*Математическая биология и биоинформатика*
*2018. Т. 13. № 1. С. 50–67. doi: 10.17537/2018.13.50*

**========== ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ ДАННЫХ ==========**

УДК 57.087.1

# Метод анализа однородности экспрессионных данных на основе теста Стьюдента

**Алиев Р.О.[1*], Борисов Н.М[2**].**

*НИЦ "Курчатовский институт"*

***Аннотация.*** Ещё в 2002 была осознана необходимость создания общедоступного хранилища результатов молекулярно-биологических экспериментов по профилированию генной экспрессии. За прошедшее время было создано несколько таких хранилищ информации, необходимых для того, чтобы результаты гибридизации мРНК на основе микрочипов (молекулярный анализ при помощи чувствительной матрицы зондов), а также высокопроизводительного секвенирования мРНК можно было анализировать и сопоставлять друг с другом. Несмотря на это, данные, депонированные в таких хранилищах, могут быть неоднородными, даже если они относятся к одинаковому типу здоровых или патологически изменённых органов и тканей, и были исследованы на одинаковой платформе. В настоящей работе предложен новый метод анализа однородности экспрессионных данных на основе теста Стьюдента. При помощи вычислительных экспериментов мы показали преимущества нашего метода по затратам времени на обработку больших массивов данных, а также разработали метод интерпретации результатов применения теста Стьюдента. С использованием нового метода анализа данных были проведены другие исследования, позволяющие визуализировать общую картину генной экспрессии и сравнить между собой экспрессионные профили при разных болезнях, либо разных стадиях одной болезни.

***Ключевые слова:*** генная экспрессия, профилирование мРНК, транскриптом, большие данные, репозитории общедоступных данных, тест Стьюдента, кластеризация.

## ВВЕДЕНИЕ

В 2000-х гг. стали чрезвычайно популярными молекулярно-биологические эксперименты профилирования генной экспрессии [1]. Такие эксперименты используют как микрочиповую гибридизацию мРНК, так и ее высокопроизводительное секвенирование [2], а также количественную масс-спектрометрию белков [3]. Результатом этих экспериментов являются большие массивы данных (профили экспрессии генов и белков).

На протяжении ряда лет наблюдался растущий интерес к тому, чтобы эти данные были доступны научному сообществу при первой публикации результатов исследования в научной литературе; еще в 2003 была осознана необходимость создания общедоступного хранилища таких наборов данных [4]. Первым ресурсом такого рода был GEO (Gene Expression Omnibus) [5]. Кроме него позднее были созданы такие большие базы данных как ArrayExpress, PRIDE, TCGA [6–8]. За несколько лет международными усилиями был создан каталог с минимальным набором информации, необходимым для того, чтобы эксперименты на основе микрочипов (молекулярный

*big_ruslik@mail.ru
**borisov@oncobox.com

анализ при помощи чувствительной матрицы зондов) можно было правильно интерпретировать и сопоставить друг с другом [4].

Несмотря на это, данные, депонированные в таких хранилищах, могут оказаться разнородными [9]. Возникли проблемы, связанные с кроссплатформенным анализом. Вопрос неоднородности данных, депонированных в разных хранилищах, существует до сих пор. Он вынуждает исследователей проводить дополнительный анализ входных данных, что накладывает некоторые ограничения на развитие данной области науки. Хотя в 2006 году и было показано, что раздельные биологические интерпретации результатов профилирования, полученных на разных платформах для схожих образцов, хорошо согласуются между собой, иногда даже дополняя друг друга [10], вопрос о том, можно ли корректно интерпретировать совмещённые кроссплатформенные данные, а не раздельные, остаётся открытым [11].

## МАТЕРИАЛЫ И МЕТОДЫ

### Материалы

В качестве источника данных для обработки была выбрана база открытого доступа GEO. Используемые материалы приведены в таблице 1, среди них есть результаты профилирования различных тканей человека как здоровых, так и имеющих патологические отклонения.

**Таблица 1.** Перечень упомянутых в статье баз данных GEO с комментариями

| Номер GEO | Тип ткани | Платформа | Тип использованных данных | № |
|---|---|---|---|---|
| GSE362 | мышечная ткань | GPL96 | только зд. тк.* | [12] |
| GSE1297 | ткань гиппокампа | GPL96 | болезнь Альцгеймера и зд. тк. | [13] |
| GSE1751 | кровь | GPL96 | только зд. тк. | [14] |
| GSE2779 | костный мозг | GPL96 | только зд. тк. | [15] |
| GSE6613 | кровь | GPL96 | только зд. тк. | [16] |
| GSE6280 | почечная ткань | GPL96 | только зд. тк. | [17] |
| GSE8397 | чёрное вещ-во | GPL96 и GPL97 | болезнь Паркинсона и зд. тк. | [18] |
| GSE9006 | кровь | GPL96 | только зд. тк. | [19] |
| GSE36980 | серое вещ-во | GPL6244 | болезнь Альцгеймера и зд. тк. | [20] |
| GSE19151 | кровь | GPL571 | венозная тромбоэмболия и зд. тк. | [21] |
| GSE48000 | кровь | GPL10558 | венозная тромбоэмболия и зд. тк. | [22] |
| GSE5107 | образец опухоли | GPL96 | глиобластома | [23] |
| GSE3446 | образец опухоли | GPL96 и GPL97 | нейробластома | [24] |
| GSE1432 | микроглия | GPL96 | только зд. тк. | [25] |

*зд. тк. – здоровые ткани

### Методы

Для оценки однородности профилей генной экспрессии мы предложили метод на основе теста Стьюдента. Нашей гипотезой было то, что значения экспрессии одного и того же гена у разных людей подчиняются нормальному распределению. В качестве первой группы используются значения экспрессии одного гена (скажем, гена $g$) в образцах, не содержащих медицинских патологий ($H$ – Healthy), в качестве второй – значения экспрессии того же гена в образцах, содержащих клинические отклонения от здоровых тканей ($D$ – Disease). Основным результатом теста является величина $P$ ($p$-value) – вероятность справедливости гипотезы о том, что сравниваемые средние значения не различаются. В качестве количественной меры отличия уровня экспрессии гена $g$ в патологическом состоянии от нормального мы использовали величину

$$f_g = sign\left(\overline{D_g} - \overline{H_g}\right) \cdot \left(-\lg\left(P_g\right)\right), \quad (1)$$

которая принимает положительные значения, если патологический ген имеет повышенный уровень экспрессии, и отрицательные, если ген имеет пониженный уровень экспрессии относительно образцов здоровых тканей. Значения $P$ невелики, они могут уменьшаться экспоненциально, поэтому в качестве абсолютной величины меры отклонения удобнее использовать логарифм от $P$.

Во время непосредственной обработки общего набора данных об экспрессии генов множество образцов случайным образом разбивается на пары таблиц так, что у них сохраняется общий набор образцов, содержащих экспрессию генов здоровых людей, при этом данные по уровням экспрессии для пациентов случайным образом делятся на две непересекающиеся группы. Всего в расчёте используется от ста до тысячи подобных случайных разбиений, за исключением случаев, когда размер входящей таблицы недостаточно велик и подобрать необходимое количество уникальных пар разбиений невозможно. Для всех генов каждой парной таблицы проводится тест Стьюдента и считается величина $f$. Для получившихся векторов значений $f_1$ и $f_2$, где $f_1$ и $f_2$ – это столбцы полученных значений для первой и второй таблицы, мы рассчитывали Пирсоновский коэффициент корреляции (назовем его $C_i$, где $i$ – номер разбиения). При

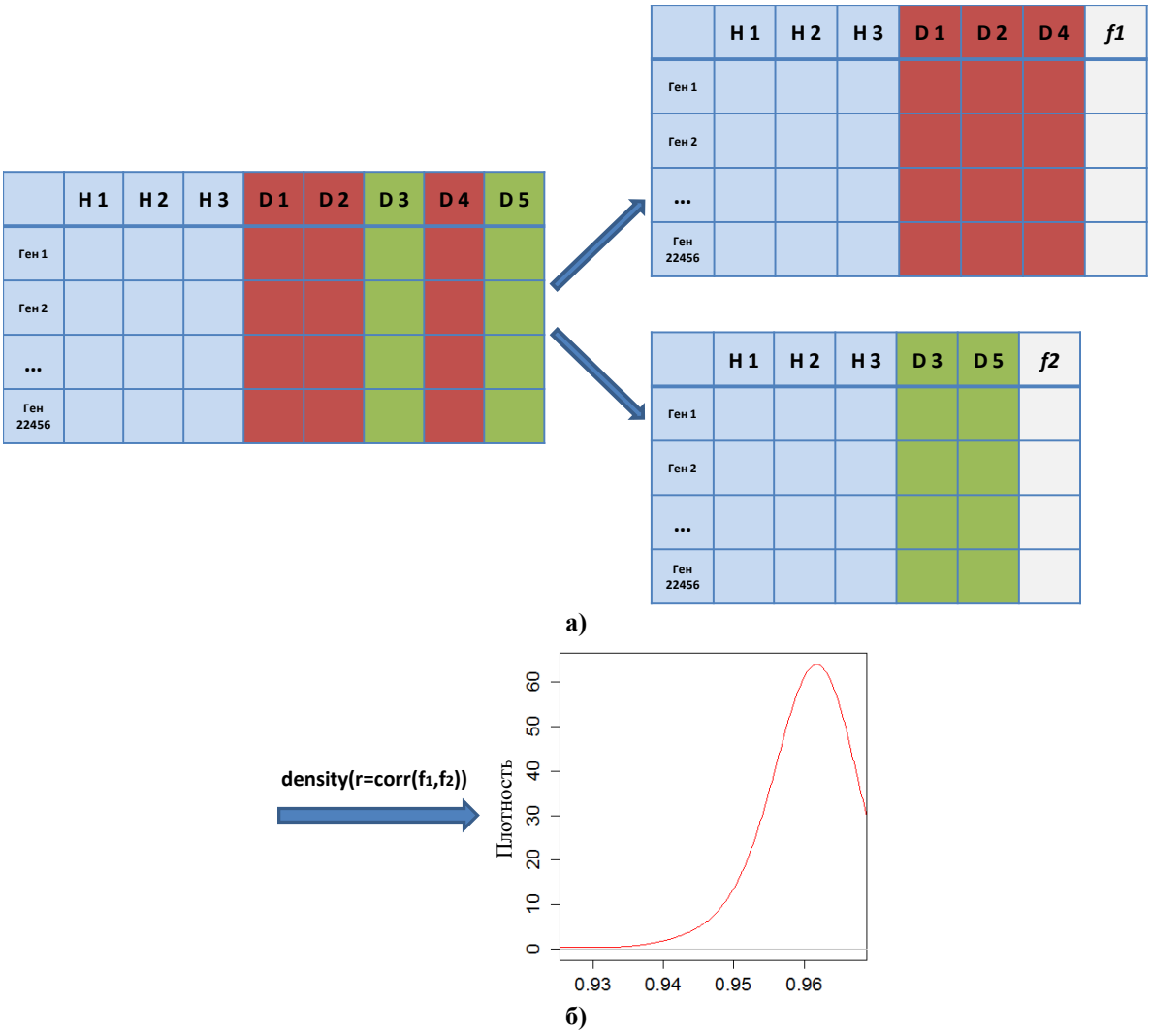



**Рис. 1.** Иллюстрация используемого метода: **а)** разбиение исходных данных, содержащих результаты транскриптомных экспериментов, на две таблицы и расчёт величины $f$ для каждого гена (см. формулу (1)); **б)** расчёт плотности коэффициентов корреляции для столбцов $f$, полученных для всех созданных разбиений.

построении графика плотности распределения коэффициентов корреляции (по оси $X$ отложена величина коэффициента корреляции, по оси $Y$ – количество точек с таким значением) можно обнаружить от одного до нескольких пиков. Если был получен один пик с центром в значении $C > 0.8$, то мы считали данные однородными. При значениях $0.7 < C < 0.8$ – условно однородными. Если $0.6 < C < 0.7$, то условно неоднородными. При $C < 0.6$ мы считаем данные неоднородными. Иллюстрация предложенного метода приведена на рисунке 1. Пороговые значения величины $C$ были получены экспериментально при проведении расчётов на данных, с заранее известной структурой, результаты этих исследований приведены далее, например, на рисунках 3,а; 9 и 10. В рамках проведенной работы коэффициент корреляции не принимал значений меньше нуля.

Мы также предложили упрощённые варианты анализа данных, которые используются для исследования поведения генов, а именно определения количества генов с пониженным или повышенным уровнем экспрессии и их степени отклонения от здоровых тканей, и для исследования уровня схожести двух болезней (или стадий одной болезни) между собой. Для этих расчётов не требуется использование разбиений. Для первого метода строится график плотности величины $f$. Для второго необходимо рассчитать $f$ для двух различных таблиц и получить Пирсоновский коэффициент корреляции между ними. Подобный подход позволяет получить количественную оценку уровня схожести двух различных баз данных. Стоит отметить, что для такого исследования необходимо, чтобы входные данные (две таблицы, содержащие результаты профилирования) были однородны сами в себе, квантильно нормализованы [26], а их образцы здоровых тканей должны быть однородны между собой.

### Используемое программное обеспечение

Для реализации метода была использована бесплатно распространяемая среда с открытым кодом – R. Использованные пакеты и их назначение приведены в таблице 2.

**Таблица 2.** Использованные пакеты R

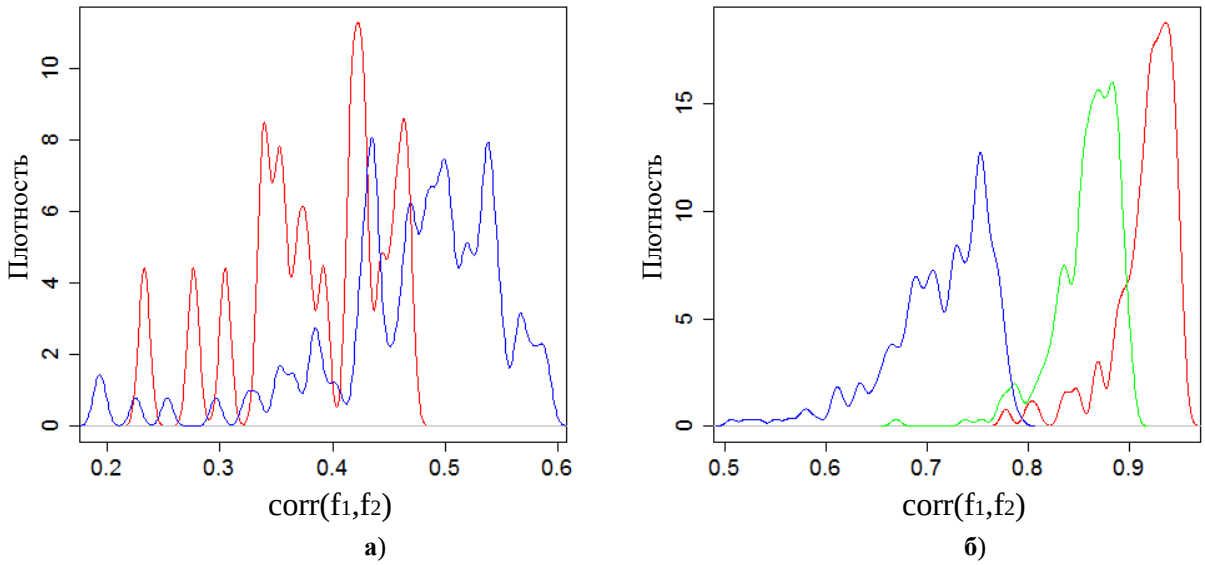
| Пакет | Решаемая задача | Ссылка |
|---|---|---|
| pvclust | Реализация метода иерархической кластеризации | [27] |
| magrittr | Использование Forward-Pipe Operator | [28] |
| BioBase | Участие в препроцессировании данных | [29] |
| gcrma | Участие в препроцессировании данных | [30] |
| affy и affyPrepr | Участие в препроцессировании данных | [31] |
| preprocessCore | Участие в препроцессировании данных | [32] |
| multtest | Участие в препроцессировании данных | [33] |
| stats | Участие в препроцессировании данных | [34] |
| AnnotationDbi | Участие в препроцессировании данных | [35] |

## РЕЗУЛЬТАТЫ

### Метод, применённый на намеренно разнородных данных, влияние здоровых тканей и предварительного логарифмирования на результаты исследования

При исследовании баз данных, содержащих образцы тканей мозга, была обнаружена многомодальность графиков плотности величин коэффициентов корреляции $C$. Подобные картины наблюдались для данных, взятых из наборов

GSE1297 [13], GSE8397 [18] и GSE36980 [20]. Иллюстрация примеров полученных графиков для GSE1297 и GSE36980 приведена на рисунке 2.




**а)** **б)**

**Рис. 2.** Иллюстрация работы метода на неоднородных данных, приведших к многомодальности: **а)** – болезнь Альцгеймера, красный: GSE1297 (9 H, 6 D, 22283 гена) – тяжёлая стадия; синий: GSE36980 (18 H, 14 D, 33297 генов) – фронтальная кора; **б)** – венозная тромбоэмболия, красный: GSE1951 (63 H, 69 D, 22215 генов), синий: GSE48000 (25 H, 39 D, 47304 гена) – тяжёлая стадия, зелёный: GSE48000 (25 H, 33 D, 47304 гена) – начальная стадия.

Результат для GSE1297 иллюстрирует, что при попытке использовать авторский метод для данных, содержащих недостаточно большую выборку образцов патологических тканей, будет получено малое число точек из-за маленького числа возможных разбиений, что приводит к многомодальности, набору узких пиков. Подобный результат покажет исследователю невозможность анализа входных данных при помощи статистических методов. Однако, как проинтерпретировать остальные графики?

Мы сделали предположение, что подобная картина возникает при кластеризации данных, например, если образец состоит из нескольких видов тканей. Так, нейроны и астроциты сопутствуют друг другу, но при этом картина генной экспрессии у них должна отличаться. Для проверки было проведено несколько расчётов, их можно свести к двум типам. Первый тип – в качестве патологически изменённых органов и тканей (Disease) использовались образцы одной и той же ткани, но из разных серий. Второй тип – в качестве патологических использовались образцы разных тканей из разных серий. Эти расчёты так же позволяют получить данные о влиянии на эксперимент не только неоднородности, но и выбора набора здоровых образцов, платформы и предварительного логарифмирования данных.

В расчёте, проиллюстрированном на рисунке 3, в качестве здоровых тканей были использованы образцы из GSE6613 [16]. В качестве исследуемых данных выступали GSE362[12], GSE2779 [15] и GSE9006 [19]. На рисунке 3,а показан результат исследования каждой ткани по отдельности. Можно обратить внимание на одномодальность полученных графиков. Это говорит в пользу однородности входных данных. На рисунке 3,б показан результат исследования данных, созданных попарным объединением образцов разных тканей. Было обнаружено, что это приводит к образованию бимодальной картины, с наиболее выраженным пиком для большей величины коэффициента корреляции. Далее была создана база данных, содержащая все три ткани одновременно. В результате был получен график (рис. 3,в), имеющий сложную картину пиков, но также содержащий наиболее выраженный пик справа. Было принято решение дополнительно исследовать этот случай. Чтобы восстановить

пропорции каждой ткани в разбиениях для каждого случая была получена величина, равная $q = \sqrt{\frac{n_1 \cdot m_1}{n \cdot m}}$, где $n_1$ и $m_1$ – наибольшее количество тканей одного типа в первой и второй группе соответственно, $n$ и $m$ – суммарное количество образцов в группах. Набор принимаемых значений $q$ для различных разбиений приведён в таблице 3. Таким образом, для каждой пропорции существует одно значение, при этом оно тем больше, чем сильнее оказались разделены ткани по разным группам. Из рисунка 3,г можно сделать вывод о том, что случаи с менее четкой кластеризацией дают большую величину коэффициента корреляции. При появлении кластеризации значение коэффициента корреляции между двумя группами падает. Также для первого графика можно найти соответствие между пиками и случаями разных пропорций в разбиениях у второго графика, однако, для случая $q > 0.5$ пики хуже различимы из-за их взаимного наложения.

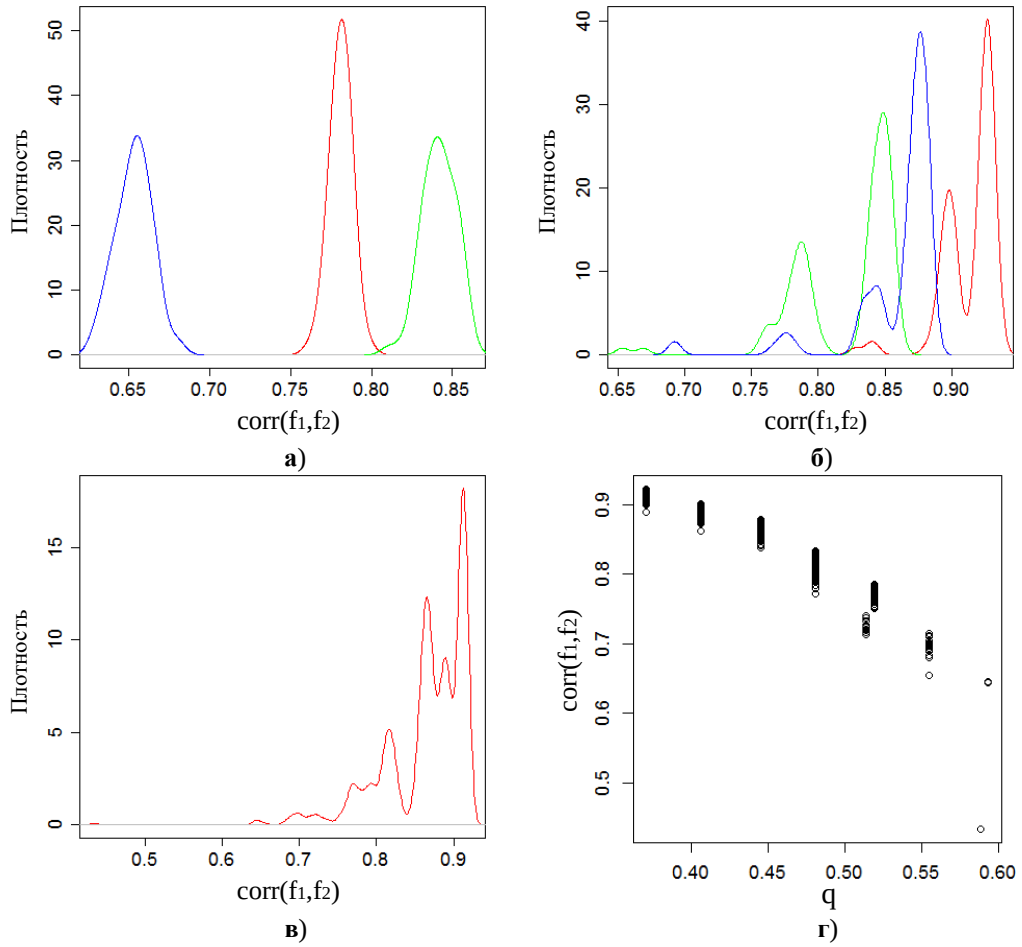



**Рис. 3. а)** Красный: GSE362 (8 D), зелёный: GSE2779 (8 D), синий: GSE9006 (8 D); **б)** красный: GSE362+GSE2779 (16 D), зелёный: GSE362+GSE9006 (16 D), синий:GSE2779+GSE9006 (16 D); **в)** GSE362+GSE2779+GSE9006 (24 D); **г)** GSE362+GSE2779+GSE9006; в качестве набора здоровых тканей выступает GSE6613 (22 H, 22215 генов).

**Таблица 3.** Соответствия значений $q$ для разбиения трёх групп по 9 образцов на две

| Разбиение | $q$ | Разбиение | $q$ |
|---|---|---|---|
| 4:4:5 | 0.3706 | 2:4:7 | 0.5189 |
| 3:5:5 | 0.406 | 1:5:7+2:3:8 | 0.5547 |
| 3:4:6 | 0.4447 | 2:2:9 | 0.5883 |
| 2:5:6 | 0.4804 | 1:4:8 | 0.593 |
| 1:6:6 | 0.5136 | | |

Для такого же набора тканей, но с другими здоровыми образцами, а именно из GSE6280 [17], был проведен аналогичный расчет (рис. 4). При замене здоровых тканей мы наблюдаем изменения (сдвинулись центры пиков) на панели (а), в случае без смешения тканей. Для образцов из набора GSE2779 была выявлена несимметричность, что может говорить о существовании второго пика, слившегося с первым. Однако общая картина по модальности оказалась схожей с результатами предыдущих расчетов. Это свидетельствует о влиянии здоровых тканей на результат расчёта, однако общая закономерность этого влияния на основании только данных, представленных на рисунке 4, остаётся неизвестной.

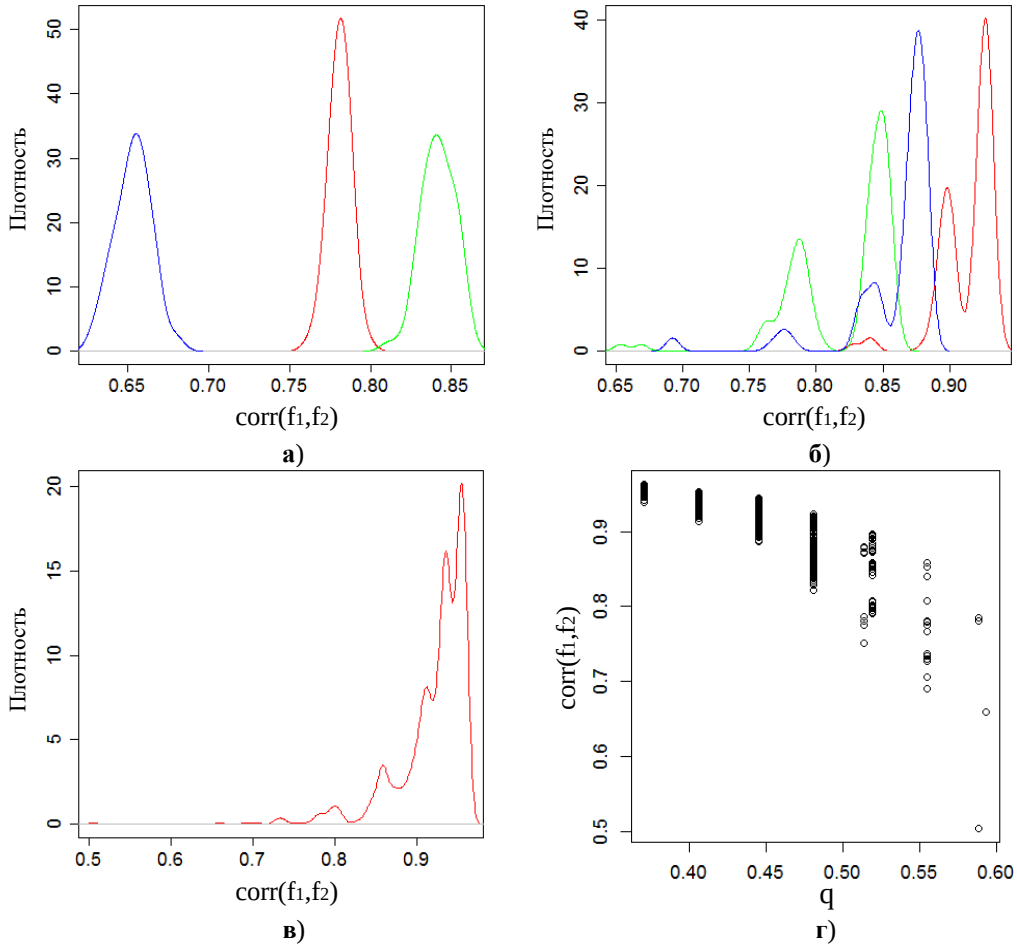



**Рис. 4. а)** Красный: GSE362 (8 Т), зелёный: GSE2779 (8 Т), синий: GSE9006 (8 Т); **б)** красный: GSE362+GSE2779 (16 Т), зелёный: GSE362+GSE9006 (16 Т), синий: GSE2779+GSE9006 (16 Т); **в)** GSE362+GSE2779+GSE9006 (24 Т); **г)** распределение GSE362+GSE2779+GSE9006; в качестве набора здоровых тканей выступает GSE6280 (6 Н, 22215 генов).

Для проверки влияния логарифмирования данных генной экспрессии на результат обработки был повторён расчёт со здоровыми тканями из GSE6613 и образцами GSE362, GSE2779, GSE9006. Результаты приведены на рисунке 5. При логарифмировании можно наблюдать смещение пиков, чаще всего значения коэффициентов корреляции увеличивается. На рисунке 4,г видно, что для каждого случая разбиения величина коэффициентов корреляции имеет более широкий набор значений, из-за этого некоторые пики накладываются друг на друга и их сложнее отделить.

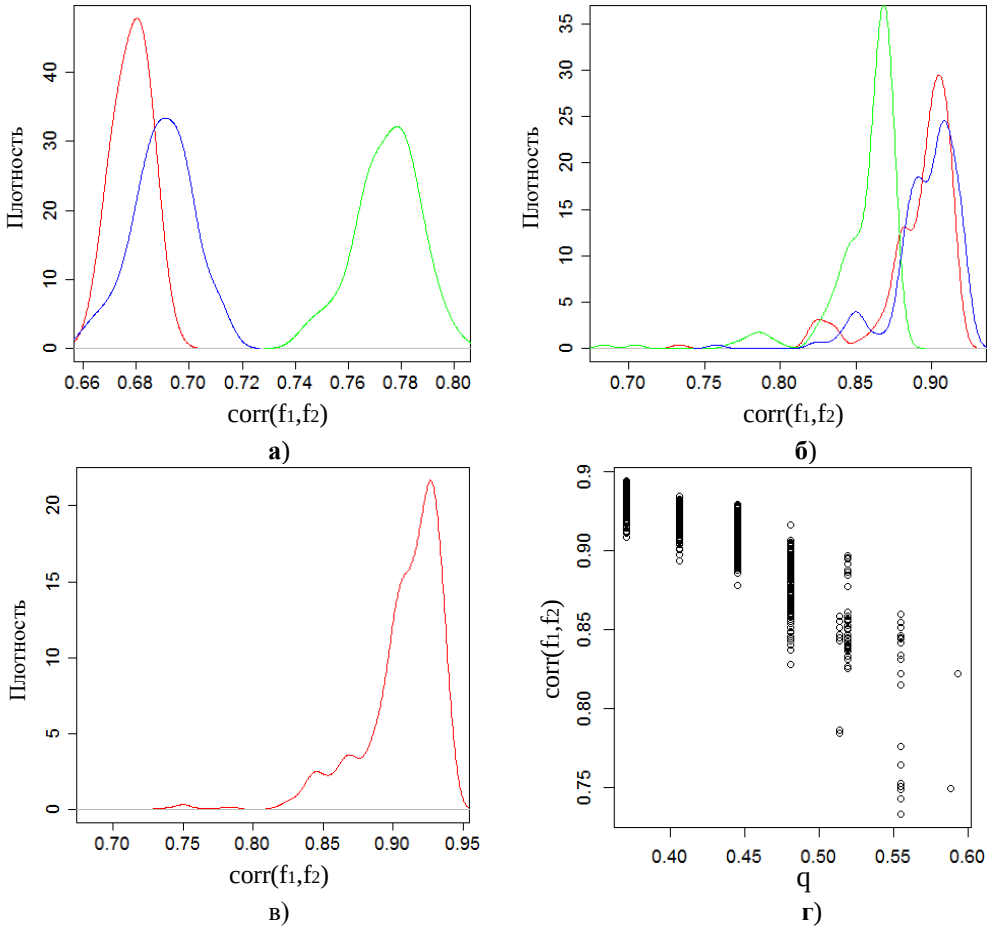



**Рис. 5. а)** Красный: GSE362, зелёный: GSE2779, синий: GSE9006; **б)** красный: GSE362+GSE2779, зелёный: GSE362+GSE9006, синий: GSE2779+GSE9006; **в)** GSE362+GSE2779+GSE9006; **г)** GSE362+GSE2779+GSE9006; в качестве набора здоровых тканей выступает GSE6613; данные предварительно прологарифмированы.

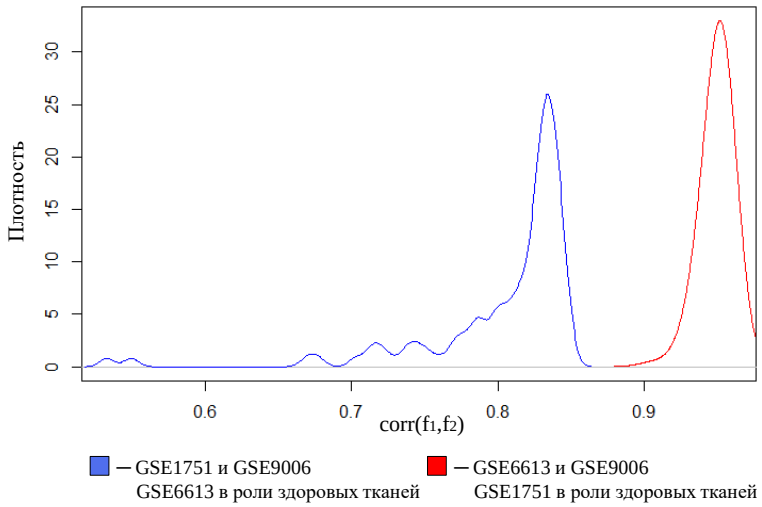



**Рис. 6.** Иллюстрация обработки баз данных GSE1751 (14 образцов), GSE6613 (21 образец), GSE9006 (19 образцов) образцов здоровых тканей при разном выборе из них набора здоровых тканей.

Использование нескольких тканей одного типа из разных исследований может привести как к однородному массиву данных, так и к не вполне однородному (пример на рис. 6). В качестве используемых данных были взяты образцы здоровых тканей крови из GSE1751 [14], GSE6613, GSE9006. Несмотря на то, что данные были получены на одной и той же платформе, а в качестве образцов выступали здоровые

ткани одного и того же типа, нами были получены разные результаты оценки однородности объединённых в один массив данных, которые зависели от выбора набора здоровых тканей, что не только подтверждает существенность влияния этого набора, но и позволяет сделать вывод о том, что подобные исследования всегда будут требовать дополнительной проверки. Таким образом, данные для одной и той же ткани и платформы могут не быть однородными между собой, если они были получены в разных экспериментах.

**Кроссплатформенный анализ неоднородного массива данных генной экспрессии полученного для болезни Паркинсона и диабета первого типа**

Использование даже схожих экспериментальных платформ для одних и тех же образцов может привести к результатам, на первый взгляд, отличающимся друг от друга. В качестве примера приведены результаты обработки базы данных GSE8397. Она содержит в себе образцы черного вещества головного мозга, разделённого на среднюю (MS) и боковые (LS) части, и лобной коры (SF) головного мозга здоровых людей и пациентов с болезнью Паркинсона. Каждый образец был обработан на двух платформах – GPL96 (A) и GPL97 (B). Данные проиллюстрированы на рисунках 7,а; 7,б и 7,в.

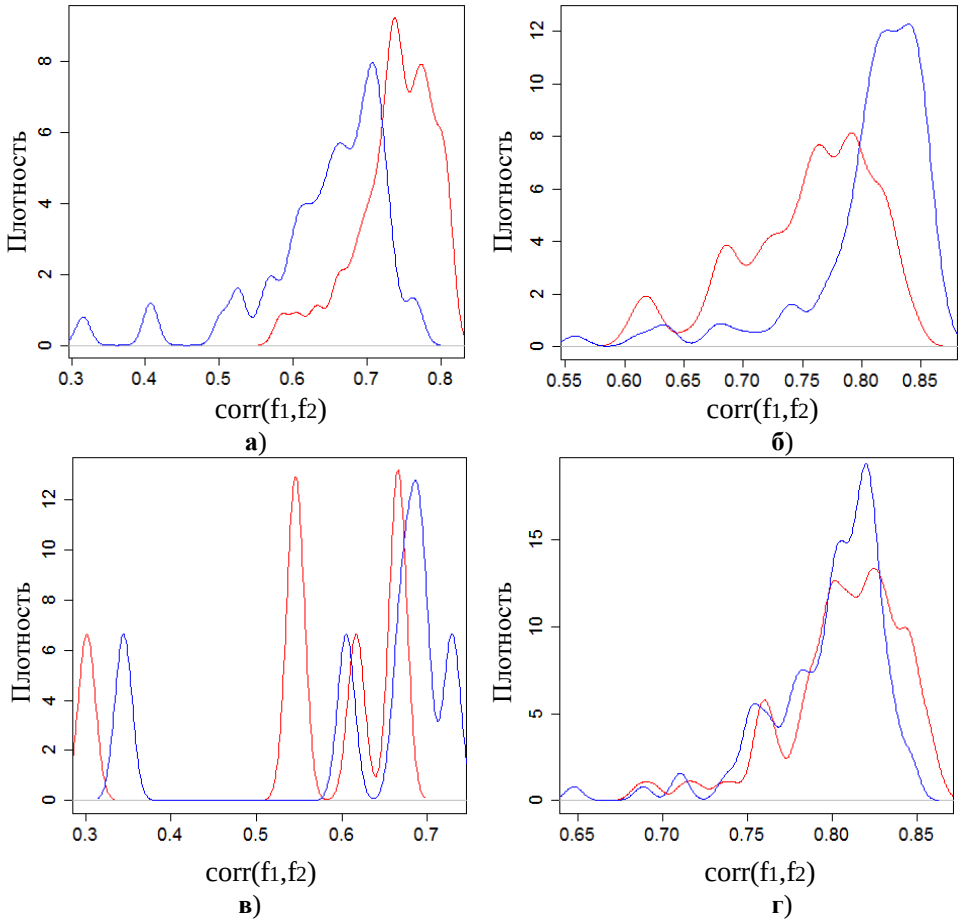



**Рис. 7.** Результат обработки базы данных GSE8397 **(а, б, в)** и GSE9006 **(г)**; красный: на платформе GPL96 (22283 гена для GSE8397, 11704 гена для GSE9006), синий: на платформе GPL97 (22645 генов для GSE837, 7669 генов для GSE9006); **а)** средняя часть чёрного вещества (7 H, 15 D); **б)** боковая часть (6 H, 9 D); **в)** лобная кора (3 H, 4 D); **г)** кровь (24 H, 90 D).

Несмотря на разный характер распределения коэффициента корреляции, можно заметить, что количество пиков для каждого случая у разных платформ одинаковое, за исключением первого случая, где два пика сливаются. Подобное отсутствие

зависимости числа пиков от исследовательской платформы возникло из-за пересекающихся наборов исследуемых генов. При уменьшении количества совпавших генов пики меняют высоту, сдвигается их центр, что, в конечном счете, приводит к видоизменениям или потере. Большое количество пиков в данном исследовании так же, как и выше, можно объяснить сложным составом тканей головного мозга.

Похожая картина была получена при анализе таблиц GSE9006, содержащих в себе данные по диабету первого и второго типа. В этом эксперименте также каждый образец был обработан на двух платформах – GPL96 (A) и GPL97 (B). Иллюстрация полученных результатов для диабета первого типа приведена на рисунке 7,г. Можно обратить внимание на то, что для платформ A и B количество пиков сохраняется.

**Пример анализа для неоднородного массива данных генной экспрессии острого венозного тромбоза, включая осложнение в виде ТЭЛА**

Стоит отметить, что данные могут оказаться неоднородными не только из-за человеческих ошибок или неправильно поставленного эксперимента. Неоднозначность природы болезни, разные варианты её течения у больных – всё это тоже оказывает сильное влияние на результаты.

При исследовании венозной тромбоэмболии были рассмотрены серии экспериментов GSE19151[21] и GSE48000[22]. Для первого набора профилей была обнаружена неоднородность данных, при этом в нём не было разделения по тяжести болезни, как это сделано для GSE48000. Было выдвинуто предположение, что данные второго набора профилей будут однородными в рамках каждого случая (лёгкое, среднее и тяжёлое течение болезни). Но было обнаружено, что данные во всех случаях неоднородны и во втором наборе профилей генной экспрессии (рис. 2,б). Это позволяет говорить об индивидуальном течении болезни и сложности систематизации общих черт картин генной экспрессии для пациентов больных острым венозным тромбозом, включая осложнение в виде ТЭЛА.

Тем не менее, наш метод позволяет провести дополнительные исследования картин генной экспрессии. На рисунке 8,а построены графики плотности значения $f$ для GSE48000 в каждом случае тяжести болезни. Можно обратить внимание, чем тяжелее протекает болезнь, тем шире график, что говорит об увеличении аномально экспрессирующих генов при развитии заболевания.

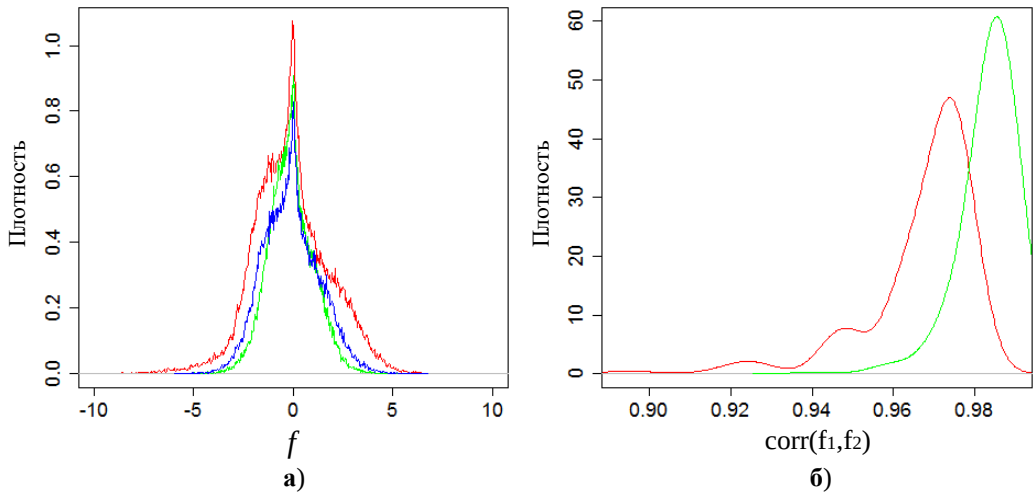



**Рис. 8. а)** Плотность величины $f$ для тромбоэмболии (GSE4800): зелёный – лёгкая форма тяжести (33 D), синий – средняя (32 D), красный – тяжёлая (39 D) (у всех 25 H, 47304 гена); **б)** – результат обработки объединённой базы данных глиобластомы (GSE5107) и нейробластомы (GSE3446) (12 H, 43 D, 20930 генов), зелёный – без предварительной квантильной нормализации, красный – с предварительной квантильной нормализацией.

Это подтверждается при дальнейшем исследовании данных. Полученные изображения оказались схожими присутствием "ступени" в левой части графика (гены с пониженным уровнем экспрессии). В составе этой ступени для лёгкой формы (L) на интервале (–2;–0.3) найдено 19049 генов, для средней тяжести (M) – 17701 ген, для тяжёлой (H) – 15906, при этом общих генов для всех случаев всего 8911 генов. Более интересным является результат для крайних случаев. Гены с сильно пониженным уровнем экспрессии увеличивают своё число при более тяжёлом течении болезни. На интервале (–7;–4) для L обнаружено 55 генов, для M – 88 генов, для H – 870 генов. Гены со значением $f < -7$ обнаружены только для H, их 69. Ближе всего к здоровым тканям, на участке (–0.3;0.3), у L – 11261 ген, M – 9207 генов, H – 8087 генов. Для генов с повышенным уровнем экспрессии картина схожа с генами, обладающими пониженным уровнем экспрессии. Для (4;6) у L – 20 генов, у M – 157 генов, у H – 758 генов. При этом генов со значением $f > 6$ у L – 1 ген, у M – 6 генов, у H – 43 гена. Всё это позволяет говорить, что чем тяжелее протекает болезнь, тем большее количество генов аномально экспрессируют, причем степень отклонений в уровнях экспрессии увеличивается.

**Анализ распределения коэффициента корреляции на заведомо неоднородных данных генной экспрессии для образцов, взятых из баз данных различных болезней; влияние предварительной квантильной нормализации**

Мы исследовали результаты профилирования образцов глиобластомы (GSE5107 [23]) и нейробластомы (GSE3446 [24]) с предварительной квантильной нормализацией [26]. В качестве набора здоровых тканей выступали экспрессионные данные микроглии из GSE1432 [25]. Полученные таблицы были признаны однородными (рис. 9).

Коэффициент корреляции величины $f$, полученной для данных глиоблостомы и нейробластомы, был равен 0.56. Аналогичные значения для разных уровней тяжести венозной тромбоэмболии (GSE48000), т.е. между заболеваниями с невысокой и средней, невысокой и высокой, а также средней и высокой степенью тяжести, оказались равными 0.80, 0.70 и 0.86 соответственно.

Стоит отметить, что сравнительный анализ результатов профилирования различных серий может быть затруднительным из-за различных эффектов (неоднородность платформы и батч-эффект [9] (систематическая погрешность измерений, обусловленная неучтёнными деталями экспериментальной процедуры), влияние которого меньше, но с которым сложнее бороться). Для борьбы с ними используются различные методы, в том числе и квантильная нормализация. Хоть этот метод и может повлиять на данные, он широко применим и активно используем. В нашем расчёте были использованы как нормализованные данные, так и ненормализованные. Так на рисунке 8,б можно увидеть, как квантильная нормализация предотвратила классификацию неоднородного набора данных за однородный.

Слияние наборов данных глиобластомы и нейробластомы дало набор, ранее распознанный как неоднородный (рис. 8,б). Полученный график плотности коэффициентов корреляций величин $f$ для разбиений имел многомодальность, что, как было показано ранее, классифицирует неоднородные данные.

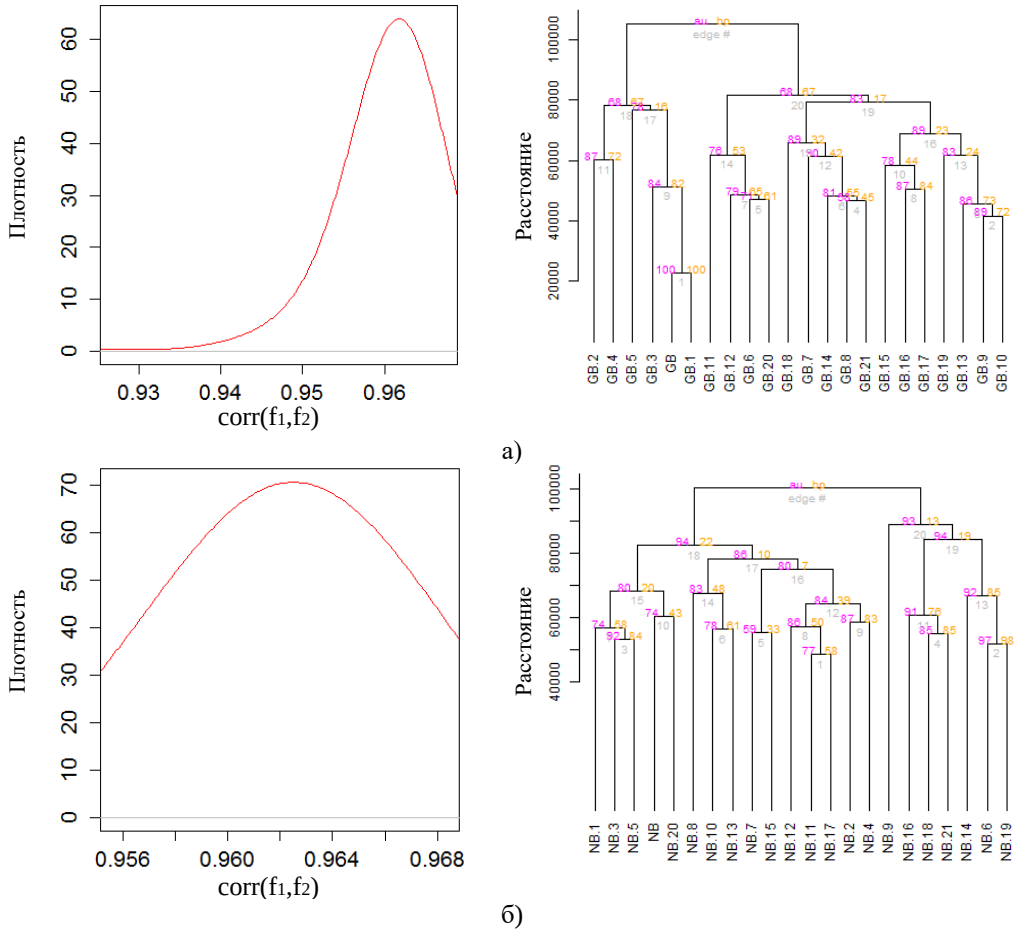



**Рис. 9.** Сравнение результатов анализа, полученных авторским методом (слева) и методом иерархической кластеризации (справа), для данных **а)** глиобластомы (GSE5107) (12 H, 21 D, 21873 гена) и **б)** нейробластомы (GSE3446) (12 H, 21 D, 21319 генов).

## Метод, применённый на случайно сгенерированных данных

Для отработки метода и набора проверенных результатов, позволяющих делать дальнейшие выводы, был проведен модельный расчёт с использованием случайно сгенерированных данных. Этот вычислительный эксперимент был проведен три раза, с использованием трех различных распределений: нормального, логнормального и равномерного. В качестве набора здоровых тканей мы использовали данные, полученные на образцах крови (GSE6613). Наибольшая однородность достигается, если добавочные уровни экспрессии подчиняются нормальному распределению: был получен узкий пик с центром в 0.99. Для добавочных уровней экспрессии, подчиняющихся логнормальному распределению, распределение величины $C$ имеет моду с центром в 0.85. Для выборки из равномерного распределения также был выявлен пик функции плотности вероятности распределения величины $C$, однако его центр находился в 0.31, что говорит об очень низком уровне корреляции. Распределение для величины $C$ является, тем не менее, одномодальным для всех трех вычислительных экспериментов со случайными добавочными значениями уровней экспрессии, так как результирующее множество уровней экспрессии оказывается неделимым на кластеры. Иллюстрация приведена на рисунке 10.



Математическая биология и биоинформатика. 2018. Т. 13. № 1. doi: 10.17537/2018.13.50

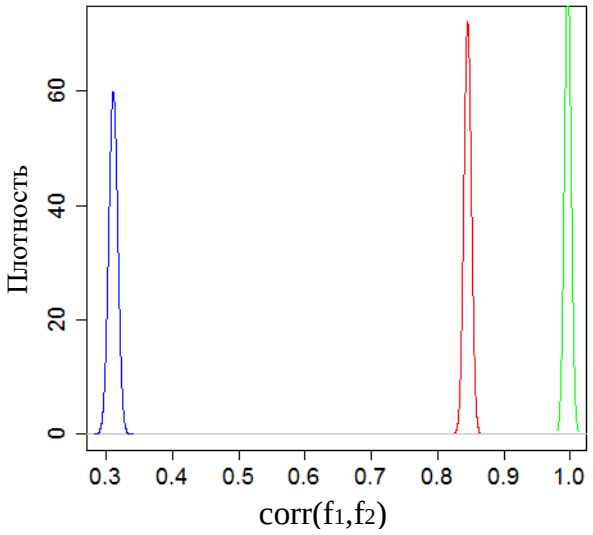



**Рис. 10.** Результат обработки случайно сгенерированных данных: красный – логнормальное распределение; зелёный – нормальное распределение; синий – равномерное распределение (22H, 24 D, 22215).

## Сравнение качества результатов и трудоёмкости предложенного нами метода анализа неоднородности с методом иерархической кластеризации

Параллельно с основным исследованием было проведен анализ с использованием метода иерархической кластеризации, реализованного в пакете pvclust [27]. На рисунке 11 проиллюстрирован результат обработки данных с набором здоровых тканей из GSE6613 и образцами GSE362, GSE2779, GSE9006. При сравнении с результатами, полученными с помощью нашего метода (рис. 3 (в)), можно увидеть, что иерархическая кластеризация даёт более полную информацию о содержащихся в таблице образцах, так все группы тканей были без ошибок разделены, в то время как наш метод позволяет говорить о факте неоднородности данных и содержании в ней трёх и более групп.

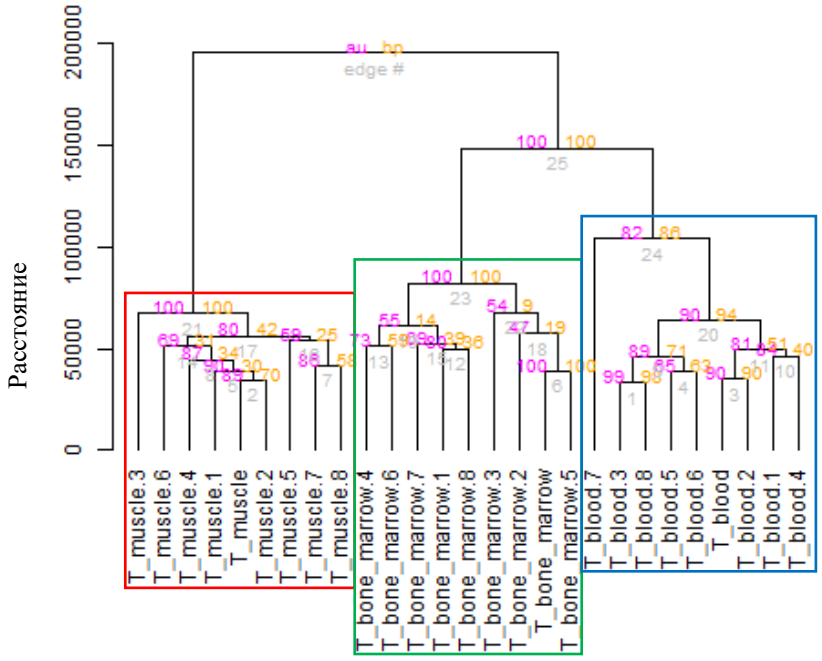



**Рис. 11.** Результат обработки данных методом иерархической кластеризации с набором здоровых тканей из GSE6613 и образцами GSE362, GSE2779, GSE9006.

На рисунке 9 приведены результаты исследования данных для нейробластомы (GSE3446) и глиобластомы (GSE5107). Из каждой базы данных были случайным образом выбраны 21 пациент, в качестве набора здоровых тканей выступала микроглиальная ткань (GSE1432, 12 образцов). Можно обратить внимание, что в то время как наш метод на основе $P$ определяет однородность данных, иерархическая кластеризация даёт результат, который можно проинтерпретировать иначе (существование кластеров). Это может говорить о том, что метод иерархической кластеризации лучше использовать тогда, когда заранее известно, что данные неоднородны.

Однако, наш метод имеет свои преимущества. Так, время обработки таблиц генной экспрессии методом на основе иерархической кластеризации сильно зависит от размера входных данных, а именно от количества патологических образцов. Мы провели расчёты для разного размера таблиц, начиная с десяти пациентов, результаты приведены на рисунке 12. Из графика можно увидеть нелинейность увеличения времени. Так при увеличении количества пациентов в 2 раза, время обработки увеличивается в 3.6 раз. При увеличении количества пациентов в 4 раза, затраченное время увеличилось в 26.5 раз. Таким образом, для наборов данных, содержащих 50 пациентов и более, предпочтительнее использовать наш метод, так как он позволяет получить выигрыш по времени.

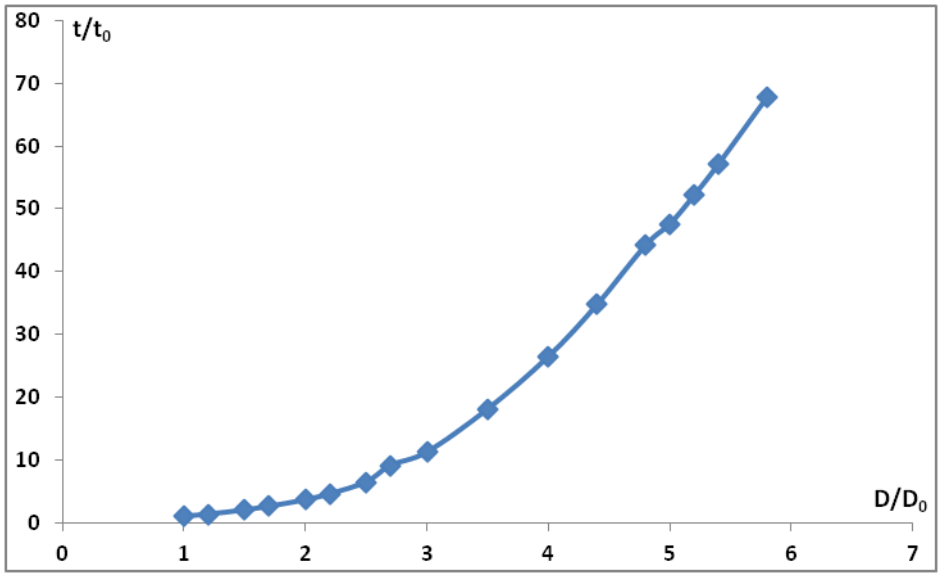



**Рис. 12.** Зависимость относительного увеличения времени обработки данных при помощи метода иерархической кластеризации от относительного увеличения количества пациентов в исследуемой таблице.

## Тест на чувствительность

Мы провели также тест на чувствительность для используемых методов по отношению к добавлению инородных данных. Для этого исследования были использованы те же данные нейробластомы и глиобластомы, что и в предыдущем разделе, с тем же набором здоровых тканей (микроглия). Полученные таблицы были признаны однородными (рис. 9). После этого последовательно добавлялись образцы крови (GSE6613), при добавлении каждого нового образца проводился перерасчёт. Было обнаружено, что добавление одного чужеродного образца в однородные данные приводит к заметному смещению центра пика влево; также он может стать асимметричным. Добавление двух образцов сразу же выливается в бимодальность, являющейся надёжным показателем неоднородности данных. Иллюстрация

результатов приведена на рисунке 13. Иерархическая кластеризация, как мы видели на примере из более чем 20 стохастических испытаний, всегда надежно свидетельствовала о присутствии хотя бы одного инородного образца.

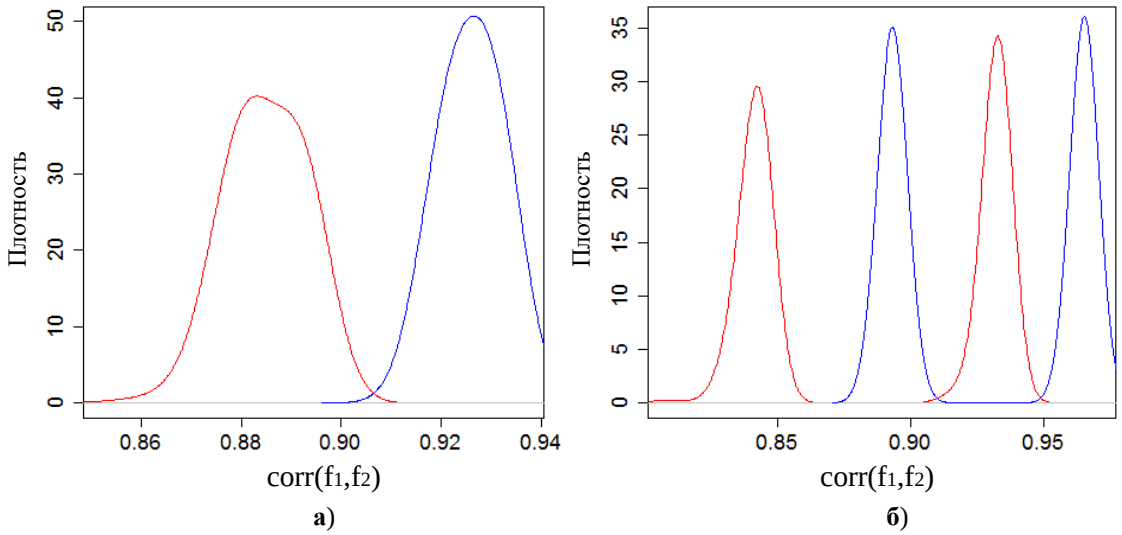



**Рис. 13.** Результаты обработки однородных данных с добавленными инородными данными: **а)** – добавление одного образца к глиобластоме (красный) и нейробластоме (синий); **б)** – добавление двух образцов к глиобластоме (красный) и нейробластоме (синий).

## ЗАКЛЮЧЕНИЕ

На настоящий момент в базе данных открытого доступа GEO содержится 1245188 профилированных человеческих образцов, объединённых в 38.264 серии [36], и эти числа будут расти. Для систематизации, очистки и объединения этих данных необходимо разработать и реализовать кроссплатформенный метод оценки однородности данных.

Нами был предложен новый метод, основанный на сравнении значений экспрессии генов в здоровых и патологических образцах с использованием теста Стьюдента для определения однородности баз данных генной экспрессии. Он позволяет не только качественно оценить уровень однородности, но и дать некоторые количественные характеристики, например число кластеров образцов, на которые разбиваются данные болезней. Был показан риск потери данных при кроссплатформенном анализе, однако метод остаётся применимым для каждой базы данных в частности. Также метод позволяет судить о схожести двух болезней при сливании данных в единый набор данных. Обнаружено влияние выбора образцов, считающихся здоровыми, предварительного логарифмирования и квантильной нормализации. Показано, метод позволяет проводить и более подробные исследования, определяя поведение генов. Дальнейшая работа позволит создать универсальный для разных платформ анализ экспериментальных данных, позволяющий давать быструю оценку для картин генной экспрессии.

Сравнение предложенного в статье метода с методом иерархической кластеризации продемонстрировало как недостатки, так и преимущества. Метод на основе анализа плотности распределения величины $P$ для теста Стьюдента (на основе анализа плотности распределения коэффициента корреляции величины $f$) является предпочтительным при работе с большим объёмом данных для анализа однородности генной экспрессии.

## СПИСОК ЛИТЕРАТУРЫ


1. Schena M., Shalon D., Davis R.W., Brown P.O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*. 1995. V. 270. № 5235. P. 467–470.
2. Zhang W., Yu Y., Hertwig F., Thierry-Mieg J., Zhang W., Thierry-Mieg D., Wang J., Furlanello C., Devanarayan V., Cheng J., Deng Y. et al. Comparison of RNA-seq and microarray-based models for clinical endpoint prediction. *Genome Biology*. 2015. V. 16. № 1. P. 133.
3. Kooken J., Foxa K., Foxa A., Altomareb D., Creekb K., Wunschelc D., Pajares-Merinod S., Martínez-Ballesterosd I., Garaizard J., Oyarzabale O., Samadpour. M. Identification of staphylococcal species based on variations in protein sequences (mass spectrometry) and DNA sequence (sodA microarray). *Molecular and cellular probes.* 2014. V. 28. № 1. P. 41–50.
4. Kellam P. Microarray gene expression database: progress towards an international repository of gene expression data. *Genome Biology*. 2001. V. 2. № 5. P. reports4011.1–4011.3. doi: 10.1186/gb-2001-2-5-reports4011
5. Edgar R., Domrachev M., Lash A. E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research.* 2002. V. 30. № 1. P. 207–210.
6. Brazma A., Parkinson H., Sarkans U., Shojatalab M., Vilo J., Abeygunawardena N., Holloway E., Kapushesky M., Kemmeren P., Garcia Lara G., Oezcimen A. et al. ArrayExpress–a public repository for microarray gene expression data at the EBI. *Nucleic Acids Research*. 2003. V. 31. № 1. P. 68–71.
7. Jones P., Côté R.G., Cho S.Y., Klie S., Martens L., Quinn A.F., Thorneycroft D., Hermjakob H. PRIDE: new developments and new datasets. *Nucleic Acids Research*. 2008. V. 36. P. D878–D883.
8. McLendon R., Bigner D., Friedman A., Van Meir E.G., Mastrogianakis G.M., Olson J.J., Brat D.J., Mikkelsen T., Lehman N., Aldape K. et al. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*. 2008. V. 455. № 7216. P. 1061–1068.
9. Demetrashvili N., Kron K., Pethe V., Bapat B., Briollais L. How to deal with batch effect in sequential microarray experiments? *Molecular Informatics*. 2010. V. 29. № 5. P. 387–393.
10. Guo L, Lobenhofer E.K., Wang C., Shippy R., Harris S.C., Zhang L., Mei1 N., Chen T., Herman D., Goodsaid F.M., et al. Rat toxicogenomic study reveals analytical consistency across microarray platforms. *Nature Biotechnology*. 2006. V. 24.
11. Borisov N., Suntsova M., Garazha A., Lezhnina K., Kovalchuk O., Aliper A., Ilnitskaya E., Sorokin M., Korzinkin M., Saenko V. et al. Data aggregation at the level of molecular pathways improves stability of experimental transcriptomic and proteomic data. *Cell Cycle*. 2017. V. 16. № 19. P. 1810–1823.
12. Welle S., Brooks A.I., Delehanty J.M., Needler N., Thornton C.A. Gene expression profile of aging in human muscle. *Physiological Genomics*. 2003. V. 14. № 2. P. 149–159.
13. Blalock E.M., Geddes J.W., Chen K.C., Porter N.M., Markesbery W.R., Landfield P.W. Incipient Alzheimer's disease: microarray correlation analyses reveal major transcriptional and tumor suppressor responses. *PNAS*. 2004. V. 101. № 7. P. 2173–2178.
14. Borovecki F., Lovrecic L., Zhou J., Jeong H., Then F., Rosas H.D., Hersch S.M., Hogarth P., Bouzou B., Jensen R.V., Krainc D. Genome-wide expression profiling of human blood reveals biomarkers for Huntington's disease. *PNAS*. 2005. V. 102. № 31. P. 11023–11028.

15. Sternberg A, Killick S., Littlewood T., Hatton C., Peniket A., Seidl T., Soneji S., Leach J., Bowen D., Chapman C. et al. Evidence for reduced B-cell progenitors in early (low-risk) myelodysplastic syndrome. *Blood*. 2005. V. 106. № 9. P. 2982–2991.
16. Scherzer C.R., Eklund A.C., Morse L.J., Liao Z., Locascio J.J., Fefer D., Schwarzschild M.A., Schlossmacher M.G., Hauser M.A., Vance J.M., Sudarsky L.R. et al. Molecular markers of early Parkinson's disease based on gene expression in blood. *PNAS*. 2007. V. 104. № 3. P. 955–960.
17. Yusenko M.V., Zubakov D., Kovacs G. Gene expression profiling of chromophobe renal cell carcinomas and renal oncocytomas by Affymetrix GeneChip using pooled and individual tumours. *International Journal of Biological Sciences*. 2009. V. 5. № 6. P. 517.
18. Duke D.C., Moran L.B., Pearce R.K.B., Graeber M.B. The medial and lateral substantia nigra in Parkinson’s disease: mRNA profiles associated with higher brain tissue vulnerability. *Neurogenetics*. 2007. V. 8. № 2. P. 83–94.
19. Kaizer E.C., Glaser C.L., Chaussabel D., Banchereau J., Pascual V., White. P.C. Gene expression in peripheral blood mononuclear cells from children with diabetes. *The Journal of Clinical Endocrinology & Metabolism*. 2007. V. 92. № 9. P. 3705–3711.
20. Hokama M., Oka S., Leon J., Ninomiya T., Honda H., Sasaki K., Iwaki T., Ohara T., Sasaki T., LaFerla F.M. et al. Altered expression of diabetes-related genes in Alzheimer's disease brains: the Hisayama study. *Cerebral Cortex*. 2013. V. 24. № 9. P. 2476–2488.
21. Lewis D.A., Stashenko G.J., Akay O.M., Price L.I., Owzar K., Ginsburg G.S., Chi J., Ortel T.L. Whole blood gene expression analyses in patients with single versus recurrent venous thromboembolism. *Thrombosis Research*. 2011. V. 128. № 6. P. 536–540.
22. Lewis D.A., Suchindran S., Beckman M.G., Hooper W.C., Grant A.M., Heit J.A., Manco-Johnson M., Moll S., Philipp C.S., Kenney K. et al. Whole blood gene expression profiles distinguish clinical phenotypes of venous thromboembolism. *Thrombosis Research*. 2015. V. 135. № 4. P. 659–665.
23. Tso C.L., Shintaku P., Chen J., Liu Q., Liu J., Chen Z., Yoshimoto K., Mischel P.S., Cloughesy T.F., Liau L.M., Nelson S.F. Primary glioblastomas express mesenchymal stem-like properties. *Molecular Cancer Research*. 2006. V. 4. № 9. P. 607–619.
24. Asgharzadeh S., Pique-Regi R., Sposto R., Wang H., Yang Y., Shimada H., Matthay K., Buckley J., Ortega A., Seeger R.C. Prognostic significance of gene expression profiles of metastatic neuroblastomas lacking MYCN gene amplification. *Journal of the National Cancer Institute*. 2006. V. 98. № 17. P. 1193–1203.
25. Rock R.B., Hu S., Deshpande A., Munir S., May B.J., Baker C.A., Peterson P.K., Kapur V. Transcriptional response of human microglial cells to interferon-[gamma]. *Genes and Immunity*. 2005. V. 6. № 8. P. 712.
26. Bolstad B.M., Irizarry R.A., Åstrand M., Speed T.P. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*. 2003. V. 19. № 2. P. 185–193.
27. Suzuki R., Shimodaira H. Pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics*. 2006. V. 22. № 12. P. 1540–1542.
28. Bache S.M., Wickham H. Magrittr: A forward-pipe operator for R. *R package version 1.5*. 2014. URL: https://CRAN.R-project.org/package=magrittr (дата обращения: 15.04.2018).
29. Gentleman R., Carey V., Morgan M., Falcon S. Biobase: base functions for Bioconductor. *R package version 2.34.0*. 2016. URL: https://www.bioconductor.org/packages/3.4/bioc/html/Biobase.html (дата обращения: 27.03.2018).

30. Wu Z., Irizarry R.A., Gentleman R., Martinez-Murillo F., Spencer F. A model-based background adjustment for oligonucleotide expression arrays. *Journal of the American statistical Association*. 2004. V. 99. № 468. P. 909–917.
31. Irizarry R.A., Gautier L., Bolstad B.M., Miller C. Methods for affymetrix oligonucleotide arrays. *R package version 1.52*. 2016. URL: https://www.bioconductor.org/packages/3.4/bioc/html/affy.html (дата обращения: 27.03.2018).
32. Bolstad B.M. preprocessCore: A collection of pre-processing functions. *R package version 1.36.0*. 2016. URL: https://www.bioconductor.org/packages/3.4/bioc/html/preprocessCore.html (дата обращения: 27.03.2018).
33. Pollard K.S., Gilbert H.N., Ge Y., Taylor S., Dudoit S. Resampling-based multiple hypothesis testing. *R package version 2.30.0*. 2016. URL: https://www.bioconductor.org/packages/3.4/bioc/html/multtest.html (дата обращения: 27.03.2018).
34. Team R.C., Worldwide C. R Foundation for Statistical Computing. *R package version 3.6.0* 2017. URL: https://stat.ethz.ch/R-manual/R-devel/library/stats/html/00Index.html дата обращения: 07.01.2018).
35. Pages H., Carlson M., Falcon S., Li N.A., AnnotationDbi P.R.S., SQLForge P.R.S. Annotation Database Interface. *R package version 1.36.2*. 2016. URL: https://bioconductor.org/packages/3.4/bioc/html/AnnotationDbi.html (дата обращения: 27.03.2017).
36. *Gene Expression Omnibus*. URL: https://www.ncbi.nlm.nih.gov/geo/ (дата обращения: 29.03.2018).

Рукопись поступила в редакцию 05.12.2017.
Дата опубликования 06.04.2018.