

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ НАУЧНОЕ УЧРЕЖДЕНИЕ
«НАУЧНО-ИССЛЕДОВАТЕЛЬСКИЙ ИНСТИТУТ БИОМЕДИЦИНСКОЙ ХИМИИ
ИМЕНИ В.Н. ОРЕХОВИЧА»

На правах рукописи

КИСЕЛЕВА ОЛЬГА ИГОРЕВНА

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

03.01.04 – биохимия

Диссертация
на соискание ученой степени кандидата биологических наук

Научный руководитель:
к.б.н. Поверенная Екатерина Владимировна

Научный консультант:
академик РАН, д.б.н. Лисица Андрей Валерьевич

Москва, 2018

СОДЕРЖАНИЕ

СПИСОК СОКРАЩЕНИЙ	3
ВВЕДЕНИЕ	4
ГЛАВА 1. ОБЗОР ЛИТЕРАТУРЫ	9
1.1. Источники гетерогенности протеома	9
1.1.1. Полиморфизм единичных аминокислот	9
1.1.2. Альтернативный сплайсинг	11
1.1.3. Посттрансляционные модификации	12
1.1.4. Другие источники протеоформ	13
1.2. Современные подходы к идентификации протеоформ	16
1.2.1. Геномные и транскриптомные подходы	16
1.2.2. Протеомные подходы	21
1.3. Ресурсы, позволяющие оценить разнообразие протеоформ	34
1.3.1. Геномные и транскриптомные ресурсы	34
1.3.2. Протеомные ресурсы	38
ГЛАВА 2. МАТЕРИАЛЫ И МЕТОДЫ	44
2.1. Транскриптомные данные: источники и метод обработки	44
2.2. Протеомные данные	46
2.3. Анализ результатов масс-спектрометрического исследования протеома клеточной линии HepG2	50
2.4. Функциональное описание протеома клеточной линии HepG2	52
ГЛАВА 3. РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ	53
3.1. Транскрибируемые протеоформы клеточной линии HepG2	53
3.2. Поиск пептидных фрагментов для идентификации протеоформ	57
3.3. Протеом клеточной линии HepG2	64
3.4. Аннотация выявленных протеоформ с учетом их физико-химических параметров	71
3.4.1. Белковые продукты одного гена, идентифицируемые в разных областях двумерной электрофореграммы	77
3.4.2. Аннотация протеоформ, входящих в состав белковой группы	82
3.5. Протеоформы клеточной линии HepG2, ассоциированные с опухолевыми процессами	88
ЗАКЛЮЧЕНИЕ	93
ВЫВОДЫ	95
СПИСОК ЛИТЕРАТУРЫ	96
БЛАГОДАРНОСТИ	106
Приложение 1. Перечень масс-спектрометрически идентифицированных протеоформ	107
Приложение 2. Перечень аннотированных протеоформ	114

СПИСОК СОКРАЩЕНИЙ

а.о. – аминокислотный остаток

АС – (англ. *alternative splicing*) альтернативный сплайсинг

БКГ – белок-кодирующий ген

ИФА – (англ. *immunoenzyme analysis*) иммуноферментный анализ

МС/МС – (англ. *shotgun LC-MS/MS*) панорамный масс-спектрометрический анализ, сопряженный с жидкостной хроматографией

нсОНП – (англ. *nonsynonymous single nucleotide polymorphism*) несинонимичная однонуклеотидная замена

ОАП – (англ. *single amino acid polymorphism*) одноаминокислотный полиморфизм

ПТМ – (англ. *posttranslational modification*) посттрансляционная модификация

ПЦР – (англ. *polymerase chain reaction*) полимеразная цепная реакция

СО – сплайс-опосредованный вариант

2DE – (англ. *two-dimensional gel electrophoresis*) двумерный гель-электрофорез

FDR – (англ. *false discovery rate*) уровень ложноположительных результатов

FPKM/RPKM – (англ. *fragments (reads) per kilobase million*) программно рассчитываемый уровень экспрессии транскрипта/гена

SRM – (англ. *selected reaction monitoring*) масс-спектрометрический метод направленного мониторинга выбранных реакций

ВВЕДЕНИЕ

Актуальность выбранной темы

Биохимические процессы, протекающие в живых организмах, в значительной степени обусловлены геномными директивами, однако не ограничены ими. Геномные и транскриптомные исследования зачастую носят вероятностный характер, а потому фрагментарные сведения, которые они предоставляют, нуждаются в валидации на уровне протеома. Белок-кодирующая часть ДНК транскрибируется и далее транслируется в белковый продукт. Интеграция данных, полученных на нескольких комплементарных уровнях передачи биологической информации, позволяет приблизиться к пониманию сути процессов, протекающих в живых системах.

Сложность живых систем при сравнительно малом количестве белок-кодирующих генов (БКГ) достигается за счет регуляции транскрипции, трансляции и посттрансляционных модификаций, обеспечивающих многообразие белковых форм. Доказательством этому являются результаты проекта «Геном человека»: по итогам выполнения этого проекта было выявлено около 20 тысяч БКГ, что ниже первоначальных оценок примерно в 50 тысяч раз [1]. Вследствие альтернативного сплайсинга (АС) один БКГ продуцирует в среднем около четырех сплайс-опосредованных форм белков (СО-вариантов). С учетом возможных одноаминокислотных полиморфизмов (ОАП) и посттрансляционных модификаций (ПТМ) суммарное количество белковых вариантов, или протеоформ, кодируемых одним геном, может составлять сотни и даже тысячи [2,3].

Знания о многообразии протеоформ позволят сформировать полное представление о гетерогенном протеоме и в будущем будут способствовать снятию напряженности, растущей с постоянно увеличивающимся числом потенциальных биомаркеров, которые не переходят в категорию применяемых в клинике.

Степень разработанности темы

На настоящий момент исследования в области протеомики сосредоточены на канонических¹ формах белков. Накопленный массив геномных, транскриптомных и протеомных данных наряду с высоким уровнем современных

¹ Последовательность считается канонической, если она чаще других встречается среди нескольких альтернативных транскрипций, имеет наибольшее сходство с белковыми формами ортологами, или аминокислотный состав или длина данной формы позволяет наиболее полно описать домены, полиморфизмы и т.д. (http://www.uniprot.org/help/canonical_and_isoforms).

постгеномных технологий дает возможность приступить к идентификации протеоформ.

При возникновении и развитии патологии в органах и тканях аккумулируется множество белковых вариантов (абберрантных протеоформ), в норме отсутствующих вовсе или экспрессирующихся в незначительных количествах [4]. Накапливаясь по мере развития заболевания, такие варианты прямо или косвенно могут влиять на регуляцию клеточного цикла, апоптоз или репарацию ДНК. Особый интерес вызывают протеомные профили злокачественных опухолей, обладающих наибольшей гетерогенностью протеома – как по сравнению с другими аналогичными опухолями, так и с нормальной тканью. В опухоли количество мутантных вариантов белка в расчете на один ген может в десятки раз превышать их количество в нормальной ткани [5].

В 2014 году были опубликованы результаты двух крупномасштабных проектов, представивших «черновики» протеома человека и рапортовавших об идентификации на белковом уровне более 90% БКГ [6,7]. Тем не менее, доподлинно неизвестно, сколько всего протеоформ составляют протеом организма или его отдельных органов и тканей в норме и при патологии. Недостаток информации во многом связан с методическими сложностями экспериментального подтверждения трансляции протеоформ. Эти сложности обусловлены динамической природой протеома и ограничениями по чувствительности существующих аналитических методов [8]. По прогнозам, учитывающим различные сценарии возникновения протеоформ, протеом человека может содержать от сотен тысяч до нескольких миллиардов уникальных белковых продуктов [3,9].

Идентификация протеоформ, влияющих на фенотип заболевания, является важной частью исследования молекулярного профиля конкретного пациента. Каталогизация гетерогенного протеома закладывает основу для понимания принципов функционирования сложных биохимических систем. Персонализированный подход к терапии, учитывающий уникальные aberrации каждого пациента на геномном, транскриптомном и протеомном уровнях, становится одним из приоритетных векторов развития молекулярной медицины.

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.

Научная новизна работы и практическая ценность

В работе показаны возможности протеоформики – нового раздела протеомики, направленного на выявление протеоформ, т.е. различных белковых форм, кодируемых одним геном. Термин «протеоформика» был предложен российскими авторами в 2014 году [10]. Нами впервые дано экспериментальное обоснование существования множественности протеоформ, специфичных для исследуемого образца, и сформулирован методический подход к их единовременному выявлению. В результатах масс-спектрометрического анализа, следующего за двумерным разделением белковой смеси в геле, на основании транскриптомных данных выявлены протеоформы, образованные вследствие

альтернативного сплайсинга, однонуклеотидных замен и посттрансляционных модификаций.

Впервые получена карта протеоформ конкретного биологического образца, характеризующая трансляцию 20% генов в геноме человека.

Впервые проведен анализ различных поисковых алгоритмов и их комбинаций в составе платформы SearchGUI для выбора оптимального сочетания для обработки массива масс-спектров. Разработано приложение MSCinderella, обеспечивающее потоковый запуск и обработку множества файлов масс-спектров поисковыми алгоритмами платформы SearchGUI.

Полученные результаты позволяют по-новому использовать метод двумерного гель-электрофореза и усовершенствовать экспериментальные подходы к протеотипированию клеток и тканей. Кроме того, в перспективе, за счет сопоставления данных протеотипирования образца в норме и при патологии, полученные результаты позволят сократить время и трудозатраты на разработку биомаркеров – особенно учитывая, что в качестве биомаркеров могут быть более перспективны комбинации транслируемых протеоформ, а не отдельные aberrантные белковые варианты.

Методология и методы диссертационного исследования

В работе использованы результаты транскриптомно-протеомного профилирования, выполненного современными экспериментальными методами [11,12]. Для обработки данных применены современные биоинформатические подходы и соблюдены требования, сформулированные международным сообществом «Протеом человека» [13], а также учтены рекомендации ведущих научных школ в области протеомики [14].

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.

Степень достоверности и апробация результатов

Достоверность полученных результатов подтверждается 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 фактическим материалом.

Основные положения диссертационной работы были представлены в виде стендовых 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).

По теме диссертации опубликовано 17 работ, из которых 7 статей в рецензируемых научных журналах и 10 публикаций в трудах конференций.

ГЛАВА 1. ОБЗОР ЛИТЕРАТУРЫ

1.1. *Источники гетерогенности протеома*

Протеом является наиболее близким к фенотипу уровнем реализации геномной информации. При этом количество форм белков значительно превосходит количество кодирующих их генов. Основными процессами, обуславливающими кодирование одним геном нескольких форм белков, являются несинонимичные однонуклеотидные замены на уровне генома, альтернативный сплайсинг пре-мРНК на транскриптомном уровне и посттрансляционные модификации на уровне протеома.

1.1.1. *Полиморфизм единичных аминокислот*

Несинонимичный однонуклеотидный полиморфизм (нсОНП) наблюдается примерно в каждой 860-ой паре оснований, что делает его одним из основных инструментов создания гетерогенности на уровне генов. Распространенность этого события иллюстрирует тот факт, что геномы двух здоровых человек в среднем отличаются друг от друга на 4 млн нуклеотидных замен [15]. К настоящему моменту известно о более чем 150 млн нсОНП в человеческих геномах.

Несинонимичные замены нуклеотидов на уровне генов влекут не только возникновение старт- или стоп-кодонов, но и одноаминокислотные полиморфизмы (ОАП) в белках, финальных акторах биохимических процессов (рис. 1). Изменение функции белка и сценариев его взаимодействия с другими белками обусловлено изменением состава, стабильности и конформации протеоформы, несущей полиморфизм [16].

Биохимические реакции в живых системах чувствительны к стереометрическим параметрам взаимодействующих соединений. Даже при сохранении продуктов реакции, протекающей с участием модифицированного белка, возможны изменения кинетических параметров процесса. К примеру, одноаминокислотные замены в белке NIS, отвечающем за транспорт иодида натрия, становятся причиной нарушения электростатических взаимодействий в его трансмембранных доменах и снижения активности транспорта ионов иода, что приводит к гипотиреозу [17]. Замена аминокислотного остатка (например, объемного бокового радикала на более компактный) может повлиять на физико-

химические свойства протеоформы: так, только 20% ассоциированных с заболеваниями ОАП не изменяют стабильность белка [18].



Рисунок 1. Схематическое представление эволюции мутации, возникшей в кодирующей области ДНК: через транскриптомный и протеомный уровни к фенотипическому проявлению.

Далеко не каждая точечная мутация приводит к заметному изменению свойств белка. Условно одноаминокислотные замены можно разделить на нейтральные (англ. *passenger*) и ассоциированные с патологией (англ. *driver*). Первые возникают спонтанно на протяжении всей жизни и не играют значимой роли в формировании фенотипа патологии. Вторые происходят в ассоциированных с болезнью генах, влияющих на функциональную активность белка. В случае онкозаболеваний мутации могут внести значительный вклад в развитие патологии, предоставив преимущество опухолевым клеткам в росте [19]. Разделение мутаций на ассоциированные с патологией и нейтральные наиболее актуально в исследованиях онкозаболеваний, но применимо и для описания мутаций, сопровождающих заболевания другой природы.

В большинстве случаев даже в онкогенах и онкосупрессорах значимыми являются не более 40% ОАП, причем зачастую критическое влияние оказывает не одиночная замена, а паттерн из полиморфизмов [20]. Однако, известны единичные мутации, которые значительно преобразовывают весь белковый профиль. Классическим примером такой мутации является замена аспарагина на валин в шестой позиции β -субъединицы гемоглобина, которая обеспечивает устойчивость к малярийному плазмодию за счет изменения его третичной структуры [21].

Наиболее изучены случаи существенного изменения свойств белка за счет замен, не обеспечивающих преимуществ в выживании, а, напротив, ассоциированных с возникновением патологии. «Эффект бабочки» в развитии онкологического заболевания, когда единственная замена преобразует профиль

экспрессии нормальной клетки в профиль злокачественной опухоли, описан на примере замены H1047R белка PIK3CA, который в линии здоровых клеток молочной железы MCF-10A отвечает за процессы клеточного роста, пролиферации и выживания [22].

Ассоциированные с патологией однонуклеотидные замены и транскрибируемые с них замены аминокислот представляют интерес для поиска биомаркеров и терапевтических мишеней.

1.1.2. Альтернативный сплайсинг

Аннотирование транскриптомов различных организмов показало, что практически каждый мультиэкзонный ген подвержен альтернативному сплайсингу (АС), т.е. вычленению «немых» интронов и сшивке информационно-насыщенных экзонов по нескольким сценариям. В результате сшивки комбинации экзонов, входящих в состав БКГ, образуются различные варианты уникальных мРНК, которые впоследствии могут быть транслированы в протеоформы [23].

Вклад АС в увеличение гетерогенности протеома – предмет множества дискуссий [2,24]. Одна из популярных гипотез определяет роль АС как ключевого инструмента тканеспецифичных белок-белковых взаимодействий [25]. Противоположная гипотеза основана на предположении о стохастическом характере АС и том, что далеко не все сплайс-транскрипты кодируют сплайс-опосредованные (СО) белки [26]. Помимо оценки вклада альтернативного сплайсинга в гетерогенность протеома, актуальность подобных исследований определяется еще и тем, что продуцируемые одним геном сплайс-формы с равным успехом могут быть как не замеченными организмом, так и критически важными: жизненно необходимыми или губительными. Баланс между различными сплайс-формами принципиален для функционирования здоровой клетки: например, нарушение соотношения между сплайс-формами 3R и 4R τ -белка, которые в норме содержатся в равных количествах, ассоциировано с развитием болезни Альцгеймера [27].

Несбалансированная трансляция сплайс-форм обеспечивает почву для развития онкозаболеваний как в начале опухолеобразования, так и на стадии прогрессирования болезни [28]. Иллюстрацией связанного с раком АС может

служить онкосупрессор *TP53*. Известно, что *TP53* порождает 12 сплайс-вариантов. Все они детектируются и в нормальной ткани, однако именно в опухолевых клетках наблюдается нарушение соотношения между канонической и альтернативными протеоформами, которое коррелирует со снижением выживаемости [29,30].

Важность учета АС при разработке стратегии лечения показана на примере вемурафениба, ингибирующего белок BRAF(V600E), который отвечает за пролиферацию клетки. Клинические исследования эффективности вемурафениба при меланоме продемонстрировали положительный отклик на терапию у 80% пациентов, который сопровождался последующим снижением эффективности препарата. Приобретаемая резистентность объясняется повышенной экспрессией склонного к димеризации сплайс-варианта белка BRAF. Этот сплайс-вариант лишен нескольких участков аминокислотной цепи, которые окружают домен, взаимодействующий с G-белками. Примечательно, что пациенты, у которых не было обнаружено склонных к димеризации вариантов BRAF, чувствительность к вемурафенибу не теряли [31].

Роль альтернативного сплайсинга в возникновении различных патологий открывает вопрос о применимости знаний об этом явлении для создания новых терапевтических стратегий: при разработке лекарств, воздействующих на конкретную сплайс-форму или поддерживающих баланс между ней и каноническим вариантом.

1.1.3. Посттрансляционные модификации

Еще одним базовым источником гетерогенности протеома являются посттрансляционные модификации, увеличивающие функциональное разнообразие белков ковалентным присоединением к ним функциональных групп (фосфо-, глико- и пр.) или целых белков (например, убиквитина). ПТМ возникают на всех этапах жизненного цикла клетки и сопровождаются изменением свойств, функций и пространственной структуры белков. Посттрансляционные модификации лежат в основе механизма контроля ряда ключевых событий в клетке – от передачи сигнала в ядро и регуляции транскрипции до запуска деградации белков.

Около 5% протеома представлено ферментами, которые непосредственно продуцируют более 400 видов ПТМ [32]. В нормальной клетке большинство модификаций происходит обратимо и используется в качестве сигнала к запуску и остановке процессов пролиферации. Аномальные ПТМ могут исключить клетку из процессов нормальной регуляции или повлиять на уровень экспрессии генов, активацию сигнальных путей и ряд других процессов, поэтому сведения о модифицированных белках важны при диагностике заболеваний. Около 10% биомаркеров плазмы крови и сыворотки, одобренных Управлением по контролю пищевых продуктов и лекарственных препаратов США (англ. US Food and Drug Administration), ориентированы на анализ ПТМ [33].

К хорошо изученным типам ПТМ относится убиквитинирование, которое участвует в системе распознавания и утилизации поврежденных белков [34]. Аномальное убиквитинирование нарушает механизм расщепления неправильно сложенных или отработавших свой срок белков. Потеря функциональной целостности протеома – один из главных факторов старения и риска развития различных патологий. Так, исследование степени модификации τ -белка в тканях пораженного болезнью Альцгеймера мозга выявило 30 сайтов фосфорилирования и три убиквитинированных остатка лизина. Гиперфосфорилированный и убиквитинированный белок, не поддающийся протеасомной деградации, предложено использовать в качестве маркера болезни Альцгеймера [35].

Профиль aberrantных модификаций, возникающих вследствие ПТМ, может рассматриваться в качестве биомаркера [36]. Степень модифицированности белка также может использоваться для персонализации стратегии лечения и прогнозирования ответа опухоли на получаемую терапию. К примеру, высокая степень фосфорилирования протеинкиназы PKB/Akt у пациентов с глиобластомой снижает эффективность эрлотиниба, ингибитора рецепторов тирозинкиназы, воздействующего на рецепторы эпидермального фактора роста EGFR [37].

1.1.4. Другие источники протеоформ

Редактирование РНК. В 1986 г. в митохондриях паразитов были впервые зафиксированы изменения нуклеотидов в последовательности мРНК. Ферментативная модификация транскриптов получила название редактирования

РНК [38]. Редактирование РНК включает в себя ряд посттранскрипционных модификаций рибонуклеотидов, приводящих к возникновению отличной от закодированной генами последовательности с измененными функциями и кодирующей способностью. Согласно предварительным оценкам, по меньшей мере 1,5% мРНК человека подвергаются редактированию [39]. Классические примеры редактирования РНК – модификации цитозина и аденина.

Наиболее подробно изучен случай редактирования РНК в энтероцитах, эпителиальных клетках тонкого кишечника. Эти клетки продуцируют протеоформу аполипопротеина В АроВ-48. Тот же ген экспрессируется и в клетках печени, однако в гепатоцитах последовательность его белкового продукта АроВ-100 почти вдвое длиннее варианта АроВ-48. Причина такой значительной разницы в структуре и свойствах вариантов аполипопротеина В – это замена С→U, переводящая кодон глутамин САА в стоп-кодон UAA. Эта модификация играет важную роль в метаболизме липопротеинов и определяет разницу механизмов транспорта жиров в кишечнике и печени [40].

Широко распространена модификация гидролитического дезаминирования аденозина в инозин (редактирование А→I), катализируемая семейством ферментов ADAR. Редактирование А→I наиболее обширно представлено в клетках центральной нервной системы [41]. Частоту событий редактирования РНК именно в этих клетках можно объяснить потребностью ЦНС в разнообразии протеоформ для гибкого функционирования.

Нарушение регуляции редактирования А→I ассоциировано и с онкологическими патологиями. Примером связи редактирования РНК с онкологическим заболеванием может служить модификация транскрипта тирозинфосфатазы, кодируемой онкосупрессором RPTN6. Множественные модификации нуклеотидов приводят к потере функций онкосупрессора и способствуют развитию миелолейкоза. Исследование биоматериала страдающих лейкозом пациентов выявило, что снижение частоты редактирования РНК у таких пациентов коррелирует с возрастанием периода ремиссии [42].

Редактирование РНК – важный, но недостаточно изученный механизм расширения генетического разнообразия, функциональную связь которого с патологическими состояниями еще только придется установить.

Химеризм. Гибридные (химерные) белки – это белки, закодированные продуктом слияния нескольких генов. Многие из таких генов, возникающих в результате хромосомных перегруппировок, ассоциированы с различными патологиями [43].

Одной из первых хромосомных перестроек, обнаруженных в опухолевых клетках, была так называемая филадельфийская хромосома Pb, характерная для миелолейкоза. Дефектная хромосома возникает в результате транслокации, объединяющей ген тирозинкиназы *ABL1* с геном *BCR*. Химерный белок BCR-ABL обладает онкогенными свойствами: его гиперэкспрессия активирует пути внутриклеточной передачи сигнала и вызывает избыточную пролиферацию клетки. BCR-ABL – перспективная терапевтическая мишень и маркер развития миелолейкоза [44].

Химеризм интересен и с позиции разработки лекарств: химерные молекулы сочетают в себе функциональные свойства белка-эффектора, повышающие распознавание и токсичность, со стабильностью белка-транспортера [45].

Нарушенный процессинг белков. Созревание белка подразумевает преобразование полипептида в активную молекулу с «правильной» пространственной структурой. Изменения в полипептиде могут повлиять на его конечную трехмерную структуру. Например, некоторые типы катаракты ассоциированы с мутацией R116C белка α -А-кристаллина, которая изменяет структуру белка и снижает его эффективность как шаперона в четыре раза по сравнению с белком дикого типа [46].

Не ликвидированные в эндоплазматическом ретикулуме ошибки процессинга связывают с различными заболеваниями – от болезни Альцгеймера до диабета [47,48]. Так, ОАП в ферментах, катализирующих расщепление белка в лизосомах, снижают активность лизосом, что приводит к лизосомальным болезням накопления – тяжелым генетическим расстройствам, характеризующимся аккумулярованием внутри клетки промежуточных протеоформ, которые в норме подвергаются деградации [49].

Основные источники гетерогенности протеома представлены на рисунке 2.

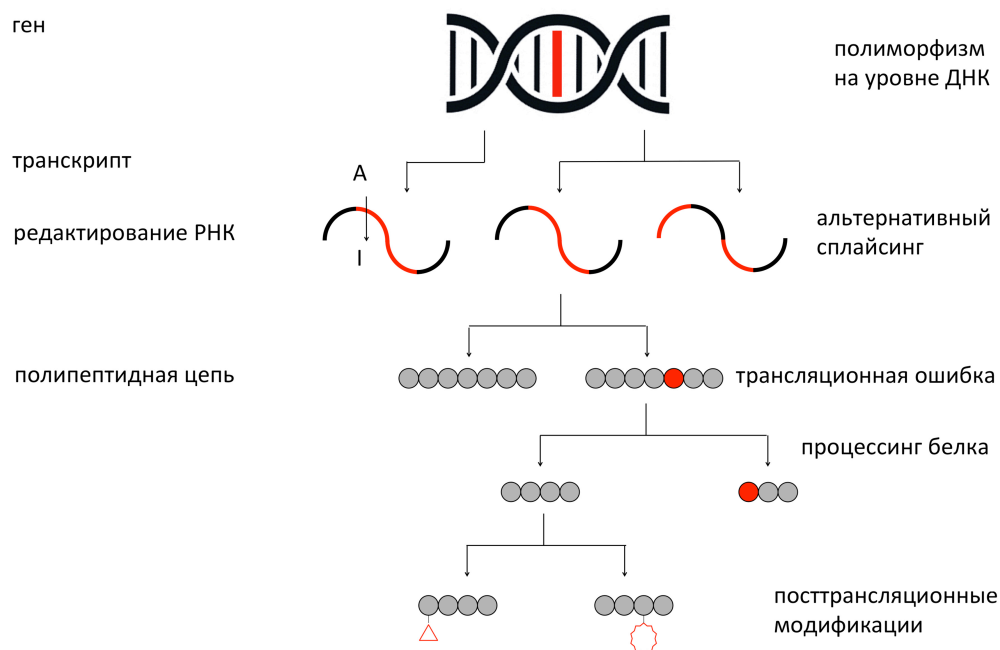


Рисунок 2. Источники гетерогенности протеома на уровнях гена, транскрипта и белка [50].

1.2. Современные подходы к идентификации протеоформ

Персонализированный подход к терапии, признающий уникальность каждого пациента на геномном, транскриптомном и протеомном уровнях, становится приоритетным направлением развития молекулярной медицины. Идентификация уникального набора генетических мутаций, транскриптов и протеоформ, определяющих фенотип заболевания, является важной задачей исследования персонального молекулярного профиля [51].

1.2.1. Геномные и транскриптомные подходы

Большая часть наследственной информации проявляет свою биологическую функцию посредством транскрипции ДНК в мРНК. Регуляция транскрипции генов – сложный механизм, определяющий фундаментальные процессы роста, дифференцировки, патогенеза и ответа на терапию. В отличие от весьма ограниченной вариабельности на уровне ДНК, гетерогенность продуктов транскрипции генов огромна – и в том числе из-за реализации АС и нсОНП.

Несмотря на то, что ПТМ находятся вне компетенций геномики и транскриптомики, а мутации нуклеотидных последовательностей не всегда удается зафиксировать на протеомном уровне, геномика и транскриптомика позволяют выявлять кандидатов для последующей протеомной валидации. Основными

инструментами анализа генома и транскриптома являются полимеразная цепная реакция, микрочипирование и секвенирование РНК.

Полимеразная цепная реакция. Базовым методом исследования изменений нуклеотидных последовательностей является полимеразная цепная реакция (ПЦР) в различных вариантах исполнения. Этот метод основан на последовательных процессах температурной денатурации, присоединения праймеров и репликации, экспоненциально увеличивающей число копий целевой нуклеотидной последовательности.

ПЦР в реальном времени (англ. real-time PCR) позволяет в автоматизированном режиме производить мультиплексный количественный анализ нуклеотидной последовательности в широком концентрационном диапазоне с высокой чувствительностью [52]. Основное отличие данной модификации от классической ПЦР состоит в том, что количество амплифицированной нуклеотидной последовательности измеряется в реальном времени после каждого цикла амплификации. ПЦР с обратной транскрипцией (ОТ-ПЦР, англ. reverse transcription PCR, RT-PCR) позволяет из химически лабильных молекул мРНК получить более стабильные кДНК, чтобы включить их в стандартный цикл ПЦР.

Метод ПЦР применим для качественного и количественного анализа модификаций в генах и транскриптах в направленном режиме. Этим методом удалось зафиксировать гиперэкспрессию сплайс-транскриптов *MDM2*, ассоциированную со снижением показателя выживаемости [53], и повышение уровня сплайс-варианта сурвивина, коррелирующего с прогрессией колоректальной и других карцином [54,55]. Замены отдельных нуклеотидов также можно изучать с помощью ПЦР – так была выявлена корреляция между возникновением точечной мутации протоонкогена *HER2* и прогрессированием рака молочной железы [56].

Успех ПЦР-анализа сильно зависит от чистоты исследуемых образцов и соблюдения протоколов пробоподготовки: в силу высокой чувствительности метода даже минимальные примеси могут привести к неудовлетворительным результатам. Метод позволяет анализировать в рамках одного эксперимента лишь несколько генов, поэтому ПЦР-тесты требуют больших время- и трудозатрат на разработку, что ограничивает их применение для общих скринингов [57].

Микрочипирование. Возникшая в 1990-х технология микрочипирования сразу завоевала популярность при исследованиях генов, профили экспрессии которых различаются в нормальных и пораженных болезнью тканях [58].

Микрочип представляет собой небольшую поверхность, на которую с высокой плотностью (до 1 млн на 1 см²) в определенном порядке нанесены олигонуклеотиды с известной последовательностью. Олигонуклеотиды выступают в роли зондов, с которыми гибридизуются комплементарные им цепи нуклеотидов из исследуемого образца, меченые флуоресцентным красителем. Содержание целевого транскрипта или гена определяется по интенсивности свечения флуоресцентных меток, закрепленных на зондах.

Для исследования альтернативного сплайсинга или полиморфизма отдельных нуклеотидов с помощью микрочипов необходимо исследовать не один участок последовательности, а все участки, где возможно изменение, и их комбинации, поэтому изучение генов/транскриптов с последовательностями, отличными от канонических, требует особой тщательности в подборе зондов для обеспечения надлежащей точности анализа.

Возможность применения микрочипов для анализа АС и ОНП была показана в широкомасштабных исследованиях генома и транскриптома [59,60]. Анализ 52 типов тканей человека позволил апробировать зонды, комплементарные к более чем 10 тыс. мультиэкзонных генов, и выяснить, что по меньшей мере $\frac{3}{4}$ из них подвержены АС. Справедливости ради следует отметить, что только треть выявленных случаев пропуска экзонов была валидирована методом ПЦР [59].

Для мультиплексного исследования нсОНП был разработан целеногеномный подход микрочипирования [61]. При таком подходе геном фракционируется рестриктазами и полученные фрагменты лигируются к адаптерам, после чего происходит реакция амплификации. Параллельно для определения сложности амплифицированных фрагментов и идентификации содержащихся в них полиморфизмов проводится фрагментация *in silico*. Эта модификация метода в сочетании с биоинформатическими алгоритмами позволила обнаружить более 70 нсОНП в генах-онкосупрессорах образцов эпителиальных и мезенхимальных опухолей [62]. С помощью микрочипирования были выявлены «лидеры» по частоте полиморфизма нуклеотидов при раке легкого [63] и

кишечника [64]: ими оказались компоненты цитоскелета, а также факторы контроля клеточного цикла и апоптоза.

Для исследования гетерогенности нуклеотидных последовательностей микрочипирование имеет ряд ограничений: метод характеризуется узким динамическим диапазоном исследуемых концентраций и высоким уровнем шума вследствие частой кросс-гибридизации [65].

Секвенирование нуклеотидных последовательностей. Первая декада нового тысячелетия было ознаменовано открытием параллельного секвенирования, известного как секвенирование нового поколения, революционной техники в анализе цепочек ДНК и РНК, которая позволяет быстро анализировать большие массивы биологических данных [66].

Несмотря на разные методы получения копий участков нуклеотидной последовательности и способы идентификации прочтенных нуклеотидов, общая схема работы для всех секвенаторов нового поколения может быть представлена в виде трех этапов. Первый этап секвенирования — это создание библиотеки случайных фрагментов исследуемой нуклеотидной последовательности, каждый из которых с двух сторон модифицирован адаптерами, т.е. заранее известными короткими последовательностями нуклеотидов. Адаптеры содержат праймеры, к которым присоединяется полимеразы – фермент, осуществляющий репликацию. Второй этап — многократное клонирование нуклеотидных фрагментов с помощью ПЦР. Третий этап заключается в определении первичной структуры всех фрагментов – например, по интенсивности свечения специфичной нуклеотиду флуоресцентной метки.

Прогресс биоинформатических алгоритмов сделал метод секвенирования РНК приоритетным при исследовании экспрессии генов и выявлении вариантов транскриптов. Данный метод имеет ряд серьезных преимуществ перед микрочиповыми технологиями: секвенирование РНК характеризуется меньшим фоновым сигналом и превосходит микрочипирование в ширине динамического диапазона определяемых концентраций. Современные секвенаторы позволяют за одно сканирование детектировать около миллиона вариантов нуклеотидных последовательностей [67] с высокой точностью (частота ошибочного прочтения нуклеотида колеблется в пределах $10^{-2} - 10^{-8}$) [68]. Метод применим не только для

поиска однонуклеотидных полиморфизмов [69], но и при изучении альтернативного сплайсинга без априорных знаний [70].

Несмотря на описанные преимущества, повсеместное использование метода ограничено в силу его дороговизны, сложности пробоподготовки и больших вычислительных мощностей, необходимых для обработки результатов [71]. Как и другие транскриптомные методы, высокопроизводительное секвенирование не является «универсальным солдатом» в поиске сплайс-транскриптов или транскриптов с ОНП и требует дополнения на уровне протеома [72]. Например, сопоставление генетически-детерминированных и наблюдаемых на белковом уровне полиморфизмов продемонстрировало, что из 13 выявленных несинонимичных ОНП человеческой эндонуклеазы G только три оказались полиморфными на белковом уровне [73].

Метод рибосомального профилирования. Корреляция между уровнями мРНК и транслируемого белка невысока, однако это не умаляет ценность уровня экспрессии транскрипта как индикатора возможности синтеза белка [74]. Для выявления доли транскриптов, подвергающихся трансляции, используется направленный метод рибосомального профилирования, позволяющий получить более точные и полные сведения об экспрессии генов.

Рибосомальное профилирование основано на глубоком секвенировании фрагментов мРНК, взаимодействующих с рибосомой. Метод заключается в ПЦР с обратной транскрипцией защищенных от нуклеаз комплексов мРНК-рибосома, результаты которой депонируются в геномную библиотеку для глубокого секвенирования. Метод позволяет *in vivo* зафиксировать активные в конкретный момент рибосомы, оценить экспрессию генов, определить, какие белковые продукты будут транслироваться в клетке, и спрогнозировать их представленность. Рибосомальное профилирование сокращает разрыв между геномом, транскриптомом и протеомом [75].

Рибосомальное профилирование в сочетании с протеомными методами особенно ценно при исследовании труднодетектируемых или короткоживущих белков. Так, сочетание рибосомального профилирования и хромато-масс-спектрометрии опухолевой линии HCT116 позволило обнаружить на несколько десятков больше белков, чем каждым методом по отдельности [76].

Несмотря на преимущества метода рибосомального профилирования, его применение весьма ограничено: метод требует большого количества биологического образца и подразумевает трудоемкий процесс его очистки.

1.2.2. Протеомные подходы

Существенный недостаток геномных и транскриптомных методов обусловлен стохастической природой протеома: только 10% мутаций, наблюдаемых в раковых клеточных линиях на геномном и транскриптомном уровнях, удастся зафиксировать в протеоме. Причиной этого может быть не только низкая приборная чувствительность, но и гетерогенность опухоли [77].

Протеом – это результат реализации процессов наследственной информации, а потому он предоставляет данные об уже осуществившихся, а не теоретически возможных белках и их характеристиках. Важную роль в регуляции активности белка играют ПТМ, не предсказуемые на транскриптомном и геномном уровнях, потому именно протеомные подходы способны в полной мере отразить функциональный портрет заболевания.

Иммуноферментный анализ. Иммуноферментный анализ (ИФА) основан на двух принципах: первый заключается в способности иммобилизованных белков, в том числе антител, сохранять свою функциональную активность, второй – в специфичности образования комплекса антитело-фермент.

Современные диагностические ИФА-тесты позволяют точно диагностировать как болезни, определяемые одним фактором (например, выявлять вирус папилломы человека через онкобелок Е6), так и заболевания мультипараметрической природы, которые требуют анализа нескольких маркерных молекул [78]. Прогресс мультиплексных ИФА дополняет тренды персонализированной медицины широким профилем мониторируемых маркеров, которые обладают большей разрешающей способностью при диагностике болезней со сходными фенотипами [79].

Расширение каталога антител, специфичных к конкретным участкам молекулы белка, способствовало разработке новых подходов к детекции протеоформ методом ИФА. Так, ИФА был использован для детекции замены K196E в S-белке – генетическом факторе риска развития венозного

тромбоэмболизма [80]. Применение специфичных моноклональных антител позволило детектировать СО-варианты белка TNFR2, который распознает и связывает цитокины семейства факторов некроза опухоли [81]. Особых успехов удалось добиться в решении нетривиальной и по определению недоступной для транскриптомных методов задачи – исследованию ПТМ сигнальных белков [82].

Результаты ИФА используются при разработке персонализированных стратегий лечения и контроле ответа пациента на терапию. Так, определение статуса фосфорилирования белков генов *EGFR* и *MEK* в клетках линии колоректального рака позволило получить важные сведения о развитии онкологического заболевания в организме конкретного пациента и способствовало персонализации лечения [83].

Несмотря на то, что ИФА часто называют «золотым стандартом» детекции белка, этот метод имеет ряд недостатков, связанных со специфичностью гибридизации, которые проявляются преимущественно при исследованиях протеоформ одного или нескольких родственных генов. Так, обнаружена протеоформа поверхностного антигена гепатита В с заменой T118M в гидрофильном экспонированном на поверхности участке белка. Этот участок «ускользал» от иммунного ответа, но являлся целевым для диагностических наборов ИФА. Ранее до 80% ИФА-наборов, применяемых для анализа крови доноров, не детектировали эту протеоформу и давали ложноотрицательные результаты для пациентов с гепатитом В, но теперь эта протеоформа учтена в тестовых наборах [84].

Другой пример связан с популярным биомаркером – простатоспецифичным антигеном PSA, применяемым для выявления рака простаты при скрининге плазмы крови. Последние исследования продемонстрировали неэффективность этого биомаркера из-за большой доли ложных результатов: исследование биопсии простаты подтверждало диагноз только у 25% пациентов с повышенным уровнем PSA, и напротив, возможность рака не исключалась даже при уровне PSA в пределах нормы. Применяемый для детекции простатоспецифичного антигена ИФА-тест предоставлял лишь обобщенную информацию о мастерной последовательности белка, который в действительности существовал в более чем 20 протеоформах [85]. Неспецифичность этого подхода, не учитывающего

различные белковые варианты, привела ко множеству необоснованных биопсий и операций. Этот пример геномного биомаркера подтверждает концепцию того, что даже хорошо зарекомендовавшие себя тесты зачастую не учитывают гетерогенность протеома, а потому получаемые результаты не всегда соответствуют истине.

Поиск потенциальных антител затратный, при этом не всегда приводит к обнаружению агентов, которые бы селективно взаимодействовали с конкретной протеоформой, а не каноническим вариантом [86]. Применимость метода ограничена константой образования комплекса антиген/антитело: белки с концентрацией ниже 10^{-12} М обнаружить не удастся [87].

Двумерный гель-электрофорез. Двумерный гель-электрофорез (англ. two dimensional gel-electrophoresis, 2DE) используется для разделения сложной белковой смеси по двум независимым физико-химическим параметрам: изоэлектрической точке pI и молекулярному весу MW белка (рис. 3).

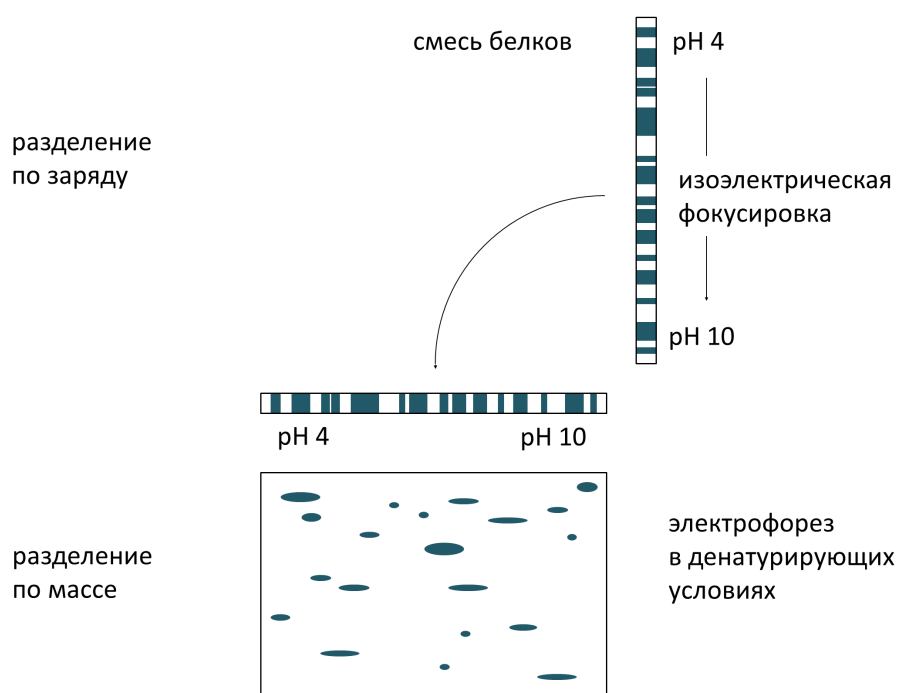


Рисунок 3. Схема проведения двумерного гель-электрофореза, результатом которого является белковая карта, где по локализации каждого окрашенного пятна на геле можно судить о физико-химических свойствах белкового варианта, а по интенсивности его окраски – о его концентрации.

Изоэлектрическая фокусировка достигается пропусканием тока через электролит, в котором устанавливается градиент потенциалов и формируется электрическое поле. Под действием поля и в соответствии с собственным зарядом

белки мигрируют к участку электролита, рН которого совпадает с их pI . Вторая ось разделения белков, ортогональная первой, – это их молекулярная масса. Разделение по MW осуществляется путем электрофореза в полиакриламидном геле и основывается на разнице в скоростях миграции белков сквозь гель под действием электрического поля. Для визуализации разделения в препарат добавляют краситель, обладающий сходной с белками электрической подвижностью [88].

Технология 2DE занимает одну из лидирующих позиций среди методов анализа и фракционирования белков. В сочетании с методами обогащения, специфичной окраски и последующей идентификации (например, с помощью масс-спектрометрии) этот метод позволяет фракционировать сложные белковые смеси, а также различать протеоформы на структурном уровне по их физико-химическим свойствам [89].

Исследование протеоформ с помощью 2DE можно проводить в нековалентном и ковалентном режимах. Реализованный в подходе DIGE (англ. Difference Gel Electrophoresis) принцип использования ковалентно связанных флуоресцентных меток, различающихся по длинам волн возбуждения и испускания, обеспечивает воспроизводимость анализа белкового состава нескольких образцов в формате одного геля [90].

Возможности 2DE в тандеме с панорамной масс-спектрометрией для идентификации протеоформ продемонстрированы как в узконаправленных исследованиях, ориентированных на детекцию конкретных белковых продуктов [91,92], так и в широкомасштабных поисковых проектах [12]. Так, например, с использованием одномерного и двумерного гель-электрофорезов с последующей масс-спектрометрией фрагментов гелевых пластин был исследован протеомный профиль культивируемых миобластов человека после индукции дифференцировки. В ходе исследования были зафиксированы изменения содержания различных сплайс-форм тропомиозинов и других белков, которые могли повлиять на дифференцировку клеток скелетной мускулатуры [93].

Среди недостатков метода двумерного электрофореза следует отметить невысокую воспроизводимость. Метод не позволяет надежно разделить низкокопийные белки или белки с «предельными» физико-химическими характеристиками: кислые или щелочные, гидрофобные, слишком объемные или,

напротив, компактные. Несмотря на перечисленные ограничения, 2DE остается мощным инструментом белкового анализа и успешно сочетается с другими протеомными технологиями.

Масс-спектрометрия. Сегодня высокопроизводительные стратегии хромато-масс-спектрометрического анализа, адаптированные под конкретную научную задачу, уверенно занимают центральное место среди протеомных методов. В типичном протеомном исследовании масс-спектрометрию предваряет хроматографическое фракционирование, позволяющее распределить пептиды в системе двух контактирующих несмешивающихся фаз. Хроматография используется для разделения по времени удержания смеси макромолекул, обладающих различными физико-химическими свойствами, а потому с разной силой взаимодействующими с поверхностью сорбента.

После хроматографической стадии ионизированный и фрагментированный пептид или белок характеризуется масс-спектрами, отражающими стадии фильтрации ионов по величине отношения их массы к заряду. Классификация масс-спектрометрических протеомных подходов по цели исследования включает две традиционные стратегии: панорамную (англ. *shotgun*), подходящую для поисковых экспериментов, и направленную (англ. *targeted*) – для анализа конкретной протеоформы.

Направленная масс-спектрометрия. Направленная масс-спектрометрия ориентирована на количественное определение конкретных протеоформ с известной аминокислотной последовательностью. Основной инструмент направленной масс-спектрометрии – метод мониторинга выбранных реакций (англ. *Selected Reaction Monitoring, SRM*). Метод SRM основан на масс-спектрометрической регистрации протеотипических пептидов. Такие пептиды картируются на белок (т.е. их аминокислотная последовательность полностью содержится в аминокислотной последовательности белка) и при этом встречаются в одном и только одном белке среди всех кодируемых геномом. Протеотипические пептиды могут служить «прототипом белка», его воспроизводимо детектируемым эквивалентом после ферментативного гидролиза.

Двойной фильтр на массово-зарядовые числа m/z протеотипического пептида-прекурсора и его дочерние ионы-фрагменты обеспечивает высокую

селективность метода. Мониторинг заранее выбранных переходов между родительским и фрагментным ионом приводит к повышению чувствительности на два порядка по сравнению с методом панорамного сканирования [94]. Направленная хромато-масс-спектрометрия все больше приобретает черты универсального инструмента воспроизводимого и точного анализа белков для решения широкого спектра медико-биологических задач: от верификации кандидатных биомаркеров патологических процессов [95] до масштабных проектов по каталогизации всех белков протеома и исследованию его гетерогенности [96].

Метод SRM эффективно используется для обнаружения единичных аминокислотных замен, связанных с различными заболеваниями. Так, SRM-анализ образцов плазмы представительной когорты пациентов с диабетом второго типа показал, что существует дисбаланс между количеством аберрантных и канонических протеоформ генов C7 и CFH, детектируемых в гомозиготных и гетерозиготных образцах [97]: диагностически значимые соотношения между уровнем экспрессии аллельных генов позволяют надежно различать между собой группы пациентов с ожирением и диабетом. В работе [98] методом SRM была выявлена онкоассоциированная мутация белка PAC, вовлеченного в различные онкологические процессы. По результатам исследования опухолевых и нормальных тканей и клеточных линий удалось обнаружить в протеотипическом пептиде мутацию G12V, специфичную для онкологических процессов. Аналогично, сочетая стратегию SRM с мечением пептидов стабильными изотопами и поиском по базе данных вариантов единичных аминокислотных замен, удалось выяснить, что содержание варианта карбамоилфосфатсинтазы T1406N в клетках гепатоцеллюлярной карциномы в 10 раз больше, чем в нормальных клетках [99].

Wu и соавторы [100] использовали SRM для исследования СО-вариантов остеопонтина. Количественный анализ трех сплайс-форм остеопонтина (OPNa, OPNb и OPNc) показал, что изоформа «а» специфична для рака легкого, где ее содержание примерно в 5 раз выше, чем в нормальных клетках.

Метод SRM также применялся для выявления и определения ПТМ. Например, в плазме крови человека определен статус гликозилирования пептид-N-гликозидазы F. Чувствительность этого анализа составила 10^{-10} М, что не уступает чувствительности иммуноферментных методов [101].

Несмотря на чувствительность, селективность и гибкость направленной масс-спектрометрии, есть факторы, ограничивающие ее применение для исследования протеоформ. Скорость-лимитирующей стадией разработки метода SRM является синтез протеотипических пептидов, выступающих в роли внутренних стандартов. Пептид, подобно антителу в методе ИФА, может оказаться неудачным – плохо ионизируемым и дающим невоспроизводимые результаты. Другой ключевой проблемой как метода SRM, так и протеомной масс-спектрометрии в целом, является выбор оптимальной референсной базы данных. Идентификация протеотипического пептида, характеристичного для конкретной протеоформы, происходит при совпадении экспериментального масс-спектра и теоретического масс-спектра из некоторой базы данных. Такие базы содержат неполную информацию о возможных протеоформах. Использование такой нерепрезентативной референсной библиотеки не позволяет зафиксировать новые ранее не аннотированные мутации в режиме направленной масс-спектрометрии, именно потому, что их нет в базах данных. Расширение библиотеки за счет невалидированных вариантов белков не является панацеей, поскольку повышает долю ложноположительных идентификаций и обесценивает качество анализа [102].

Выбор уникальных триптических пептидов также представляет сложную задачу из-за высокой гомологичности аминокислотных последовательностей и ограниченного числа протеотипических пептидов, которые могут различать протеоформы одного гена. Так, в ходе компьютерного моделирования протеотипических пептидов для сплайс-вариантов и сравнения с данными UniProt было установлено, что только 6% из всех возможных протеотипических пептидов могут быть использованы для обнаружения СО-протеоформ [10].

Панорамная масс-спектрометрия. Превалирующую долю протеомных данных получают методами панорамной протеомики, сочетающей жидкостную хроматографию с tandemной масс-спектрометрией [103]. Задача панорамной протеомики состоит в идентификации максимального числа белков в исследуемом образце. Эксперименты в панорамной протеомике сконструированы таким образом, что каждый попавший в масс-спектрометр, а после ионизированный и фрагментированный пептид или белок на выходе будет охарактеризован MS/MS-спектрами, отражающими два этапа масс-зарядной фильтрации. Ключевой этап в

панорамных анализах – это идентификация пептидов в полученных масс-спектрах. В отличие от геномных и транскриптомных подходов, в которых фрагменты нуклеотидных последовательностей прочитываются побуквенно, в протеомике пептиды идентифицируются по своего рода пептидному «отпечатку пальцев» – по совпадению полученных на практике масс-спектров с теоретическими спектрами всевозможных кандидатных пептидов, содержащихся в базе данных.

Наиболее популярны две комплементарные стратегии панорамного масс-спектрометрического анализа: «bottom-up» и «top-down», которые сходны в хромато-масс-спектрометрической части анализа, но различаются на стадии пробоподготовки. При использовании подхода «bottom-up» белки подвергают протеолизу, и полученную смесь пептидов после разделения на хроматографической колонке анализируют в масс-спектрометре. В случае «top-down» в масс-спектрометре анализируют целые белки – без предварительного расщепления.

Bottom-up. Последние два десятилетия масс-спектрометрия в режиме «bottom-up», в котором предусмотрен протеолиз белков на пептиды длиной 9–25 аминокислотных остатков, является приоритетным методом белкового анализа [104] (рис. 4а). Простота разделения пептидов и возможность прогнозирования фрагментации позволили использовать эту технологию для высокопроизводительного анализа сложных белковых смесей и целых протеомов.

В литературе описан ряд экспериментов по изучению протеоформ методом «bottom-up», ориентированных как на исследование конкретной мутации (например, замены, определяющей тип человеческого гемоглобина [105]), так и на каталогизацию всех вариантов целевого биоматериала [106]. Широкомасштабные «bottom-up» исследования клеточных линий или тканей человека позволяют детектировать десятки СО-вариантов, сотни ОАП и ПТМ, некоторые из которых удастся ассоциировать с заболеваниями [107,108]. Так, в нескольких клеточных линиях удалось детектировать замену T186R в белке BCAT2, связанную с развитием глиомы [109].

Метод «bottom-up» пользуется большой популярностью, хоть и обладает существенными недостатками. При анализе ПТМ эта технология не позволяет определять, например, был ли белок модифицирован в двух сайтах, или же

исследователь имеет дело с двумя популяциями белковых молекул, в каждой из которых по одной модификации. Для анализа сплайс-вариантов применимость «bottom-up» также ограничена, поскольку отличная от канонической протеоформы часть аминокислотной последовательности может оказаться слишком короткой и не содержать сайты протеолиза [110]. Вплоть до 95% некартируемых масс-спектров попадают в неинтерпретируемую зону аминокислотной последовательности протеоформы [111] – этот факт частично объясняется недостаточным качеством анализируемых масс-спектров, несовершенными поисковыми алгоритмами и неполнотой референсных библиотек аминокислотных последовательностей, не учитывающих возможные модификации. Идентификация модифицированных пептидов относится к непростым задачам. Если исследователь не укажет в настройках поиска конкретный полиморфизм или модификацию, то такой измененный пептид ожидаемо не будет идентифицирован, при этом учет всех возможных вариаций в поиске может привести к комбинаторному взрыву. Корректная интерпретация протеомных данных и использование особых «группоспецифичных» подходов к оценке уровня ложноположительных идентификаций позволяют надежно выявлять неканонические протеоформы [112].

Top-down. Масс-спектрометрия в режиме «top-down» позволяет исследовать протеоформы без предварительного протеолиза [113] (рис. 4б).

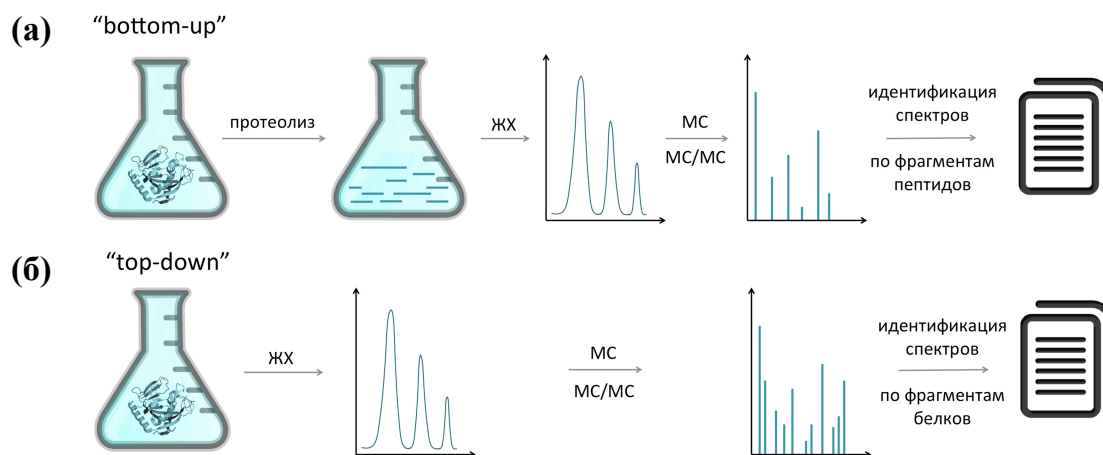


Рисунок 4. Принцип панорамной масс-спектрометрии с предварительным хроматографическим фракционированием в исполнении «bottom-up» (а) и «top-down» (б).

Для правильной идентификации белковых продуктов одного или нескольких родственных генов, часто сходных между собой по физико-

химическим свойствам, требуется высокоточное оборудование и специальные программные алгоритмы. К примеру, триметилирование лизина изменяет массу белка на 42,0470 Да, а ацетилирование – на 42,0106 Да. Разница составляет 0,0364 Да, и надежно ее зафиксировать масс-спектрометрией в режиме «top-down» не всегда представляется возможным. Более того, без предварительного протеолиза удастся исследовать только относительно короткие белки (до 50 кДа). Наряду с ограниченной чувствительностью и сложной приборной базой это ограничивает применение метода «top-down» для исследования модификаций, но не так сильно, как в случае «bottom-up».

В общем случае масс-спектрометрический анализ целых белков имеет достаточно узкую зону применения и ориентирован на тщательный анализ отдельных коротких белков или простых смесей с объективно невысокой пропускной способностью. Несмотря на серьезный прогресс в анализе целых белковых молекул, его чувствительность уступает чувствительности пептидного «bottom-up» анализа [114], однако даже с учетом несовершенств метод «top-down» незаменим в исследовании коротких протеоформ.

Стратегии обработки результатов панорамных масс-спектрометрических экспериментов. Прогресс высокопроизводительных технологий и биоинформатических подходов обработки экспериментальных данных позволил повысить качество и надежность протеомного анализа. Качественный скачок в развитии высокоэффективных протеомных технологий вывел на первый план нетривиальную задачу хранения и интерпретации результатов эксперимента: скорость генерирования масс-спектрометрических данных сегодня значительно превышает скорость их обработки до уровня воспринимаемых человеком результатов. Сложность и объем массива получаемых в рамках одного измерения масс-спектрометрических данных требуют их автоматизированной интерпретации посредством специализированных инструментов.

Ключевыми инструментами обработки масс-спектрометрических данных являются поисковые машины (англ. search engines). Принципиальная схема поискового алгоритма, обрабатывающего файлы выдачи масс-спектрометра,

включает несколько стадий. Первый этап заключается в определении области поиска, которая включает референсную библиотеку пептидных последовательностей, а также сведения о ферментативном гидролизе и возможных ПТМ. Второй этап подразумевает переход от формата файла масс-спектров, определяемого производителем оборудования, к «открытому» формату [115]. Далее следует стадия высокопроизводительной интерпретации tandemных MS/MS-спектров. Заключительный этап – статистическая валидация полученных результатов с установлением уровня их достоверности.

Наиболее насыщенная вычислениями стадия биоинформатического анализа протеомных данных заключается в использовании поискового алгоритма, который позволил бы проинтерпретировать полученные масс-спектры для того, чтобы поставить в пару к спектру пептид (англ. peptide-spectrum match, PSM). Многие спектры, особенно в результатах исследования сложных биологических смесей, способны образовывать пары с несколькими пептидами, что усложняет процесс идентификации.

С каждым годом для обработки масс-спектрометрических данных появляется все больше коммерческих и общедоступных поисковых алгоритмов, призванных сделать протеомный анализ достовернее, быстрее и проще.

Наиболее широко распространенная стратегия идентификации белковых молекул состоит в сопоставлении определенных экспериментально масс-зарядных характеристик фрагментов белковых молекул с референсными значениями из баз данных [116–118]. В основе работы таких поисковых машин лежит алгоритм сопоставления зафиксированных в ходе эксперимента спектров пептидных фрагментов со сгенерированными *in silico* спектрами пептидов, последовательности которых получены виртуальным протеолизом белковой библиотеки (рис. 5).

Несмотря на прогресс в развитии масс-спектрометрических технологий, в отсутствие стандартизованных вычислительных и статистических методов анализ полученных данных предоставляет значительную долю ложно-положительных результатов, а потому остается «ахиллесовой пятой» протеомики [119,120].

Каждый поисковый алгоритм уникален и обладает преимуществами и недостатками на фоне остальных алгоритмов. Поисковики, сопоставляющие

спектры с проанализированными спектрами ранее проведенных экспериментов, гораздо чувствительнее и специфичнее поисковых машин, сравнивающих экспериментальные данные с теоретическими, однако такие характеристики распространяются только на присутствующие в библиотеке пептиды.

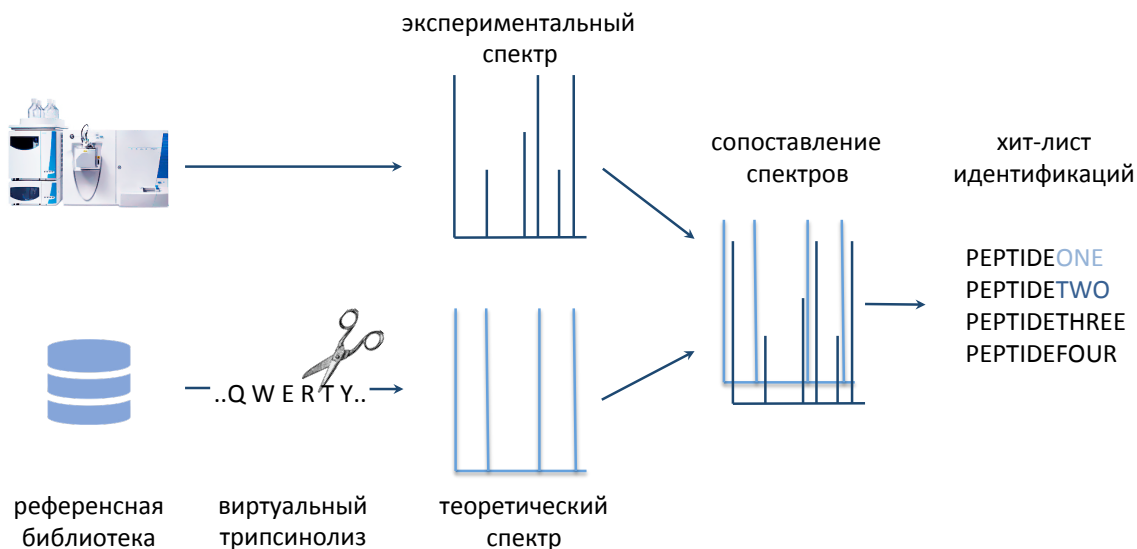


Рисунок 5. Принципиальная схема работы поисковых алгоритмов, использующих для идентификации пептидов и белков референсную библиотеку аминокислотных последовательностей.

Помимо идеологических различий, заложенных в алгоритм идентификации пептидов, поисковые алгоритмы различаются по набору доступных для варьирования параметров поиска, способу оценки достоверности результатов и производительности. Несмотря на то, что в научной литературе представлены работы по сравнению поисковых программ, использующих различные стратегии идентификации [121], не удастся сделать однозначный вывод о конкурентном преимуществе одного из алгоритмов, поскольку наблюдается значительное перекрытие между результатами, полученными с помощью различных поисковых машин [122].

Поскольку разные поисковые машины позволяют идентифицировать разные наборы пептидов, очевидным шагом была бы комбинация нескольких алгоритмов. Подход к принятию научного решения, основанный на обобщении выводов нескольких экспертных мнений или алгоритмов, полезен при решении научных вопросов, в которых отсутствует эталонная парадигма.

Комбинация популярных поисковых алгоритмов уже реализована в некоторых программных пакетах. К примеру, программа IPeak комбинирует

результаты мультипоточного процессирования масс-спектрометрических данных и использует алгоритм Percolator для повышения качества распознавания истинных и ложных identификаций [123]. Также популярен алгоритм машинного обучения PepArML [124], интегрирующий результаты поиска несколькими алгоритмами, среди которых Mascot, X!Tandem, OMSSA и др. PepArML использует собственные вычислительные кластеры для обработки данных, которые пользователи загружают через web-интерфейс. Достоинство PepArML состоит в том, что сложные вычислительные процессы по интерпретации экспериментальных результатов производятся с привлечением внешних вычислительных ресурсов. Тем не менее, интерпретация данных с использованием PepArML не отличается экспрессностью, поскольку требует дополнительного взаимодействия пользователя с командой администраторов по ряду вопросов, включающих, например, загрузку оптимизированной базы данных.

Одним из наиболее удобных в эксплуатации интеграторов, позволяющих интерпретировать результаты поиска нескольких алгоритмов, является PeptideShaker, встроенный в графический интерфейс SearchGUI [125,126]. PeptideShaker обеспечивает переход от необработанных масс-спектров к наглядной интерпретации результатов, учитывающей особенности стадий пробоподготовки и самого масс-спектрометрического измерения. Эта платформа адаптирована к файлам выдачи поисковых алгоритмов X!Tandem, MS-GF+, MS Amanda, MyriMatch, Comet, Tide, Andromeda, OMSSA, а также *de novo* стратегий Novor и DirecTag. Принципиальным отличием PeptideShaker от аналогичных инструментов, интерпретирующих и интегрирующих файлы масс-спектров, является его адаптированность к аппаратному обеспечению и программному окружению стандартного компьютера.

При идентификации пептидов и белков в PeptideShaker используется стратегия истинных и ложных identификаций (англ. target-decoy approach, TDA) за счет оценки уровня ложноположительных identификаций (англ. false discovery rate, FDR) и апостериорной вероятности ошибки (англ. posterior error probability, PEP) результатов нескольких поисковых алгоритмов [122]. Обе статистические оценки, PEP и FDR, комплементарно дополняют друг друга, по сути являясь

сторонами одной медали: PER характеризует качество совпадения пептида и спектра, а FDR – качество группы таких совпадений (рис. 6).

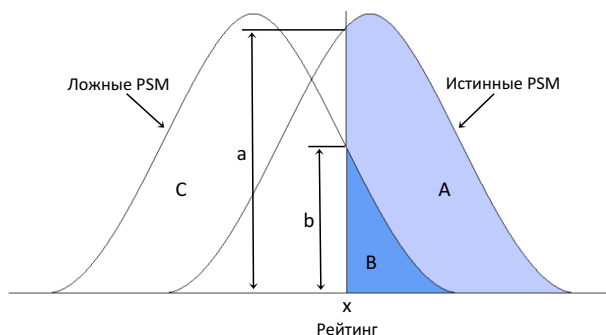


Рисунок 6. Величина FDR – это отношение величины В (количества ложных идентификаций, преодолевших некоторый уровень отсечения x) к сумме $A+B$ (суммарному количеству PSM, т.е. соответствующих спектрам пептидов, преодолевших x). Величина PER, в свою очередь, является отношением соответствующих высот областей распределения: b , характеризующего идентификации с рейтингом x , к $a+b$, то есть количеству идентификаций с рейтингом x .

Взвешенная комбинация нескольких поисковых алгоритмов позволяет получить более полные и достоверные сведения об исследуемом объекте, однако влечет за собой значительное увеличение вычислительного времени. С демократизацией вычислительных кластеров и облачных ресурсов следует ожидать популяризации использования нескольких поисковых алгоритмов для интерпретации протеомных данных, однако выбор сочетания поисковиков в условиях ограниченных вычислительных ресурсов актуален уже сейчас. Как показывает практика, оптимальное сочетание поисковых алгоритмов варьируется от задачи к задаче и во многом определяется параметрами оборудования, используемого для анализа, настройками поиска с особым акцентом на мониторинг посттрансляционных модификаций, а также референсной библиотекой пептидных последовательностей.

1.3. Ресурсы, позволяющие оценить разнообразие протеоформ

1.3.1. Геномные и транскриптомные ресурсы

Технологии секвенирования нового поколения предоставляют информацию о нуклеотидных последовательностях с увеличивающейся производительностью и снижающимися затратами. Архивирование экспериментальных результатов обеспечивает задел для развития воспроизводимых методов, а потому необходимость создания геномных и транскриптомных репозиториях приобретает

все большую значимость, в том числе и для идентификации новых, ранее нигде не аннотированных протеоформ.

1000 Genomes Project. Выполнение проекта 1000 Genomes (1KG), направленного на идентификацию и аннотацию генетических полиморфизмов у людей разных рас, позволило создать обширный каталог генетической variability. Глубина секвенирования, выбранная в 1KG, не позволяет зафиксировать все вариации, но ее достаточно для детекции наиболее представленных вариантов с частотой встречаемости $\geq 1\%$. Секвенирование геномов более 2 тыс. человек выявило свыше 80 миллионов нсОМП.

За девять лет выполнения проекта 1KG появлялись аналогичные полногеномные исследования с фокусом на медицинские задачи, причем некоторые из них превышали 1KG по масштабу [127]. Однако в силу ряда причин именно 1KG остается приоритетным источником сведений о генетической variability. Во-первых, образцы, исследованные в 1KG, обеспечивают более полную картину генетических вариантов по сравнению с основной массой крупных проектов, обычно сфокусированных на конкретной патологии и расе. Во-вторых, 1KG объединяет несколько стратегий обработки получаемой генетической информации: такой более трудоемкий подход позволяет повысить качество результатов. Наконец, в-третьих, доступность результатов проекта и самих исследованных образцов обеспечивает возможность их многократного анализа.

Ensembl. Создание геномной базы Ensembl, приуроченное к окончанию проекта «Геном человека», акцентировало тот факт, что интерпретация генетического кода не менее важна, чем его прочтение [128]. Платформа Ensembl ориентирована на автоматизированную аннотацию геномов, т.к. обработка всех последовательностей генома в ручном режиме не представляется возможной.

Ensembl позволяет визуализировать выравнивание нуклеотидных последовательностей человека и других организмов и сопоставить их с референсными геномами для оптимизации поиска к конкретным научным задачам. Преимуществом Ensembl по сравнению с другими геномными базами данных является доступ к интегрированным сведениям об нуклеотидной последовательности гена и соответствующих ей транскриптах и протеоформах.

Несмотря на то, что использование геномных баз для исследования гетерогенности протеома имеет ряд ограничений (например, несовершенство автоматической аннотации прочтений), по прошествии почти 20 лет Ensembl остается одним из центральных ресурсов для анализа нуклеотидных и аминокислотных последовательностей.

SRA. Целью создания архива SRA (Sequence Read Archive) стало предоставление исследователям доступа к необработанным результатам секвенирования нуклеотидных последовательностей различных организмов. Этот архив является ядром инфраструктуры, хранящей геномные и транскриптомные данные до момента их публикации. Балансируя между необходимостью снизить затраты на долговременное хранение информации и потребностью в репозитории, позволяющем вернуться к повторному анализу данных, SRA стал одним из мощнейших ресурсов, предоставляющим доступ к 10 Пб геномных и транскриптомных данных для различных органов и тканей [129].

ClinVar. Ресурс ClinVar содержит более 100 тыс. записей о фенотипах и генетических вариациях генеративной и соматической природы, ассоциированных с заболеваниями. Каждая запись в репозитории содержит данные о клинической значимости конкретного варианта. Ресурс интегрирован в платформу ClinGen, позволяющую повысить качество интерпретации посредством распространения данных и экспертной оценки. Достоверность и клиническая значимость каждого варианта зависит от степени подтверждения, предоставляемого исследователем в виде публикаций или результатов клинических исследований [130].

The Cancer Genome Atlas. The Cancer Genome Atlas (TCGA) – это результат коллаборации Национального центра изучения рака США, Национального института исследования генома США и еще 18 организаций, нацеленной на создание обширных многомерных карт геномных вариаций. Общедоступный репозиторий TCGA к настоящему моменту содержит более 2 Пб геномных данных, которые характеризуют более 30 типов онкологических заболеваний.

Сведения, представленные в TCGA, используются для определения молекулярных основ онкозаболеваний и усовершенствования терапевтических стратегий: так, один из проектов TCGA позволил выявить на молекулярном уровне

подтип рака молочной железы, сходный с раком яичников по механизму развития и реакции на лекарственные препараты [131].

COSMIC. Одной из причин возникновения онкологических заболеваний является накопление на протяжении всей жизни соматических мутаций. Репозиторий COSMIC был запущен в 2004 г. и на момент открытия содержал информацию о соматических мутациях лишь в 4 генах человека. По прошествии лет этот экспертируемый ресурс приобрел важнейшее значение для биохимиков и онкологов. В репозитории приведены данные о 4 млн случаях редактирования генома в кодирующих областях, выявленных в 30 тыс. образцах опухолей. Помимо однонуклеотидных замен, вставок и делеций, COSMIC содержит сведения о химерных генах, генных перестройках и химических модификациях нуклеотидов.

OMIM. База OMIM (Online Mendelian Inheritance in Man) содержит ежедневно обновляемые сведения обо всех известных менделевских заболеваниях и более чем 15 тыс. генов, изменения в которых ассоциированы с почти 400 патологиями [132]. Источниками данных базы являются научные статьи: ресурс проводит поиск информации о связи между геномом и фенотипом по базе научной литературы PubMed.

GTEx. За последние 10 лет было проведено более 2 тыс. полногеномных исследований по аннотации вариантов в геноме человека в норме и при патологии. Прогресс полногеномных методов позволил идентифицировать ассоциированные с болезнями генетические вариации, однако большинство из них расположено вне БКГ, что затрудняет диагностику и понимание механизмов развития патологии. Для определения влияния генетической вариации на экспрессию гена платформа GTEx (Genotype-Tissue Expression) использует анализ экспрессии локусов количественных признаков eQTLs [133]. Технология eQTLs призвана сократить информационный пробел между редактированием нуклеотидной последовательности и болезнью.

Помимо обширной базы данных, включающей сведения о редактировании нуклеотидной последовательности в более чем 8 тыс. образцов 53 типов, платформа GTEx располагает биобанком, содержащим генотипированные биообразцы от 900 доноров.

dbSNP. Архив полиморфизмов единичных нуклеотидов *dbSNP* предоставляет доступ к сведениям о генетических вариациях. Несмотря на то, что в название *dbSNP* вынесен только полиморфизм отдельных нуклеотидов, ресурс располагает данными о делециях, вставках и других мутациях [134]. В *dbSNP* описана подавляющая часть из существующих на сегодняшний день аллелей генома человека и представлено значительное число мутаций, наблюдаемых в различных клеточных линиях. Ресурс объединяет информацию о полиморфизмах, полученную из разных источников: исследовательских лабораторий, масштабных проектов по поиску полиморфизмов и других баз данных.

При использовании сведений из базы *dbSNP* (равно как и данных из других геномных репозиториях) следует учитывать высокую долю ложноположительных identifications – на уровне 15-17% [135].

1.3.2. Протеомные ресурсы

Успех идентификации белков в условиях наиболее популярного подхода «bottom-up» определяется используемой для поиска референсной библиотекой аминокислотных последовательностей. Библиотека формируется на основе результатов собственных экспериментов или данных из web-ресурсов, которые различаются по полноте представляемой информации, избыточности данных и качеству аннотации последовательностей.

UniProtKB. База знаний *UniProtKB* – это наиболее обширная протеомная платформа, интегрирующая в себе сведения из множества ресурсов. К 2017 году количество аминокислотных последовательностей в *UniProtKB* превысило 75 млн и продолжает удваиваться с каждым годом.

Аминокислотные последовательности, содержащиеся в разделе *UniProt/Swiss-Prot*, получены посредством автоматической трансляции кодирующих последовательностей нуклеотидов в белки с последующей экспертной проверкой. Раздел *UniProt/Swiss-Prot* предоставляет данные с минимальной избыточностью: экспертная аннотация каждой записи подразумевает тщательный анализ экспериментально обоснованных или *in silico* спрогнозированных последовательностей протеоформ одного гена (включая сплайс-варианты, замены отдельных аминокислот и посттрансляционные

модификации), упоминаний о них в литературных источниках, поиска гомологов, а также оценки уровня достоверности идентификации и контроля качества проведенной аннотации.

Поток экспериментальных данных, накапливаемых в ходе выполнения крупных геномных проектов, спровоцировал создание автоматически аннотируемой библиотеки данных нуклеотидных последовательностей TrEMBL как дополнения базы Swiss-Prot, возможностей которой было недостаточно для трудоемкого процесса экспертной проверки всех поступающих сведений. Превалирующее большинство (> 95%) записей в UniProt приходится на раздел TrEMBL и представляет результат автоматизированной трансляции БКГ, содержащихся в базах данных нуклеотидных последовательностей EMBL-Bank/GenBank/DDBJ. Помимо ресурсов Swiss-Prot и TrEMBL, UniProtKB предоставляет доступ к архиву неаннотированных белков UniParc и кластерам UniRef, объединяющим родственные аннотированные и неаннотированные записи о белковых последовательностях [136].

Главным преимуществом ресурса UniProt является стабильное качество аннотации данных, покрывающих широкое таксономическое разнообразие.

neXtProt. На основе базы Swiss-Prot и результатов крупномасштабных протеомных проектов была создана онлайн платформа neXtProt, содержащая сведения о белках человека, их функциях, структуре, взаимодействиях с другими белками, внутриклеточной локализации, ОАП и роли в развитии заболевания [137]. Записи в neXtProt подвергаются экспертной оценке, по итогу которой последовательности присваивается статус «gold», «silver» или «bronze», который соответствует уровню достоверности зафиксированного белка.

В каждую запись платформы интегрированы кросс-ссылки на страницы в других ресурсах, среди которых Gene Ontology, Entrez Gene, MINT, ENZYME, PDB, и статьи в библиотеке PubMed.

Human Protein Atlas. Основной базой данных, хранящей результаты ИФА белков в нормальных и опухолевых тканях человека и клеточных линиях, является портал Human Protein Atlas (HPA) [138]. Кластер HPA подразделяется на 3 раздела:

1. The Tissue Atlas – сведения о профилях экспрессии свыше 15 тыс. генов человека на транскриптомном и протеомном уровнях;

2. The Cell Atlas – информация о пространственном распределении белков и экспрессии РНК свыше 12 тыс. генов в панели из 56 клеточных линий человека;

3. The Cancer Atlas – данные об экспрессии генов в более чем 200 опухолевых образцах 20 распространенных онкозаболеваний.

Peptide Atlas/PASSEL. Каталог PeptideAtlas охватывает обработанные и валидированные стандартизованным образом результаты более тысячи масс-спектрометрических экспериментов, которые исследователи могут загружать в каталог с помощью компонента PASSEL. Библиотека PASSEL, содержащая результаты свыше 100 протеомных экспериментов по направленной масс-спектрометрии, позволяет распространять и повторно анализировать загруженные данные [139]. PeptideAtlas и PASSEL активно применяются не только для перекрестного анализа протеомных данных, но и для планирования масс-спектрометрических экспериментов.

PRIDE. База данных PRIDE (PRoteomics IDentifications Archive) - это ресурс, содержащий результаты масс-спектрометрических экспериментов в тканях более 60 биологических видов [140]. PRIDE участвует в консорциуме ProteomeXchange и ежемесячно принимает около 150 пакетов результатов панорамных масс-спектрометрических экспериментов. К настоящему моменту архив содержит около 700 млн спектров. Архив хранит не только результаты детекции пептидов и белков, но и сведения об их содержании в исследуемых образцах, сами масс-спектры, изображения 2DE-гелей и другие сведения, предоставленные исследователями.

GPMdb. База данных GPMdb (Global Proteome Machine database) – это архив сведений о белках и пептидах, которые были детектированы масс-спектрометрически. GPMdb позволяет сопоставлять собственные данные с результатами других исследовательских групп с помощью поисковой машины GPM [141]. К 2018 году база GPMdb содержала данные о более чем 16 млн пептидов и 3 млн протеоформ 22 модельных организмов.

Базы данных протеоформ. При использовании референсной библиотеки, содержащей только канонические варианты аминокислотной последовательности, другие протеоформы естественным образом не могут быть идентифицированы [142]. Один из наиболее популярных сценариев при выборе референсной базы

аминокислотных последовательностей состоит в обогащении библиотеки канонических протеоформ сведениями из ресурсов, описывающих варианты изменений белкового сиквенса.

Базы данных об одноаминокислотных полиморфизмах

dbSAP. База dbSAP является одним из немногих поддерживаемых узкоспециализированных репозиториях сведений о заменах нуклеотидов или аминокислот. База содержит сведения о более чем 16 тыс. уникальных пептидов с одноаминокислотными полиморфизмами, которые были детектированы масс-спектрометрически, и позволяет учитывать вклад ПТМ полиморфных остатков в сдвиг на спектре характеристичного замене пика [143].

MS-CanProVar. Итогом выполнения проекта CanProVar, нацеленного на аннотацию онкоассоциированных мутантных форм белков, стало создание библиотеки MS-CanProVar [144]. Библиотека хранит экспериментально подтвержденные сведения о детекции ОАП. Помимо экспериментальных данных, полученных в ходе выполнения проекта CanProVar, репозиторий содержит сведения из баз TCGA, COSMIC и OMIM: свыше 40 тыс. записей о неспецифичных ОАП и более 11 тыс. – о полиморфизмах, связанных с возникновением и развитием онкозаболеваний.

Другие узкоспециальные, ориентированные на каталогизацию ОАП репозитории, например, PMD (Protein Mutant Database [145]), как и MS-CanProVar послужили «плечами гигантов» для dbSAP, но не обновляются уже более 10 лет.

Базы данных о ПТМ. Развитие высокопроизводительных протеомных технологий коррелирует с увеличением числа и масштаба репозиториях, содержащих сведения о ПТМ белков различных организмов и при различных состояниях [146]. Некоторые репозитории сфокусированы на накоплении сведений о различных типах ПТМ, другие содержат информацию о конкретной модификации.

Наиболее мощным ресурсом, предоставляющим информацию о ПТМ различных типов, является dbPTM [147]. В этом ресурсе собраны сведения об экспериментально валидированных и предсказанных ПТМ из узкоспециальных источников. Репозиторий содержит информацию о 600 тыс. модифицированных

пептидов и влиянии модификации на физико-химические свойства и пространственную структуру белков.

Среди репозиториев, ориентированных на исследование конкретной модификации, особого упоминания заслуживают платформы, хранящие сведения о фосфорилировании. Фосфорилирование – часто детектируемая и хорошо изученная модификация. По прогнозам, протеом человека содержит более полумиллиона сайтов фосфорилирования [148]. Крупнейшими архивами в области фосфопротеомики являются Phospho.ELM [149] и PhosphoSitePlus [150].

Базы данных о сплайс-протеоформах. За последние 15 лет было запущено несколько платформ, ориентированных на хранение сведений о реализации альтернативного сплайсинга на протеомном уровне (в их числе FAST-DB [151], ProSAS [152] и др.). К сожалению, несмотря на актуальность исследования альтернативного сплайсинга как одного из ключевых источников гетерогенности протеома, большинство перечисленных репозиториев перестали существовать как отдельные информационные единицы и были инкорпорированы в более крупные и общие ресурсы, такие как Ensembl и UniProt.

Транскриптомные данные для создания референсной библиотеки. Секвенирование нуклеотидных последовательностей позволяет сформировать геномный или транскриптомный профиль исследуемого образца и создать на его основе специфичную для конкретного образца референсную библиотеку, учитывающую мутации и сплайс-формы [153].

Одна из стратегий создания библиотеки учитывает все возможные варианты сдвига открытой рамки считывания. Такой подход, зарекомендовавший себя при работе с геномами прокариотов, не подверженными альтернативному сплайсингу, позволяет охватить все вариации аминокислотной последовательности [154]. Геном человека на 98% состоит из не кодирующих белки генов, потому применение этого подхода для исследования гетерогенности протеома существенно увеличило бы поисковое пространство, которое не распространяется на продукты АС и ОНП. Более того, поскольку из шести возможных нуклеотидных

последовательностей только одна истинна, не удастся избежать повышения уровня ложных идентификаций [155].

Референсная библиотека меньшего размера может быть создана на основе транскриптомных данных (рис. 7).

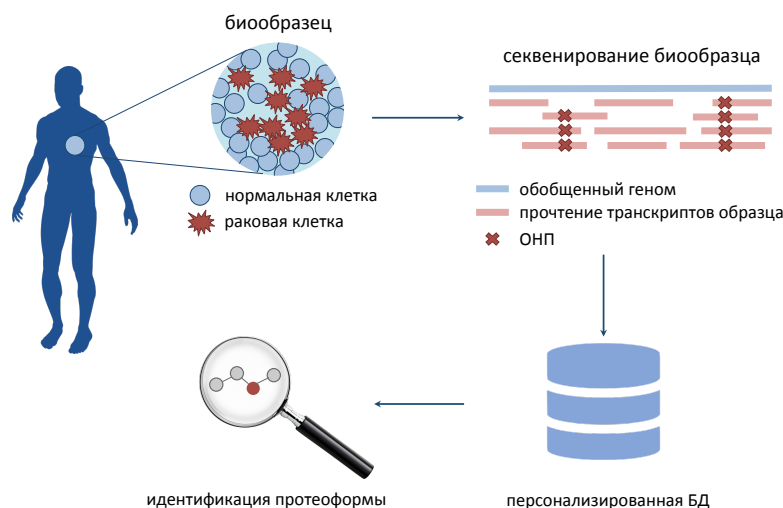


Рисунок 7. Создание и использование персонализированной референсной библиотеки, основанной на транскриптомных данных: пример поиска онкоассоциированных протеоформ.

Такой подход к исследованиям протеома человека снижает размер поискового пространства более чем в 30 раз по сравнению с “шестирамочной” стратегией [156]. Специфичная библиотека не ограничена обобщенными геномными или транскриптомными данными, что открывает возможности для обнаружения новых протеоформ [157]. Пример транскриптомно-протеомного исследования двух клеточных линий колоректального рака, описанного в [158], демонстрирует эффективность использования транскриптом-специфичной библиотеки: мультиомный подход позволил детектировать в 1,5 раза больше пептидов по сравнению со стандартным поиском по библиотеке Ensembl.

Популярным объектом применения омиксных технологий являются клеточные линии. В данной работе был проанализирован протеом опухолевой клеточной линии HepG2. Эта клеточная линия заслуживает особого внимания, поскольку до сих пор ведутся дискуссии по поводу типа первичной опухоли (гепатобластома или гепатоцеллюлярная карцинома), из которой произошла линия HepG2. Особенности опухолевой линии HepG2 подробно изложены при обсуждении результатов работы.

ГЛАВА 2. МАТЕРИАЛЫ И МЕТОДЫ

2.1. Транскриптомные данные: источники и метод обработки

В работе был применен транскриптно-протеомный подход к поиску протеоформ, который состоит в использовании оптимизированной референсной библиотеки аминокислотных последовательностей, сгенерированной на основе результатов высокопроизводительного секвенирования мРНК исследуемого образца – в нашем случае, клеток опухолевой линии HepG2.

Для составления такой библиотеки были использованы данные высокопроизводительного секвенирования транскриптома клеток линии HepG2, которые были независимо получены на двух платформах (Illumina Genome HiSeq 2000 и Applied Biosystems SOLiD 4) в трех технических повторениях (ID экспериментов в NCBI: SRX395473 и SRX390071). Результаты секвенирования мРНК на платформе Illumina содержали 41, 43 и 45 млн прочтений в каждом из технических повторений. На платформе SOLiD, обеспечивающей двойное прочтение каждого нуклеотида последовательности, так же, в трех технических повторениях, было зафиксировано, соответственно, 142, 149 и 198 млн прочтений последовательности нуклеотидов. Полученные на двух платформах данные были объединены на основании ранее показанной высокой корреляции экспрессии генов, детектируемых с использованием разных платформ [159]. Дальнейшей обработке подвергали транскрипты со статусом «Reviewed» в протеомном ресурсе UniProt при условии, что среднее значение программно-рассчитываемого показателя экспрессии транскрипта (англ. fragments per kilobase of exon per million reads mapped, FPKM) превышало 0,1. Среднее значение FPKM рассчитывали по данным трех технических повторов, как минимум в двух из которых коэффициент корреляции не превышал 30%.

Обработка транскриптомных данных была произведена с помощью программного продукта PPLine [160]. Схема автоматизации процесса исследования постгеномных данных на предмет выявления транскрибируемых протеоформ с помощью PPLine представлена на рисунке 8.

Предварительный этап обработки данных секвенирования мРНК включал удаление областей с низким значением параметра Phred Quality ($Q < 25$), характеризующего качество прочтения нуклеотидной последовательности. После

удаления участков с низким качеством (около 6% данных, полученных на платформе Illumina, и 9% результатов SOLiD-секвенирования) из всего массива полученных результатов для дальнейшего анализа были отсортированы прочтения длиной минимум 40 пар нуклеотидов.

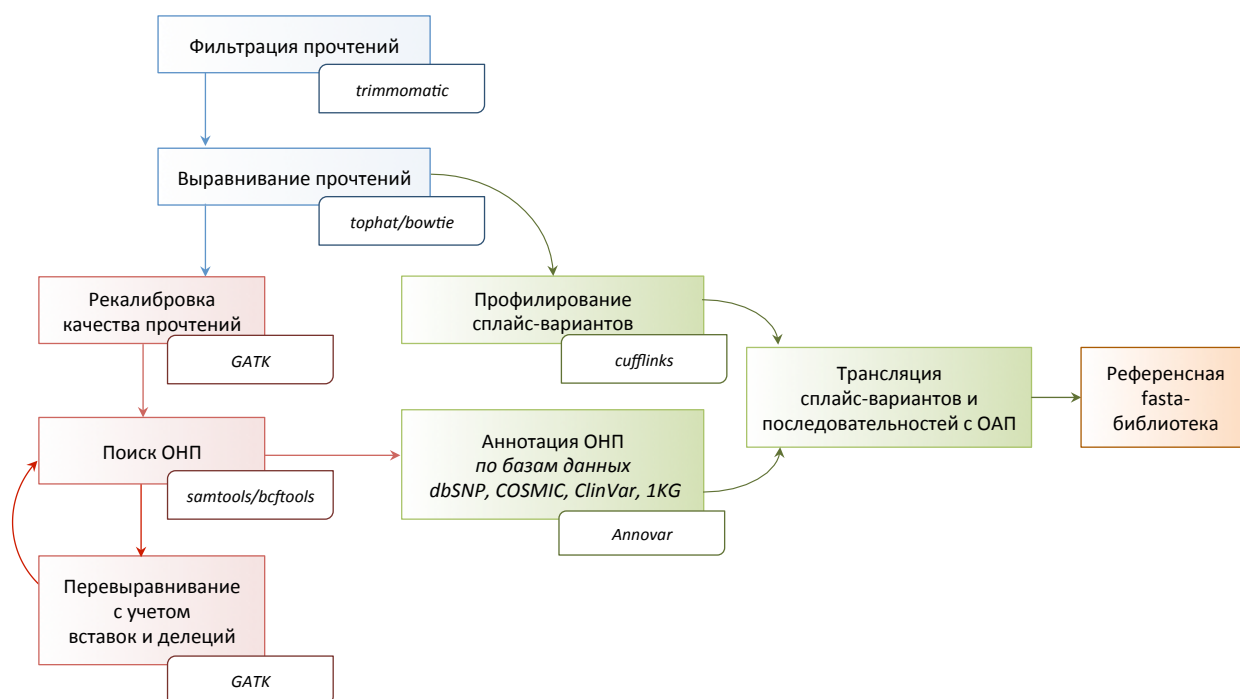


Рисунок 8. Последовательность операций, реализованных в программном продукте PPLine для обработки данных секвенирования РНК и интеграции полученных после их обработки результатов в единую референсную библиотеку для дальнейшего протеомного анализа [160].

Картирование объединенного по всем техническим повторениям пула прочтений осуществлялось с помощью алгоритма TopHat. После картирования были определены кандидатные положения с заменами нуклеотидов. Допустимое число несоответствий нуклеотидов было определено как 2 для Illumina и 3 для SOLiD. В качестве референсного генома были использованы координаты экзонов в генах, их комбинации, составляющие различные транскрипты, а также координаты кодирующих областей геномной модели Ensembl (сборка GRCh38, p.6).

Для контроля качества картирования прочтений на референсный геном с фокусом на обнаружение полиморфных нуклеотидов была проведена рекалибровка (BQSR) [161]. Это позволило исключить возможные ошибки, обусловленные

дизайном эксперимента или технической спецификой конкретной методики секвенирования.

Процесс рекалибровки (рис. 8) заключался в сопоставлении необработанных результатов секвенирования с известными последовательностями, хранящимися в базах нуклеотидных замен, в частности dbSNP. Совпадение однонуклеотидной замены, выявленной в собственных результатах, с валидированными сведениями из репозитория позволило сформировать две тестовые группы. Первая группа состоит из последовательностей, дополнительно валидированных с помощью тематических репозиториях, в которой, предположительно, будут находиться действительные мутации. Вторая группа включала все прочие случаи и ожидаемо содержала значительное количество ошибок. Сравнение двух групп между собой позволило выявить достоверные различия в особенностях прочтений полиморфных последовательностей, определить подверженные ошибкам прочтения нуклеотиды и провести корректировку. Рекалибровка позволила существенно улучшить результаты поиска мутаций: доля замен, аннотированных в базе dbSNP, включающей в себя данные проекта 1000 Genomes и Exome Sequencing Project, возросла на 20% по сравнению с первичным поиском.

Картирование объединенного по техническим повторам пула прочтений осуществлялось с помощью алгоритма TopHat. Для идентификации сплайс-транскриптов, не аннотированных в базах Ensembl и RefSeq, было использовано приложение Cufflinks (<http://cole-trapnell-lab.github.io/cufflinks>), ориентированное на сборку полных транскриптов. Алгоритм, включающий стадии предварительной обработки прочтений, их картирование на геном, поиск и аннотацию полиморфных вариантов, представлен в работе [160].

2.2. Протеомные данные

Протеомный анализ был выполнен для приблизительно 10 млн клеток опухолевой линии HerG2, образец которой был использован для проведения транскриптомного анализа. Общее содержание белка в исследуемом образце клеток было оценено по методу Брэдфорда и составило около 2 мг.

Фракционирование белковой смеси с помощью двумерного геле-электрофореза в полиакриламидном геле было проведено научной группой под руководством д.б.н. С.Н. Нарыжного по стандартному протоколу Клозе и О'Фаррелла [162,163] с использованием иммобилизованного градиента pH на приборах Ettan TM IPGphor3 и Ettan TM DALTsix (GE Healthcare). Гель, содержащий фракционированную смесь белков линии HepG2, разрезали на 96 ячеек с координатами по молекулярному весу (MW) и изоэлектрической точке (pI). Детали эксперимента представлены в работах [12,164].

Каждый фрагмент геля (~0,7 см²) был измельчен, обработан трипсином и подвергнут хромато-масс-спектрометрическому анализу в панорамном режиме на приборе ESI LC-MS/MS (LTQ-Orbitrap Q-Exactive, Thermo) в двух технических повторениях. Методические особенности проведения масс-спектрометрического эксперимента изложены в работах [11,159,165].

Общий объем 192-х необработанных файлов масс-спектров в формате raw превышал 114 Гб. После перекодирования в формат mgf (текстовый формат хранения сведений о массах и зарядах пептидных фрагментов и времени удержания) файлы анализировали с помощью пакета SearchGUI (версия 3.2.20), установленного на ЭВМ с операционной системой macOS Sierra (CPU 3,5 GHz, RAM 8 Gb) [126].

Карбамидометилирование цистеина было указано в качестве фиксированной химической модификации, окисление метионина – в качестве вариабельной. Допустимая погрешность (mass tolerance) для масс-зарядных характеристик родительских и дочерних ионов была установлена на уровне 20 ppm и 0,01 Da, соответственно. При интерпретации файлов масс-спектров допускали возможность пропуска одного сайта трипсинолиза. В соответствии с международными критериями [13] использовали результаты идентификации, удовлетворяющие уровню показателя достоверности $FDR \leq 1\%$.

В качестве референсной базы данных, по которой производили биоинформатический поиск, использовали транскриптом-специфичную библиотеку, созданную на основе транскриптомного секвенирования и расширенную с учетом заведомо ложных обращенных последовательностей (decoy) и загрязняющих белков-контаминантов из базы cRAP [166].

Оптимизация параметров анализа файлов масс-спектров

Графический интерфейс SearchGUI, использованный для анализа экспериментальных результатов, применяли, чтобы перейти от необработанных масс-спектров к наглядной интерпретации результатов, учитывающей особенности стадий пробоподготовки, ионизации, фрагментации и масс-спектрометрического измерения. Платформа SearchGUI включает поисковые алгоритмы X!Tandem, MS-GF+, MS Amanda, MyriMatch, Comet, Tide, Andromeda, OMSSA, а также *de novo* стратегии Novor и DirecTag. SearchGUI может использоваться для комбинирования результатов интерпретации масс-спектров, полученных с помощью нескольких поисковых алгоритмов.

Для определения оптимального сочетания поисковых алгоритмов при решении базовой протеомной задачи – протеотипировании клеточной линии – исследовали влияние количества используемых поисковых машин, объема референсной базы данных и параметров поиска на производительность анализа. Выбор сочетания поисковых алгоритмов валидировали с использованием результатов панорамного масс-спектрометрического исследования стандартной смеси UPS2 (Merck, Германия), состоящей из 48 белков в динамическом диапазоне концентраций, охватывающем 5 порядков (500 амоль – 50 пмоль).

Сведения из информационных ресурсов

В качестве дополнительных источников сведений о гетерогенности протеома использовали различные геномные, транскриптомные и протеомные репозитории, а также биоинформатические инструменты для анализа аминокислотных последовательностей. Перечень информационных ресурсов и инструментов, использованных в работе, приведен в таблице 1.

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

№	Название ресурса и электронный адрес	Тип предоставляемых сведений
1	UniProtKB, 2017_04 www.uniprot.org	Сведения об аминокислотной последовательности протеоформ, являющихся результатом АС, ОНП и ПТМ, уровне доказательства существования и их ассоциации с патологией

2	PeptideAtlas, 2017_01 www.peptideatlas.org	Обработанные результаты более тысячи экспериментов по направленной масс-спектрометрии
3	PRIDE, 2017_12 www.ebi.ac.uk/pride/archive/	Результаты панорамного масс-спектрометрического анализа белков и пептидов
4	GPMdb, 2017_10 www.gpmdb.thegpm.org	Сведения о пептидах и их модификациях, детектированных в протеомных экспериментах 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 в 26 исследуемых популяциях
7	dbSNP, v. 151 www.ncbi.nlm.nih.gov/SNP/	Сведения об обнаруженных на уровне генома ОНП, для которых не показана взаимосвязь с развитием патологий
8	iPTMnet, v. 4.1. www.research.bioinformatics.ucdel.edu/iptmnet/	Информация о предсказанных и экспериментально валидированных ПТМ
9	MalaCards, v.4.6.0.11 www.malacards.org	Информация о взаимосвязи гена с развитием патологии
10	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
11	The Cancer Genome Atlas, 2017 www.cancergenome.nih.gov	Систематизированные данные о генетических мутациях, ассоциированных с 20 онкологическими заболеваниями
12	TumorPortal, 2017 www.tumorportal.org	Сведения о точечных соматических мутациях генов в опухолевых и нормальных тканях
13	ClinVar, 2017_05 www.ncbi.nlm.nih.gov/clinvar/	Репозиторий данных об ассоциации генетической вариации и конкретной патологии
14	OMIM, 2017_11 www.omim.org	Сведения о связи свыше 15 тыс. генов человека с 380 патологиями
15	ExPASy Compute pI/MW, 2017_12 www.web.expasy.org/compute_pi/	Теоретические значения изоэлектрической точки и молекулярной массы, рассчитываемые для аминокислотных последовательностей
16	IPC Isoelectric Point Calculator, 2016_10 www.isoelectric.ovh.org	Теоретические значения изоэлектрической точки полипептидов
17	GOrilla http://cbl-gorilla.cs.technion.ac.il	Отранжированный и визуализированный список генов/белков с обогащенными терминами Gene Ontology (функция, участие в биологическом процессе и локализация белка)

2.3. Анализ результатов масс-спектрометрического исследования протеома клеточной линии HepG2

По результатам анализа масс-спектрометрического исследования белков опухолевой клеточной линии HepG2 в программе PeptideShaker сгенерировано 192 отчета – по отчету на каждое из двух технических повторений эксперимента для всех 96 ячеек геля. Отчет представлял собой таблицу в формате xls и содержал столбцы, описание которых представлено в таблице 2.

Таблица 2. Сведения, представляемые в файле выдачи по результатам интерпретации масс-спектрометрических данных

№	Название столбца	Тип сведений в данном столбце
1	Protein(s)	Список идентификаторов белков в формате AC UniProt, на аминокислотные последовательности которых может быть картирован детектированный пептид
2	Protein Inference	Степень уникальности детектированного пептида по отношению к белку: принадлежность детектированного пептида к единственной протеоформе или к нескольким белковым продуктам одного гена/нескольких родственных генов
3	Sequence	Аминокислотная последовательность детектированного пептида
4	#Missed Cleavages	Количество пропущенных сайтов трипсинолиза
5	Modified Sequence	Аминокислотная последовательность детектированного пептида, учитывающая ПТМ
6	#PSM	Число совпадающих пар теоретического и экспериментального спектров (peptide-spectra match)
7	Proteoform Identified	Протеоформа или группа протеоформ одного гена, присутствие которых в исследуемом образце можно предполагать по детектированному пептиду
8	Comment	Сведения о типе детектированной протеоформы/группы протеоформ и однозначности детекции

Каждый отчет был в автоматическом режиме проанализирован согласно алгоритму, приведенному на рисунке 9. Используемый для анализа алгоритм реализован с помощью интерпретатора Python (версия 2.7.7).

Использованный алгоритм состоял из трех частей. Первая часть обработки отчетных данных заключалась в предварительной фильтрации полученных сведений – удалении из поля рассмотрения нетриптических пептидов. Пептиды, картированные на единственную протеоформу, также подвергались дополнительной проверке на уникальность относительно всех белков человека в базе UniProt. Второй этап подразумевал картирование пептидов на

соответствующие протеоформы. Заключительной стадией алгоритма являлось определение типа детектированной протеоформы и класса однозначности идентификации.

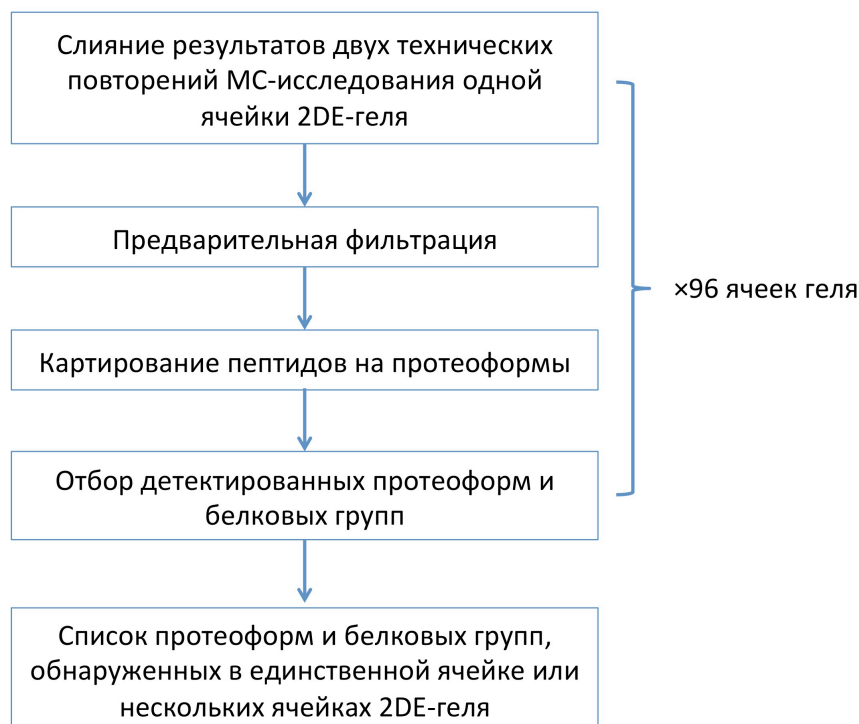


Рисунок 9. Основные стадии анализа результатов масс-спектрометрического исследования, направленного на выявление протеоформ и белковых групп.

В своем исследовании мы выделяли случаи однозначной детекции протеоформ [96] и неоднозначной детекции т.н. «мастерной» формы, когда обнаруженные пептиды позволяют констатировать лишь факт продукции гена без конкретизации возможного события.

Если протеоформе (каноническому варианту, СО-варианту, последовательности с ОАП или комбинации этих событий) соответствовало не менее двух пептидов, хотя бы один из которых являлся протеотипическим, то есть картирующимся исключительно на один белковый продукт, идентификацию считали однозначной. Прочие случаи, когда пептиды картировались одновременно на несколько протеоформ одного гена или белковых продуктов нескольких генов, классифицировали как неоднозначные. В результатах поиска игнорировали протеоформы, содержащие одноаминокислотные замены, трудноразличимые в результатах масс-спектрометрических экспериментов (I/L, Q/E, Q/K, G/N, F/M, N/D, M/T).

Для конкретизации детектированных протеоформ и белковых групп использовали результат фракционирования протеома опухолевой линии HepG2 с помощью 2DE. На двумерных электрофореграммах протеоформы одного гена могут иметь разную локализацию, определяемую их физико-химическими характеристиками – pI и MW , которые зависят от аминокислотной последовательности и наличия ПТМ. Для каждой детектированной протеоформы или нескольких протеоформ в рамках белковой группы сопоставляли экспериментальные и теоретически рассчитанные с помощью пакета Pyteomics [167], реализованного для Python (версия 2.7.7), координаты локализации протеоформы на геле. Несовпадение теоретических и экспериментальных координат локализации было поводом для дополнительного выяснения причин сдвига: рассматривали смещение на геле из-за альтернативного сплайсинга, делеций, вставок, деградации и разрывов полипептидных цепей, химических и посттрансляционных модификаций, изомеризации, а также совместной миграции комплекса из нескольких белковых партнеров.

2.4. Функциональное описание протеома клеточной линии HepG2

Для описания белков, аннотированных в клеточной линии HepG2, их молекулярной функции и участия в биологических процессах был использован структурированный словарь Gene Ontology [168] и веб-сервис GOrilla (<http://cbl-gorilla.cs.technion.ac.il>). При аннотации обнаруженных белков в терминах Gene Ontology в качестве контрольной выборки использовали полный набор канонических форм белков человека (20 316 белков по данным Swiss-Prot). Отсечение по значимости (p -value) было установлено на уровне 10^{-9} .

Чтобы выявить белки клеточной линии HepG2, которые потенциально вовлечены в онкологические процессы, сопоставляли перечень выявленных в результате транскриптомно-протеомного анализа со списком из 320 ассоциированных с онкологическими заболеваниями белков согласно ресурсу TumorPortal [169].

ГЛАВА 3. РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

3.1. Транскрибируемые протеоформы клеточной линии HepG2

Корреляция между представленностью транскрипта и белка является предметом множества научных дискуссий [74,170]. Тем не менее, нет сомнений, что транскриптомные и протеомные данные комплементарны. Экспрессия транскрипта – это необходимое требование для трансляции белка, а потому детекция транскрипта в конкретном образце позволяет с большей (хоть и не абсолютной) вероятностью ожидать его обнаружения на белковом уровне в том же биоматериале.

Секвенирование РНК предоставляет уникальные сведения о транскриптоме конкретного образца, которые могут быть использованы для создания транскриптом-специфичной библиотеки аминокислотных последовательностей. В протеомном анализе формирование референсной библиотеки имеет стратегическое значение, поскольку включение аминокислотной последовательности в библиотеку является условием дальнейшей идентификации этой последовательности или ее части в масс-спектрометрическом эксперименте. Особую важность вопрос формирования библиотеки приобретает при поиске новых, ранее не идентифицированных протеоформ.

По результатам обработки транскриптомных данных для клеточной линии гепатобластомы HepG2 нами была сформирована библиотека в fasta-формате для анализа результатов масс-спектрометрических экспериментов. Транскриптом-специфичная библиотека состояла из 51 809 последовательностей, кодируемых 12 148 генами. Согласно транскриптомным данным, в среднем на один БКГ приходится 4,3 протеоформы. В созданной библиотеке 11 347 записей относится к каноническим вариантам белков, 21 925 последовательностей представлено сплайс-формами; 31 972 последовательности содержат полиморфные сайты (ОАП, вставки и делеции) как для канонических белковых продуктов, так и сплайс-форм.

Таблица 3 иллюстрирует результат сравнения транскриптом-специфичной библиотеки, сформированной по результатам секвенирования мРНК клеток линии HepG2, со стандартной библиотекой, загруженной из базы UniProt, которая помимо канонических последовательностей содержит сплайс-протеоформы.

Размеры собственной библиотеки белков HerG2 и стандартной библиотеки UniProt различаются между собой незначительно, однако собственная библиотека описывает результат экспрессии 12 тыс. генов, а стандартная библиотека охватывает весь геном, но при этом не учитывает ОАП. Учет ОАП, вставок и делеций выравнивает характер распределения БКГ по количеству транскрибируемых вариантов в транскриптом-специфичной библиотеке (рис. 10).

Таблица 3. Сопоставление собственной библиотеки со стандартной библиотекой референсных последовательностей, загружаемой из UniProt

Библиотека	Кол-во а.к. последовательностей	Кол-во генов	Частота встречаемости сплайс-вариантов (на 1 БКГ с АС)	Частота встречаемости ОАП (на 1 БКГ с нсОМП)
Транскрибируемые в HerG2 протеоформы	51 809	12 148	1,5	2,6
Протеоформы человека (UniProt)	51 885	20 197	2,6	-

* – библиотека всех возможных протеоформ, в т.ч. последовательностей с делециями и вставками

** – стандартная библиотека, состоящая из канонических форм и сплайс-форм человека

*** – популяционные несинонимичные однонуклеотидные полиморфизмы

Согласно собственным транскриптомным данным для клеточной линии HerG2, максимальной гетерогенностью обладает ген *PLEC* – ему может соответствовать 171 протеоформа. Этот ген кодирует белок плектин (Q15149), структурно связывающий между собой части цитоскелета. Внушительная по длине белковая последовательность (4684 аминокислотных остатков) – это площадка для реализации различных комбинаций 9 сплайс-вариантов, 8 ОАП и 9 вставок и делеций.

По данным базы UniProt, больше всего – 37 – протеоформ можно ожидать в случае гена *CACIC*, кодирующего α -1-субъединицу потенциал-зависимого кальциевого канала (Q13936). Примечательно, что собственные исследования не выявили транскрипцию этого гена в клетках HerG2, а потому аминокислотные последовательности его протеоформ не были включены в референсную библиотеку.

Качество библиотеки определяется равновесием между стремлением включить все белковые последовательности, которые потенциально могут быть

обнаружены в исследуемом образце, и стремлением максимально ограничить пространство ложных идентификаций. Смещение фокуса со сведений популяционного формата на конкретный организм, орган или ткань позволяет оптимизировать размер библиотеки за счет удаления неэкспрессирующихся или малоэкспрессирующихся генов и при этом учесть индивидуальные изменения, в том числе замены отдельных аминокислот, вставки и делеции, которые не включены в список последовательностей из UniProt. Учет популяционных данных, характеризующих аминокислотные полиморфизмы и схожие с ними молекулярные события редактирования белковой последовательности, может привести к многократному расширению поискового пространства, взрывному росту временной сложности интерпретации и неконтролируемому увеличению количества ложноположительных идентификаций [144,171].

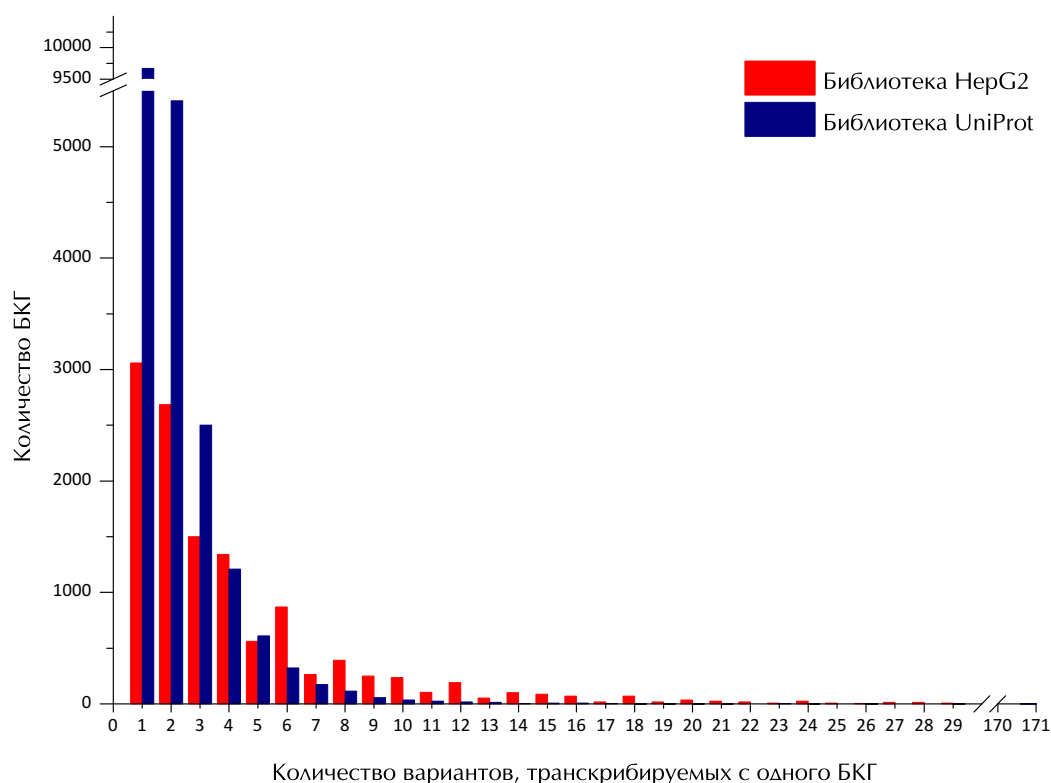


Рисунок 10. Гистограммы распределения белок-кодирующих генов по количеству транскрибируемых с них вариантов – потенциальных протеоформ – в библиотеке, сформированной по собственным транскриптомным данным для образца клеток линии HepG2, и стандартной библиотеке, загруженной из UniProt.

Иллюстрацией такого лавинообразного роста размера библиотеки популяционной природы может быть пример протеоформ, кодируемых геном *TP53* – известным онкосупрессором, множественные мутации которого были выявлены в

большинстве типов злокачественных опухолей [172]. Транскриптомное профилирование клеток HepG2 позволяет предполагать экспрессию 9 различных протеоформ (канонической формы и четырех СО-вариантов, в каждом из которых может произойти замена остатка пролина на остаток аргинина).

По данным UniProt, ген *TP53* кодирует 9 СО-вариантов белка P04637. Помимо сведений о продуктах альтернативного сплайсинга, страница белка P04637 в ресурсе UniProt содержит записи о почти 1,5 тыс. различных вариантов, в числе которых нейтральные полиморфизмы, вставки, делеции, ассоциированные с патологиями мутации и продукты редактирования РНК. Методические ограничения современных протеомных методов не позволяют оценить частоту совместной встречаемости таких событий, однако, если предположить, что точечные изменения аминокислотной последовательности могут возникать как в канонических, так и в альтернативных вариантах, то описание всех возможных протеоформ только одного этого гена – по самым скромным подсчетам – потребует учета свыше 15 тыс. последовательностей.

Секвенирование РНК предоставляет полную и неангажированную другими исследованиями картину транскриптома для тщательного изучения протеома целевого образца – не только с точки зрения учета уникальных и возможно ранее не аннотированных протеоформ, но и с позиции получения априорных знаний о путях экспрессии генов.

Результаты выполнения проекта ENCODE (Encyclopedia of DNA Elements) подтвердили тот факт, что гены в большинстве своем не «придерживаются» стратегии минимальной экспрессии различных протеоформ: в среднем для одного гена было обнаружено 10-12 сплайс-транскриптов в каждом исследованном типе биоматериала [173]. Выявление превалирующего транскрипта и определение степени его доминирования над другими транскриптами того же гена позволит сформировать представление о вкладе транскриптомной гетерогенности в гетерогенность протеомную [174].

Для выявления случаев количественного превалирования альтернативных транскриптов над каноническими для каждого из вариантов нами было рассчитано число прочтений. В качестве условной количественной оценки экспрессии транскрипта мы использовали значение параметра FPKM. Анализ экспрессии

экзонов продемонстрировал, что в 25% БКГ клеточноц линии HepG2 превалирующей является отличная от канонической сплайс-форма.

Знания о превалирующей в конкретном образце СО-протеоформе позволяют внести корректировки в алгоритм поиска и идентификации белков в результате масс-спектрометрического анализа. Выявление превалирующего сплайс-варианта может быть первым шагом в разработке методики направленного поиска конкретной протеоформы.

3.2. Поиск пептидных фрагментов для идентификации протеоформ

Свободный доступ к результатам протеомных экспериментов открывает широкие возможности для получения новых знаний [170,175] и смещает фокус с накопления экспериментальных данных на их биоинформационную интерпретацию. Прогресс биоинформатики, адаптировавший задачу анализа протеомных данных к базовым вычислительным мощностям [176], стимулировал возникновение множества поисковых алгоритмов. Тем не менее, в настоящий момент не существует универсального подхода к обработке результатов масс-спектрометрических экспериментов, и потому эффективным решением может быть использование комбинации нескольких поисковых алгоритмов, позволяющих получить более полные сведения об исследуемом объекте.

Для выбора сочетания алгоритмов, оптимального для решения типичной протеомной задачи – исследования состава протеома клеточной линии, мы проанализировали влияние поисковых настроек и размера референсной библиотеки, относительно которой производится поиск, на производительность обработки результатов масс-спектрометрического эксперимента.

Тестовый MS-файл среднего размера (78 Мб), полученный для одной из ячеек двумерного геля, был обработан по отдельности восемью поисковыми алгоритмами (использовали X!Tandem, MS-GF+, MSAmada, Myrimatch, Comet, Tide, Andromeda и OMSSA) и объединенным комплексом из перечисленных алгоритмов. Алгоритмы MSAmada и Andromeda демонстрировали нестабильный результат, и потому были исключены из тестирования. Поиск каждым алгоритмом осуществляли по библиотеке канонических и сплайс-опосредованных белковых последовательностей человека UniProt (версия 2017_04) с поисковыми

настройками, описанными в разделе «Источники и процессинг протеомных данных» главы «Материалы и методы». Рисунок 11а, содержащий сведения о длительности обработки МС-данных, демонстрирует высокую экспрессность алгоритмов X!Tandem, OMSSA и MyriMatch.

Другим базовым параметром выбора поисковой машины является количество детектированных пептидов. Из шести исследуемых поисковых машин наибольшее количество пептидов идентифицируется при использовании алгоритма OMSSA, остальные поисковые алгоритмы демонстрируют на 20-25% меньший результат. Минимальное количество пептидов было детектировано с помощью алгоритма MyriMatch (Рис. 11б).

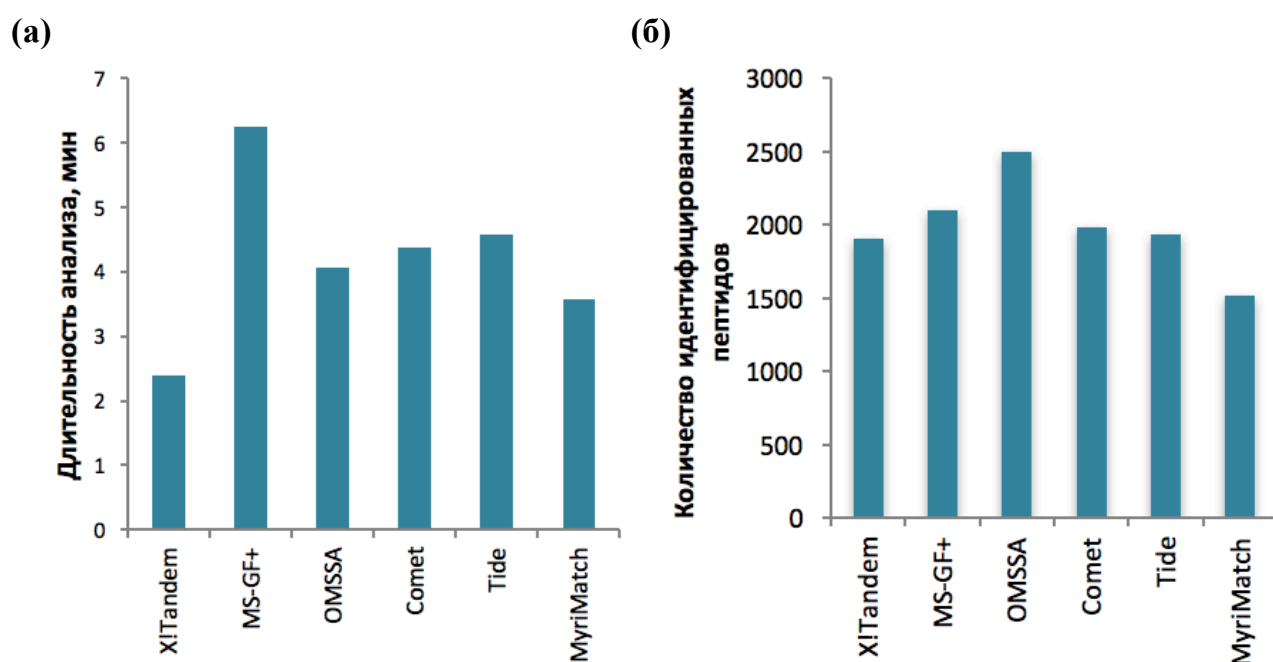


Рисунок 11. Результаты тестирования поисковых машин на предмет экспрессности анализа (а) и количества идентифицированных пептидов (б).

Помимо производительности поисковые алгоритмы различаются набором доступных для варьирования параметров поиска и способом оценки достоверности результатов: применение разных поисковых алгоритмов к одному и тому же МС-файлу ожидаемо даст разные результаты. Для выбора оптимальной комбинации поисковых машин масс-спектрометрический файл среднего размера был проанализирован всеми возможными комбинациями поисковых машин – от одной машины в комбинации до одновременного подключения всех шести рассматриваемых поисковых алгоритмов. Результат тестирования

продемонстрировал экспоненциальный рост вычислительного времени интерпретации масс-спектров с увеличением комбинации поисковых машин.

На рисунке 12 приведены усредненные данные по времени и количеству идентификаций при различных вариантах комбинаций машин. Включение все большего числа поисковых машин в поисковую комбинацию не приводит к значительному росту идентификаций – зависимость количества детектированных пептидов от размера комбинации алгоритмов носит логарифмический характер (рис. 12а). Небольшой прирост в идентификациях может быть объяснен изменением соотношения между ложными и истинными идентификациями, зафиксированными разными поисковыми алгоритмами.

Расширение арсенала поисковых машин ожидаемо увеличивает временные затраты – зависимость вычислительного времени от размера комбинации алгоритмов носит экспоненциальный характер (рис. 12б). Такое явление нелинейного роста вычислительного времени с увеличением числа используемых для анализа поисковых машин может быть связано с конкуренцией поисковых алгоритмов за ограниченные процессорные ресурсы. Мы полагаем, что перераспределение поисковых алгоритмов между несколькими десинхронизированными процессорами позволит добиться линейной зависимости вычислительного времени от количества используемых поисковых инструментов.

Сопоставление перечня белков, идентифицированных при различных комбинациях поисковых машин, показывает, что перекрытие между результатами варьируется между 45% и 95%. При одновременном использовании нескольких машин количество идентификаций будет отличаться как для отдельных машин, так и для их комбинаций, поскольку интегратор PeptideShaker не объединяет, а «взвешивает» вклад идентификации, выявленной каждым из алгоритмов в отдельности.

На основании полученных данных было выдвинуто предложение, что оптимальная для интерпретации результатов протеомного профилирования комбинация включает 3-5 поисковых машин. Поскольку количество пептидов, детектированных при использовании пяти алгоритмов, не оправдывает временные затраты по сравнению с использованием трех алгоритмов, то комбинация из трех алгоритмов была выбрана в качестве оптимальной.

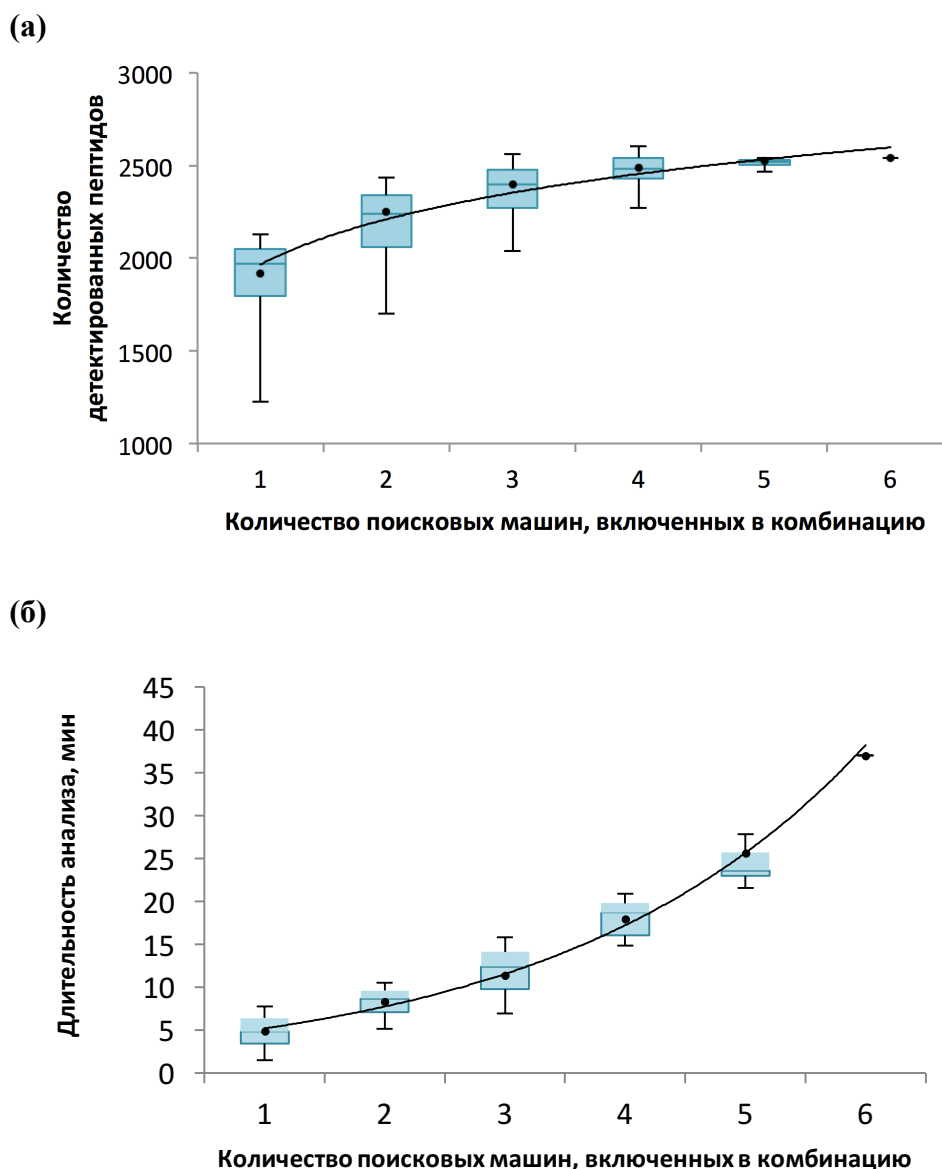


Рисунок 12. Изменение числа идентифицированных белков (а) и времени обработки масс-спектрометрических данных (б) при расширении арсенала используемых поисковых алгоритмов.

Чтобы выбрать из интегрированных в платформу SearchGUI три поисковых машины, которые могли войти в состав оптимальной комбинации, было проанализировано 10 возможных сочетаний. Из сопоставления были исключены Andromeda, MS Amanda, а также Tide. Последний схож с алгоритмом, реализованным в Comet, но при этом позволяет детектировать меньшее количество пептидов за большее время (рис. 12 а,б).

Усредненные показатели времени и количества пептидов, идентифицированных при использовании различных комбинаций поисковиков, приведены на рисунке 13. Количество идентификаций незначительно варьируется:

как при обработке разных по размеру файлов масс-спектров, так и при использовании различных библиотек. Временные затраты, напротив, значительно увеличиваются в случае максимальной нагрузки: самого большого файла масс-спектров и наиболее обширной библиотеки.

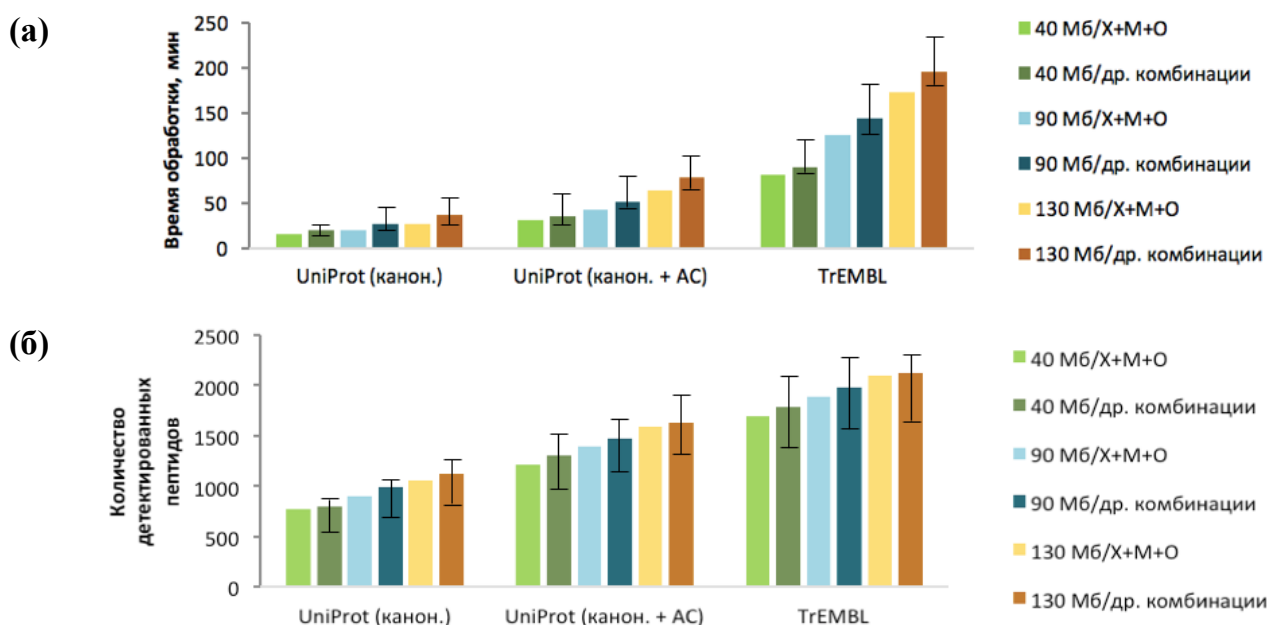


Рисунок 13. Зависимость длительности интерпретации (а) и количества детектированных пептидов (б) от размера референсной библиотеки и анализируемого файла панорамных масс-спектров. В каждой паре столбцов гистограммы светлый столбец относится к комбинации X!Tandem + MS-GF + OMSSA (X+M+O), а темный характеризует усредненное по всем остальным комбинациям время анализа и количество детектированных пептидов.

По результатам анализа сочетание X!Tandem, MS-GF+ и OMSSA было признано оптимальным по соотношению скорости анализа и количества идентификаций. Поисковый алгоритм X!Tandem является наиболее быстрым среди представленных, а MS-GF+ and OMSSA – наиболее результативными.

При обработке MC-файла среднего размера и со стандартными параметрами выбранная комбинация показывает схожие результаты с другими комбинациями поисковых машин, однако при необходимости обработки множества файлов разного размера комбинация X!Tandem, MS-GF+ и OMSSA демонстрирует преимущество над остальными возможными вариантами.

В рамках решения задачи поиска протеоформ, которая связана с обработкой большого массива масс-спектрометрических файлов, была показана эффективность выбора комбинации X!Tandem, MS-GF+ и OMSSA. При использовании выбранной комбинации поисковых машин обработка 192 масс-

спектров фракционированных белков опухолевой клеточной линии HerG2 заняла около 20 часов. Примечательно, что в случае обработки файлов всеми доступными в пакете SearchGUI алгоритмами, процесс интерпретации мог бы растянуться на 11 дней. Такая экономия времени особенно важна потому, что зачастую исследователь в рамках одной задачи вынужден многократно интерпретировать результаты одного и того же масс-спектрометрического эксперимента. Подобного сценария следует ожидать, если исследование сфокусировано, например, на выявлении нескольких типов посттрансляционных модификаций. Так, получение сведений только о фосфорилированных и ацетилированных протеоформах опухолевой линии HerG2 могло занять больше месяца.

Вычислительное время, необходимое для обработки масс-спектров, – это одна из самых критичных характеристик рабочего процесса интерпретации результатов, особенно в условиях обработки больших массивов данных. Заметное время тратится на прописывание параметров поиска, пути к анализируемому файлу и папке, хранящей результаты обработки. Автоматизация этих рутинных задач разработанным алгоритмом, реализованном на языке java (<https://yadi.sk/d/ldZGMovy3KJPqn>), позволила сэкономить около 5-10% времени, которое тратится на интерпретацию масс-спектров: почти минута на каждый файл или более 3 часов работы оператора, которые потребовались бы для настройки параметров поиска всех 192-х технических повторений.

Валидация комбинации поисковых алгоритмов на наборе UPS2

Для валидации выбранной комбинации поисковых машин (X!Tandem, OMSSA и MS-GF+) был проанализирован результат масс-спектрометрического исследования белков, входящих в состав тестового набора UPS2. Набор UPS2 состоит из 48 белков человека в диапазоне от 500 аттомолей до 50 пикомолей. Методические детали панорамного масс-спектрометрического анализа набора UPS2 описаны в работе [177].

Результаты масс-спектрометрического анализа смеси белков набора UPS2 были обработаны алгоритмами X!Tandem, OMSSA и MS-GF+ по отдельности, а также поисковым инструментом Mascot – одним из наиболее популярных коммерческих продуктов для интерпретации масс-спектрометрических файлов. В

качестве референсной библиотеки был использован перечень канонических и сплайс-опосредованных последовательностей из репозитория UniProt.

На рисунке 14 представлены результаты сопоставления белковых identifications, полученных с помощью Mascot и выбранного сочетания поисковых машин. Оба подхода к обработке масс-спектрометрических файлов позволили обнаружить только половину белков набора UPS2, концентрация которых выше 10^{-8} М, при этом комбинация поисковых алгоритмов позволила выявить на один белок больше. Примечательно, что данный белок – лизоцим (P61616, 500 фемтомоляр) – был идентифицирован только при использовании поисковой комбинации из трех алгоритмов.

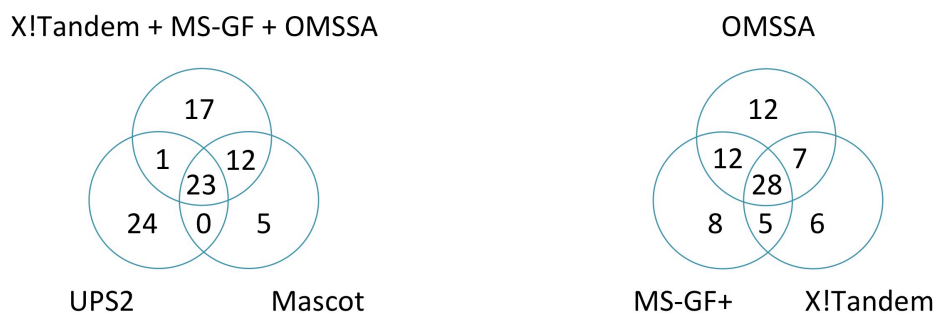


Рисунок 14. Сопоставление результатов обработки масс-спектрометрического файла коммерческим инструментом Mascot, поисковым сочетанием из трех алгоритмов и по отдельности каждым алгоритмом, входящим в комбинацию. Анализ белков набора UPS2 был проведен на масс-спектрометре Orbitrap Fusion (Thermo). Поиск проводили по референсной библиотеке, включающей канонические последовательности белков человека из ресурса Swiss-Prot, при интерпретации учитывали возможность пропуска одного сайта трипсинолиза и стандартные фиксированные и переменные химические модификации.

Следует отметить, что поисковая комбинация позволила идентифицировать 29 неспецифичных для набора UPS2 белков – значительную часть этого массива составляют загрязняющие образец кератины человека. Поисковым алгоритмом Mascot было детектировано 17 таких контаминирующих белков.

Итогом решения данной задачи стало определение оптимальной комбинации поисковых алгоритмов, позволяющих быстро обрабатывать большие массивы файлов масс-спектрометрических экспериментов: в качестве такой комбинации было выбрано сочетание X!Tandem, MS-GF+ и OMSSA.

Это исследование показало, что рациональный подход к выбору поисковых машин должен учитывать экспоненциальный рост временных затрат при

расширении поисковой комбинации. Оптимальный результат интерпретации файлов масс-спектров несколькими поисковыми машинами достигается с использованием комбинации наиболее разноплановых алгоритмов. К примеру, поисковые машины Tide и Comet «эволюционировали» от общего предшественника CruX, а алгоритмы MSAmada и Andromeda значительно напоминают Mascot. Одновременное использование нескольких «родственных» алгоритмов приводит к одинаковым истинным и ошибочным обнаружениям пептидов, поэтому включение поисковых алгоритмов одного генеза зачастую приводит лишь к увеличению времени анализа.

3.3. Протеом клеточной линии HepG2

В основу работы положен анализ протеоформ с использованием электрофореграммы, или протеомной карты. Протеомная карта, показанная на рисунке 15, была получена в работе С.Н. Нарыжного и соавторов путем применения метода двумерного электрофореза в сочетании с панорамной масс-спектрометрией [12]. Существенной методической особенностью работы являлось то, что двумерная электрофореграмма была подвергнута молекулярному сканированию [178]. Полиакриламидный гель после фракционирования был разрезан на 96 ячеек. Каждая из ячеек, содержащих фракционированные белки опухолевой линии HepG2, была трипсинолизирована и исследована с применением высокоразрешающей масс-спектрометрии пептидных фрагментов.

Для целей данной работы полученная коллегами протеомная карта была конвертирована в информационный массив, состоящий из 96 записей. Каждая запись содержала информацию о координатах ячейки на двумерной электрофореграмме и два масс-спектра. Каждый из масс-спектров был получен в результате анализа белкового гидролизата методом tandemной масс-спектрометрии, сопряженной с высокоэффективной жидкостной хроматографией. Таким образом, исходную протеомную карту представили в виде записей из координат «молекулярная масса» и «изоэлектрическая точка» и соотнесенных с этими координатами файлами двух технических повторений масс-спектрометрического эксперимента. Такое сочетанное представление о физико-химических (координаты ячейки на геле), биохимических (доступность для триптического гидролиза в геле) и физических данных (измерение массово-

зарядовых характеристик пептидных фрагментов белков), позволило автоматизировать процедуру выявления протеоформ в составе протеома клеточной линии HepG2.

Алгоритм, разработанный для автоматической обработки файлы отчетов, которые генерирует программа PeptideShaker, был применен к результатам протеомного профилирования фракционированных белков опухолевой линии HepG2.

В качестве иллюстрации работы алгоритма рассмотрим ячейку E5 – одну из 96 ячеек геля, наиболее насыщенную по количеству детектированных пептидов (рис. 15). Она охватывает участок геля с координатами по pI от 6,30 до 6,88 и по MW с координатами 40 – 52 кДа. Средние значения изоэлектрической точки и молекулярной массы, рассчитанные для протеома человека, составляют 7,23 и 61 кДа, соответственно. Таким образом, белковые молекулы, локализующиеся в ячейке E5, обладают «средними» для протеома физико-химическими свойствами.

В первом техническом повторении масс-спектрометрического исследования этой ячейки было детектировано 4 017 пептидов, во втором – 4 080, пересечение составило 3 145 пептидов (74% от всех детектированных). Сведения о детекции пептидов в первом и втором повторениях были объединены. Объединенный массив был отфильтрован от пептидов, длина которых не попадала в интервал 9-25 а.о. Из дальнейшего рассмотрения были удалены нетриптические пептиды, равно как и пептиды, содержащие два и более сайтов трипсинолиза. Также были отфильтрованы пептиды контаминирующих белков: таких «загрязняющих» белков было 9, в их числе кератины человека и свиной трипсин, использованный в ходе пробоподготовки для ферментативного протеолиза.

После технической фильтрации объединенный по двум техническим повторам результат детекции включал 3170 пептидов, 560 из которых были протеотипическими, то есть соответствовали единственной форме белка в исследуемом протеоме.

Остальные пептиды картировались на несколько протеоформ одного гена или несколько родственных белков. Была проведена оценка уникальных наборов пептидов (т.е. групп пептидов, которые картируются на несколько белковых продуктов в рамках белковой группы и не позволяют сделать вывод об

однозначной детекции протеоформы, а свидетельствуют только об экспрессии гена) – таких групп в рассматриваемой ячейке было 514. В среднем, на одну группу приходилось 6,2 пептидов. Количество уникальных наборов пептидов позволяет произвести предварительную, ориентировочную оценку гетерогенности протеома «снизу», но не предоставляет сведений о разнообразии белкового состава.

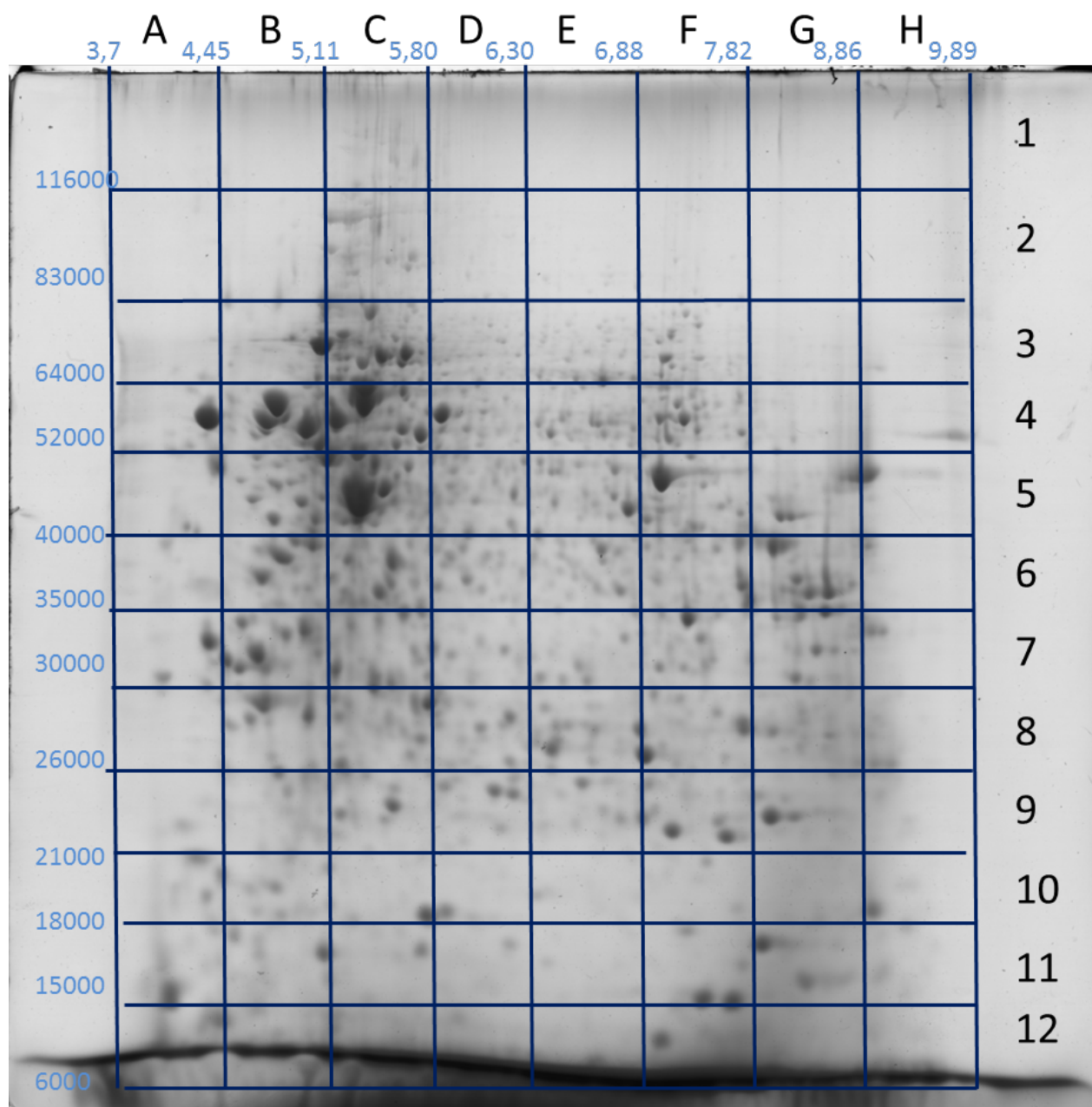


Рисунок 15. 2DE-карта протеома клеточной линии HepG2: по оси ординат – значения молекулярной массы, Да, по оси абсцисс – значения изоэлектрической точки [12].

Международные стандарты надежной детекции протеоформы требуют зафиксировать не менее двух пептидов, минимум один из которых является протеотипическим [13]. В исследуемой ячейке надежно было детектировано 117

протеоформ, закодированных 115 генами, в том числе восемь сплайс-вариантов, три последовательности с ОАП и один случай комбинации этих событий. Аминокислотные последовательности пептидов, характеризующих протеоформы, были дополнительно проверены на уникальность среди всех белков человека, аннотированных в UniProt.

Выявленные аминокислотные замены были исследованы на предмет степени влияния на функциональные свойства белка с помощью веб-приложения MutationAssessor [179], валидированного на массиве из 60 тыс. нейтральных и ассоциированных с патологией аминокислотных полиморфизмов. Уровень вредоносности аминокислотной замены (англ. functional impact score, FIS) определялся эволюционной консервативностью полиморфной позиции во всем семействе белков и специфичностью этой позиции в рамках подсемейства белков-ортологов.

С особым вниманием были проанализированы масс-спектры протеотипических пептидов неканонических протеоформ. Детекция сплайс-вариантов и протеоформ с аминокислотными заменами зачастую затруднена в силу малого количества протеотипических пептидов, которые могли однозначно свидетельствовать о детекции такого варианта. В случае точечной замены характеристичный пептид, который различается между каноническим и полиморфным вариантами, и вовсе единственный.

Детекция сплайс-опосредованной протеоформы может быть описана на примере варианта глутаминазы O94925-3, катализирующей реакцию гидролиза глутамина – одного из основных механизмов регуляции кислотно-щелочного равновесия в организме. Недавние исследования позволяют предположить, что данная сплайс-форма глутаминазы в силу своих структурных особенностей эффективнее, чем канонический вариант, справляется с ключевыми задачами поддержания повышенных метаболических потребностей опухоли. Таким образом, дисбаланс содержания канонической и сплайс-форм может рассматриваться в качестве потенциального биомаркера опухолевого процесса [180]. Сплайс-вариант глутаминазы O94925-3 был детектирован по двум протеотипическим и 14 неуникальным пептидам, с равным успехом картирующимся на аминокислотные последовательности нескольких родственных протеоформ. Спектр фрагментации

одного из двух детектированных протеотипических пептидов (VSPESNEDISTTVVYR) представлен на рисунке 16.

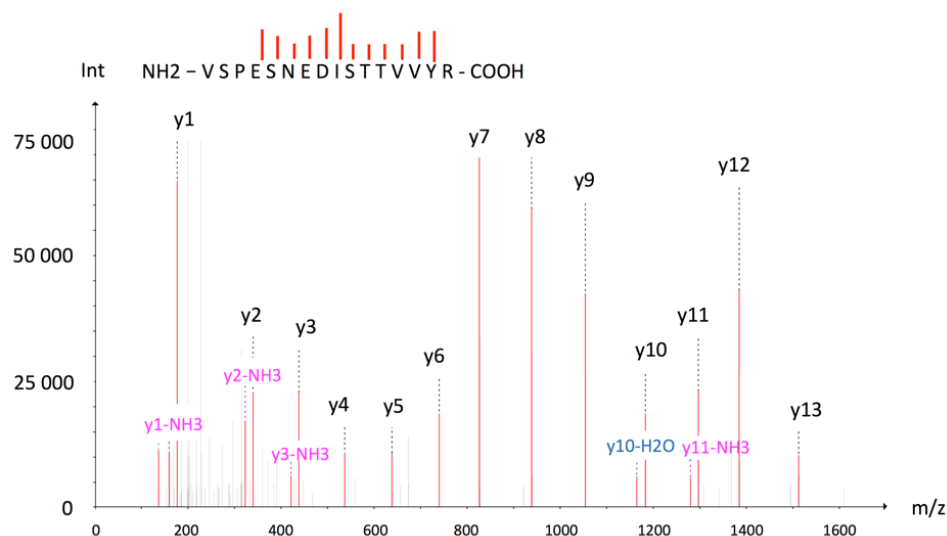


Рисунок 16. Масс-спектр и фрагментные ионы пептида VSPESNEDISTTVVYR, однозначно характеризующего сплайс-опосредованный вариант O94925-3. По данным репозитория GPMdb, этот пептид был детектирован более 28 тыс. раз. По оси абсцисс спектра отмечены значения массово-зарядового параметра m/z , по оси ординат – значения интенсивности.

Полное покрытие аминокислотной последовательности С-концевыми у-ионами подтверждает надежную детекцию данного пептида, и, как следствие, всей сплайс-протеоформы O94925-3.

Рассмотрим случай детекции протеоформы митохондриальной пептидазы O75439, которая содержит замену остатка глутаминовой кислоты на остаток аспарагиновой в 396-ом положении. Этот полиморфизм детектирован как на транскриптомном (по данным dbSNP), так и на протеомном уровнях (по данным GPMdb) и считается естественным популяционным вариантом.

Детекция серии у-ионов пептидных фрагментов (y5-y10), содержащих полиморфное положение, позволяет с уверенностью говорить о обнаруженной аминокислотной замене (рис. 17а).

Помимо протеотипического пептида в исследуемой ячейке было детектировано еще 11 пептидов. Их аминокислотные последовательности полностью картируются не только на протеоформу с аминокислотной заменой, но и на каноническую протеоформу, а также на вариант с делецией. Пептид с канонической глутаминовой кислотой в 396-ом полиморфном положении также был детектирован в ячейке E5 (рис. 17б). Согласно данным GPMdb, канонический

вариант встречается в результатах масс-спектрометрических экспериментов в 20 раз чаще.

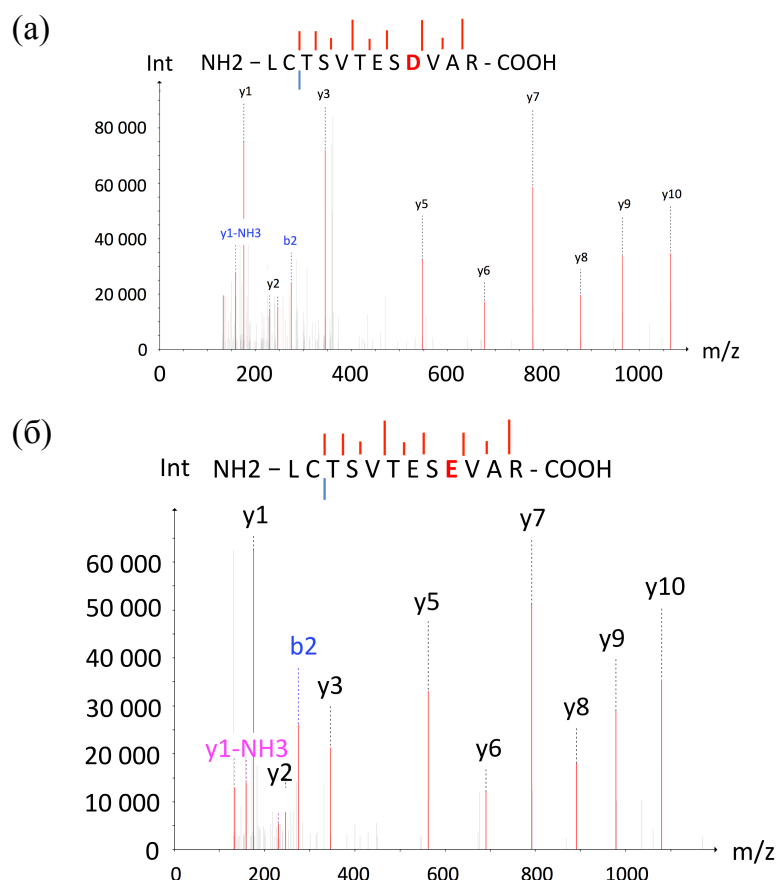


Рисунок 17. Масс-спектры и фрагментные ионы пептида LCTSVTESDVAR (а), специфичного протеоформе O75439 с заменой остатка глутаминовой кислоты на остаток аспарагиновой кислоты в 396-ом положении, и пептида LCTSVTESEVAR (б), протеотипического для канонического варианта белка.

Детекция замены аминокислоты в сплайс-варианте белка является весьма редким событием. Среди всех уникальных пептидов, соответствующих последовательностям белков человека в базе UniProt, только около 0,03% потенциально могут быть использованы для детекции такого типа протеоформ [10]. В анализируемой ячейке была детектирована сплайс-протеоформа лиазы Q96I15-2, в которой остаток аланина был заменен на остаток треонина в 175-ом положении. Эта протеоформа была идентифицирована по восьми пептидам, семь из которых картировались еще и на сплайс-форму Q96I15-2 без полиморфизма, и только один пептид содержал полиморфное положение. Спектр фрагментации пептида VSGQTEVDDILAAVRPTTR представлен на рисунке 18.

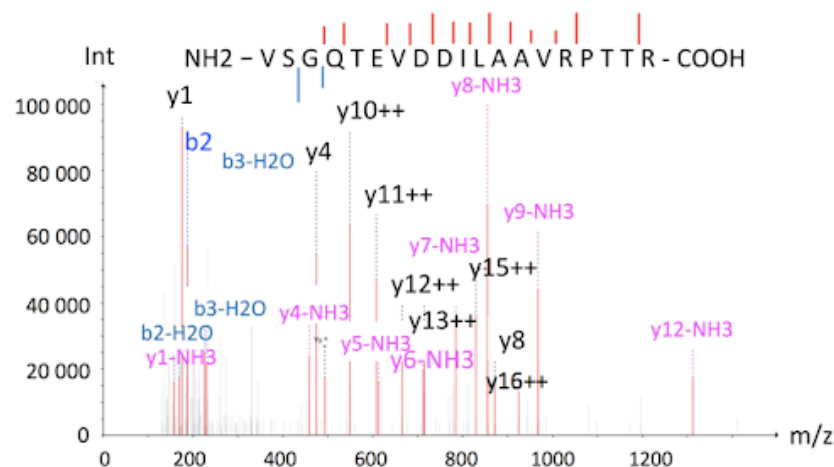


Рисунок 18. Масс-спектр и фрагментные ионы пептида VSGQTEVDDILA AVRPTTR, специфичного протеоформе Q96I15-2, в которой остаток аланина был замещен на остаток треонина в 175-ом положении. Протеотипический пептид был детектирован в 265 экспериментах, согласно данным репозитория GPMdb. Пептид с «каноническим» остатком аланина в полиморфном положении был детектирован в 6 раз чаще.

Алгоритм, направленный на интерпретацию электронных отчетов, был применен к результатам масс-спектрометрического исследования всех 96 ячеек геля. Суммарно во всех 96 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.

Большая часть из обнаруженных протеоформ (998) локализовалась в нескольких ячейках геля одновременно (в среднем – в шести) 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: фосфорилирования по остаткам серина, треонина и тирозина, ацетилирования лизина, метилирования лизина и аргинина, амидирования С-концевых остатков и деамидирования остатков аспарагина и глутамина. Дальнейший анализ заключался в сопоставлении данных о физико-химических параметрах белков, химических и посттрансляционных модификациях с координатами молекулярной массы и изоэлектрической точки, характерными для тех ячеек, в которых протеоформы были детектированы.

Оставшиеся 530 протеоформ были детектированы в единственной ячейке. Среди детектированных протеоформ было 76 сплайс-вариантов, 76 протеоформ с одноаминокислотными заменами, пять протеоформ, в которых замена аминокислоты возникла одновременно с альтернативным сплайсингом, и один случай делеции.

Фактически обнаружено было 6 950 белковых групп, каждая из которых включает несколько белковых продуктов. На данном этапе, заручившись только результатом масс-спектрометрического эксперимента, невозможно по детектированным для белковых групп пептидам конкретизировать обнаруженную протеоформу. Тем не менее, даже такие данные в ряде случаев не являются тупиковыми и могут быть эффективно использованы при дальнейшем анализе.

3.4. Аннотация выявленных протеоформ с учетом их физико-химических параметров

Двумерный гель-электрофорез долгие годы стабильно занимает одну из лидирующих позиций среди биохимических методов. Годы использования этого метода позволили не только по достоинству оценить его преимущества, но и выявить ограничения: в стандартном исполнении метод не подходит для фракционирования и исследования сильно-кислых, сильно-основных, гидрофобных и мембранных белков.

Ограничения протеомных методов в целом, и двумерного гель-электрофореза в частности, приводят к тому, что исследователи от эксперимента к эксперименту испытывают чувство дежавю, наблюдая одни и те же высокопредставленные гидрофильные белки со средними значениями pI практически в каждом геле сложного биообразца. Таких «повторяющихся» из раза в раз белков около тысячи: ограниченный круг их функций включает поддержание общего метаболизма, синтез белка и ответ на стрессовые воздействия [181]. Учитывая эти наблюдения, можно заключить, что метод 2DE применительно к биологическим образцам – это не способ обозреть весь «объем» протеома, а скорее техника фракционирования, которая не только приближает «сложность» исследуемого объекта к разрешающей мощности самого метода, но и позволяет сделать выводы о транслируемых протеоформах. Актуален вопрос – какая доля протеома обладает физико-химическими параметрами, приемлемыми для исследования методом 2DE?

Для оценки доли протеома клеточной линии HepG2 с приемлемыми для исследования методом 2DE физико-химическими параметрами было построено распределение экспрессируемого массива белков по MW и pI в соответствии с результатами транскриптомного анализа образца. Для предсказания значений изоэлектрической точки был применен ресурс Proteome- pI , позволяющий рассчитать значение pI белка по 18 алгоритмам [182]. Построения на рисунке 19 были выполнены по усредненному значению pI . Несмотря на то, что использованные формулы берут за основу аминокислотную последовательность и не учитывают вклад ПТМ, которые способны значительно изменить общий заряд белка [183], даже такие отрывочные сведения могут быть использованы в качестве стартовой точки на пути к получению новых знаний о белках.

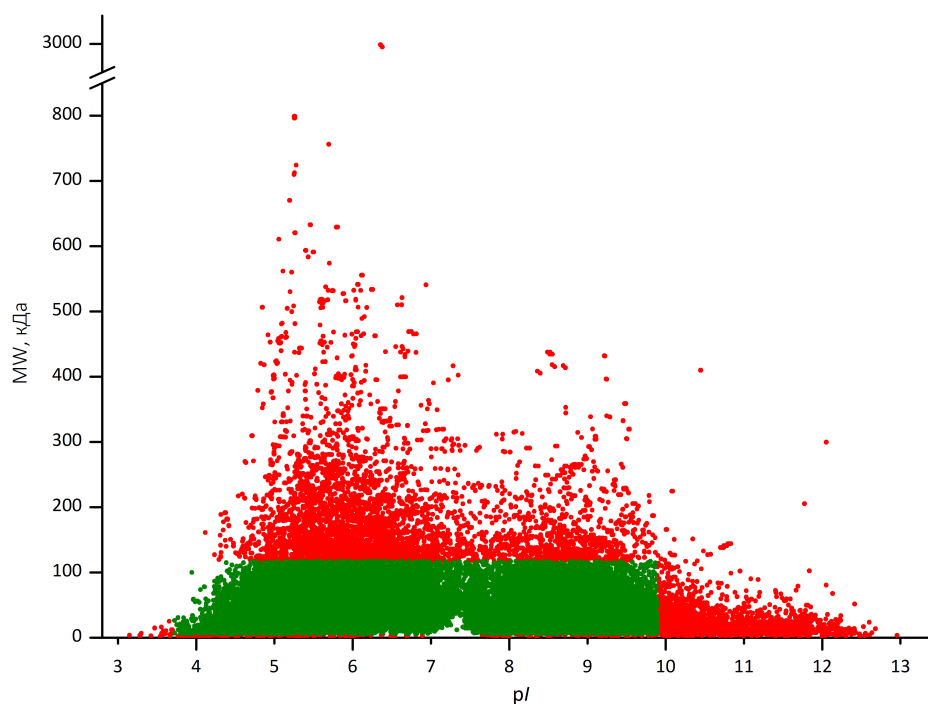


Рисунок 19. Распределение белков человека по молекулярной массе и изоэлектрической точке. Концентрированная область зеленых точек относится к белкам, молекулярная масса и изоэлектрическая точка которых попадают в надежно мониторируемый интервал 3,70 – 9,89 по pI и 6 – 116 кДа по MW. Красные точки соответствуют белкам, параметры которых выходят за пределы геля.

Примечательно, что, согласно расчетам, 22% ожидаемого протеома HepG2 находится за пределами надежно мониторируемого интервала pI и MW. Аналогичная оценка (18%) была получена и для всего массива белков человека. Сопоставление теоретических и экспериментальных координат локализации белка на геле может быть эффективно использовано в качестве дополнительной формы

контроля. Выявление протеоформы в ячейке, координаты MW и pI которой значительно отличаются от теоретически рассчитанных, может свидетельствовать о ПТМ, протеолизе или деградации белка и других молекулярных событиях, если иные варианты протеоформ (альтернативный сплайсинг или нуклеотидные замены) не подкреплены транскриптомными прогнозами.

Локализация протеоформы на геле зависит от множества факторов. Координаты локализации белка в равной степени определяются его физико-химическими характеристиками и особенностями технического проведения конкретного эксперимента. Сложность интерпретации результатов 2DE-эксперимента, сопряженного с панорамной масс-спектрометрией, была продемонстрирована на тестовом массиве из 530 надежно идентифицированных протеоформ клеток HepG2. Все протеоформы, входящие в этот массив, были обнаружены только в одной ячейке геля. Для каждой протеоформы было масс-спектрометрически детектировано не менее двух пептидов, минимум один из которых был протеотипическим, причем протеотипичность дополнительно проверялась на всех белках базы UniProt. Тестовый массив включал 473 канонических варианта белков, 29 протеоформ с одноаминокислотной заменой, 27 сплайс-вариантов и одну сплайс-протеоформу с заменой аминокислоты.

Для каждой из детектированных протеоформ были сопоставлены теоретически рассчитанные с помощью инструмента Pyteomics [167] параметры pI и MW с координатами ячейки, в которой пептиды протеоформы были детектированы. Почти для половины массива протеоформ (243 варианта) теоретические и экспериментально установленные координаты локализации совпали. Для оставшейся части массива была проанализирована величина и направление сдвига экспериментальных координат относительно теоретически рассчитанных для того, чтобы установить причину, вызвавшую сдвиг.

Треть тестового массива (163 протеоформы) была детектирована в ячейках, координаты которых соответствуют меньшим значениям молекулярного веса. Причиной миграции протеоформ в области геля, соответствующие более легким белковым молекулам, может быть протеолиз белка на несколько фрагментов. Последовательности таких фрагментов могут содержать протеотипические пептиды, необходимые для идентификации «родительской» белковой молекулы,

при этом молекулярный вес фрагментов естественным образом оказывается ниже веса интактного белка. Анализ последовательности пептидов показал, что в большинстве случаев (71%), пептиды компактно картируются на один участок, а не распределяются по всей белковой последовательности. Также аналогичный эффект на локализацию протеоформы может оказать липидирование – липофильные группы, прикрепленные к белковой молекуле, повышают ее гидрофобность и, как следствие, миграция липидированного белка в геле может облегчиться.

Напротив, сдвиг координат локализации в область больших масс может быть вызван олигомеризацией протеоформы или ковалентной сшивкой с белковыми партнерами или контаминирующими пептидами, нуклеиновыми кислотами и другими загрязняющими веществами. Сшивка протеоформы с другими белковыми продуктами, которая привела к обнаружению специфичных ей пептидов в нехарактерной для их массы и заряда области геля, рассмотрена на примере миграции комплекса простат-специфичного антигена (P07288) с пептидом-ингибитором SPAAT, также обладающим свойствами сериновых протеаз, в область больших молекулярных масс [184]. Подобный сдвиг наблюдался для 68 протеоформ исследуемого массива. Помимо ковалентной сшивки или олигомеризации, причиной сдвига такого рода могут быть особенности структуры и аминокислотного состава белка. Аналогичным эффектом обладает и множественное гликозилирование. Сдвиг в область больших масс мы наблюдали для 68 протеоформ тестового массива.

При интеграции масс-спектрометрических данных и результатов двумерного гель-электрофореза необходимо также учитывать особенности структуры и аминокислотного состава белка. Например, значительно снизить подвижность молекулы может высокое содержание остатков пролина. Остатки этой аминокислоты способны нарушить структуру α -спиралей или β -складок и во многом определяют «угловатую» форму полипептидной цепи. Каноническим примером влияния пролинов на локализацию белка на геле является опухолевый антиген P53 (P04637). Молекулярная масса канонического варианта P53, самого длинного из всех возможных сплайс-вариантов, составляет 43,5 кДа. В целом по протеому, частота встречаемости пролина в аминокислотных последовательностях белков человека составляет около 6%. Последовательность белка P53 состоит из

пролинов на 11%. Высокое содержание остатков пролина затрудняет миграцию белка в геле, поэтому белок обнаруживается в области более высоких молекулярных масс - приблизительно на 10 кДа выше по оси молекулярной массы, чем собственный вес. «Аномальная» локализация опухолевого антигена в области 53 кДа послужила причиной его названия [185]. Предположительно, несовпадение теоретических и экспериментальных координат для 18 протеоформ тестового массива связано именно с высокой (> 10%) встречаемостью пролина в их последовательностях.

Примечательно, что неполная денатурация также может стать причиной аномальной локализации белка на геле. Закономерности в данном вопросе окончательно не выяснены, однако, согласно отдельным исследованиям, даже после денатурации белков около 20% β -складок и α -спиралей сохраняют свою структуру [186]. Существует целый класс т.н. кинетически стабильных белков, невосприимчивых к воздействию протеаз и денатурирующих агентов. Конформация таких белков (например, транстиретина и авидина) «заблокирована» из-за их необычайно низкой структурной гибкости [187]. На сегодняшний день нет специализированного репозитория, который бы свел воедино неструктурированные сведения о кинетически стабильных белках, поэтому гипотеза о влиянии высокой кинетической стабильности на локализацию белка на геле может стать объектом будущих исследований.

Для 41 протеоформы тестового массива было зафиксировано смещение по оси pI в «кислую» область относительно теоретически рассчитанных координат. Предположительно, такой сдвиг мог быть вызван множественными химическими и посттрансляционными модификациями, которые незначительно меняют общий вес белковой молекулы, но существенно влияют на ее суммарный заряд [183]. Для 24 протеоформ это предположение было подтверждено данными из протеомных репозиторий GPMdb и iPTMnet. Также был проведен повторный анализ файлов масс-спектров с учетом возможных модификаций с фокусом на наиболее часто встречающиеся и хорошо изученные типы ПТМ: фосфорилирование серина, треонина и тирозина, ацетилирование лизина или N-концевое ацетилирование, а также деамидирование остатков аспарагина и глутамина. Были учтены и химические модификации – окисление метионина и карбамидометилирование

цистеина [188]. Стоит отметить, что примерно 85% пептидов были детектированы с химическими модификациями – окислением метионина или карбамидометилированием цистеина, возникающими в ходе пробоподготовки клеток опухолевой линии HepG2. Осознанно узкий круг рассматриваемых химических и посттрансляционных модификаций связан с экспоненциальным ростом вычислительных затрат при увеличении количества рассматриваемых аминокислотных остатков, подверженных модификациям (и с линейным ростом при увеличении количества типов ПТМ) [189]. Если предположить возможность всего лишь двух ПТМ на пептид (из более чем 400 типов), то поисковое пространство увеличится в 160 тыс. раз, а вместе с ним увеличится и FDR. Поиск всех возможных паттернов ПТМ и химических модификаций увеличивает поисковое пространство до неконтролируемых размеров. Масс-спектрометрическая детекция модификации в большинстве случаев не представляет проблемы, гораздо более трудную задачу представляет интерпретация полученных данных.

Одну из самых сложных для интерпретации фракций представляют случаи сдвига локализации протеоформы в область щелочных pI , которых в тестовом массиве было пять. К примеру, пептиды «кислого» импорта O95373 (pI 4,7) были детектированы в ячейке H2 (интервал pI 8,9 – 10,0). Преимущественно ПТМ смещают локализацию протеоформы в кислую область относительно немодифицированной белковой молекул. Тем не менее, есть немногочисленный класс модификаций, которые могут привести к щелочному сдвигу. Одной из наиболее распространенных ПТМ, смещающих локализацию протеоформы в область основных значений pI , является амидирование С-концевых остатков [190]. Часто детектируемое метилирование лизина также способно перераспределить заряд на атоме азота аминогруппы, что приводит к небольшому увеличению значения изоэлектрической точки [191]. 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 [12,192]. Например, часть молекул «кислых» белков, попав в область высоких значений pH на стадии изоэлектрофокусировки, могла вступить в реакцию нейтрализации и выпасть в осадок. Как следствие, миграция осажденных белковых молекул в соответствующую их кислотности область геля могла быть существенно затруднена, из-за чего протеоформа обнаруживалась в одной или нескольких ячейках, характерных для более щелочных белков.

Несмотря на описанные выше методические особенности, тандем комплементарно дополняющих друг друга методов 2DE и панорамной масс-спектрометрии позволяет в ряде случаев различить между собой даже высокогомологичные протеоформы. Значения изоэлектрической точки и молекулярной массы нативного белка являются важными характеристиками, позволяющими подтвердить достоверность идентификации протеоформы. Анализ локализации идентифицированных протеоформ на геле способствовал исследованию белковых групп, а также аннотированию белковых продуктов, обнаруженных на пространственном удалении друг от друга.

3.4.1. Белковые продукты одного гена, идентифицируемые в разных областях двумерной электрофореграммы

Первый тип неоднозначных идентификаций относится к случаям, когда в нескольких пространственно удаленных ячейках были детектированы пептиды, относящиеся к одной и той же протеоформе. Помимо локализации на границе нескольких ячеек, детекция одной и той же протеоформы в различных зонах геля может быть вызвана целым спектром причин, который включает как посттрансляционные модификации, так и протеолиз, денатурацию, изомеризацию, ковалентную сшивку белковой молекулы с ее интерактивным партнером и другие описанные выше события.

Согласно молекулярной массе белка теплового шока P04792 (23 кДа) и расчетам значения pI (5,98), детекцию этого белка следует ожидать в ячейке D9 или D8, если принять к сведению высокое содержание пролиновых остатков. Однако, пептиды P04792 были также детектированы еще в восьми ячейках.

Смещение белка P04792 в область меньших масс обусловлено протеолизом (ячейка H12, MW 6-15 кДа), а сдвиг в область больших масс, одновременно с значимым изменением *pI* (ячейки A7, B6 и C2, MW 30-116 кДа) – гомо- или гетероолигомеризацией [193]. Также допустима артефактная С-С сшивка белка P04792 с интерактомыми партнерами – α - и β -тубулинами, каждый из которых весит около 50 кДа [194].

Один из самых распространенных типов ПТМ – фосфорилирование – заключается в замене нейтральных гидроксильных групп серина, треонина и тирозина отрицательно заряженными фосфо-группами. Как следствие, *pI* модифицированной молекулы снижается, причем изменение может быть значительным: даже одиночное фосфорилирование может снизить *pI* белка на 1-2 единицы [195]. Смещение белка теплового шока в область меньших *pI*, предположительно, было вызвано множественным фосфорилированием остатков серина. Согласно литературным данным, остатки серина S15, S78 и S82 белка P04792 фосфорилируются MAP-киназами, что приводит к отщеплению белка P04792 от олигомера, состоящего из нескольких небольших белков теплового шока [196]. В ячейке C8, соответствующей интервалу *pI* 5,11-5,80, было детектировано фосфорилирование в положении S82, причем в «оригинальных» ячейках D8 и D9 ПТМ не были обнаружены (рис. 20). Возможно, белок теплового шока несет и другие модификации, которые способствовали смещению его *pI* в кислую область, однако пептиды, которые позволили бы подтвердить другие сайты фосфорилирования, детектировать не удалось.

Таким образом можно заключить, что только в двух из 10 ячеек детектирована истинная протеоформа (канонический вариант без модификаций и с фосфорилированием), а детекция протеотипических для белка P04792 пептидов в остальных восьми случаях – результат артефактов пробоподготовки.

Характеристичные конкретной протеоформе пептиды в редких случаях могут быть обнаружены в нескольких пространственно удаленных участках геля при ковалентной сшивке исследуемой белковой молекулы с ее интерактивным партнером.

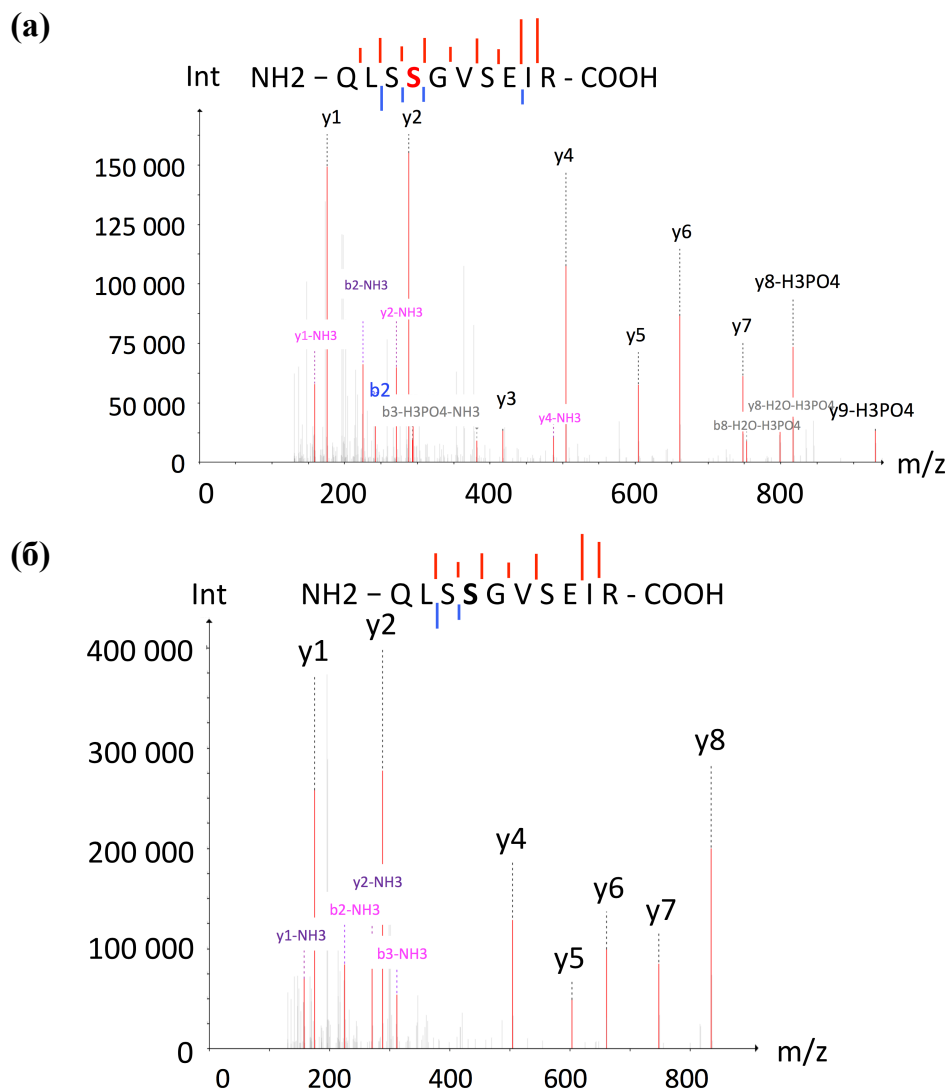


Рисунок 20. Масс-спектры фосфорилированного (а) и нефосфорилированного (б) пептида QLS(Phospho)SGVSEIR, который был детектирован в ячейках C8 и D8, соответственно. Серия пиков у-ионов (y8, y9 – с C-конца) и b-ионов (b3, b4, b8, N-конец) пептида свидетельствует о фосфорилировании серина в положении 82. Для нефосфорилированного пептида также надежно детектирована серия фрагментных ионов, которая позволяет сделать вывод об отсутствии в данном пептиде посттрансляционных модификаций.

Рассмотрим случай миграции комплекса на примере оксидоредуктазы Q96HP4. Протеотипические пептиды этого белка были обнаружены в ячейках F1 и F6. Предположительно, в ячейке F6 обнаружена каноническая форма белка Q96HP4, поскольку его масса (35 кДа) и заряд (8,86) соответствуют параметрам ячейки. Ячейка F1 соответствует гораздо менее подвижным белкам с большой молекулярной массой (> 100 кДа). Согласно базе данных BioGRID, белок Q96HP4 взаимодействует с семью белковыми партнерами, выявленными методом аффинной пробоподготовки в сочетании с масс-спектрометрией (рис. 21). Пептиды, протеотипические для одного из партнеров оксидоредуктазы – шаперона P10809 –

также были детектированы в ячейке F1. Молекулярная масса шаперона P10809 (около 61 кДа) тоже недостаточна велика, чтобы он самостоятельно мигрировал в область ячейки F1. Таким образом, можно полагать, что в ходе пробоподготовки или непосредственно 2DE-эксперимента произошла ковалентная сшивка дегидрогеназы и шаперона.

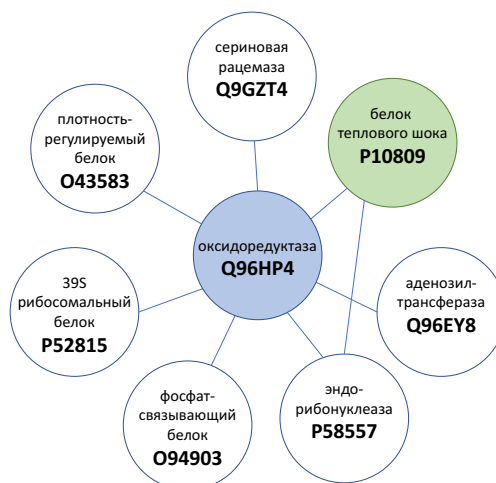


Рисунок 21. Интерактивная сеть, построенная для оксидоредуктазы Q96HP4.

Обнаружение одной и той же протеоформы в нескольких ячейках также может быть объяснено объединением мономеров в олигомеры. Такую ситуацию иллюстрирует пример глутарил-КоА дегидрогеназы Q92947. Пептиды, картирующиеся на аминокислотную последовательность этого белка, были детектированы в двух пространственно удаленных ячейках геля – F1 (*pI* 6,88-7,82, *MW* > 116 кДа) и F5 (тот же интервал *pI*, *MW* 10-52 кДа). Такое пространственное разделение объясняется существованием во фрагментируемой денатурированной белковой смеси двух вариантов белка: мономера с молекулярной массой около 48 кДа, который мигрировал в соответствующую его массе ячейку F5, и описанного в литературе тетрамера с соответствующей ячейке F1 молекулярной массой около 192 кДа [197], межсубъединичные контакты в котором были не полностью разрушены в ходе пробоподготовки.

Значительное изменение состава аминокислотной последовательности вследствие протеолиза полипептидной связи ожидаемо приводит к сдвигу локализации белка на геле в область низких молекулярных масс. Более того, такие события могут повлиять и на значение изоэлектрической точки «труктурированного» белка. Изменения изоэлектрической точки следует ожидать, если в результате

протеолиза сместилось соотношение между положительно- и отрицательно-заряженными остатками аминокислот белка.

Согласно полученным данным, протеолиз наблюдается примерно в 25% случаев обнаружения пептидов протеоформы в нескольких ячейках геля. Так, например, пептиды белка P61011, входящего в частицу узнавания сигнала, были обнаружены в двух ячейках геля. Параметры ячейки G4, в которой было детектировано более 20 пептидов, равномерно картирующихся на всю длину аминокислотной последовательности P61011 (около 40% покрытия), соответствуют значению молекулярной массы белка (рис. 22а). В ячейке G10, расположенной в зоне низких масс, было детектировано всего два пептида, которые соответствуют первым α -спиралям белка (рис. 22б).

Возможно, на этапе пробоподготовки или непосредственно во время двумерного гель-электрофореза часть молекул белка P61011 была расщеплена на фрагменты с низкой молекулярной массой, которые мигрировали в буфер и потому не представлены на геле.

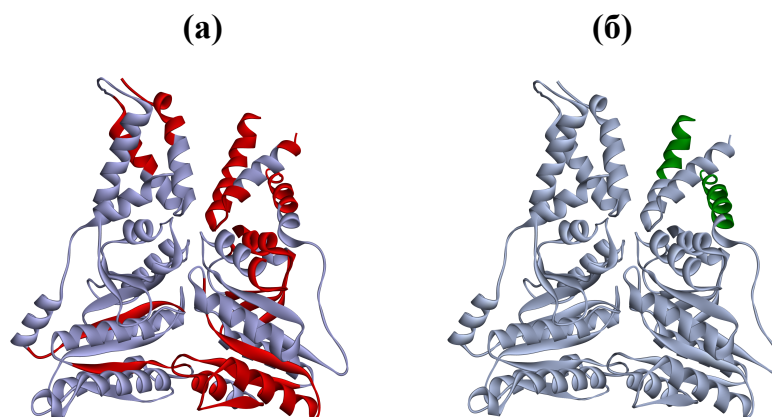


Рисунок 22. Экспериментально полученная модель трехмерной структуры белка P61011 из ресурса PDB. Красные участки структуры (а) соответствуют пептидам, детектированным в ячейке G4, зеленые (б) – в ячейке G10.

Закономерности процесса деградации белков на настоящий момент представляют больше вопросов, чем ответов. При обнаружении протеоформы в нескольких ячейках, координаты массы которых ниже ее молекулярного веса, мы полагали, что произошло расщепление белковой молекулы, и не рассматривали деградированные и фрагментированные в ходе пробоподготовки варианты белков как самостоятельные протеоформы.

3.4.2. Аннотация протеоформ, входящих в состав белковой группы

Выше мы рассмотрели пример аннотации белковых продуктов, обнаруженных в нехарактерных участках электрофореграммы. Мы выяснили, что смещение локализации протеоформы относительно теоретически рассчитанных координат может быть вызвано рядом причин, в их числе посттрансляционные модификации, протеолиз, денатурация, ковалентная сшивка белковой молекулы с ее белковыми партнерами и др.

Другой тип неоднозначной идентификации протеоформ объединяет случаи, когда в одной ячейке обнаружена белковая группа. Белковая группа включает протеоформы одного или нескольких родственных генов. Детектированный для такой группы набор пептидов картируется сразу на несколько белковых продуктов и не позволяет однозначно идентифицировать конкретный белковый вариант. Сопоставление теоретических характеристик заряда и молекулярного веса протеоформ белковой группы с границами ячейки, в которых эти протеоформы были обнаружены, дает возможность сделать вывод о надежности идентификаций или различить даже высокогомлогичные белки.

Так, на аминокислотные последовательности двух белков семейства ацил-КоА-тиоэстераз (P49753 и Q86TX2) картировались одни и те же пептиды в ячейке F5 (pI : 6,88-7,82 и MW : 40-52 кДа). 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).

Зачастую альтернативный сплайсинг значительно изменяет аминокислотную последовательность протеоформы. Сопоставление теоретических и экспериментальных pI и MW позволяет конкретизировать белок вплоть до сплайс-формы. Примером тому являются семь триптических пептидов, которые картировались на пять сплайс-форм белковой группы ацил-КоА-тиоэстеразы (O00154, O00154-4, O00154-5, O00154-6, O00154-7). Сопоставление молекулярных масс и значений изоэлектрических точек, рассчитанных с помощью инструмента Pyteomics [167] для каждой из возможных СО-форм, с параметрами ячейки показало, что из пяти возможных вариантов только у одного (O00154-7) параметры массы и заряда попадают внутрь границ ячейки, а потому мы предполагаем, что в данной ячейке детектирована именно седьмая сплайс-форма белка O00154 с молекулярным весом 41 кДа.

Аналогично, из двух протеоформ фактора процессинга пре-мРНК (каноническая протеоформа Q9UMS4 и вариант, в котором фрагмент последовательности из 230 а.к. заменен на фрагмент из 36 а.к.) физико-химическим параметрам ячейки удовлетворяет только «короткая» протеоформа.

В сложном по молекулярному составу биологическом образце зачастую не удается надежно различить между собой одноаминокислотные полиморфизмы, входящие в состав белковой группы. В общем случае, если не рассматривать полиморфизмы, которые критическим образом влияют на стабильность, конформацию и электрофоретическую подвижность белковой молекулы [92], выявление ОАП по положению в довольно крупной ячейке (около 0,7 пунктов pI и 10 тыс. Да по MW) на 2DE-геле не представляется возможным: изменения по массе и заряду между модифицированной и канонической протеоформой слишком малы.

Для аннотации протеоформ, обнаруженных как в единственной ячейке, так и сразу в нескольких ячейках геля, был разработан следующий алгоритм (рис. 23). Стартовой точкой работы алгоритма являлся перечень детектированных протеоформ и белковых групп, объединяющих несколько форм белков одного гена. Для каждой отдельной протеоформы и протеоформы, входящей в состав белковой группы, были сопоставлены теоретические и экспериментальные координаты локализации на геле. Если экспериментальные MW и pI совпадали с рассчитанными значениями, протеоформу считали аннотированной. В противном случае рассматривались возможные причины, которые могли привести к сдвигу: химические и посттрансляционные модификации, изомеризация, ко-миграция с белковыми партнерами, протеолиз, деградация, особенности аминокислотного состава белка, артефакты технического характера.

В результате, в 6 649 белковых группах, состоящих из белковых продуктов одного белок-кодирующего гена, удалось идентифицировать 751 протеоформ. Стоит отметить, что 10% данных групп содержали последовательности с ОАП, которые можно отличить в случае детекции пептида с заменой. Для 301 белковой группы, объединяющей белковые продукты разных генов удалось различить 25 протеоформ.

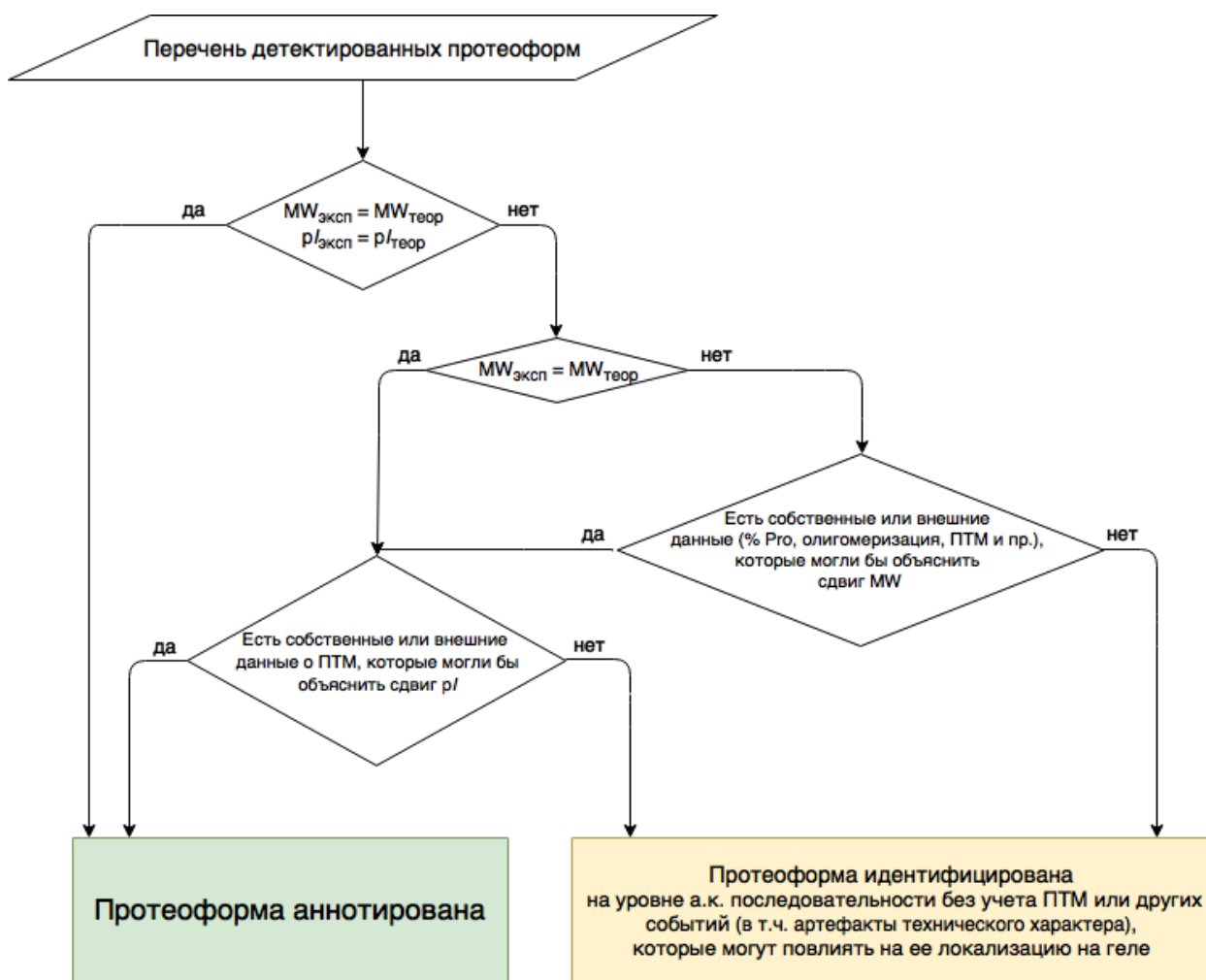


Рисунок 23. Блок-схема алгоритма, реализованного для аннотации протеоформ, детектированных после фракционирования на 2DE-геле с использованием панорамного масс-спектрометрического анализа. Масс-спектрометрическая детекция протеотипических пептидов использовалась для идентификации протеоформы. Протеоформа считалась аннотированной, если одновременно 1) были надежно детектированы протеотипические пептиды, однозначно характеризующие протеоформу, и 2) совпадали теоретические и экспериментально рассчитанные координаты локализации протеоформы на геле.

С использованием данного алгоритма суммарно для клеток НерG2 было аннотировано 2 358 протеоформ 1904 генов, в их числе 1 658 канонических последовательностей, 224 сплайс-опосредованных протеоформы и 115 случаев аминокислотных замен, а также 31 протеоформа, в которой одноаминокислотная замена произошла в сплайс-варианте. Для 990 протеоформ на основании результатов масс-спектрометрического анализа и данных 2DE идентифицированы посттрансляционные модификации (табл. 4).

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 [198].

Таблица 4. Протеом клеточной линии НерG2, детектированный с применением комбинации двумерного гель-электрофореза и панорамной масс-спектрометрии

Тип протеоформы	Совпад. теор. и эксп. pI и MW		Химическая модификация		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	10	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 им БКГ

Итоги анализа протеома клеток НерG2 методом 2DE с последующей масс-спектрометрией были сопоставлены с результатом панорамного МС-исследования тех же клеток без фракционирования. Показано, что фракционирование с помощью 2DE, предваряющее масс-спектрометрический анализ, позволяет увеличить количество идентифицированных протеоформ. При включении 2DE-фракционирования в протеомный эксперимент число идентификаций возросло на 76% сразу в силу двух причин: во-первых, за счет снижения «сложности» анализируемой смеси, а, во-вторых, за счет дополнительных сведений о физико-химических параметрах целевых молекул, которые позволяют анализировать вектор смещения области локализации протеоформы на геле.

В соответствии с собственными транскриптомными данными и формулами оценки размера протеома, допускающими возникновение ПТМ и ОАП как в канонических, так и в сплайс-вариантах [3], количество протеоформ клетках линии

НерG2 может варьировать от 124 до 306 тыс. Сочетание двумерного гелевого электрофореза и панорамной масс-спектрометрии позволило детектировать около 3-7% теоретически возможного многообразия протеоформ клеточной линии НерG2. При этом было экспериментально подтверждено, что причины возникновения протеоформ (альтернативный сплайсинг, одноаминокислотные замены и посттрансляционные модификации) появляются независимо друг от друга.

Скромное по сравнению с числом экспрессирующихся транскриптов количество детектированных белковых вариантов наглядно иллюстрирует сложности, с которыми сталкиваются исследователи, работающие в области транскриптопротеомики и протеогеномики и, в особенности, фокусирующие свое внимание на протеоформах. Есть ряд как биологических, так и исключительно технических причин, ограничивающих шансы на детекцию отличной от канонической формы белка.

Во-первых, следует отметить комбинаторную природу референсной библиотеки, построенной по результатам транскриптомных экспериментов. Созданная библиотека учитывает все возможные варианты изменений аминокислотной последовательности. Так, например, если один из фрагментов сиквенса РНК содержит несинонимичный полиморфизм и при этом картируется как на конститутивный, так и на сплайс-транскрипт, то в библиотеку войдут оба варианта: сплайс-форма с ОАП и каноническая форма с ОАП.

Во-вторых, не все варианты транскриптов транслируются в белки: например, попадание стоп-кодона на альтернативно сплайсированный фрагмент приведет к деградации более короткого транскрипта посредством нонсенс-опосредованного распада – механизма предотвращения клетки от синтеза усеченных белков с потенциально вредоносными функциями. Не исключена возможность такого сплайс-события, как удержанный интрон, который не покидает ядро и, соответственно, не подвергается трансляции. Другая причина кроется в разнице кинетических характеристик транскрипции и трансляции: когда транскрипты уже присутствуют на уровне РНК, но еще не успели транслироваться в белки на момент изоляции РНК или белка. Протеоформы могут не подвергаться детекции,

поскольку зачастую они обладают меньшей стабильностью по сравнению с каноническим вариантом или присутствуют в крайне низких концентрациях.

Для надежного обнаружения протеоформы необходим протеотипический пептид, обладающий физико-химическими свойствами, подходящими для воспроизводимой масс-спектрометрической детекции. В связи с этим особые сложности наблюдаются при детекции укороченных сплайс-форм, из последовательности которых вычленен фрагмент.

Степень покрытия протеома не является константой. Доля идентифицированных протеоформ среди всего закодированного геномом многообразия находится в зависимости от чувствительности и специфичности протеомных технологий. Прогресс аналитических методов естественным образом позволит обнаружить больше белковых продуктов. Фракционирование с помощью 2DE, предваряющее масс-спектрометрический анализ, позволяет заметно повысить чувствительность и специфичность протеомного исследования. Двумерный гель-электрофорез не только снижает сложность анализируемого образца, но и предоставляет исследователю дополнительные возможности для конкретизации протеоформы в соответствии с ее физико-химическими свойствами. Массивы результатов протеомных экспериментов цементируют новую парадигму системной биологии, которая принципиально отличается от классической догмы «один ген – один белок» [199]. В свете новой парадигмы биологические процессы представляются как интерактомные сети, в которых взаимодействия «моделируют» функцию каждой участвующей в сети протеоформы. Так, вероятно, следующим шагом в мультиомных исследованиях станет переход от каталогизации белковых молекул, обнаруживаемых в различных биообразцах и при различных условиях, к определению их функций с помощью прогрессивных аналитических инструментов и протеомных карт, специфичных для конкретных тканей, органов и целых организмов.

3.5. Протеоформы клеточной линии HepG2, ассоциированные с опухолевыми процессами

Клеточные линии – модельные системы для *in vitro* изучения молекулярных механизмов возникновения и развития онкологических патологий через призму биологической гетерогенности первичных опухолей.

Клеточная линия HepG2, протеом которой был исследован в рамках данной работы, была выделена из опухоли печени пятнадцатилетнего юноши почти 40 лет назад. Тем не менее, до недавнего времени не существовало единого мнения по поводу природы этой линии, несмотря на ее активные исследования: почти 8 тыс. публикаций в ресурсе PubMed относят линию HepG2 к гепатоцеллюлярной карциноме, около 700 публикаций – к гепатобластоме. Расследование, проведенное под руководством доктора Lopez-Terrada [200], потребовавшее в т.ч. обращения к гистопатологическим данным первичного образца, выявило, что линия HepG2 является примером гепатобластомы, а не более агрессивной гепатокарциномы, как это считалось ранее. Различия в природе и функционировании линий гепатобластомы и гепатокарциномы были показаны при сопоставлении их метаболомных профилей [201] и паттернов экспрессии белков [202].

Массив генов, белковые продукты которых мы обнаружили в клеточной линии HepG2, был проанализирован с помощью веб-сервиса GOrilla. Сервис позволяет выявлять значимые характеристики набора исследуемых белков в соответствии с терминами Gene Ontology (GO) в трех категориях: молекулярной функции, участия в биологических процессах и локализации в клетке. Согласно проведенной GO-аннотации, основными функциями белков HepG2 являются связывание (с белками, рибонуклеотидами, нуклеозидтрифосфатами и др.) и катализ (с фокусом на АТФазы, лигазы, изомеразы и хеликазы). Эти функции типичны для «бессмертной» клеточной линии [203]. Преимущественно, обнаруженные белковые продукты HepG2 вовлечены в клеточные метаболические процессы и сплайсинг пре-мРНК. Привлекает внимание факт, что локализация обнаруженных нами протеоформ равномерно распределяется между компартментами клетки, что характерно для бластом, к которым относится линия HepG2 [204]. Это существенным образом отличается от характера распределения для карцином, где аберрантные белки локализованы преимущественно в ядре.

Для выявления протеоформ клеточной линии гепатобластомы, ассоциированных с онкологическими процессами, мы проанализировали полученный перечень белков с использованием сведений ресурса TumorPortal [169]. Этот ресурс содержит данные о 320 генах, выявленных в результате сопоставления пар экзотов здоровых тканей и почти 5 тыс. опухолевых образцов, соответствующих 21 типу раковых патологий. Белковые продукты всех 320 генов, связанных с раком, 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 онкоассоциированных генов в случайных выборках генов человека такого же размера: $1,5 \pm 0,2\%$ (для 1000 выборок с $p\text{-value} < 0,0001$). Значительное обогащение линии HepG2 белками, гены которых ассоциированы с опухолевыми процессами, дополнительно подчеркивает, что данная линия получена из неконтролируемо пролиферирующего очага гепатобластомы [204].

Среди 320 генов, аннотированных в ресурсе TumorPortal, для 25 генов было надежно детектировано 27 протеоформ (табл. 5), а остальные 295 генов были выявлены в виде белковых групп, содержащих неразличимые по набору пептидов соответствующие протеоформы. В среднем, каждая протеоформа была идентифицирована по семи триптическим пептидам, что составило примерно 22% от всей аминокислотной последовательности. Анализ физико-химических свойств 27 протеоформ онкоассоциированных генов показал, что распределение детектированных протеоформ по молекулярному весу и изоэлектрической точке повторяет распределение всех транскрибируемых протеоформ клеточной линии HepG2. Большинство детектированных белковых продуктов обладают средними по протеому человека характеристиками pI и MW . Исключение составили четыре массивных белка с $MW > 100$ кДа (продукты генов *ATM*, *IPO7*, *MTOR*, *SMC1A*) и один ультра-щелочной рибосомальный белок P62841.

Для 21 онкоассоциированного гена рассматриваемой выборки была детектирована протеоформа с канонической аминокислотной последовательностью, три гена представлены последовательностями с одноаминокислотными заменами и три гена – сплайс-вариантами (табл. 5).

Таблица 5. Характеристики детектированных в клеточной линии HepG2 протеоформ, которые кодируются значительно мутированными в опухолевых образцах генами (согласно репозиторию Tumor Portal [169])

Ген	Детектированная протеоформа (AC UniProt)	MW, кДа/pI протеоформы	Количество детектированных пептидов	Покрытие а.к. последовательности, %
ATM ^{ИО}	Q13315:V:2906:I	351/6,39	3	1
AZGP1	P25311	34/5,71	4	18
CBFB ^{ИО}	Q13951	22/6,23	4	22
	Q13951-2 (деамидирован)	22/5,47	6	35
CDK4 ^{ИО}	P11802	34/6,51	11	44
CCND1 ^{ИО}	P24385	34/4,97	2	10
DDX5	P17844 (деамидирован)	69/9,06	24	40
EIF2S2	P20042	38/5,61	3	12
ELOC ^{ИО}	Q15369	12/4,74	6	63
EZR	P15311	69/5,94	2	3
FUBP1 ^{ИО}	Q96AE4	68/7,18	18	35
	Q96AE4-2	69/6,85	18	35
GOT1	P17174	46/6,53	6	23
IDHC ^{ИО}	O75874:I:367:V	47/6,53	20	56
IPO7	O95373	120/4,69	3	3
KRAS ^{ИО}	P01116-2	21/8,24	4	24
MTOR ^{ИО}	P42345	289/6,73	6	4
PCBP1 ^{ПК}	Q15365	37/6,66	4	15
PCK1 ^{ПК}	P35558:V:184:L (амидирован)	69/5,79	12	24
PDSS2	Q86YN6	44/8,83	6	17
RAD21 ^{ПК}	O60216	72/4,54	9	18
RHEB ^{ПК}	Q15382	21/5,65	2	13
RPS15	P62841	17/10,39	4	39
PPP2R1A ^{ИО}	P30153	65/5,01	3	7
SMAD4 ^{ИО}	Q13485	60/6,49	2	3
SMC1A ^{ИО}	Q14683	143/7,51	9	12
TBL1XR1 ^{ИО}	Q9BZK7	56/5,28	12	23

^{ИО} – гены, для которых подтверждена связь с онкологическим процессом;

^{ПК} – перспективный кандидат в онкоассоциированные гены.

В числе трех детектированных сплайс-вариантов была гуанозинтрифосфотаза P01116-2, кодируемая протоонкогеном *KRAS* [205]. Данный ген является регулятором сигнальной системы клетки, функция которого заключается в контроле клеточной пролиферации. Нарушение механизма

пролиферации в большинстве случаев приводит к развитию опухолевых процессов, исходной причиной которого может служить множественные мутации в гене *KRAS*.

Для гена *KRAS* известно две сплайс-формы: канонический вариант K-Ras4A (P0116-2) и альтернативный – K-Ras4B (P01116-2), отличающиеся по своим функциям. Каноническая форма активизирует апоптоз, а сплайс-вариант, напротив, его ингибирует; гиперэкспрессия K-Ras4A ассоциирована с лучшей выживаемостью, а гиперэкспрессию K-Ras4B связывают с увеличением размера опухоли [206]. При этом, разница между протеоформами заключается в том, что K-Ras4B в отличие от K-Ras4A лишена «якорной» изопреновой группы в С-концевой части белковой последовательности, отвечающей за прикрепление белка к клеточной мембране. В результатах транскриптомного профилирования клеточной линии HepG2 были выявлены оба варианта последовательностей, кодируемых протоонкогеном *KRAS* [205], 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. Стоит отметить, что масс-спектрометрическая детекция канонического варианта затруднена в силу слишком коротких триптических пептидов, характерных для K-Ras4A, которые могут использоваться в качестве надежного «прототипа» белка.

Одноаминокислотные замены, обнаруженные в белках онкоассоциированных генов, вероятно, носят нейтральный характер. Мутации произошли между близкими по размеру неполярными аминокислотами: к примеру, в серин-треонин-киназе Q13315 валин в положении 2906 был замещен изолейцином. В работе Цуканова и др. патогенные замены этого белка были обнаружены при исследовании рака молочной железы (L1420F, D298H, D1853V, S49C), в связи с восприимчивостью опухоли к химиотерапии [207]. Этот пример подчеркивает специфичность паттерна аминокислотных замен в каждом типе биоматериала и в каждом индивидуальном исследуемом образце.

В пептидах трех протеоформ (P17844, Q13951-2 и P35558:V:184:L) были выявлены посттрансляционные модификации: деамидирование остатков аспарагина и глутамина и амидирование С-концевых остатков. Однако, о взаимосвязи таких

ПТМ с нарушением и/или изменением биологических процессов на данный момент не известно.

Безусловно, текущий уровень протеомных технологий ограничивает исследование всех вариантов изменений и модификаций аминокислотной последовательности белков, что выражается в низкой информативности современных моделей молекулярных процессов. Так, из более 20 тыс. генов только 320 ассоциированы с опухолевыми процессами, при этом, выявление таких генов было выполнено на основе геномных данных [169]. Функциональная аннотация протеоформ является необходимым условием понимания происходящих в клетке событий, где важным шагом является биосинтез протеоформ.

Как было показано в данной работе, комбинация транскриптомных и протеомных данных – это инструмент создания белкового портрета, учитывающего уникальность молекулярного профиля исследуемого объекта. Сопоставление результатов транскриптомно-протеомного профилирования опухолевого образца с результатами персонализированного исследования нормальных и пораженных другими патологиями тканей позволит выявить аномальные каскады реакций и биомаркеры, а также определить мишени для терапевтического вмешательства [208].

ЗАКЛЮЧЕНИЕ

Высокопроизводительные технологии исследования нуклеотидных и аминокислотных последовательностей с высокой скоростью наполняют репозитории данными о случаях нарушения конститутивного сплайсинга, заменах единичных нуклеотидов или аминокислот и посттрансляционных модификациях. Удельная значимость каждой следующей обнаруженной мутации важна для понимания механизмов функционирования живых систем и роли конкретной aberrации в возникновении и развитии патологии. Отсутствие однозначных критериев вредоносности молекулярных событий отягощается запутанностью биологических процессов и сложной многофакторной природой заболеваний. Терабайты данных, относящихся к aberrантным последовательностям, естественным образом объясняются самой природой заболеваний, в особенности онкологических. Гетерогенность каждой опухоли достигается в ходе спонтанного мутагенеза под влиянием индивидуальных для каждого случая факторов, что существенно осложняет экстраполирование единичных артефактов на популяцию. Выявление aberrантных протеоформ требует представительной выборки биообразцов, а также технологий, которые позволят отличать значимые молекулярные события от статистического шума [209].

Высокая информативная ценность протеомных исследований, предоставляющих уникальные сведения о реализовавшемся, а не вероятностном статусе генетической информации исследуемого образца, зачастую компенсируется сложностью экспериментов в силу ряда факторов. Среди них нестабильность белковых молекул, усугубляемая на каждом этапе лабораторного исследования, динамическая природа и разнообразие протеоформ, определяющая огромное информационное пространство протеома, отсутствие эквивалента ПЦР, позволившего амплифицировать целевую белковую молекулу, недостаточная чувствительность аналитических приборов, не покрывающая весь динамический диапазон белков в сложных биологических образцах.

Тем не менее, использование транскриптомно-протеомного подхода для анализа состава протеома позволяет минимизировать влияние некоторых факторов. В данной работе мы показали, что использование характерной для объекта исследования библиотеки аминокислотных последовательностей в качестве

референса позволяет сфокусироваться на наиболее вероятных для конкретного образца протеоформах. Выполнение протеомного профилирования путем фракционирования белковой смеси с помощью двумерного гель-электрофореза, разделение получаемого геля на секторы и их дальнейший анализ с помощью масс-спектрометрии позволяет выявлять в разы больше протеоформ, чем при использовании только одного из методов анализа. Интерпретация получаемых данных позволяет не только выявлять случаи одноаминокислотных замен, вариантов альтернативного сплайсинга и посттрансляционных модификаций, а также их комбинации, но и получать данные о взаимодействиях белков.

Мы полагаем, что эффективность анализа результатов крупных протеомных экспериментов будет максимальной при выполнении ряда условий, которые послужили отправными точками при выполнении данной работы. Во-первых, необходимо использовать адаптированную к объекту исследования библиотеку аминокислотных последовательностей в качестве референса. Такая библиотека даст возможность сфокусироваться на наиболее вероятных для конкретного образца протеоформах и учесть аберрантные варианты. Во-вторых, автоматизация процесса анализа экспериментальных результатов с возможностью обращения к интуитивно понятному графическому интерфейсу позволит исследователю обработать большой массив информации и при этом подробно изучить случаи идентификации аберрантных протеоформ. В-третьих, использование сведений о возможных вариантах из геномных, транскриптомных и протеомных ресурсов и консолидация экспериментальных результатов, полученных несколькими методами, обеспечит более полный протеомный портрет исследуемого образца.

ВЫВОДЫ

1. По результатам транскриптного профилирования клеточной линии HepG2 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 HepG2 методом двумерного гелевого 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 HepG2 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.

СПИСОК ЛИТЕРАТУРЫ

1. Smith L.M., Kelleher N.L. Proteoform: a single term describing protein complexity. // Nat. Methods. 2013. Vol. 10, № 3. P. 186–187.
2. Ezkurdia I. et al. Most highly expressed protein-coding genes have a single dominant isoform // J. Proteome Res. 2015. Vol. 14, № 4. P. 1880–1887.
3. Ponomarenko E.A. et al. The size of the human proteome: the width and depth. // Int. J. Anal. Chem. 2016. Vol. 2016. P. 7436849.
4. Hanahan D., Weinberg R.A. Hallmarks of cancer: the next generation // Cell. 2011. Vol. 144, № 5. P. 646–674.
5. Martincorena I., Campbell P.J. Somatic mutation in cancer and normal cells // Science. 2015. Vol. 349, № 6255. P. 1483–1489.
6. Kim M.-S. et al. A draft map of the human proteome // Nature. 2014. Vol. 509, № 7502. P. 575–581.
7. Wilhelm M. et al. Mass-spectrometry-based draft of the human proteome. // Nature. 2014. Vol. 509, № 7502. P. 582–587.
8. Archakov A. et al. Chromosome-centric approach to overcoming bottlenecks in the Human Proteome Project. // Expert Rev. Proteomics. 2012. Vol. 9, № 6. P. 667–676.
9. Lisitsa A. V et al. The Width of the Human Plasma Proteome Compared With a Cancer Cell Line and Bacteria. 2015. Vol. 4, № 4. P. 10–13.
10. Lisitsa A. et al. Profiling proteoforms: promising follow-up of proteomics for biomarker discovery. // Expert Rev. Proteomics. 2014. Vol. 11, № 1. P. 121–129.
11. Zgoda V.G. et al. Chromosome 18 transcriptome profiling and targeted proteome mapping in depleted plasma, liver tissue and HepG2 cells // J. Proteome Res. 2013. Vol. 12, № 1. P. 123–134.
12. Naryzhny S.N. et al. Combination of virtual and experimental 2DE together with ESI LC-MS/MS gives a clearer view about proteomes of human cells and plasma. // Electrophoresis. 2016. Vol. 37, № 2. P. 302–309.
13. Deutsch E.W. et al. Human Proteome Project mass spectrometry data interpretation guidelines. // J. Proteome Res. 2016. Vol. 15, № 11. P. 3961–3970.
14. Nesvizhskii A.I. Proteogenomics: concepts, applications and computational strategies // Nat. Methods. 2014. Vol. 11, № 11. P. 1114–1125.
15. 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. // Nature. 2010. Vol. 467, № 7319. P. 1061–1073.
16. Zhang Z. et al. Analyzing effects of naturally occurring missense mutations // Comput. Math. Meth. Med. 2012. Vol. 2012. P. 805827.
17. Fujiwara H. et al. A novel V59E missense mutation in the sodium iodide symporter gene in a family with iodide transport defect iodide. // Thyroid. 2000. Vol. 10, № 6. P. 471–474.
18. Wang Z., Moulton J. SNPs, protein structure, and disease // Hum. Mutat. 2001. Vol. 17, № 4. P. 263–270.
19. Vogelstein B. et al. Cancer genome landscapes. 2013. Vol. 339, № 6127. P. 1546–1558.
20. Abhishek N., Vihinen M. Harmful somatic amino acid substitutions affect key pathways in cancers // BMC Med. Genomics. BMC Medical Genomics, 2015. Vol. 8, № 53. P. 1–12.
21. Kutlar A. Sick cell disease: a multigenic perspective of a single gene disorder // Hemoglobin. 2007. Vol. 31, № 2. P. 209–224.
22. Hart J.R. et al. The butterfly effect in cancer : a single base mutation can remodel the cell // Proc. Natl. Acad. Sci. U.S.A. 2014. Vol. 112, № 4. P. 1131–1136.
23. Wang E.T. et al. Alternative isoform regulation in human tissue transcriptomes. // Nature. 2008. Vol. 456, № 7221. P. 470–476.
24. Tress M.L., Abascal F., Valencia A. Alternative splicing may not be the key to proteome complexity // Trends Biochem. Sci. 2017. Vol. 42, № 2. P. 98–110.

25. Buljan M. et al. Tissue-specific splicing of disordered segments that embed binding motifs rewires protein interaction networks. // *Mol. Cell*. 2012. Vol. 46, № 6. P. 871–883.
26. Melamud E., Moult J. Stochastic noise in splicing machinery. // *Nucleic Acids Res.* 2009. Vol. 37, № 14. P. 4873–4886.
27. Lacovich V. et al. Tau isoforms imbalance impairs the axonal transport of the amyloid precursor protein in human neurons. // *J. Neurosci.* 2017. Vol. 37, № 1. P. 58–69.
28. Bonomi S. et al. Oncogenic alternative splicing switches: role in cancer progression and prospects for therapy. // *Int. J. Cell Biol.* 2013. Vol. 2013.
29. Wei J., Zaika E., Zaika A. P53 family: Role of protein isoforms in human cancer. // *J. Nucleic Acids*. 2012. Vol. 2012. P. 687359.
30. Marabesea M. et al. Expression levels of p53 and p73 isoforms in stage I and stage III ovarian cancer. // *Eur. J. Cancer*. 2011. Vol. 4, № 164. P. 131–141.
31. Poulikakos P.I. et al. RAF inhibitor resistance is mediated by dimerization of aberrantly spliced BRAF(V600E). // *Nature*. 2011. Vol. 480, № 7377. P. 387–390.
32. Walsh C.T. Posttranslational Modification of Proteins. Expanding Nature's Inventory by Roberts & Co, 2005. 576 p.
33. Anderson N.L. The clinical plasma proteome: a survey of clinical assays for proteins in plasma and serum. // *Clin. Chem.* 2010. Vol. 56, № 2. P. 177–185.
34. Weissman A.M. Themes and variations on ubiquitylation. // *Nat Rev Mol Cell Biol.* 2001. Vol. 2, № 3. P. 169–178.
35. Thomas S.N., Cripps D., Yang A.J. Proteomic analysis of protein phosphorylation and ubiquitination in Alzheimer's disease. // *Meth. Mol. Biol.* 2009. Vol. 566. P. 109–121.
36. Pagel O. et al. Current strategies and findings in clinically relevant post-translational modification-specific proteomics. // *Expert Rev. Proteomics*. 2015. Vol. 12, № 3. P. 235–253.
37. Haas-Kogan D.A. et al. Epidermal growth factor receptor, protein kinase B/Akt, and glioma response to erlotinib. // *J. Natl. Cancer Inst.* 2005. Vol. 97, № 12. P. 880–887.
38. Benne R. et al. Major transcript of the frameshifted coxII gene from trypanosome mitochondria contains four nucleotides that are not encoded in the DNA. // *Cell*. 1986. Vol. 46, № 6. P. 819–826.
39. Athanasiadis A., Rich A., Maas S. Widespread A-to-I RNA editing of Alu-containing mRNAs in the human transcriptome. // *PLoS Biol.* 2004. Vol. 2, № 12. P. e391.
40. Anant S., Davidson N.O. Molecular mechanisms of apolipoprotein B mRNA editing. // *Curr. Opin. Lipidol.* 2001. Vol. 12, № 2. P. 159–165.
41. Tariq A., Jantsch M.F. Transcript diversification in the nervous system: A to I RNA editing in CNS function and disease development. // *Front. Neurosci.* 2012. Vol. 6. P. 99.
42. Beghini A. et al. RNA hyperediting and alternative splicing of hematopoietic cell phosphatase (PTPN6) gene in acute myeloid leukemia. // *Hum. Mol. Genet.* 2000. Vol. 9, № 15. P. 2297–2304.
43. Ranieri M. et al. Mitochondrial fusion proteins and human diseases. // *Neurol. Res. Int.* 2013. Vol. 2013. P. 293893.
44. Modi H. et al. Role of BCR/ABL gene-expression levels in determining the phenotype and imatinib sensitivity of transformed human hematopoietic cells. // *Blood*. 2007. Vol. 109, № 12. P. 5411–5421.
45. Baldo B.A. Chimeric fusion proteins used for therapy: indications, mechanisms, and safety. // *Drug Saf.* 2015. Vol. 38, № 5. P. 455–479.
46. Cobb B.A., Petrash J.M. Structural and functional changes in the alpha A-crystallin R116C mutant in hereditary cataracts. // *Biochemistry*. 2000. Vol. 39, № 51. P. 15791–15798.
47. Fiorelli T., Kirouac L., Padmanabhan J. Altered processing of amyloid precursor protein in cells undergoing apoptosis. // *PLoS One*. 2013. Vol. 8, № 2. P. e57979.
48. Iwawaki T., Oikawa D. The role of the unfolded protein response in diabetes mellitus. // *Semin. Immunopathol.* 2013. Vol. 35, № 3. P. 333–350.

49. Schmidt B. et al. A novel amino acid modification in sulfatases that is defective in multiple sulfatase deficiency. // *Cell*. 1995. Vol. 82, № 2. P. 271–278.
50. Киселева О.И., Лисица А.В., Поверенная Е.В. Протеоформы: методы исследования и клинические перспективы. // *Молекулярная биология*. 2018. Т. 52, № 3. С. 1–17.
51. Planchard D. Identification of driver mutations in lung cancer : first step in personalized cancer. // *Target. Oncol.* 2013. Vol. 8. P. 3–14.
52. Mayer G., Müller J., Lünse C.E. RNA diagnostics: real-time RT-PCR strategies and promising novel target RNAs. // *Wiley Interdiscip. Rev. RNA*. 2011. Vol. 2, № 1. P. 32–41.
53. Kondo T. Casting doubt on the traditional approach of cancer biomarker discovery through proteomics. // *Expert Rev. Proteomics*. 2014. Vol. 11, № 1. P. 9–12.
54. Davis-taber R. et al. Molecular characterization of human SUR2-containing K channels ATP. // *Cloning*. 2000. Vol. 256. P. 261–270.
55. Suga K. et al. Correlation between transcriptional expression of survivin isoforms and clinicopathological findings in human colorectal carcinomas. // *Oncol. Rep.* 2005. Vol. 13, № 5. P. 891–897.
56. Chen X., Sullivan P.F. Single nucleotide polymorphism genotyping: biochemistry, protocol, cost and throughput. // *Pharmacogenomics J.* 2003. Vol. 3, № 2. P. 77–96.
57. Tsoi D.T. et al. Cost-effectiveness analysis of recurrence score-guided treatment using a 21-gene assay in early breast cancer. // *Oncologist*. 2010. Vol. 15, № 5. P. 457–465.
58. Kerr G. et al. Techniques for clustering gene expression data. // *Comput. Biol. Med.* 2008. Vol. 38, № 3. P. 283–293.
59. Johnson J.M. et al. Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. // *Science*. 2003. Vol. 302, № 5653. P. 2141–2144.
60. Wang D.G. et al. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. // *Science*. 2012. Vol. 280, № 5366. P. 1077–1082.
61. Kennedy G.C. et al. Large-scale genotyping of complex DNA. // *Nat. Biotech.* 2003. Vol. 21, № 10. P. 1233–1237.
62. Beroukhim R. et al. Inferring loss-of-heterozygosity from unpaired tumors using high-density oligonucleotide SNP arrays. // *PLoS Comp. Biol.* 2006. Vol. 2, № 5. P. e40.
63. Xi L. et al. Whole genome exon arrays identify differential expression of alternatively spliced, cancer-related genes in lung cancer. // *Nucleic Acids Res.* 2008. Vol. 36, № 20. P. 6535–6547.
64. Gardina P.J. et al. Alternative splicing and differential gene expression in colon cancer detected by a whole genome exon array. // *BMC Genomics*. 2006. Vol. 7, № 325. P. 1–18.
65. Royce T.E., Rozowsky J.S., Gerstein M.B. Toward a universal microarray: prediction of gene expression through nearest-neighbor probe sequence identification. // *Nucleic Acids Res.* 2007. Vol. 35, № 15. P. 1–10.
66. Maekawa S. et al. RNA sequencing: from sample preparation to analysis. // *Meth. Mol. Biol.* 2014. Vol. 1164. P. 51–65.
67. The Cancer Genome Atlas Research Network. Comprehensive molecular portraits of human breast tumours. // *Nature*. 2012. Vol. 490, № 7418. P. 61–70.
68. Fox E.J. et al. Accuracy of Next Generation Sequencing Platforms. // *Next Gener. Seq. Appl.* 2014. Vol. 1. P. 1000106.
69. Shah S.P. et al. The clonal and mutational evolution spectrum of primary triple-negative breast cancers. // *Nature*. 2012. Vol. 486, № 7403. P. 395–399.
70. Shargunov A. V et al. Tissue-specific alternative splicing analysis reveals the diversity of chromosome 18 transcriptome. // *J. Proteome Res.* 2014. Vol. 13, № 1. P. 173–182.
71. Chen R. et al. Personal omics profiling reveals dynamic molecular and medical phenotypes // *Cell*. 2013. Vol. 148, № 6. P. 1293–1307.
72. Wu J., Zeng R. Molecular basis for population variation: from SNPs to SAPs // *FEBS Lett.* 2012. Vol. 586, № 18. P. 2841–2845.
73. Takeshita H. et al. Nonsynonymous single-nucleotide polymorphisms of the human

- apoptosis-related endonuclease-DNA fragmentation factor beta polypeptide, endonuclease G, and Flap endonuclease-1-genes show a low degree of genetic heterogeneity. // *DNA Cell Biol.* 2012. Vol. 31, № 1. P. 36–42.
74. Fortelny N. et al. Can we predict protein from mRNA levels? // *Nature.* 2017. Vol. 547, № 7664. P. 19–20.
 75. Gonzalez C. et al. Ribosome profiling reveals a cell-type-specific translational landscape in brain tumors. // *J. Neurosci.* 2014. Vol. 34, № 33. P. 10924–10936.
 76. Koch A. et al. A proteogenomics approach integrating proteomics and ribosome profiling increases the efficiency of protein identification and enables the discovery of alternative translation start sites. // *Proteomics.* 2014. Vol. 14. P. 2688.
 77. Ruggles K. V. et al. An analysis of the sensitivity of proteogenomic mapping of somatic mutations and novel splicing events in cancer. // *Mol. Cell. Proteomics.* 2015. P. M115.056226.
 78. Frampton J.P. et al. Aqueous two-phase system patterning of detection antibody solutions for cross-reaction-free multiplex ELISA. // *Sci. Rep.* 2014. Vol. 4. P. 871–874.
 79. Tighe P.J. et al. ELISA in the multiplex era: potentials and pitfalls. // *Proteomics Clin. Appl.* 2015. Vol. 9, № 3–4. P. 406–422.
 80. Maruyama K. et al. ELISA-based detection system for protein S K196E mutation, a genetic risk factor for venous thromboembolism. // *PLoS One.* 2015. Vol. 10, № 7. P. 1–10.
 81. Romero X., Cañete J.D., Engel P. Determination of soluble tumor necrosis factor receptor 2 produced by alternative splicing. // *Methods Mol. Biol.* 2014. Vol. 1155. P. 187–199.
 82. Sheehan K.M. et al. Use of reverse phase protein microarrays and reference standard development for molecular network analysis of metastatic ovarian carcinoma. // *Mol. Cell. Proteomics.* 2005. Vol. 4, № 4. P. 346–355.
 83. Petricoin E.F. et al. Mapping molecular networks using proteomics: a vision for patient-tailored combination therapy. // *J. Clin. Oncol.* 2005. Vol. 23, № 15. P. 3614–3621.
 84. Yang Y. et al. The Thr to Met substitution of amino acid 118 in hepatitis B virus surface antigen escapes from immune-assay-based screening of blood donors. // *J. Gen. Virol.* 2016. Vol. 97, № 5. P. 1210–1217.
 85. Heidegger I. et al. PSA isoforms velocities for early diagnosis of prostate cancer. // *Anticancer Res.* 2015. Vol. 35, № 6. P. 3567–3570.
 86. Juncker D. et al. Cross-reactivity in antibody microarrays and multiplexed sandwich assays: shedding light on the dark side of multiplexing. // *Curr. Opin. Chem. Biol.* 2014. Vol. 18. P. 29–37.
 87. Archakov A. et al. Biospecific irreversible fishing coupled with atomic force microscopy for detection of extremely low-abundant proteins. // *Proteomics.* 2009. Vol. 9, № 5. P. 1326–1343.
 88. Rabilloud T., Lelong C. Two-dimensional gel electrophoresis in proteomics: a tutorial. // *J. Proteomics.* 2011. Vol. 74, № 10. P. 1829–1841.
 89. Rogowska-Wrzesinska A. et al. 2D gels still have a niche in proteomics. // *J. Proteomics.* 2013. Vol. 88. P. 4–13.
 90. Arentz G. et al. State of the art of 2D DIGE. // *Proteomics Clin. Appl.* 2015. Vol. 9, № 3. P. 277–288.
 91. Chung H.Y. Effects of SNPs using differentially expressed serum proteins at growth stages on average daily gain in pig. // *Mol. Biol. Rep.* 2011. Vol. 38, № 6. P. 3777–3785.
 92. Huang H.-L. et al. Biomarker discovery in breast cancer serum using 2-D differential gel electrophoresis/ MALDI-TOF/TOF and data validation by routine clinical assays. // *Electrophoresis.* 2006. Vol. 27, № 8. P. 1641–1650.
 93. Макаров А.А. и др. Изучение изменений белкового профиля в дифференцирующихся миобластах человека // *Онтогенез.* 2009. Т. 39, № 2. С. 112–119.
 94. Lange V. et al. Selected reaction monitoring for quantitative proteomics: a tutorial. // *Mol. Syst. Biol.* 2008. Vol. 4, № 1. P. 222.

95. Whiteaker J.R. et al. A targeted proteomics-based pipeline for verification of biomarkers in plasma. // *Nat. Biotechnol.* 2011. Vol. 29, № 7. P. 625–634.
96. Paik Y.K. et al. Standard guidelines for the chromosome-centric human proteome project. // *J. Proteome Res.* 2012. Vol. 11, № 4. P. 2005–2013.
97. Su Z.-D. et al. Quantitative detection of single amino acid polymorphisms by targeted proteomics. // *J. Mol. Cell Biol.* 2011. Vol. 3, № 5. P. 309–315.
98. Wang Q. et al. Mutant proteins as cancer-specific biomarkers. // *Proc. Natl. Acad. Sci. U.S.A.* 2011. Vol. 108, № 6. P. 2444–2449.
99. Song C. et al. Large-scale quantification of single amino-acid variations by a variation-associated database search strategy. // *J. Proteome Res.* 2014. Vol. 13, № 1. P. 241–248.
100. Wu J. et al. Identification and quantification of osteopontin splice variants in the plasma of lung cancer patients using immunoaffinity capture and targeted mass spectrometry. // *Biomarkers.* 2012. Vol. 17, № 2. P. 125–133.
101. Ossola R. et al. Biomarker validation in blood specimens by selected reaction monitoring mass spectrometry of N-glycosites. // *Methods Mol. Biol.* 2011. Vol. 728. P. 179–194.
102. Lobas A.A. et al. Exome-based proteogenomics of HEK-293 human cell line: coding genomic variants identified at the level of shotgun proteome. // *Proteomics.* 2016. Vol. 16, № 14. P. 1980–1991.
103. Bantscheff M. et al. Quantitative mass spectrometry in proteomics: a critical review. // *Anal. Bioanal. Chem.* 2007. Vol. 389, № 4. P. 1017–1031.
104. Wiśniewski J.R., Zougman A., Mann M. Combination of FASP and StageTip-based fractionation allows in-depth analysis of the hippocampal membrane proteome. // *J. Proteome Res.* 2009. Vol. 8, № 12. P. 5674–5678.
105. Gatlin C.L. et al. Automated identification of amino acid sequence variations in proteins by HPLC/microspray tandem mass spectrometry. // *Anal. Chem.* 2000. Vol. 72, № 4. P. 757–763.
106. Brosch M. et al. Shotgun proteomics aids discovery of novel protein-coding genes, alternative splicing, and resurrected pseudogenes in the mouse genome. // *Genome Res.* 2011. Vol. 21, № 5. P. 756–767.
107. Cesnik A.J. et al. Human proteomic variation revealed by combining RNA-Seq proteogenomics and global post-translational modification (G-PTM) search strategy. // *J. Proteome Res.* 2016. Vol. 15, № 3. P. 800–808.
108. Zhang B. et al. Proteogenomic characterization of human colon and rectal cancer. // *Nature.* 2014. Vol. 513, № 7518. P. 382–387.
109. Lichti C.F. et al. Systematic identification of single amino acid polymorphisms in glioma stem cell-derived chromosome 19 proteins. // *J. Proteome Res.* 2015. Vol. 14. P. 778–786.
110. Liu X. et al. Constrained selected reaction monitoring: quantification of selected post-translational modifications and protein isoforms. // *Methods.* 2013. Vol. 61, № 3. P. 304–312.
111. Veenstra T.D. Where are all the biomarkers? // *Expert Rev. Proteomics.* 2011. Vol. 8, № 6. P. 681–683.
112. Мошковский С.А. и др. Идентификация единичных аминокислотных замен в протеогеномике. // *Биохимия.* 2018. Т. 83, № 3. С. 368–378.
113. Nedelkov D. et al. Investigating diversity in human plasma proteins. // *Proc. Natl. Acad. Sci. U.S.A.* 2005. Vol. 102, № 31. P. 10852–10857.
114. Ntai I. et al. Integrated bottom-up and top-down proteomics of patient-derived breast tumor xenografts. // *Mol. Cell. Proteomics.* 2016. Vol. 15, № 1. P. 45–56.
115. Kohlbacher O. et al. TOPP-the OpenMS proteomics pipeline // *Bioinformatics.* 2007. Vol. 23, № 2. P. 191–197.
116. Craig R., Beavis R.C. TANDEM: matching proteins with tandem mass spectra. // *Bioinformatics.* 2004. Vol. 20, № 9. P. 1466–1467.

- 117.Perkins D.N. et al. Probability-based protein identification by searching sequence databases using mass spectrometry data. // *Electrophoresis*. 1999. Vol. 20, № 18. P. 3551–3567.
- 118.Geer L.Y. et al. Open mass spectrometry search algorithm. // *J. Proteome Res.* 2004. Vol. 3, № 5. P. 958–964.
- 119.Nesvizhskii A.I., Aebersold R. Analysis, statistical validation and dissemination of large-scale proteomics datasets generated by tandem MS. // *Drug Discov. Today*. 2004. Vol. 9, № 4. P. 173–181.
- 120.Patterson S.D. Data analysis--the Achilles heel of proteomics. // *Nat. Biotechnol.* 2003. Vol. 21, № 3. P. 221–222.
- 121.Brosch M. et al. Comparison of Mascot and X!Tandem performance for low and high accuracy mass spectrometry and the development of an adjusted Mascot threshold. // *Mol. Cell. Proteomics*. 2008. Vol. 7, № 5. P. 962–970.
- 122.Shteynberg D. et al. Combining results of multiple search engines in proteomics. // *Mol. Cell. Proteomics*. 2013. Vol. 12, № 9. P. 2383–2393.
- 123.Wen B. et al. IPeak: An open source tool to combine results from multiple MS/MS search engines. // *Proteomics*. 2015. Vol. 15, № 17. P. 2916–2920.
- 124.Edwards N.J. PepArML: A meta-search peptide identification platform for tandem mass spectra. // *Curr. Prot. Bioinform.* 2013. Vol. 44. P.1-23.
- 125.Vaudel M. et al. PeptideShaker enables reanalysis of MS-derived proteomics data sets. // *Nat. Biotechnol.* 2015. Vol. 33, № 1. P. 22–24.
- 126.Vaudel M. et al. SearchGUI: An open-source graphical user interface for simultaneous OMSSA and X!Tandem searches. // *Proteomics*. 2011. Vol. 11, № 5. P. 996–999.
- 127.Gudbjartsson D.F. et al. Large-scale whole-genome sequencing of the Icelandic population // *Nat. Genet.* 2015. Vol. 47, № 5. P. 435–444.
- 128.Hubbard T. et al. The Ensembl genome database project. // *Nucleic Acids Res.* 2002. Vol. 30, № 1. P. 38–41.
- 129.Kodama Y. et al. The Sequence Read Archive: explosive growth of sequencing data. // *Nucleic Acids Res.* 2012. Vol. 40. P. 54-56.
- 130.Landrum M.J. et al. ClinVar: public archive of interpretations of clinically relevant variants. // *Nucleic Acids Res.* 2016. Vol. 44, № D1. P. D862-D868.
- 131.Tomczak K., Czerwińska P., Wiznerowicz M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. // *Contemp. Oncol.* 2015. Vol. 19, № 1A. P. 68-77.
- 132.Maglott D., Amberger J.S., Hamosh A. Online Mendelian Inheritance in Man (OMIM): a directory of human genes and genetic disorders // *Nucleic Acids Res.* 2002. Vol. 30, № 1. P. 1–7.
- 133.Lonsdale J. et al. The Genotype-Tissue Expression (GTEx) project // *Nat. Genet.* 2013. Vol. 45, № 6. P. 580–585.
- 134.Bhagwat M. Searching NCBI's dbSNP database. // *Curr. Prot. Bioinform.* 2010. Vol. 1. P. 1.19.
- 135.Musumeci L. et al. Single nucleotide differences (SNDs) in the dbSNP database may lead to errors in genotyping and haplotyping studies. // *Hum. Mutat.* 2010. Vol. 31, № 1. P. 67–73.
- 136.Leinonen R. et al. UniProt archive. // *Bioinformatics*. 2004. Vol. 20, № 17. P. 3236–3237.
- 137.Gaudet P. et al. The neXtProt knowledgebase on human proteins: current status. // *Nucleic Acids Res.* 2015. Vol. 43, № D1. P. D764–D770.
- 138.Uhlen M. et al. Towards a knowledge-based Human Protein Atlas // *Nat. Biotechnol.* 2010. Vol. 28, № 12. P. 1248–1250.
- 139.Farrah T. et al. PASSEL: The PeptideAtlas SRMexperiment library // *Proteomics*. 2012. Vol. 12, № 8. P. 1170–1175.
- 140.Jones P. et al. PRIDE: a public repository of protein and peptide identifications for the proteomics community. // *Nucleic Acids Res.* 2006. Vol. 34, № 90001. P. D659–D663.
- 141.Beavis R.C. Using the Global Proteome Machine for Protein Identification. // *Meth. Mol. Biol.* 2006. Vol. 328. P. 217–228.

- 142.Ning K., Nesvizhskii A.I. The utility of mass spectrometry-based proteomic data for validation of novel alternative splice forms reconstructed from RNA-Seq data: a preliminary assessment. // BMC Bioinformatics. 2010. Vol. 11. P. S14.
- 143.Cao R. et al. dbSAP: single amino-acid polymorphism database for protein variation detection. // Nucleic Acids Res. 2017. Vol. 45, № D1. P. D827–D832.
- 144.Li J. et al. A bioinformatics workflow for variant peptide detection in shotgun proteomics. // Mol. Cell. Proteomics. 2011. Vol. 10, № 5. P. M110.006536.
- 145.Kawabata T., Ota M., Nishikawa K. The Protein Mutant Database. // Nucleic Acids Res. 1999. Vol. 27, № 1. P. 355–357.
- 146.Kamath K.S., Vasavada M.S., Srivastava S. Proteomic databases and tools to decipher post-translational modifications. // J. Proteomics. 2011. Vol. 75, № 1. P. 127–144.
- 147.Lee T.-Y. et al. dbPTM: an information repository of protein post-translational modification. // Nucleic Acids Res. 2006. Vol. 34, № 90001. P. D622–D627.
- 148.Lemeer S., Heck A.J. The phosphoproteomics data explosion. // Curr. Opin. Chem. Biol. 2009. Vol. 13, № 4. P. 414–420.
- 149.Diella F. et al. Phospho.ELM: a database of experimentally verified phosphorylation sites in eukaryotic proteins. // BMC Bioinformatics. 2004. Vol. 5, № 1. P. 79.
- 150.Hornbeck P. V. et al. PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. // Nucleic Acids Res. 2012. Vol. 40, № D1. P. D261–D270.
- 151.De la Grange P. et al. A new advance in alternative splicing databases: from catalogue to detailed analysis of regulation of expression and function of human alternative splicing variants. // BMC Bioinform. 2007. Vol. 8, № 1. P. 180.
- 152.Birzele F. et al. ProSAS: a database for analyzing alternative splicing in the context of protein structures. // Nucleic Acids Res. 2008. Vol. 36. P. D63–D68.
- 153.Woo S. et al. Proteogenomic database construction driven from large scale RNA-seq data. // J. Proteome Res. 2014. Vol. 13, № 1. P. 21–28.
- 154.Baudet C. et al. Cassis: detection of genomic rearrangement breakpoints. // Bioinform. 2010. Vol. 26, № 15. P. 1897–1898.
- 155.Hernandez C., Waridel P., Quadroni M. Database construction and peptide identification strategies for proteogenomic studies on sequenced genomes. // Curr. Top. Med. Chem. 2014. Vol. 14, № 3. P. 425–434.
- 156.Edwards N.J. Novel peptide identification from tandem mass spectra using ESTs and sequence database compression. // Mol. Syst. Biol. 2007. Vol. 3. P. 102.
- 157.De Souza G.A. et al. Proteogenomic analysis of polymorphisms and gene annotation divergences in prokaryotes using a clustered mass spectrometry-friendly database. // Mol. Cell. Proteomics. 2011. Vol. 10, № 1. P. M110.002527.
- 158.Wang X. et al. Protein identification using customized protein sequence databases derived from RNA-Seq data. // J. Proteome Res. NIH Public Access, 2012. Vol. 11, № 2. P. 1009–1017.
- 159.Ponomarenko E.A. et al. Chromosome 18 transcriptoproteome of liver tissue and HepG2 cells and targeted proteome mapping in depleted plasma: update 2013 // J. Proteome Res. 2014. Vol. 13. P. 183–190.
- 160.Krasnov G.S. et al. PPLine: an automated pipeline for SNP, SAP, and splice variant detection in the context of proteogenomics. // J. Proteome Res. 2015. Vol. 14, № 9. P. 3729–3737.
- 161.Warden C.D. et al. Detailed comparison of two popular variant calling packages for exome and targeted exon studies. // PeerJ. 2014. Vol. 2. P. e600.
- 162.Klose J. Protein mapping by combined isoelectric focusing and electrophoresis of mouse tissues. A novel approach to testing for induced point mutations in mammals. // Humangenetik. 1975. Vol. 26, № 3. P. 231–243.
- 163.O'Farrell P.H. High resolution two-dimensional electrophoresis of proteins. // J. Biol. Chem.

1975. Vol. 250, № 10. P. 4007–4021.
164. Naryzhny S.N. et al. Virtual-experimental 2DE approach in Chromosome-Centric Human Proteome Project. // J. Proteome Res. 2016. Vol. 15, № 2. P. 525–530.
 165. Poverennaya E.V. et al. 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.
 166. Mellacheruvu D. et al. The CRAPome: a contaminant repository for affinity purification-mass spectrometry data. // Nat. Methods. 2013. Vol. 10, № 8. P. 730–736.
 167. Goloborodko A.A. et al. Pyteomics—a Python framework for exploratory data analysis and rapid software prototyping in proteomics. // J. Am. Soc. Mass Spectrom. 2013. Vol. 24, № 2. P. 301–304.
 168. Gene Ontology Consortium. Gene Ontology Consortium: going forward. // Nucleic Acids Res. 2015. Vol. 43. P. D1049–D1056.
 169. Lawrence M.S. et al. Discovery and saturation analysis of cancer genes across 21 tumour types. // Nature. 2014. Vol. 505, № 7484. P. 495–501.
 170. Wilhelm M. et al. Mass-spectrometry-based draft of the human proteome. // Nature. 2014. Vol. 509, № 7502. P. 582–587.
 171. Colinge J., Bennett K.L. Introduction to computational proteomics. // PLoS Comput. Biol. 2007. Vol. 3, № 7. P. e114.
 172. Hollstein M. et al. P53 mutations in human cancers. // Science. 1991. Vol. 253, № 5015. P. 49–53.
 173. Djebali S. et al. Landscape of transcription in human cells. // Nature. 2012. Vol. 489, № 7414. P. 101–108.
 174. Muñoz J., Heck A.J.R. From the human genome to the human proteome. // Angew. Chem. Int. Ed. Engl. 2014. Vol. 53, № 41. P. 10864–10866.
 175. Kim M.-S. et al. A draft map of the human proteome. // Nature. 2014. Vol. 509, № 7502. P. 575–581.
 176. Gallegos D. et al. Label-free biodetection using a smartphone. // Lab Chip. 2013. Vol. 13, № 11. P. 2124–2132.
 177. Скворцов В.С., Микурова А.В., Рыбина А.В., Применение методов *de novo* секвенирования для идентификации белков. // Биомед. Химия. 2017. Т. 63, № 4. С. 341–350.
 178. Поверенная Е. и др. Мультиомная стратегия исследования протеома клеточной линии гепатоцеллюлярной карциномы HepG2. // Биомед. Химия. 2017. Т. 63, № 5. С. 373–378.
 179. Reva B., Antipin Y., Sander C. Predicting the functional impact of protein mutations: application to cancer genomics. // Nucleic Acids Res. Vol. 39, № 17. P. e118.
 180. Cassago A. et al. Mitochondrial localization and structure-based phosphate activation mechanism of Glutaminase C with implications for cancer metabolism. // Proc. Natl. Acad. Sci. U.S.A. 2012. Vol. 109, № 4. P. 1092–1097.
 181. Petrak J. et al. Déjà vu in proteomics. A hit parade of repeatedly identified differentially expressed proteins. // Proteomics. 2008. Vol. 8, № 9. P. 1744–1749.
 182. Kozlowski L.P. Proteome-pI: proteome isoelectric point database. // Nucleic Acids Res. 2017. Vol. 45, № D1. P. D1112–D1116.
 183. Zhu K. et al. Protein pI Shifts due to Posttranslational Modifications in the Separation and Characterization of Proteins. // Anal. Chem. 2005. Vol. 77, № 9. P. 2745–2755.
 184. Ковалева М.А. и др. Определение “аминокислотных конфликтов” и аминокислотных замен в первичных структурах 41 белка человека протеомными технологиями. // Биомед. Химия. 2008. Т. 54, № 4. С. 420–434.
 185. Naryzhny S. Towards the full realization of 2DE power. // Proteomes. 2016. Vol. 4, № 4. P. 33.
 186. Matsuo K. et al. Secondary-structure analysis of denatured proteins by vacuum-ultraviolet circular dichroism spectroscopy. // Biophys. J. Biophys. Soc. 2007. Vol. 92, № 11. P. 4088–

187. Manning M., Colón W. Structural basis of protein kinetic stability: esistance to sodium dodecyl sulfate suggests a central role for rigidity and a bias toward β -sheet structure. // *Biochemistry*. 2004. Vol. 43, № 35. P. 11248–11254.
188. Prabakaran S. et al. Post-translational modification: nature's escape from genetic imprisonment and the basis for dynamic information encoding // *Wiley Interdiscip. Rev. Syst. Biol. Med.* 2012. Vol. 4, № 6. P. 565–583.
189. Yu F., Li N., Yu W. PIPI: PTM-invariant peptide identification using coding method. // *J. Proteome Res.* 2016. Vol. 15, № 12. P. 4423–4435.
190. Kuyama H. et al. A new approach for detecting C-terminal amidation of proteins and peptides by mass spectrometry in conjunction with chemical derivatization. // *Proteomics*. 2009. Vol. 9, № 16. P. 4063–4070.
191. Evich M. et al. Effect of methylation on the side-chain pK_a value of arginine. // *Protein Sci.* 2016. Vol. 25, № 2. P. 479–486.
192. Thiede B. et al. High resolution quantitative proteomics of HeLa cells protein species using stable isotope labeling with amino acids in cell culture (SILAC), two-dimensional gel electrophoresis (2DE) and nano-liquid chromatograpohy coupled to an LTQ-OrbitrapMass spectrometer. // *Mol. Cell. Proteomics*. 2013. Vol. 12, № 2. P. 529–538.
193. Nefedova V.V. et al. Structure and properties of G84R and L99M mutants of human small heat shock protein HspB1 correlating with motor neuropathy. // *Arch. Biochem. Biophys.* 2013. Vol. 538, № 1. P. 16–24.
194. Hino M. et al. Small heat shock protein 27 (HSP27) associates with tubulin/microtubules in HeLa cells. // *Biochem. Biophys. Res. Commun.* 2000. Vol. 271, № 1. P. 164–169.
195. Halligan B.D. ProMoST: a tool for calculating the pI and molecular mass of phosphorylated and modified proteins on two-dimensional gels. // *Meth. Mol. Biol.* 2009. Vol. 527. P. 283–298.
196. Stokoe D. et al. Identification of MAPKAP kinase 2 as a major enzyme responsible for the phosphorylation of the small mammalian heat shock proteins. // *FEBS Lett.* 1992. Vol. 313, № 3. P. 307–313.
197. Keyser B. et al. Disease-causing missense mutations affect enzymatic activity, stability and oligomerization of glutaryl-CoA dehydrogenase (GCDH). // *Hum. Mol. Genet.* 2008. Vol. 17, № 24. P. 3854–3863.
198. Poverennaya E. V. et al. Why are the correlations between mRNA and protein levels so low among the 275 predicted protein-coding genes on human chromosome 18? // *J. Proteome Res.* 2017. Vol. 16, № 12. P. 4311–4318.
199. Beadle G.W., Tatum E.L. Genetic control of biochemical reactions in neurospora. // *Proc. Natl. Acad. Sci. U.S.A.* 1941. Vol. 27, № 11. P. 499–506.
200. López-Terrada D. et al. HepG2 is a hepatoblastoma-derived cell line // *Hum. Pathol.* 2009. Vol. 40, № 10. P. 1512–1515.
201. Ghaffari P. et al. Identifying anti-growth factors for human cancer cell lines through genome-scale metabolic modeling. // *Sci. Rep.* 2015. Vol. 5. P. 8183.
202. Pang R.T. et al. Comparison of protein expression patterns between hepatocellular carcinoma cell lines and a hepatoblastoma cell line. // *Clin. Proteomics*. 2004. Vol. 1. T. 3-4. P. 313–331.
203. Gillet J.-P., Varma S., Gottesman M.M. The clinical relevance of cancer cell lines. // *J. Natl. Cancer Inst.* 2013. Vol. 105, № 7. P. 452–458.
204. Rikhi R.R. et al. Hepatoblastoma: a need for cell lines and tissue banks to develop targeted drug therapies. // *Front. Pediatr.* 2016. Vol. 4. P. 22.
205. Tsuchida N., Murugan A.K., Grieco M. Kirsten Ras oncogene: significance of its discovery in human cancer research. // *Oncotarget*. 2016. Vol. 7, № 29. P. 46717–46733.
206. Miura K., Fujibuchi W., Unno M. Splice isoforms as therapeutic targets for colorectal cancer. // *Carcinogenesis*. 2012. Vol. 33, № 12. P. 2311–2319.

- 207.Цуканов К.Ю. и др. Биоинформатический протокол для обработки NGS-данных и идентификации мутаций в солидных опухолях человека // Биомед. Химия. 2017. Т. 63, № 5. С. 413–417.
- 208.Aebersold R. et al. How many human proteoforms are there? // Nat. Chem. Biol. 2018. Vol. 14, № 3. P. 206–214.
- 209.Лисица А.В. и др. Постгеномная медицина: альтернатива биомаркерам. // Вестник Рос. Акад. Мед. Наук. 2016. Т. 71, № 3. С. 255–260.

БЛАГОДАРНОСТИ

Автор выражает глубокую благодарность д.б.н. Станиславу Николаевичу Нарыжному за предоставление результатов экспериментов по двумерному электрофорезу белков опухолевой линии HerG2, а также д.б.н., проф. Виктору Гавриловичу Згоде за предоставление результатов протеомного масс-спектрометрического анализа фракционированных и нефракционированных белков линии HerG2. 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 РФФИ № 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).

Приложение 1. Перечень масс-спектрометрически идентифицированных протеоформ

Абсолютно уникальные пептиды (относительно как транскриптом-специфичной библиотеки для клеточной линии HepG2, так и всего протеома человека по данным репозитория UniProt)

A6NCS6	O15498	O75131	O95292	P02649
A6NDG6	O15511	O75191	O95299	P02652
A6NHQ2	O15514	O75208	O95336	P02656
A6NIH7	O43169	O75223	O95372	P02766
A6NJ78	O43175	O75323	O95373	P02771
A6ZKI3	O43181	O75347	O95396	P02787:I:448:V
C9JLW8	O43242	O75348	O95400	P02792
O00148	O43324	O75351	O95445	P02794
O00151	O43396	O75356	O95470	P02795
O00161	O43402	O75368	O95571	P03928
O00186	O43447	O75380	O95573:F:551:S	P04080
O00193	O43464	O75414	O95613:S:3122:I	P04181
O00214:R:184S	O43488	O75436	O95619	P04183
O00217	O43491	O75438	O95721	P04350
O00217:A:200:T	O43583	O75439	O95747	P04733
O00244	O43598	O75439:E:396:D	O95749	P04792
O00273	O43615	O75449	O95777	P04843
O00299	O43617	O75489	O95816	P05114
O00303	O43633	O75531	O95825	P05161
O00399	O43676	O75569	O95865	P05162
O00422	O43678	O75575	O95881	P05198
O00425	O43681	O75616	O95926	P05204
O00461	O43708-2:M:27:T	O75663	O95954	P05387
O00483	O43741	O75817	O95954:A:438:V	P05423
O00487	O43765	O75821	O95989	P05455
O00505	O43776	O75874:I:367:V	O95994	P05556
O00515:K:323:E	O43809	O75880	O95999	P05783
O00584	O43813	O75884	O96000	P06132
O00629	O43818	O75934	O96007	P06493
O14548	O43819	O75937	O96008	P06702
O14562	O43823	O75940	O96019	P06746
O14618	O43920	O75947	P00352	P06865:I:436:V
O14657	O43924	O75964	P00374	P07148:T:94:A
O14734	O60216	O76003:P:123:S	P00439	P07148:V:42:M
O14737	O60218:N:313:D	O76071	P00491	P07237
O14745	O60220	O94817	P00492	P07311
O14907	O60262	O94826	P00558:L:64:I	P07384
O14908	O60443	O94832	P00568	P07711
O14949	O60508	O94905	P00813	P07737
O15116	O60568	O94925	P00966	P07737:G:94:R
O15143	O60671	O94925-3	P01009	P07738
O15145	O60762	O95139	P01011	P07741
O15212	O60783	O95140	P01023:L:158:M	P07902
O15264	O60826	O95159	P01024:P:314:L	P07919
O15270	O60869	O95163	P01040	P07942:Q:1022:R
O15347	O60884	O95167	P01116-2	P07947
O15372	O60925	O95182	P01833	P08069
O15427	O60930	O95249	P02647	P08134

P08579	P15880	P25311	P35558:V:184:L	P49006
P08833	P15907	P25398	P35754	P49247
P08865	P15954	P25685	P35900	P49327
P09001	P16219	P25789	P35914	P49368
P09012	P16401	P26885	P35998	P49368:N:61:T
P09132	P16930	P27144	P36404	P49406
P09234	P16930-2	P27348	P36405	P49427
P09382	P16949	P27694	P36507	P49458
P09429	P16989-2	P27707	P36542	P49459
P09455	P17152	P28065	P36957	P49585
P09601	P17568	P28070:Q:114:L	P37198	P49591
P09669	P17812	P28072	P37235	P49593:L:420:R
P09972	P17844	P29083	P38646	P49642
P0CG38	P17900:V:153:A	P29279:H:83:D	P38919	P49675
P0CG39	P17980	P29373	P39019	P49720
P0CW22	P17987	P29966	P39687	P49721
P10109	P18031	P29992	P40121-2:H:320:R	P49755
P10253:H:199:R	P18077	P30040	P40222	P49756
P10606	P18085	P30041	P40306	P49770
P10644	P18124	P30042	P40855	P49773
P10746	P18669	P30043	P40938	P49789
P10768	P19338	P30046	P41208	P49914
P10809	P19387	P30049	P41214	P50135
P11021	P19388:S:44:F	P30050	P41223	P50150
P11166	P19623	P30084:T:75:I	P41236	P50225:V:223:M
P11216:A:303:S	P20042	P30086	P41567	P50226
P11233	P20290	P30101	P42126	P50336
P11387	P20290-2	P30153	P42166	P50395
P11441	P20338	P30405	P42167	P50402
P11717:G:573:S	P20674	P30419	P42224	P50416
P11766	P20962	P30520	P42285	P50453
P11802	P21266	P30533	P42345	P50552
P12004	P21266:V:224:I	P30740	P42574	P50583
P12235	P21281	P31146	P42677	P50914
P12236	P21283	P31150	P42765	P50990
P12429	P21291	P31153	P43034	P51149
P13073	P21549	P31321	P43243	P51452
P13489	P21796	P31689	P43490	P51571
P13639	P21912	P31937	P45973	P51580
P13640	P21953	P31943	P46060	P51665
P13674	P21980	P32119	P46063	P51688
P13693	P22033	P32969	P46109	P51812
P13929:N:71:S	P22090	P33176	P46459	P51956
P13929:V:85:A	P23246	P33316	P46776	P51991
P13984	P23284	P33552	P46778	P51991-2
P13987	P23297	P33992	P46783	P52435
P14174	P23378	P34059	P46940	P52597
P14550	P23434	P34932:N:506:D	P47224	P52758
P14621	P23528	P35080	P47755	P52888
P14649	P23610	P35080-2	P47813	P52907
P14735	P23786:V:368:I	P35218	P47929	P52926
P14854	P23919	P35244	P48047	P53582
P14927	P24385	P35249	P48147	P53597
P15121	P24539	P35250	P48507	P53602
P15311	P24539:I:84:V	P35268	P48637	P53611
P15374	P25205:E:777:K	P35270	P48723	P53618

P53634:I:453:V	P62280	P82979	Q13151	Q15397
P53701	P62304	P82980	Q13158	Q15404
P53999	P62312	P83436	Q13162	Q15428
P54105	P62316	P83731	Q13185	Q15493
P54652	P62330	P83876	Q13217	Q15545
P54920	P62333	P83881	Q13232	Q15643
P55036	P62424	P83916	Q13242	Q15645
P55081	P62487	P84095	Q13287	Q15649
P55157:D:384:A	P62701	P84098	Q13315:V:2906:I	Q15650
P55199	P62714	P84157-2	Q13347	Q15691
P55212	P62750	P98179	Q13451	Q15758
P55769	P62753	Q00059	Q13485	Q15773
P55789	P62760	Q00266	Q13501	Q15776
P55795	P62826	Q00403	Q13526	Q15785
P56385	P62829	Q00534	Q13547:L:271:I	Q15819
P56556	P62834	Q00535	Q13572	Q16378:Q:120:R
P56962	P62841	Q00577	Q13630	Q16378:R:96:Q
P57105	P62851	Q00688	Q13643	Q16539-2
P58546	P62854	Q00796	Q13685	Q16543
P60510	P62857	Q01085	Q13686	Q16595
P60891	P62875	Q01085-2	Q13825	Q16658
P60903	P62888	Q01105	Q13951	Q16706
P60953	P62906	Q01105-2	Q13951-2	Q16740
P60983	P62910	Q01469	Q14011	Q16762
P61006	P62917	Q01658	Q14012	Q16795
P61011	P62937-2	Q01850	Q14019	Q2KHR2:I:204:T
P61024	P62942	Q01995	Q14061	Q2NL82
P61026	P62987	Q02127	Q14137	Q2Q1W2
P61081	P62993	Q02338	Q14192	Q2TAA5
P61088	P63104	Q02543	Q14197	Q3ZCW2
P61106	P63146	Q02790	Q14232	Q4VC31
P61158	P63167	Q02818	Q14241	Q53F19
P61160	P63172	Q02878	Q14258	Q53FT3
P61225	P63173	Q02978	Q14320	Q53H12
P61247	P63208	Q03393	Q14554	Q53H82
P61254	P63220	Q03701:V:102:I	Q14653	Q53HC9
P61326-2	P63244	Q04837	Q14657	Q53RE8
P61421	P63272	Q04917	Q14677	Q53S33
P61457	P67809	Q05639	Q14683	Q53T59
P61513	P67936	Q06323	Q14690:L:1216:F	Q5BKU9
P61599	P68402	Q06520	Q14692	Q5BKZ1
P61604	P78318	Q06546	Q14696	Q5EBL4
P61803	P78330	Q07021	Q14739:S:154:N	Q5HYI8
P61970	P78406	Q07960	Q15007	Q5HYK3
P61981	P78540	Q08380	Q15018	Q5JS37
P62072	P78560	Q08945	Q15031	Q5JVS0
P62081	P80297	Q09666:G:2527:D	Q15061	Q5R3I4:F:243:L
P62136	P80404	Q0VDF9	Q15102	Q5RKV6
P62140	P82650	Q10472	Q15126	Q5SRE7
P62241	P82663	Q12788	Q15155	Q5T653
P62244	P82664	Q12874	Q15170-2	Q5T6V5
P62249	P82673	Q12907	Q15293	Q5U5X0
P62253	P82675	Q12965	Q15363	Q5VIR6-4
P62256	P82909	Q12972	Q15365	Q5VT66:T:165:A
P62266	P82932	Q13123	Q15369	Q5W111
P62277	P82933	Q13148	Q15382	Q5XKP0

Q69YN2	Q8IU81	Q8TEA8	Q96DV4	Q96SL4
Q6DKJ4	Q8IUR0:S:52:A	Q8WUD4	Q96E39	Q96SZ5
Q6GMV3	Q8IV08	Q8WUJ0	Q96EH3	Q99417
Q6I9Y2	Q8IVD9	Q8WUM0:I:294:V	Q96EK6	Q99442
Q6IAA8	Q8IVS2	Q8WUX2	Q96EK9	Q99470
Q6IN84	Q8IWA0	Q8WUX9	Q96EL3	Q99471
Q6N069	Q8IWB7	Q8WV22	Q96EQ0	Q99496
Q6NSJ5	Q8IWE2:G:84:R	Q8WV92	Q96EX1	Q99497
Q6NVY1:T:46:A	Q8IWL3	Q8WVE0	Q96EY7	Q99519
Q6P1J9	Q8IWX8	Q8WVJ2	Q96F10	Q99541
Q6P1L8	Q8IXM3	Q8WVM0	Q96F63	Q99567
Q6P1N9	Q8IXT5	Q8WVY7	Q96FJ2	Q99570
Q6P1X6	Q8IYB8	Q8WX92	Q96FK6	Q99584
Q6P2I3	Q8IYK4	Q8WXA9-2	Q96FN9	Q99598
Q6P587	Q8IYS1	Q8WXX5	Q96FQ6	Q99614
Q6PI48	Q8N1Q1	Q8WZA9	Q96FW1	Q99653
Q6PL18	Q8N2Z9	Q92520	Q96FX7	Q99747
Q6PL24	Q8N335	Q92572	Q96FZ7	Q99757
Q6PUV4	Q8N490-2	Q92575	Q96G03:E:488:D	Q99805
Q6QNY0	Q8N4P3	Q92665	Q96GA7	Q99816
Q6UW78:G:89:S	Q8N4Q0	Q92747	Q96GG9	Q99829
Q6UWS5	Q8N543	Q92785	Q96GW9	Q99986
Q6UXH0	Q8N5K1	Q92786	Q96GX9	Q9BPW8
Q6UXN9	Q8N5M1	Q92820	Q96H79	Q9BPX5
Q6ZN17	Q8N5M4	Q92882	Q96HE7	Q9BQ04
Q70UQ0-4	Q8N5N7	Q92905	Q96HP4	Q9BQ61
Q7L0Y3	Q8N668	Q92922	Q96HQ2	Q9BQ67
Q7L3T8	Q8N806	Q92947	Q96HY6	Q9BQ69
Q7L5N1	Q8N8N7	Q92979	Q96HY6:A:303:T	Q9BQP7
Q7LBR1	Q8N8S7	Q93034	Q96I15-2:A:175:T	Q9BR61
Q7RTV0	Q8N9N2-2	Q93052	Q96I25	Q9BR76
Q7Z2K6	Q8N9N7	Q969G6	Q96I99	Q9BRF8
Q7Z2W9	Q8N9N8	Q969H8	Q96IU4:G:27:S	Q9BRF8:K:241:R
Q7Z417	Q8NAV1	Q969S3	Q96IY1	Q9BRG1
Q7Z434:Q:93:E	Q8NBJ5	Q969T9	Q96IZ0	Q9BRJ6
Q7Z4G1	Q8NBN7	Q96A35	Q96J01	Q9BRJ7
Q7Z4H8	Q8NBX0	Q96A49	Q96JH7	Q9BRP4
Q7Z4W1	Q8NC60	Q96AB6	Q96JM3	Q9BRP8
Q7Z5K2:V:124:I	Q8NEZ5	Q96AE4	Q96K17	Q9BRQ6
Q86SX6:A:146:T	Q8NFB3	Q96AE4-2	Q96KP1	Q9BRQ6:A:95:S
Q86SZ2	Q8NFB4	Q96AG4	Q96LA8	Q9BRT9
Q86TX2	Q8NFB8	Q96B26	Q96LJ7	Q9BRX5
Q86U28	Q8NFW8	Q96B54	Q96MF7	Q9BS26
Q86U90	Q8NHP8	Q96BJ3	Q96MH2	Q9BSB4
Q86UE4	Q8NI36	Q96BN2	Q96MW1	Q9BSD7
Q86UY8	Q8TAA5	Q96BN8	Q96MW5	Q9BSE5:G:105:R
Q86V88	Q8TAE8	Q96BR5	Q96NB3:H:99:Q	Q9BSF4
Q86VS8	Q8TB37	Q96C01	Q96NC0	Q9BSH4
Q86VU5	Q8TBC4	Q96C19	Q96NT0	Q9BSH5
Q86WR0	Q8TBF4	Q96C86	Q96NU7	Q9BSH5:G:146:E
Q86WW8	Q8TCD5	Q96C90	Q96P11-2	Q9BSY4
Q86Y39	Q8TCS8:I:121:V	Q96CD2	Q96PE2	Q9BT09
Q86Y82	Q8TDL5	Q96CT7	Q96PE7	Q9BT30
Q86YH6	Q8TDN6	Q96D46	Q96PZ0	Q9BT73
Q86YM7	Q8TDP1	Q96DG6	Q96QR8	Q9BT78
Q86YS6	Q8TEA1	Q96DH6	Q96S44	Q9BTE7

Q9BTL3	Q9H1E3	Q9NPD3	Q9NX63	Q9UJC5
Q9BTT6	Q9H1Y0	Q9NPD8	Q9NYB0	Q9UK22
Q9BU02	Q9H204	Q9NPE3	Q9NYG5	Q9UK41
Q9BU61	Q9H299	Q9NPF4	Q9NYL9	Q9UKB3
Q9BU89	Q9H2H8	Q9NQ88	Q9NYP9	Q9UKK6
Q9BUE0	Q9H2I8	Q9NQE9	Q9NZ45	Q9UKK9
Q9BUF5	Q9H2W6	Q9NQG5	Q9NZD8	Q9UKL6
Q9BUH6	Q9H3G5:A:435:V	Q9NQG6	Q9NZJ9	Q9UKU7
Q9BUI4	Q9H3P2	Q9NQP4	Q9NZW5	Q9UKY7
Q9BUL8	Q9H3P7	Q9NQR4	Q9P000	Q9UL25
Q9BUR4	Q9H3R5	Q9NQS1	Q9P015	Q9ULR3
Q9BUR5	Q9H477	Q9NQ4T:T:5:M	Q9P016	Q9ULX9
Q9BV19	Q9H479	Q9NQ4T	Q9P032	Q9UMS4
Q9BV20	Q9H488	Q9NQZ5	Q9P035	Q9UMX5
Q9BV86	Q9H4A6	Q9NR33:G:17:V	Q9P0J7	Q9UN37
Q9BVA1	Q9H4M9	Q9NR45	Q9P0M6	Q9UNE7
Q9BVI4	Q9H583:N:1694:S	Q9NR56	Q9P0P8	Q9UNI6
Q9BW61	Q9H5N1	Q9NRF8	Q9P0R6	Q9UNM6:N:13:S
Q9BW85	Q9H5X1	Q9NRF9	Q9P0S9	Q9UNQ2
Q9BW92	Q9H6E4	Q9NRN7	Q9P0U1	Q9UQ80
Q9BWJ5	Q9H6S3	Q9NRP2	Q9P0Z9	Q9Y221
Q9BWM7	Q9H773	Q9NRS6	Q9P1F3	Q9Y244
Q9BWT6	Q9H7C9	Q9NRV9	Q9P258	Q9Y248
Q9BWU0:P:139:T	Q9H7Z7	Q9NRW7	Q9P287	Q9Y265
Q9BX68	Q9H845	Q9NRX2	Q9P2E9-3:L:610:H	Q9Y266
Q9BXB4	Q9H871	Q9NRX4	Q9P2E9-3:S:766:L	Q9Y276
Q9BXJ9	Q9H8H0	Q9NS69	Q9P2G3	Q9Y294
Q9BXR0	Q9H8H3	Q9NSE4	Q9P2J9	Q9Y2B0
Q9BY49	Q9H8W4	Q9NTG7	Q9P2N5	Q9Y2G5
Q9BYC8	Q9H936	Q9NTM9	Q9P2W9	Q9Y2P8
Q9BYD1	Q9H944	Q9NTZ6	Q9UBE0	Q9Y2Q9
Q9BYD3	Q9H993	Q9NUG6	Q9UBF6	Q9Y2R0
Q9BYD6	Q9H9A6	Q9NUP1	Q9UBP6	Q9Y2S0
Q9BYG3	Q9H9B4	Q9NUP9	Q9UBQ0	Q9Y2S6
Q9BYN0	Q9H9H4	Q9NV31	Q9UBQ0-2	Q9Y2S7
Q9BYN8	Q9H9J2	Q9NV56	Q9UBQ5	Q9Y2T2
Q9BZE1:C:366:S	Q9H9Q4	Q9NVI7-2:G:15:D	Q9UBQ7	Q9Y2V2
Q9BZK7	Q9HA64:K:182:R	Q9NVM6	Q9UBU9	Q9Y2W1:A:201:V
Q9C005	Q9HA77	Q9NVP1	Q9UDR5	Q9Y2W6
Q9C075	Q9HAN9	Q9NVT9	Q9UDX5	Q9Y2Z4
Q9GZL7	Q9HAS0	Q9NWM8	Q9UEU0	Q9Y314
Q9GZP4	Q9HAV7	Q9NWS6	Q9UFC0	Q9Y315
Q9GZS3	Q9HB40	Q9NWU2	Q9UFG5	Q9Y333
Q9GZU8	Q9HBH1	Q9NWV4	Q9UFW8	Q9Y365
Q9GZZ1	Q9HBL8	Q9NWV8	Q9UGK8:K:429:E	Q9Y375
Q9GZZ9	Q9HBM1	Q9NWX6	Q9UGM6	Q9Y376
Q9H061	Q9HC36:G:45:E	Q9NWX4	Q9UGV2	Q9Y399
Q9H0A8	Q9HC52	Q9NWZ8	Q9UH62	Q9Y3A6
Q9H0C8	Q9HCD5	Q9NX01	Q9UHI6	Q9Y3B4
Q9H0F6	Q9HCE1	Q9NX08	Q9UHQ9	Q9Y3B8
Q9H0N5	Q9HCN8	Q9NX18	Q9UHV9	Q9Y3B9
Q9H0R4	Q9HD15	Q9NX24	Q9UHY1	Q9Y3C0
Q9H0T7	Q9NP66	Q9NX38	Q9UHY7	Q9Y3C6
Q9H0U4	Q9NP92	Q9NX40	Q9UI09	Q9Y3C7
Q9H173	Q9NP97	Q9NX55	Q9UII2	Q9Y3C8
Q9H1D9	Q9NPA0	Q9NX62	Q9UIJ7	Q9Y3D0

Q9Y3D2	Q9Y4L1	Q9Y5J7	Q9Y5Y2	Q9Y6G3
Q9Y3D5	Q9Y4W6	Q9Y5K6	Q9Y5Z4	Q9Y6G5
Q9Y3D6	Q9Y4Z0	Q9Y5L4	Q9Y619	Q9Y6G9
Q9Y3D9	Q9Y512:I:345:V	Q9Y5M8	Q9Y676	Q9Y6H1
Q9Y3E5	Q9Y530	Q9Y5R8	Q9Y678	Q9Y6J9
Q9Y3I0	Q9Y547	Q9Y5S9	Q9Y696	Q9Y6N1
Q9Y3X0	Q9Y584	Q9Y5U8	Q9Y6A4	
Q9Y478	Q9Y5J1	Q9Y5X3	Q9Y6B6	

Протеоформы, масс-спектрометрически идентифицированные по протеотипическим пептидам, уникальным относительно транскриптом-специфичной библиотеки для клеточной линии HepG2, но картирующимся на несколько протеоформ одного белок-кодирующего гена в масштабе всего протеома человека по данным репозитория UniProt

P49588	Q8N6H7	Q9P260-2	P30048	P36915
P48739	Q9BUT1	P60842	Q53GG5-2	P43897-4
P22314	Q13247	Q6L8Q7	Q9UH92-3	P43897-2
P13861	Q9H8Y8	Q96GC5	Q8TCA0-2	P33240
P51570	P16870	Q9BQ95	Q06547	Q9Y3D3
P09960	Q6FI81-3	P46926	Q6UX04	P62913
Q13884	Q8IW45-4	P04179-4	Q9H6R4	Q9H832
P14868	P35232-2	P10909	Q96DI7	Q9GZT9
Q9HCC0	P27824	Q43837	Q86X83	Q9BX40
P09493-10	Q8TCA0-3	Q9Y2Z9-3	Q9NZB2	O95219
Q5JTV8-3	Q9NYR9	P82912	P15289-2	P40925
P06753-6	Q13838	P11940	Q5TFE4	Q9Y508
Q4G0N4-2	P55209	Q14247	Q14155-1	Q92890
P62191-2	P17174	Q14919	Q86VM9	Q9Y697-2
O75431	Q9UJU6-5	Q9UJ41-2	Q9NP79	Q9H6R7
P34949-2	Q5TDH0	Q86TG7-2	O95218-2	P78537
Q9H9Q2	P05026	Q9NVS2	Q96AT9	Q99436
Q9H0W9	Q6NZY4-2	Q9UGI8	O94842	Q96S59
Q8IY67	Q8NFB4-4	P28161	Q15526	Q8IVS8
Q6NUK1-2	Q9NY27	Q9BW27	Q8TDH9	P33121
P12694-2	P17858	Q92783	Q9NYB9-2	O14732
Q9NUJ1	Q9P0J1	Q96TC7	O75306	Q9UEW8
O75381	Q02750	Q8NCA5-2	Q9NQH7	Q8N163
O94874	P19971	Q8N3X1	O43929	Q8WW12
Q9NQX3-2	Q8IUF8	Q8IUI8	Q86Y37	Q9NWU5-3
Q9Y3B7-2	Q3MHD2	Q4G0N4-3	Q8WUW1	Q15287
Q9NUI1	Q5T2W1	P11172	P07954	P24928-2
Q14558-2	Q6P6C2	Q9H6V9	Q9NUW8	Q9H098
Q13505-3	Q9UHB9	Q16718-2	Q13438-4	Q9BQQ3-2
Q15717	P37802	Q53GS9	Q86VP6	Q9BQQ3-3
P06733	A0AVT1	P78549	O75608	Q9NZN8
P31946	P28288	Q9NPH0-2	Q9BRQ8	Q9UNN5
Q9H2P9	Q8TE02	Q8TCT9	P12931	Q9H2U2
Q8NFB5	Q01415-2	Q86WQ0	Q9Y2Y0	Q9BX93
Q99873-4	Q9BRR6-2	P20823-5	P51948-2	Q2TAY7
Q8WZA0	P49354	Q9UBK9	P39748	Q9H4A5

Q15119	Q8WWV3	Q9UMX0	P35749-2	P0DMV9
Q16514	Q86XA9-3	Q86W56	Q9NR31	P62158
Q9UJA5	Q5EBL8	Q86W50	P04732	P0DJD1
Q9P0J0	Q8N954-2	P31947	Q13620-1	
O00231	Q9BW66	P00367	Q8NB90	
Q9BW71	Q8IYB3	P11234	Q8WUD1	
Q96BW1	Q9UK59	Q06587	P49753	
Q9NY12	Q8IXQ4	C9JRZ8-2	Q6KB66	

Протеоформы, масс спектрометрически идентифицированные по пептидам, уникальным относительно транскриптом-специфичной библиотеки для клеточной линии HerG2, но картирующимся на протеоформы нескольких белок-кодирующих генов в масштабе всего протеома человека по данным репозитория UniProt

Q8TCE6	P68104	A6NNZ2
P50502	O60812	P62308
Q9Y3D7	O14925	P01889:I:218:V

Приложение 2. Перечень аннотированных протеоформ

Перечень содержит список идентификаторов UniProt для протеоформ, аннотированных в клеточной линии HepG2.

В скобках приведены сведения о детектированных посттрансляционных модификациях: *deam* – деамидирование остатков аспарагина и глутамина, *am* – амидирование С-концевых остатков, *ac* – ацетилирование лизина, *meth* – метилирование лизина и аргинина, *ph* – фосфорелирование остатков серина, треонина и тирозина. Аминокислотные полиморфизмы обозначены как A:n:B, где A – аминокислотный остаток в канонической последовательности, B – остаток аминокислоты, который заместил исходный остаток в результате несинонимичного однонуклеотидного полиморфизма, n – полиморфное положение. Вставки и делеции обозначены как AA...AA[a]:x:BB...BB[b], где AA...AA – фрагмент нативной аминокислотной последовательности, который заменяется на фрагмент BB...BB, а и b – количество аминокислотных остатков в заменяемом и заменяющем фрагментах, а n – положение в последовательности, с которого начинаются изменения.

A0AVT1 (ac)	O00487	2787:TAA..QS[32aa]	O43707-
A2VDF0	O00487 (deam*3, am)	O15264	2:PFN..DL[373aa]:320:RP
A6NCS6	O00505	O15270	S..RA[28aa]
A6NDG6 (deam, am)	O00515:K:323:E	O15347 (deam)	O43707:PFN..DL[373aa]:
A6NDG6 (deam*2)	(deam*2, am)	O15372 (am)	539:RPS..RA[28aa]
A6NHQ2	O00515:RKE..EV[57aa]:4	O15400-2	O43741
A6NIH7 (deam, am)	61:QGRRT	O15427	O43765
A6NJ78 (deam)	O00571	O15498	O43765 (am)
A6ZKI3	O00571:GRS..GN[77aa]:5	O15511 (ac)	O43768-3
A9UHW6-2	86:GHV..GV[77aa]	O15514 (deam, am)	O43776
C9JLW8	O00584 (deam*2, am)	O15523:EEA..GN[379aa]:	O43776 (deam*3, am)
O00148 (deam)	O00629	282:GRK..LV[16aa]	O43809 (deam)
O00148 (deam*4)	O00629 (deam, am)	O43169 (ac)	O43813
O00151 (deam*4, am*2)	O14530	O43175	O43815-2
O00154-4	O14548	O43175 (deam*3, am)	O43818
O00154-6	O14562 (deam*3, am)	O43181 (deam)	O43819 (deam)
O00154-7	O14602	O43237-2:N:140:H	O43823
O00161 (deam, am, ac*2)	O14618	O43242 (deam)	O43903
O00178:ERM..GC[313aa]	O14657	O43324 (deam)	O43920
:357:EED..AC[39aa]	O14734 (deam, am)	O43390	O43924 (deam)
O00186	O14737 (deam, am)	O43390-2	O43929 (deam, am)
O00193 (deam, ph)	O14737 (deam*2, am)	O43390-3	O60216 (deam*2, am)
O00214:R:184:S	O14745 (deam*6, am*3)	O43396 (deam*3, am*2)	O60220
O00217 (deam*2)	O14773	O43402	O60262
O00217:A:200:T (deam)	O14907 (deam, am)	O43402 (deam*3, am*2)	O60443
O00231 (deam*3, am)	O14908	O43414-2:G:189:R	O60493-4
O00244 (deam)	O14949 (deam, am)	O43447 (deam, am)	O60506
O00264:LSD..ND[43aa]:1	O14964:QEQ..FD[241aa]:	O43464 (deam)	O60506-3
53:LRV..NC[19aa]	537:QRA..CG[91aa]	O43488 (deam*3)	O60508
O00273 (deam*2, am)	O14979-2	O43491-3	O60547-2
O00299 (ac)	O14979-2:LK:107:LKK	O43491(deam)	O60547-2:T:125:R
O00303	O15067:DNS..SC[1068aa]	O43583	O60568 (deam, am)
O00303 (deam*3, am)	:271:DTA..KG[30aa]	O43598	O60573-2
O00330-2:R:23:C	O15091-3:F:146:C	O43615	O60671
O00399	O15116 (am)	O43617 (deam)	O60701
O00422 (deam*4, am*3)	O15116 (deam*2, am, ac)	O43633	O60701-
O00425	O15143 (deam)	O43676 (deam, am)	2:FDG..KV[56aa]:372:LS
O00425 (deam)	O15145 (deam*2)	O43678 (deam, am)	M..LK[56aa]
O00461 (deam, am)	O15212 (deam, am)	O43681 (deam*3, am)	O60701-
O00483	O15230:SRQ..AA[909aa]:	O43707-2	3:FDG..KV[56aa]:342:LS

M..LK[56aa]	O75822	O95721	4:SVG..LP[59aa]:516:TC
O60762 (deam)	O75874:DMV..KL[162aa]	O95747	G..PS[59aa]
O60783 (deam, am)	:253:DYG..GL[17aa]	O95747 (deam*2, am)	P02545-
O60826 (am)	O75874:I:367:V (deam*3,	O95749	5:NKS..IM[209aa]:357:T
O60830	am, ac)	O95777 (deam)	QVQ
O60869 (deam)	O75874:MVA..KL[161aa]	O95793-	P02545-6
O60869:KIE..AK[26aa]:1	:254:WWPKL	2:REF..RC[261aa]:236:M	P02647 (deam, am)
23:NRA..GA[22aa]	O75880 (deam, am)	GV..LF[70aa]	P02649 (deam*2, am*2)
O60884 (deam, am)	O75884	O95793-	P02652 (deam)
O60925	O75884 (deam*2, am)	3:REF..RC[261aa]:242:M	P02656
O60925 (deam, am)	O75934 (deam*3, am*2)	GV..LF[70aa]	P02671-2
O60930	O75935	O95816	P02679-2:D:390:V
O75083-	O75937	O95816 (ac)	P02766
3:AVH..TY[155aa]:312:A	O75940	O95825	P02768-3:E:84:D
QC..PR[28aa]	O75947	O95865 (deam, am)	P02768-3:E:84:V
O75083:DMM..TY[54aa]:	O75964	O95881 (deam*2, am)	P02771 (deam*7, am*4)
553:A	O76003:P:123:S (deam*2,	O95926	P02786:VWDIDNEF:753:
O75083:DMM..TY[54aa]:	am)	O95954 (deam*3, am)	CLGH
553:AHDGVCLDPE	O76071 (deam, am)	O95954:A:438:V	P02787:I:448:V
O75131 (deam*3, am)	O76094:RKK..GW[210aa	(deam*2, am)	P02787:I:448:V (deam*4,
O75179-	]:462:RGR..HL[33aa]	O95983-2	am)
7:NQS..LG[1803aa]:688:	O94817	O95989 (deam, am)	P02787:QEH..RP[409aa]:
NVSLQRAL	O94826 (ac)	O95994 (deam, am)	290:LGT..AV[25aa]
O75190-2	O94832	O95999	P02792
O75191 (deam, am)	O94842 (deam)	O96000 (deam, am)	P02794
O75208	O94888	O96007	P02795 (deam, ac*2)
O75208 (deam*2, am)	O94905	O96008	P02795 (deam, ac*4,
O75223 (deam, am)	O94905 (deam, am)	O96008-2	meth*2, ph)
O75323	O94925 (deam*4, am*2)	O96019 (deam*5, am)	P03928 (deam*2, am*2)
O75347 (deam*2, am)	O94925-3	P00338-2	P04075:I:107:T
O75348	O94925-3 (am)	P00338-5	P04080 (deam*4, am*2,
O75351	O95139	P00352 (deam*5, 2*am)	ac*5, ph, meth)
O75356	O95140	P00367 (deam*5, 2*am)	P04181 (ac)
O75368 (deam*2, am)	O95159	P00367:TIG..FT[316aa]:2	P04181-2
O75380	O95163	43:THRAL	P04183 (ac)
O75390:SGV..SG[62aa]:4	O95167	P00374 (deam*2, am)	P04350 (deam*2)
05:EWG..GL[59aa]	O95182 (deam, am)	P00439 (deam, am)	P04406-2
O75391	O95219	P00491 (deam*6, am*2,	P04406-2:K:173:N
O75391:CVP..TS[37aa]:1	O95249 (deam, am)	ac*4)	P04439:D:98:H
91:VCARGQ	O95292 (deam, am)	P00492	P04632:GNM..YS[47aa]:2
O75410-6	O95292 (deam, am)	P00492 (ac)	22:GTW..SR[34aa]
O75414 (deam, am)	O95299 (deam*4, meth)	P00568 (deam*3, am)	P04732 (deam, ac*4,
O75436	O95336 (deam, am)	P00813 (deam, am)	meth*3, ph)
O75438	O95347:EAE..EV[275aa]:	P00966	P04733 (meth)
O75439 (deam)	923:ERLKMVLQRYPKC	P01009 (deam*3, am)	P04792
O75439:E:396:D (deam,	O95372	P01009-3	P04843 (deam*3, am)
am)	O95372 (deam, am)	P01011	P04899
O75449	O95373 (deam, am)	P01023:L:158:M (deam,	P05091-2
O75489 (deam)	O95394-4	am)	P05114 (deam)
O75531 (ac)	O95396 (deam*2, am)	P01024:GDH..PN[1077aa	P05114 (deam*3)
O75569 (deam)	O95400 (deam, am)	]:587:GLT..SC[19aa]	P05161 (deam*2, am*2)
O75575	O95433:GGK..LF[92aa]:2	P01024:P:314:L (deam*2,	P05162 (deam*3, am)
O75607	47:GEESSTW	am)	P05166
O75616	O95445 (deam)	P01040	P05166-2
O75643:WLQ..SD[1870aa	O95453-4	P01116-2	P05166-
]:267:CAAAAQSF	O95470	P01833	2:FVD..PL[149aa]:411:L
O75663 (deam, am)	O95571 (deam*2, am)	P01889:I:218:V (deam,	C
O75688	O95573:F:551:S (deam,	am)	P05166:FVD..PL[149aa]:
O75688:KVE..KI[58aa]:4	am)	P02545-	391:LC
22:KSRGRRKPC	O95613:S:3122:I	4:NKS..LP[231aa]:344:T	P05166:KNI..PL[202aa]:3
O75817	(deam*4, am*2, meth)	QVQ	38:KQH..PT[22aa]
O75821 (ph)	O95619	P02545-	

P05198 (ac)	2:KKA..ED[229aa]:413:R	2:PSGGASSGPTIEEVD:	P13987
P05204	EGHGGSRHR	479:PLWWCFLRAHH	P14174
P05387 (ac)	P08133:KKA..ED[229aa]:	P11166 (deam)	P14314-
P05423 (ac)	445:REGHGGSRHR	P11172	2:AMK..EL[67aa]:459:AT
P05455 (deam, am)	P08134	P11172:ELS..GV[257aa]:	P14324
P05556 (deam*2)	P08195	224:ELS..CS[46aa]	P14324-2
P05783 (deam, am)	P08238:EAD..VD[82aa]:6	P11177	P14373
P05783:IDL..RH[123aa]:3	43:DRP..IA[41aa]	P11177-2	P14550 (deam*3, am)
08:TRP..AP[38aa]	P08579	P11216:A:303:S (deam)	P14618:TSD..VP[127aa]:
P05783:RKN..RH[170aa]:	P08579 (deam)	P11233	405:KPA..GP[24aa]
261:QEPRGARQVLVSA	P08833 (ph)	P11234	P14621 (deam)
D	P08865 (deam*4, am)	P11387 (deam*2)	P14649
P06132	P09001	P11441 (deam, am)	P14649 (deam)
P06132 (am)	P09012	P11498:KEV..IE[291aa]:8	P14735
P06396:AKR..AA[56aa]:7	P09132 (deam*3, am)	88:RRS..SR[19aa]	P14854 (deam*3, am*2)
27:D	P09234	P11717:G:573:S	P14868 (deam*3, am*2)
P06493 (deam, am)	P09234 (deam, am)	P11766 (deam*2, am)	P14927 (deam*2, am*2)
P06702	P09382 (deam, am)	P11802	P15121
P06730-3	P09429	P12004	P15121 (deam)
P06733 (deam*7, am*3)	P09455 (deam*2, am*2)	P12235 (deam)	P15170
P06744	P09493-10 (deam)	P12236	P15311 (deam)
P06744-	P09543-2:T:22:S	P12268	P15311:MRE..AL[246aa]:
2:AKD..VQ[326aa]:244:A	P09543:T:42:S	P12268:K:205:M	341:IAREGGVDAAAAG
GSFCSGEALCCPVY	P09601	P12429 (deam*2, am*2)	L
P06746	P09651	P12931	P15374 (deam*3, am*2)
P06748	P09651-2	P12956-	P15531
P06748-2	P09651-	2:VEK..QD[144aa]:425:V	P15531-2
P06753-6 (deam*4, am)	2:GGN..RF[43aa]:278:GE	RS..WI[31aa]	P15880
P06865:I:436:V (am)	E..ED[43aa]	P12956:DLR..QD[382aa]:	P15907
P07148:T:94:A (deam*4, am)	P09651-3	228:DSGFTLRNPAS	P15924:KEL..GH[1308aa]
P07148:V:42:M (deam*3, am)	P09651:GGN..RF[43aa]:3	P13010:AYK..MI[470aa]:	:1564:KRA..HC[62aa]
P07237 (deam, am)	30:GEE..ED[43aa]	263:AV	P15924:LEE..GH[1278aa]
P07311 (deam, am)	P09669 (deam)	P13073	:1594:LERKSWKA
P07355	P09960	P13489 (deam*5, am*2)	P15924:RRA..GH[1604aa]
P07384	P09960-3	P13639	:1268:SAS
P07711	P09972 (deam*2)	P13639 (deam)	P15924:RRA..GH[1604aa]
P07737 (deam*4, am)	P0C870	P13640 (ac*2, meth*2)	:1268:SEL..LR[16aa]
P07737:G:94:R (deam*5, am*2, meth*2)	P0CG38 (deam*2, meth)	P13645:H:487:Y	P15954
P07738	P0CG39 (deam)	P13645:I:101:S	P16118
P07741	P0CG47	P13667:KKD..EL[103aa]:	P16219 (deam, am)
P07814:NSK..SY[865aa]:	P0CG48	543:KGR..SS[19aa]	P16401 (deam, am)
648:TVSMKS	P0CG48:G:380:E	P13674 (deam*3, am)	P16870 (deam*2)
P07900-	P0CG48:NIQ..GV[550aa]:	P13693	P16930 (deam*4, am*3, meth)
2:QQS..VD[329aa]:526:H	136:NTS..SS[22aa]	P13693 (deam*2)	P16930-2 (deam*6, am*3, meth)
TK..LY[22aa]	P0CW22 (deam*2)	P13716:WHG..EE[45aa]:2	P16949 (deam, am)
P07902	P10109	86:CAWSPGRGI	P16949-
P07910-3	P10253:H:199:R (deam, am)	P13797-	2:EKM..HH[50aa]:125:V
P07910-4	P10253:LAV..WC[110aa]	2:DII..RV[109aa]:509:DY	ED..LC[19aa]
P07910-	:843:HWLWP	HCELGEQNVE	P16989
4:KME..DS[57aa]:194:M	P10606	P13804-2	P16989-2 (deam)
RW..WS[33aa]	P10606 (deam*2, am)	P13804-2:K:167:T	P17152
P07910:KME..DS[57aa]:2	P10644 (deam*2, am, ac, ph)	P13804-	P17174
50:MRW..WS[33aa]	P10746 (deam)	2:VGA..KK[89aa]:196:C	P17174 (deam*2, am*2)
P07919	P10768 (deam*2, am*2)	WCFPCCC	P17568
P07942:Q:1022:R	P10809 (deam*12, am*5, ac, meth*2)	P13804:VGA..KK[89aa]:2	P17812 (deam, am)
(deam*5, am)	P11021 (deam*15, am*7)	45:CWCFPCCC	P17844
P07947 (deam)	P11117	P13929-3	P17844 (deam)
P08069	P11117:Q:368:H	P13929:N:71:S (deam*3, am)	P17844-2
P08133-	P11142-	P13929:V:85:A (deam*2, am)	P17900:V:153:A
		P13984 (deam, am)	P17980 (deam)

P17987 (deam*9, am*2, ac*3)	P22102:D:752:G	2:YRC..EF[472aa]:285:Y	P31146
P18031	P22234-2	EMWPFDRSLPGSSC	P31150 (deam, am)
P18031 (deam, am)	P22234-	P26639:YRC..EF[472aa]:	P31153 (deam*2)
P18077	2:YEG..NL[110aa]:323:Y	252:YEMWPFDRSLPGS	P31153:SER..KY[55aa]:3
P18085 (deam)	LK..DQ[23aa]	SC	41:SERESY
P18124 (deam, am)	P22234:YEG..NL[110aa]:	P26641:KEE..FK[189aa]:	P31321 (deam, am)
P18669 (deam*10, am*2, ac*2)	316:YLK..DQ[23aa]	249:RGGEKGGCPCS	P31483
P18754	P22307-7	P26885	P31483-2
P19338 (deam*5, am)	P22307-	P26885	P31483-
P19387 (deam*2, am)	8:KMG..KL[144aa]:380:R	P27144	2:AYG..TQ[56aa]:320:AI
P19388:S:44:F (deam)	DGFSGSRQFF	P27348 (deam*4, ac)	WNVWPGMEPARI
P19623 (deam*5, am, ac)	P22307:KMG..KL[144aa]	P27361	P31483:AYG..TQ[56aa]:3
P19971 (deam, am)	:404:RDGFSGSRQFF	P27694 (deam*2, am)	31:AIWNVWPGMEPARI
P20042	P22570-	P27695:TAY..AL[150aa]:	P31689 (deam)
P20073-2	6:LQD..GH[70aa]:382:LA	169:THM..SS[27aa]	P31689:QKN..TS[269aa]:
P20290	G..GP[70aa]	P27707	129:QKECDL
P20290-2	P23246	P27707 (deam)	P31937 (deam*2, am)
P20290-2 (deam, am)	P23246 (deam*2, am, meth*2)	P28065 (deam*2, am*2)	P31943
P20338	P23258:DCV..EQ[252aa]:	P28070:G:10:R	P31943 (deam, am)
P20339	200:DVC..EQ[52aa]	P28070:Q:114:L (deam*3, am)	P31947 (deam*2, am)
P20339-2	P23284	P28072 (deam*2, am)	P31948
P20674 (deam*2)	P23284 (deam*3)	P28074-3	P31948-
P20700:SVR..IM[179aa]:	P23297 (deam*2, am)	P28331-4	3:EFQ..IR[135aa]:385:VV
408:STYN	P23378 (deam, am)	P28482-2	PAGTQGL
P20823-5 (deam*2)	P23381-	P28838	P31948:EFQ..IR[135aa]:4
P20839-5	2:FPA..FQ[132aa]:299:FP	P28838-2	09:VVPAGTQGL
P20839-7	S..GQ[34aa]	P29083 (deam)	P31948:QIA..IR[246aa]:2
P20962	P23396	P29279:H:83:D	98:RDC..FL[26aa]
P21266 (deam*2, am, meth*2)	P23396-2	P29373	P32119 (deam*3, am)
P21266:V:224:I (deam)	P23434 (deam)	P29966	P32321
P21281 (deam*5, am)	P23528	P29966 (am)	P32321-2
P21283	P23588	P29992 (deam*3, am*2)	P32969 (deam*2, am)
P21291 (deam)	P23588-2	P30038-2	P33121
P21333:EED..VP[2490aa]	P23588-	P30040 (deam, am)	P33176 (deam*6, am*2)
:158:RGG	2:LQR..AE[225aa]:348:L	P30041	P33316 (deam, am)
P21333:KVE..VP[1330aa]	SVSWMSQN	P30041 (deam, am)	P33316-2
:1318:TSG..DL[18aa]	P23588:LQR..AE[225aa]:	P30042	P33552
P21333:VRV..VP[1290aa]	387:LSVSWMSQN	P30043 (deam)	P33991:EEL..LL[185aa]:6
:1358:VVS..WL[34aa]	P23610	P30046	79:ERA..AK[25aa]
P21549	P23786:V:368:I	P30049	P33991:PRL..LL[164aa]:7
P21549:IKW..KL[149aa]:	P23919 (deam)	P30050 (deam*2)	00:PRG
244:TQVAGQLLGL	P24385	P30084:T:75:I (deam)	P33992
P21549:SWR..KL[106aa]:	P24539 (deam*2, am)	P30086 (deam)	P33992:TQA..LK[188aa]:
287:TLA..WL[44aa]	P24539:I:84:V (deam*3, am)	P30101 (deam*10, am*4, ac, meth*2)	547:TLRLWRARLTWPS
P21695	P24844	P30153 (deam*6, am*2)	P34059
P21695-2	P24844:EVD..DD[35aa]:1	P30154-4	P34896
P21796 (deam*8, am*2, meth*2)	38:ASGRDVPGGTH	P30405 (deam*2, am)	P34896-2
P21912 (deam)	P24928-2	P30419	P34896-2:K:309:*
P21953 (deam)	P25205:E:777:K	P30520 (deam*2, am*2)	P34896:K:348:*
P21964-	P25311	P30533 (deam)	P34897-
2:HVR..GP[40aa]:182:RR	P25398 (deam*2, am, ac)	P30613	3:QGG..EH[181aa]:303:Q
AREQLL	P25398 (deam*2)	P30613-2	WGAPTIMPLLQ
P21964:HVR..GP[40aa]:2	P25685 (deam*2, am)	P30613-2:A:123:AN	P35080 (deam, am)
32:RRAREQLL	P25685-2	P30613:A:154:AN	P35080-2 (deam*2, am)
P21980	P25786	P30740 (deam*3, am*2, ac)	P35218
P21980 (deam)	P25786-2	P30740:NKK..SP[192aa]:	P35241
P22033 (deam*2, am)	P25787:G:5:E	188:I	P35241-4
P22090	P25789 (deam)	P31040	P35241-
	P26639	P31040:GK:181:E	4:ERR..AM[192aa]:256:R
	P26639-		ASSC
			P35241:ERR..AM[192aa]:
			392:RASSC

P35244 (deam)	P41223	P48444	2:KRA..TR[122aa]:388:R
P35249	P41236 (deam)	P48444-2	EG..IN[20aa]
P35250 (deam, am)	P41567	P48449	P50991:KRA..TR[122aa]:
P35268	P42126 (deam*2)	P48507	418:REG..IN[20aa]
P35270 (deam ,am)	P42166 (deam*2, am)	P48507 (deam, am)	P50995
P35558:V:184:L (deam*2, am)	P42166:QAL..KH[82aa]:6	P48637	P50995:R:230:C
P35579:SFR...AE[621aa]:1	13:QCAWDSEQNI	P48723	P51148
P340:SPS..PT[30aa]	P42167 (deam, am)	P48735	P51148-2
P35606:AWA..DD[512aa]	P42167-3	P48735:R:451:W	P51149
:395:DMG..DI[21aa]	P42224 (deam)	P48739-2	P51452 (deam)
P35754 (deam*2, am, meth)	P42285	P49006 (deam)	P51571
P35813	P42345	P49247 (deam*3, am)	P51580 (deam, am)
P35900	P42574 (ac*4)	P49327 (deam*7, am)	P51610:VSW..GQ[1654aa]
P35914	P42574 (deam, am)	P49368	:382:SEL..EI[19aa]
P35998 (deam)	P42677	P49368:N:61:T	P51659-
P36404 (deam)	P42704	P49406 (deam*3, am)	2:GLF..KL[502aa]:260:G
P36405 (deam*4, am)	P42704:R:450:L	P49427	CL..LL[22aa]
P36405:GEG..KK[20aa]:1	P42765 (deam*6, am*2)	P49441	P51659:NIG..KL[78aa]:65
63:ERG..EE[20aa]	P42785	P49441:RSR..HT[24aa]:3	9:IYWG
P36507 (deam*3, am*2)	P42785-2	76:SDPGSGWRHS	P51665
P36542	P42785:DDC..QH[105aa]:	P49448:DKI..FT[145aa]:4	P51665 (deam*2, am)
P36639-2	392:DLTVFNSGV	14:ERS..ST[16aa]	P51687:NVQ..SP[30aa]:5
P36639-4	P43034 (deam)	P49458	16:TCA..LP[30aa]
P36873	P43243 (deam*6, am*2)	P49459	P51688
P36873-2	P43490 (deam*3, am, ac*4)	P49585	P51812
P36915 (deam, am)	P43686	P49588 (deam*4, am)	P51948-2 (deam)
P36957 (deam, am)	P43686-	P49589-	P51956
P37198 (deam*2, am)	2:LED..YK[59aa]:329:LR	2:LQY..RS[274aa]:453:L	P51991
P37235 (deam)	RLCGPAR	SI	P51991 (deam*4, am*4, ac*3, ph, meth*3)
P37268	P45973	P49591	P51991-2 (deam*4, am*4, ac*2, meth*2, ph)
P37837:TERMFNAENG	P46060	P49593:L:420:R (deam, am)	P52209
K:327:TRTNVQCREWK	P46063	P49642 (deam, am)	P52209:AK:47:AKK
P38117:AGV..RI[21aa]:2	P46087-3	P49662-2	P52272-2
35:ARRQGGDH	P46108-2	P49675	P52435 (ac*4)
P38117:R:233:P	P46108-2:N:144:D	P49720 (deam*5, am*2)	P52597 (deam*3, am)
P38117:V:208:M	P46109	P49721 (deam*2, am*2, ac*3)	P52758 (deam*2)
P38646 (deam*15, am*3)	P46379-	P49753	P52888 (deam*3, am)
P38919 (deam)	4:PVV..DP[457aa]:447:H	P49755	P52907
P39019	RG..GS[115aa]	P49756	P52926 (deam*3, am)
P39019 (deam*3, am*2)	P46459 (deam*2, am)	P49756 (deam, am)	P52948-
P39687 (deam, am)	P46734-2	P49770	3:GTT..GQ[554aa]:384:V
P39748	P46736	P49770 (deam*2)	QPVAGSLVTNP
P40121-2:H:320:R (deam, am)	P46776 (deam, am)	P49773 (deam*3, am)	P53384
P40222	P46778	P49789 (deam)	P53396-
P40222 (deam*2)	P46783	P49914 (deam*2, am)	2:AIG..SM[416aa]:676:A
P40227-2:Q:389:R	P46783 (deam, am)	P50135 (deam*4)	SL..SK[27aa]
P40227:Q:434:R	P46926 (deam*2, am)	P50150	P53582
P40306 (deam)	P46940	P50225:V:223:M (am)	P53597 (deam)
P40855	P46940:MOV..GK[604aa]	P50226 (deam*3, am)	P53602 (deam*3, am)
P40926-	:1054:IWL	P50336	P53611
2:GKK..LK[44aa]:253:GE	P47224	P50395 (deam*7, am*2)	P53618 (deam*2, am)
K..LF[17aa]	P47755 (deam, am)	P50402 (deam)	P53634:I:453:V (deam)
P40926:GKK..LK[44aa]:2	P47813 (deam)	P50416	P53701 (deam)
95:GEK..LF[17aa]	P47929	P50453 (deam*3, am)	P53999 (deam)
P40926:LGK..LK[45aa]:2	P48047 (deam, am)	P50542-3	P54105
94:LVG..RR[22aa]	P48147	P50552 (deam*4, am)	P54252-2
P40938 (deam ,am)	P48147:KEG..IP[172aa]:5	P50583	P54577:KLA..IS[194aa]:3
P41208 (deam*2, am)	39:TRKVTHTLPRG	P50914 (deam, am)	35:TTG..ED[58aa]
P41214 (deam, am)	P48147:LDG..IP[249aa]:4	P50990 (deam*8, am)	P54578:EIMEESEQ:486
	62:FWM..FL[31aa]	P50991-	:EYNGRGRK

P54652 (deam*4, am)	P61026	P62826 (deam*3, am*2,	P82664
P54727	P61081	meth)	P82673
	P61081 (deam)	P62829	P82675 (deam)
P54727:QEA..ED[77aa]:3	P61088	P62834	P82909 (deam, am)
33:QKL..VI[24aa]	P61106 (deam*2, am)	P62834 (deam, am)	P82932 (deam, am)
P54819:L:20:P	P61158 (deam*2, am)	P62841	P82933
P54920 (deam*4, am)	P61158:KLS..MS[71aa]:3	P62851	P82979
P55036 (deam*2, am)	48:NN	P62854	P82980 (deam, am)
P55060-	P61160 (deam*2, am)	P62857 (deam)	P83436
4:QPK..LL[158aa]:758:H	P61201	P62875 (deam, am)	P83436 (deam*2)
TKNVWNGFGKNYYYS	P61225 (deam, am)	P62888 (deam, am)	P83731
P55072:FLI..YG[540aa]:2	P61247 (deam, am)	P62906	P83876 (deam)
67:SLDQWS	P61254	P62910 (deam, am)	P83881
P55081 (deam*3, am)	P61289-2	P62917	P83916
P55084:ANQ..PK[334aa]:	P61289-3	P62937-2 (deam*4, am)	P84095
141:APS..PV[16aa]	P61326	P62942 (deam, am)	P84098
P55157:D:384:A (deam)	P61326-2 (deam)	P62987 (deam, am)	P84157-2
P55199 (deam, am)	P61421 (deam)	P62993	P98179 (deam*2, meth)
P55212	P61457 (deam)	P63000	Q00059
P55263-3:N:148:H	P61513	P63104	Q00266 (deam*2)
P55735	P61599	P63104 (deam*6, am*2,	Q00403 (deam)
P55735-3	P61604 (deam)	ac*4)	Q00534
P55769	P61803 (deam, am)	P63146	Q00535 (deam)
P55769 (am)	P61916	P63151	Q00577 (deam*2)
P55786:LAR..TV[295aa]:	P61923-5	P63151-2	Q00688 (deam, meth)
625:LLELESLAL	P61970	P63167 (deam ,am)	Q00796
P55789	P61981 (deam*5, am)	P63172 (deam*2, am,	Q00839-
P55795 (ph)	P62072 (deam, am)	ac*3)	2:RGS..YY[64aa]:743:R
P56192:EYL..KK[316aa]:	P62081	P63173	W
585:DST	P62136 (deam*3, am)	P63208 (deam, am)	Q01082:AVV..KK[1885a
P56192:RIK..KK[174aa]:	P62140 (deam*3, am)	P63220 (deam)	a]:480:ACG..DG[66aa]
727:RILKAVRLTGNGQ	P62195	P63241	Q01085 (deam, am)
EQ	P62195-2	P63244 (deam*4, am*2)	Q01085-2 (deam*2, am)
P56385	P62241 (deam*3, am)	P63272	Q01105 (deam*2, am)
P56556	P62244 (deam*2, am)	P67809	Q01105-2
P56962	P62249 (deam*2)	P67809 (deam)	Q01415-2 (deam*2, am)
P57105 (deam)	P62253 (deam, am, ac)	P67809:PEG..AE[147aa]:	Q01433-
P58546	P62256 (deam, am)	178:PRP..GC[43aa]	2:KRA..PQ[445aa]:354:N
P59998	P62263:RKGGRGRRL:	P67809:QNS..AE[162aa]:	GQ
P59998-3	142:REGGSPWSPS	163:RE	Q01469
P60174-1	P62266	P67936 (deam)	Q01469 (deam*2, am,
P60174-4	P62277	P68036	meth*2)
P60510 (deam*3, am)	P62280	P68363	Q01518-
P60604-2	P62304	P68402	2:NTT..AG[118aa]:357:T
P60763	P62312 (deam, am)	P68402 (deam, am)	HDIANQGQN
P60842 (deam*4, am)	P62316 (deam, am)	P78318 (deam)	Q01518:NTT..AG[118aa]:
P60842:RRK..LI[125aa]:2	P62330 (deam*2, am)	P78330	358:THDIANQGQN
82:RKG..ET[32aa]	P62333 (deam*3, am)	P78330 (am)	Q01658
P60866-2	P62424 (deam*2, am)	P78346	Q01850
P60891 (deam, am)	P62487 (deam, am)	P78347-4	Q01995
P60903 (deam)	P62491:E:186:D	P78406	Q02127 (deam*2, am)
P60953 (deam*3, am)	P62633-4	P78406 (deam*4, am)	Q02338
P60981-2	P62633-7	P78417-3:A:112:D	Q02543
P60983	P62701	P78540 (am)	Q02790 (deam, am)
P61006	P62701 (deam)	P78540 (deam)	Q02809:DYE..DP[54aa]:6
P61011 (deam)	P62714 (deam*3)	P78549 (deam*3, am)	74:EIT..SI[54aa]
P61019	P62750	P78560 (deam, am)	Q02818 (deam*3, am)
P61019-	P62753	P80297 (ac*3, meth*2,	Q02878
2:QAGGGCC:182:QRLG	P62760 (deam, am)	ph)	Q02978 (am*2)
AAA	P62820:QSGGGCC:199:L	P80404	Q03393 (am)
P61020	SQVEVA	P82650 (deam*3, am)	Q03701:V:102:I
P61024 (deam)	P62826 (deam*2, am)	P82663 (deam)	Q04637:MAK..HN[576aa

]:1024:MMP..GD[68aa]	Q12907	Q13724-2:GS:152:GSS	Q14697-2
Q04828:RNV..EY[23aa]:3	Q12907 (deam, am)	Q13724-	Q14697-2:A:784:D
01:RMCDI	Q12931-2:R:254:G	2:LHA..DY[297aa]:435:L	Q14697-
Q04837 (deam*2)	Q12931:KEE..RH[145aa]:	RM..PY[33aa]	2:KRR..LR[486aa]:481:R
Q04917 (deam*4, am)	560:TRR..PG[27aa]	Q13724-2:N:176:Y	EA..VC[32aa]
Q05639	Q12965 (deam, am)	Q13795	Q14697-
Q05682-	Q12972	Q13813-	2:REP..LR[261aa]:706:PR
4:KLK..KV[241aa]:298:R	Q13045-	3:LQL..VN[1026aa]:1427:	AMAVTISAQ
AQ..RD[56aa]	3:NNL..LA[943aa]:316:N	PAVPSGL	Q14697-
Q05682-	TT..QS[16aa]	Q13825 (deam)	2:RVV..LR[54aa]:913:R
5:KLK..KV[241aa]:292:R	Q13045:NNL..LA[943aa]:	Q13867	WW
AQ..RD[56aa]	327:NTT..QS[16aa]	Q13867:VDR..AE[32aa]:4	Q14697:EHA..LR[514aa]:
Q06203:LGA..EW[66aa]:	Q13098-	24:CGQEACP	431:DTC
452:LKEQTVLCICQ	7:AAG..NM[240aa]:288:	Q13884	Q14697:KRR..LR[486aa]:
Q06323 (deam*2)	AVQ..TR[97aa]	Q13884 (am)	459:REA..VC[32aa]
Q06323 (deam*2)	Q13123 (am)	Q13895:KKA..VE[161aa]	Q14697:REP..LR[261aa]:
Q06520 (deam*3, am)	Q13148 (deam*6, am*3,	:277:REG..CG[33aa]	684:PRAMAVTISAQ
Q06546 (deam*2, am)	meth*3)	Q13951	Q14697:RVV..LR[54aa]:8
Q06587	Q13151	Q13951-2 (deam*2)	91:RWW
Q07021	Q13151 (am)	Q14011 (deam, am)	Q14739:S:154:N
Q07021 (deam)	Q13158	Q14012	Q14847:GWM..AI[24aa]:
Q07666-	Q13162	Q14012 (deam*3)	238:ALD..GH[24aa]
3:TVR..RY[106aa]:299:T	Q13162 (deam*3, ac,	Q14019 (deam, am)	Q14974:AWT..QA[448aa]
CE..NI[29aa]	meth)	Q14061 (am)	:429:AMDCRQNL
Q07666:TVR..RY[106aa]:	Q13185 (deam, am, meth,	Q14103	Q15003-
338:TCE..NI[29aa]	ph)	Q14103-3	2:PQK..GD[117aa]:489:P
Q07866-6	Q13185 (deam, am)	Q14137	EGK
Q07866-6:HE:412:H	Q13200-	Q14152:DRV..RR[428aa]:	Q15007 (am)
Q07960	3:DVC..DL[313aa]:466:D	955:RSGSPAWHG	Q15018 (deam, am)
Q07960 (am)	VCVHMQALGMC	Q14152:PRD..RR[141aa]:	Q15019
Q08117	Q13217 (deam*3)	1242:L	Q15019-3
Q08257-2	Q13232	Q14192 (deam)	Q15019-
Q08380 (ac)	Q13242 (deam*2, am, ph)	Q14197	3:KRG..HV[57aa]:315:RE
Q08380:SSW..VD[165aa]	Q13263-	Q14203-	RRQESGE
:421:SLL..FF[24aa]	2:AEA..GP[434aa]:320:D	3:KKI..IS[475aa]:762:RE	Q15019:KRG..HV[57aa]:
Q08426:EVG..KL[478aa]:	RGLWQDCGRASWH	D..GI[26aa]	305:RERRQESGE
246:ESG..AG[20aa]	Q13263-	Q14203-5	Q15031
Q08945 (deam)	2:GSS..GP[398aa]:356:A	Q14203-	Q15046
Q08J23	QQ..RR[18aa]	5:KKI..IS[475aa]:665:RE	Q15046-2
Q08J23-2	Q13263-	D..GI[26aa]	Q15061
Q08J23-2:R:195:M	2:LTE..GP[61aa]:693:LT	Q14204:MSH..TE[4012aa]	Q15061:SKR..EE[251aa]:
Q09666:G:2527:D	L..RA[22aa]	:635:MRV..GS[24aa]	427:SRG..WI[20aa]
Q0VDF9 (deam*2, am)	Q13263:GSS..GP[398aa]:	Q14232	Q15084-
Q10472 (deam)	438:AQQ..RR[18aa]	Q14232 (deam*2)	2:DGR..EL[28aa]:465:DR
Q10713:NVL..FR[160aa]:	Q13283	Q14241	QGWRASRGG
366:TRAQQAPLDV	Q13287 (deam, am, ac)	Q14247 (am)	Q15102
Q12788	Q13315:V:2906:I	Q14258	Q15102:RAA..AP[72aa]:1
Q12797-	Q13347 (deam*6, am*2)	Q14320	60:QGG..PS[29aa]
10:QQF..AI[329aa]:401:H	Q13404	Q14376	Q15119
TL..SQ[23aa]	Q13409-2	Q14444-2	Q15126 (deam, am)
Q12824-	Q13451 (deam*2, am)	Q14457:KV:214:KEV	Q15155 (deam)
2:DQW..AW[39aa]:338:D	Q13485	Q14554 (deam*2)	Q15170-2
AVVPTAGDSRR	Q13501	Q14653 (deam, am)	Q15185
Q12840:E:881:V	Q13526	Q14657	Q15185-4
Q12874 (deam*6, am*3,	Q13564-2	Q14677 (deam)	Q15257-2
ac*4)	Q13572	Q14678-	Q15257-2:S:322:L
Q12904	Q13596:EAE..IS[91aa]:43	2:ETR..FD[640aa]:555:E	Q15293 (deam*4, am)
Q12904-2	2:DRP..LG[29aa]	YA..WV[36aa]	Q15363
Q12904-2:V:270:G	Q13630	Q14683 (deam*2)	Q15365
Q12904:N:195:K	Q13643 (deam, am)	Q14690:L:1216:F (deam)	Q15365 (deam)
Q12904:V:246:G	Q13685 (deam)	Q14692	Q15366-5
Q12906-2	Q13686 (deam)	Q14696 (deam*3, am*2)	Q15366-

8:VVM..SS[159aa]:164:V	Q16740 (deam, am)	Q66LE6	Q7Z2W9 (deam)
CH.FS[84aa]	Q16762 (deam*3)	Q66LE6:KVVW..IN[139aa]	Q7Z3B4
Q15369 (deam)	Q16762:FEK..KA[76aa]:2	:315:KCG..PS[49aa]	Q7Z3B4-3
Q15382 (deam, am)	22:FRE..QA[46aa]	Q676U5:ALC..QY[294aa]	Q7Z3J3:EKV..EE[405aa]:
Q15397 (deam*2)	Q16774	:314:AIVCLRCT	1354:EEK..VI[29aa]
Q15404 (deam)	Q16775-2	Q69YN2	Q7Z417 (deam, am)
Q15404:KYL..NR[35aa]:2	Q16775-2:AI:26:AAI	Q6DKJ4 (deam)	Q7Z4G1 (deam)
43:NIP..EE[18aa]	Q16795	Q6DKJ4-3	Q7Z4H8 (deam*2)
Q15404:KYL..NR[35aa]:2	Q16836	Q6GMV2	Q7Z4W1 (deam, am)
43:NTS..RT[34aa]	Q16836-3	Q6GMV3 (deam, am)	Q7Z4W1 (deam*3, am)
Q15427	Q16851-2	Q6I9Y2 (deam, am)	Q7Z5K2:V:124:I (deam)
Q15428 (deam)	Q16851-2:G:104:GL	Q6IAA8	Q86SX6:A:146:T
Q15493 (deam)	Q16891-2:Q:520:R	Q6IBS0	(deam*2)
Q15528-2	Q16891:Q:531:R	Q6ICL3	Q86SZ2 (deam)
Q15545 (deam*2, am)	Q2KHR2:I:204:T	Q6IN84	Q86TX2 (deam*3, am)
Q15555-2	Q2NL82 (deam, am)	Q6N069 (deam)	Q86U28
Q15599	Q2Q1W2 (deam, am)	Q6NSJ5 (deam)	Q86U90
Q15628	Q2TAA5	Q6NUK1-2 (deam)	Q86UE4 (deam)
Q15637-	Q2TAY7 (deam*2, am)	Q6NVY1:T:46:A	Q86UY8
7:PPS..QN[285aa]:329:H	Q2VIR3:LAF..DD[265aa]	Q6NXE6	Q86V88
RLSCLPPRAAHPG	:208:LRLSKVQ	Q6P1J9 (deam)	Q86VM9
Q15643	Q32P51:TVK..RF[152aa]:	Q6P1L8	Q86VP6
Q15645	169:TEGPQL	Q6P1N9 (deam*2)	Q86VS8
Q15649	Q3ZCW2 (deam*2)	Q6P1X6 (deam*2)	Q86VU5 (deam)
Q15650 (deam*2, am)6	Q3ZCW2:QITKLG:167:Q	Q6P2E9:RPR..LP[620aa]:	Q86W56
Q15691	DHQA W	782:RPP..QP[34aa]	Q86W92-
Q15691 (ac)	Q4VC31 (deam*2, am)	Q6P2I3	4:VLA..NV[668aa]:313:V
Q15750	Q4VXU2:TGM..MH[47aa	Q6P587 (am)	HW..LQ[40aa]
Q15750:GED..AP[32aa]:4	]:568:TRHAAGD	Q6P587-3	Q86WQ0 (deam)
73:ARGRSC	Q53F19	Q6PI48	Q86WR0
Q15758	Q53FT3 (deam)	Q6PL18	Q86WW8
Q15773	Q53GG5-2	Q6PL24	Q86X76-5
Q15776 (deam, am)	Q53H12	Q6PUV4 (deam, am)	Q86XA9-3 (deam, am, ac)
Q15785 (deam*4, am*3)	Q53H82 (deam)	Q6QNY0	Q86XP3:EEL..DS[427aa]:
Q15813	Q53H96:K:150:N	Q6UW78:G:89:S	512:EES
Q15813-	Q53HC9 (deam*2, am,	Q6UWP2	Q86XR7-2
2:IGQ..RW[166aa]:413:I	ac*3)	Q6UWS5 (deam)	Q86Y37 (deam)
AAS	Q53RE8	Q6UX04	Q86Y39
Q15819 (deam*2, am)	Q53S33 (deam)	Q6UXH0	Q86Y82
Q15833-3:P:228:PR	Q53T59 (am)	Q6UXN9 (deam*2, am)	Q86YH6
Q15907	Q5BKU9	Q6Y7W6-	Q86YM7
Q15907-2	Q5BKZ1 (deam, am)	3:KEL..DY[415aa]:906:K	Q86YS6
Q15942	Q5EBL4	RA..VG[63aa]	Q8IU81 (deam, am)
Q16186	Q5HYI8	Q6YP21-3	Q8IUF8 (deam, am)
Q16186:PAG..LD[131aa]:	Q5HYK3 (deam, am)	Q6YP21-3:I:57:V	Q8IU18 (deam, am)
277:PRP..VC[16aa]	Q5JS37 (deam)	Q6ZN17 (deam)	Q8IUR0:S:52:A (deam,
Q16378:Q:120:R	Q5JVS0	Q70UQ0-4 (deam*4, am)	am)
Q16378:R:96:Q (deam,	Q5R3I4:F:243:L (am)	Q7KZF4:ADI..SR[406aa]:	Q8IV08 (deam, am)
am)	Q5RKV6	505:ARY..SF[23aa]	Q8IVD9
Q16401	Q5SRE7 (deam*2, am)	Q7KZF4:GKK..SR[499aa	Q8IVS2 (deam)
Q16401-2	Q5T0D9	]:412:GIRRS M	Q8IVS8
Q16514	Q5T2W1 (deam, am)	Q7L0Y3	Q8IWA0 (deam)
Q16531	Q5T653	Q7L0Y3 (deam*3, am)	Q8IWB7
Q16531:DFL..IH[222aa]:9	Q5T6V5 (deam*2, am)	Q7L1Q6-3	Q8IWE2-2
19:DLH..KL[24aa]	Q5TFE4	Q7L2H7:DCK..DT[201aa	Q8IWE2:G:84:R
Q16539-2 (deam, am)	Q5U5X0 (deam, am)	]:174:DV	(deam*3, am)
Q16543 (deam*3, am)	Q5VIR6-4 (deam, am)	Q7L3T8	Q8IWL3 (deam, am)
Q16576-2	Q5VT66:T:165:A	Q7L5N1 (deam*2)	Q8IWX8
Q16595 (deam, am)	(deam*2, am)	Q7LBR1	Q8IWX8:ERL..CK[233aa
Q16658 (deam*2, am)	Q5VZF2-2	Q7LG56	]:684:GEA..EQ[22aa]
Q16698-2	Q5W111 (deam, am)	Q7RTV0	Q8IX12:RRL..SV[858aa]:
Q16706 (deam*2, am)	Q5XKP0	Q7Z2K6	293:RHD

Q8IXJ6-2	Q8TBE9	Q92905 (deam*2, am)	Q96EK6
Q8IXM3	Q8TBF4	Q92922	Q96EK6 (deam, am)
Q8IXT5	Q8TBX8	Q92947 (deam*3, am)	Q96EK9 (ac)
Q8IYB8	Q8TBX8-3	Q92979 (am)	Q96EL3
Q8IYK4	Q8TCD5 (deam, am)	Q93009:KER..HN[547aa]:	Q96EQ0 (deam*2, am)
Q8IYS1	Q8TCS8:I:121:V	556:LRSGRKPISICKCR	Q96ER9:TFV..AS[16aa]:3
Q8IYS1 (deam)	(deam*2, am)	Q93034	96:TLC..SQ[16aa]
Q8IZP0-	Q8TD30	Q93052 (deam*2)	Q96EX1
12:MSG..TD[294aa]:203:I	Q8TD30-2	Q93062	Q96EY7
SQ..TI[29aa]	Q8TD30:REG..YA[37aa]:	Q969G6	Q96F10 (deam*2, am)
Q8N128-2	487:RRH..VR[36aa]	Q969H8	Q96F63
Q8N163 (am)	Q8TDL5	Q969M7	Q96FJ2 (deam*2)
Q8N163:RRD..SN[202aa]	Q8TDN6	Q969S3 (deam)	Q96FK6 (deam, am)
:722:RET	Q8TDP1 (deam, am, ac)	Q969S9-2	Q96FN9 (deam, am)
Q8N1Q1	Q8TDQ7-3:D:177:Y	Q969T9 (deam, am)	Q96FQ6 (deam*2, am*2)
Q8N2Z9	Q8TE02 (deam)	Q969Y2	Q96FW1 (deam*4, am)
Q8N335	Q8TEA1 (deam, am)	Q969Y2:V:250:A	Q96FX7 (deam, am)
Q8N392:HAL..PL[174aa]:	Q8TEA8	Q96A35	Q96FX7 (deam*3, am)
490:HMHW	Q8TEA8 (ph)	Q96A49	Q96FZ7
Q8N392:TLK..PL[211aa]:	Q8WUD4	Q96A72	Q96G03:E:298:EK
453:NTEGPS	Q8WUD4 (deam*2, am,	Q96A72:D:83:Y	Q96G03:E:488:D (deam,
Q8N490-2 (deam, am)	ac)	Q96AB6 (deam*2, am)	am)
Q8N4P3	Q8WUF8-3	Q96AE4 (am)	Q96GA7
Q8N4Q0	Q8WUJ0	Q96AE4-2 (deam*3,	Q96GD0
Q8N543	Q8WUM0:I:294:V	am*2)	Q96GD0:TEGLED:291:T
Q8N543	(deam*2, am)	Q96AG4 (deam, am)	RGVGG
Q8N5K1	Q8WUW1 (deam, am)	Q96B26 (deam)	Q96GG9
Q8N5M1 (deam*3, am)	Q8WUX2 (deam, am)	Q96B54	Q96GW9 (deam)
Q8N5M4 (deam, am)	Q8WUX9	Q96BJ3	Q96GX9 (deam*2, am)
Q8N5N7 (deam, am)	Q8WV22	Q96BN2 (deam*2, am)	Q96H79 (deam)
Q8N668 (deam, am)	Q8WV74	Q96BN8 (deam)	Q96HE7 (deam*4, am)
Q8N6T3-5	Q8WV92	Q96BR5	Q96HP4 (deam)
Q8N806	Q8WVE0 (deam, am)	Q96BR5 (deam, am)	Q96HQ2 (am)
Q8N8N7	Q8WVJ2	Q96BW1 (deam, am)	Q96HY6 (deam, am)
Q8N8S7	Q8WVM0	Q96BW5	Q96HY6:A:303:T (deam,
Q8N8S7 (deam)	Q8WVY7	Q96BW5:EEG..FK[64aa]:	am)
Q8N9N2	Q8WWV3	286:VKRAVKIEFW	Q96I15-2:A:175:T (deam,
Q8N9N2-2	Q8WX92	Q96C01 (deam, meth*2)	am, meth)
Q8N9N7 (deam)	Q8WXA9-2	Q96C19	Q96I25
Q8N9N8 (deam)	Q8WXX5 (deam)	Q96C86 (ac)	Q96I99
Q8NAV1 (deam)	Q92520	Q96C90 (deam)	Q96IU4
Q8NBJ5	Q92530:TSR..YL[94aa]:1	Q96CD2	Q96IU4:G:27:S
Q8NBK3-4	78:TQSAASLV	Q96CM8-	Q96IU4:G:27:S (deam)
Q8NBN7 (deam, am)	Q92572	2:KKV..NL[448aa]:193:R	Q96IU4:IMK..LQ[29aa]:1
Q8NBX0	Q92575	EG..PG[63aa]	82:IHE..AP[29aa]
Q8NC60	Q92599	Q96CN7	Q96IU4:IMK..LQ[29aa]:1
Q8NE71:GRR..RE[517aa]	Q92665	Q96CN7:R:96:C	82:IT
:329:GAATGW	Q92665 (deam*2)	Q96CT7	Q96IY1 (deam)
Q8NEV1:AME..QQ[74aa]	Q92688	Q96CU9-	Q96IY1 (deam*2)
:318:DHG..FI[20aa]	Q92747	3:FQD..II[104aa]:369:FL	Q96IZ0 (deam, am)
Q8NEZ5	Q92769	RTRCGPIWP	Q96J01
Q8NFH3 (deam*3, am)	Q92769:R:50:I	Q96D46 (deam)	Q96J01:PKR..DS[35aa]:3
Q8NFH4 (deam)	Q92785	Q96DG6	17:HQKASAGICL
Q8NFAQ (deam, am)	Q92786	Q96DH6	Q96JH7
Q8NFAV4-4	Q92820 (deam)	Q96DH6 (deam, ac*2)	Q96JM3
Q8NFW8 (deam, am)	Q92879-	Q96DH6-2	Q96K17
Q8NHP8 (deam)	4:TME..PY[314aa]:199:T	Q96DV4 (deam, am)	Q96KP1 (deam, am)
Q8NI36 (deam, am)	HGGLLITHGGKIC	Q96E11-3	Q96KP4:KNV..KD[46aa]:
Q8TAA5 (deam*3, am)	Q92882 (deam)	Q96E39	430:KAT..TT[28aa]
Q8TAE8 (deam*4, am*2)	Q92888:FFR..CT[669aa]:	Q96EH3 (deam)	Q96LA8
Q8TB37 (deam)	244:FSSGKR	Q96EK5:IIT..LT[65aa]:55	Q96LJ7
Q8TBC4 (deam*2, am*2)	Q92890	7:IKL..RK[37aa]	Q96MF7

Q96MH2	148:TKWSRGSCWKSVP	Q9BUE0	Q9BZX2:S:26:R
Q96MW1 (deam, am)	Q99961	Q9BUF5 (am)	Q9BZZ5-2
Q96MW5 (deam)	Q99986 (deam)	Q9BUF5:RHG..DG[141aa	Q9C005
Q96NB3:H:99:Q	Q9BPW8	]:306:RMAAT	Q9C075 (deam)
Q96NB3:H:99:Q	Q9BPW8 (deam*3)	Q9BUH6	Q9C0B1
(deam*2)	Q9BPX5	Q9BUI4 (deam, am)	Q9C0B1:GQLLEAKP:49
Q96NC0	Q9BQ04 (deam, am)	Q9BUJ2-	8:VSASGSKT
Q96NT0 (deam, am)	Q9BQ61	3:DEE..TQ[320aa]:433:D	Q9GZL7
Q96NU7	Q9BQ67 (deam*2, am)	ERKGRMSQIMRS	Q9GZN8
Q96P11-2 (deam, am)	Q9BQ69 (am)	Q9BUJ2-	Q9GZP4 (deam)
Q96PE2	Q9BQ87	4:MDK..TQ[398aa]:359:	Q9GZS3 (deam)
Q96PE7 (deam*2, am)	Q9BQ95 (deam)	MIRCG	Q9GZT6-2
Q96PU8	Q9BQA1:TVG..TE[28aa]:	Q9BUL8	Q9GZT8
Q96PU8-3	315:TSG..CY[27aa]	Q9BUR4	Q9GZT8-2
Q96PZ0	Q9BQE3:IYD..EY[241aa]	Q9BUR5	Q9GZU8
Q96PZ0:NHF..LR[319aa]:	:209:TMT..LP[33aa]	Q9BV19 (deam, am, ac)	Q9GZZ1 (deam*4, am*2)
343:TTTSLLFSEI	Q9BQP7 (deam*2, am)	Q9BV20	Q9GZZ9
Q96QR8 (deam*3)	Q9BR61	Q9BV20 (deam*2, am)	Q9H061
Q96RS6	Q9BR76	Q9BV86 (deam*2, am)	Q9H078-5
Q96S19	Q9BRF8 (deam*3, am)	Q9BVA1 (deam*4, am)	Q9H098 (deam, am)
Q96S19-4	Q9BRF8:K:241:R	Q9BVI4 (deam, am)	Q9H0A8
Q96S44 (deam)	(deam*2, am)	Q9BVK6:WSI..LV[30aa]:	Q9H0C8 (deam, am, ac)
Q96S59	Q9BRG1	206:CVH..LL[24aa]	Q9H0E2
Q96SI9-2	Q9BRJ6 (deam, am)	Q9BW61	Q9H0E2:H:92:RH
Q96SL4 (deam, am)	Q9BRJ7	Q9BW66	Q9H0E2:ITG..EP[73aa]:2
Q96SZ5	Q9BRJ7 (deam, am)	Q9BW71	02:IHRDARCL
Q96T23:QKE..QL[713aa]	Q9BRP4 (ac)	Q9BW85	Q9H0F6 (deam, am)
:729:QER..NI[54aa]	Q9BRP8 (deam, am, ac)	Q9BW92 (deam, am)	Q9H0N5
Q96TC7	Q9BRQ6 (deam, am)	Q9BWF3:QYADRARYS	Q9H0R4
Q99417	Q9BRQ6:A:95:S	AF:354:QCRSGAVLSL	Q9H0R6
Q99436 (deam*2, am)	(deam*2, am)	Q9BWJ5 (deam, am)	Q9H0T7 (deam, am)
Q99442 (deam, am)	Q9BRQ8	Q9BWM7 (deam, am)	Q9H0U4
Q99447	Q9BRT9	Q9BWT6 (deam, am)	Q9H0U4 (deam, am, ac)
Q99447-	Q9BRX5 (deam*2)	Q9BWU0:P:139:T (deam)	Q9H115
3:IYV..DF[174aa]:234:TL	Q9BS26 (deam*4, am)	Q9BX40	Q9H173 (deam, am)
C..TL[36aa]	Q9BSB4	Q9BX68 (deam)	Q9H1D9 (deam, am)
Q99447:K:158:N	Q9BSD7 (deam*2, am)	Q9BXB4 (deam, am)	Q9H1E3
Q99470 (deam*3, am*2)	Q9BSE5:G:105:R	Q9BXJ9 (deam, am)	Q9H1Y0 (deam, am)
Q99471 (deam*3, am*2)	Q9BSF4	Q9BXR0	Q9H204 (deam, am)
Q99496 (deam, am)	Q9BSH4 (deam)	Q9BY44-	Q9H299 (am)
Q99497 (deam*2, am)	Q9BSH5	3:KIW..GI[198aa]:363:KL	Q9H2H8 (deam*3, am*2)
Q99519	Q9BSH5:G:146:E	ALYWLYLAQV	Q9H2I8
Q99541 (deam)	(deam*3, am)	Q9BY44-	Q9H2P9
Q99567 (deam, am)	Q9BSY4 (am)	3:LWN..GI[359aa]:202:L	Q9H2W6 (deam*2, am)
Q99570	Q9BT09	E	Q9H361:QNK..TV[190aa]
Q99584	Q9BT30 (deam)	Q9BY49 (ac)	:442:QPI..VL[19aa]
Q99584 (ac)	Q9BT73 (deam)	Q9BYC8 (deam, am)	Q9H3G5:A:435:V
Q99598 (deam*6, am*2)	Q9BT78 (deam, am)	Q9BYD1	Q9H3P2
Q99614 (deam*3, am*3)	Q9BTE7 (deam)	Q9BYD3	Q9H3P7
Q99627	Q9BTL3	Q9BYD6	Q9H3R5
Q99653 (deam, am)	Q9BTT0	Q9BYG3	Q9H410-4
Q99714-2	Q9BTT0-2	Q9BYN0 (deam, am)	Q9H410-4:C:49:W
Q99729-2	Q9BTT6 (deam*3, am*3)	Q9BYN8 (deam, am)	Q9H477 (deam*2)
Q99747	Q9BTT6:IRF..SV[67aa]:4	Q9BZE1	Q9H479 (deam*2, am)
Q99757 (deam)	58:MPICGG	Q9BZE1:C:366:S (deam)	Q9H488 (deam*2, am)
Q99805 (deam, am)	Q9BTY2	Q9BZE9-	Q9H488:GKV..EF[64aa]:
Q99816 (deam)	Q9BTY2:K:456:I	2:KRS..KR[18aa]:630:RE	325:EGR
Q99828	Q9BTZ2	E..QE[18aa]	Q9H4A5
Q99829 (deam*2, am)	Q9BU02 (deam, am)	Q9BZK7 (deam*2, am)	Q9H4A6 (deam, am)
Q99832:DKV..PH[397aa]:	Q9BU61 (deam*3, am)	Q9BZL4:EDL..SK[467aa]	Q9H4B7:NNV..GH[105aa
147:E	Q9BU89	:316:EGP..AL[21aa]	]:347:NPTSRWLSATSRP
Q99832:KVE..PH[396aa]:	Q9BU89 (ac)	Q9BZX2	GG

Q9H4M9 (deam, am)	Q9NP79 (ac)	Q9NVP1 (deam)	Q9P0V9:S:115:I
Q9H583:N:1694:S	Q9NP92 (deam*2)	Q9NVQ4	Q9P0Z9
Q9H5N1 (deam*4, am*3)	Q9NP97 (deam*3)	Q9NVT9 (ac)	Q9P0Z9 (deam*3)
Q9H5X1	Q9NPA0	Q9NW13-	Q9P1F3 (ac)
Q9H6E4 (deam*4, am)	Q9NPA0 (deam)	2:TSG..DS[249aa]:370:H	Q9P1U1
Q9H6R7	Q9NPA0 (deam*2)	Q9NW13:EHS..DS[387aa	Q9P1U1-3
Q9H6S3 (deam, am)	Q9NPD3 (am)	] :373:ECILKVVHLPSS	Q9P258
Q9H6V9	Q9NPD3 (deam*3, am)	Q9NWM8	Q9P287 (deam*2)
Q9H6Z4-	Q9NPD8 (deam. Am)	Q9NWS6 (deam)	Q9P287-4
3:LVM..ST[125aa]:375:H	Q9NPE3 (deam)	Q9NWU2 (deam*3, am)	Q9P289-2
SD..RA[84aa]	Q9NPF4 (deam, am)	Q9NWW4 (deam)	Q9P2E9-3:L:610:H
Q9H773	Q9NPH2	Q9NWW8 (deam, am)	(deam*5, am*2)
Q9H7C9	Q9NQ88	Q9NWX6	Q9P2E9-
Q9H7D7:NCV..DS[45aa]:	Q9NQE9	Q9NWWY4	3:RMQ..SV[650aa]:328:R
617:TLC..LR[18aa]	Q9NQG5 (deam*3, am*2)	Q9NWX8	ACRPATGST
Q9H7Z7 (deam*4)	Q9NQG6 (deam, am)	Q9NX01	Q9P2E9-3:S:766:L
Q9H845 (deam, am)	Q9NQP4 (deam*2, am*2)	Q9NX08 (deam*3, am)	(deam*8, am*2)
Q9H871	Q9NQR4 (deam*3, am)	Q9NX18 (deam, am)	Q9P2G3 (deam, am)
Q9H8H0	Q9NQS1	Q9NX24	Q9P2J5-
Q9H8H3	Q9NQT5 (ac)	Q9NX38	2:KLS..VH[899aa]:224:N
Q9H8W4	Q9NQX3-2 (deam)	Q9NX40	N
Q9H8Y8	Q9NQY0	Q9NX40 (deam, am)	Q9P2J5:KLS..VH[899aa]:
Q9H8Y8 (deam*3)	Q9NQZ5	Q9NX55	278:NN
Q9H910-3	Q9NR12-5	Q9NX62 (deam*3, am*2)	Q9P2J9 (deam, am)
Q9H936	Q9NR33:G:17:V	Q9NX63 (am)	Q9P2N5
Q9H944 (deam, am, meth)	Q9NR45 (deam*2, am)	Q9NY27 (deam)	Q9P2T1
Q9H974	Q9NR56	Q9NY33:LLA..QA[187aa	Q9P2T1-2
Q9H974-4	Q9NRF8	] :551:LSL..GL[23aa]	Q9P2W9 (deam, am)
Q9H993 (deam*2)	Q9NRF9 (deam, am)	Q9NYB0	Q9UBE0 (deam*4, am,
Q9H9A6 (deam*2, am)	Q9NRN7 (deam, am)	Q9NYG5	ac*4)
Q9H9B4 (deam*2, am)	Q9NRP2 (deam*2)	Q9NYL9	Q9UBF6 (deam, am)
Q9H9H4	Q9NRS6	Q9NYL9 (deam*3)	Q9UBK9 (deam, am)
Q9H9J2 (deam*2, am)	Q9NRV9 (deam*3, am)	Q9NYP9	Q9UBL3
Q9H9Q4	Q9NRW7 (deam)	Q9NYR9	Q9UBP6 (deam, am)
Q9HA64:K:182:R (deam,	Q9NRX2	Q9NYR9 (deam*3)	Q9UBQ0 (deam*2, am)
am)	Q9NRX4 (deam*2, am)	Q9NZ45	Q9UBQ0-2 (deam*3, am)
Q9HA77	Q9NS69 (deam*2, am)	Q9NZ45 (ac)	Q9UBQ5 (deam*2, am)
Q9HAN9 (ac)	Q9NSE4	Q9NZD8 (deam)	Q9UBQ7 (deam*3, am)
Q9HAS0	Q9NSE4:KLV..GK[555aa	Q9NZI9 (deam, am)	Q9UBQ7:VIL..KL[40aa]:
Q9HAV7	] :458:RNW..GL[26aa]	Q9NZL9-2:S:184:R	289:D
Q9HAV7 (deam*3, am)	Q9NSK0-5	Q9NZW5 (deam, am)	Q9UBT2-2
Q9HB40	Q9NSK0:R:72:H	Q9P000 (deam)	Q9UBT2-2:K:527:E
Q9HB40-2	Q9NTG7 (deam, am)	Q9P015	Q9UBT2-
Q9HBH1 (deam)	Q9NTM9	Q9P015 (deam*2)	2:RSR..LD[17aa]:528:RV
Q9HBL8	Q9NTM9 (deam*3)	Q9P016	TYRTEGRA
Q9HBM1	Q9NTZ6	Q9P032	Q9UBT2:FDI..LD[267aa]:
Q9HC35:SID..SS[298aa]:	Q9NTZ6 (deam)	Q9P035 (deam)	374:LYQ..YY[16aa]
684:SYRWYLPGRIS	Q9NUG6	Q9P0J0 (am)	Q9UBT2:K:623:E
Q9HC36:G:45:E (deam,	Q9NUP1	Q9P0J1	Q9UBU9 (deam, am)
am)	Q9NUP9 (deam*2, am)	Q9P0J7	Q9UBV8
Q9HC52 (deam, ac)	Q9NUQ6-	Q9P0J7:DPI..PL[185aa]:1	Q9UDR5 (deam)
Q9HCC0 (deam*4, am*2)	2:VDS..VA[313aa]:177:V	97:DSYS	Q9UDX5
Q9HCD5 (deam*3, am)	R	Q9P0L0	Q9UEU0
Q9HCE1	Q9NUW8 (deam)	Q9P0L0:ETR..IL[79aa]:1	Q9UEY8
Q9HCN8 (deam)	Q9NV31 (deam)	71:EHKETNGRV	Q9UEY8-2
Q9HCU5:DVS..FL[111aa	Q9NV56	Q9P0M6	Q9UFC0
] :307:DQSVNPAPS	Q9NVI7-	Q9P0P8	Q9UFG5
Q9HD15	2:EAR..PS[311aa]:276:EL	Q9P0R6 (deam, am)	Q9UFW8
Q9HDC9	GWGSRP	Q9P0S9	Q9UGK8:K:429:E
Q9HDC9:GLV..AV[35aa]	Q9NVI7-2:G:15:D	Q9P0U1	(deam*2, am)
:382:GAG..GC[35aa]	(deam*2)	Q9P0V9	Q9UGM6 (am)
Q9NP66	Q9NVM6	Q9P0V9-3	Q9UGV2

Q9UGV2 (am*2)	Q9ULR0-1	Q9Y2S7 (deam, am)	ERI..LY[23aa]
Q9UGV2-2	Q9ULR3 (deam, am)	Q9Y2T2	Q9Y4B6:GEY..NE[1170a
Q9UH62 (deam*2, am)	Q9ULV4	Q9Y2T3	a]:338:ERI..LY[23aa]
Q9UH65:KKQ..TE[258aa	Q9ULV4:ASV..AA[42aa]	Q9Y2V2 (deam, am, ph)	Q9Y4L1 (deam*3, am)
] :328:REAAAG	:433:AQCAK	Q9Y2W1:A:201:V (deam,	Q9Y4L1:DTV..EL[401aa]
Q9UH92-3	Q9ULX9	am)	:599:VYC..GS[28aa]
Q9UH99-	Q9ULZ3	Q9Y2W6 (deam, am)	Q9Y4W6 (deam)
2:GGG..AH[249aa]:490:E	Q9UMR2	Q9Y2Z4	Q9Y4Z0
RG..LQ[72aa]	Q9UMR2-4	Q9Y2Z9-3	Q9Y508
Q9UHB9-4	Q9UMS4	Q9Y2Z9-3 (am)	Q9Y512:I:345:V
Q9UHD1-2:A:310:D	Q9UMS4	Q9Y314 (deam, am)	Q9Y530 (deam*2, am)
Q9UHD9:MVQ..PS[62aa]	Q9UMS4:VVF..SL[235aa	Q9Y315	Q9Y547
:563:MGA..KP[35aa]	] :270:GGV..GP[36aa]	Q9Y333 (deam)	Q9Y584
Q9UHI6 (deam)	Q9UMX5	Q9Y365	Q9Y5J1
Q9UHQ9 (deam)	Q9UN36-6	Q9Y365 (deam)	Q9Y5J7 (deam*2, am)
Q9UHV9 (deam*4, am)	Q9UN37 (am)	Q9Y375 (deam)	Q9Y5K3
Q9UHX1-	Q9UNE7	Q9Y376	Q9Y5K3-2
2:QEH..SA[123aa]:420:R	Q9UNE7 (deam, am)	Q9Y399	Q9Y5K6 (deam)
GAHEHLGQ	Q9UNI6	Q9Y3A3	Q9Y5L4 (deam*3, am, ac)
Q9UHY1	Q9UNM6:N:13:S	Q9Y3A5:QQL..FE[47aa]:	Q9Y5M8
Q9UHY7 (deam*3, am)	Q9UNM6:N:13:S	204:HTVRNRMSD	Q9Y5P6
Q9UI09 (deam*2)	(deam*3, am)	Q9Y3A6	Q9Y5R8
Q9UIC8	Q9UNN5 (deam, am)	Q9Y3B4	Q9Y5S9 (deam, am)
Q9UII2	Q9UNQ2 (deam*2)	Q9Y3B8	Q9Y5U8
Q9UIJ7	Q9UQ80 (deam*3, am)	Q9Y3B9	Q9Y5X3
Q9UIJ7 (deam*3)	Q9Y221 (deam*2)	Q9Y3C0	Q9Y5Y2 (deam)
Q9UIV1	Q9Y244	Q9Y3C4	Q9Y5Z4 (deam, am)
Q9UJ41-2	Q9Y248 (ac*2)	Q9Y3C6 (deam, am)	Q9Y619
Q9UJ68-3	Q9Y263	Q9Y3C7 (deam)	Q9Y678
Q9UJ70	Q9Y263:HAI..LL[110aa]:	Q9Y3C8	Q9Y678 (deam, am)
Q9UJ70-	686:HMQ	Q9Y3D0	Q9Y678:CSA..VG[595aa]
2:FRK..FS[110aa]:281:FE	Q9Y265 (deam*4, am)	Q9Y3D0 (deam*2, am)	:280:CKC..PQ[33aa]
G..RD[18aa]	Q9Y266	Q9Y3D2	Q9Y696 (deam)
Q9UJ70:FRK..FS[110aa]:	Q9Y266 (deam*3, am)	Q9Y3D5	Q9Y697-3
235:FEG..RD[18aa]	Q9Y276	Q9Y3D6	Q9Y6A4 (deam)
Q9UJA5-4	Q9Y277	Q9Y3D7 (deam)	Q9Y6B6 (deam, am)
Q9UJC5 (deam, am)	Q9Y277-2	Q9Y3D9	Q9Y6E0-
Q9UJU6-4	Q9Y277-	Q9Y3D9 (deam*2)	2:RGE..SH[218aa]:214:E
Q9UK22	2:KVGLGFELEA:275:K	Q9Y3E5	RG..LQ[31aa]
Q9UK41	WLA WDLNWK	Q9Y3F4-2	Q9Y6G3
Q9UK59	Q9Y277:KVGLGFELEA:	Q9Y3I0 (deam*2, am)	Q9Y6G5
Q9UKB3 (deam, am)	274:KWLA WDLNWK	Q9Y3L3:K:86:*	Q9Y6G9 (deam*2, am)
Q9UKK6 (deam)	Q9Y281	Q9Y3X0 (deam)	Q9Y6H1
Q9UKK9 (deam*4, am*2,	Q9Y281-3	Q9Y3X0 (ph)	Q9Y6I3-1
ac*3)	Q9Y294	Q9Y478	Q9Y6J9
Q9UKL6	Q9Y2B0 (deam, am)	Q9Y490:ISA..EH[148aa]:	Q9Y6K9
Q9UKU7	Q9Y2G5	2394:ILL..SR[37aa]	Q9Y6M1
Q9UKV3-	Q9Y2P8	Q9Y490:SAA..EH[147aa]	Q9Y6M1-1
2:ERE..RR[52aa]:563:RE	Q9Y2P8-2	:2395:LCCPDGGCGHQQ	Q9Y6M1-
R..SG[16aa]	Q9Y2Q9 (deam, am)	SV	3:KIA..SK[138aa]:399:R
Q9UKV3-3	Q9Y2R0 (deam)	Q9Y490:VDN..EH[1333a	DC..LL[41aa]
Q9UKX7	Q9Y2S0 (deam*3, am,	a]:1209:VIMP	Q9Y6M1:KIA..SK[138aa]
Q9UKY7 (deam, am)	ac*2)	Q9Y4B6-	:462:RDC..LL[41aa]
Q9UL25 (deam, am)	Q9Y2S6	2:GEY..NE[1170aa]:337:	Q9Y6N1 (deam*2, am)