, также называемое функцией соседства
- $\alpha(s)$ - ограничение обучения, связанное с ростом итераций

Алгоритм генерации карты:

1. Случайный выбор весовых векторов для узлов карты
2. Случайно выбирается входной вектор *D(t)*
3. Перебор каждого узла карты
   (a) Используется формула для поиска евклидового расстояния для нахождения схожих значний во входном векторе и векторе весов узла карты
   (b) Запоминается узел, который производит наименьшее расстояние (это НПЭ)
4. Обновляются весовые векторы для узлов, соседствующих с НПЭ, сдвигая их ближе к входному вектору
   (a) $W_v(s+1) = W_v(s) + \theta(u, v, s) \cdot \alpha(s) \cdot (D(t) - W_v(s))$
5. Увеличивается s и повторяется с шага 2 пока $s < \lambda$

Heatmap является методом визуализации кластеризации[14]. Значения данных представляются в виде цвета. Сначала проводится кластеризация столбцов и колонок матрицы данных. Столбцы и строки переставляются в соответствии с кластеризацией. Последним шагом является визуализация данных. Результатом является информация о переменных, характерных для каждого кластера.

Замеченные аутлаеры – W3 (агрессивный) и L2 (подавляемый). Эта конкретная выборка была не очень стабильная по словам ее создателей, поэтому отличия на уровне транскриптома были существенные.

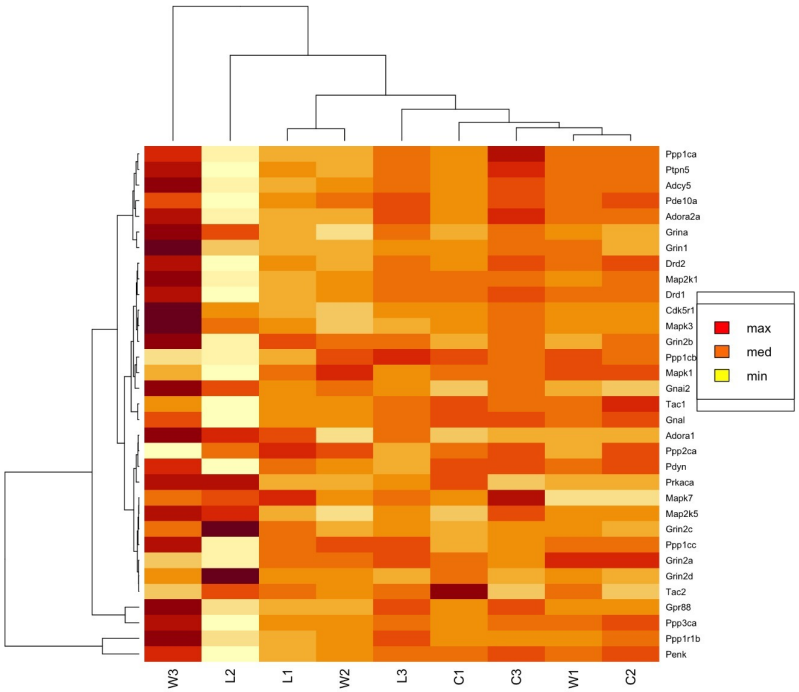

Рис. 2: Построение тепловой карты методом кластеризации 33 генов с помощью языка программирования для статистических вычислений R. По вертикали – кластеры 33 генов, по горизонтали – выборки (3 особи контроля (C), 3 депрессивных (L), и 3 агрессивных (W)

5.3. Метод главных компонент

Метод главных компонент является эвристическим методом, позволяющим уменьшить размерность данных. Он используется для выявления зависимостей между имеющимися признаками. Метод работает с минимальной потерей информации[15].

В качестве исходных данных имеем матрицу 9x33 со значениями fpkm для каждого гена по каждой из девяти особей. Рассматриваем её как многомерный вектор.

Вектор должен проходить через центр имеющейся выборки для минимизации потерь информации. Поэтому мы отцентрируем выборку для упрощения дальнейших вычислений, то есть сделаем линейный сдвиг таким образом, чтобы средние значения признаков были равны 0.

Для описания формы случайного вектора необходима ковариационная матрица[16]. Это матрица, у которой (i, j) - элемент является корреляцией признаков (X_i, X_j).

$$Cov(X_i, X_j) = \mathbf{E}\big[(X_i - \mathbf{E}(X_i) \cdot ((X_j - \mathbf{E}(X_j))\big] = \mathbf{E}(X_i X_j) - \mathbf{E}(X_i) \cdot \mathbf{E}(X_j) \tag{12}$$

После этого надо найти вектор, при котором максимизируется размер (дисперсия) проекции нашей выборки на него. Возьмём ежиничный вектор, на который будем проецировать наш вектор. Проекция на него будет равна $\vec{v}^T X$. Дисперсия проекции:

$$Var(X^*) = \mathbf{E}(X^* \cdot X^{*T}) = \mathbf{E}\big((\vec{v}^T X) \cdot (\vec{v}^T X)^T\big) = \mathbf{E}(\vec{v}^T X \cdot X^T \vec{v}) = \vec{v}^T \mathbf{E}(X \cdot X^T)\vec{v} \tag{13}$$

Дисперсия максимизируется при максимальном значении $\vec{v}^T \mathbf{E}(X \cdot X^T)\vec{v}$. Воспользуемся отношением Рэлея[17] для случая ковариационных матриц:

$$R(M, \vec{x}) = \frac{\vec{x}^T M \vec{x}}{\vec{x}^T \vec{x}} = \lambda \frac{\vec{x}^T \vec{x}}{\vec{x}^T \vec{x}} = \lambda, \quad M\vec{x} = \lambda\vec{x} \tag{14}$$

Направление максимальной дисперсии у проекции всегда совпадает с собственными векторами, имеющими максимальное собственное значение.

Максимальные вектора имеют направление, схожее с линией регрессией, проекция выборки на него приводит к потере информации, сравнимой с суммой остаточных членов регрессии. Для проведения проекции надо провести операцию $\vec{v}TX$ (вектор должен быть длины 1). В нашем случае не вектор, а гиперплоскость, поэтому берем матрицу базисных векторов V^T. Полученная матрица является массивом проекций наших наблюдений.

Для результатов, полученных методом главных компонент, потребовалась дополнительная биологическая интерпретация, за которой мы обратились к научному руководителю[18].

Метод главных компонент располагает W3 в зоне с генами, отвечающими за выработку дофамина. Потребление дофамина связано с экспрессией эндогенного (внутриклеточного) опиоида в стриатуме (наркотик), поэтому агрессивные мыши ждут схваток с нетерпением. В качестве метрики сходства экспрессии генов использовался коэффициент корреляции Пирсона.

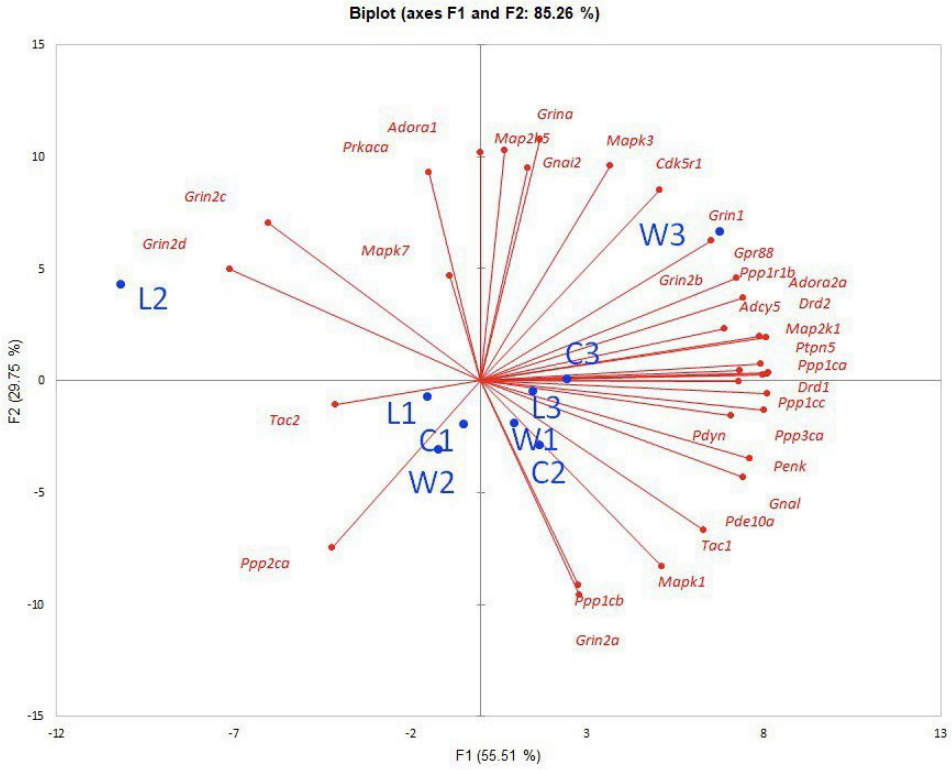

Рис. 3: Построение графика методом главных компонент, синим обозначены особи, красным - гены каскада реакций

6. Заключение

Были изучены методы, необходимые для работы с данными, полученными в процессе RNA-Seq, и выявлены принципы их работы.

Проведено картирование и анализ экспрессии генов дорсального стриатума 9 особей модели социальных конфликтов на мышах.

Выявлены 1021 дифференциально экспрессирующихся генов между группами L и C, 386 между W и C, и 720 между W и L.

Выявлены 33 гена cAMP цикла с помощью генной онтологии, и на базе их двумерной проекции произведено картирование особей.

Две особи (W2 и L2) оказались достоверно отличающимися от контроля и располагающимися в противоположных зонах нейромедиаторов: в зоне W2 выявлено активное потребление дофамина, который запускает сеть cAMP и характеризуется высокой двигательной активностью. Особь L2 оказалась в зоне нейромедиатора глутамата, который блокирует потребление дофамина и является характерным для депрессивного состояния.

Сделано заключение, что вариация профилей генов групп L, W достаточно широкая (см. рис. 3), и не дает четкой кластеризации групп, что говорит о сложностях с качественным анализом отражения поведения на генном уровне. Тем не менее, 2 особи L находятся

в зоне глутаматного нейромедиатора, 2 особи W находятся в зоне экспозиции дофамина (см. рис. 3). Особи контроля (C) находятся в области умеренного потребления дофамина.

Выдвигается рекомендация увеличения выборки до 6-10 особей в группах.

7. Список литературы

1. Wang Z., Gerstein M., Snyder M. RNA-Seq: a revolutionary tool for transcriptomics // Nat Rev Genet. 2009. №10(1). P. 57-63.

2. Wu T. D. Bitpacking techniques for indexing genomes: I. Hash tables // Algorithms Mol Biol. 2016. №11(5).

3. Кормен Т., Лейзерсон Ч., Ривест Р., Штайн К. Глава 11. Хеш-таблицы. // Алгоритмы: построение и анализ = Introduction to Algorithms / Под ред. И. В. Красикова. 2005. 1296 с.

4. Takenaka Y., Seno S., Matsuda H. Perfect Hamming code with a hash table for faster genome mapping // BMC genomics. 2011. №12(3). P. 8-10.

5. Dobin A., Davis C. A., Schlesinger F., Drenkow J., Zaleski C., Jha S., Batut P., Chaisson M., Gingeras T. R. STAR: ultrafast universal RNA-seq aligner // Bioinformatics. 2013. №29(1). P. 15–21.

6. Delcher AL, et al. Alignment of whole genomes // Nucleic Acids Res. 1999. №27. P. 2369–2376.

7. Manber U., Myers G. Suffix arrays — a new method for online string searches // SIAM J. Comput. 1993. №22. P. 935–948.

8. Хоукинс Дж. Структура и экспрессия гена = Gene Structure and Expression / Джон Хоукинс; Пер. с англ. С. Б. Серебряного; Под ред. В. К. Кибирева. — Киев: Наукова думка, 1991. — 168 с.

9. Trapnell C., Hendrickson D. G., Sauvageau M., Goff L., Rinn J. L., Pachter L. Differential analysis of gene regulation at transcript resolution with RNA-seq // Nat Biotechnol. 2013. №31(1). P. 10.

10. Trapnell C., Williams B., Pertea G., Mortazavi A., Kwan G., van Baren M. J., Salzberg S. L., Wold B. J., Pachter L. Transcript assembly and abundance estimation from RNA-Seq reveals thousands of new transcripts and switching among isoforms // Nat Biotechnol. 2010. №28(5). P. 511–515.

11. Trapnell C., Roberts A., Goff L., Pertea G., Kim D., Kelley D., Pimentel H., Salzberg S., Rinn J., Pachter L. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks // Nat Protoc. 2012. №7(3). P. 562–578.

12. Anjum A., Jaggi S., Varghese E., Lall S., Bhowmik A., Rai A. Identification of Differentially Expressed Genes in RNA-seq Data of Arabidopsis thaliana: A Compound Distribution Approach // J Comput Biol. 2016. №23(4). P. 239–247.

13. Kohonen T., HonkelaT. Kohonen Network // Scholarpedia. 2007. №2 (1). P. 1568.

14. Wilkinson L., Friendly M. The History of the Cluster Heat Map // The American Statistician. №63 (2). P. 179–184.

15. Айвазян С. А., Бухштабер В. М., Енюков И. С., Мешалкин Л. Д. Прикладная статистика. Классификация и снижение размерности. Финансы и статистика. 1989. 607 с.

16. Ledoit O., Wolf M. Nonlinear Shrinkage Estimation of Large-Dimensional Covariance Matrices // The Annals of Statistics. 2012. №40(2). P. 1024–1060.

17. Б. Парлетт. Симметричная проблема собственных значений. Численные методы. 1983. 87 с.

18. Babenko V. N., Galyamina A. G., Rogozin I. B., Smagin D. A., Kudryavtseva N. N. Dopamine response gene pathways in dorsal striatum MSNs from a gene expression viewpoint: cAMP-mediated gene networks // BMC Neurosci. 2020. №21. P. 12.