РАЗМЕР ПРОТЕОМА ЧЕЛОВЕКА КАК ФУНКЦИЯ РАЗВИТИЯ ЭКСПЕРИМЕНТАЛЬНЫХ ТЕХНОЛОГИЙ И МЕТОДОВ БИОИНФОРМАТИКИ

Е.В. Сарыгина*, А.С. Козлова, Е.А. Пономаренко, Е.В. Ильгисонис

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

Представлен ретроспективный анализ изменений сведений о количестве протеоформ человека, событий посттрансляционных модификаций (ПТМ), альтернативного сплайсинга (АС), одноаминокислотных полиморфизмов (ОАП), ассоциированных с белок-кодирующими генами в базе данных neXtProt. В 2016 году нашей группой были предложены три математические модели для предсказания количества различных белков (протеоформ) в протеоме человека. Спустя восемь лет мы сравнили исходные данные информационных ресурсов и их вклад в результаты предсказаний, сопоставив различия с новыми подходами экспериментального и биоинформатического анализа модификаций белков. Цель данной работы — актуализировать информацию о статусах записей в базах данных о выявленных протеоформах с 2016 года, а также выявить тренды изменений количеств этих записей. Согласно различным информационным моделям, современные экспериментальные методы позволяют выявить от 5 до 125 млн различных протеоформ — белков, образованных в результате альтернативного сплайсинга, реализации на протеомном уровне однонуклеотидных полиморфизмов и посттрансляционных модификаций в различных комбинациях. Данный результат отражает увеличение размера человеческого протеома на 20 и более раз за последние 8 лет.

Ключевые слова: протеомика; протеоформы; посттрансляционные модификации; одноаминокислотные замены; альтернативный сплайсинг; neXtProt

DOI: 10.18097/PBMC20247005364

ВВЕДЕНИЕ

Реальное число различных видов белков значительно превышает количество кодирующих их генов [1]. Многообразие белков возникает вследствие различных причин: альтернативного сплайсинга (АС), единичных нуклеотидных замен в геноме, реализующихся на протеомном уровне в виде одноаминокислотных полиморфизмов (ОАП), а также посттрансляционных модификаций (ПТМ) белков [1].

Многообразие белков, кодируемых одним геном, авторы называют “формы белка”, “изоформы белка” [2], “модифицированные варианты” [3], “виды белка” [4]. В 2013 году группа под руководством проф. Kelleher ввела для группы белков, кодируемых одним геном, обозначение “протеоформы” [5] — совокупность форм белков, возникающих вследствие модификаций на геномном, транскриптомном и протеомном уровнях, а также в результате эндогенного протеолиза [5]. Протеом человека является совокупностью всех белков тканей и органов, поэтому многообразие протеоформ может быть обозначено как “ширина” протеома и рассчитано с использованием моделей, учитывающих частоты возникновения модификаций на геномном, транскриптомном и протеомном уровнях [6].

Анализу протеома биологических объектов препятствует широкий динамический диапазон концентраций аналитов [7], из-за которого масс-спектрометрический сигнал от конкретного белка может теряться среди “шумов” от других молекул, а также ограниченная чувствительность аналитического метода [8]: чем выше чувствительность метода, тем большее количество белков (как канонических, так и протеоформ) может быть определено экспериментально. Существующие методические решения не позволяют проводить экспериментальный анализ полного спектра протеоформ в высокопроизводительном режиме [9, 10]. Экспериментально показано, что ограничение аналитической чувствительности современной масс-спектрометрии — лимитирующий фактор применения протеомных технологий в медицине.

В 2016 году нашей группой были предложены три математические модели для предсказания возможного количества протеоформ человека [11]. Для использования моделей необходима информация о количестве белок-кодирующих генов, для которых экспериментально показана кодируемая протеоформа, и общем количестве сплайс-вариантов, ОАП и ПТМ. В статье 2016 года предсказания базировались на данных версии neXtProt (ver. 2015_06) и предсказано существование 0,62, или 0,88, или 6,13 миллиона видов белков человека (без учёта комбинаторных вариантов, [11]. В настоящем исследовании мы актуализируем эти оценки с учётом обновления сведений в базе данных neXtProt и развития экспериментальных и биоинформатических подходов для детекции белков.

В работе Aebersold и соавт. [1] в 2018 году была произведена аналогичная биоинформатическая оценка. Исследователи оценивали теоретически возможное количество протеоформ как произведение вероятностей возникновения всех существующих (есть данные с экспериментальным подтверждением в информационных ресурсах) ПТМ в каждом аминокислотном остатке белок кодирующего гена.

Учитывая только бинарные модификации, число теоретических протеоформ оценили как астрономически большое (1×10^{27}). В то же время, разнообразие протеоформ, которое мы реально наблюдаем в биологических системах, оказывается значительно ниже теоретически предсказанного по причине ограниченных возможностей существующих технологий детекции. В связи с этим для оценки многообразия протеоформ целесообразно использовать информационные ресурсы, в которых консолидированы результаты экспериментального изучения протеома различными научными группами во всем мире. Наиболее полным ресурсом по протеому человека является система neXtProt [12].

neXtProt — информационная платформа, созданная в 2011 году для объединения данных, полученных при исследовании протеома человека [12]. Основу для neXtProt обеспечивают данные в записях UniProtKB/Swiss-Prot (Reviewed) для *Homo sapiens* (TaxID: 9606) с ключевым словом “Complete proteome” (KW-0181). Ресурс обогащён также и внешними источниками биологических данных, количество которых с 2016 года интегрируется всё больше. Помимо Uniprot, предоставляющего основной объём знаний об изоформах транскриптов, а также локализации генов на хромосоме и белковых продуктов в клетке, на 2023 год в neXtProt включены также и базы данных однонуклеотидных замен (UniProtKB, COSMIC, dbSNP, neXtProt, gnomAD), и ПТМ (UniProtKB, neXtProt, PeptideAtlas, GlyConnect), а также другие (Ensembl, PeptideAtlas, PDB via UniProtKB).

После завершения проекта “Геном человека” [13] были созданы технологические предпосылки для высокопроизводительного исследования продуктов генов на транскриптомном (секвенирование нового поколения) и протеомном (масс-спектрометрия, например Orbitrap) уровнях, тем самым способствуя существенному увеличению количества данных экспериментов.

Данная работа представляет собой ретроспективный анализ изменений сведений о количестве протеоформ в базе данных neXtProt спустя 8 лет после работы нашей группы в 2016 году [11]. Проведение ретроспективной оценки количества протеоформ помогает выявить тенденции в изучении протеомов, а также идентифицировать области, требующие дополнительных исследований или обновлений. Проведение подобного анализа позволяет оценить, какие изменения произошли в базе данных в связи с развитием методов анализа протеомов и развитием технологий, а также оценить качество данных, содержащихся в базе neXtProt. Ранее подобный подход позволил выявить наиболее часто используемые экспериментальные и биоинформатические методы анализа функций белков [14].

МЕТОДИКА

Использовали раздел аннотации neXtProt, описывающий кодируемые геном аминокислотные последовательности для здорового человека.

Для каждого года из FTP-сервера neXtProt [15] загружали xml-файл, содержащий сведения обо всех белках. NeXtProt предоставляет обновленные версии несколько раз в год, поэтому выбрали по одной версии из каждого года в интервале с 2016 по 2023 год, а именно: [nextprot_release_]2016-12-02, 2017-08-01, 2018-09-03, 2019-08-22, 2020-11-26, 2021-11-19, 2022-10-31, 2023_09.

Из каждого xml-файла с использованием собственных скриптов, написанных на языке python, выгружали сведения из следующих полей: (i) <entry accession>, (ii) <annotation-category> с атрибутом category равными “modified-residue”, “cross-link”, “disulfide-bond” и “glycosylation-site”, “variant”, (iii) <property-list>, (iv) <chromosomal-location-list>, (v) <transcript-mapping>.

Каждая аннотация записи белка имеет степень достоверности данных [16] “Gold” (экспериментально доказанные данные высочайшего качества, соответствующие уровню ошибок <1% методами иммуногистохимии, тандемной масс-спектрометрии и др.), “Silver” (данные предсказанные инструментами биоинформатики, но не получившие подтверждения экспериментальными методами, соответствующие уровню ошибок <5%) или “Bronze” (данные низкого качества). В данной работе рассматривали только данные с Gold и Silver уровнями, полученными из тега <annotation quality>. Таким образом устанавливали какое количество ПТМ, ОАП и АС соответствует каждому белок-кодирующему гену. После этого полученные сведения суммировали и человеческий геном каждой из отобранных для анализа версий (с 2016 по 2023 год) характеризовали с использованием набора следующих дескрипторов:

[N] — количество белок-кодирующих генов;

[AS], ([SAP], [PTM]) — количество аминокислотных последовательностей, предполагаемых исходя из данных об альтернативном сплайсинге мРНК (содержащих ОАП или ПТМ);

[ASav], ([SAPav], [PTMav]) — количество кодируемых одним геном вариантов аминокислотных последовательностей, образованных в результате альтернативного сплайсинга (наличия ОАП или ПТМ); дескриптор рассчитывается как отношение [AS] ([SAP], [PTM]) к [N].

Для оценки потенциального числа протеоформ рассматривались три различных случая сочетания событий ПТМ, ОАП и АС согласно следующим уравнениям информационных моделей [11] (1)–(3):

$$Nps = N\times(ASav+SAPav+PTMav) \quad (1);$$

$$Nps = (N+AS)\times(SAPav+PTMav) \quad (2);$$

$$Nps = N\times Asav\times SAPav\times PTMav \quad (3).$$

Уравнение (1) предполагает, что ПТМ появляются исключительно в канонических последовательностях белков, но не в сплайс-вариантах. Уравнение (2) предполагает, что ПТМ и ОАП могут встречаться как в белках, кодируемых каноническими последовательностями, так и в сплайс-вариантах.

Уравнение (3) предполагает, что все типы модификаций (ПТМ, ОАП и АС) происходят независимо друг от друга.

Визуализацию полученных результатов проводили с использованием собственных скриптов на языке python (v3.8), включающих библиотеки pandas (v2.2.2), matplotlib (v3.8.4) и seaborn (v 0.13.2).

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

Мы проанализировали динамику изменения в период 2016–2023 г.г. ключевых параметров моделей, предложенных ранее для оценки количества протеоформ — количество белок-кодирующих генов (БКГ), ПТМ, АС, ОАП.

Основной параметр — количество БКГ — начиная с 2018 года остаётся неизменным и составляет 20,3 тыс. (рис. 1). Это свидетельствует о том, что биоинформатические алгоритмы аннотации генома, которые помогают идентифицировать новые гены и их продукты, а также уточнять сведения об уже известных генах, существенным образом не менялись. Основные источники информации о БКГ, которые используют кураторы neXtProt, — это базы данных последовательностей геномов и их аннотаций, такие как EMBL-EBI (EMBL's European Bioinformatics Institute), в частности Ensembl, KeGG и другие [12].

Анализ динамики изменения количества ПТМ (рис. 2) показывает, что фактически количество экспериментально подтверждённых ПТМ вышло

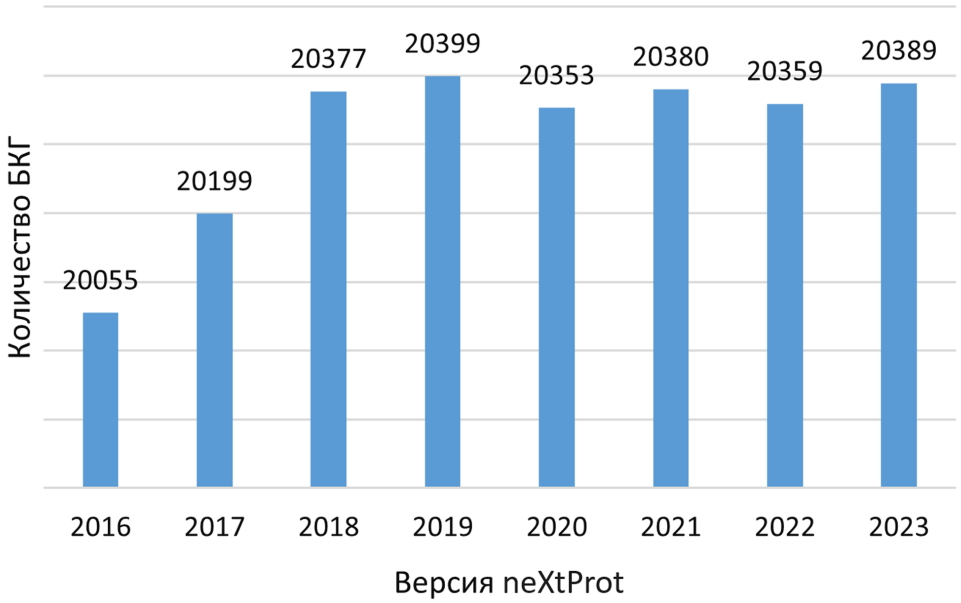

Рисунок 1. Количество БКГ генома человека, согласно данным neXtProt.

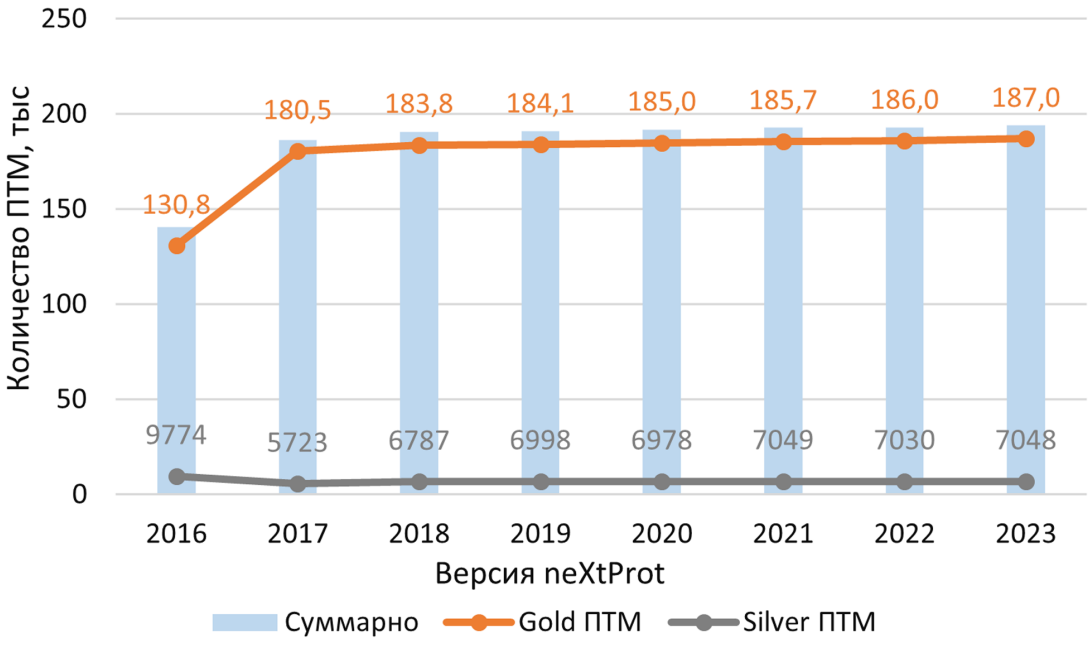

Рисунок 2. Количество ПТМ, ассоциированных с БКГ человека, согласно данным neXtProt.

на плато в 2017 году, тогда как подходы *in silico* совершенствуются и позволяют предсказывать всё больше (свыше 1000 ПТМ) возможных вариантов [17]. С одной стороны, выход на плато может говорить о достижении пределов существующих на сегодняшний день технологий детекции ПТМ (в основном, для этого используют масс-спектрометрические методы) [18]. С другой стороны, многие ПТМ зависят от регуляторных сигналов клетки, влияющих на активность ферментов, которые осуществляют ПТМ аминокислотных остатков (киназы, фосфатазы, ацетилтрансферазы и др.), а также от внешних или внутренних факторов, например, окисления ароматических аминокислот в результате оксидативного стресса [19]. Наличие или отсутствие этих факторов в конкретном моменте времени определяет возможность возникновения модификации и должно учитываться при проведении профилирования ПТМ [20].

Не все БКГ человека содержат аннотации с информацией о наличии ПТМ, АС или ОАП — для части генов такие события либо не характерны, либо пока не обнаружены. Доля БКГ, для которых есть информация о кодируемых белках со сплайс-вариантами или ОАП, не изменилась с 2016 года и составляет 48,3% и 96% соответственно. Это означает, что чуть более половины генов генома человека не охарактеризованы по наличию вариантов АС. В то же время, почти все БКГ снабжены записями neXtProt с описанием вариантов (как минимум, одного) ОАП. Интересно, что постепенно увеличивается доля БКГ, для которых показано наличие ПТМ в кодируемых белках: в 2016 году таких генов было 74,2%, а в 2023 — 76,4%. Вероятно, это связано с развитием методов масс-спектрометрического анализа, позволяющих характеризовать ПТМ белков в высокопроизводительном режиме [18]. Скорее всего, со временем доля таких БКГ будет расти, в отличие от доли БКГ с аннотированными АС и ОАП — изменений в этом случае следует ожидать в случае прорывного изменения технологий секвенирования.

Для количества вариантов АС (рис. 3) также наблюдается стабилизация количества записей. Этот параметр может меняться сравнительно медленно по нескольким причинам. Во-первых, часть событий АС происходит в регуляторных регионах транскриптов (5′/3′-UTR) и не приводит к разнообразию белковых структур [21]. Во-вторых, профили альтернативного сплайсинга могут меняться в зависимости от внешних и внутренних стимулов, а также при онкотрансформации [22, 23]. В настоящее время совершенствуются методы и алгоритмы поиска новых сплайс-форм в экспериментах РНК-секвенирования [24, 25]. С 2018 года стартовал проект MANE, направленный на интеграцию двух основных аннотаций генома человека (RefSeq и Ensembl/GENCODE) и новых сплайс-форм, а также их валидацию; однако проект всё ещё не завершён [26].

События АС широко подтверждены на уровне мРНК (транскриптов). Подтверждений существования сплайс-вариантов белков существенно меньше из-за низкой степени покрытия пептидов в большинстве панорамных протеомных экспериментов [27]. Это связано с общепризнанными ограничениями протеомики для обнаружения таких событий: общепринятой практикой является разделение изоформ белков на группы, содержащие один и тот же пептид, поскольку в результате АС полипептидная последовательность меняется не полностью, а лишь в отдельных участках [28].

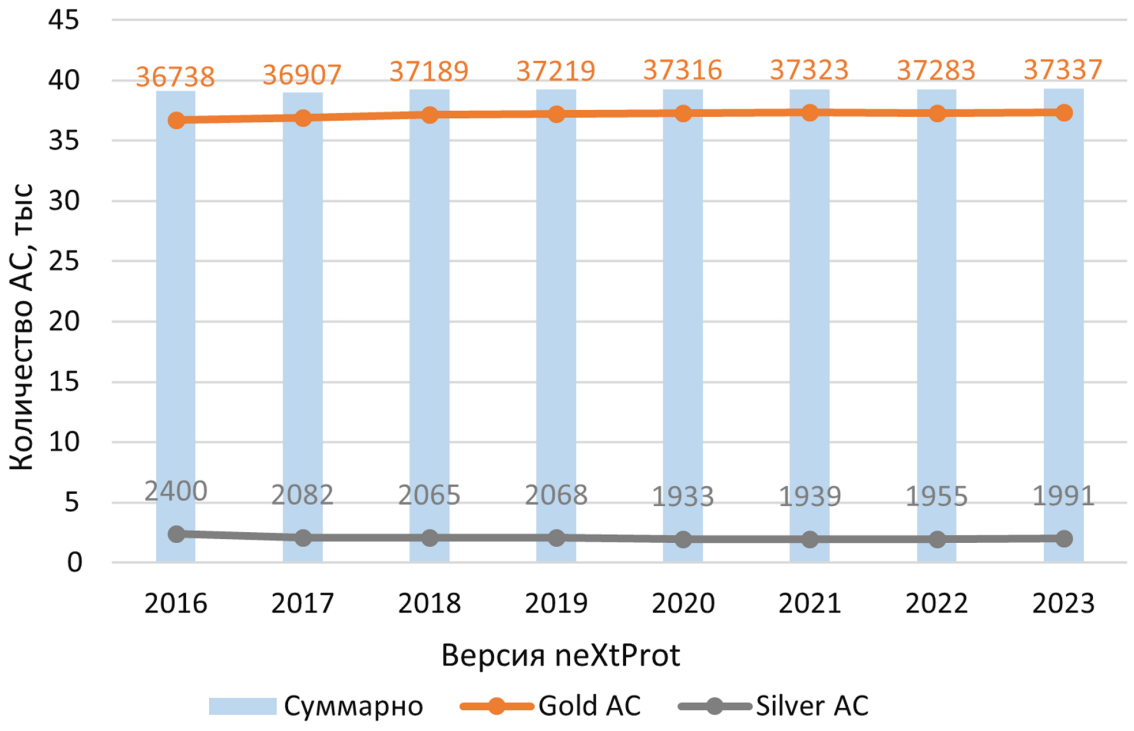

Рисунок 3. Количество АС, ассоциированных с БКГ человека, согласно данным neXtProt.

В 2023 году Sinitcyn и соавт. показали, что более половины (около 64%) событий сплайсинга высокоэкспрессируемых генов, обнаруженных с помощью транскриптомики, действительно транслируются и присутствуют на уровне белка [29]. На протеомном уровне удалось обнаружить только 22% сплайс-форм со средним уровнем экспрессии [29]. По всей видимости, эта оценка занижена, учитывая высокодинамичный характер экспрессии белка и проблемы обнаружения дифференциально экспрессируемых сплайс-форм.

Наибольший интерес представляет анализ изменений количества известных ОАП по годам (рис. 4). В 2019 году отмечен взрывной рост ОАП со статусом “Gold”, поскольку в neXtProt были интегрированы данные о частоте вариантов из Базы данных агрегации генома (gnomAD) [30], расширяющие информацию о вариациях последовательностей на уровне белка.

Аналогичный рост количества экспериментально детектированных ОАП в 2023 году, по всей видимости, связан с исследованием Sinitcyn и соавт. [29]. Это самое глубокое на сегодняшний день протеогеномное исследование, свидетельствующее о том, что примерно 73% несинонимичных SNP (то есть ОАП) транслируются и присутствуют в протеоме.

Хромосомоцентричный анализ (из аннотаций отбирали сведения о номере хромосомы, на которой расположен БКГ) ОАП БКГ на основе версии neXtProt 2023 года показал медианное значение присутствия 81 ОАП для вариантов, ассоциированных с заболеваниями, и 300 для остальных ОАП (рис. 5).

Рисунок 5 подтверждает одинаковую плотность распределения ОАП для всех хромосом. Выбросы достигают более 2000 ОАП-вариантов на БКГ (например ген TTN, SYNE1 и PCLO) для ОАП, ассоциированных с развитием патологических состояний, и более восьми тысяч (например ген MUC4, OBSCN и AHNAK2) среди остальных вариантов.

В 2016 году были получены оценки предположительного количества видов белков человека равные 0,62, или 0,88, или 6,13 миллиона (при наличии 20043 белок-кодирующих генов), в зависимости от используемой модели [11].

На 2023 год расчётное количество белков, оцениваемое по тем же уравнениям, достигало 5,8, 16,3 и 124,3 миллиона белков соответственно (рис. 6).

За восемь лет предполагаемые оценки размера ширины человеческого протеома выросли более, чем в 20 раз. Данные показатели рассчитаны только для экспериментально подтверждённых, “Gold” аннотаций ПТМ, ОАП (не включая варианты замен, ассоциированных с заболеваниями) и АС и для генов, их кодирующих, а также не включая варианты замен. Учитывая биоинформатически предсказанные данные (имеющие статус “Silver”), количество потенциальных видов белков в организме человека превышает более, чем в 40 раз значения, полученные в 2016 году.

Наибольший вклад в разнообразие белков вносит наличие ОАП в геноме, которые затем реализуются на протеомном уровне в виде замен аминокислотных остатков в белке. Вклад в многообразие белков альтернативного сплайсинга и пост-трансляционных модификаций существенно более скромный (рис. 7).

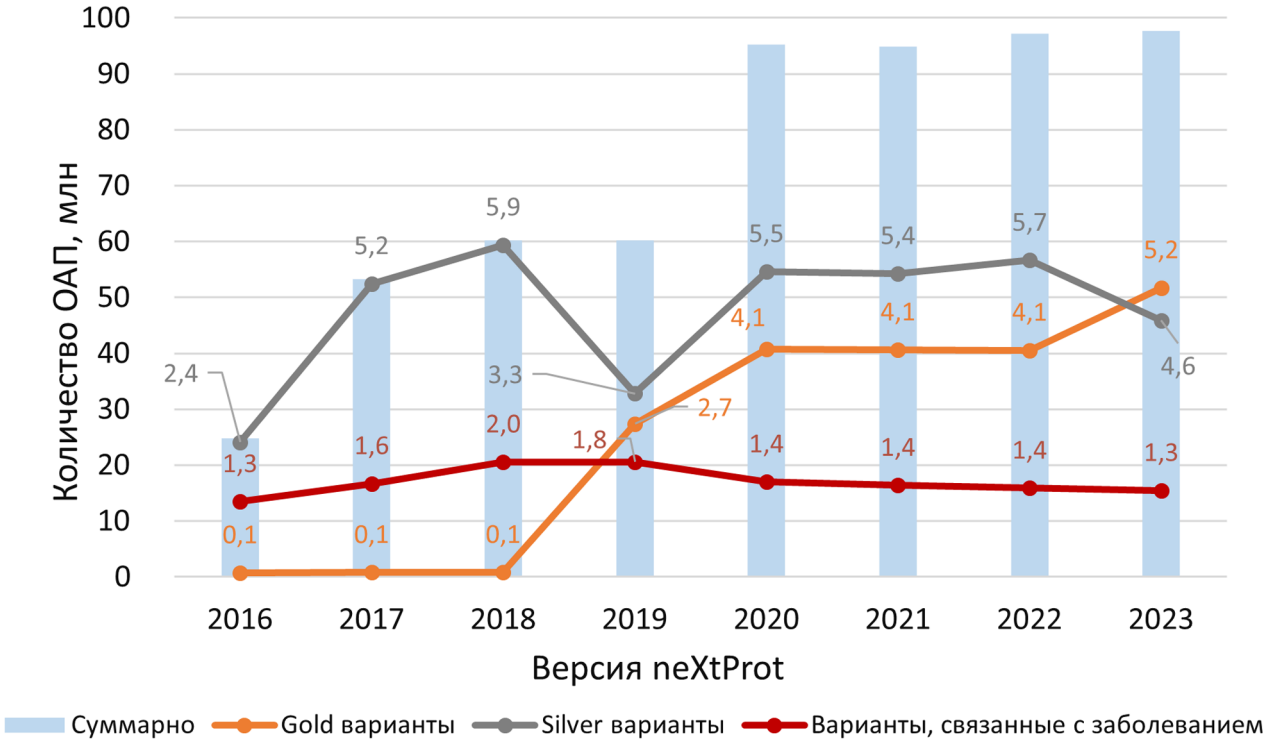

Рисунок 4. Количество одноаминокислотных полиморфизмов (ОАП), ассоциированных с БКГ человека, согласно данным neXtProt.

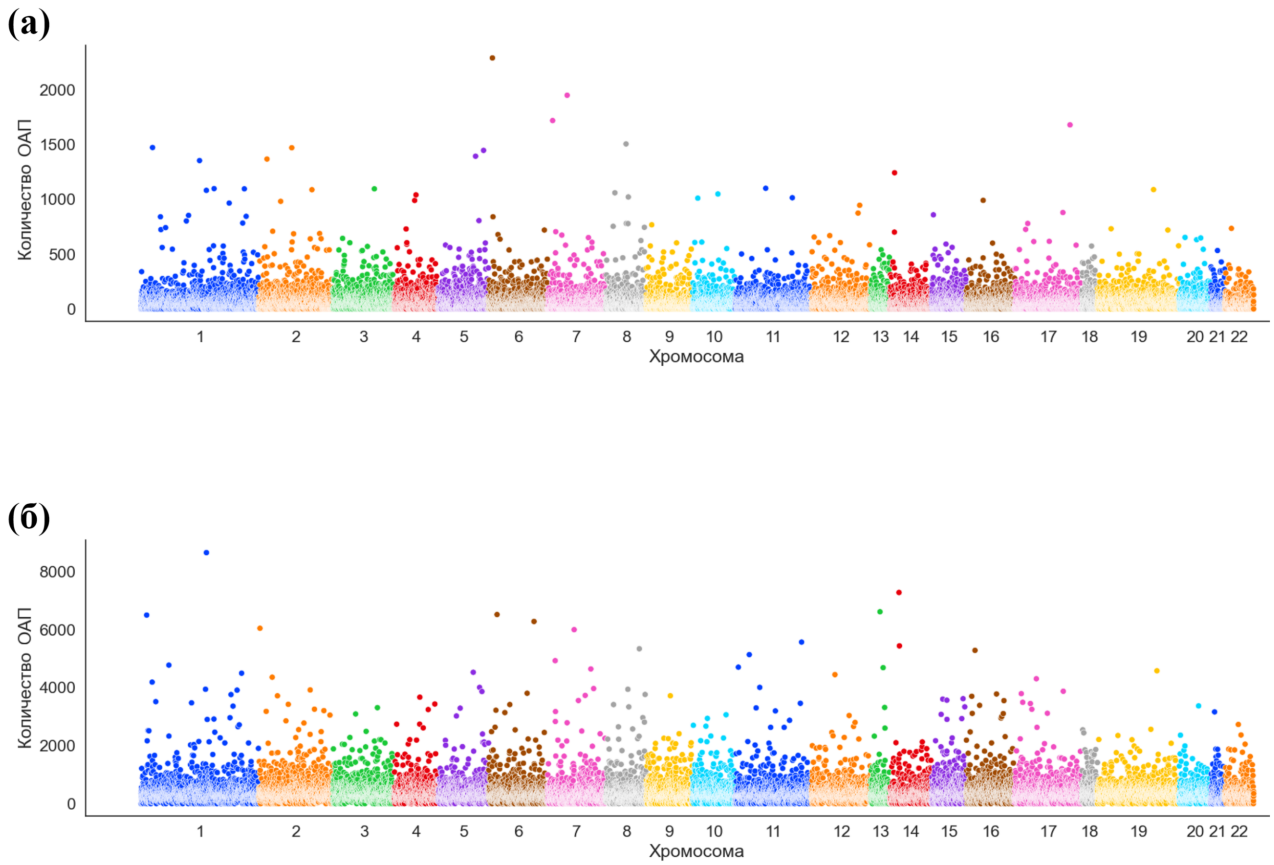

Рисунок 5. Хромосомоцентричное распределение одноаминокислотных полиморфизмов (ОАП), ассоциированных с БКГ человека, согласно последним данным версии neXtProt 2023 года для вариантов ОАП (**а**) связанных с заболеванием; (**б**) не связанных с патологическим процессом. Точкой показаны БКГ, цвет точки кодирует номер хромосомы.

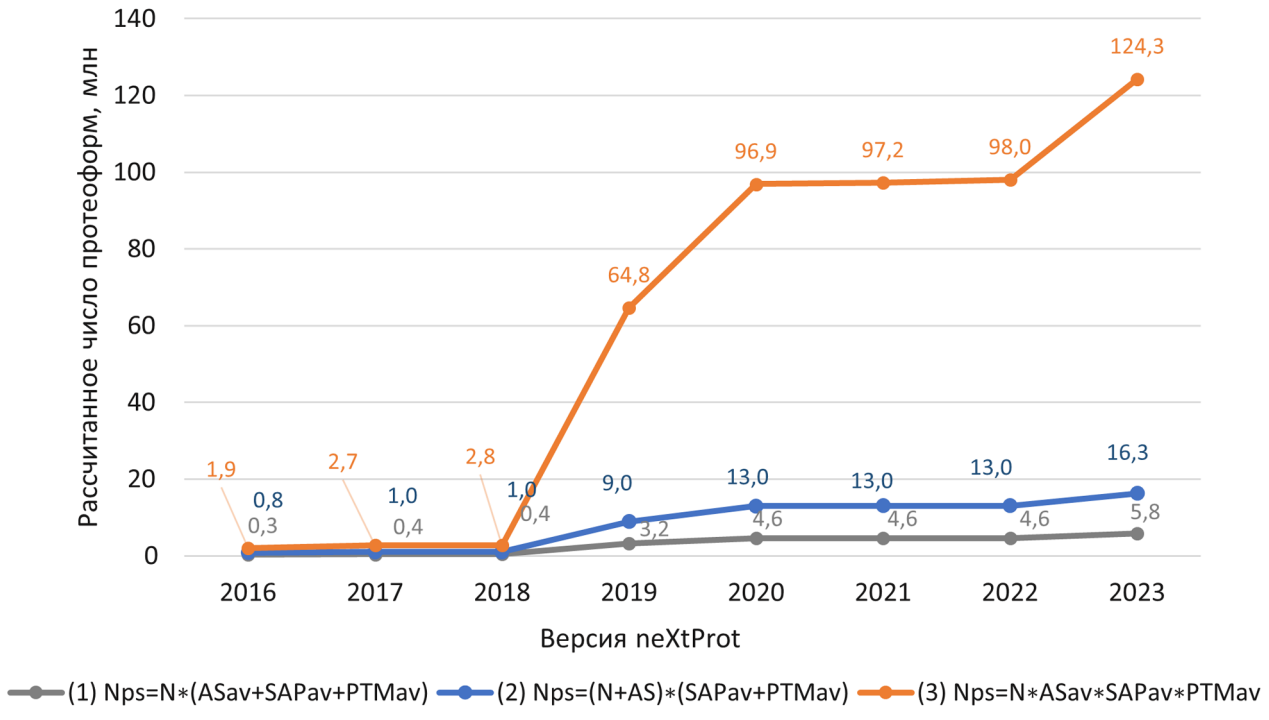

Рисунок 6. Расчётное количество протеоформ человека (согласно данным ресурса neXtProt с 2016 по 2023 год).

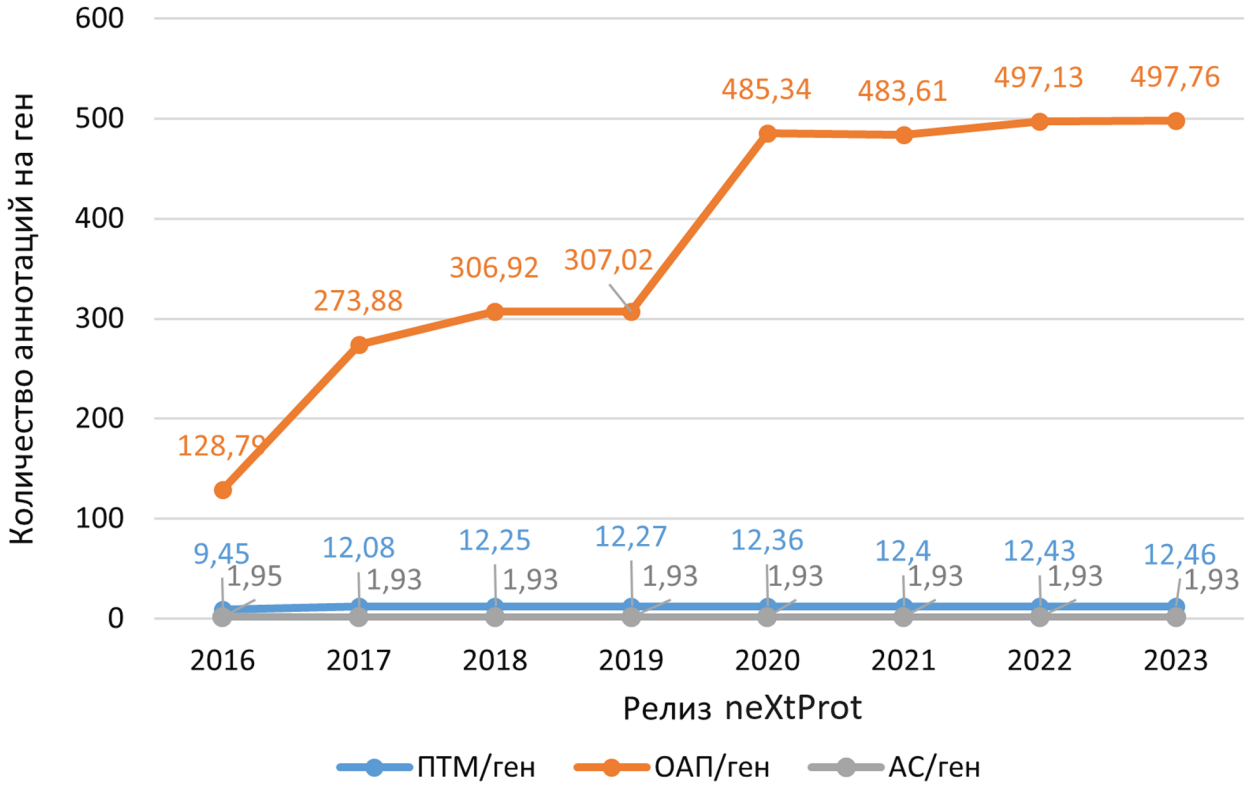

Рисунок 7. Изменение количества аннотаций на БКГ человека в ресурсе neXtProt в период 2016-2023 гг.

Это особенно видно для модели, рассчитанной по уравнению (3) (рис. 6), которая основана на предположении о том, что ПТМ возникают в любых аминокислотных последовательностях, и совместное возникновение ПТМ и ОАП происходит во всех сплайс-вариантах и канонических последовательностях.

За анализируемый период произошли значительные усовершенствования в методах исследования протеома, таких как масс-спектрометрия и биоинформатика, что позволило обнаруживать и идентифицировать белки, ранее недоступные для обнаружения. Во многом, возможно, это результат выполнения проекта “Протеом человека” в части поиска missing белков [31]. Такие масс-спектрометрические подходы, как тандемная масс-спектрометрия высокого разрешения получили широкое распространение, что отразилось на количестве детектируемых белков. С точки зрения биоинформатики, стали развиваться алгоритмы, основанные на машинном обучении для решения различных задач, таких как предсказание структуры белков [32], классификация генов [33], анализ последовательностей ДНК и многое другое.

Новые методы секвенирования генома и анализа генетических вариаций позволили более точно идентифицировать различные варианты белков, которые могут быть обусловлены генетическими различиями между индивидами и альтернативным сплайсингом. Появление метода секвенирования РНК с длинными чтениями, разработанного Oxford Nanopore Technologies (ONT), охватывающими несколько экзонов, открыло новые возможности для изучения АС путём прямой идентификации и количественной оценки изоформ транскриптов [34].

Интеграция данных из большего количества различных источников и применение междисциплинарных подходов, таких как системная биология и сетевой анализ, также помогли расширить предполагаемый размер протеома.

Увеличение предполагаемых оценок размера ширины человеческого протеома в 20 и более раз за последние 8 лет является результатом совокупности технологических прорывов, улучшения методов анализа данных и более глубокого понимания биологической сложности протеома. При этом тот факт, что количество детектируемых ПТМ и вариантов альтернативного сплайсинга фактически вышли на плато свидетельствует о том, что геномные и транскриптомные технологии развиваются существенно быстрее протеомных. Подобное различие вызвано масштабами и сложностью самих объектов исследования. Геномы и транскриптомы являются более статическими структурами по сравнению с белками. Кроме того, стоимость оборудования для детекции белков всё ещё существенно превышает стоимость приборов для секвенирования. Исследование протеома требует преодоления многих технических и методологических препятствий, включая сложность анализа белковой структуры, идентификации и количественной оценки белковых компонентов, а также анализа их функций и взаимодействий внутри клетки [35].

ЗАКЛЮЧЕНИЕ

Спустя 8 лет после расчётной оценки размера протеома [11] мы провели ретроспективный анализ изменения тренда количества протеоформ человека. Мы использовали экспериментальные результаты различных научных групп, агрегированные в ресурсе neXtProt — самом полном хранилище данных о протеоме человека. По различным информационным моделям, современные экспериментальные методы позволяют выявить от 5 до 125 млн различных протеоформ — белков, образованных в результате альтернативного сплайсинга, реализации на протеомном уровне однонуклеотидных полиморфизмов и посттрансляционных модификаций в различных комбинациях. Данный результат отражает увеличение размера человеческого протеома на 20 и более раз за последние 8 лет. Динамика накопления данных свидетельствует о развитии методических подходов: мы не наблюдаем за период с 2016 по 2023 год увеличение количества белок-кодирующих генов или генов, для которых впервые показано наличие альтернативного сплайсинга или одноаминокислотных замен. Это свидетельствует о насыщении в части такого рода геномных характеристик; в отличие от генов, кодирующих белки с вариантами ОАП, их количество возрастает, но они исследуются не методами секвенирования, а протеомными методами на основе масс-спектрометрии. Важным достижением последних лет является работа [29], показавшая, что теоретически предсказанное многообразие протеоформ подтверждается экспериментально на уровне протеома. Вследствие ограничений методов чувствительности, присутствующие в сравнительно (с каноническими, мастерными формами) небольших концентрациях протеоформы достаточно сложно зарегистрировать экспериментально. Один из возможных вариантов преодоления этой проблемы — использование различных типов биоматериала, различных условий в надежде, что где-то концентрации вида конкретной протеоформы будет достаточно для срабатывания детектора, ведь протеом динамичен.

Наша работа открывает перспективы исследования протеома с учётом многообразия видов белков — протеоформ. Впервые на основе консолидированного массива экспериментальных данных получен протеом протеоформ, достигающий до 125 млн на 2023 год. В дальнейшем можно ожидать роста найденных вариантов — за счёт идентификации на протеомном уровне однонуклеотидных замен в различных типах биоматериала (в норме и при патологиях) и пост-трансляционных модификаций. Вероятно, именно совокупность таких незначительных структурных изменений в целом является биомаркером ответа организма на патологические процессы.

ФИНАНСИРОВАНИЕ

Работа выполнена в рамках Программы фундаментальных научных исследований в Российской Федерации на долгосрочный период 2021–2030 годы (№ 122030100168-2).

СОБЛЮДЕНИЕ ЭТИЧЕСКИХ СТАНДАРТОВ

Настоящая статья не содержит каких-либо исследований с участием людей или с использованием животных в качестве объектов.

КОНФЛИКТ ИНТЕРЕСОВ

Авторы заявляют об отсутствии конфликта интересов.

ЛИТЕРАТУРА

1. *Aebersold R., Agar J.N., Amster I.J., Baker M.S., Bertozzi C.R., Boja E.S., Costello C.E., Cravatt B.F., Fenselau C., Garcia B.A., Ge Y., Gunawardena J., Hendrickson R.C., Hergenrother P.J., Huber C.G., Ivanov A.R., Jensen O.N., Jewett M.C., Kelleher N.L., Kiessling L.L., Krogan N.J., Larsen M.R., Loo J.A., Ogorzalek Loo R.R., Lundberg E., MacCoss M.J., Mallick P., Mootha V.K., Mrksich M., Muir T.W., Patrie S.M., Pesavento J.J., Pitteri S.J., Rodriguez H., Saghatelian A., Sandoval W., Schlüter H., Sechi S., Slavoff S.A., Smith L.M., Snyder M.P., Thomas P.M., Uhlén M., van Eyk J.E., Vidal M., Walt D.R., White F.M., Williams E.R., Wohlschlager T., Wysocki V.H., Yates N.A., Young N.L., Zhang B.* (2018) How many human proteoforms are there? Nat. Chem. Biol., **14**(3), 206–214. DOI: 10.1038/nchembio.2576
2. *Zhang F., Chen J.Y.* (2016) A method for identifying discriminative isoform-specific peptides for clinical proteomics application. BMC Genomics, **17**(Suppl 7), 522. DOI: 10.1186/s12864-016-2907-8
3. *Prabakaran S., Lippens G., Steen H., Gunawardena J.* (2012) Post-translational modification: Nature's escape from genetic imprisonment and the basis for dynamic information encoding. Wiley Interdiscip. Rev. Syst. Biol. Med., **4**(6), 565–583. DOI: 10.1002/wsbm.1185
4. *Schlüter H., Apweiler R., Holzhьtter H.G., Jungblut P.R.* (2009) Finding one's way in proteomics: A protein species nomenclature. Chem. Cent. J., **3**, 11. DOI: 10.1186/1752-153X-3-11
5. *Smith L.M., Kelleher N.L., Consortium for Top Down Proteomics* (2013) Proteoform: A single term describing protein complexity. Nat. Methods, **10**(3), 186–187. DOI: 10.1038/nmeth.2369
6. *Semba R.D., Enghild J.J., Venkatraman V., Dyrlund T.F., van Eyk J.E.* (2013) The human eye proteome project: Perspectives on an emerging proteome. Proteomics, **13**(16), 2500–2511. DOI: 10.1002/pmic.201300075
7. *Wasinger V.C., Locke V.L., Raftery M.J., Larance M., Rothemund D., Liew A., Bate I., Guilhaus M.* (2005) Two-dimensional liquid chromatography/tandem mass spectrometry analysis of Gradiflow fractionated native human plasma. Proteomics, **5**(13), 3397–3401. DOI: 10.1002/pmic.200401160
8. *Vavilov N., Ilgisonis E., Lisitsa A., Ponomarenko E., Farafonova T., Tikhonova O., Zgoda V., Archakov A.* (2022) Number of detected proteins as the function of the sensitivity of proteomic technology in human liver cells. Curr. Protein Pept. Sci., **23**(4), 290–298. DOI: 10.2174/1389203723666220526092941
9. *Po A., Eyers C.E.* (2023) Top-down proteomics and the challenges of true proteoform characterization. J. Proteome Res., **22**(12), 3663–3675. DOI: 10.1021/acs.jproteome.3c00416

10. *Carvalho A.S., Penque D., Matthiesen R.* (2015) Bottom up proteomics data analysis strategies to explore protein modifications and genomic variants. Proteomics, **15**(11), 1789–1792. DOI: 10.1002/pmic.201400186
11. *Ponomarenko E.A., Poverennaya E.V., Ilgisonis E.V., Pyatnitskiy M.A., Kopylov A.T., Zgoda V.G., Lisitsa A.V., Archakov A.I.* (2016) The size of the human proteome: The width and depth. Int. J. Anal. Chem., **2016**, 7436849. DOI: 10.1155/2016/7436849
12. *Lane L., Argoud-Puy G., Britan A., Cusin I., Duek P.D., Evalet O., Gateau A., Gaudet P., Gleizes A., Masselot A., Zwahlen C., Bairoch A.* (2012) neXtProt: A knowledge platform for human proteins. Nucleic Acids Res., **40**(Database issue), D76-D83. DOI: 10.1093/nar/gkr1179
13. *Lander E.S., Linton L.M., Birren B., Nusbaum C., Zody M.C., Baldwin J., Devon K., Dewar K., Doyle M., FitzHugh W., Funke R., Gage D., Harris K., Heaford A., Howland J., Kann L., Lehoczky J., LeVine R., McEwan P., McKernan K., Meldrim J., Mesirov J.P., Miranda C., Morris W., Naylor J., Raymond C., Rosetti M., Santos R., Sheridan A., Sougnez C., Stange-Thomann Y., Stojanovic N., Subramanian A., Wyman D., Rogers J., Sulston J., Ainscough R., Beck S., Bentley D., Burton J., Clee C., Carter N., Chen Y.J., Szustakowki J., International Human Genome Sequencing Consortium* (2001) Initial sequencing and analysis of the human genome. Nature, **409**(6822), 860–921. DOI: 10.1038/35057062
14. *Ilgisonis E.V., Pogodin P.V., Kiseleva O.I., Tarbeeva S.N., Ponomarenko E.A.* (2022) Evolution of protein functional annotation: Text mining study. J. Pers. Med., **12**(3), 479. DOI: 10.3390/jpm12030479
15. neXtProt downloads. FTP-server. Retrieved August 6, 2024, from: https://download.nextprot.org/pub/previous_releases
16. *Gaudet P., Argoud-Puy G., Cusin I., Duek P., Evalet O., Gateau A., Gleizes A., Pereira M., Zahn-Zabal M., Zwahlen C., Bairoch A., Lane L.* (2013) neXtProt: Organizing protein knowledge in the context of human proteome projects. J. Proteome Res., **12**(1), 293–298. DOI: 10.1021/pr300830v
17. *Li Z., Li S., Luo M., Jhong J.H., Li W., Yao L., Pang Y., Wang Z., Wang R.., Ma R., Yu J., Huang Y., Zhu X., Cheng Q., Feng H., Zhang J., Wang C., Hsu J.B., Chang W.C., Wei F.X., Huang H.D., Lee T.Y.* (2022) dbPTM in 2022: An updated database for exploring regulatory networks and functional associations of protein post-translational modifications. Nucleic Acids Res., **50**(D1), D471–D479. DOI: 10.1093/nar/gkab1017
18. *Yang F., Wang C.* (2020) Profiling of post-translational modifications by chemical and computational proteomics. Chem. Commun. (Cambridge), **56**(88), 13506–13519. DOI: 10.1039/d0cc05447j
19. *Santos A.L., Lindner A.B.* (2017) Protein posttranslational modifications: roles in aging and age-related disease. Oxid. Med. Cell. Longev., **2017**, 5716409. DOI: 10.1155/2017/5716409
20. *Basak S., Lu C., Basak A.* (2016) Post-translational protein modifications of rare and unconventional types: Implications in functions and diseases. Curr. Med. Chem., **23**(7), 714–745. DOI: 10.2174/0929867323666160118095620
21. *Lim C.S., Wardell S.J.T., Kleffmann T., Brown C.M.* (2018) The exon-intron gene structure upstream of the initiation codon predicts translation efficiency. Nucleic Acids Res., **46**(9), 4575–4591. DOI: 10.1093/nar/gky282
22. *Sciarrillo R., Wojtuszkiewicz A., Kooi I.E., Gómez V.E., Boggi U., Jansen G., Kaspers G.J., Cloos J., Giovannetti E.* (2016) Using RNA-sequencing to detect novel splice variants related to drug resistance in *in vitro* cancer models. J. Vis. Exp., **9**(118), 54714. DOI: 10.3791/54714
23. *Roy M., Xu Q., Lee C.* (2005) Evidence that public database records for many cancer-associated genes reflect a splice form found in tumors and lack normal splice forms. Nucleic Acids Res., **33**(16), 5026–5033. DOI: 10.1093/nar/gki792
24. *Cmero M., Schmidt B., Majewski I.J., Ekert P.G., Oshlack A., Davidson N.M.* (2021) MINTIE: Identifying novel structural and splice variants in transcriptomes using RNA-seq data. Genome Biol., **22**, 296. DOI: 10.1186/s13059-021-02507-8
25. *Adamopoulos P.G., Kontos C.K., Scorilas A., Sideris D.C.* (2020) Identification of novel alternative transcripts of the human Ribonuclease κ (RNASEK) gene using 3′ RACE and high-throughput sequencing approaches. Genomics, **112**(1), 943–951. DOI: 10.1016/j.ygeno.2019.06.010
26. *Morales J., Pujar S., Loveland J.E., Astashyn A., Bennett R., Berry A., Cox E., Davidson C., Ermolaeva O., Farrell C.M., Fatima R., Gil L., Goldfarb T., Gonzalez J.M., Haddad D., Hardy M., Hunt T., Jackson J., Joardar V.S., Kay M., Kodali V.K., McGarvey K.M., McMahon A., Mudge J.M., Murphy D.N., Murphy M.R., Rajput B., Rangwala S.H., Riddick L.D., Thibaud-Nissen F., Threadgold G., Vatsan A.R., Wallin C., Webb D., Flicek P., Birney E., Pruitt K.D., Frankish A., Cunningham F., Murphy T.D.* (2022) A joint NCBI and EMBL-EBI transcript set for clinical genomics and research. Nature, **604**(7905), 310–315. DOI: 10.1038/s41586-022-04558-8
27. *Reixachs-Solé M., Eyras E.* (2022) Uncovering the impacts of alternative splicing on the proteome with current omics techniques. Wiley Interdiscip. Rev. RNA, **13**(4), e1707. DOI: 10.1002/wrna.1707
28. *Nesvizhskii A.I., Keller A., Kolker E., Aebersold R.* (2003) A statistical model for identifying proteins by tandem mass spectrometry. Anal. Chem., **75**(17), 4646–4658. DOI: 10.1021/ac0341261
29. *Sinitcyn P., Richards A.L., Weatheritt R.J., Brademan D.R., Marx H., Shishkova E., Meyer J.G, Hebert A.S., Westphall M.S., Blencowe B.J., Cox J., Coon J.J.* (2023) Global detection of human variants and isoforms by deep proteome sequencing. Nat. Biotechnol., **41**(12), 1776–1786. DOI: 10.1038/s41587-023-01714-x
30. *Lek M., Karczewski K.J., Minikel E.V., Samocha K.E., Banks E., Fennell T., O'Donnell-Luria A.H., Ware J.S., Hill A.J., Cumming B.B., Tukiainen T., Birnbaum D.P., Kosmicki J.A., Duncan L.E., Estrada K., Zhao F., Zou J., Pierce-Hoffman E., Berghout J., Cooper D.N., Deflaux N., de Pristo M., Do R., Flannick J., Fromer M., Gauthier L., Goldstein J., Gupta N., Howrigan D., Kiezun A., Kurki M.I., Moonshine A.L., Natarajan P., Orozco L., Peloso G.M., Poplin R., Rivas M.A., Ruano-Rubio V., Rose S.A., Ruderfer D.M., Shakir K., Stenson P.D., Stevens C., Thomas B.P., Tiao G, Tusie-Luna M.T., Weisburd B., Won H.H., Yu D., Altshuler D.M., Ardissino D., Boehnke M., Danesh J., Donnelly S., Elosua R., Florez J.C., Gabriel S.B., Getz G, Glatt S.J., Hultman C.M., Kathiresan S., Laakso M., McCarroll S., McCarthy M.I., McGovern D., McPherson R., Neale B.M., Palotie A., Purcell S.M., Saleheen D., Scharf J.M., Sklar P., Sullivan P.F., Tuomilehto J., Tsuang M.T., Watkins H.C., Wilson J.G., Daly M.J., MacArthur D.G., Exome Aggregation Consortium* (2016) Analysis of protein-coding genetic variation in 60,706 humans. Nature, **536**(7616), 285–291. DOI: 10.1038/nature19057
31. *Omenn G.S., Lane L., Overall C.M., Corrales F.J., Schwenk J.M., Paik Y.K., van Eyk J.E., Liu S., Snyder M., Baker M.S., Deutsch E.W.* (2018) Progress on identifying and characterizing the human proteome: 2018 metrics from the HUPO human proteome project. J. Proteome Res., **17**(12), 4031–4041. DOI: 10.1021/acs.jproteome.8b00441

32. *Senior A.W., Evans R., Jumper J., Kirkpatrick J., Sifre L., Green T., Qin C., Žídek A., Nelson A.W.R., Bridgland A., Penedones H., Petersen S., Simonyan K., Crossan S., Kohli P., Jones D.T., Silver D., Kavukcuoglu K., Hassabis D.* (2020) Improved protein structure prediction using potentials from deep learning. Nature, **577**(7792), 706–710. DOI: 10.1038/s41586-019-1923-7
33. *Walker A.S., Clardy J.* (2021) A machine learning bioinformatics method to predict biological activity from biosynthetic gene clusters. J. Chem. Inf. Model., **61**(6), 2560–2571. DOI: 10.1021/acs.jcim.0c01304
34. *Wright C.J., Smith C.W.J., Jiggins C.D.* (2022) Alternative splicing as a source of phenotypic diversity. Nat. Rev. Genet., **23**(11), 697–710. DOI: 10.1038/s41576-022-00514-4
35. *Chandramouli K., Qian P.-Y.* (2009) Proteomics: Challenges, techniques and possibilities to overcome biological sample complexity. Human Genomics Proteomics, **2009**, 239204. DOI: 10.4061/2009/239204

Поступила в редакцию: 01. 05. 2024.
После доработки: 23. 07. 2024.
Принята к печати: 08. 08. 2024.

THE HUMAN PROTEOME SIZE AS A TECHNOLOGICAL DEVELOPMENT FUNCTION

E.V. Sarygina*, A.S. Kozlova, E.A. Ponomarenko, E.V. Ilgisonis

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

Changes in information on the number of human proteoforms, post-translational modification (PTM) events, alternative splicing (AS), single-amino acid polymorphisms (SAP) associated with protein-coding genes in the neXtProt database have been retrospectively analyzed. In 2016, our group proposed three mathematical models for predicting the number of different proteins (proteoforms) in the human proteome. Eight years later, we compared the original data of the information resources and their contribution to the prediction results, correlating the differences with new approaches to experimental and bioinformatic analysis of protein modifications. The aim of this work is to update information on the status of records in the databases of identified proteoforms since 2016, as well as to identify trends in changes in the quantities of these records. According to various information models, modern experimental methods may identify from 5 to 125 million different proteoforms: the proteins formed due to alternative splicing, the implementation of single nucleotide polymorphisms at the proteomic level, and post-translational modifications in various combinations. This result reflects an increase in the size of the human proteome by 20 or more times over the past 8 years.

The whole English version is available at http://pbmc.ibmc.msk.ru.

Key words: proteomics; proteoforms; post-translational modifications; single-amino acid substitutions; alternative splicing; neXtProt

Funding. The work was carried out within the framework of the Program of Fundamental Research in the Russian Federation for the long-term period (2021–2030) (No. 122030100168-2).

Received: 01.05.2024; revised: 23.07.2024; accepted: 08.08.2024.