БИОХИМИЯ, 2021, том 86, вып. 3, с. 395 – 408

УДК 577.2

КЛЕТОЧНЫЙ ОТВЕТ НА СТРЕСС В ПАНОРАМНОЙ ПРОТЕОМИКЕ: КОНТРОЛЬ ЛОЖНОПОЛОЖИТЕЛЬНЫХ РЕЗУЛЬТАТОВ

© 2021 И.Т. Габдрахманов[1], М.В. Горшков[2,3], И.А. Тарасова[3]*

[1] *Сколковский институт науки и технологий, 121205 Москва, Россия*

[2] *Московский физико-технический институт (национальный исследовательский университет), 141701 Долгопрудный, Московская обл., Россия*

[3] *Институт энергетических проблем химической физики им. В.Л. Тальрозе ФГБУН Федерального исследовательского центра химической физики им. Н.Н. Семенова РАН; 119334 Москва, Россия; электронная почта: iatarasova@yandex.ru*

Поступила в редакцию 09.08.2020
После доработки 03.11.2020
Принята к публикации 01.12.2020

Одной из основных задач количественной протеомики является определение молекулярных изменений на уровне белков в клеточном ответе на стресс. При этом в основе биоинформатической обработки получаемых экспериментальных данных лежит статистический анализ, в котором многократно тестируется гипотеза о равенстве концентраций белков в стрессе относительно нормы. Возникает классическая проблема множественных сравнений, когда повышается вероятность получения ложноположительных результатов. На сегодняшний день известно множество подходов для решения этой проблемы. Однако их применение с исторически принятыми фиксированными порогами статистической значимости может приводить к потере потенциально ценной биологической информации. Используя протеомные данные, полученные ранее для модельных образцов дрожжей, содержащих белки в известных концентрациях, а также данные для биологических моделей раннего и позднего ответа клеток на стресс, были исследованы распределения ложноположительных и ложноотрицательных результатов в зависимости от значений кратных изменений концентраций и порога статистической значимости. На основе анализа плотности распределения точек на диаграммах рассеяния, метода Бенджамини–Хохберга и анализа обогащений генных онтологий предложен наглядный протокол оптимизации статистического порога и отбора дифференциально регулированных белков, который будет полезен исследователям, работающим в области количественного анализа протеомных данных.

КЛЮЧЕВЫЕ СЛОВА: протеомика, биоинформатика, клеточный ответ, масс-спектрометрия.

DOI: 10.31857/S0320972521030088

ВВЕДЕНИЕ

В современных протеомных исследованиях широкое распространение получил так называемый метод протеомики «снизу-вверх» («bottom-up proteomics»), в основе которого лежит использование тандемной масс-спектрометрии в сочетании с высокоэффективной жидкостной хроматографией (ВЭЖХ-МС/МС) [1]. В таком эксперименте цельноклеточные или субклеточные белковые фракции ферментативно расщепляются на пептиды, которые затем хроматографически разделяются и переводятся из жидкой фазы в газовую с помощью источника ионизации. После ионизации аналит попадает в масс-спектрометр, где измеряются массы пептидов, а также осуществляется их фрагментация. Благодаря спектрам фрагментации становится возможным восстановить первичную структуру пептидов, соотнести их с последовательностями белков и таким образом идентифицировать белковый состав образца [2].

Идентификация пептидов осуществляется с использованием так называемых протеомных поисковых машин (proteomic search engines) – специальных программ, основной задачей которых является сопоставление экспериментальных и теоретических спектров фрагментации пептидов в пределах заданной точности измерений массы. Для заданного правила ферментативного гидролиза из баз данных белков автоматически создаются списки теоретически возможных пептидов, для каждого из которых рас-

Принятые сокращения: ГО – генные онтологии; FC – кратное изменение концентрации белка (англ. fold change); fdr – доля ложно положительных результатов (англ. false discovery rate); MBR – выравнивание хроматомасс-спектрометрических данных по времени удерживания и массе пептида между экспериментами (англ. match-between-runs); NSAF – нормализованные спектральные интенсивности белков (англ. normalized spectral abundance factor); SC – подсчёт спектров фрагментации на белок (англ. spectral count); SI – нормализованный спектральный индекс (англ. spectral index normalized).

* Адресат для корреспонденции.

считываются массы пептида и их теоретически возможных фрагментов. Каждое совпадение экспериментального и теоретического спектра (пептид-спектральное совпадение, peptide-spectrum match, PSM) характеризуется параметрами, определяющими степень доверия такому событию. На этапе последующей обработки результатов поиска с помощью программного обеспечения осуществляется контроль уровня ложноположительных идентификаций и сборка белков из идентифицированных пептидов [3].

В количественной протеомике ставится цель: определить дифференциально регулированные белки при сравнении нескольких групп образцов [4]. Методы количественной протеомики делят на две группы. В первую группу входят так называемые безметочные методы, в основе которых лежит либо подсчёт числа спектральных идентификаций на белок, либо анализ хроматографических интенсивностей пептидов. Вторую группу составляют методы, в которых аминокислоты пептидов или белков метят изотопными метками. Главное отличие методов второй группы заключается в том, что они позволяют делать мультиплексный эксперимент, т.е. смешивать в одной пробе образцы с разными метками. Последнее позволяет минимизировать время анализа и повысить техническую воспроизводимость количественных экспериментов и, соответственно, получить более точные полуколичественные оценки.

С тех пор как была показана линейная зависимость между числом спектральных идентификаций и концентрацией белка в образце, этот параметр (т.е. число спектров фрагментации на белок) используется в качестве простой количественной оценки [3]. Это наблюдение дало начало целой серии методов, основанных на подсчёте спектров, которые различаются между собой способами нормирования [5–7]. Например, в методе нормализованных спектральных интенсивностей белков (Normalized Spectral Abundance Factor, NSAF) количественный индекс представляет собой отношение числа пептид-спектральных совпадений (Spectral Count, *SC*) на белок к его длине *L*, которое затем нормализуется на сумму значений *SC/L* для всех белков, идентифицированных в сложной смеси [6].

$$\mathrm{NSAF} = \left(\frac{SC}{L}\right)_k \frac{1}{\sum_{i=1}^{n}\left(\frac{SC}{L}\right)_i}.$$

Другой метод, известный как нормализованный спектральный индекс (Spectral Index Normalized, SI), сочетает в себе три параметра: число уникальных пептидов на белок, число спектров фрагментации на пептид и суммарную интенсивность ионов-фрагментов, совпавших с теоретическими фрагментами [7]:

$$\mathrm{SI} = \sum_{k=1}^{pn}\left(\sum_{j=1}^{sc} i_j\right),$$

где i_j – интенсивность пика j-го фрагмента, SC – число пептид-спектральных совпадений на пептид k, а pn – число пептидов, идентифицированных для данного белка. Чтобы рассчитать относительное содержание белка в образце, значения SI нормализуют по длине белка L и сумме SI всех идентифицированных белков.

$$\mathrm{SI_N} = \frac{SI}{\sum_{i=1}^{n} SI_i}\frac{1}{L}.$$

Методы, основанные на подсчёте спектров, до сих пор являются одними из наиболее распространённых подходов в количественном анализе белков.

Наиболее серьёзным недостатком программных конвейеров, в которых используется подсчёт спектров фрагментации, является неудовлетворительная воспроизводимость результатов идентификации пептидов и белков между экспериментами. Основной причиной является плохое покрытие белковой последовательности идентифицированными пептидами, что связано со стохастической природой масс-спектрометрических данных. Так, в различных наборах данных 20–50% всех белков могут быть идентифицированы лишь по одному пептиду, причём в разных экспериментах это могут быть разные белки [8]. Возникает проблема так называемых «отсутствующих значений» (missing value problem), которые необходимо чем-то замещать, прежде чем выполнять статистический анализ. Строго говоря, отсутствующие значения могут быть исключены из анализа, или анализ может быть выполнен без их замещения. Однако в методах Spectral Count, основанных на подсчёте спектров фрагментации, белки, регуляция которых сильно изменяется при воздействии на клетки, могут находиться за пределами обнаружения в образцах, не подвергавшихся стрессу. В результате исключение белков с отсутствующими значениями приведёт к потере существенной доли отклика на стресс, а анализ без замещения для значительной части гипотез приведёт к нарушению условий применимости статистических тестов (например, требование на минимальное число точек в выборке, распределение по Гауссу). В свою очередь, замена отсут-

ствующих значений становится источником больших погрешностей при оценке кратных изменений количественного содержания белков в программных конвейерах, где количественному анализу предшествует идентификация белков. Существует большое количество различных стратегий замещения отсутствующих значений и даже классификация источников их происхождения [9]. В самом простом случае отсутствующее значение замещается, например минимальным значением интенсивности, измеренным в конкретном ВЭЖХ-МС эксперименте. Строго говоря, такое произвольное замещение гарантирует не только некорректный расчёт кратных изменений (Fold Change, FC), но будет отражаться и на статистической значимости: чем сильнее отличаются значения средних и дисперсии, тем меньше *p*-value, и наоборот. Как следствие, алгоритмы восстановления данных могут влиять на количество и ложноположительных, и ложноотрицательных признаков в результатах количественного анализа за счёт белков с восстановленными значениями.

В последнее время наиболее популярным решением является стратегия замещения отсутствующих значений на этапе, предшествующем идентификации пептидов. Стратегия основана на выравнивании наборов экспериментальных данных по времени хроматографического удерживания и массы пептида в пределах приборной погрешности (Match-Between-Runs, MBR) [10]. Ионы пептидов одинаковой первичной структуры детектируются масс-спектрометром в один и тот же момент времени. В хроматограмме, построенной по ионному току для конкретного соотношения *m/z* ионов пептида в зависимости от времени t, будет наблюдаться пик, высота и площадь которого коррелирует с концентрацией пептида. Данный факт широко используется для количественных оценок концентраций пептидов и белков [1, 11]. Совокупность параметров (t, *m/z*) представляет собой признак детектирования пептида (Peptide Feature). Если в наборе экспериментальных данных присутствует такой признак, и он имеет высокую степень сходства с пептидными признаками из других экспериментов этой же серии, но ему не была сопоставлена пептидная идентификация, то отсутствующему значению присваивается интенсивность этого схожего признака. Такая стратегия позволяет свести долю отсутствующих значений в данных к 2–3%, что позволяет безболезненно исключить их из дальнейшего рассмотрения [12]. В ряде исследований было показано, что такая процедура способна внести в данные существенное число ошибочных замещений, которые тем не менее успешно отфильтровываются на последующих стадиях анализа [13]. Методы, основанные на использовании хроматографических интенсивностей, позиционируют в безметочной протеомике как более точные и чувствительные, особенно при использовании масс-спектрометров высокого разрешения 14–16]. В их основе, как правило, лежит использование байесовской статистики для определения количественных изменений в наборах пептидов, которые предположительно связаны с одной протеоформой [15], и вводится концепция апостериорных вероятностей как замена статистически необоснованному выбору порога кратных изменений концентраций белков [16]. Отметим, что источниками ошибок для этих методов могут быть несовершенства алгоритмов выравнивания экспериментальных данных и детектирования пептидных признаков, а также технические особенности работы прибора. В связи с этим необходимы нормировка данных, эффективный контроль качества пептидной выборки и уровня ложноположительных результатов в количественном анализе.

Известно, что вероятность совершения ложных статистических выводов значительно возрастает при одновременном тестировании большого числа гипотез [17]. В статистике определяют ошибки I-го и II-го рода. Ошибка I-го рода совершается при отклонении верной нулевой гипотезы (ложноположительный результат), тогда как ошибка II-го рода возникает при принятии неверной нулевой гипотезы (ложноотрицательный результат). Смысл всех существующих методов контроля ложноположительных результатов сводится к установлению более жёсткого статистического порога. Так, в поправке Бонферрони пороговая статистическая значимость для каждого теста находится делением стандартного значения порога на количество тестов. Коррекция Бонферрони является консервативной и приводит к высокому уровню ложных отрицаний при большом числе сравнений. Другой способ, метод Холма, пошагово вычисляет уровни значимости $\alpha'(i)$ в зависимости от ранга гипотезы i, $\alpha'(i) = \alpha/(m\text{-}i\text{+}1)$, где α – статистическая значимость, m – число гипотез. Наиболее популярной мерой контроля ошибок первого рода является частота ложноположительных результатов (false discovery rate, fdr), которая определяется как ожидаемая доля неправильно отклоненных нулевых гипотез среди всех отклонений. В методе Бенджамини–Хохберга для каждого номера гипотезы i и заданного значения q проверяется условие $p(i) \leq iq/m$, где m – общее количество тестов. Метод приме-

няется повсеместно в биологических исследованиях, в том числе и в задачах количественного анализа протеомных данных.

Целесообразность использования фиксированных статистических порогов при анализе данных до сих пор горячо обсуждается научным сообществом [18]. За время нашей работы в области количественной протеомики [19–21] мы также пришли к выводу, что использование фиксированного порога $p < 0,05$ для любого набора данных может приводить к потере потенциально ценной биологической информации или вносить в данные статистический шум.

В данной работе мы рассмотрели методы безметочного количественного анализа белков, интегрированные в различные протеомные поисковые конвейеры для эффективного определения уровня ошибок I-го рода в протеомных данных. В частности, мы показываем, что ошибки I-го рода представляют серьезную проблему при анализе раннего отклика на стресс: в случаях, когда протеомный ответ слабо выражен. В работе также предложен простой и наглядный способ оптимизации статистического порога на основе метода Бенджамини–Хохберга, анализа плотности распределения точек на диаграммах рассеяния и генных онтологий (ГО). Предложенная методика оптимизации статистического порога будет полезна исследователям, занимающимся количественным анализом протеомных данных.

МАТЕРИАЛЫ И МЕТОДЫ

Образцы. Были использованы экспериментальные данные для модельных образцов *Saccharomyces cerevisiae* из исследования iPRG-2015 (https://abrf.org/sites/default/files/temp/RGs/iPRG/iprg2015_study_instructions_final.pdf). В этом исследовании четыре триптических гидролизата шести белков в различных известных концентрациях были смешаны с одинаковым количеством триптического гидролизата белков дрожжей. Этот набор данных является количественным стандартом сравнительной протеомики.

Для сравнения безметочных методов использовались клеточная модель нормальных астроцитов (предоставлена для протеомного анализа сотрудниками Института молекулярной биологии им. Энгельгардта) и дикий тип *S. cerevisiae*. Клетки астроцитов подвергались обработке интерфероном-α в течение 24 ч, согласно стандартному протоколу [20], и являлись в данном исследовании моделью позднего отклика клеток на стресс. Клетки дрожжей дикого типа обрабатывали этанолом в течение 1,5 ч и являлись моделью раннего отклика клеток на стресс [19].

Идентификация пептидов и белков. Экспериментальные данные формата (raw) конвертировались в форматы (mgf) и (mzML) с помощью программы MSConvert, используя параметры по умолчанию [22]. Для идентификации пептидов использовались базы данных белков человека и дрожжей SwissProt (sp) (загружены 16.10.2019). Для работы с модельными образцами дрожжей использовалась база из исследования iPRG-2015. Поиски выполнялись по объединенным базам (sp+sp_reverse); базы ложных белков создавались считыванием sp справа налево (sp_reverse). Для поиска пептид-спектральных совпадений использовались поисковые машины MSFragger [23] и IdentiPy [24] со следующими параметрами: относительная точность измерения масс ионов-предшественников – 15 и 20 ppm, ошибка определения масс ионов-продуктов – 0,01 Да и 20 ppm для модельного образца дрожжей из iPRG-2015 и для остальных данных соответственно. Остальные параметры – по умолчанию. Для пост-обработки результатов поиска использовалось программное обеспечение Scavager [25] и Percolator [26]. Расчёт индексов NSAF реализован в Scavager, статистический анализ выполнялся, как опубликовано ранее [19, 20]. Отсутствующие значения количественных индексов белков, NSAF, замещались минимальным значением индекса в реплике и преобразовывались в $\log_{10}$-шкалу с последующим выравниванием глобального распределения относительно нуля и нормировки на стандартное отклонение. Для количественных алгоритмов Diffacto [15] и Triqler [16] интенсивности пептидов извлекались автоматически в программном конвейере IdentiPy/Scavager. Использование Triqler было адаптировано для программы Scavager. Отсутствующие значения хроматографических интенсивностей пептидов замещались с использованием подхода MBR после идентификации пептид-спектральных совпадений, преобразовывались в $\log_{10}$-шкалу и нормализовались с помощью методов, реализованных в Diffacto и Triqler. Таблицы с идентификациями пептидов и белков поисковым конвейером X!Tandem/MPscore взяты из ранее опубликованной работы [21]. Анализ генных онтологий выполнен с помощью программы GOrilla; порогом статистической значимости обогащения терминов ГО было принято $\alpha = 0,05$ с учётом корректировки на множественные сравнения по Бенджамини–Хохбергу [27].

РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЯ

Использование модельных образцов *S. cerevisiae* для экспериментального определения ошибок I-го рода. Для экспериментального определения ошибок I-го рода были использованы четыре модельных образца, каждый из которых содержал одинаковое количество триптического гидролизата *S. cerevisiae* (фоновые белки), смешанного с шестью белками других организмов в известных и различных концентрациях. Эти шесть белков в дальнейшем будут называться контрольными. Различные поисковые машины (MSFragger [23], IdentiPy [24]) и программы для пост-обработки идентификаций (Scavager [25], Percolator [26]) были объединены в пять программных конвейеров, чтобы оценить, как изменяются результаты количественного анализа в зависимости от использования того или иного программного продукта.

На рисунке 1, *а* для разных протеомных конвейеров показаны доли *k* правильно определенных контрольных белков (true positive, tp*)* среди общего числа белков N, которые преодолели статистический порог. Доля таких белков варьируется от 0% до 100% для разных конвейеров, а также имеет существенный разброс в пределах одного конвейера для разных попарных сравнений образцов (макс. ± 40%). Логично предположить, что такой разброс может быть связан с соотношениями концентраций контрольных белков в образцах. На рис. 1, *а* ошибкам I-го рода соответствуют значения 1–tp. На рис. 1, *б* показаны ошибки 2-го рода – доли X–k контрольных белков (false negatives, fn), не преодолевших статистический порог среди ожидаемых контрольных белков X. Ожидаемыми контрольными белками считались белки, для которых отношения фактических концентраций FC были вне диапазона 0,846 < FC < 1,182. Граничные значения интервала соответствуют фактическому соотношению концентраций контрольных белков в модельных образцах. Было принято допущение, что данные значения могут представлять собой предельные значения чувствительности количественных методов. Результаты показали, что действительно методы, основанные на анализе хроматографических интенсивностей, демон-

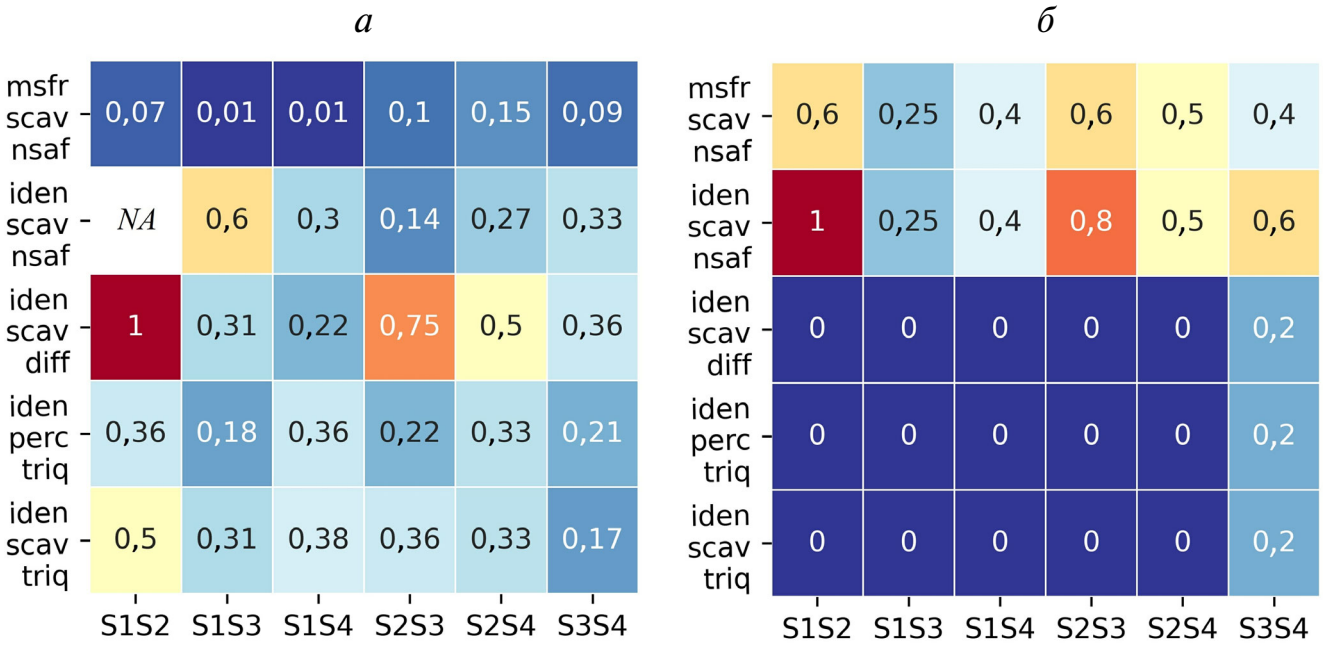

Рис. 1. Доли правильных (true positive, tp) белков (*а*) и ложно отрицательных (false negative, fn) результатов (*б*) для разных программных конвейеров в попарных сравнениях модельных образцов дрожжей. *а* – Доля tp контрольных белков k в общем числе белков N, которые преодолели статистический порог fdr_{BH} (*q*-value) = 0,05 и порог кратных изменений FC = 1,0; tp = k/N. «*NA*» соответствует k, N = 0. Значения 1–tp соответствуют ошибкам I-го рода. *б* – Доля fn ложноотрицательных результатов X–k относительно ожидаемых белков X; fn = (X–k)/X. Значения fn соответствуют ошибкам II-го рода. Ожидаемыми дифференциально регулированными белками считались белки, для которых отношения фактических концентраций FC были вне диапазона 0,846 < FC < 1,182. Обозначения: поисковые машины MSFragger (msfr), IdentiPy (iden); программы для пост-обработки пептид-спектральных совпадений Scavager (scav), Percolator (perc); количественные методы NSAF (nsaf), Triqler (triq), Diffacto (diff); попарные сравнения модельных образцов *SiSj*, *i*≠*j*. Цветовая шкала соответствует изменению значений от 0,0 до 1,0. (С цветными вариантами рис. 1–6 можно ознакомиться в электронной версии статьи на сайте: http://sciencejournals.ru/journal/biokhsm/.)

стрируют как высокую чувствительность к слабым изменениям концентраций (FC = 1,2), так и более высокую точность определения множества белков с изменениями концентраций.

Следует учитывать, что в общем случае белки фона, преодолевшие критерии отбора, могут быть идентифицированы по пептидам, часто встречаемым в протеомных образцах контаминантов, либо совпадать с пептидами контрольных белков. Проверка по базе данных таких контаминантов (ftp://ftp.thegpm.org/fasta/cRAP) показала, что пептиды, по которым идентифицированы фоновые белки, в белках-контаминантах отсутствуют. Пересечений с последовательностями контрольных белков также обнаружено не было.

Одним из самых простых способов оценить точность определения кратных изменений концентрации белка является использование модельных образцов с белками известных концентраций. На рис. 2 показаны корреляции между рассчитанным и фактическим соотношениями концентраций для каждого белка во всех парах образцов из исследования iPRG-2015. Наилучшие результаты с угловыми коэффициентами близкими к единице были получены при использовании программных конвейеров, в которые интегрированы алгоритмы количественного анализа интенсивностей пептидов, Diffacto и Triqler. Использование индекса NSAF приводит к двукратному занижению расчётных значений $|\log_2 FC|$, что необходимо учитывать в дальнейшем анализе. Отметим, что для всех конвейеров наблюдался разброс белков фона в широком диапазоне значений $|\log_2 FC|$, что согласуется с результатами на рис. 1 и диктует необходимость выбора более строгих статистических критериев.

Для оптимизации критериев отбора исследуем зависимость доли контрольных белков k в общем числе белков N, которые преодолевают статистический порог, от значения этого порога, fdr_{BH} (или *q*-value) (рис. 3, *а*). Согласно полученным зависимостям, статистические пороги в некоторых случаях могут быть выбраны так, чтобы отсеять фоновые (ложные) белки, максимально сохранив контрольные (правильные): оптимальным считается порог, при котором доля контрольных белков $tp \rightarrow max$ при $fn \rightarrow min$. Так оптимальными значениями статистических порогов, усреднёнными по попарным сравнениям образцов, были приняты 0,003, 0,03 и 0,05 для конвейеров, в которые интегрированы, соответственно, алгоритмы Triqler, Diffacto и NSAF.

Была исследована также зависимость доли контрольных белков от порогового значения кратных изменений $|\log_2 FC|$ для оптимальных порогов статистической значимости (рис. 3, *б*). Наблюдалась ожидаемая зависимость: чем выше порог отбора кратных изменений концентраций FC, тем меньше ложноположительных результатов в количественном анализе. Так, например, при $|\log_2 FC| > 3{,}0$ доля ложноположительных результатов в количественном анализе для большинства рассмотренных конвейеров близка к нулю, тогда как для диапазона $0{,}5 < |\log_2 FC| < 2{,}0$ она может составлять от 10% до 90%.

Поздний и ранний отклик на стресс на примере анализа количественных изменений белков в клеточных моделях астроцитов и *S. cerevisiae*. Для исследования общих закономерностей в распределениях ошибок I-го рода были использованы протеомные данные для клеточных моделей астроцитов и дрожжей. Модель нормальных астроцитов являлась примером позднего отклика на стресс, когда изменения на трансляционном уровне сильно выражены (24 ч интерферон-α). Клетки *S. cerevisiae* дикого типа, инкубированные с этанолом в течение 1,5 ч, моделировали слабые изменения на уровне протеома (ранний отклик на стресс). На рис. 4 для данных моделей показано сравнение классических методов корректировки на множественные сравнения. Для панорамного количественного анализа протеомов (с глубиной идентификации более 2000 белков) применение метода Бенджамини–Хохберга является лучшим выбором. Однако обратим внимание, что в случае раннего отклика стандартные статистические критерии отбора $fdr_{BH} < 0{,}05$ проходит в 2–3 раза больше белков, чем в случае позднего. При этом 40% таких белков сосредоточено в области слабых количественных изменений $0{,}5 < |\log_2 FC| < 2{,}0$, где ожидается высокий уровень ложно отвергнутых гипотез (согласно рис. 3). При использовании алгоритма Triqler (рис. 4, *б*) этот эффект выражен слабее (206 и 227 белков, соответственно, в позднем и раннем откликах). Такое наблюдение также согласуется с результатами анализа модельных образцов дрожжей и подтверждает, что в раннем отклике ожидается большее число ложноположительных результатов.

Оптимизация критериев отбора и оценка биологической релевантности дифференциально регулированных белков на основе анализа генных онтологий. Для оценки биологической релевантности молекулярных изменений в биологических моделях в ответ на стресс был использован анализ ГО. Как видно из рис. 5, ужесточение статистического порога приводит к увеличению относительного числа дифференциально экспрессированных генов, вовлеченных в обогащенные биологические процессы. Этот эффект наиболее заметен при анализе белков с не-

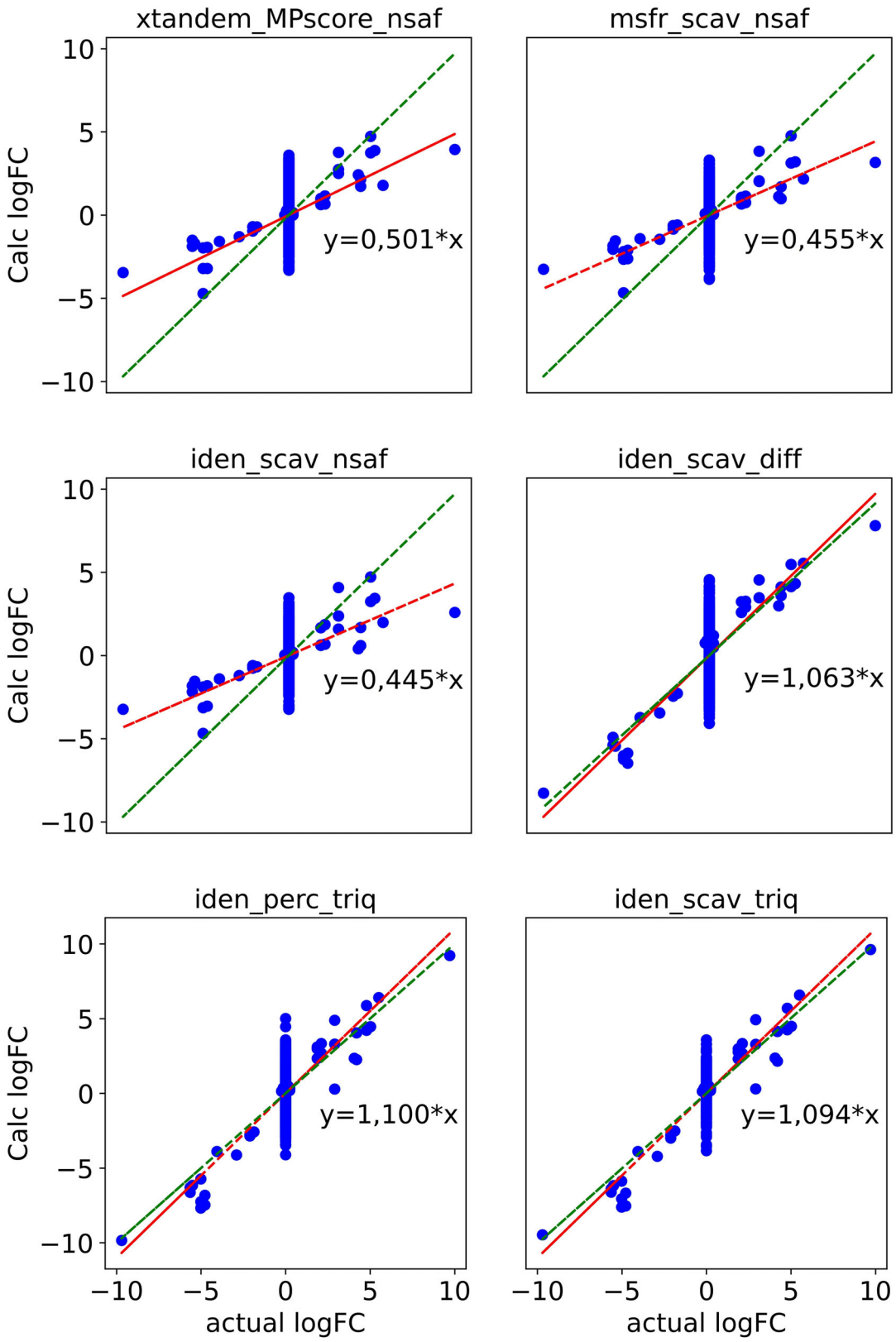

Рис. 2. Сравнение расчётных и фактических соотношений концентраций контрольных и фоновых белков. Расчётные значения для каждого белка усреднены по техническим репликам. Для фоновых белков фактическое отношение концентраций принято за единицу. Красный и зелёный пунктир соответствуют линейной аппроксимации точек и ожидаемой зависимости y = x соответственно

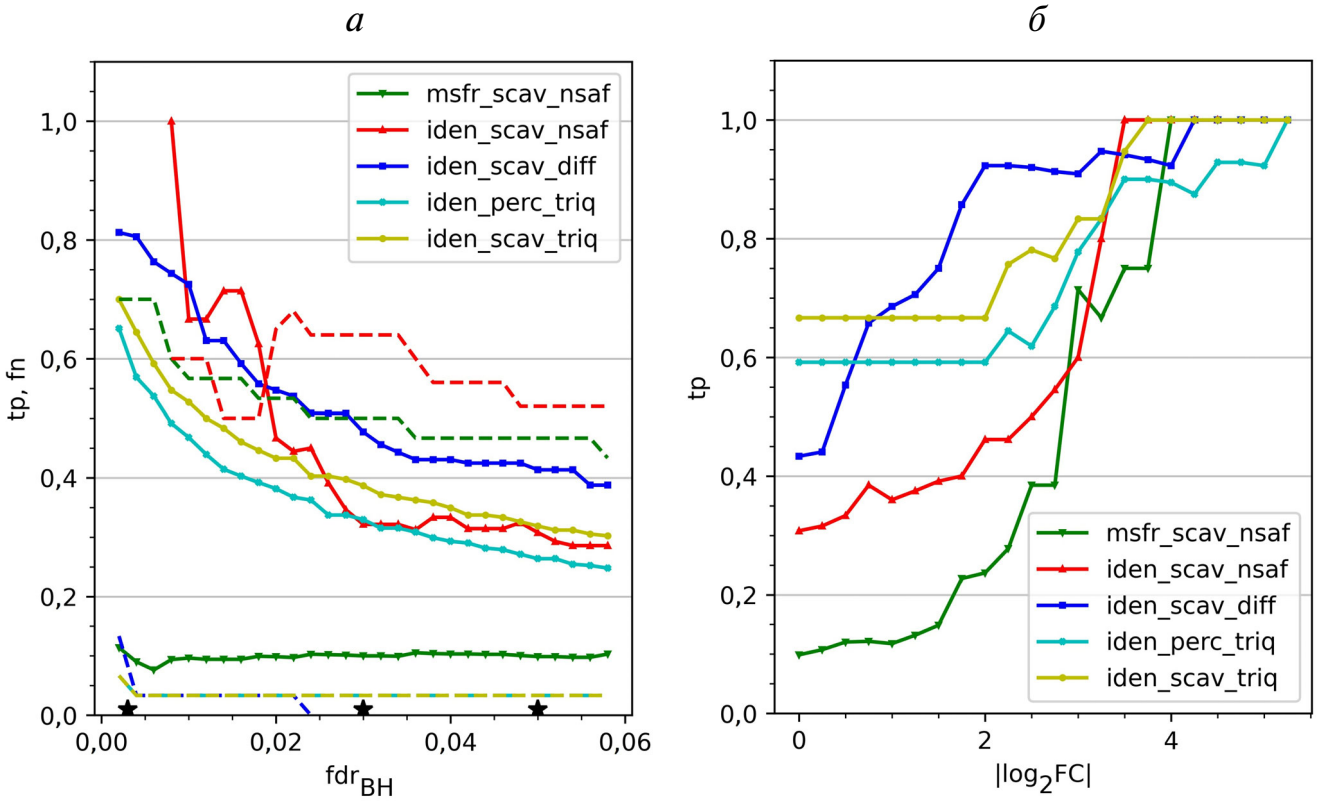

Рис. 3. Зависимость доли правильно определенных (true positive, tp) белков и ложноотрицательных (false negative, fn) результатов от критериев отбора дифференциально регулированных белков: *а* – статистический порог, fdr_{BH} (q-value); *б* – пороговое значение кратных изменений, $|\log_2 FC|$, при фиксированном (оптимизированном) пороге статистической значимости. Оптимизированные пороги обозначены ★ и равны 0,003 (Triqler), 0,03 (Diffacto), 0,05 (NSAF). Доля контрольных белков k среди общего числа статистически значимых белков N: tp = k/N (сплошная линия). Доля ложноотрицательных результатов X–k относительно ожидаемых белков X: fn = (X–k)/X (пунктир). Значения tp и fn усреднены по попарным сравнениям образцов из iPRG-2015

большими изменениями концентраций (группы down – в астроцитах, up и down – в пекарских дрожжах рис. 5). Данный результат хорошо согласуется с результатами анализа модельных образцов дрожжей и подтверждает, что ужесточением статистического порога возможно сократить число ложноположительных результатов в области малых кратных изменений. Этой тенденции не подчиняются результаты анализа раннего отклика методом Triqler, для которого, очевидно, в условиях раннего стресса в выбранном диапазоне пороговых значений статистической значимости такое правило не работает (рис. 5, *б*). Также в выбранном диапазоне наблюдается определенная несогласованность и для результатов, полученных методом NSAF.

ОБСУЖДЕНИЕ РЕЗУЛЬТАТОВ

Анализ модельных образцов дрожжей показал, что ошибки I-го рода неравномерно распределены по значениям кратных изменений концентраций белков и могут представлять серьёзную проблему для диапазона слабых изменений концентраций, $0{,}5 < |\log_2 FC| < 2{,}0$. Это ожидаемый результат. На практике такой эффект обычно компенсируют отбором по пороговому значению кратных изменений (FC), который, как правило, варьируется в интервале 1,5–2,5 и выбирается произвольным образом. Однако после такого отбора рассчитанный уровень ложноположительных результатов фактически следует признавать недействительным [16]. Такой подход также будет негативным образом сказываться на результатах анализа раннего отклика на стресс, когда изменения на уровне протеома слабо выражены. Фактически, в случае раннего отклика такой отбор отсеет не только ложноположительные результаты, но и существенную часть правильных. Более эффективной стратегией с этой точки зрения является оптимизация статистического критерия, характеризующего степень доверия рассчитанным значениям кратных изменений.

Метод Бенджамини–Хохберга, несомненно, является оптимальным выбором для задач

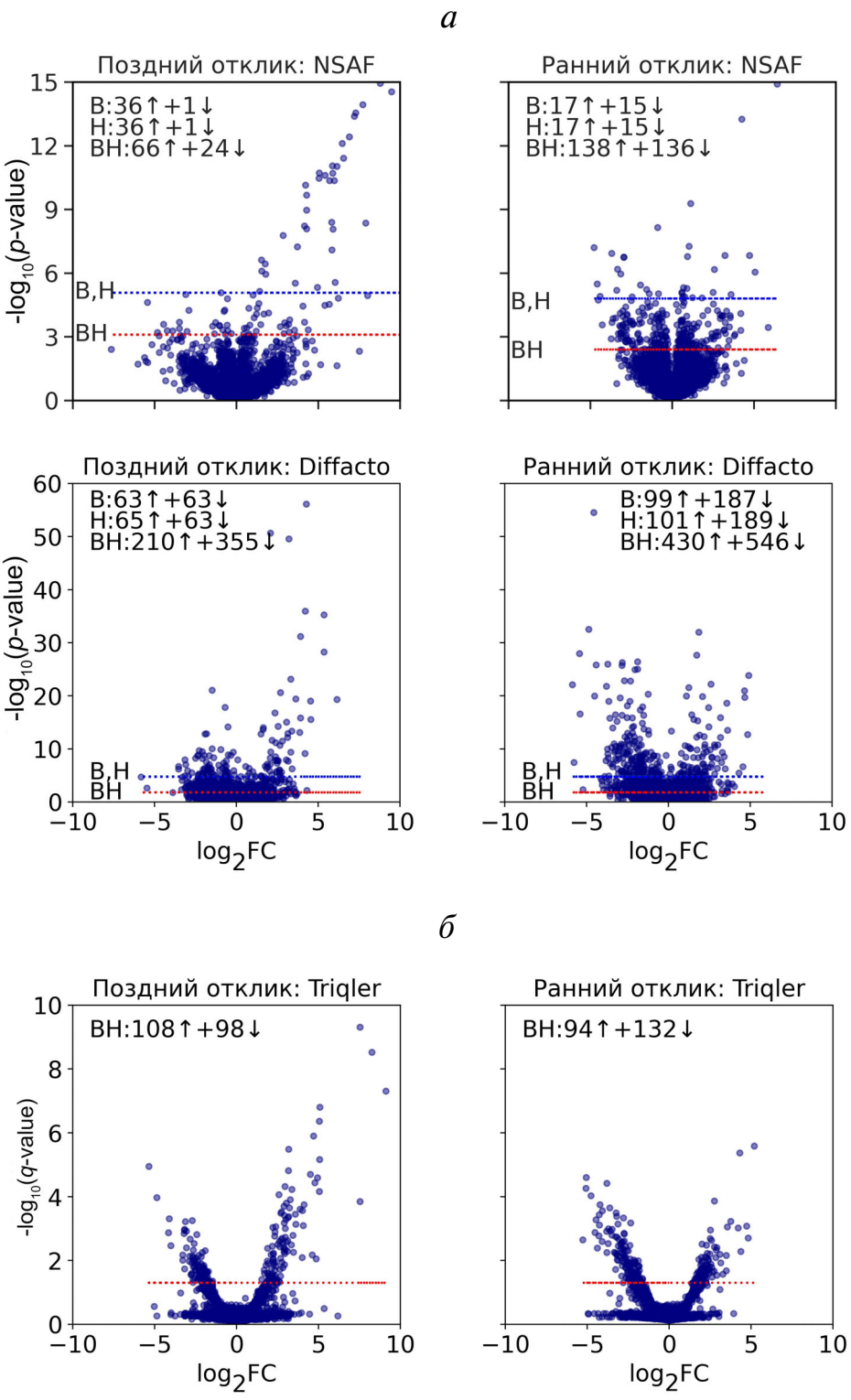

Рис. 4. Сравнение классических методов контроля уровня ложноположительных результатов в случае позднего (число тестов = 3500) и раннего (число тестов = 2300) отклика клеток на внешнее воздействие (*а*). Результаты количественного анализа с использованием метода Triqler для случаев позднего и раннего отклика клеток на стресс (*б*). Идентификация и пост-поисковая валидация белков выполнены программным конвейером Identipy/Scavager. Горизонтальный пунктир соответствует порогу 0,05; B – поправка Бонферрони, H – метод Холма, BH – метод Бенджамини–Хохберга. Стрелки ↑↓ соответствуют направлению дифференциальной регуляции белков. Для NSAF значения FC скорректированы согласно угловым коэффициентам (рис. 2). Ось X: кратное изменение концентраций белков FC в $\log_2$-шкале. Ось Y: статистическая значимость, *p*-value или *q*-value, в отрицательной $\log_{10}$-шкале

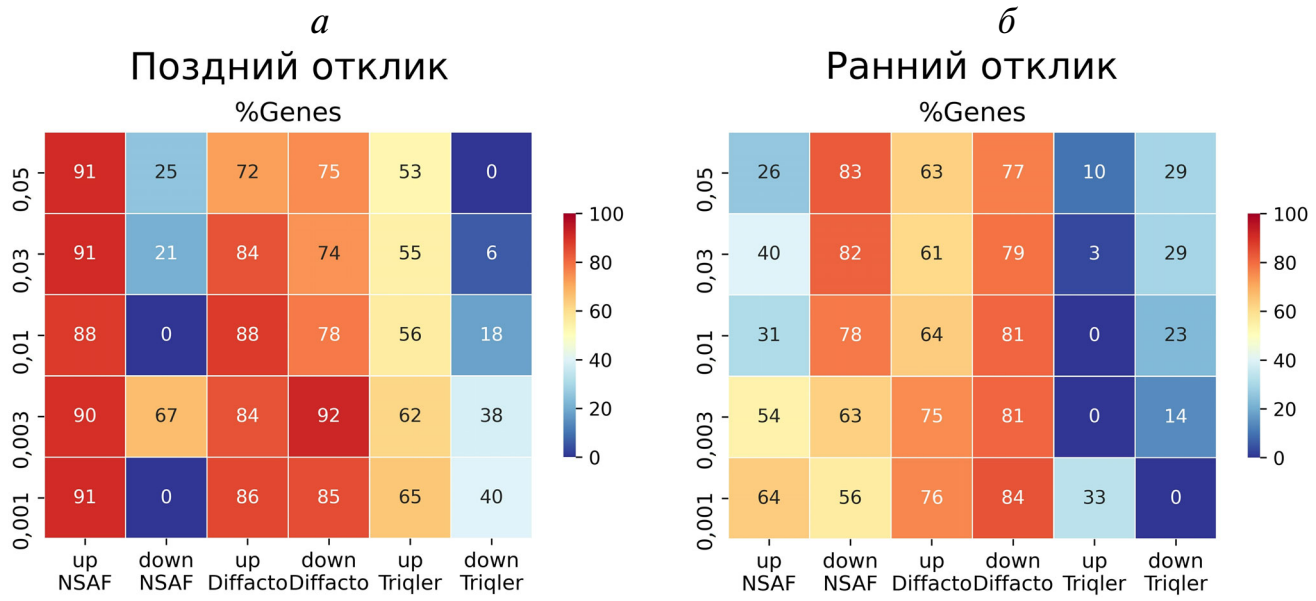

Рис. 5. Зависимость числа генов, вовлечённых в обогащённые биологические процессы, от количества ошибок I-го рода в позднем и раннем отклике клеток на стресс. Для NSAF значения FC скорректированы, согласно угловым коэффициентам на рис. 2. Поздний отклик: инкубация астроцитов 24 ч с IFNα (*а*), ранний отклик: инкубация *S. cerevisiae* 1,5 ч с этанолом (*б*). Обозначения: %Genes – число генов, вовлечённых в обогащённые ГО, относительно общего числа дифференциально экспрессированных генов

панорамной количественной протеомики [28, 29], однако использование одинакового статистического порога для отбора дифференциально регулированных белков способно привести к потере потенциально ценной биологической информации. В нашей работе показано, что оптимальный статистический порог не только позволяет сократить число ложноположительных результатов в области малых кратных изменений концентраций (рис. 3), но и коррелирует с увеличением относительного числа дифференциально экспрессированных генов, вовлечённых в обогащение ГО (рис. 5). Использование различных методов и стратегий количественного анализа, в том числе поисковых машин, программ пост-поисковой обработки данных, стратегий восстановления отсутствующих значений и, в наибольшей степени, различных подходов для статистической обработки количественных данных существенным образом отражается как на точности рассчитанных кратных изменений концентраций белков, так и на значениях статистической значимости (рис. 2 и 4). В случае раннего отклика клеток на стресс стандартный статистический порог может проходить в 2–3 раза больше белков, чем в случае позднего, причём до 40% таких белков сосредоточено в области малых изменений концентрации белков (рис. 4), где ожидается больше ложноположительных результатов (рис. 2). Возникает вопрос, как отбирать дифференциально регулированные признаки в условиях разной чувствительности количественных методов и программных конвейеров, и как выбирать оптимальный статистический порог в случае анализа реальных биологических моделей. На основе полученных результатов мы предлагаем дополнять протокол отбора дифференциально экспрессированных признаков анализом распределения плотности точек на диаграммах рассеяния (рис. 6). В основе такого анализа лежат два предположения: (1) плотное множество точек на диаграмме рассеяния ($\log_2$FC, $\log_{10}$ *p*-value) соответствует неизменной части протеома; и (2) граница между «плотным» и «разреженным» множеством точек определяется как сумма третьего квартиля и полутора межквартильных расстояний распределения плотности точек (отклонение +2,698σ от медианы). Если порог 0,05 fdr_{BH} проходит ниже границы, то в данных ожидается статистический шум, а уровень ошибок I-го рода занижен (что актуально для метода Diffacto, рис. 4, *а*). Если порог 0,05 fdr_{BH} проходит выше границы, то предполагается высокий уровень ошибок II-го рода (как в случае с методами NSAF и Triqler, рис. 4, *б*). В таких случаях целесообразно будет изменить порог отбора в сторону ужесточения (для Diffacto) или расслабления (для NSAF и Triqler). Новый порог можно считать оптимальным и обоснованным, если его использование приводит к увеличению доли дифференци-

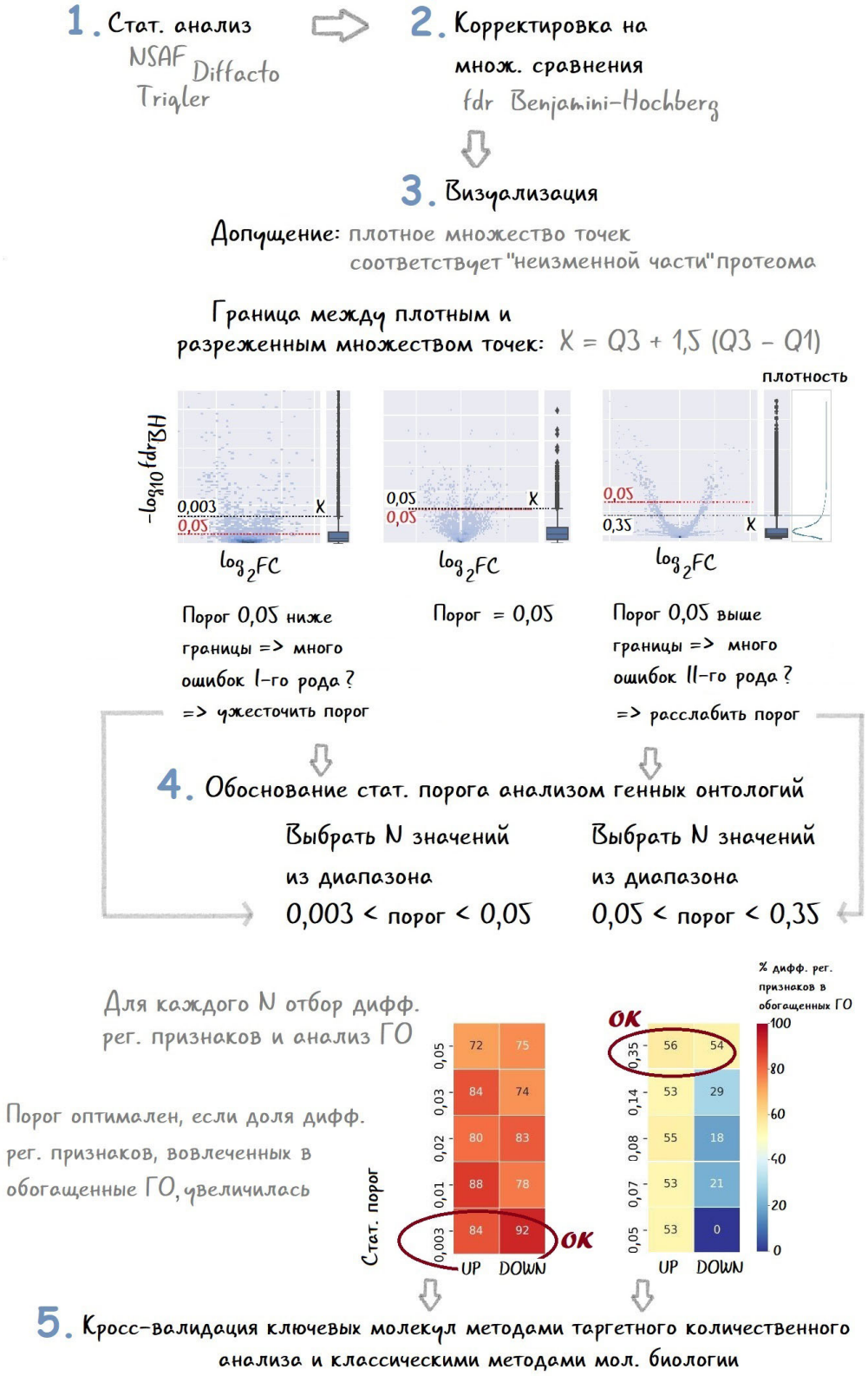

Рис. 6. Схема отбора дифференциально регулированных белков и оптимизации статистического порога на основе анализа плотности распределения точек на диаграммах рассеяния и анализа генных онтологий для количественной протеомики. Обозначения: NSAF, Diffacto, Triqler – методы безметочного количественного анализа; fdr_{BH} – ожидаемая доля ложных отклонений гипотез, скорректированная на множественные сравнения методом Бенджамини–Хохберга; *Q1*, *Q3* – первый и третий квартили распределения плотности точек по оси $-\log_{10}fdr_{BH}$; FC – кратное изменение количественного содержания белка

ально регулированных признаков, вовлечённых в обогащённые термины ГО, что, в свою очередь, должно способствовать исчерпывающей интерпретации количественных данных. Согласно предложенной методике, были определены новые статистические пороги для моделей раннего и позднего отклика, а анализ обогащений ГО дифференциально регулированных признаков, отобранных согласно оптимизированным порогам, показал увеличение доли дифференциально регулированных белков в обогащённых ГО (рис. S1 в Приложении). Такой подход предлагается интерпретировать как способ качественной визуализации уровня ошибок I-го и II-го рода в результатах, полученных в условиях отличающихся статистических дизайнов с помощью методов, обладающих различной статистической мощностью.

Применение такого подхода не бесспорно. Во-первых, в работе мы используем статистический анализ обогащения ГО как источник данных, комплементарных результатам протеомного анализа, считая положительную корреляцию подтверждением релевантности отобранных дифференциально регулированных признаков. В то же время статистический анализ обогащений ГО сам содержит ошибки I-го и II-го рода. Во-вторых, если к использованию более жёсткого порога можно отнестись снисходительно, то идея его расслабления вызывает обоснованный протест, так как уменьшая уровень ошибок II-го рода, мы можем существенно увеличить долю ложноположительных результатов (рис. S1 в Приложении). Поэтому мы предлагаем использовать анализ плотности распределения точек, как дополнительную качественную характеристику мощности статистического комплекса, применяемого для количественного анализа в протеомике. В сочетании с анализом ГО такой подход позволит обосновать выбор другого статистического порога, способствующего более полной интерпретации количественных данных. Однако следует понимать, что последнее подразумевает привлечение дополнительных статистических и/или экспериментальных методов для валидации результатов такого анализа. На этапе генерации и проверки ключевой гипотезы для этих целей целесообразно использовать методы таргетного количественного анализа, такие как метод мониторинга множественных переходов (Multiple reaction monitoring, MRM), иммунологические методы и полимеразную цепную реакцию.

В заключение отметим, что анализ генных онтологий довольно давно служит простой и интуитивно понятной метрикой биологической релевантности найденных количественных изменений [30, 31]. Естественным ограничением данного метода интерпретации количественных данных являются новые и ранее не описанные регуляторы и участники биологических процессов – это ситуация аналогичная «поиску под фонарем» – мы интерпретируем ту часть количественных данных, для которой известна классификация/аннотация (экспериментально установлена, предсказана с некоторой точностью). Однако это не препятствует применению анализа ГО в рамках предложенного в работе подхода. Мы считаем, что анализ плотности распределения количественных изменений может быть полезен и в других ОМИКС-подходах, где возникает схожая проблематика.

Финансирование. Исследование выполнено при поддержке Российского научного фонда (грант №20-14-00229).

Благодарности. Культура астроцитов для протеомного анализа была предоставлена сотрудниками ИМБ РАН, проф. П.М. Чумаковым и А.В. Соболевой в рамках работ, поддержанных Российским фондом фундаментальных исследований (грант №18-29-01059-мк).

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Соблюдение этических норм. Настоящая статья не содержит описания выполненных авторами исследований с участием людей или использованием животных в качестве объектов.

Дополнительные материалы. Приложение к статье на английском языке опубликовано на сайте журнала «Biochemistry» (Moscow) (http://protein.bio.msu.ru/biokhimiya/) и на сайте издательства Springer (https://link.springer.com/journal/10541), том 86, вып. 3, 2021.

СПИСОК ЛИТЕРАТУРЫ

1. Nikolov, M., Schmidt, C., and Urlaub, H. (2012) Quantitative mass spectrometry-based proteomics: an overview, in *Quantitative Methods in Proteomics* (Marcus, K., ed.) Humana Press, Totowa, NJ, pp. 85-100, doi: 10.1007/978-1-61779-885-6_7.
2. Zhang, X., Fang, A., Riley, C. P., Wang, M., Regnier, F. E., and Buck, C. (2010) Multi-dimensional liquid chromatography in proteomics – a review, *Anal. Chimica Acta*, **664**, 101-113, doi: 10.1016/j.aca.2010.02.001.
3. Podwojski, K., Stephan, C., and Eisenacher, M. (2012) Important issues in planning a proteomics experiment: statistical considerations of quantitative proteomic data, in *Quantitative Methods in Proteomics*, (Marcus, K., ed.), Humana Press, Totowa, NJ, pp. 3-21, doi: 10.1007/978-1-61779-885-6_1.

4. Tuli, L., and Ressom, H. W. (2009) LC–MS based detection of differential protein expression, *J. Proteomics Bioinform.*, **02**, 416-438, doi: 10.4172/jpb.1000102.
5. Ishihama, Y., Oda, Y., Tabata, T., Sato, T., Nagasu, T., et al. (2005) Exponentially modified protein abundance index (EmPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein, *Mol. Cell. Proteomics*, **4**, 1265-1272, doi: 10.1074/mcp.M500061-MCP200.
6. Griffin, N. M., Yu, J., Long, F., Oh, P., Shore, S., et al. (2010) Label-free, normalized quantification of complex mass spectrometry data for proteomic analysis, *Nat. Biotechnol.*, **28**, 83-89, doi: 10.1038/nbt.1592.
7. Trudgian, D. C., Ridlova, G., Fischer, R., Mackeen, M. M., Ternette, N., et al. (2011) Comparative evaluation of label-free SINQ normalized spectral index quantitation in the central proteomics facilities pipeline, *Proteomics*, **11**, 2790-2797, doi: 10.1002/pmic.201000800.
8. Webb-Robertson, B.-J. M., Wiberg, H. K., Matzke, M. M., Brown, J. N., Wang, J., et al. (2015) Review, evaluation, and discussion of the challenges of missing value imputation for mass spectrometry-based label-free global proteomics, *J. Proteome Res.*, **14**, 1993-2001, doi: 10.1021/pr501138h.
9. Karpievitch, Y. V., Dabney, A. R., and Smith, R. D. (2012) Normalization and missing value imputation for label-free LC-MS analysis, *BMC Bioinformatics*, **13**, S5, doi: 10.1186/1471-2105-13-S16-S5.
10. Nagaraj, N., Kulak, N. A., Cox, J., Neuhauser, N., Mayr, K., et al. (2012) System-wide perturbation analysis with nearly complete coverage of the yeast proteome by single-shot ultra HPLC runs on a bench top orbitrap, *Mol. Cell. Proteomics*, **11**, M111.013722, doi: 10.1074/mcp.M111.013722.
11. Wiener, M. C., Sachs, J. R., Deyanova, E. G., and Yates, N. A. (2004) Differential mass spectrometry: a label-free LC–MS method for finding significant differences in complex peptide and protein mixtures, *Anal. Chem.*, **76**, 6085-6096, doi: 10.1021/ac0493875.
12. Zhang, B., Käll, L., and Zubarev, R. A. (2016) DeMix-Q: quantification-centered data processing workflow, *Mol. Cell. Proteomics*, **15**, 1467-1478, doi: 10.1074/mcp.O115.055475.
13. Lim, M. Y., Paulo, J. A., and Gygi, S. P. (2019) Evaluating false transfer rates from the match-between-runs algorithm with a two-proteome model, *J. Proteome Res.*, **18**, 4020-4026, doi: 10.1021/acs.jproteome.9b00492.
14. Cox, J., and Mann, M. (2008) MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification, *Nat. Biotechnol.*, **26**, 1367-1372, doi: 10.1038/nbt.1511.
15. Zhang, B., Pirmoradian, M., Zubarev, R., and Käll, L. (2017) Covariation of peptide abundances accurately reflects protein concentration differences, *Mol. Cell. Proteomics*, **16**, 936-948, doi: 10.1074/mcp.O117.067728.
16. The, M., and Käll, L. (2019) Integrated identification and quantification error probabilities for shotgun proteomics, *Mol. Cell. Proteomics*, **18**, 561-570, doi: 10.1074/mcp.RA118.001018.
17. Chen, S.-Y., Feng, Z., and Yi, X. (2017) A general introduction to adjustment for multiple comparisons, *J. Thorac. Dis.*, **9**, 1725-1729, doi: 10.21037/jtd.2017.05.34.
18. Kennedy-Shaffer, L. (2019) Before $p < 0.05$ to beyond $p < 0.05$: using history to contextualize p-values and significance testing, *Am. Stat.*, **73**, 82-90, doi: 10.1080/00031305.2018.1537891.
19. Bubis, J. A., Spasskaya, D. S., Gorshkov, V. A., Kjeldsen, F., Kofanova, A. M., et al. (2020) Rpn4 and proteasome-mediated yeast resistance to ethanol includes regulation of autophagy, *Appl. Microbiol. Biot.*, **104**, 4027-4041, doi: 10.1007/s00253-020-10518-x.
20. Tarasova, I. A., Tereshkova, A. V., Lobas, A. A., Solovyeva, E. M., Sidorenko, A. S., et al. (2018) Comparative proteomics as a tool for identifying specific alterations within interferon response pathways in human glioblastoma multiforme cells, *Oncotarget*, **9**, 1785-1802, doi: 10.18632/oncotarget.22751.
21. Bubis, J. A., Levitsky, L. I., Ivanov, M. V., Tarasova, I. A., and Gorshkov, M. V. (2017) Comparative evaluation of label-free quantification methods for shotgun proteomics: LFQ methods for proteomics, *Rapid Commun. Mass Spectrom.*, **31**, 606-612, doi: 10.1002/rcm.7829.
22. Kessner, D., Chambers, M., Burke, R., Agus, D., and Mallick, P. (2008) ProteoWizard: open source software for rapid proteomics tools development, *Bioinformatics*, **24**, 2534-2536, doi: 10.1093/bioinformatics/btn323.
23. Kong, A. T., Leprevost, F. V., Avtonomov, D. M., Mellacheruvu, D., and Nesvizhskii, A. I. (2017) MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics, *Nat. Methods*, **14**, 513-520, doi: 10.1038/nmeth.4256.
24. Levitsky, L. I., Ivanov, M. V., Lobas, A. A., Bubis, J. A., Tarasova, I. A., et al. (2018) IdentiPy: an extensible search engine for protein identification in shotgun proteomics, *J. Proteome Res.*, **17**, 2249-2255, doi: 10.1021/acs.jproteome.7b00640.
25. Ivanov, M. V., Levitsky, L. I., Bubis, J. A., and Gorshkov, M. V. (2019) Scavager: a versatile postsearch validation algorithm for shotgun proteomics based on gradient boosting, *Proteomics*, **19**, 1800280, doi: 10.1002/pmic.201800280.
26. The, M., MacCoss, M. J., Noble, W. S., and Käll, L. (2016) Fast and accurate protein false discovery rates on large-scale proteomics data sets with Percolator 3.0, *J. Am. Soc. Mass Spectrom.*, **27**, 1719-1727, doi: 10.1007/s13361-016-1460-7.
27. Eden, E., Navon, R., Steinfeld, I., Lipson, D., and Yakhini, Z. (2009) GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists, *BMC Bioinformatics*, **10**, 48, doi: 10.1186/1471-2105-10-48.
28. Lualdi, M., and Fasano, M. (2019) Statistical analysis of proteomics data: a review on feature selection, *J. Proteomics*, **198**, 18-26, doi: 10.1016/j.jprot.2018.12.004.
29. Diz, A. P., Carvajal-Rodríguez, A., and Skibinski, D. O. F. (2011) Multiple hypothesis testing in proteomics: A strategy for experimental work, *Mol. Cell. Proteomics*, **10**, M110.004374, doi: 10.1074/mcp.M110.004374.
30. Fruzangohar, M., Ebrahimie, E., and Adelson, D. L. (2017) A novel hypothesis-unbiased method for gene ontology enrichment based on transcriptome data, *PLoS One*, **12**, e0170486, doi: 10.1371/journal.pone.0170486.
31. Gong, H., Wu, T. T., and Clarke, E. M. (2014) Pathway-gene identification for pancreatic cancer survival via doubly regularized cox regression, *BMC Syst. Biol.*, **8**, S3, doi: 10.1186/1752-0509-8-S1-S3.

PROTEOMICS OF CELLULAR RESPONSE TO STRESS: TAKING CONTROL OF FALSE POSITIVE RESULTS

I. T. Gabdrakhmanov[1], M. V. Gorshkov[2,3], and I. A. Tarasova[3]*

[1] *Skolkovo Institute of Science and Technology, 121205 Moscow, Russia*

[2] *Moscow Institute of Physics and Technology (National Research University), 141701 Dolgoprudny, Moscow Region, Russia*

[3] *Talrose Institute for Energy Problems of Chemical Physics, Semenov Federal Research Center for Chemical Physics, Russian Academy of Sciences, 119334 Moscow, Russia; E-mail: iatarasova@yandex.ru*

One of the main goals of quantitative proteomics is molecular profiling of cellular response to stress at the protein level. To perform this profiling, statistical analysis of experimental data involves multiple testing of a hypothesis about the equality of protein concentrations between the cells under normal and stress conditions. This analysis is then associated with the multiple testing problem dealing with the increased chance of obtaining false positive results. A number of solutions to this problem are known, yet, they may lead to the loss of potentially important biological information when applied with commonly accepted thresholds of statistical significance. Using the proteomic data obtained earlier for the yeast samples containing proteins at known concentrations and the biological models of early and late cellular responses to stress, we analyzed dependences of distributions of false positive and false negative rates on the protein fold changes and thresholds of statistical significance. Based on the analysis of the density of data points in the volcano plots, Benjamini–Hochberg method, and gene ontology analysis, visual approach for optimization of the statistical threshold and selection of the differentially regulated proteins has been suggested, which could be useful for researchers working in the field of quantitative proteomics.

Keywords: proteomics, bioinformatics, cell response, mass spectrometry