Биомедицинская химия, 2017 том 63, вып. 5, с. 373-378.

МУЛЬТИОМНАЯ СТРАТЕГИЯ ИССЛЕДОВАНИЯ ПРОТЕОМА КЛЕТОЧНОЙ ЛИНИИ ГЕПАТОЦЕЛЛЮЛЯРНОЙ КАРЦИНОМЫ HepG2

Е.В. Поверенная, О.И. Киселева*, Е.А. Пономаренко, С.Н. Нарыжный, В.Г. Згода, А.В. Лисица

Научно-исследовательский институт биомедицинской химии имени В.Н. Ореховича,
119121, Москва, ул. Погодинская, 10; эл. почта: olly.kiseleva@gmail.com

На сегодняшний момент исследования в области протеомики сосредоточены в основном вокруг наиболее представленных форм белков, зачастую кодируемых каноническими (неизмененными) нуклеотидными последовательностями. Накопленный массив транскриптомных и протеомных данных наряду с высоким уровнем современных технологических возможностей постгеномных исследований позволяет приступить к идентификации аберрантных форм белков. Данная работа была нацелена на оценку гетерогенности протеома HepG2, возникающей в результате реализации аберраций на белковом уровне. В качестве перспективного инструмента исследования протеома был предложен комплекс транскриптомных (RNAseq) и протеомных (2DE и MS/MS) методов.

Ключевые слова: протеом, транскриптом, транскриптопротеом, протеоформа, альтернативный сплайсинг, полиморфизм единичных аминокислот, посттрансляционные модификации

DOI: 10.18097/PBMC20176305373

ВВЕДЕНИЕ

Ключевым вопросом научных исследований в области молекулярной биологии является механизм реализации геномной информации на уровне фенотипа. С появлением высокопроизводительных методов молекулярного профилирования изучение этого вопроса и вклада транскрипционных, трансляционных и посттрансляционных процессов стало более достижимым.

В геноме человека определено около 20 тыс. белок-кодирующих генов [1], при этом половина из них характеризуются минимум одной аберрантной формой белка (протеоформой), являющейся продуктом альтернативного сплайсинга, мутаций или посттрансляционных модификаций. По приблизительным оценкам размер протеома составляет от 600 тыс. до 6 млн протеоформ [2]. Однако, несовершенство протеомных подходов зачастую позволяет констатировать лишь детекцию мастерного белка* [3]: в силу малого покрытия аминокислотной последовательности детектируемым протеотипическим пептидом идентификация конкретной протеоформы представляет собой нетривиальную задачу. Ещё одним узким местом протеомики является задача отличить белки, гены которых, в принципе, не экспрессируются в конкретном типе биологического материала при заданных условиях, от белков, которые экспрессируются в количестве, недостаточном для детекции [4]. Сложность исследования протеома усугубляется механизмами тканеспецифичной экспрессии, белок–белковыми взаимодействиями, а также субклеточной локализацией белков. Кроме того, протеомные профили динамически изменяются под влиянием внешних факторов и при развитии физиологических/патофизиологических состояний.

Одним из путей исследования гетерогенности протеома является направленный поиск только тех белков, для кодирующих генов которых показана экспрессия на уровне транскриптома. Даже принимая во внимание отсутствие корреляции между содержанием мРНК и белка вследствие ряда факторов [5], наличие в образце транскрипта указывает, что конкретный белок-кодирующий ген экспрессируется конкретно в этом образце и, следовательно, вероятность детекции продукта экспрессии этого гена на протеомном уровне повышается. Результаты транскриптомного профилирования позволяют сфокусироваться на поиске конкретной протеоформы (канонической, сплайс-опосредованной или содержащей замены отдельных аминокислот), для которой была детектирована нуклеотидная последовательность. Для выявления посттрансляционных модификаций (предсказание наличия которых неосуществимо на основе транскриптомных или геномных данных) необходим протеомный анализ: так, например, изменение физико-химических свойств, обнаруживаемое с помощью 2DE, позволяет предположить ту или иную посттрансляционную модификацию.

Данная работа была нацелена на оценку размера протеома гепатоцеллюлярной клеточной линии HepG2 посредством идентификации протеоформ, возникающих вследствие реализации на белковом уровне альтернативного сплайсинга (АС) и однонуклеотидного полиморфизма (ОАП),

* Мастерный белок – форма белка, аминокислотная последовательность которой позволяет обобщить набор протеоформ, кодируемых одним геном. Необходимость использования этого термина обусловлена особенностью масс-спектрометрического анализа в режиме bottom-up, в ходе которого идентифицируется не весь белок, а только его фрагмент – пептид, который в ряде случаев может с равным успехом быть картирован на несколько белковых продуктов одного гена.

* - адресат для переписки

а также посттрансляционных модификаций (ПТМ) путём интеграции транскриптомных и протеомных данных одного и того же биологического образца.

МАТЕРИАЛЫ И МЕТОДЫ

Анализ данных RNAseq

Ключевая идея транскриптопротеомного подхода к поиску протеоформ состоит в использовании оптимизированной референсной библиотеки, сгенерированной на основе результатов высокопроизводительного секвенирования РНК исследуемых образцов – в нашем случае, клеток линии гепатоцеллюлярной карциномы HepG2. Для составления такой библиотеки были использованы данные высокопроизводительного секвенирования транскриптома клеток HepG2, которые были независимо получены на двух платформах (Illumina Genome HiSeq 2000 и Applied Biosystems SOLiD 4) в трёх технических повторениях (ID экспериментов в NCBI: SRX395473 и SRX390071). Полученные результаты секвенирования были объединены на основании ранее продемонстрированной высокой корреляции экспрессии генов, детектируемой на разных платформах [6].

При анализе транскриптомных данных была использована референсная сборка генома Ensembl GRCh38 (релиз 80), применены программные пакеты и алгоритмы Trimmomatic, Tophat, bowtie2, HTSeq, Cufflinks, GATK, picard-tools, samtools, bcftools, GTFplus и Annovar. Алгоритм, включающий стадии предварительной обработки прочтений, их картирование на геном, поиск и аннотацию полиморфных вариантов, представлен в работе [7].

Протеомное профилирование с помощью двумерного гель-электрофореза с последующей панорамной масс-спектрометрией

Для протеомного анализа использовали тот же образец клеток HepG2, что и для транскриптомного профилирования. Фракционирование белковой смеси проводили с помощью 2DE по стандартному протоколу Клозе и О'Фаррелла [8, 9] с использованием иммобилизованного градиента pH на приборах Ettan TM IPGphor3 и Ettan TM DALTsix ("GE Healthcare", США). Гель, содержащий фракционированную смесь белков HepG2, разрезали на 96 ячеек с координатами по молекулярному весу (MW) и изоэлектрической точке (pI). Каждый фрагмент геля был обработан трипсином и подвергнут хромато-масс-спектрометрическому исследованию в панорамном режиме на приборе ESI LC-MS/MS (LTQ-Orbitrap Q-Exactive) в двух технических повторениях. Далее файлы масс-спектров (192 файла с общим объёмом, превышающим 114 Гб) проанализировали в программе Mascot 2.4.1 ("Matrix Science", Великобритания). В качестве возможных посттрансляционных модификаций в настройках поиска были указаны карбамидометилирование цистеина и окисление метионина, допустимая погрешность (mass tolerance) для масс родительского и дочернего ионов была установлена на уровне 20 ppm и 0,01 Да, соответственно, при интерпретации масс-спектров учитывали возможность двух трипсиновых "недорезов". Минимальный приемлемый уровень показателя достоверности (Mascot score) составил 30, FDR ≤1%. В качестве референсной базы данных, против которой производили биоинформационный поиск, использовали экзом-специфичную библиотеку, созданную на предыдущей стадии работы и расширенную с учётом заведомо ложных (decoy) последовательностей. В результатах поиска игнорировались протеоформы, несущие одноаминокислотные замены, которые неразличимы или трудноразличимы в результатах масс-спектрометрических экспериментов (I/L, Q/E, Q/K, G/N, F/M, N/D, M/T). Помимо оптимизированной библиотеки аминокислотных последовательностей в качестве дополнительных источников сведений о гетерогенности протеома использовали геномные ресурсы dbSNP, The Cancer Genome Atlas, пост-трансляционные базы данных dbPTM, PTMcode, PhosphoSitePlus, протеомные репозитории PRIDE, GPMdb, PeptideAtlas; информационные ресурсы UniProt, OMIM, GeneCards и пр. Детали эксперимента описаны в работе [10].

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

Интерпретация результатов информационно-зависимого масс-спектрометрического анализа состоит в сопоставлении экспериментально полученных спектров исследуемого образца с теоретическими спектрами входящих в референсную библиотеку аминокислотных последовательностей. Несмотря на распространённость такой стратегии, у неё есть ряд существенных ограничений в поиске аберрантных протеоформ. Так, отсутствие специфичного протеотипического пептида, надежно детектируемого масс-спектрометрически, который позволил бы различить между собой несколько кодируемых одним геном белковых продуктов [11], или неполнота референсных баз данных [12] затрудняют исследование протеома.

Протеоформы, кодируемые одним геном, обладают различными аминокислотными последовательностями и различаются между собой по физико-химическим свойствам. Вследствие невысокого уровня покрытия аминокислотной последовательности, которым характеризуется панорамный масс-спектрометрический анализ, зачастую удаётся детектировать только мастерную форму белка. При этом, детекция мастерной формы белка в нескольких пространственно удалённых ячейках геля может быть объяснена наличием нескольких протеоформ одного гена. Анализ величины сдвига по координатной плоскости pI-MW позволяет предположить его причину (конкретизировать тип изменения белковой последовательности или посттрансляционные модификации) и сопоставить её с результатом транскриптомного профилирования.

Таким образом, для точечного поиска и аннотирования протеоформ была разработана схема эксперимента, представленная на рисунке 1.

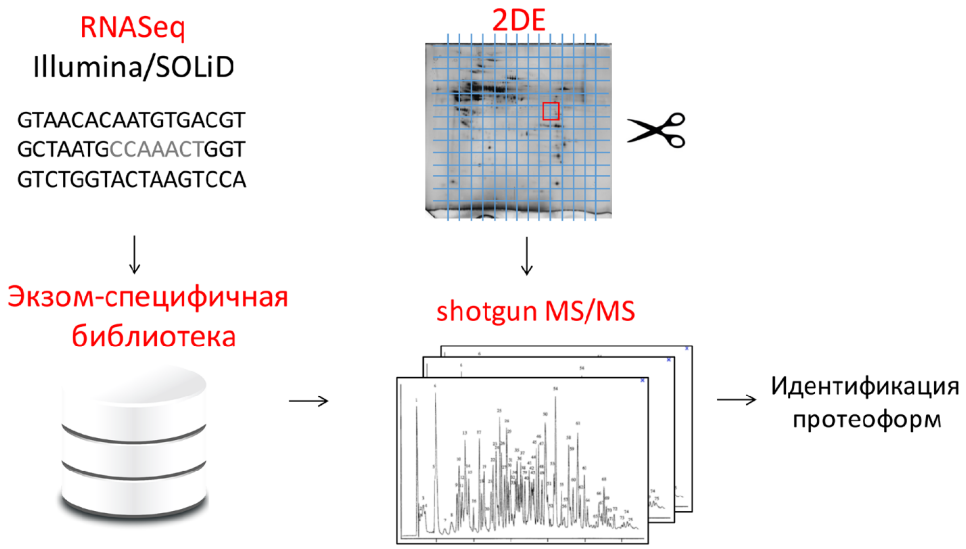

Рисунок 1. Схема поиска и аннотирования представленных в исследуемом образце протеоформ, интегрирующая результаты транскриптомного анализа, двумерного гель-электрофореза и панорамной масс-спектрометрии.

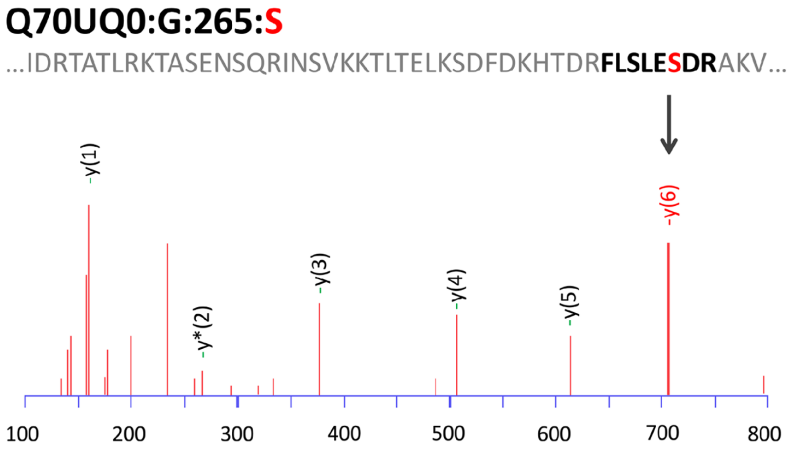

Рисунок 2. Пример идентификации протеоформы с заменой единичной аминокислоты Достоверность идентификации белкового варианта Q70UQ0:G:265:S определяется детекцией характеристичных родительских и соответствующих дочерних фрагментов (y6).

По итогам анализа результатов секвенирования была создана экзом-специфичная библиотека аминокислотных последовательностей, включающая 52 тыс. аминокислотных последовательностей, соответствующих 12 тыс. генов. Суммарно было выявлено 32 тыс. последовательностей с полиморфизмами (ОАП, вставками и делециями), 22 тыс. последовательностей, образованных в результате альтернативного сплайсинга и 11 тыс. канонических сиквенсов. В библиотеку были включены транскрипты, белок-кодирующие гены, которых имеют в ресурсе UniProt статус “Reviewed”, а также, если уровень экспрессии FPKM >0,1.

Протеомное профилирование клеточной линии HepG2 методом 2DE с последующим масс-спектрометрическим анализом является самым эффективным на настоящий момент подходом к исследованию гетерогенности протеома: тандем 2DE-LC-MS/MS позволил детектировать более 30 тыс. протеоформ HepG2, кодируемых 4 тыс. генов, при этом для того же типа биоматериала визуализировать на геле удалось не более 18 тыс. протеоформ [10, 13].

В первую очередь были определены протеоформы, для которых детектировали пептиды, однозначно характеризующие данную протеоформу, то есть сплайс-опосредованные протеотипические пептиды или содержащие замену (рис. 2). Всего, таким образом, было определено 126 протеоформ с аминокислотными заменами и 37 сплайс-опосредованных белковых вариантов.

Детальное сопоставление транскриптомных и протеомных данных в рамках предлагаемого мультиомного алгоритма представляет собой гораздо более длительный процесс и на данном этапе не обладает автоматизацией, однако предоставляет возможность по косвенным признакам (например, по сдвигу пятна белка по координатной плоскости pI-MW на геле) предположить наличие конкретных протеоформ.

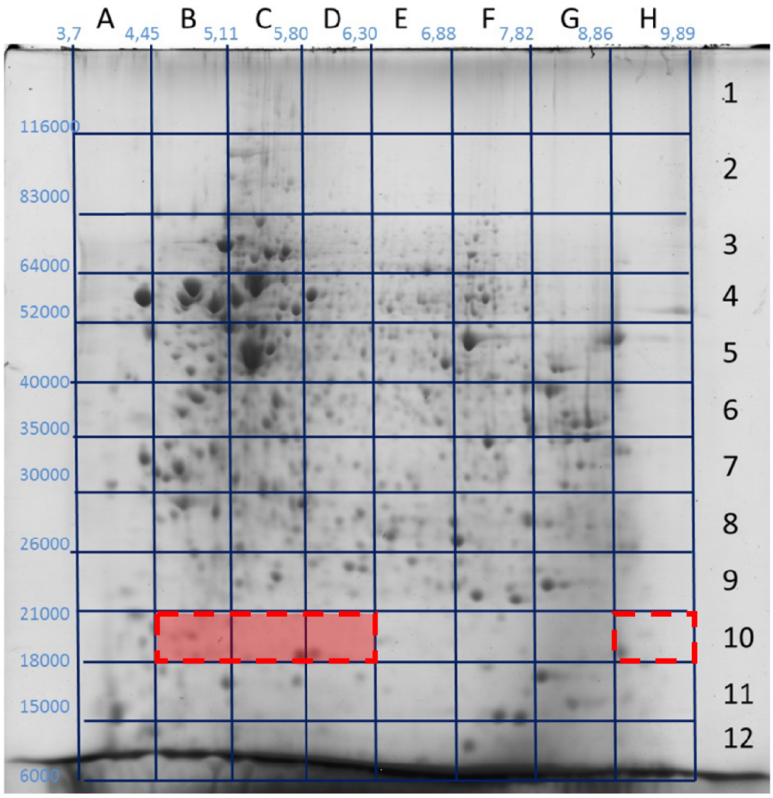

Рисунок 3. Пример исследования посттрансляционных модификаций. Для белка P62280, помимо канонической протеоформы (детектирована в ячейке H10), было обнаружено 6 протеоформ (детектированы в ячейках B10, C10, D10), предположительно несущих модификации аланина (N-ацетилаланин), аргинина (цитруллин) и цистеина (S-пальмитирование).

Большинство генов, транскрипты которых были детектированы в клетках HepG2, характеризуются несколькими сплайс-формами в UniProt KB. Однако, в исследуемом типе биоматериала соответствующие транскрипты были выявлены только для части из них (60%). В случае генов, которые были представлены только одним видом транскриптов, 1302 белков детектировали только в одной ячейке геля, имеющие параметры pI-MW, соответствующие канонической аминокислотной последовательности. При этом 1985 белков были детектированы в нескольких ячейках (рис. 3), и, согласно характеру сдвига пятен, а также данных об известных модификациях для части из них, можно заключить об идентификации белков с ПТМ. Выявление протеоформ с ПТМ производилось следующим образом. Если мастерная форма белка детектировалась в нескольких удалённых друг от друга ячейках геля, и при этом ей соответствовал только один транскрипт, мы предполагали, что дополнительные пятна на геле вызваны посттрансляционными модификациями. По данным статьи [14] мы выбрали 10 наиболее часто встречающихся посттрансляционных модификаций (Phosphoserine, Phosphotreonine, N-linked glycosylation, N6-acetyllysine, Glycyl lysine isopeptide, Phosphotyrosine, O-linked glycosylation, N-acetylalanine, 4-hydroxyproline, Pyrrolidone carboxylic acid) и каждой сопоставили сдвиг на координатной плоскости pI-MW, который может вызвать такая модификация. В своей работе мы также учитывали возможность множественных модификаций одного вида или суперпозиции нескольких ПТМ разных типов.

Самый сложный этап анализа – идентификация протеоформ при детекции соответствующих протеотипических пептидов в соседних ячейках. Безусловно, получаемые сдвиги пятен могут быть обусловлены как независимым ОАП, альтернативным сплайсингом и ПТМ, так и комбинацией этих событий, которые могут совпадать по своим характеристикам в координатах pI-MW. Другим немаловажным аспектом является то, что детекция белка в соседних ячейках, в частности горизонтальных, может быть следствием деградации белка в процессе электрофореза. Такой детальный анализ был проведён для белков, кодируемых генами хромосомы 18 человека, где удалось описать 36 протеоформ (для 15 генов) из 116 детектированных согласно MC-данным (таблица).

Для оценки потенциального многообразия протеоформ, образующихся в результате реализации на белковом уровне альтернативного сплайсинга, полиморфизма единичных нуклеотидов и различных посттрансляционных модификаций в исследуемых образцах клеточной линии HepG2, были использованы три расчётные модели, которые учитывают сценарии совместного и независимого возникновения аберраций. Согласно данным расчётных моделей и размеру экзома человека (около 20 тыс. белок кодирующих генов), протеом человека может содержать от 600 тыс.

Таблица. Результат протеомного профилирования клеточной линии HepG2 методом 2DE с последующим масс-спектрометрическим анализом с использованием персонализированной базы данных при идентификации протеоформ, кодируемых генами хромосомы 18 человека

Характер протеоформы	БКГ	Суммарное количество
Всего	32	116
Представлены только канонической последовательностью	9	9
Наличие ОАП в канонической последовательности	3	3
Наличие микроделеции в канонической последовательности	2	2
Наличие ПТМ в канонической последовательности	5	11
Представлены альтернативной последовательностью (сплайс-форма)	3	5
Наличие ОАП в альтернативной последовательности (сплайс-форме)	1	1
Наличие микроделеции в альтернативной последовательности (сплайс-форме)	2	2
Наличие ПТМ в альтернативной последовательности (сплайс-форме)	3	3
Невыясненный характер протеоформ	16	80

до 6 млн протеоформ [2]. Использование в качестве расчётных параметров собственных результатов транскриптомного профилирования позволяет предположить наличие от 135 до 500 тыс. протеоформ в составе протеома клеток HepG2.

ЗАКЛЮЧЕНИЕ

В рамках данного проекта мультиомный алгоритм позволил идентифицировать 2399 канонических протеоформ, 37 сплайс-опосредованных вариантов белков, 126 протеоформ с единичными аминокислотными заменами (см. Доп. табл. 1) и 734 протеоформы с посттрансляционными модификациями (на основе анализа позиции пятна протеоформы на геле).

Успех дальнейших исследований гетерогенного протеома требует совместных усилий биоинформатиков и биохимиков по консолидации накопленных транскриптомных и протеомных данных на информационном и интерпретационном уровнях. Полученные знания о белковом разнообразии отдельных органов, тканей, организмов и целых популяций позволят сформировать более полную картину функционирования живой клетки и ассоциировать те или иные аберрации с возникновением или развитием патологических процессов.

БЛАГОДАРНОСТИ

Работа выполнена при финансовой поддержке Гранта РНФ № 15-15-30041. Экспериментальная часть работ выполнена с использованием оборудования ЦКП “Протеом человека”, поддержанного Минобрнауки России в рамках выполнения соглашения №14.621.21.0017 (уникальный идентификатор проекта RFMEFI62117X0017).

ЛИТЕРАТУРА

1. *Gregory S.G., Barlow K.F., McLay K.E., Kaul R., Swarbreck D., Dunham A., Scott C.E., Howe K.L., Woodfine K., Spencer C.C et al.* (2006) Nature, **441**, 315-321.
2. *Ponomarenko E.A., Poverennaya E.V., Ilgisonis E.V., Pyatnitskiy M.A., Kopylov A.T., Zgoda V.G., Lisitsa A.V., Archakov A.I.* (2016) Int. J. Anal. Chem., **2016**, 7436849. DOI: 10.1155/2016/7436849
3. *Poverennaya E.V., Kopylov A.T., Ponomarenko E.A., Ilgisonis E.V., Zgoda V.G., Tikhonova O.V., Novikova S.E., Farafonova T.E., Kiseleva Y.Y., Radko S.P. et al.* (2016) J. Proteome Res., **15**, 4030-4038.
4. *Archakov A., Aseev A., Bykov V., Grigoriev A., Govorun V., Ivanov V., Khlunov A., Lisitsa A., Mazurenko S., Makarov A., Ponomarenko E., Sagdeev R., Skryabin K.* (2011) Proteomics, **11**, 1853-1856. DOI: 10.1002/pmic.201000540
5. *Liu Y., Beyer A., Aebersold R.* (2016) Cell, **165**, 535-550.
6. *Ponomarenko E.A., Kopylov A.T., Lisitsa A.V., Radko S.P., Kiseleva Y.Y., Kurbatov L.K., Ptitsyn K.G., Tikhonova O.V., Moisa A.A., Novikova S.E. et al.* (2014) J. Proteome Res., **13**, 183-190.
7. *Krasnov G.S., Dmitriev A.A., Kudryavtseva A.V., Shargunov A.V., Karpov D.S., Uroshlev L.A., Melnikova N.V., Blinov V.M., Poverennaya E.V., Archakov A.I., Lisitsa A.V., Ponomarenko E.A.* (2015) J. Proteome Res., **14**, 3729-3737.
8. *Klose J.* (1975) Humangenetik, **26**, 231-243.
9. *O'Farrell P.H.* (1975) J. Biol. Chem., **250**, 4007-4021.
10. *Naryzhny S.N., Maynskova M.A., Zgoda V.G., Ronzhina N.L., Kleyst O.A., Vakhrushev I.V., Archakov A.I.* (2016) J. Proteome Res., **15**, 525-530.
11. *Veenstra T.D.* (2011) Exp. Rev. Proteomics, **8**, 681-683.
12. *Stevens S.M.Jr., Prokai-Tatrai K., Prokai L.* (2008) Mol. Cell Proteomics, **7**, 2442-2451.
13. *Naryzhny S.* (2016) Proteomes, **4**, e33. DOI: 10.3390/proteomes4040033
14. *Prabakaran S., Lippens G., Steen H., Gunawardena J.* (2012) Wiley Interdiscip. Rev. Syst. Biol. Med., **4**, 565-583.

Поступила: 31. 08. 2017.
Принята к печати: 29. 09. 2017.

MULTIOMICS STUDY OF HepG2 CELL LINE PROTEOME

E.V. Poverennaya, O.I. Kiseleva, E.A. Ponomarenko, S.N. Naryzhny, V.G. Zgoda, A.V. Lisitsa

Institute of Biomedical Chemistry,
10 Pogodinskaya str., Moscow, 119121 Russia; e-mail: olly.kiseleva@gmail.com

Current proteomic studies are generally focused on the most abundant proteoforms encoded by canonical nucleic sequences. Transcriptomic and proteomic data, accumulated in a variety of postgenome sources and coupled with state-of-art analytical technologies, allow to start the identification of aberrant (non-canonical) proteoforms. The main sources of aberrant proteoforms are alternative splicing, single nucleotide polymorphism, and post-translational modifications. The aim of this work was to estimate the heterogeneity of HepG2 proteome. We suggested multiomics approach, which combines transcriptomic (RNAseq) and proteomic (2DE-MS/MS) methods, as a promising strategy to explore the proteome.

Key words: proteome, transcriptome, transcriptoproteome, proteoform, alternative splicing, single amino acid polymorphism, post-translational modification