

Киселева Ольга Игоревна

**ТРАНСКРИПТОМНО-ПРОТЕОМНЫЙ ПОДХОД
ДЛЯ АНАЛИЗА ПРОТЕОФОРМ КЛЕТОЧНОЙ ЛИНИИ HEPG2**

03.01.04 – биохимия

АВТОРЕФЕРАТ

диссертации на соискание ученой степени
кандидата биологических наук

Москва, 2018

Работа выполнена в Федеральном государственном бюджетном научном учреждении «Научно-исследовательский институт биомедицинской химии имени В.Н. Ореховича».

- Научный руководитель:** кандидат биологических наук
Поверенная Екатерина Владимировна
- Научный консультант:** доктор биологических наук, академик РАН
Лисица Андрей Валерьевич
- Официальные оппоненты:** **Шишкин Сергей Сергеевич**
доктор биологических наук, профессор
Федеральный исследовательский центр
«Фундаментальные основы биотехнологии» Российской
академии наук (Институт биохимии им. А.Н. Баха РАН),
лаборатория биомедицинских исследований,
заведующий лабораторией
- Шевченко Валерий Евгеньевич**
доктор биологических наук, профессор, ФГБУ
«Национальный медицинский исследовательский центр
онкологии им. Н.Н. Блохина» Минздрава России,
Научно-исследовательский институт канцерогенеза,
лаборатория онкопротеомики,
руководитель лаборатории
- Ведущая организация:** ФГБУН Институт молекулярной биологии имени
В.А. Энгельгардта РАН

Защита состоится «27» сентября 2018 года в 11 часов на заседании Диссертационного совета Д 001.010.01 при Федеральном государственном бюджетном научном учреждении «Научно-исследовательский институт биомедицинской химии имени В.Н. Ореховича» (ИБМХ) по адресу: 119121, г. Москва, Погодинская ул., д. 10, стр. 8.

С диссертацией можно ознакомиться в библиотеке Федерального государственного бюджетного научного учреждения «Научно-исследовательский институт биомедицинской химии имени В.Н. Ореховича» и на сайте www.ibmc.msk.ru.

Автореферат разослан «___» _____ 2018 г.

Ученый секретарь
Диссертационного совета,
кандидат химических наук



Е.А. Карпова

1. ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

1.1. Актуальность проблемы, цель и задачи

Сложность живых систем при сравнительно малом количестве белок-кодирующих генов (БКГ) достигается за счет регуляции транскрипции, трансляции и посттрансляционных модификаций, обеспечивающих многообразие белковых форм (протеоформ). Доказательством этого являются результаты проекта «Геном человека», согласно которым у человека было выявлено около 20 тысяч БКГ, что ниже первоначальных оценок приблизительно в 50 тысяч раз (Smith & Kelleher, 2013).

На транскриптомном уровне, который является первой стадией реализации генетической информации, количество вариантов транскриптов, считанных с одного гена, обеспечивает биосинтез 12-30 белковых продуктов, а на протеомном уровне их число достигает уже 20-80 (Ponomarenko et al., 2017). Формы белков (протеоформы), кодируемые одним геном, отличаются по своей аминокислотной последовательности, свойствам, активности и содержанию в различных органах и тканях от канонической формы белка, которая соответствует обобщенной сборке генома человека. Основными источниками гетерогенности являются одноаминокислотные полиморфизмы (ОАП), возникающие вследствие несинонимичных однонуклеотидных замен в ДНК, альтернативный сплайсинг пре-мРНК (АС) и посттрансляционные модификации белков (ПТМ).

Протеомные подходы позволяют идентифицировать все типы протеоформ. Чувствительность современных методов масс-спектрометрии, нацеленных на детекцию конкретной протеоформы, достигает 10^{-16} М, что эквивалентно содержанию 600 молекул в 1 мкл образца. Значителен прогресс методов панорамной масс-спектрометрии, направленной на идентификацию сложной белковой смеси и позволяющей детектировать более 10 000 различных белков в образце (Doll et al., 2017). Однако, несмотря на развитие технологий, детекция всех возможных протеоформ, транслируемых в исследуемом образце, к настоящему моменту ограничена.

Задача прочтения протеома сложнее, чем расшифровка генома: число протеоформ может достигать 2 млрд (Kelleher, 2012; Ponomarenko et al., 2016), а исследователь имеет дело с «алфавитом» из 21 аминокислоты против четырехнуклеотидного кода ДНК. Несмотря на значительный прогресс в детекции канонических белков – обычно это наиболее длинные или часто детектируемые последовательности белка – использовать результаты протеомного профилирования в практике пока затруднительно. Одна из основных проблем связана с недостатком данных о референсных характеристиках протеоформ: неизвестно, какие канонические и аберрантные протеоформы транслируются в различных органах и тканях и каково их количественное содержание.

Развитие патологических процессов в организме характеризуется изменениями протеома, которые заключаются в появлении в клетке ассоциированных с болезнью протеоформ. Исследование протеоформ –

протеоформика – становится основой мониторинга развития заболеваний, в особенности онкологических, для которых определена наибольшая гетерогенность протеома (Nedelkov, 2008; Lisitsa et al., 2014). В опухолях количество мутантных вариантов белка в расчете на один ген может в десятки раз превышать их количество в нормальной ткани (Martincorena & Campbell, 2015).

Необходимость исследования альтернативного сплайсинга, полиморфизма отдельных аминокислот и посттрансляционных модификаций обусловлена влиянием данных процессов на уровень экспрессии и функциональные свойства белков. Отсутствие сведений о белковом составе тканей и органов человека в норме существенно осложняет определение диагностической значимости той или иной протеоформы.

Актуальность темы исследования обозначена потребностью в разработке подходов анализа полного состава протеома органов и тканей, включающих выявление протеоформ, образованных вследствие альтернативного сплайсинга, одноаминокислотных полиморфизмов или посттрансляционных модификаций.

Целью данной работы являлось выявление совокупности протеоформ, образующихся в результате несинонимичных однонуклеотидных замен, альтернативного сплайсинга и посттрансляционных модификаций, на основе сопоставления экспериментальных данных транскриптомного и протеомного профилирования клеточной линии HepG2 как модельной системы.

Для достижения указанной цели были поставлены следующие **задачи**:

1. Сформировать перечень транскрибируемых протеоформ на основе анализа результатов высокопроизводительного секвенирования мРНК в образце клеточной линии HepG2;
2. Оптимизировать параметры биоинформатического анализа файлов масс-спектров пептидных фрагментов для идентификации протеоформ линии HepG2, образующихся в результате альтернативного сплайсинга, однонуклеотидного полиморфизма, а также протеоформ, содержащих посттрансляционные модификации;
3. Сопоставить данные о транскрибируемых протеоформах с результатами протеомного исследования, проведенного с использованием фракционирования в двумерном гель-электрофорезе и последующей масс-спектрометрии;
4. Определить состав протеома клеточной линии HepG2 путем сопоставления результатов масс-спектрометрического анализа и положения протеоформы на двумерной электрофореграмме.

1.2. Научная новизна и практическая значимость работы

В данной работе показаны возможности протеоформики – нового раздела протеомики, базирующегося на исследовании протеома с использованием транскриптомного и протеомного анализа биологических образцов и направленного на выявление протеоформ, т.е. различных белковых форм, кодируемых одним геном. Термин «протеоформика» был предложен в работе

российских авторов в 2014 году (Lisitsa et al., 2014) на основе теоретических предпосылок. В настоящей работе впервые дано экспериментальное обоснование существования множественности протеоформ и предложен методический подход к выявлению генетически детерминированного полиморфизма единичных аминокислот, сплайс-опосредованных вариантов, а также посттрансляционных модификаций, влияющих на физико-химические свойства белковых молекул. В результатах масс-спектрометрического анализа, следующего за двумерным разделением белковой смеси в геле, идентифицированы протеоформы, образованные вследствие АС, ОАП или ПТМ. Получена карта протеоформ, характеризующая трансляцию более 20% генов в геноме человека.

С практической стороны, полученные результаты позволяют усовершенствовать экспериментальные подходы к протеомному анализу клеток и тканей. Сопоставление профиля протеоформ в норме и при патологии является перспективным источником нового поколения биомаркеров заболеваний.

1.3. Методология и методы исследования

В работе использованы результаты транскриптомно-протеомного профилирования, выполненного современными экспериментальными методами (Zgoda et al., 2013, Naryzhny et al., 2016). Для обработки данных применены современные биоинформатические подходы и соблюдены требования, сформулированные международным сообществом «Протеом человека» (Deutsch et al., 2016), а также учтены рекомендации ведущих научных школ в области протеомики (Nesvizhskii, 2014; Aebersold, 2003).

1.4. Участие соискателя в получении результатов диссертации

Соискатель проанализировал научную литературу по теме диссертации, создал схему алгоритма интерпретации масс-спектрометрических данных и провел аннотацию идентифицированных протеоформ в соответствии с их физико-химическими свойствами. Представленные в работе результаты были получены либо лично автором, либо при непосредственном участии автора на всех этапах исследования, включая планирование, проведение экспериментов, интерпретацию и аннотацию полученных данных, подготовку публикаций.

1.5. Основные положения, выносимые на защиту

1. Результаты транскриптомно-протеомного профилирования, выполненного в высокопроизводительном режиме для одного и того же образца, позволяют описать уникальный состав его протеома.
2. Сопоставление результатов масс-спектрометрической идентификации белков с экспериментальными и расчетными данными о физико-химических характеристиках белковых молекул (параметров молекулярной массы и изоэлектрической точки) позволяет дифференцировать протеоформы, образующиеся в результате биологически обусловленного изменения

первичной структуры белка, посттрансляционных модификаций, а также артефактов пробоподготовки.

3. Протеоформы могут возникнуть как в результате независимых, так и совместных событий альтернативного сплайсинга, одноаминокислотных замен и посттрансляционных модификаций, усиливающих гетерогенность протеома.

1.6. Степень достоверности и апробация результатов

Достоверность полученных результатов подтверждена согласованностью результатов современных экспериментальных и вычислительных методов, использованных в работе. Обсуждение результатов проведено с учетом современных данных биологической науки. Научные положения и выводы, изложенные в диссертации, обоснованы и подтверждены статистическими выкладками.

Основные положения диссертационной работы были представлены в виде постерных докладов на конгрессе «Протеомный форум» (Proteomic Forum 2015, Берлин), конгрессе Европейской Протеомной Ассоциации (EuPA 2016, Стамбул), конгрессе международной организации «Протеом человека» (HUPO 2017, Дублин) и конференции «Клиническая протеомика. Постгеномная медицина» (ClinProt 2017, Москва). Устные сообщения представлялись на конференции «Ломоносов» (Москва, 2015), конгрессе «Человек и лекарство» (Москва, 2016) и на конференции «Высокопроизводительное секвенирование в геномике» (Новосибирск, 2017).

1.7. Публикации

По теме диссертационной работы опубликовано 17 работ, из которых 7 статей в международных рецензируемых изданиях и 10 публикаций в трудах конференций.

1.8. Объем и структура диссертации

Диссертация изложена на 125 страницах машинописного текста, содержит 5 таблиц и 23 рисунка. Состоит из следующих глав: Список Сокращений, Введение, Обзор Литературы, Материалы и Методы, Результаты и Обсуждение, Заключение, Выводы, Список литературы, а также включает Благодарности и два Приложения.

2. МАТЕРИАЛЫ И МЕТОДЫ ИССЛЕДОВАНИЯ

2.1. Источники и обработка транскриптомных данных

Поиск протеоформ осуществляли с использованием референсной библиотеки, сгенерированной на основе результатов высокопроизводительного секвенирования мРНК клеток опухолевой линии HepG2.

Для составления библиотеки использовали данные высокопроизводительного секвенирования транскриптома клеток линии HepG2, полученные независимо на

платформах Illumina Genome HiSeq 2000 и Applied Biosystems SOLiD 4 (ID экспериментов в NCBI: SRX395473 и SRX390071). Результаты секвенирования объединяли на основании ранее продемонстрированной высокой корреляции уровней экспрессии генов, детектируемых на разных платформах (Ponomarenko et al., 2014). Анализировали транскрипты генов со статусом «Reviewed» в базе UniProt, из них отбирали транскрипты с уровнем значения программно-рассчитываемого показателя экспрессии (англ. fragments per kilobase of exon per million reads mapped, FPKM) $> 0,1$. Транскриптомные данные обрабатывали с помощью программы PPLine (Krasnov et al., 2015).

2.2. Источники и обработка протеомных данных

Протеомному анализу подвергли около 10 млн клеток линии HepG2, образец которой был ранее использован для проведения транскриптомного анализа. Фракционирование белковой смеси было проведено с помощью двумерного гелелектрофореза по стандартному протоколу (Klose, 1975; O'Farrell, 1975) под руководством д.б.н. С.Н. Нарыжного. Гель, содержащий фракционированную смесь белков HepG2, разрезали на 96 ячеек. Каждый фрагмент геля ($\sim 0,7 \text{ см}^2$) измельчили, обработали трипсином и исследовали хромато-масс-спектрометрически в панорамном режиме на приборе LTQ-Orbitrap Q-Exactive (Thermo) в двух технических повторениях (Naryzhny et al., 2014; Naryzhny et al., 2016). Отдельно были проанализированы результаты панорамного масс-спектрометрического профилирования белков линии HepG2 без предварительного фракционирования с помощью двумерного гелелектрофореза. Методические особенности проведения белкового фракционирования и масс-спектрометрического эксперимента изложены в работах (Zgoda et al., 2013; Poverennaya et al., 2016). Файлы масс-спектров анализировали с помощью платформы SearchGUI (v. 3.2.20, Vaudel et al., 2015). Карбамидометилирование цистеина указывали в качестве фиксированной химической модификации, окисление метионина – в качестве варибельной. При интерпретации масс-спектров учитывали возможность одного пропущенного сайта трипсинолиза. В качестве референсной базы данных использовали транскриптом-специфичную библиотеку, расширенную с учетом заведомо ложных последовательностей и белков-контаминантов из базы cRAP (115 белков, Mellacheruvu et al., 2013). Для дальнейшего анализа использованы результаты идентификации, удовлетворяющие уровню показателя достоверности (англ. false discovery rate) $\text{FDR} < 1\%$.

2.3. Оптимизация параметров анализа масс-спектров

Для интерпретации результатов масс-спектрометрического эксперимента использовали платформу SearchGUI, которая включает поисковые алгоритмы X!Tandem, MS-GF+, MS Amanda, MyriMatch, Comet, Tide, Andromeda и OMSSA. Входящий в состав платформы интегратор PeptideShaker применен для

комбинирования результатов интерпретации файлов масс-спектров ячеек геля, полученных с помощью нескольких поисковых алгоритмов. Для выбора оптимального сочетания поисковых алгоритмов исследовали влияние количества используемых поисковых машин и объема референсной базы данных на производительность анализа. Эффективность сочетания поисковых алгоритмов оценивали путем сопоставления результатов масс-спектрометрического анализа стандартной смеси UPS2 (Merck) 48 белков (500 амоль – 50 пмоль).

2.4. Анализ результатов масс-спектрометрического исследования протеома клеточной линии HepG2 и аннотирование протеоформ

По результатам масс-спектрометрического анализа фракционированных с помощью 2DE белков линии HepG2 в программе PeptideShaker сгенерировали 192 отчета. Каждый из отчетов проанализировали согласно разработанному алгоритму, реализованному с помощью интерпретатора Python (v. 2.7.7). Принципиально алгоритм может быть разделен на три стадии: 1) предварительная фильтрация полученных сведений, включающая удаление из поля рассмотрения нетриптических пептидов и проверку пептидов на уникальность относительно всех белков человека в UniProt; 2) картирование пептидов на соответствующие протеоформы; 3) определение типа детектированной протеоформы и достоверности идентификации.

Выделяли случаи однозначной идентификации канонической или aberrантной протеоформы. Неоднозначной идентификацией считали обнаружение белковой группы, то есть группы из нескольких протеоформ, кодируемых одним или несколькими генами, для которых был детектирован одинаковый набор пептидов. Если протеоформе соответствовало не менее двух пептидов, хотя бы один из которых – протеотипический, т.е. характеризующий единственный белковый продукт, то идентификация считалась однозначной, все прочие случаи классифицировались как неоднозначные.

2.5. Web-ресурсы

В качестве дополнительных источников сведений о гетерогенности протеома использовали геномные, транскриптомные и протеомные ресурсы и вычислительные инструменты (табл. 1).

Таблица 1. Список информационных ресурсов и инструментов, использованных для исследования протеоформ

№	Название	Сведения в ресурсе
1	UniProtKB, 2017_04 www.uniprot.org	Сведения о протеоформах (СО, ОАП, ПТМ), уровне доказательства их существования и связи с патологией
2	PeptideAtlas, 2017_01 www.peptideatlas.org	Результаты более 1000 протеомных экспериментов, выполненных в режиме направленной масс-спектрометрии
3	PRIDE, 2017_12 www.ebi.ac.uk/pride/archive/	Результаты панорамного анализа белков и пептидов

4	GPMdb, 2017_10 www.gpmdb.thegpm.org	Сведения о пептидах и их модификациях, детектированных методом панорамной масс-спектрометрии
5	The Human Protein Atlas, v.18 www.proteinatlas.org	Информация об экспрессии генов и транскриптов, сведения о соответствующих белковых продуктах, детектированных в клеточных линиях, нормальных тканях и опухолевых образцах
6	1000 Genomes, 2015_09 www.internationalgenome.org	Сведения о генетических вариациях (ОНП, вставках, делециях), детектированных не реже, чем у 1% индивидуумов
7	dbSNP, v.151 www.ncbi.nlm.nih.gov/SNP/	Сведения об обнаруженных на уровне генома ОНП
8	iPTMnet, v.4.1. http://research.bioinformatics.udel.edu/iptmnet/	Информация о предсказанных и экспериментально валидированных ПТМ
9	COSMIC (в т.ч. Cell lines project), v.83 www.cancer.sanger.ac.uk/cosmic	Сведения о соматических мутациях, возникающих на генетическом уровне при онкологических заболеваниях
10	TumorPortal, 2017 www.tumorportal.org	Сведения о точечных мутациях генов в опухолевых и нормальных тканях
11	ClinVar, 2017_05 www.ncbi.nlm.nih.gov/clinvar	Данные об ассоциации генетической вариации и конкретной патологии
12	IPC Isoelectric Point Calculator, 2016_10 www.isoelectric.ovh.org	Теоретические значения молекулярной массы и изоэлектрической точки полипептидов
13	Gorilla http://cbl-gorilla.cs.technion.ac.il	Отранжированный список генов/белков в терминах Gene Ontology (функция, участие в биологическом процессе и локализация)

3. ОСНОВНЫЕ РЕЗУЛЬТАТЫ

3.1. Транскрипция в клеточной линии НерG2

По результатам обработки транскриптомных данных для клеточной линии НерG2 была сформирована библиотека аминокислотных последовательностей, включающая 51 809 последовательностей, кодируемых 12 148 генами. В библиотеке 11 347 записей относятся к каноническим формам, 21 925 последовательностей представлено сплайс-опосредованными (СО) формами; 31 972 последовательности содержат ОАП, вставки и делеции как для канонических белковых продуктов, так и сплайс-опосредованных форм. Таблица 2 отражает сравнение собственной библиотеки с библиотекой из ресурса UniProt.

Согласно данным таблицы 2, количества белковых последовательностей в созданной нами библиотеке белков НерG2 и стандартной библиотеки UniProt различаются незначительно. Собственная библиотека описывает результат экспрессии 12 тыс. генов, а стандартная библиотека охватывает весь геном, но не учитывает полиморфизмы отдельных аминокислот. Переход от популяционных данных к данным о транскриптоме конкретного организма, органа или ткани

позволяет исключить из рассмотрения малоэкспрессирующиеся гены, но при этом спрогнозировать индивидуальные протеоформы.

Таблица 2. Сопоставление транскриптом-специфичной библиотеки со стандартной библиотекой последовательностей (UniProt, v. 2017_04)

Библиотека	Кол-во а.к. Послед.	Кол-во генов	Частота встречаемости сплайс-вариантов	Частота встречаемо- сти ОАП (ОНП)
Транскрибируемые протеоформы HepG2	51 809*	12 148	1,5	2,6
Протеоформы человека (UniProt)	51 885**	20 197	2,6	2,1***

* – библиотека всех возможных аминокислотных последовательностей, в т.ч. с учетом делеций и вставок;

** – стандартная библиотека, состоящая из канонических и сплайс-опосредованных протеоформ человека;

*** – популяционные несинонимичные однонуклеотидные полиморфизмы.

Для выявления случаев количественного превалирования аберрантных транскриптов над каноническими для каждого варианта было рассчитано количество прочтений. В качестве условной оценки экспрессии транскрипта использовано значение параметра FPKM. Анализ экспрессии экзонов продемонстрировал, что в 25% БКГ линии HepG2 превалирует сплайс-транскрипт.

3.2. Анализ файлов масс-спектров для идентификации протеоформ

Платформа SearchGUI обеспечивает переход от необработанных файлов масс-спектров к наглядной интерпретации результатов протеомного эксперимента. SearchGUI интегрирует восемь популярных алгоритмов (MSAmanda, Andromeda, X!Tandem, MS-GF+, MyriMatch, Comet, Tide и OMSSA) и консолидирует результаты интерпретации масс-спектров несколькими поисковыми машинами, идентифицирующих пептидные последовательности по совпадению экспериментальных пиков молекулярных ионов с теоретическими пиками из референсной библиотеки.

Для определения оптимального сочетания алгоритмов на примере файла среднего размера одной из ячеек 2DE-геля проанализировали влияние параметров поиска на производительность обработки результатов масс-спектрометрического эксперимента. Поиск осуществляли по библиотеке канонических и сплайс-опосредованных последовательностей белков человека UniProt. Алгоритмы MSAmanda и Andromeda продемонстрировали нестабильный результат и были

исключены из дальнейшего анализа. Среди поисковых машин самыми производительными являются алгоритмы X!Tandem, OMSSA и MyriMatch. По количеству детектированных пептидов лучший результат показал алгоритм OMSSA.

Следующим шагом стал анализ результатов сочетанного применения поисковых машин. Тестирование показало, что увеличение комбинации сопровождается экспоненциальным ростом времени компьютерной обработки масс-спектров (рис. 1а). При этом, значительного роста количества идентификаций не происходит (рис. 1б).

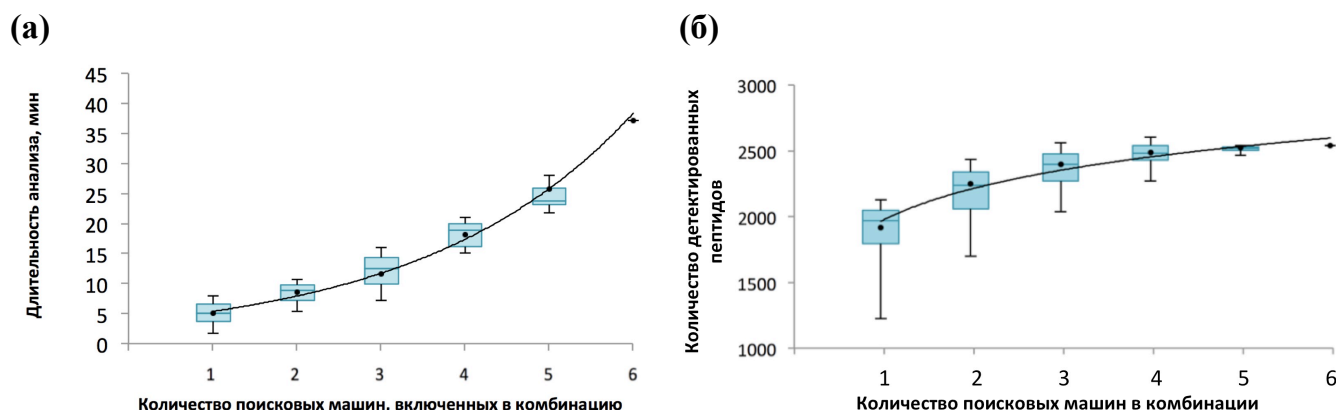


Рисунок 1. Изменение времени обработки масс-спектрометрических данных (а) и количества детектированных пептидов при расширении арсенала используемых поисковых машин (б).

Оптимальной для интерпретации результатов протеомного профилирования было признано сочетание, включающее три поисковых машины. Прирост количества пептидов, обнаруженных при использовании пяти алгоритмов, не оправдывает временные затраты по сравнению с применением трех алгоритмов.

Для выбора трех поисковых машин были сопоставлены различные сочетания пяти алгоритмов. Было определено, что вариант объединения алгоритмов X!Tandem + MS-GF + OMSSA является оптимальным для потоковой обработки масс-спектров (рис. 2).

Дополнительной проверкой эффективности выбранного сочетания алгоритмов послужило сопоставление результатов масс-спектрометрического анализа тестового набора UPS2, состоящего из 48 белков. Интерпретацию результатов масс-спектрометрического эксперимента, выполненного поисковым инструментом Mascot, сопоставили с выбранной комбинацией алгоритмов X!Tandem, OMSSA и MS-GF+, а также с каждым алгоритмом по отдельности. Использование трех отобранных нами алгоритмов позволило обнаружить 24 белка, что на один белок набора UPS2 больше, чем было идентифицировано с помощью Mascot.

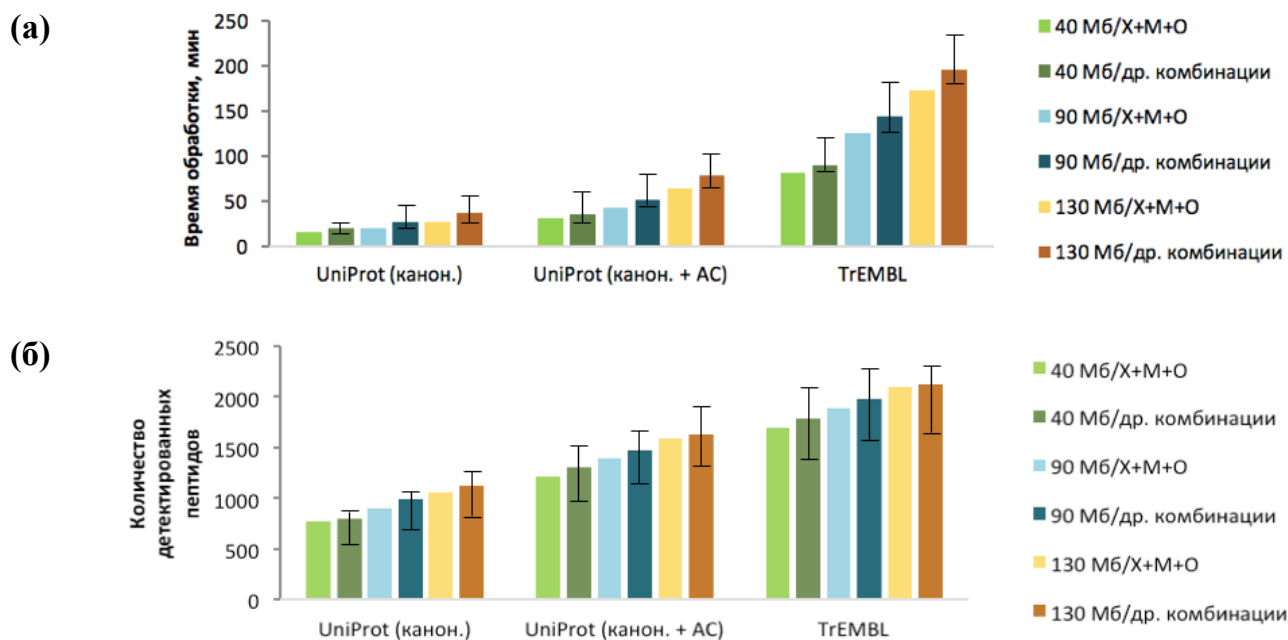


Рисунок 2. Зависимость времени интерпретации (а) и количества детектированных пептидов (б) от размеров библиотеки и файла масс-спектров (рассматривали файлы масс-спектров 40 Мб, 90 Мб и 130 Мб). В паре столбцов гистограммы светлый столбец характеризует комбинацию X!Tandem + MS-GF + OMSSA (X + M + O), темный относится к усредненному по остальным комбинациям времени анализа и количеству детектированных пептидов.

3.3. Протеом клеточной линии HepG2

Для аннотации протеоформ линии HepG2 был разработан алгоритм, позволяющий в автоматическом режиме обрабатывать файлы отчетов поисковых машин. Для иллюстрации логики работы алгоритма приведен анализ одной из наиболее насыщенных по количеству детектированных пептидов ячеек геля – E5 (рис. 3).

Суммарно в двух технических повторениях масс-спектрометрического исследования ячейки E5 было детектировано 4952 пептида, пересечение идентификаций между повторениями составило 74%. Полученный массив данных был отфильтрован от пептидов, длина которых выходила за рамки интервала 9-25 аминокислотных остатков. Из дальнейшего рассмотрения были исключены нетриптические пептиды, а также пептиды, картирующиеся на белки-контаминанты. После технической фильтрации массив содержал 3 170 пептидов, 560 из которых были протеотипическими, а остальные картировались на 514 белковых групп, включающих протеоформы одного или нескольких генов.

Всего в ячейке E5 по уникальным пептидам было детектировано 117 протеоформ, относящихся к 115 генам, в том числе восемь сплайс-вариантов, три последовательности с заменами отдельных аминокислот и один случай, когда эти события встретились одновременно.

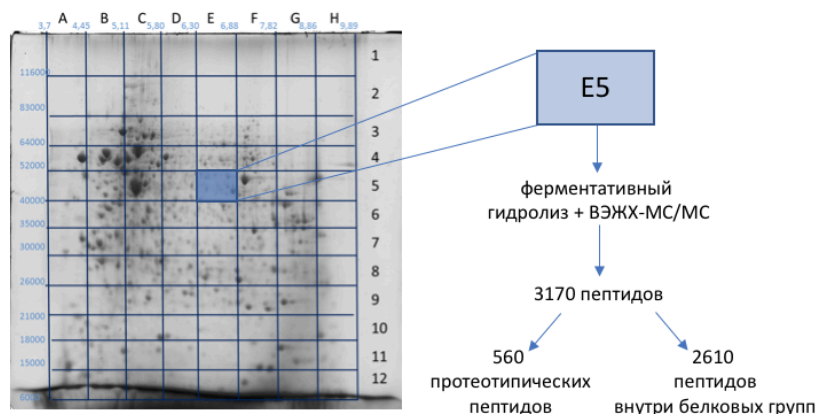


Рисунок 3. Двумерная электрофореграмма протеома клеточной линии HepG2: по оси ординат – значения молекулярной массы MW, по оси абсцисс – значения изоэлектрической точки pI (Naryzhny et al., 2016).

Детекция сплайс-опосредованной (СО) протеоформы иллюстрируется на примере варианта глутаминазы O94925-3, детектированной по двум пептидам (рис. 4). Предположительно, эта СО-форма эффективнее канонического варианта справляется с поддержанием повышенных метаболических потребностей опухоли. Дисбаланс между канонической и СО-формами может рассматриваться в качестве маркера опухолевого процесса (Cassago et al., 2012). Каноническая форма O94925 была также детектирована, но в ячейке E3, согласующейся с расчетными параметрами MW и pI для данной протеоформы.

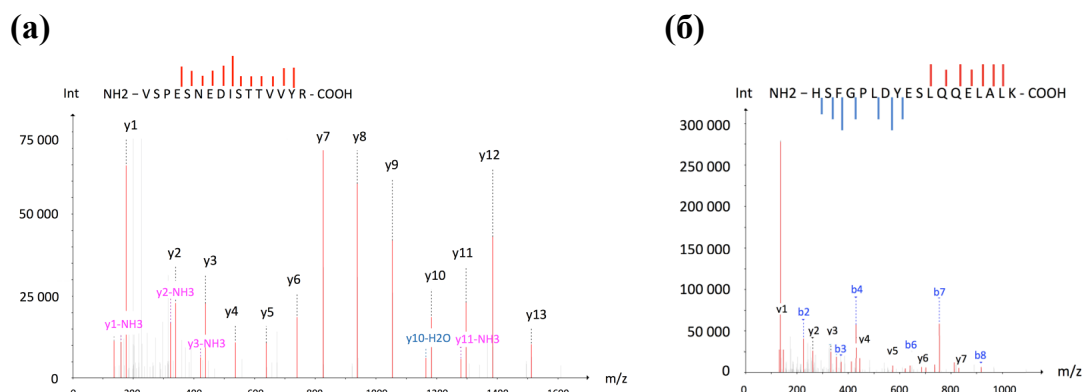


Рисунок 4. Масс-спектры и фрагментные ионы пептидов VSPESNEDISTTVVYR (а) и HSFGPLDYESLQQELALK (б), протеотипических для сплайс-опосредованного варианта глутаминазы O94925-3. По оси абсцисс отмечены значения отношения массы к заряду m/z , по оси ординат – значение интенсивности Int.

Для иллюстрации детекции полиморфизма аминокислот на рисунке 5а приведена протеоформа пептидазы O75439, где в позиции 396 остаток глутаминовой кислоты заменен на остаток аспарагиновой кислоты. Детекция серии у-ионов пептидных фрагментов (y5-y10), содержащих полиморфное положение, позволяет с уверенностью говорить о обнаруженной аминокислотной замене. Этот

полиморфизм был также ранее детектирован на транскриптомном (по данным dbSNP) и на протеомном (по данным GPMdb) уровнях.

Детекция аминокислотного полиморфизма в сплайс-опосредованной протеоформе – редкое событие. Среди всех уникальных пептидов белков человека в базе UniProt только около 0,03% подходят для детекции продуктов сплайсинга, в которых произошла точечная замена (Lisitsa et al., 2014). Протеоформа лиазы Q96I15-2 с заменой аланина на треонин в 175-ом положении была детектирована по семи пептидам, которые картировались на сплайс-форму Q96I15-2. Один пептид содержал еще и полиморфное положение. Пики ионов y15 и y16, которые относятся к пептиду VSGQTEVDDILA AVRPTTR и содержат полиморфное положение с С-конца, свидетельствует о детекции ОАП-варианта (рис. 5б).

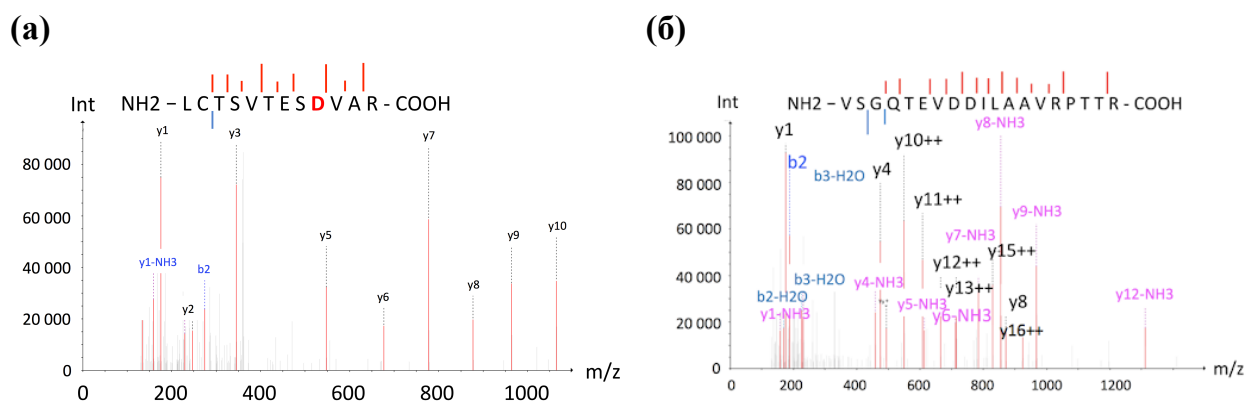


Рисунок 5. Масс-спектр и фрагментные ионы пептида LCTSVTESDVAR, специфичного протеоформе пептидазы O75439 с заменой остатка глутаминовой кислоты в позиции 396 на остаток аспарагиновой кислоты (а) и пептида VSGQTEVDDILA AVRPTTR, специфичного сплайс-варианту Q96I15-2, в котором произошла замена остатка аланина на треонин в 175-ой позиции.

Суммарно в 96-и ячейках геля было выявлено 38 704 уникальных пептидов, из которых 8 412 – протеотипические, т.е. специфично характеризующие 2 156 протеоформ, кодируемых 2 115 генами. После технической фильтрации исходного массива надежно детектированными являлись 1 618 уникальные протеоформы 1 589 генов. Среди детектированных протеоформ было 76 СО-вариантов, 81 протеоформа с ОАП (из которых пять неразличимы масс-спектрометрически), пять протеоформ, в которых замена аминокислоты возникла одновременно с альтернативным сплайсингом, и один случай делеции. Помимо однозначно детектированных протеоформ, также было обнаружено 6 649 белковых групп, состоящих из нескольких протеоформ, кодируемых одним геном, и 301 группа – объединяющая белковые продукты разных генов.

Большая часть из обнаруженных протеоформ (998) локализовалась в нескольких ячейках геля одновременно (в среднем в шести ячейках) вследствие

технических артефактов или из-за протеолиза и посттрансляционных модификаций. Для подтверждения модификаций файлы масс-спектров были проанализированы повторно с указанием наиболее часто встречающихся ПТМ: фосфорилирование по остаткам серина, треонина и тирозина, ацетилирование лизина, метилирование лизина и аргинина, амидирование С-концевых остатков и деамидирование остатков аспарагина и аспарагиновой кислоты. Дальнейший анализ заключался в сопоставлении данных о физико-химических параметрах белков, химических и посттрансляционных модификациях с координатами молекулярной массы и изоэлектрической точки, характерным для тех ячеек, в которых протеоформы были детектированы.

3.4. Аннотация протеоформ с учетом их физико-химических параметров

Метод 2DE применительно к биообразцам – это способ сделать выводы о возможных абберациях в составе протеома (Petrak et al., 2008). Выявление протеоформы в ячейке геля, координаты MW и pI которой отличаются от теоретических, свидетельствует о модификациях, протеолизе белка или иных молекулярных событиях. Для оценки влияния каждого из возможных факторов был проанализирован массив из 530-и достоверно идентифицированных в единственной ячейке геля протеоформ клеток опухолевой линии HepG2. Отобранный массив данных включал 473 канонических варианта белков, 29 протеоформ с одноаминокислотными полиморфизмами, 27 сплайс-вариантов и одну сплайс-протеоформу с аминокислотной заменой.

Для каждой из детектированных протеоформ были сопоставлены теоретически рассчитанные с помощью инструмента Pyteomics (Goloborodko et al., 2013) параметры pI и MW с координатами ячейки, в которой протеоформа была идентифицирована. Для половины массива протеоформ (243 варианта) теоретические и экспериментальные координаты локализации совпали. Для оставшихся протеоформ был проанализирован вектор сдвига экспериментальных координат относительно теоретических.

Треть тестового массива (163 протеоформы) была идентифицирована в ячейках, координаты которых соответствуют меньшим по сравнению с расчетными значениям MW. Причиной такой миграции может быть протеолиз белка на несколько фрагментов. Последовательность фрагментов может содержать протеотипические пептиды, необходимые для идентификации «родительской» молекулы, при этом молекулярный вес фрагментов естественным образом оказывается ниже веса целой протеоформы. Анализ белковых последовательностей показал, что в большинстве случаев (71%) пептиды картируются компактно, а не распределяются по всей последовательности.

Напротив, сдвиг координат локализации в область больших масс может быть вызван олигомеризацией протеоформы или ковалентной сшивкой с интерактивными партнерами или контаминантами. Такой сдвиг наблюдался для 68-и протеоформ массива. Помимо ковалентной сшивки или олигомеризации

причиной сдвига такого рода могут быть особенности структуры и аминокислотного состава белка. Значительно снижает подвижность молекулы высокое содержание остатков пролина, придающих полипептидной цепи «угловатую» форму. Примером влияния остатков пролина на подвижность белка является опухолевый антиген Р53 (Р04637). Молекулярная масса белка Р53, который на 11% состоит из остатков пролина, составляет 43,5 кДа, но он локализуется на 10 кДа выше по оси молекулярной массы. Для 18-и протеоформ, аминокислотная последовательность которых на 10% и более состояла из остатков пролина, было обнаружено смещение в область больших масс.

Для 41 протеоформы тестового массива было зафиксировано смещение по оси pI в «кислую» область вследствие деградации или множественных химических и посттрансляционных модификаций, которые незначительно меняют молекулярный вес белка, но существенно влияют на его суммарный заряд (Zhu et al., 2005). Стоит отметить, что примерно 85% пептидов были детектированы с химическими модификациями – окислением метионина или карбоксилированием цистеина, возникающих в ходе пробоподготовки клеток HepG2. Посттрансляционные модификации, смещающие локализацию протеоформы в область меньших значений pI (фосфорилирование, ацетилирование, деамидирование и др.), были выявлены в 23-х случаях.

Одну из самых сложных для интерпретации ситуаций представляли случаи сдвига локализации протеоформы в область щелочных pI , которых в тестовом массиве было пять. К примеру, пептиды импорта О95373 (pI 4,7) были детектированы в ячейке Н2 (pI 8,9 – 10,0). Преимущественно ПТМ смещают локализацию протеоформы в кислую область относительно немодифицированной белковой молекул. Однако, есть немногочисленный класс модификаций, которые могут привести к щелочному сдвигу, например, является амидирование С-концевых остатков (Kuyama et al., 2009). Метилирование лизина (Evich et al., 2016) также способно перераспределить заряд на атоме азота аминокислотной группы, что приводит к небольшому увеличению значения изоэлектрической точки. Эти модификации были рассмотрены в качестве возможной причины несовпадения экспериментальных и теоретических координат локализации протеоформы на геле. Выяснилось, что две из пяти протеоформ, смещенных в щелочную область, были метилированы и амидированы. Причиной сдвига остальных трех протеоформ могли быть проблемы на этапе пробоподготовки или при проведении 2DE-эксперимента: недостаточная очистка от нуклеиновых кислот, полисахаридов и других контаминантов, неполная денатурация или технические артефакты (Naryzhny et al., 2016, Thiede et al., 2013).

3.4.1. Детекция пептидов одной протеоформы в нескольких ячейках геля

Самый распространенный тип посттрансляционных модификаций белков – фосфорилирование – заключается в замене нейтральных гидроксильных групп серина, треонина и тирозина отрицательно заряженными фосфо-группами. Как

следствие, изоэлектрическая точка модифицированной молекулы снижается, причем в отдельных случаях изменение может быть значительным: одиночное фосфорилирование может снизить pI белка на 1-2 единицы (Halligan, 2009).

Согласно молекулярной массе белка теплового шока P04792 (22783 Да) и расчетам значения pI (5,98), детекцию белка следовало бы ожидать в ячейке D9, или D8, если принять к сведению высокое содержание пролиновых остатков. Однако, пептиды данного белка были также детектированы еще в 8-и ячейках геля. Вероятно, сдвиг в область меньших масс был обусловлен протеолизом (ячейка H12, MW 6-15 кДа), а сдвиг в область больших масс, одновременно с изменением pI (ячейки A7, B6 и C2, 30-116 кДа) – гомо- или гетероолигомеризацией (Nefedova et al., 2013) или артефактной С-С сшивкой с белковыми партнерами – α - и β -тубулинами, молекулярный вес каждого из них составляет около 50 кДа.

Смещение в область меньших pI могло произойти из-за множественного фосфорилирования. По литературным данным, при фосфорилировании MAP-киназами остатка серина S82 белок теплового шока отщепляется от олигомера, состоящего из нескольких небольших белков (Stokoe et al., 1992). В ячейке C8, соответствующей интервалу 5,11-5,80 pI , было детектировано фосфорилирование в положении S82, причем в «оригинальных» ячейках D8 и D9 модификаций не обнаружили (рис. 6). Таким образом, можно заключить, что только в двух из 10-и ячеек детектирована истинная протеоформа (канонический вариант и вариант с фосфорилированием), а детекция протеотипических для белка P04792 пептидов в остальных восьми случаях – результат артефактов пробоподготовки.

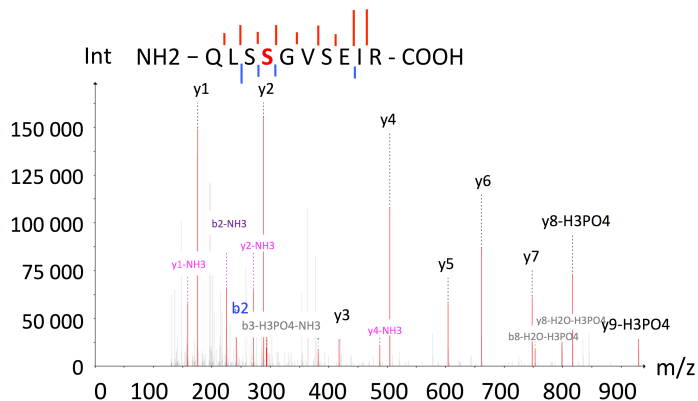


Рисунок 6. Масс-спектр пептида QLS(Phospho)SGVSEIR, который был детектирован в ячейке геля C8 (pI 5,11–5,80; MW 26–30 кДа). Серия пиков у-ионов (y8, y9 с С-конца) и b-ионов (b2, b3, b4, b8 с N-конца) пептида свидетельствует о фосфорилировании серина в положении 82.

Олигомеризацию как причину обнаружения одной протеоформы в разных ячейках иллюстрирует пример глутарил-КоА дегидрогеназы Q92947. Пептиды, картирующиеся на аминокислотную последовательность дегидрогеназы, были детектированы в двух удаленных ячейках геля – F1 (pI 6,88-7,82, MW > 116 кДа) и F5 (тот же интервал pI , MW 10-52 кДа). Пространственное разделение в геле обусловлено существованием в белковой смеси двух макромолекул дегидрогеназы: мономера с массой около 48 кДа, который мигрировал в соответствующую его

массе ячейку F5, и тетрамера с соответствующей ячейке F1 массой около 192 кДа (Keyser et al., 2008), где не были разрушены межсубъединичные контакты.

Характеристичные для протеоформы пептиды могут быть обнаружены в удаленных участках геля при ковалентной сшивке исследуемой молекулы с ее белковым партнером. Такой случай может быть представлен на примере оксидоредуктазы Q96HP4, пептиды которой были обнаружены в удаленных друг от друга ячейках F1 и F6. Предположительно, в ячейке F6 был обнаружен целевой белок Q96HP4, т.к. его масса (около 35 кДа) соответствует параметрам ячейки. Ячейка F1 соответствует менее подвижным белкам с массой более 100 кДа. Согласно ресурсу BioGRID, дегидрогеназа Q96HP4 взаимодействует с семьей белками-партнерами. Пептиды одного из белков-партнеров – шаперона P10809 – также были детектированы в ячейке F1, при этом и его масса (около 61 кДа) недостаточно велика, чтобы в ходе электрофореза этот шаперон самостоятельно мигрировал в ячейку F1. Можно полагать, что в ходе пробоподготовки или непосредственно 2DE-эксперимента произошла ковалентная сшивка между дегидрогеназой и шапероном.

Укорочение аминокислотной последовательности белка вследствие протеолиза приводит к сдвигу в область низких молекулярных масс и может повлиять на уровень pI , смещая баланс положительно- и отрицательно-заряженных остатков аминокислот. Согласно полученным данным, примерно 25% случаев обнаружения пептидов протеоформы в нескольких ячейках геля определяются протеолизом. Так, пептиды сигнал-узнающей частицы P61011 были детектированы в двух ячейках геля. Параметры ячейки G4, в которой было детектировано более 20 пептидов, распределенных по всей последовательности белка (около 40% покрытия), соответствуют его расчетному молекулярному весу (56 кДа, рис. 7а).

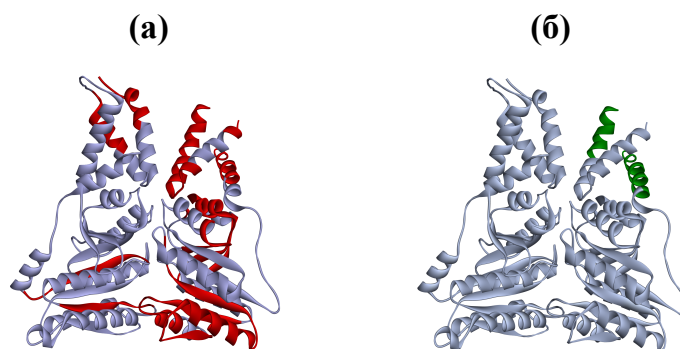


Рисунок 7. Экспериментально полученная модель трехмерной структуры белка P61011 из ресурса PDB. Красные участки структуры (а) соответствуют пептидам, детектированным в ячейке G4, зеленые (б) – в ячейке G10.

В ячейке G10, расположенной в зоне низких масс, детектировано всего два пептида первых α -спиралей белка (рис. 7б). Возможно, в ходе пробоподготовки или 2DE-эксперимента часть молекул белка P61011 была расщеплена на

фрагменты с низкой молекулярной массой, которые мигрировали в подвижный буфер и потому не представлены на геле.

3.4.2. Детекция пептидов, картирующихся на белковую группу

Детектированный для белковой группы набор пептидов картируется сразу на несколько протеоформ, кодируемых одним или несколькими генами, что не позволяет идентифицировать белковый вариант. Сопоставление теоретических характеристик заряда и массы протеоформ белковой группы с границами ячейки, в которых эти протеоформы были обнаружены, дает возможность сделать вывод о надежности идентификаций или различить даже высокоомологичные белки. Так, с аминокислотными последовательностями белков P49753 и Q86TX2 семейства ацил-КоА-тиоэстераз совпадали одни и те же пептиды. Однако, параметрам ячейки F5 (pI 6,88-7,82 и MW 40-52 кДа), в которой были детектированы эти пептиды, соответствует только протеоформа Q86TX2 (pI 6,9 и MW 46 кДа).

Сопоставление теоретических и экспериментальных pI и MW в ряде случаев позволяет установить сплайс-форму белка. Примером могут служить белковая группа, состоящая из пяти сплайс-форм ацил-КоА-тиоэстеразы (O00154, O00154-4, O00154-5, O00154-6, O00154-7): пептиды, детектированные для этой группы сплайс-вариантов, не позволяют установить, какая именно форма белка присутствует в биообразце. Однако параметрам ячейки E5 (pI 6,30-6,88, MW 40-52 кДа), в которой эти пептиды были детектированы, соответствует только вариант O00154-7 (pI 6,8 и, MW 40 кДа), поэтому можно предположить, что в рассматриваемой ячейке был детектирован именно этот сплайс-вариант.

Для аннотации протеоформ, обнаруженных как в единственной ячейке, так и в нескольких ячейках геля, был разработан алгоритм. Стартовой точкой алгоритма являлся перечень детектированных протеоформ (в т.ч. входящих в состав белковых групп), для которых сопоставляли теоретические и экспериментальные координаты локализации протеоформы на геле. Если экспериментальные MW и pI совпадали с рассчитанными значениями, протеоформу считали аннотированной. В противном случае рассматривали возможные причины, которые могли привести к сдвигу (химические и посттрансляционные модификации, изомеризация, протеолиз, деградация, особенности аминокислотного состава белка, технические артефакты).

В семи тыс. белковых групп, состоящих из белковых продуктов одного гена, удалось идентифицировать 751-у протеоформу. Стоит отметить, что около 10% данных групп составляли последовательности с ОАП. Для 301-й белковой группы, объединяющей продукты разных генов, удалось различить 25 протеоформ.

Для клеток линии HepG2 было аннотировано 2 358 протеоформ 1 904 генов, в их числе 1 658 канонических последовательностей, 224 СО-протеоформы и 115 ОАП, а также была аннотирована 31 протеоформа, в которой замена аминокислоты произошла в сплайс-варианте. Для 990 протеоформ на основании результатов масс-спектрометрического анализа и данных 2DE идентифицированы ПТМ (табл. 3).

Таблица 3. Протеом клеточной линии HepG2, детектированный с применением комбинации 2DE и панорамной масс-спектрометрии. СО – сплайс-опосредованная протеоформа, ОАП – одноаминокислотный полиморфизм, ПТМ – посттрансляционная модификация

Тип протео- формы	Совпаден ие теор. и эксп. pI и MW		Химическая модифи- кация		Протеолиз		Прочее (агрегация, высокая доля а.о. Pro и т.д.)		ВСЕГО	
	П*	Г**	П*	Г**	П*	Г**	П*	Г**	П*	Г**
Канон.вариант	481	481	57	57	107	107	107	107	752	732
СО-вариант	185	174	5	5	5	5	5	5	200	184
ОАП-вариант	42	39	6	6	6	6	5	5	59	55
СО+ОАП	24	23	2	2	1	1	0	0	27	25
ПТМ	527	527	124	124	163	163	92	92	906	858
СО+ПТМ	17	17	1	1	3	3	3	3	24	24
ОАП+ПТМ	33	31	6	6	8	8	9	8	56	50
СО+ОАП+ПТМ	2	1	0	0	1	1	1	1	4	3
Вставки, делеции	329	254	0	0	1	1	0	0	330	255

*П – количество протеоформ

**Г – количество соответствующих протеоформам белок-кодирующих генов

Из почти 52 тыс. транскрибируемых протеоформ для 12 тыс. генов, составивших референсную библиотеку, на протеомном уровне было аннотировано 2 358 протеоформ и выявлено 6 950 белковых групп, суммарно характеризующих трансляцию 8 512 генов. Зависимости между уровнем экспрессии транскрипта и вероятностью детекции соответствующей протеоформы не было выявлено, что подтверждается описанными в литературе результатами исследований транскрипто-протеома (Poverennaya et al., 2017, Fortelny et al., 2017).

Итоги инвентаризации протеома клеток опухолевой линии HepG2 методом 2DE с последующей масс-спектрометрией были сопоставлены с результатом панорамного исследования тех же клеток без фракционирования. Фракционирование, предваряющее масс-спектрометрический анализ, позволяет увеличить количество идентифицированных протеоформ на 76%, в т.ч. числе аннотировать aberrantные последовательности по вектору смещения локализации протеоформы на геле.

В соответствии с формулами оценки размера протеома, допускающими возникновение ПТМ и ОАП как в канонических, так и в СО-вариантах (Ponomarenko et al., 2017), на основании транскриптомных данных количество протеоформ в клетках линии HepG2 может варьироваться от 124 до 306 тыс. Сочетание двумерного гель-электрофореза и панорамной масс-спектрометрии с фокусом на наиболее часто регистрируемые посттрансляционные модификации позволило детектировать около 3-7% теоретически возможного многообразия протеоформ. При этом было экспериментально подтверждено, что источники (альтернативный сплайсинг, одноаминокислотные замены и посттрансляционные модификации) возникают независимо друг от друга.

3.4.3. Протеоформы линии HepG2, ассоциированные с опухолевыми процессами

Гены, белковые продукты которых были обнаружены в клеточной линии HepG2, были проанализированы с помощью веб-сервиса GOrilla, который предоставляет доступ к сведениям о локализации белков, их молекулярной функции и участии в биологических процессах. Согласно проведенной аннотации, основными функциями идентифицированных белков клеточной линии HepG2 являются связывание (с белками, рибонуклеотидами и нуклеозидтрифосфатами) и катализ (с фокусом на АТФазы, лигазы, изомеразы и хеликазы). Преимущественно, обнаруженные белковые продукты вовлечены в метаболические процессы и сплайсинг мРНК и равномерно распределены по большинству компартментов клетки.

Для выявления протеоформ клеточной линии гепатобластомы, ассоциированных с онкологическими процессами, мы проанализировали полученный перечень белков с использованием сведений ресурса TumorPortal (Lawrence et al., 2014). Этот ресурс содержит сведения о 320-и генах, значительно мутировавших в опухолях. Белковые продукты всех 320-и генов, связанных с онкологическими процессами, были детектированы в исследованном нами образце клеточной линии HepG2, что составило ~9% от всех выявленных генов (3 640). Данная величина в разы превышает вероятность наличия онкогенов в случайных выборках генов человека такого же размера: $1,5 \pm 0,2\%$ (для 1000 выборок с $p\text{-value} = 0,0001$).

Суммарно, для 25 генов, аннотированных в ресурсе TumorPortal, было детектировано 27 протеоформ. Продукты остальных 295 генов были выявлены по результатам анализа белковых групп. Этот массив содержит 13 досконально изученных онкоассоциированных генов. В их числе, например, белок семейства внутриклеточной сигнализации SMAD4. Также были детектированы белковые продукты четырех перспективных кандидатов в онкоассоциированные гены: среди них ген *RHEB* из суперсемейства гуанозинтрифосфатаз, вовлеченный в процессы регуляции роста и клеточного цикла (табл. 4).

В среднем, каждая протеоформа была идентифицирована по семи триптическим пептидам, что составило примерно 22% от всей аминокислотной последовательности. Выявленные протеоформы онкоассоциированных генов были представлены 21-ой канонической и тремя сплайс-опосредованными последовательностями, а также тремя последовательностями с ОАП.

В числе детектированных сплайс-вариантов была ГТФаза P01116-2, кодируемая протоонкогеном *KRAS*. Для данного гена известно две сплайс-формы: канонический вариант K-Ras4A (P01116-2) и альтернативный – K-Ras4B (P01116-2), отличающиеся по своим функциям. Каноническая форма активизирует апоптоз, а сплайс-вариант, напротив, его ингибирует.

Таблица 4. Характеристики детектированных в клеточной линии HepG2 протеоформ, которые кодируются значительно мутированными в опухолевых образцах генами

Ген	Детектиров. протеоформа	Покрывтне послед., %	Ген	Детектиров. протеоформа	Покрывтне послед., %
ATM ^{ИО}	Q13315:V:2906:I	1	IPO7	O95373	3
AZGP	P25311	18	KRAS ^{ИО}	P01116-2	24
CBFB ^{ИО}	Q13951	22	FUBP1 ^{ИО}	Q96AE4	35
	Q13951-2	35		Q96AE4-2	35
	(деамидирован)				
CDK4 ^{ИО}	P11802	44	PCK1 ^{ПК}	P35558:V:184:L (амидирован)	24
CCND1 ^{ИО}	P24385	10	PDSS2	Q86YN6	17
DDX5	P17844	40	RAD21 ^{ПК}	O60216	18
	(деамидирован)				
EIF2S2	P20042	12	RHEB ^{ПК}	Q15382	13
ELOC ^{ИО}	Q15369	63	RPS15	P62841	39
EZR	P15311	3	PPP2R1A ^{ИО}	P30153	7
GOT1	P17174	23	SMAD4 ^{ИО}	Q13485	3
IDHC ^{ИО}	O75874:I:367:V	56	SMC1A ^{ИО}	Q14683	12
MTOR ^{ИО}	P42345	4	TBL1XR1 ^{ИО}	Q9BZK7	23
PCBP1 ^{ПК}	Q15365	15			

^{ИО} – ген, для которого подтверждена связь с онкологическим процессом;

^{ПК} – перспективный кандидат в онкоассоциированные гены.

Из литературных данных известно, что гиперэкспрессия K-Ras4A ассоциирована с лучшей выживаемостью пациентов, а гиперэкспрессию K-Ras4B связывают с увеличением размера опухоли (Miura et al., 2012). В результатах транскриптомного профилирования клеточной линии HepG2 были выявлены оба варианта последовательностей, кодируемых протонкогеном KRAS, причем экспрессия сплайс-варианта (FPKM = 3,94) почти в 5 раз превышает экспрессию канонического транскрипта (FPKM = 0,82). На белковом уровне была надежно детектирована только протеоформа K-Ras4B путем идентификации характеристичного протеотипического пептида QGVDDAFYTLVR.

Как было показано на примере клеточной линии HepG2, использование транскриптомных и протеомных данных – это инструмент создания наиболее полного белкового портрета, учитывающего уникальность исследуемого объекта. Следующим шагом на пути к выявлению перспективных биомаркеров является сравнительный транскриптомно-протеомный анализ опухолевых образцов и нормальных тканей. Сопоставление результатов транскриптомно-протеомного профилирования опухолевого образца с результатами аналогичного персонализированного исследования нормальных и пораженных другими патологиями тканей, позволит выявить аномальные каскады реакций и биомаркеры, а также определить мишени для терапевтического вмешательства.

4. ВЫВОДЫ

1. По результатам транскриптомного профилирования клеточной линии НерG2 создана референсная библиотека транскрибируемых протеоформ, включающая 51 809 последовательностей, кодируемых 12 148 генами, содержащая 31 972 последовательности с одноаминокислотными полиморфизмами и 21 925 сплайс-опосредованных вариантов белков.
2. Сочетание поисковых машин X!Tandem, MS-GF+ и OMSSA, интегрированных в платформу SearchGUI, является оптимальным для эффективной обработки больших массивов результатов протеомных масс-спектрометрических экспериментов.
3. В результате анализа масс-спектрометрических данных, полученных при исследовании протеома клеточной линии НерG2 методом двумерного гель-электрофореза с последующей масс-спектрометрией, идентифицировано 1 618 протеоформ и 6 950 белковых групп, содержащих протеоформы одного или нескольких белок-кодирующих генов.
4. На основе сопоставления физико-химических свойств детектированных протеоформ и белковых групп, а также координат их локализации на двумерной электрофореграмме аннотировано 2 358 протеоформ. Среди этих протеоформ выявлены белковые продукты 26 генов, ассоциированных с опухолевыми процессами.
5. На протеомном уровне для клеточной линии НерG2 выявлено 3 640 из 12 148 транскрибирующихся генов, при этом для 1 904 генов обнаружено 1 658 канонических и 224 сплайс-опосредованных последовательностей, 115 случаев аминокислотных замен, а также 31 протеоформа, в которой одноаминокислотная замена произошла в сплайс-варианте. Для 990 протеоформ идентифицированы посттрансляционные модификации.

5. СПИСОК ОПУБЛИКОВАННЫХ РАБОТ ПО ТЕМЕ ДИССЕРТАЦИИ

Статьи:

1. Kiseleva O., Romashova Yu., Moskaleva N., Petushkova N., Teryaeva N., Belyaev A., Lisitsa A. Plasma preparation to measure FDA-approved protein markers by selected reaction monitoring. // J. Clin. Transl. Med. 2015. Vol. 4, № 1. P. 32.
2. Lisitsa A., Ponomarenko E., Kiseleva O., Poverennaya E., Archakov A. Molar concentration welcomes Avogadro in postgenomic analytics. // Biochem. Anal. Biochem. 2015. Vol. 4. P. 216.
3. Poverennaya E., Kopylov A., Ponomarenko E., Ilgisonis E., Zgoda V., Tikhonova O., Novikova S., Farafonova T., Kiseleva Y., Radko S., Vakhrushev I., Yarygin K., Moshkovskii S., Kiseleva O., Lisitsa A., Sokolov A., Mazur A., Prokhortchouk E., Skryabin K., Kostjukova E., Tyakht A., Gorbachev A., Ilina E. Govorun V., Archakov A. State of the art of chromosome 18-centric HPP in 2016: transcriptome and proteome profiling of liver tissue and HepG2 cells. // J. Proteome Res. 2016. Vol. 15, № 11. P. 4030-4038.
4. Поверенная Е.В., Киселева О.И., Пономаренко Е.А., Нарыжный С.Н., Згода В.Г., Лисица А.В. Мультиомная стратегия исследования протеома клеточной линии гепатоцеллюлярной карциномы HepG2. // Биомед. Хим. 2017. Т. 63, № 5. С. 373-378.
5. Kiseleva O., Poverennaya E., Shargunov A., Lisitsa A. Proteomic Cinderella: Customized analysis of bulky MS/MS data in one night. // J. Bioinf. Comput. Biol. 2017. Vol. 13. P. 1740011.
6. Kaysheva A., Kopylov A., Ponomarenko E., Kiseleva O., Teryaeva N., Potapov A., Izotov A., Morozov S., Archakov A. Relative abundance of proteins in blood plasma samples from patients with chronic cerebral ischemia. // J. Mol. Med. 2018. Vol. 64, №3. P. 440-448.
7. Киселева О.И., Лисица А.В., Поверенная Е.В. Протеоформы: методы исследования и клинические перспективы. // Мол. Биол. 2018. Т. 52, №3. С. 394-410.

Материалы конференций:

8. Kiseleva O., Lisitsa A., Krasnov G., Kudryavtseva A. Selection of SRM-compatible peptides to detect single amino acid variations. // Abstract Book of Proteomic Forum-2015. Berlin. 2015. P. 59.
9. Киселева О.И. Прогнозирование и выбор пригодных для SRM-анализа протеотипических пептидов вариантов альтернативного сплайсинга и аминокислотного полиморфизма белков хромосомы 18 по результатам высокопроизводительного секвенирования образцов ткани печени и клеточной линии HepG2, // Сборник тезисов международной конференции «Ломоносов». Москва. 2015. P. UID81624 (https://lomonosov-msu.ru/archive/Lomonosov_2015/7090/uid81624_report.pdf).

10. Киселева О., Поверенная Е. Анализ транскрипто-протеома для выявления aberrантных форм белков на примере клеточной линии гепатоцеллюлярной карциномы. // Сборник тезисов конгресса «Человек и Лекарство». Москва. 2016. С. 236.
11. Kiseleva O., Romashova Yu., Poverennaya E., Lisitsa A. Saturation point of one-type biological samples pool towards proteome heterogeneity cataloging. // Abstract Book of EuPA 2016. Istanbul. 2016. P. 125.
12. Poverennaya E., Kiseleva O., Ilgisonis E., Krasnov G., Zgoda V., Naryzhny S., Lisitsa A. Black box of transcriptoproteome: proteoforms scouting. // EuPA 2016. Istanbul. 2016. P. 41.
13. Kiseleva O., Poverennaya E. Proteoforms scouting in HepG2 cell line. // Moscow Conference on Computational Molecular Biology. – Moscow, 2017. P. 158 (<http://mccmb.belozersky.msu.ru/2017/proceedings/abstracts/158.pdf>).
14. Киселева О., Поверенная Е. «Мультиомикс» для исследования гетерогенности протеома. // Всероссийская конференция с международным участием «Высокопроизводительное секвенирование в геномике». Новосибирск. 2017. С. 20.
15. Kiseleva O., Poverennaya E.V. Proteomic Cinderella: customized analysis of large amount of MS/MS data in one night. // 16th Human Proteome Organization World Congress. Dublin. 2017. P. C-031.
16. Poverennaya K., Kiseleva O., Krasnov G., Zgoda V., Naryzhny S., Lisitsa A. Proteomics space odyssey: Strategy of proteoforms scouting // 16th Human Proteome Organization World Congress. – Dublin. 2017. C-052.
17. Kiseleva O., Poverennaya E. Proteomic Cinderella: customized analysis of large amount of MS/MS data in one night. // International Conference “Clinical Proteomics. Postgenome Medicine”. Moscow. 2017. P. 171.

БЛАГОДАРНОСТИ

Автор выражает глубокую благодарность д.б.н. Станиславу Николаевичу Нарыжному за предоставление результатов экспериментов по двумерному электрофорезу белков опухолевой линии HepG2, а также д.б.н., проф. Виктору Гавриловичу Згоде за предоставление результатов протеомного масс-спектрометрического анализа фракционированных и нефракционированных белков линии HepG2. За плодотворную дискуссию и содержательные комментарии по представлению результатов работы автор признателен д.б.н., проф. Алексею Евгеньевичу Медведеву, д.б.н. Елене Александровне Пономаренко и д.б.н., проф. Сергею Александровичу Мошковскому.

В основу данной работы легли результаты выполнения российской части международного проекта «Протеом человека», в частности полученные при реализации гранта РНФ № 15-15-30041 «Исследование гетерогенности протеома для решения задач клинической протеомики».

Автор благодарит ведущую научную школу академика РАН Андрея Валерьевича Лисицы (грант Президента Российской Федерации НШ-6313.2018.4).