На правах рукописи

Пономаренко Елена Александровна

ТРАНСКРИПТОМ И ПРОТЕОМ ХРОМОСОМЫ 18: ЭКСТРАПОЛЯЦИЯ РЕЗУЛЬТАТОВ АНАЛИЗА НА ГЕНОМЫ ЧЕЛОВЕКА И МОДЕЛЬНЫХ ОБЪЕКТОВ

03.01.09 – математическая биология, биоинформатика

03.01.04 – биохимия

АВТОРЕФЕРАТ

диссертации на соискание ученой степени

доктора биологических наук

Москва – 2017

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

Научные консультанты:

доктор биологических наук, профессор, академик РАН **АРЧАКОВ Александр Иванович**,

доктор биологических наук, академик РАН **ЛИСИЦА Андрей Валерьевич**

Официальные оппоненты:

ШАЙТАН Константин Вольдемарович, доктор физико-математических наук, профессор, Московский государственный университет им. М.В.Ломоносова, Биологический факультет, профессор кафедры

ОРЛОВ Юрий Львович, доктор биологических наук, профессор, Федеральное государственное бюджетное научное учреждение «Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук», старший научный сотрудник

ЖАРКОВ Дмитрий Олегович, доктор биологических наук, профессор, Федеральное государственное бюджетное учреждение науки Институт химической биологии и фундаментальной медицины Сибирского отделения Российской академии наук, заведующий лабораторией

Ведущая организация:

Федеральное государственное учреждение «Федеральный исследовательский центр «Фундаментальные основы биотехнологии» Российской академии наук

Защита состоится **« 02 » ноября 2017 г. в 11.00** часов на заседании диссертационного совета Д 001.010.01 на базе Федерального государственного бюджетного научного учреждения «Научно-исследовательский институт биомедицинской химии имени В. Н. Ореховича» по адресу: 119121, Москва, ул. Погодинская, д. 10, стр. 8.

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

Автореферат разослан «___» _____ 2017 г.

Ученый секретарь
диссертационного совета
кандидат химических наук — Карпова Елена Анатольевна

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность проблемы. Выполнение проекта «Геном человека» способствовало развитию методических подходов к исследованию человеческого генома и позволило лучше понять эволюционные отношения между человеком и другими видами. В ходе секвенирования генома (International Human Genome Sequencing Consortium, 2004) были идентифицированы тысячи нуклеотидных последовательностей белок-кодирующих генов. Проект «Геном человека» позволил систематизировать знания о структуре генома, в то время как состав протеома человека до сих пор остается неизвестным (Legrain et al., 2011). Открытым остается вопрос, какое количество генов действительно реализует свою информацию в виде белкового продукта в конкретной клетке, ткани, органе, в организме человека и даже в масштабе всей популяции.

Если предположить, что каждый ген кодирует минимум один белок, то протеом человека должен содержать по крайней мере 20 тыс. немодифицированных (канонических) белков, согласно количеству белок-кодирующих генов в геноме. Количество различных видов белков (далее *протеоформ*, Smith et al., 2013) является следствием альтернативного сплайсинга, реализации в виде одноаминокислотных полиморфизмов единичных нуклеотидных замен в геноме и посттрансляционных модификаций (Roth et al., 2005). За счет комбинации этих и других типов молекулярных событий количество протеоформ существенно превышает число генов в геноме. По некоторым данным, однонуклеотидные замены и альтернативный сплайсинг приводят к образованию до нескольких сотен тысяч различных транскриптов (Altmäe et al., 2014). Наличие посттрансляционных модификаций на протеомном уровне еще больше увеличивает разнообразие белков. Для сравнения, метаболом человека (совокупность низкомолекулярных соединений) состоит всего из нескольких тысяч метаболитов (Botros et al., 2008).

Тематика исследования протеома человека признана актуальной на международном уровне. В 2010 году был начат крупнейший проект «Протеом человека», в котором Россия принимает активное участие. Целью проекта является исследование продуктов экспрессии генов – транскриптов и белков – в хромосомоцентричном формате. Российская часть проекта заключается в исследовании транскриптов и белков, кодируемых хромосомой 18 человека.

В отличие от генома состав протеома не постоянен во времени, меняется в зависимости от внешних факторов и отличается в различных типах биологического материала. Изменения затрагивают как разницу в типах протеоформ, так и концентрацию каждой формы (Kopylov et al., 2016). Многообразие протеоформ может быть обозначено как *ширина* протеома, а количественное содержание каждой протеоформы в образце – как его *глубина* (Ponomarenko et al., 2016).

В 2014 году были опубликованы работы с результатами предварительного анализа полного протеома человека (Wilhelm et al., 2014; Kim et al., 2014). В этих работах протеом представлен в форме каталога – перечня детектированных белков без указания количественного содержания в биоматериале. Отсутствие информации о количестве белков и их концентрации в образце затрудняет интерпретацию протеомных данных и препятствует использованию этих данных в области постгеномной медицины.

Исследователи возлагали на протеомику большие надежды в области использования результатов для диагностики (Ozdemir et al., 2017) и поиска специфичных для заболеваний биологических маркеров – белков. Однако за 15 лет развития постгеномных методов практических результатов в области медицины и диагностики не так много (Veenstra, 2011). Сам подход поиска биологических маркеров заболеваний в протеомном поле служит предметом множества дискуссий (Kondo, 2014; Anderson, 2010). Несмотря на то что показана взаимосвязь между развитием заболеваний и генами (Hall et al., 2010), практическое применение полученных данных остается ограниченным ввиду отсутствия функциональной информации о кодируемых белках (Lek et al., 2016; Lewis et al., 2015). Возможно, ограничения протеомики с точки зрения диагностики заболеваний связаны с отсутствием универсального методического решения для измерения всего многообразия белков. В тоже время динамическая природа протеома и существование протеоформ открывают перспективу исследования протеома человека с учетом многообразия белков.

Цель работы. На примере белков, кодируемых хромосомой 18 человека, создать информационно-аналитическую модель протеома для оценки количества протеоформ с учетом взаимосвязей между геномным, транскриптомным и протеомным уровнями организации, текущего уровня развития аналитических методов и баз данных.

Задачи исследования:

1) выбрать дескрипторы для описания хромосом человека, генома человека и геномов модельных организмов в виде информационных профилей;

2) рассчитать информационные профили и провести сравнительный мета-анализ хромосом человека на основе сведений, предоставляемых молекулярно-биологическими интернет-ресурсами с учетом динамики их наполнения в период с 2011 года;

3) на основе результатов транскриптомных и протеомных экспериментов, выполненных на образцах ткани печени человека и клеточной линии гепатоцитов HepG2, провести анализ корреляции между количественным содержанием транскриптов и белков в образце (оценка *глубины* протеома человека);

4) предложить расчетные модели и оценить многообразие протеоформ, образующихся в результате альтернативного сплайсинга мРНК, реализации на протеомном уровне несинонимичных замен нуклеотидных остатков в геноме и/или вследствие посттрансляционных модификаций (оценка *ширины* протеома человека и модельных объектов);

5) определить условия применимости предложенных расчетных моделей для оценки количества протеоформ путем сопоставления аннотаций генов человека и аннотаций наиболее популярных модельных объектов исследования – представителей разных царств живой природы;

6) предложить количественный способ оценки результатов протеомного анализа и определить показатели чувствительности, специфичности и точности аналитических методов, применяемых для обнаружения (детекции), идентификации и количественного анализа белков в составе протеома человека.

Научная новизна. В данной работе впервые проведен сравнительный анализ хромосом человека на основе накопленных геномных аннотаций. Результаты позволили обосновать выбор хромосомы 18 для российской части международного проекта «Протеом человека» (Ponomarenko et al., 2012), как оптимальной по соотношению количества белок-кодирующих генов и их медицинской значимости. Впервые показано, что информационный профиль выборки случайным образом отобранных генов человека совпадает с полногеномным.

В хромосомоцентричном формате проведено сопоставление экспериментальных результатов направленного и панорамного транскриптомного и протеомного исследований биологических образцов и показано отсутствие количественной взаимосвязи между транскриптами и белками, кодируемыми генами одной хромосомы. Впервые предложен способ количественной оценки результатов протеомного анализа на основе показателей чувствительности, специфичности и точности (по аналогии с показателями диагностических тестов).

Для оценки многообразия протеоформ в составе протеома предложены методы расчета, позволяющие предсказать количество видов белков (в конкретном образце ткани, организме или в масштабе всей популяции) и возможность экспериментальной детекции протеоформы с учетом ранее накопленных в молекулярно-биологических интернет-ресурсах знаний. Проведенная работа позволяет по-новому взглянуть на аналитическую биохимию, как на совокупность технологий обнаружения, идентификации и количественного анализа белков с учетом многообразия их протеоформ.

Основные положения, выносимые на защиту:

1. Хромосомы человека тождественны по *информационному профилю*, за исключением наиболее коротких хромосом – Y и митохондриальной. *Информационный профиль* выборки, содержащей двести и более случайным образом отобранных генов человека, совпадает с полногеномным.

2. Результаты транскриптомного анализа могут применяться в качестве стандарта для оценки чувствительности, специфичности и точности метода протеомного анализа. Для этого в геноцентричном формате количество копий белков необходимо сопоставлять

с количеством копий транскриптов, измеренных в том же образце методами со сравнимой аналитической чувствительностью.

3. Для оценки количества протеоформ может быть использована информационная модель и соответствующая расчетная формула, согласно которой события альтернативного сплайсинга, возникновение одноаминокислотных полиморфизмов и посттрансляционных модификаций являются независимыми. Для оценки *ширины* популяционного протеома могут быть использованы частоты таких событий из протеомных баз данных, при оценке индивидуального протеома – из результатов транскриптомного анализа образца биоматериала конкретного индивидуума.

Научно-практическая значимость работы. Работа открывает новый этап использования протеомики и постгеномных технологий для молекулярного профилирования человека и поиска биологических маркеров. В перспективе для оценки адаптационного потенциала человека может применяться мониторинг стабильности во времени индивидуального набора протеоформ человека при заданном уровне *глубины*, доступном для анализа высокопроизводительными аналитическими методами. Прикладным аспектом оценки многообразия протеоформ является создание специализированных библиотек аминокислотных последовательностей, используемых при интерпретации масс-спектрометрических данных.

Результаты работы обеспечивают современную методологию обработки молекулярно-биологической информации, способы интеграции данных и проекции данных на персонализированный молекулярный профиль индивидуума. В работе продемонстрирована возможность перехода от популяционного уровня обобщения данных к спектру специфичных для конкретного организма молекулярных изменений первичной структуры белков. Практическое значение заключается в характеристике *ширины* и *глубины* протеомов человека и других организмов.

Личный вклад автора. Автором разработан дизайн хромосомоцентричного исследования транскриптома и протеома хромосомы 18 (в рамках международного проекта «Протеом человека»), обеспечено планирование экспериментальных работ и обработка полученных данных. Также автором предложена система описания набора генов с учетом текущего уровня знаний, отраженного в постгеномных информационных ресурсах. В геноцентричном режиме на основе сопоставления экспериментальных результатов и опубликованных данных формируется интегральная характеристика выборки генов – *информационный профиль*, совпадающий для случайным образом отобранной группы генов и полного генома человека. Для оценки количества протеоформ предложено три информационные модели, а также соответствующие моделям расчетные формулы, учитывающие частоты молекулярных событий, источник данных и тип биологического материала. На основе предложенных моделей автором проведена оценка

количества протеоформ, составляющих протеомы модельных объектов, хромосом человека, ткани печени и клеточной линии HepG2.

Апробация работы. Основные результаты работы доложены и обсуждены на ежегодных итоговых конференциях и конгрессах «Мир биотехнологий» (Москва, 2017), HUPO (Taipei, 2016), HUPO (Vancouver, 2015), EUPA (Milano, 2015), 7th AOHUPO Congress and 9th International Symposium of the Protein Society of Thailand (Bangkok, 2014), HUPO (Madrid, 2014) и др. Результаты работы доложены и обсуждены на заседании Бюро секции медико-биологических наук отделения медицинских наук Российской академии наук 20 декабря 2016 года.

Публикации. Материалы диссертационной работы отражены в 76 публикациях: в 34 статьях (в 13 российских и 21 международных научных журналах), 42 материалах российских и международных научных конференций. Индекс Хирша соискателя ученой степени составляет 10 по данным системы Scopus.

Объем и структура диссертации. Диссертационная работа изложена на 270 страницах машинописного текста, включая 18 таблиц, 29 рисунков. Состоит из введения, обзора литературы, описания исходных данных и методов обработки, результатов и обсуждения, заключения, выводов и списка литературы, включающего 345 источников.

СОДЕРЖАНИЕ РАБОТЫ

В обзоре литературы рассматриваются цели, задачи и сложности реализации международного проекта «Протеом человека», дается описание основных информационных ресурсов для хранения и обработки постгеномных данных. Отдельная глава посвящена описанию экспериментальных методов, используемых для анализа транскриптома и протеома биообразцов в направленном и панорамном режимах. Возможности постгеномных технологий и перспективы их использования для сохранения здоровья человека раскрыты в заключительной части обзора литературы.

ИСХОДНЫЕ ДАННЫЕ И МЕТОДЫ ОБРАБОТКИ

Сравнительный анализ хромосом человека

Информацию о количестве белок-кодирующих генов, локализованных на каждой хромосоме человека, получали из интернет-ресурса NextProt (Gaudet et al., 2015). NextProt для каждого белок-кодирующего гена предоставляет описание (аннотацию) – результат обработки информационных ресурсов, экспериментальных данных, биоинформатических предсказаний и анализа текстов научных публикаций. Аннотация содержит описание функций, свойств аминокислотных последовательностей, взаимосвязи с развитием заболеваний, ссылок на другие информационные ресурсы и др. Использовали раздел аннотации NextProt (UniProt), описывающий кодируемые геном белки здорового человека – информацию о вариантах белковой последовательности, возникающих в результате альтернативного сплайсинга, реализации на протеомном уровне однонуклеотидных замен в геноме и посттрансляционных модификаций.

Сведения о детектированных транскриптах получали из базы данных ProteinAtlas (Uhlen et al., 2015), количество детектированных белков – из ресурсов PeptideAtlas (все типы биологического материала, Desiere et al., 2006) и Plasma Proteome Database (количество белков, детектированных в плазме крови человека, Nanjappa et al., 2013). Для каждой хромосомы рассчитывали долю генов, для которых согласно аннотациям MalaCards (Rappaport et al., 2014) показана взаимосвязь с возникновением патологических состояний. С использованием модуля автоматического анализа резюме научных публикаций (Пономаренко и соавт., 2010) для каждой хромосомы подсчитывали количество генов, названия которых упоминаются в статьях в системе реферирования статей PubMed. Запросы к информационным ресурсам формировали в автоматическом режиме.

Каждая хромосома человека была охарактеризована *информационным профилем* – совокупностью доступных в информационных ресурсах описаний, где параметру (*дескриптору*) соответствует доля аннотированных генов (*коэффициент дескриптора*). Для каждого дескриптора информационного профиля рассчитывали среднее значение – *вес дескриптора* – по всему геному, а также границы доверительного интервала и уровень отклонения по этому показателю для конкретной хромосомы и случайным образом сформированных выборок генов.

Дизайн хромосомоцентричного исследования

В работе анализировали транскриптомный состав клеток ткани печени человека и клеточной линии HepG2, протеом тех же образцов, а также протеом образцов плазмы крови здорового человека. Образцы плазмы крови здоровых добровольцев получали от ГНЦ ИМБП РАН. В банке биоматериала *ILSBioBiobank* приобретали образцы ткани

печени доноров, смерть которых наступила в результате несчастного случая. Клетки линии HepG2 выращивали согласно стандартному протоколу. Образцы ткани печени и линии HepG2 разделяли на аликвоты и исследовали транскриптомный и протеомный состав с использованием направленных и панорамных методов.

Транскриптомный анализ. мРНК, выделенная из образцов ткани печени и клеток HepG2, исследована параллельно с использованием двух платформ РНК-секвенирования – SOLiD4 и Illumina, в трех технических повторах для каждого образца. Полученный уровень экспрессии сопоставляли с данными ПЦР-анализа тех же образцов. Использовали объединенные результаты ПЦР в реальном времени ($ПЦР_{рв}$, qRT-PCR) и цифровой капельной ПЦР ($ПЦР_{цк}$, ddPCR).

Протеомный анализ. Детекцию и количественное измерение белков в трех типах биоматериала проводили параллельно методами направленной масс-спектрометрии методом мониторинга множественных реакций (ММР, использовали прибор QQQ 6490, Agilent) и с помощью панорамного масс-спектрометрического анализа (MC/MC) на приборе Q-Exactive HF (Thermo Scientific, Germany). Панорамную масс-спектрометрию использовали для идентификации белков без проведения количественного анализа. Количественные измерения проводили в серии экспериментов с использованием в качестве стандарта пептидов с изотопной меткой (далее обозначали как $ММР_{мс}$). Результаты измерений сопоставляли с данными о количественном содержании белка, полученными при использовании в качестве стандарта пептидов без изотопной метки (направленное детектирование эндогенных пептидов без использования стандартов, обозначение метода – $ММР_{нс}$). Выбор изотопно-меченых пептидов соответствовал международным стандартам постановки масс-спектрометрических измерений для количественного анализа (применяли уровни достоверности данных согласно Carr et al., 2014). Пептиды без изотопных меток рекомендованы для использования на начальном этапе протеомного исследования согласно той же статье.

Методические особенности проведения экспериментов изложены в статьях (Zgoda et al., 2013; Ponomarenko et al., 2014; Kopylov et al., 2016; Poverennaya et al., 2016). Сведения о количественном содержании транскриптов и белков хромосомы 18 человека получали из дополнительных материалов к указанным выше статьям.

Анализ транскриптомных и протеомных данных

Анализировали данные по значениям RPKM (программно рассчитываемая характеристика экспрессии гена (транскрипта) – *Read counts per kilobase per million*), полученные в трех технических повторах, при этом отбирали только те значения, коэффициент вариации между которыми не превышал 30 % (CV < 30 %) в двух из трех повторов. Использовали среднее значение уровня PRKM между техническими повторами для количественного представления уровня экспрессии каждого гена.

Детектированными считали транскрипты, зарегистрированные методом РНК-секвенирования ($РНК_{сек}$) на уровне RPKM выше порогового, или обнаруженные методами ПЦР ($C_t \leq 30$, C_t – пороговый цикл, т. е. количество циклов амплификации после проведения которых флуоресценция расщепленного зонда превысила фоновые значения).

Зависимость количества сплайс-вариантов, или транскриптов, содержащих несинонимичные полиморфизмы, от количества образцов, исследовали на основе материалов ресурса *NCBI Sequence Read Archieve* (Kodama et al., 2012), содержащего данные экзомного профилирования образцов ткани печени здоровых людей. Для аннотации использовали референсную сборку генома hg19/GRCh37.p10.

Полученные с использованием направленной масс-спектрометрии без изотопно-меченых стандартов результаты измерений белков распределили по двум категориям. Первая категория экспериментов имеет статус $ММР_{нс}$* – белки детектированы с использованием двух протеотипических пептидов на один белок, соотношение концентраций которых находится в пределах одного порядка. Ко второй категории, которая обозначена как $ММР_{нс}$, отнесли всю совокупность полученных таким методом экспериментальных результатов.

Для сопоставления экспериментальных результатов анализа транскриптома и протеома рассчитывали показатели чувствительности, специфичности и точности (Guidance for Industry and FDA Staff, 2007). *Чувствительность (Se)* – способность аналитического (диагностического) метода давать правильный результат, который определяется как доля истинно-положительных результатов среди всех проведенных измерений (доля детектированных белков, для которых в том же образце наблюдали транскрипт). *Специфичность (Sp)* – способность метода не давать ложноположительных результатов (доля недетектированных белков, для которых в том же образце отсутствовал транскрипт). *Точность (Ac)* – доля правильных результатов (сумма истинно-положительных и истинно-отрицательных случаев):

Чувствительность (Se):
$$\mathrm{Se} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \times 100\ \% \quad (1)$$

Специфичность (Sp):
$$Sp = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} \times 100\ \% \quad (2)$$

Точность (Ac):
$$Ac = \frac{(\mathrm{TP} + \mathrm{TN})}{\mathrm{N}} \times 100\ \% \quad (3)$$

где *TP* (*true positive, истинно-положительные*) – количество белок-кодирующих генов, для которых детектирован транскрипт и белок;

FP (*false positive, ложно-положительные*) – белок детектирован при отсутствии транскрипта;

TN (*true negative, истинно-отрицательные*) – транскрипт и белок отсутствуют;

FN (*false negative, ложно-отрицательные*) – детектирован только транскрипт, а белок – нет;

N = 275, количество генов на хромосоме 18 человека.

Методика оценки количества протеоформ человека и модельных объектов

Используемый в работе подход основан на предположении о том, что количество протеоформ (*Np*) может быть рассчитано как *Np* (*H) = f (p; H*), где *p* – набор параметров, загружаемых из молекулярно-биологических информационных ресурсов, а *H* – информационная модель, отражающая взаимосвязи между параметрами *p*. Предположительно, значения параметров *p* не являются постоянными во времени, что следует из наблюдения: базы данных постоянно пополняются. Таким образом, *p =f (t; D),* где *D* – базы данных, выбираемые для оценки *p*. Зависимость от времени *t* и от ресурса *D* позволяет выполнить исследование анализа соотношений между p_1 и p_2: $p_1/p_2=f(t_1, D)/f(t_2/ D)$ или $p_1/p_2= f(t, D_1)/f(t/D_2)$, где *t* без индекса и *D* без индекса условно являются константами. В таблице 1 приведен перечень параметров *p*, используемых для оценки количества протеоформ.

Расчет количества протеоформ проводили для человека и ряда модельных объектов. В перечень модельных организмов были включены 18 видов, для которых секвенированы полные геномы и которые чаще упоминаются в текстах научных публикаций. Среди отобранных организмов были представители царства бактерий, грибов, растений и животных.

Ширина протеома как параметрическая функция. В работе рассматривали несколько расчетных информационных моделей, отражающих гипотетические варианты комбинации молекулярных событий, приводящих к возникновению протеоформ. Рассматривали процессы альтернативного сплайсинга (АС), возникновения одноаминокислотных полиморфизмов (несинонимичных замен, ОАП) и посттрансляционных модификаций (ПТМ). Предложенные оценки не учитывали сведения о совместных частотах встречаемости событий разного типа и комбинаторные варианты ввиду отсутствия экспериментально подтвержденных данных.

Согласно **модели 1** предположим, что изменения первичной структуры белка (одноаминокислотные замены (ОАП), либо посттрансляционные модификации (ПТМ)) возникают только в условно канонических (согласно определению, данному в базе UniProt[1]) вариантах последовательностей аминокислотных остатков. Предполагая, что альтернативный сплайсинг (АС), возникновение ПТМ и ОАП происходит во всех генах, рассматривали следующие варианты (4):

[1] Консорциум UniProt определяет последовательность как каноническую в случае, если это (а) наиболее распространенная форма белка, (б) наиболее сходная с ортологичными последовательностями других видов или, в случае отсутствия такой информации, (в) наиболее длинная из всех вариантов аминокислотных последовательностей гена. (http://www.uniprot.org/faq/30).

$$Nps_{1.1} = N \times (1 + ASav + SAPav + PTMav) \qquad (4)$$

где *Nps* – количество протеоформ, при этом индекс 1.1 отражает используемую расчетную модель; *N* – количество белок-кодирующих генов; *ASav/SAPav /PTMav* – среднее количество АС/ОАП/ПМТ на один ген.

Таблица 1 – Перечень параметров, используемых при оценке количества протеоформ (*ширины* протеома)

Обозначение	Расшифровка	Источник (способ расчета)
N	Количество белок-кодирующих генов в геноме	Загружали из UniProt, NeXtProt или Ensembl
ASg	Количество генов, подвергающихся альтернативному сплайсингу	Загружали из UniProt или NeXtProt
ASd	Доля генов в геноме, подвергающаяся альтернативному сплайсингу	Рассчитывали как отношение ASg к N
AS	Количество аминокислотных последовательностей, образованных в результате альтернативного сплайсинга	Загружали из UniProt или NeXtProt
ASav	Среднее количество сплайс-вариантов на один ген	Рассчитывали как отношение AS к ASg
SAPg	Количество генов, содержащих однонуклеотидные замены, реализуемые в одноаминокислотные полиморфизмы (ОАП)	Загружали из UniProt или NeXtProt
SAPd	Доля генов в геноме, содержащих однонуклеотидные замены, реализуемые в одноаминокислотные полиморфизмы (ОАП)	Рассчитывали как отношение SAPg к N
SAP	Количество аминокислотных последовательностей, содержащих ОАП	Загружали из UniProt или NeXtProt
SAPav	Среднее количество вариантов последовательностей, содержащих ОАП, на один ген	Рассчитывали как отношение SAP к SAPg
PTMg	Количество генов, соответствующих белкам с ПТМ	Загружали из UniProt или NeXtProt
PTMd	Доля генов в геноме, кодирующих ПТМ-содержащие белки	Рассчитывали как отношение PTMg к N
PTM	Количество аминокислотных последовательностей, содержащих ПТМ	Загружали из UniProt или NeXtProt
PTMav	Среднее количество вариантов последовательностей, содержащих ПТМ, на один ген	Рассчитывали как отношение PTM к PTMg
Np (H)	Количество протеоформ (*ширина* протеома)	Формула расчета количества протеоформ, где в качестве аргумента указывается информационная модель, например (1.2)

Формула (5) учитывает долю генов (кодируемых белков), подвергающихся каждому из рассматриваемых типов модификаций (условные обозначения долей генов, кодирующих белки: ASd – со сплайс-вариантами, $SAPd$ – с одноаминокислотными заменами и $PTMd$ – с посттрансляционными модификациями):

$$Nps_{1.2} = N \times (1 + ASd \times ASav + SAPd \times SAPav + PTMd \times PTMav) \quad (5)$$

Модель 2. Предполагали, что изменения аминокислотной последовательности (одноаминокислотные замены (ОАП), либо посттрансляционные модификации (ПТМ)) возможны также и в сплайс-вариантах, известных из данных РНК-секвенирования. Тогда возникновение ПТМ и ОАП происходит во всех сплайс-вариантах независимо, согласно формуле (6):

$$Nps_{2.1} = Nps_{1.2} + AS \times (SAPav + PTMav) \quad (6)$$

Модель 3. Предположим, возникновение ПТМ происходит в любых аминокислотных последовательностях. Это означает, что совместное возникновение ПТМ и ОАП происходит во всех сплайс-вариантах и канонических последовательностях (7):

$$Nps_{3.1} = Nps_{2.1} + N \times SAPav \times PTMav + AS \times SAPav \times PTMav \quad (7)$$

Расчетные значения многообразия протеоформ, полученные с использованием предложенных трех различных моделей, были сопоставлены с другими информационными категориями, а также степенью изученности объекта и количеством сведений в информационных ресурсах. На основе полученных данных определяли вклад различных молекулярных событий в разнообразие спектра протеоформ и границы применимости предложенных моделей.

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

Сравнительный анализ хромосом человека

Развитие постгеномных информационных ресурсов дает возможность описать каждую хромосому в виде совокупности числовых характеристик входящих в нее дискретных объектов – белок-кодирующих генов (БКГ). При этом используется гено-(или хромосомо-) центричный подход (Paik et al., 2012; Lane et al., 2014), заключающийся в том, что результаты анализа методами биоинформатики и результаты экспериментального исследования транскриптов и белков картируются на гены определенной хромосомы. В основе хромосомоцентричного подхода лежит гипотеза о том, что в информационном плане каждая хромосома тождественна другим хромосомам. Это означает, что совпадают *информационные профили* хромосом, т. е. доли генов, охарактеризованных в четких категориях: например, доля генов, для которых есть экспериментально детектированные транскрипты, совпадает между хромосомами.

Было проведено сравнение полногеномного информационного профиля и профилей отдельных хромосом человека, и установлено минимальное количество генов в выборке, для которых выполняется условие тождественности с полногеномным информационным профилем. Сопоставление хромосом проводили по критериям: 1) количество генов на хромосоме; 2) доля генов, кодирующих белки, содержащие сплайс-варианты (АС), одноаминокислотные полиморфизмы (ОАП) или ПТМ; 3) доля генов, для которых экспериментально детектированы транскрипты и белки, показана взаимосвязь с развитием патологий или названия которых встречаются в резюме научных публикаций.

Геном человека включает около 20 тыс. белок-кодирующих генов, расположенных на 22 соматических хромосомах, половых хромосомах (Х, У) и митохондриальной хромосоме. Наименьшее количество генов известно для митохондриальной хромосомы (14 генов согласно базе NeXtProt, v.2016_01 (Gaudet et al., 2015)) и Y хромосомы (47 генов, см. таблицу 2). Среди соматических хромосом меньше всего генов на хромосоме 21, всего 250. Лидирует по количеству генов хромосома 1, содержащая более 2 тыс. генов. В среднем по геному количество генов составляет 802,3 ± 189,5 на хромосому (среднее ± ст. ошибка, 95 % – доверительный интервал). Разница в количестве генов между хромосомами достигает одного порядка величины.

В таблице 2 представлены сведения об информационных профилях хромосом человека и средние значения, отражающие частоту встречаемости сплайс-вариантов, аминокислотных полиморфизмов и посттрансляционных модификаций. Значения, полученные на основе сведений базы данных NeXtProt (v.2016_01), нормировали на количество генов на хромосоме. Оказалось, что для хромосомы 18 человека описано всего 326 сплайс-вариантов транскриптов, соответствующих 164 генам. Отношение этих показателей (326 / 164) дает величину 1,99 варианта альтернативного сплайсинга (АС) на

один ген (см. таблицу 2). Аналогично, величина 50,24 ОАП на один ген получается как отношение 1198 ОАП, упомянутых в аннотациях 195 генов этой хромосомы.

Исходя из данных, представленных в таблице 2, можно заключить, что большинство хромосом человека не различается по среднему количеству аннотированных сплайс-вариантов, полиморфизмов или модификаций на один ген. Исключение составляют наиболее короткие хромосомы – Y и митохондриальная. Данные, приведенные в таблице 2, указывают, что наибольший вклад в разнообразие белков вносит наличие однонуклеотидных полиморфизмов в геноме, которые затем реализуются на протеомном уровне в виде замен аминокислотных остатков в белке, а не альтернативный сплайсинг или пост-трансляционные модификации.

Таблица 2 – Частота встречаемости* сплайс-вариантов (АС), одноаминокислотных полиморфизмов (ОАП) и посттрансляционных модификаций (ПТМ) для хромосом человека (CHR) в протеомной базе NeXtProt (v. 2016_01)

CHR	БКГ	АС	ОАП	ПТМ	CHR	БКГ	АС	ОАП	ПТМ
CHR_1	2056	2,19	45,51	7,54	**CHR_15**	602	1,94	44,92	7,52
CHR_2	1231	2,19	46,03	7,39	**CHR_16**	837	1,90	42,48	6,18
CHR_3	1070	2,21	47,51	6,97	**CHR_17**	1166	2,02	42,82	6,80
CHR_4	761	2,09	45,02	7,29	**CHR_18**	276	1,99	50,24	6,14
CHR_5	868	2,00	44,82	7,29	**CHR_19**	1428	1,83	46,18	6,80
CHR_6	1110	2,10	44,20	7,45	**CHR_20**	550	2,00	44,13	6,31
CHR_7	934	2,11	46,16	6,96	**CHR_21**	250	2,57	42,46	8,90
CHR_8	701	2,03	44,06	6,78	**CHR_22**	460	1,89	46,50	6,89
CHR_9	810	2,17	44,65	7,24	**CHR_X**	824	2,04	44,77	6,88
CHR_10	761	2,30	44,77	7,25	**CHR_Y**	47	1,68	16,42	3,31
CHR_11	1322	2,09	45,94	6,59	**CHR_MT**	14		37,00	2,00
CHR_12	1029	2,19	46,82	7,35	**Среднее**	802,3	2,1	43,7	6,7
CHR_13	327	1,99	43,54	7,37	**Ст. откл.**	459,0	0,2	6,1	1,4
CHR_14	623	2,11	45,47	6,79	**Дов. инт. (p = 0,95)**	802,3 ± 189,5	2,1 ± 0,3	43,7 ± 10,0	6,7 ± 2,2

* Отношение общего количества аннотаций в базе NeXtProt (v.2016_01) к количеству аннотированных белок-кодирующих генов (БКГ). Цветом обозначены значения, выходящие за границы интервала распределения 90 % выборки (среднее ± 1,64 × стандартное отклонение, доверительный уровень – 90 %).

Основными информационными ресурсами, консолидирующими сведения о белках человека, являются UniProt (Boutet et al., 2016) и NeXtProt (Gaudet et al., 2015). Оба ресурса интегрируют информацию из ряда источников, при этом в UniProt собрана информация о белках многих видов, в том числе растений, грибов и бактерий. С другой стороны, NeXtProt является ресурсом, сфокусированным исключительно на белках человека. В рамках работы было сопоставлено количество данных о многообразии белков человека (сведениях о АС, ОАП или ПТМ), накопленных в этих ресурсах. Стояла задача проанализировать, какие информационные категории совпадают между этими

системами, а какие – существенно различаются. Это имеет значение для оценки результатов, полученных при сопоставлении человека и модельных объектов. Анализировали, как изменяется количество накопленных данных в зависимости от времени, т. е. оценивали тенденции накопления данных о сплайс-вариантах, одноаминокислотных заменах и посттрансляционных модификациях в UniProt и NeXtProt.

Анализ изменения количества аннотаций для одного гена в зависимости от версии ресурса показал, что среднее количество обнаруженных сплайс-вариантов, одноаминокислотных замен и посттрансляционных модификаций в расчете на один ген человека стабилизировалось в 2013 году (см. рисунок 1). Можно заключить, что сопоставление хромосом человека на основе сведений информационных ресурсов в текущий момент времени правомерно, поскольку не наблюдается тенденций к резкому изменению количества информации о вариативности белков человека в анализируемых ресурсах. Для исследования многообразия форм белков человека могут быть использованы данные как базы UniProt, так и NeXtProt, принимая во внимание стабильность анализируемых величин во времени.

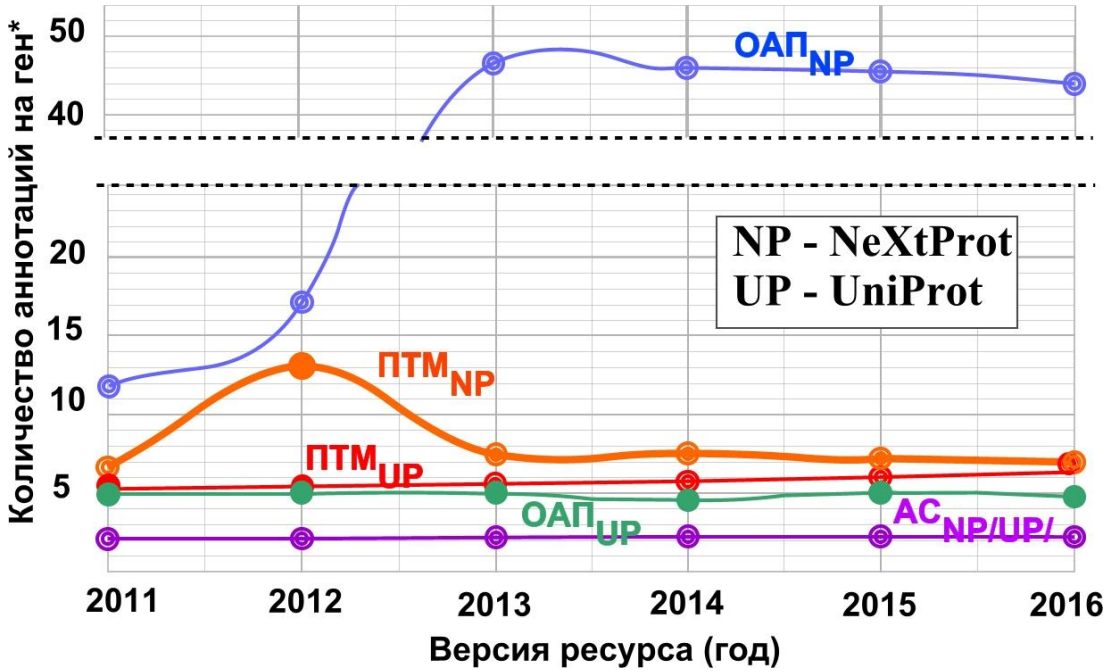

Рисунок 1. Изменение количества аннотаций белков человека в ресурсах UniProt и NeXtProt в период 2011–2016 гг. (*) Для генов человека, имеющих минимум одну аннотацию в категориях: наличие одноаминокислотных полиморфизмов (ОАП), сплайс-вариантов (АС), посттрансляционных модификаций (ПТМ)

Обнаружение одноаминокислотных замен и сплайс-вариантов белков не привязано к чувствительности методов аналитической биохимии. Информация об этих событиях накапливается в ходе крупномасштабных проектов по секвенированию, поскольку есть возможность амплификации молекул. Напротив, для обнаружения ПТМ амплификация аналита пока невозможна. Несмотря на качественные различия между

экспериментальными возможностями исследования ОАП/АС и ПТМ, все три показателя синхронно вышли на уровни насыщения. Это указывает на баланс между данными, полученными с использованием стандартных методов белковой химии (накопленными за последние 50 лет), и данными, полученными с использованием высокопроизводительных методов, таких как РНК-секвенирование нового поколения и протеомная масс-спектрометрия.

Для сопоставления хромосом человека использовали интегральную характеристику – *информационный профиль*. Информационный профиль отражает соотношение между долями генов выборки, имеющими аннотации определенной категории в постгеномных базах знаний или научных публикациях. Было показано, что доля генов, для которых имеются описания (например, количество детектированных транскриптов, белков, взаимосвязь с развитием патологических состояний, упоминание в публикациях и др.), в целом постоянна для большинства хромосом и генома человека (Ponomarenko et al., 2012).

В контексте данной работы в состав информационного профиля были включены дескрипторы, по которым сопоставляли хромосомы человека. Последовательность аннотаций (дескрипторов) в информационном профиле была выбрана следующая:

N п/п	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Дескриптор	[ОАП]	[Транскр]	[Статьи]	[Белки]	[ПТМ]	[АС]	[Патология]	[Плазма]
Вес	95	85	84	70	69	53	48	1

где -

1, 5, 6 – [ОАП] ([ПТМ], [АС]) – доля генов, в аннотациях к которым, согласно базе знаний UniProt, есть информация как минимум об одном одноаминокислотном полиморфизме в последовательности кодируемого белка (по аналогии – о наличии посттрансляционной модификации [ПТМ] или последовательности, образованной в результате альтернативного сплайсинга [АС]);

2 – [Транскр] – доля генов, для которых экспериментально детектированы транскрипты, согласно информации базы данных GTEx (Lonsdale et al., 2013);

3 – [Статьи] – доля генов, названия которых встречаются в текстах научных публикаций в библиотеке биомедицинских текстов PubMed/MEDLINE (NCBI Resource Coordinators, 2014);

4 – [Белки] – доля генов, для которых экспериментально детектированы белки, согласно информации из ресурса PeptideAtlas (Desiere et al., 2006);

7 – [Патология] – доля генов, для которых есть сведения о взаимосвязи с возникновением или развитием патологических процессов в базе данных MalaCards (Rappaport et al., 2016);

8 – [Плазма] – доля генов, для которых в базе данных Plasma Proteome DB (Nanjappa et al., 2014) имеется информация о концентрации белка в плазме крови человека.

Для создания информационного профиля не имеет значения порядок следования дескрипторов: важно сохранение этого порядка при сопоставлении информационных профилей различных хромосом или наборов генов. В нашей работе дескрипторы упорядочены по убыванию доли генов с соответствующими характеристиками (весовые коэффициенты дескриптора). При мета-анализе постгеномных аннотаций генов в геноме человека были получены следующие весовые коэффициенты дескрипторов информационного профиля: около 95 % (± 2) могут кодировать белки, содержащие ОАП, для 85 % (± 7) генов экспериментально обнаружены продукты экспрессии на транскриптомном уровне. Сопоставимая доля генов будет упомянута в текстах научных публикаций – 84% (± 4), для 70 % (± 5) генов методами панорамной масс-спектрометрии определены соответствующие белки. Сведения о ПТМ и АС присутствуют в аннотациях 69 % (± 6) и 53 % (± 4) генов, соответственно. Взаимосвязь с развитием патологических процессов показана примерно для половины генов – 48% (± 5), но только для 1 % (± 1) измерено содержание белков в плазме крови – самом перспективном для диагностики типе биологического материала.

На рисунке 2 представлены информационные профили хромосом человека. Рассчитанное соотношение между весовыми коэффициентами дескрипторов в составе информационного профиля нарушается для наиболее коротких хромосом человека – Y (N=47) и митохондриальной (N=14). Для этих хромосом наблюдается существенное изменение информационного профиля в сравнении с другими хромосомами человека (см. рисунок 2). Нарушение информационного профиля связано с небольшим количеством белок-кодирующих генов (N), а не с биологическими причинами, выделяющими эти хромосомы.

Случайным образом из совокупности генов человека формировали выборки разного размера с целью определить минимальное количество генов в составе выборки, для которой сохраняются указанные выше весовые коэффициенты информационного профиля. Случайным образом (n=100) отбирали выборки, содержащие N генов, при этом N варьировали в интервале от 50 до 20 000 генов. Для каждой выборки рассчитывали долю генов, аннотированных в указанных выше ресурсах (UniProt, GTEx, Pubmed, PeptideAtlas, Plasma Proteome DB и MalaCards), значения дескрипторов и весовые коэффициенты дескрипторов в составе информационного профиля выборки. Для выборок каждой группы (одинакового размера) подсчитывали величину стандартного отклонения между характеристиками группы и коэффициентами информационного профиля, полученного для всего генома человека.

Получено, что лишь незначительные изменения наблюдаются для коэффициентов информационного профиля при размере выборки N >200. При дальнейшем увеличении выборки стандартное отклонение между коэффициентами дескрипторов информационного профиля составляет менее 3 % от коэффициентов информационного профиля для всего генома человека. Для выборок объемом менее 100 генов величины стандартного отклонения достигают 10 %. Это не позволяет считать такие группы генов

репрезентативными, т. е. выборками, весовые коэффициенты дескрипторов информационного профиля которых сопоставимы с информационным профилем всего генома человека. К такой группе относятся и наиболее короткие хромосомы человека – хромосома Y и митохондриальная хромосома (Mt).

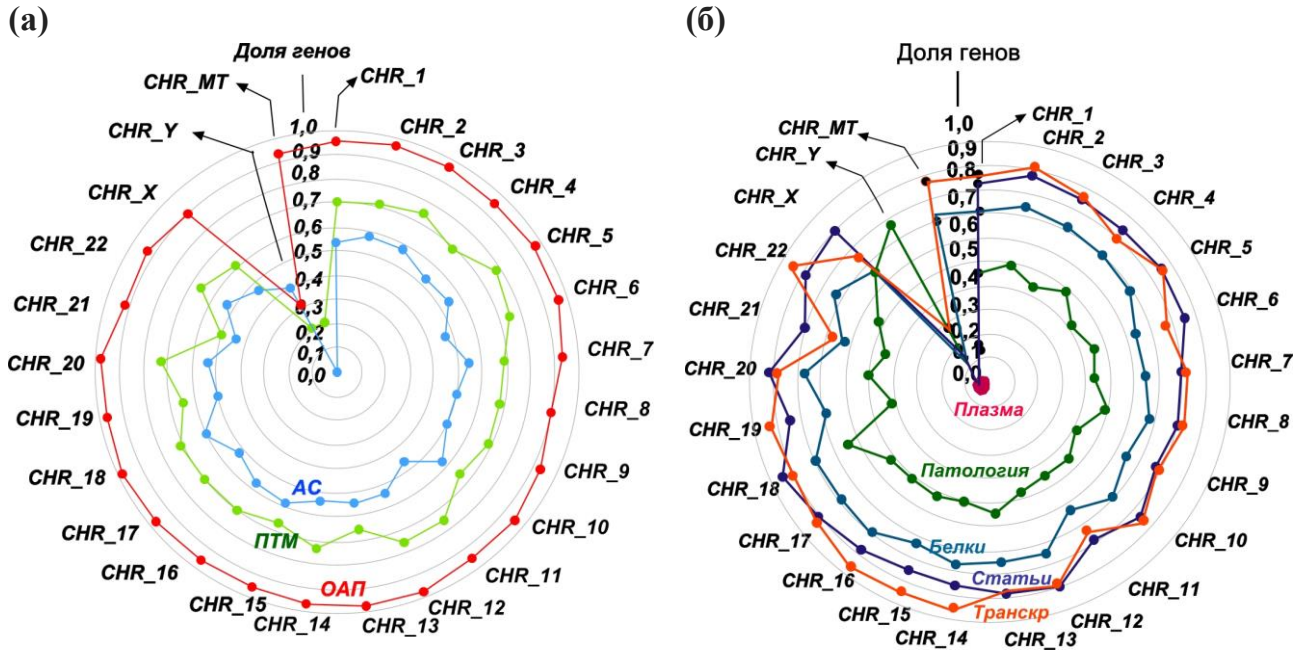

Рисунок 2. Информационные профили хромосом человека (CHR), полученные в результате мета-анализа постгеномных информационных ресурсов: **(а)** соотношение количества аннотаций (дескрипторов) [ОАП] – одноаминокислотные полиморфизмы, [ПТМ] – посттрансляционные модификации и [АС] – альтернативный сплайсинг; **(б)** сведения о наличии транскриптов – [Транскр] , публикаций – [Статьи], детектированных белков – [Белки], данных, подтверждающих взаимосвязь с развитием патологических процессов – [Патология], сведений о концентрации белка в плазме крови человека – [Плазма]

Использование *информационного профиля* позволяет оценить смещение выборки генов по каким-либо информационным категориям. Смещение по одной категории (например, выбор хорошо изученных генов или генов, связанных с развитием патологий) закономерно приводит к изменению весовых коэффициентов взаимосвязанных дескрипторов. В большинстве случаев взаимосвязь обусловлена человеческим фактором – исследовательским интересом, социальной значимостью проблематики, финансированием тематики или доступным методическим арсеналом. На основе такого описания могут быть скорректированы аналитические выводы по итогам эксперимента: действительно ли, например, найденная группа генов связана с патологией, или же представлена наиболее доступными для исследования современными аналитическими методами объектами, или чаще исследуется в силу иных факторов. Специфичный информационный профиль характерен и для различных видов организмов (см. далее).

Значения весовых коэффициентов ожидаемо будут различаться между организмами вследствие различий в степени изученности организмов.

Полученные результаты свидетельствуют о том, что измерение продуктов экспрессии (например, транскриптов или белков), локализованных на одной хромосоме или случайным образом отобранных генов (N > 200), позволит получить достаточную информацию о состоянии человека как системы. Не имеет смысла использовать в качестве дескрипторов большее количество генов, поскольку это не влияет на информационный профиль, но закономерно увеличивает количество «информационного шума», препятствующего анализу данных.

Транскрипты и белки хромосомы 18 человека

В таблице 3 обобщены результаты транскриптомного исследования клеток линии HepG2 и ткани печени человека. Детектированы транскрипты, соответствующие 234 генам хромосомы 18, что составляет 85 % от общего количества белок-кодирующих генов анализируемой хромосомы. Наибольшее количество транскриптов детектировано методом РНК-секвенирования ($РНК_{сек}$) на платформе SOLiD. Разница между количеством транскриптов, найденных в каждом из типов биоматериала, несущественна: в клеточной линии наблюдалось всего на 14 транскриптов меньше, чем в образце ткани печени.

Поскольку для одного и того же образца выполняли измерения методами $РНК_{сек}$ и методами на основе ПЦР, то на этапе обработки данных оценивали взаимосвязи между полученными разными методами количественными значениями. В геноцентричном формате сопоставляли уровень PRKM, полученный методом $РНК_{сек}$ (Illumina, RPKM>0), и уровень того же транскрипта, измеренного в том же образце методом $РНК_{сек}$ (SOLiD, RPKM>0). Корреляция между значениями $lg(_{RPKM})$ составляет $R^2 = 0,88$ в случае анализа клеточной линии HepG2 и $R^2 = 0,82$ при анализе клеток ткани печени. Высокая степень корреляции между результатами, полученными на платформах SOLiD и Illumina, позволила в дальшейшем оценивать уровень экспрессии транскрипта на основе объединения данных двух платформ.

Калибровочное уравнение для перевода значений RPKM в количество копий транскрипта на клетку получали на основе сопоставления результатов измерений методами $РНК_{сек}$ и ПЦР ($ПЦР_{рв}$+$ПЦР_{цк}$). Для построения уравнения использовали все результаты, полученные методом ПЦР (отличные от нуля значения, независимо от C_t), и данные $РНК_{сек}$ хотя бы одной из платформ (в случае наличия показателей RPKM, измеренных двумя платформами, данные усредняли). Коэффициент R^2 между этими показателями равен 0,53, что свидетельствует о хорошей степени соответствия между результатами панорамных и направленных измерений транскриптов в клетках ткани печени (для результатов анализа клеточной линии HepG2 $R^2 = 0,64$).

Таблица 3 – Количество транскриптов хромосомы 18 человека, детектированных в ткани печени и клеточной линии HepG2. Общее количество белок-кодирующих генов на данной хромосоме – 275

Метод	**Всего детектировано**	**Тип биоматериала**	
		HepG2	*Печень*
$РНК_{сек}$ (Illumina, RPKM* > 0)	220	198	190
$РНК_{сек}$ (SOLID, RPKM > 0)	227	198	224
ПЦР ($ПЦР_{рв}$+ $ПЦР_{цк}$, $C_t \leq 30$)	147	143	69
Всего	***234 (85 %)***	***214 (78 %)***	***228 (83 %)***

* RPKM (*Reads per kilobase per million mapped reads*) – программно рассчитываемый уровень экспрессии гена (транскрипта), C_t – пороговый цикл, т. е. количество циклов амплификации, в ходе которых флуоресценция расщепленного зонда превысила фоновые значения; $ПЦР_{рв}$ – ПРЦ в реальном времени (*Real Time quantitative Reverse Transcription, RT-qPCR*); $ПЦР_{цк}$ – цифровая капельная ПЦР (*Next-generation droplet digital PCR, ddPCR*).

Калибровочные уравнения использовали в случаях, если транскрипт не был детектирован методом ПЦР, но присутствовали сведения о его обнаружении методом РНК-секвенирования, а также в случае, если транскрипт детектирован методом РНК-секвенирования на обеих платформах с RPKM > 1, но методом ПЦР обнаружен при C_t > 30, поскольку в этом случае количественная оценка может быть дана с высокой ошибкой. В остальных случаях количество копий транскрипта получали по данным ПЦР-анализа. Для клеточной линии HepG2 получены значения копийности для 255 транскриптов, в клетках ткани печени были получены для 264 транскриптов, из которых 206 – измерены методом ПЦР и 58 – получены на основе калибровочного уравнения по данным РНК-секвенирования.

На рисунке 3 приведено распределение количества детектированных транскриптов по количеству копий в одной клетке. В обоих типах биоматериала (клетки линии HepG2 и ткань печени) сохраняется одинаковый диапазон распределения количества копий транскриптов на клетку – от одной копии на 10^2 – 10^3 клеток до тысяч копий транскрипта в одной клетке. Наиболее представленный транскрипт, найденный в обоих типах биоматериала, соответствует гену, кодирующему белок транстиретин (*TTR*, P02766). Транстиретин наблюдали в количестве $9{,}6\times10^3$ копий транскрипта на клетку HepG2 и $4{,}8\times10^3$ копий транскрипта на клетку печени. Около половины от общего количества детектированных транскриптов в ткани печени представлены в количестве 10–100 молекул на одну клетку. В клеточной линии максимум сдвинут вправо: большее количество транскриптов обнаружено в области 100–1000 копий молекул на одну клетку.

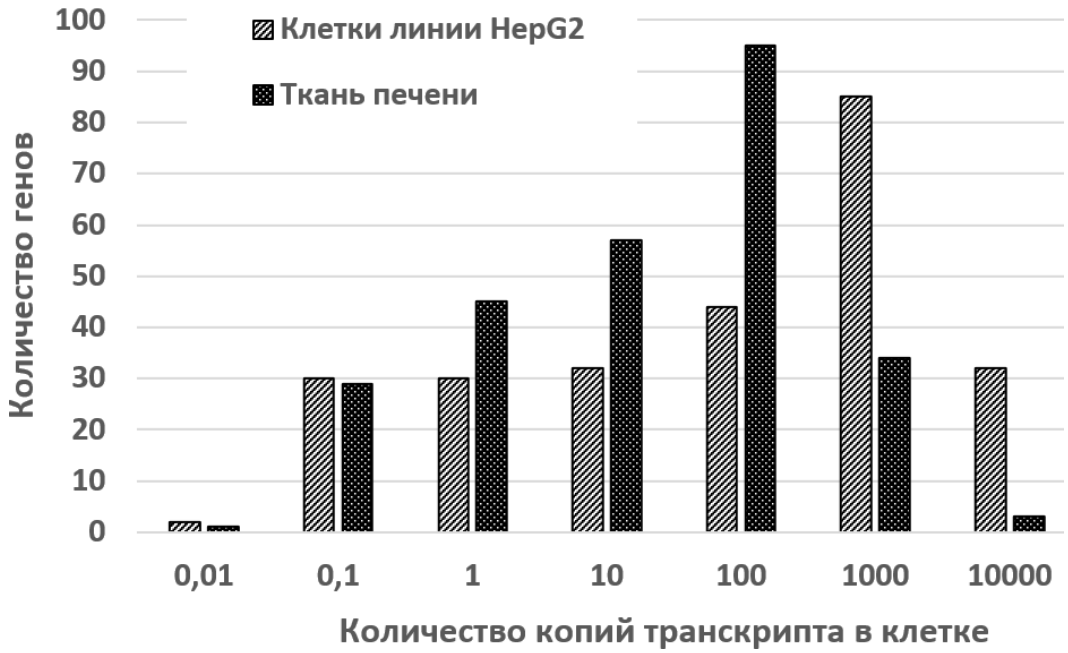

Рисунок 3. Распределение количества транскриптов, кодируемых хромосомой 18 человека, в клетках линии HepG2 (*n*=255) и клетках ткани печени человека (*n*=264, по результатам измерений методами РНК$_{сек}$ и ПЦР, см. описание в тексте)

Также хорошо согласуются данные о количественном уровне экспрессии транскриптов между исследуемыми биологическими материалами (R^2 = 0,66, результаты сопоставления количественных данных об уровне 255 транскриптов, обнаруженных в обоих типах биологического материала). Высокая степень совпадения указывает на сходство транскриптомного состава гепатоцитов и искусственной клеточной линии.

В таблице 4 обобщены данные о масс-спектрометрическом исследовании протеома хромосомы 18 в трех типах биологического материала. Всего было обнаружено 270 белков из 275, кодируемых на выбранной хромосоме. Максимальное количество белков (267) было детектировано наименее точным подходом (*Tier 3*, Carr et al., 2014) – путем направленного масс-спектрометрического метода множественных реакций (ММР) с использованием немеченых пептидных стандартов (ММР$_{нс}$). Среди обнаруженных этим методом белков выделяли группу ММР$_{нс*}$, в которую вошли белки, детектированые с использованием двух протеотипических пептидов на один белок, разница концентраций которых находится в пределах одного порядка. В литературе дискутируется надежность этого метода для детекции белков: он не обладает высокой специфичностью, поскольку вследствие интерференции может давать ложноположительные детекции (Ludwig et al., 2012). Метод может регистрировать как присутствующий в образце протеотипический пептид, так и сигнал от химически сходных, но не идентичных искомому пептиду молекул.

Использование в качестве стандарта пептида с изотопной меткой (метод ММР$_{мс}$) соответствует более высоким требованиям к качеству масс-спектрометрических данных согласно (*Tier 2*, Carr et al., 2014). В этом случае обнаружение и количественное измерение содержания белка в образце проводится на основе сопоставления формы

и положения на хроматограмме масс-спектрометрических сигналов нативного пептида в образце и идентичного по последовательности аминокислотных остатков синтетического пептида с изотопной меткой, добавленного в образец в качестве стандарта.

Таблица 4 – Результаты протеомного анализа клеток линии HepG2, ткани печени и плазмы крови человека (на примере белков, кодируемых 275 генами хромосомы 18)

		Тип биоматериала		
Метод[1]	**ВСЕГО**	*Плазма крови*	*Печень*	*Клеточная линия HepG2*
ММР$_{мс}$	128	84	73	66
МС/МС	65	36	8	32
ММР$_{нс*}$	205	155	166	158
ММР$_{нс}$	267	264	261	265
ВСЕГО	**270 (98 %)**	268 (97 %)	264 (96 %)	266 (97 %)

[1]Условные обозначения названий аналитических методов: ММР$_{нс}$ – направленный анализ масс-спектрометрическим методом мониторинга множественных реакций (ММР) с использованием немеченых пептидных стандартов (аналогично – ММР$_{мс}$ – с применением минимум одного меченого пептидного стандарта); ММР$_{нс*}$ – группа белков, детектированых с использованием двух протеотипических пептидов на один белок, соотношение концентраций которых находилось в пределах одного порядка величины; МС/МС – панорамный анализ масс-спектрометрическим методом (*Shotgun LC-MS/MS*).

Методом МС/МС детектировано 36 белков в плазме крови человека, 32 белка в клеточной линии HepG2 и только 8 белков найдено в клетках ткани печени. Результаты позволяют предположить, что в клеточной линии увеличена экспрессия части белков, нехарактерных для здоровой ткани печени.

На рисунке 4 представлена гистограмма распределения копийности белков в трех типах биологического материала, измеренных методом ММР. Нисходящая часть гистограмм отражает наличие высоко- и среднекопийных белков, в то время как восходящая часть может объясняться либо низкой представленностью некоторых видов белков, либо недостаточной чувствительностью аналитического метода, по причине чего белок не может быть обнаружен.

Для корректного сопоставления количественных транскриптомных и протеомных данных использовали одни и те же единицы измерения – количество копий транскрипта (белка), определенное в одной клетке анализируемого образца биоматериала с тем же уровнем аналитической чувствительности. Количественный метод ММР позволяет обнаружить белок, если его количество превышает 10 копий на одну клетку. Для сравнения, транскрипт может быть детектирован на уровне 1 копии на 100 клеток (см. рисунки 3 и 4). Из массива результатов транскриптомного анализа отбирали результаты, полученные при чувствительности около 10 копий на клетку, что позволило использовать одинаковый «масштаб» при сравнении с протеомными данными.

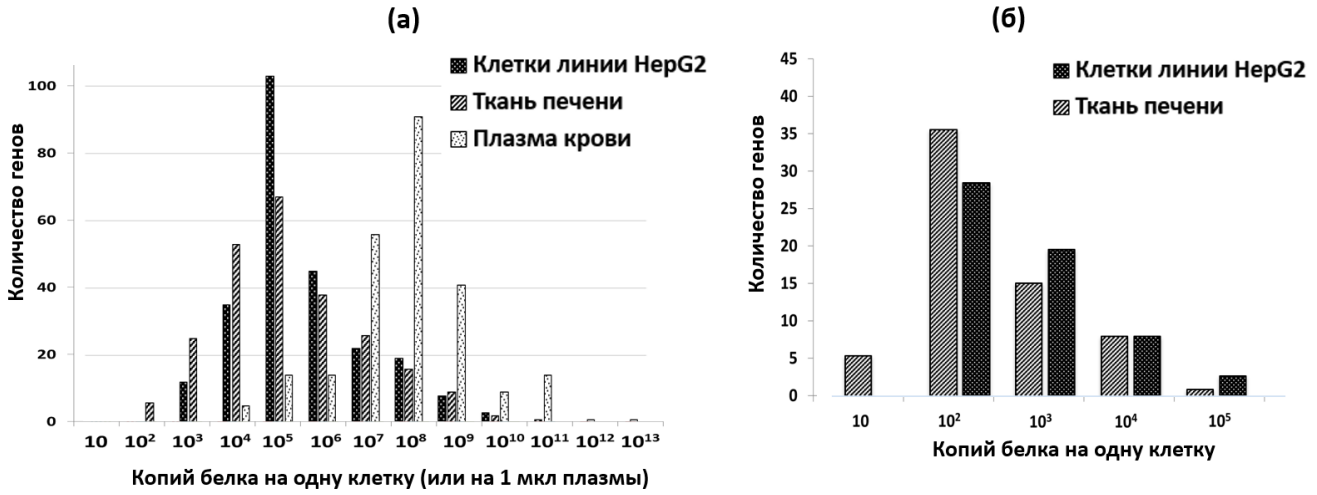

Рисунок 4. Распределение копийности белков, кодируемых хромосомой 18 человека, обнаруженных в клетках ткани печени и клеточной линии HepG2, детектированных путем направленного анализа масс-спектрометрическим методом **(а)** с использованием немеченых пептидных стандартов ($MMP_{нс}$) и **(б)** предусматривающим использование в качестве стандарта пептида с изотопной меткой ($MMP_{мс}$)

На рисунке 5 представлено распределение транскриптов и белков в зависимости от типа биоматериала, в котором они были обнаружены. Данные приведены для белков, найденных методом $MMP_{мс}$, поскольку этот метод отвечает требованиям, предъявляемым к качеству данных при количественном протеомном анализе (Carr et al., 2014). Видно, что количество детектированных белков почти в два раза меньше, чем количество транскриптов: всего в обоих типах биоматериала зарегистрировано наличие 81 белка (данные, полученные методом $MMP_{мс}$), в то время как количество экспрессируемых на уровне транскриптома генов составляет 173 (объединение результатов анализа транскриптомов двух типов биоматериала). Для 46 генов были детектированы и транскрипты, и белки, причем в обоих исследуемых образцах.

Для каждого типа анализируемого биоматериала строили зависимость между количеством копий транскриптов и детектированных в этом же образце белков. Получено, что соответствия между уровнем экспрессии транскриптов и белков не наблюдается – коэффициент R^2 ≈0,06–0,1; значения сопоставимы для обоих типов биологического материала. Возможное объяснение несоответствия протеомных и транскриптомных данных может быть связано с тем, что исследователь на молекулярном уровне работает с «фотоснимками», сделанными в определенных точках эксперимента, при этом, молекулы РНК имеют одну шкалу временных координат, а белка – другую, поскольку различается время существования этих молекул (Schwanhäusser et al., 2011).

Для сопоставления результатов направленного и панорамного масс-спектрометрического анализа были использованы показатели чувствительности, специфичности и точности каждого метода, по аналогии с оценкой диагностических тестов (*Guidance for Industry and FDA, 2007*). При данном анализе транскриптом использовали в качестве «стандарта». Перечень детектированных транскриптов

сопоставляли с перечнями найденных разными методами протеомного анализа белков, полученных для того же образца биоматериала. Результаты представлены в таблице 5.

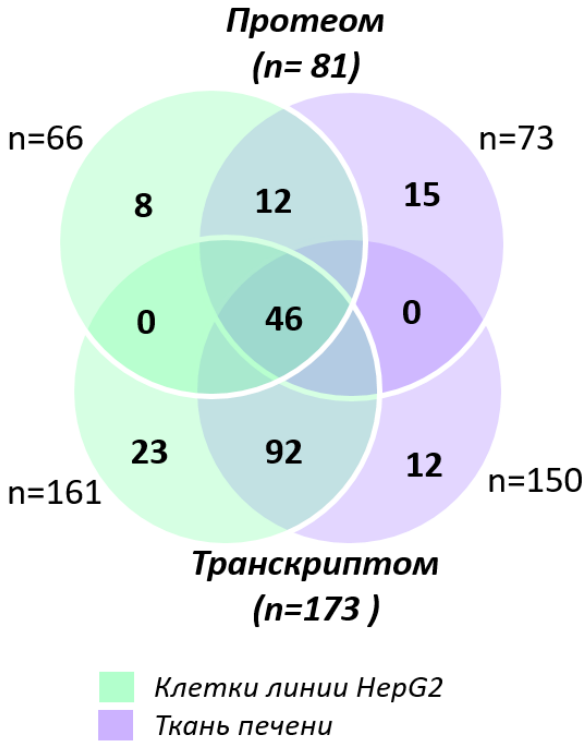

Рисунок 5. Сопоставление количества транскриптов и белков, кодируемых генами хромосомы 18 человека, измеренных в образцах клеток ткани печени человека и клеточной линии HepG2. Количественный анализ протеома выполняли направленным анализом масс-спектрометрическим методом мониторинга множественных реакций (MMP), предусматривающим использование в качестве стандарта пептида с изотопной меткой ($ММР_{мс}$), транскриптома – методом ПЦР ($ПЦР_{рв}$+$ПЦР_{цк}$) и $РНК_{сек}$ (платформы SOLiD, Illumina; для получения количества копий транскрипта на клетку использовали калибровочные уравнения, см. описание в тексте)

Наиболее чувствительным из рассматриваемых подходов является метод направленного детектирования эндогенных пептидов с использованием немеченых пептидных стандартов ($ММР_{нс}$). Наибольшая специфичность наблюдается при идентификации белков панорамным МС/МС методом. Его специфичность сопоставима с результатами направленного анализа при добавлении в пробу изотопно-меченых синтетических пептидных стандартов ($ММР_{мс}$, см. таблицу 5).

Метод $ММР_{нс}$ позволяет получить спектры, которые могут быть сопоставлены пептидам почти всех исследуемых белков, однако вследствие наличия интерференции (Gallien et al., 2012) метод обладает крайне низкой специфичностью. Специфичность повышается при использовании для детекции белка двух протеотипических пептидов с разницей концентраций в пределах одного порядка ($ММР_{нс*}$). В этом случае второй пептид служит дополнительным контролем, поскольку при существенной разнице концентраций между пептидами можно предположить, что детектированы сигналы

разных белков (вследствие интерференции), а концентрация пептида должна быть эквимолярна концентрации белка.

Максимальная точность наблюдается при сопоставлении объединенного перечня белков, обнаруженных методами $ММР_{мс}$, и белков из группы $ММР_{нс*}$ и результатов транскриптомного профилирования этого же образца.

Таблица 5 – Показатели чувствительности (*Sn),* специфичности (*Sp*) и точности (*Ac*) при сопоставлении методов протеомного анализа на примере исследования клеток линии HepG2 и ткани печени человека. В качестве референтной выборки использовали результаты транскриптомного профилирования того же образца биоматериала

Тип биоматериала	**Метод[1]**	**Чувствительность Sn, %**	**Специфичность Sp, %**	**Точность Ac, %**
Клетки линии HepG2 (данные для ткани печени указаны в скобках)	$ММР_{мс}$	34 (40)	89 (88)	57 (63)
	$ММР_{нс*}$	58 (60)	44 (39)	52 (50)
	$ММР_{мс}$ + $ММР_{нс*}$	73 (75)	39 (36)	59 (56)
	$ММР_{нс}$	97 (96)	4 (5)	54 (53)
	МС/МС	10 (8)	97 (99)	20 (18)

[1]Условные обозначения названий аналитических методов: $ММР_{нс}$ – направленный анализ масс-спектрометрическим методом мониторинга множественных реакций (ММР) с использованием немеченых пептидных стандартов (аналогично – $ММР_{мс}$ – с применением минимум одного меченого пептидного стандарта); $ММР_{нс*}$ – группа белков, детектированных с использованием двух протеотипических пептидов на один белок, разница концентраций которых находится в пределах одного порядка; МС/МС – панорамный анализ масс-спектрометрическим методом (*Shotgun LC-MS/MS*).

Многообразие протеоформ человека

Использование метода $РНК_{сек}$ позволяет не только оценить уровень экспрессии каждого гена в образце, но и получить информацию о наличии модифицированных участков в последовательности мРНК, что необходимо при исследовании модифицированных вариантов белков – *протеоформ*. Термином *протеоформы* в контексте данной работы обозначены формы белков, кодируемые одним геном, но различающиеся вследствие различий первичной структуры белка и/или наличия посттрансляционных модификаций.

Согласно предлагаемой в работе концепции, экспериментальное определение количества детектированных протеоформ (*ширины* протеома) зависит от чувствительности и специфичности аналитического метода. Вследствие технических ограничений экспериментальных методов в протеомике оценка многообразия белков без использования методов биоинформатики и данных информационных ресурсов невозможна.

Для предсказания многообразия протеоформ, кодируемых хромосомой 18, использовали результаты, полученные при исследовании транскриптома методом РНК-секвенирования. Анализировали многообразие протеоформ для клеток ткани печени и клеток линии HepG2. Подсчитывали частоту встречаемости событий – сплайс-вариантов и реализованных на уровне транскриптома однонуклеотидных замен на основе результатов РНК-секвенирования. Полученные частоты сопоставляли с результатами анализа информации, накопленной в протеомной базе знаний NeXtProt.

Полученные результаты представлены в таблице 6. Частоты встречаемости сплайс-вариантов (АС) и последовательностей с реализованными на уровне транскриптома несинонимичными однонуклеотидными полиморфизмами (ОП) указаны в расчете на один ген. Приведены данные по всем хромосомам и отдельно для хромосомы 18 человека. По данным РНК-секвенирования, среднее количество обнаруженных сплайс-вариантов в ткани печени составило 1,3 на ген при расчете на весь геном человека, или 1,4 – для одного гена хромосомы 18. Эти данные хорошо соотносятся с данными NeXtProt, согласно которым количество сплайс-вариантов в среднем для генома человека составляет 2,1.

Частоты встречаемости последовательностей с заменами как в среднем по геному, так и для белков, кодируемых на хромосоме 18, существенно различаются между результатами РНК-секвенирования и данными NeXtProt. Транскрипты с заменами в последовательности обнаружены в исследуемом образце ткани печени в среднем с частотой 1,4 варианта на один белок-кодирующий ген. Согласно данным NeXtProt, в среднем следует ожидать до 25 вариантов последовательностей с заменами, кодируемых одним геном (обобщенные сведения по всем типам биологического материала).

Сведения о вариабельности последовательностей белков, представленные в базах данных, являются результатом обработки тысяч опубликованных научных работ и экспериментально полученных наборов данных. Это исследования образцов от различных индивидуумов, разных органов и тканей. Протеомные базы знаний объединяют информацию о результатах транскриптомных и протеомных исследований множества людей, предоставляя интегральную оценку вариабельности белков на основе информации о популяционной совокупности. Полученные на основе обработки этих данных оценки количества протеоформ характеризуют возможное количество белков, которое может быть во всей человеческой популяции, т. е. *популяционный протеом*. В противоположность этому экспериментальные данные РНК-секвенирования конкретного образца биоматериала (в нашем случае – ткани печени человека и клеток линии HepG2) указывают на молекулярные события, характеризующие отдельный образец или ткань, т. е. индивидуальный протеом (см. таблицу 6).

Таблица 6 – Частоты встречаемости сплайс-вариантов и транскриптов, содержащих реализованные несинонимичные однонуклеотидные полиморфизмы, вычисленные по результатам РНК-секвенирования образца ткани печени человека и аннотациям протеомной базы NeXtProt (v.12_2015)

Источник данных	**РНК-секвенирование (индивидуальные данные)**		**Данные NeXtProt (популяционные данные)**	
Показатель, усредненное кол-во на один ген	*Все хромосомы (n=20 216)*	*Хромосома 18 (n=275)*	*Все хромосомы (n=20 216)*	*Хромосома 18 (n=275)*
Сплайс-варианты	1,3	1,4	2,1 ± 0,2	1,9
Транскрипты с полиморфизмами	1,4	1,3	**24,4 ± 1,8**	**24,6**

Для оценки количества протеоформ, кодируемых хромосомой 18 человека, использовали значения частот, рассчитанные на основе сведений базы данных NeXtProt (Gaudet et al., 2015). Согласно версии 01_2016, значения составили для хромосомы 18: N = 275; *ASav = 2,0; SAPav = 24,6; PTMav = 5,8; ASd = 0,59; SAPd = 0,97 и PTMd = 0,71*, где общее количество генов *N*, средняя частота встречаемости сплайс-варианта (*ASav*), или полиморфного варианта (*SAPav*) или модификации (*PTMav*), доля генов, кодирующих белки со сплайс-вариантами (*ASd*), одноаминокислотными заменами (*SAPd*) или посттрансляционными модификациями (*PTMd*). Полученные частоты подставляли в информационные модели, отражающие гипотетические варианты комбинаций возможных молекулярных событий, приводящих к возникновению протеоформ.

Используя для расчетов модель 1, базирующуюся на предположении о том, что одноаминокислотные замены или посттрансляционные модификации возникают только в условно канонических вариантах белков и не затрагивают сплайс-варианты, получено, что хромосома 18 может кодировать от 8,3 до 9,2 тыс. различных протеоформ.

Если применять расчетную модель 2, учитывающую возможность возникновения полиморфизмов также и в последовательностях, образованных в результате альтернативного сплайсинга (формула (6)), то оценочное количество протеоформ увеличивается почти в два раза и составляет до 18,2 тыс. различных видов белков, кодируемых на хромосоме 18 человека.

Модель 3 основана на предположении о том, что все три типа молекулярных событий происходят независимо. Используя формулу (7), получили оценочное количество протеоформ, кодируемых хромосомой 18, на уровне 103,9 тыс. Таким образом, среднее количество протеоформ, кодируемых одним геном выбранной хромосомы, в зависимости от рассматриваемых допущений, колеблется на порядок от 30-66 до 376.

Масштабировав предложенный вариант теоретической оценки *ширины* протеома на весь геном человека, получили следующие результаты. Анализ NeXtProt показал наличие 21 921 варианта альтернативного сплайсинга для белков человека в 10 519 белок-кодирующих генах (в среднем 2,1 ± 0,2 сплайс-варианта на один ген). Наибольшее количество модифицированных форм было обусловлено возникновением ОАП, появившихся в результате мутаций (в среднем 22,2 ± 3,8 варианта на ген, всего аннотировано 434 695 модификаций в 18 999 белок-кодирующих генах, не считая раковых модификаций, полученных из базы данных о генетических изменениях в раковых клетках – COSMIC (Forbes et al., 2015)). Посттрансляционные модификации возникают в среднем с частотой 6,4 ± 1,2 на ген (94 042 ПТМ аннотировано для 14 009 белок-кодирующих генов). Принимая N = 20057, в организме человека теоретически можно предполагать наличие от 0,55 (формула 5), или 1,18 (формула 6), до 7,14 (формула 7) миллионов протеоформ.

В исследуемом образце ткани печени человека можно предполагать наличие от 130 до 400 тыс. разных протеоформ (в зависимости от расчетной формулы). Общее количество предсказанных протеоформ в клеточной линии HepG2 несколько большее, чем для клеток печени, и составляет от 130 до 500 тыс. соответственно. Эти значения соотносятся с ранее опубликованными экспериментальными результатами оценки количества протеоформ (Naryzhny et al., 2014), полученными методом 2DE с красителями различной чувствительности. Таким образом, многообразие протеоформ конкретного индивидуума, рассчитанное на основе транскриптома определенного биологического материала, более чем в 10 раз меньше по масштабам, чем рассчитанный на основе базы данных NeXtProt популяционный протеом.

На рисунке 6 приведены графики зависимости количества найденных транскриптов, содержащих однонуклеотидные замены (ОП), от количества проанализированных образцов. Исходя из монотонного характера кривых, можно предполагать, что исчерпывающая инвентаризация всех транскриптов, содержащих однонуклеотидные замены, может оказаться принципиально недостижимой. Каждый человек будет нести неповторяемый набор мутаций, добавляя новые варианты транскриптов и, как следствие, протеоформ.

Для сравнения, на рисунке 7 представлены графики зависимости количества найденных сплайс-вариантов транскриптов (АС) от количества анализируемых образцов (анализировали те же данные *Sequence Read Archive*).

Полученные результаты позволили сопоставить динамику накопления уникальных вариантов транскриптов, содержащих однонуклеотидные замены или образованных вследствие альтернативного сплайсинга. Как видно из рисунка 7, половину возможных вариантов транскриптов (принимая за 100 % общее количество обнаруженных в анализируемых транскриптомах сплайс-вариантов) можно обнаружить уже в 10 случайно выбранных образцах: 90-процентное покрытие достигается при 20–30 образцах, а 99-процентное – при 50–60 образцах. Для сравнения, 50 % от общего количества

детектированных транскриптов с реализованными на уровне транскриптома заменами в геноме (ОП) достигается на 25 образцах (из 60), а 90-процентный охват – на 50 образцах.

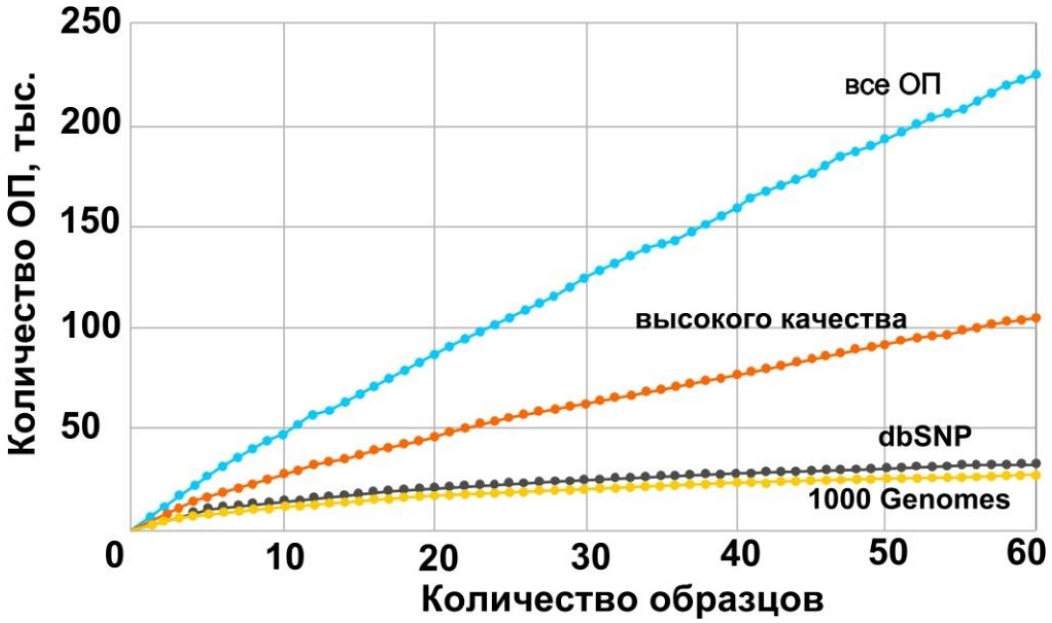

Рисунок 6. Количество транскрибируемых вариантов однонуклеотидных полиморфизмов в геноме (ОП) на основе анализа 60 транскриптомов ткани печени человека (анализ транскриптомов, доступных в *Sequence Read Archive* (Lipman et al., 2011; Kodama et al., 2012)

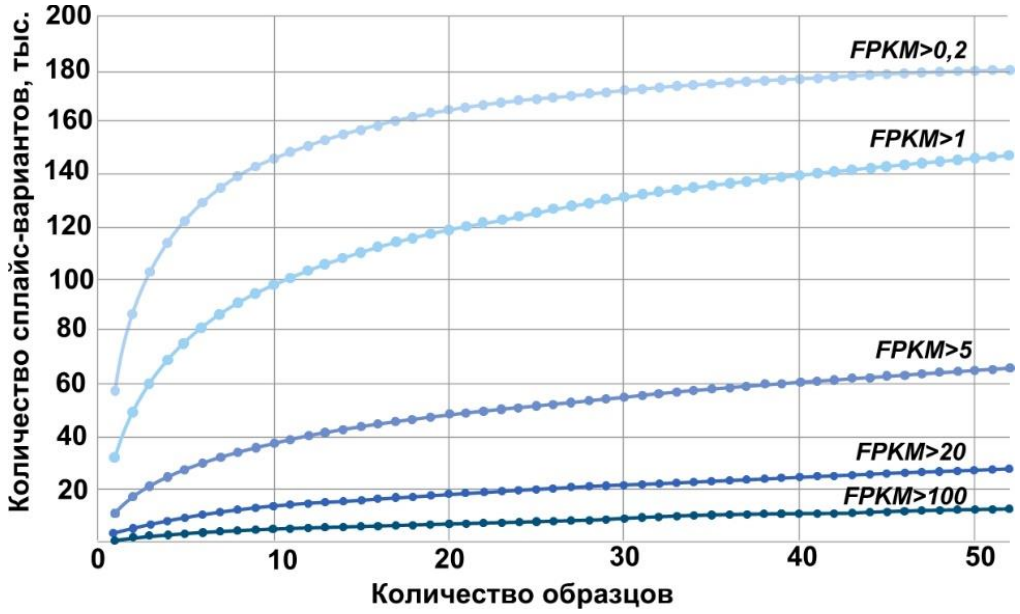

Рисунок 7. Количество сплайс-вариантов транскриптов, выявленное при анализе ткани печени человека (данные базы *Sequence Read Archive*). Графики построены для разных уровней отсечения по FPKM (программно рассчитываемый уровень экспрессии гена)

Многообразие протеоформ бактерий, растений и животных

Степень изученности геномов модельных объектов оценивали пропорционально количеству генов в геноме, для которых есть проверенные экспертом аннотации в разделе UniProtKB/SwissProt (версия 08_2016). Чем большее количество генов имеет экспертные аннотации (описания) в UniProtKB/SwissProt, тем более изученным является организм.

Более чем для 90 % организмов доля аннотированных экспертом генов (записей в разделе UniProtKB/SwissProt) составляет 10–20 %. Только 10 видов (большинство из которых – бактериальные) имеют долю аннотированных генов в геноме более 75 %. Мы сопоставили данные о количестве проверенных экспертом аннотаций для генов наиболее популярных модельных объектов (см. таблицу 7).

Данные таблицы 7 свидетельствуют о том, что геномы более примитивных видов (например, бактерий или грибов) охарактеризованы полнее, чем геномы других организмов. Если не рассматривать простые организмы, то единственный вид, для которого аннотирован полный геном (около 100 % генов аннотированы экспертами и присутствуют в разделе UniProtKB/SwissProt), – это человек. Исследование других видов организмов так или иначе направлено на получение информации о человеке, о его молекулярном функционировании, развитии патологических состояний.

От 50 до 75 % генома аннотировано для видов, представленных в пп. 10–13 таблицы 7. Наиболее полно описаны продукты экспрессии мышиного генома (для 75 % генов присутствуют экспертные аннотации в UniProtKB/SwissProt). Среди растений наиболее полно охарактеризован геном Арабидопсиса (*Arabidopsis thaliana)* – для него аннотировано около половины генов генома.

Среди организмов, для которых аннотировано менее 40 % генов, преобладают животные. Среди них такие известные модельные объекты, как *Drosophila melanogaster*, *Caenorhabditis elegans* и *Danio rerio*. Доля аннотированных генов для них составляет 24, 18 и 12 %, соответственно. Проект «Геном Зебрафиш» (Zebrafish Genome Project) был начат 15 лет назад (Howe et al., 2013). Из 26 тыс. белок-кодирующих генов, предсказанных в геноме Danio rerio (Collins et al., 2012), экспертная аннотация в UniProtKB/SwissProt есть для 3 тыс. генов; несмотря на то, что около 70 % генов человека имеют ортологи в геноме Zebrafish (Howe et al., 2013). Наиболее часто этот объект используют для моделирования генетических заболеваний человека, описывая взаимосвязи между мутацией и фенотипическими проявлениями. Интерес к исследованию молекулярной микрогетерогенности на белковом уровне только в самом начале, о чем свидетельствует сравнительно небольшое количество записей о белках Zebrafish в протеомных базах. Наименьшее количество аннотированных генов – всего 4 % от общего количества генов в геноме – у шимпанзе, примата семейства гоминид. Эта наиболее информативная модель с точки зрения исследования заболеваний человека, тем не менее, остается невостребованной.

Таблица 7 – Количество аннотированных экспертами базы UniProt генов (представленных в разделе UniProtKB/SwissProt) для человека (*Homo sapiens*) и наиболее популярных модельных объектов

	Вид	**Кол-во генов** (Ensembl)	**Аннотации** (UniProtKB/ SwissProt)	**Доля аннотиро-ванных генов,** %	**Царство**
1	*Escherichia coli*	4140	5970	144*	Бактерии
2	*Mycoplasma genitalium*	476	483	101*	Бактерии
3	*Methanocaldococcus jannas.*	1770	1787	101*	Археи
4	*Saccharomyces cerevisiae*	6692	6721	100	Грибы
5	*Buchnera aphidicola subsp.*	507	507	100	Бактерии
6	*Haemophilus influenzae*	1709	1707	100	Бактерии
7	*Mycoplasma pneumoniae*	688	687	100	Бактерии
8	*Schizosaccharomyces pombe*	5145	5120	100	Грибы
9	***Homo sapiens***	**20441**	**20198**	**99**	**Животные**
10	*Mus musculus*	22642	16813	74	Животные
11	*Rickettsia prowazekii*	834	586	70	Бактерии
12	*Arabidopsis thaliana*	27416	14754	54	Растения
13	*Aquifex aeolicus*	1553	784	50	Бактерии
14	*Rattus norvegicus*	22263	7963	36	Животные
15	*Drosophila melanogaster*	13918	3328	24	Животные
16	*Caenorhabditis elegans*	20362	3731	**18**	Животные
17	*Danio rerio*	25759	2967	**12**	Животные
18	*Oryza sativa subsp. Japonica*	35679	3517	10	Растения
19	*Pan troglodytes*	18759	693	4	Животные

* Виды, аннотированные избыточно в сравнении с Ensembl.

На рисунке 8 отображены графики изменения количества автоматически сгенерированных (TrEMBL) и прошедших экспертную проверку (UniProtKB/SwissProt) записей для видов модельных организмов, указанных в таблице 7. Количество проверенных экспертами записей (раздел UniProtKB/SwissProt, см. рисунок 8 (а)) отражает количество белок-кодирующих генов, поскольку информация о вариабельности белковых последовательностей и наличии посттрансляционных модификаций депонируется в геноцентричном формате. Количество записей раздела TrEMBL (см. рисунок 8 (б)) позволяет оценить многообразие видов белков. Сопоставляя тенденции накопления данных в этих разделах (см. рисунок 8 (а), (б)), можно заключить, что исследование многообразия видов белков пока в самом начале. Количество записей в TrEMBL имеет тенденцию к экспоненциальному росту, тогда как записи в разделе UniProtKB/SwissProt уже вышли на плато. Это означает, что внимание экспертов в большей степени направлено на уточнение уже сформированных аннотаций, чем на формирование новых записей.

Закономерности, выявленные для всей базы данных UniProt, подтверждаются при анализе трендов по изменению количества аннотированных генов модельных объектов (рисунок 8 (в), (г)). Количество автоматически сгенерированных записей увеличилось в период 2011–2015 гг. на 70 % для человека и 20 % для мыши, в то время как число

проверенных экспертами геномных аннотаций осталось на прежнем уровне (рисунок 8 (г)). С одной стороны, это свидетельствует о неравномерности исследовательского интереса к различным видам организмов, а с другой – о существенном отставании экспертной обработки данных от скорости поступления экспериментальных данных, получаемых в ходе высокопроизводительных экспериментов.

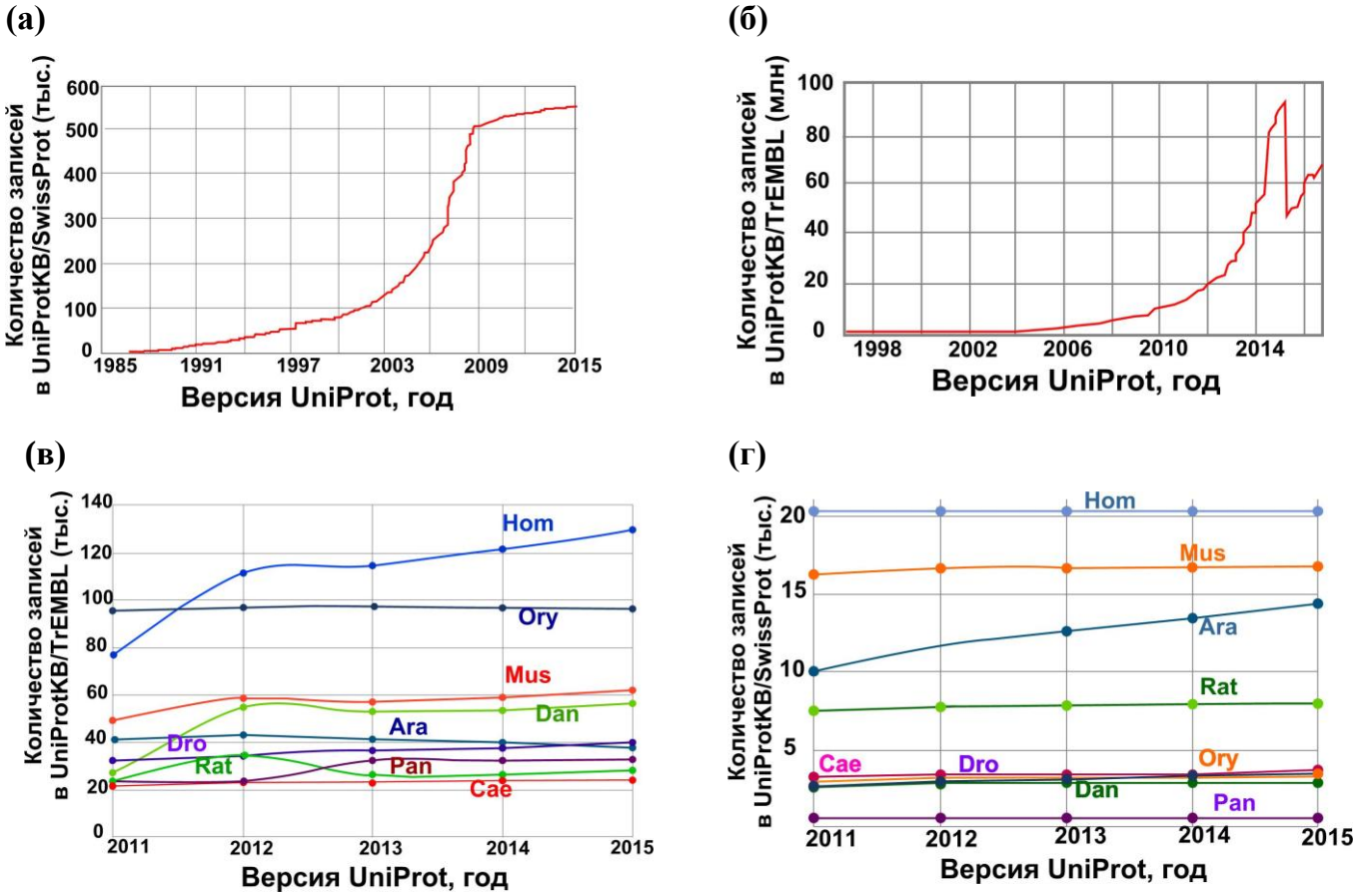

Рисунок 8. Динамика изменения количества записей, размещенных в разделах базы данных UniProt, в результате **(а)** экспертной проверки записей (UniProtKB/SwissProt), все виды организмов; **(б)** автоматической аннотации геномных последовательностей (TrEMBL), все виды организмов; **(в)** человек и модельные организмы (TrEMBL); **(г)** человек и модельные организмы (UniProtKB/SwissProt). Обозначения: **Hom** – *Homo sapiens*, **Mus –** *Mus musculus*, **Ara –** *Arabidopsis thaliana*, **Rat –** *Rattus norvegicus*, **Dro** – *Drosophila melanogaster*, **Cae** – *Caenorhabditis elegans*, **Dan –** *Danio rerio*, **Ory** – *Oryza sativa subsp. Japonica*, **Pan** – *Pan troglodytes*

Данные о многообразии протеоформ (*ширине* протеомов) выбранных модельных объектов суммированы в таблице 8. Разнообразие бактериального протеома обусловлено в основном наличием ОАП и ПТМ. Среди грибов наиболее изучены виды *Saccharomyces cerevisiae* и *Schizosaccharomyces pombe*. *Saccharomyces cerevisiae* по количеству протеоформ, кодируемых одним геном, наиболее близок к животным: предполагается,

что один ген *Saccharomyces cerevisiae* может кодировать от 2 до 14 вариантов белков, что сопоставимо с показателями для *Mus musculus* (см. таблицу 8).

Таблица 8 – Количество протеоформ (*ширина* протеома) некоторых видов модельных организмов согласно трем расчетным информационным моделям (ИМ1-ИМ3). Приведено как общее количество протеоформ для модельного объекта, так и среднее количество протеоформ, кодируемых одним белок-кодирующим геном (БКГ)

Царство	**Обозначение***	**Количество генов** *(Ensembl)*	**Количество протеоформ всего (на 1 БКГ)**		
			Информационная модель, ИМ (формула расчета)		
			ИМ1 (5)	ИМ2 (6)	ИМ (7)
Археи	Met	1 770	1 814 (1,0)	1 814 (1,0)	1 814 (1,0)
Бактерии	Esc	4 140	5 055 (1,2)	5 104 (1,0)	27 042 (7,0)
Бактерии	Myg	476	485 (1,0)	485 (1,0)	485 (1,0)
Бактерии	Buc	507	530 (1,0)	530 (1,0)	530 (1,0)
Бактерии	Hae	1 709	1 963 (1,0)	1 977 (1,0)	12 233 (7,0)
Бактерии	Myp	688	713 (1,0)	713 (1,0)	3 293 (5,0)
Бактерии	Ric	834	874 (1,0)	874 (1,0)	4 627 (6,0)
Бактерии	Aqu	1 553	1 596 (1,0)	1 596 (1,0)	4 702 (3,0)
Грибы	Sac	6 692	14 315 (2,1)	15 156 (2,0)	94 662 (14,0)
Грибы	Sch	5 145	7 743 (1,5)	7 776 (2,0)	36 845 (7,0)
Животные	*Hom*	*20441*	*160 676 (7,9)*	*401 527 (20,0)*	*1 647 510 (81,0)*
Животные	Mus	22 642	71 838 (3,2)	138 402 (6,0)	616 832 (27,0)
Животные	Rat	22 263	46 378 (2,1)	59 125 (3,0)	356 926 (16,0)
Животные	Dro	13 918	20 134 (1,4)	34 436 (2,0)	348 796 (25,0)
Животные	Cae	20 362	22 062 (1,1)	28 449 (1,0)	169 926 (8,0)
Животные	Dan	25 759	26 461 (1,0)	27 537 (1,0)	176 416 (7,0)
Животные	Pan	18 759	20 351 (1,1)	20 541 (1,0)	270 415 (14,0)
Растения	Ara	27 416	34 628 (1,3)	51 325 (2,0)	408 478 (15,0)
Растения	Ory	35 679	36 285 (1,0)	37 523 (1,0)	212 153 (6,0)

***Esc** – Escherichia coli, **Myg** – Mycoplasma genitalium, **Met** – Methanocaldococcus jannas., **Sac –** Saccharomyces cerevisiae, **Buc** – Buchnera aphidicola subsp., **Hae** – Haemophilus influenza, **Myp** – Mycoplasma pneumoniae, **Sch** – Schizosaccharomyces pombe, **Hom** – Homo sapiens, **Mus –** Mus musculus, **Ric –** Rickettsia prowazekii, **Ara** – Arabidopsis thaliana, **Aqu –** Aquifex aeolicus, **Rat –** Rattus norvegicus, **Dro** – Drosophila melanogaster, **Cae** – Caenorhabditis elegans, **Dan –** Danio rerio, **Ory** – Oryza sativa subsp. Japonica, **Pan –** Pan troglodytes

Наибольшее расчетное количество протеоформ получено для млекопитающих. У человека эта величина составляет от 160 тыс. до 1,6 млн различных форм в зависимости от принимаемой информационной модели (таблица 8). Геном мыши, содержащий около 22 тыс. генов, потенциально кодирует более 600 тыс. различных белков. Немногим меньше – около 400 тыс. протеоформ – у растений (на примере *A. thaliana*, геном которого содержит 27 тыс. кодирующих генов). Наименьшее разнообразие кодируемых белков у организмов с небольшим количеством белок-кодирующих генов – дрожжей и бактерий – около 95 и 27 тыс., соответственно. Для микоплазмы (у этого организма минимальный геном – 476 белок-кодирующих генов) расчетное количество протеоформ почти

совпадает с количеством генов и составляет около 500. Согласно данным UniProtKB/SwissProt, для этого организма показано только наличие ПТМ для белков, кодируемых девятью генами из 476.

Для бактерий, животных и грибов наблюдается соответствие между размером генома и количеством протеоформ. В целом прослеживается зависимость: чем меньше генов, тем меньше вариантов протеоформ (нормировано на размер генома). Высокая корреляция является следствием того, что исследуются либо сравнительно простые организмы (например, бактерии), либо сложные, но при этом более доступные в лабораториях (например, лабораторная мышь). Для растений эта тенденция не наблюдается: несмотря на большее (в сравнении с рассматриваемыми видами животных) количество генов в геноме, количество потенциальных протеоформ относительно невелико.

Основной вклад в многообразие белков вносит не количество генов в геноме, а молекулярные процессы, увеличивающие количество вариантов белков, кодируемых одним геном: альтернативный сплайсинг, мутации, посттрансляционные модификации, геномное редактирование, редактирование мРНК и др. Согласно результатам, полученным в данной работе, также можно высказать предположение о значимости социальной составляющей, отражающейся в количестве опубликованных результатов исследований и доли аннотированных генов в геноме. Ранее этот феномен был замечен при работе с большими массивами научно-технической информации (Yao et al., 2015).

Для оценки зависимости расчетных значений количества протеоформ от степени интереса к объекту исследования сопоставляли результаты предсказанного количества протеоформ (*ширины* протеома) и доли аннотированных для организма генов. Под аннотированными генами в данной ситуации подразумевали количество генов, для которых есть хотя бы одна запись в разделе UniProtKB/SwissProt. Анализ проводили для восьми видов (растения – *Oryza sativa subsp. Japonica, Arabidopsis thaliana*; животные – *Danio rerio, Mus musculus, Rattus norvegicus, Caenorhabditis elegans, Pan troglodytes* и *Drosophila melanogaster*). Для этих видов количество генов в геноме варьируется от 13 тыс. (*Drosophila melanogaster*) до 35 тыс. (рис, *Oryza sativa subsp. Japonica*). Степень изученности – от 4 % (шимпанзе) до 74 % (мышь). Модельные объекты для этого эксперимента подбирали таким образом, чтобы количество генов в геноме было сопоставимо, но существенно различалась степень изученности.

Из трех предложенных информационных моделей комбинации молекулярных событий (альтернативного сплайсинга, возникновения ОАП или ПТМ) при образовании протеоформ наиболее сопоставима со степенью изученности организма модель, при которой каждое из трех рассматриваемых молекулярных событий рассматривается как независимое (информационная модель 3). Коэффициент корреляции между долей аннотированных генов в геномах организмов и количеством протеоформ (согласно формуле 7) составляет $R^2 = 0{,}82$. Это означает, что чем большее количество генов прошло экспертную аннотацию, тем большее количество протеоформ можно предполагать в

составе протеома организма (для организмов, имеющих сходное количество генов в геноме). Таким образом, объективная оценка *ширины* протеома возможна только в том случае, если процесс аннотирования завершился для большей части генов в геноме организма. Для сравнения, расчетный коэффициент корреляции между степенью изученности генома организма и количеством протеоформ составляет R^2 =0,79 (модель 2, уравнение 6) и R^2=0,66 (модель 1, уравнение 5). Для модели 3 наибольший вклад в оценочные значения вносит доля аннотированных в информационных ресурсах генов, для которых указано наличие сплайс-вариантов (85 % влияния параметра на итоговую оценку количества протеоформ), посттрансляционных модификаций в кодируемых геном белках (80 %) и одноаминокислотных полиморфизмов (69 %).

ЗАКЛЮЧЕНИЕ

Вследствие динамичной природы, отличающей протеом от генома, протеом биологических объектов характеризуется многообразием протеоформ и количеством копий каждой протеоформы. Возможности современных аналитических методов не позволяют экспериментально оценить размеры протеома. Для оценки многообразия протеоформ в работе предложена информационная модель и расчетные формулы, учитывающие частоты возникновения сплайс-вариантов мРНК, одноаминокислотных полиморфизмов и посттрансляционных модификаций, а также степень изученности объекта. Предложенный подход позволяет теоретически оценить многообразие белков протеома исследуемого объекта.

Проведенный в работе анализ свидетельствует о том, что следует различать протеом (а) конкретного образца – индивидуальный протеом, специфичный для индивидуума и типа биологического материала; (б) биологического материала – объединение индивидуальных протеомов, измеренных в одном и том же типе биоматериала в различных экспериментах; (в) вида организма – популяционный протеом, совокупность сведений об индивидуальных протеомах популяции, объединенных для различных типов биологического материала в составе постгеномных информационных ресурсов.

Данная работа позволяет сформулировать новую научную задачу в области биохимии и протеомики – создание протеомного профиля человека с учетом индивидуальных особенностей экспрессии определенных видов белков (протеоформ).

ВЫВОДЫ

1. Для описания набора генов предложена интегральная количественная характеристика – *информационный профиль* (ИП), отражающая соотношение между аннотациями генов. Для описания хромосом человека выбраны дескрипторы ИП, описывающие наличие информации о полиморфизмах, альтернативном сплайсинге, посттрансляционных модификациях, экспериментально детектированных транскриптах и белках, взаимосвязи с развитием заболеваний и опубликованных исследованиях. Сведения хотя бы об одном однонуклеотидном полиморфизме, потенциально реализуемом на уровне протеома, есть для 95% генов человека; экспериментально детектированных белках – для 70% генов, сведения о количестве белка в плазме крови человека – только для 1% генов. Взаимосвязь с развитием заболеваний показана для 48% генов в геноме человека.

2. Значения весовых коэффициентов дескрипторов ИП постоянны для хромосом человека, за исключением наиболее коротких – Y хромосомы и митохондриальной, и совпадают с весовыми коэффициентами дескрипторов полногеномного ИП человека и выборки случайным образом отобранных генов (N > 200).

3. Начиная с 2013 года, весовые коэффициенты дескрипторов, отражающих количество аннотаций о сплайс-вариантах, мутациях и посттрансляционных модификациях, согласно протеомным базам NeXtProt и UniProt, оставались неизменны для генома человека, что позволяет использовать накопленные в этих ресурсах частоты указанных молекулярных событий для оценки многообразия протеоформ.

4. Объединение экспериментальных результатов анализа трех типов биологического материала показало, что панорамные ($РНК_{сек}$) и направленные методы ($ПЦР_{рв}$ + $ПЦР_{цк}$) транскриптомного анализа позволяют измерить в хромосомоцентричном эксперименте транскрипты для 85 % генов; методом направленного масс-спектрометрического анализа ($ММР_{мс}$) в том же образце может быть измерено содержание белков для 46 % генов выбранной хромосомы. Корреляция между количеством копий транскрипта и соответствующего белка отсутствует ($R^2 \approx 0.1$).

5. Протеом биологических объектов характеризуется количеством протеоформ и количеством копий каждой протеоформы. Следует разделять протеом конкретного образца (индивидуальный протеом) и протеом биологического вида (популяционный протеом), характеризующиеся различным количеством протеоформ. Для теоретической оценки многообразия протеоформ в составе протеома предложено три информационных модели и расчетных формулы, учитывающие частоты возникновения протеоформ в результате альтернативного сплайсинга, одноаминокислотных замен и посттрансляционных модификаций.

6. С учетом независимой природы указанных молекулярных событий, индивидуальный протеом ткани печени может состоять из ≈ 400 тыс. протеоформ, опухолевая клеточная линия HepG2 ≈ 500 тыс., популяционный протеом человека до 7,14 млн протеоформ (согласно частотам, предоставляемым базой NeXtProt). Накопление

многообразия протеформ в составе популяционного протеома обусловлено наличием уникальных транскрибируемых вариантов однонуклеотидных полиморфизмов в геноме.

7. Изученность объекта (доля аннотированных генов в геноме) вносит наибольший вклад в оценку количества протеоформ. Следствием является сравнительно небольшое количество предсказанных протеоформ для протеома растений (около 400 тыс.) при существенном размере генома (27,4 тыс. белок-кодирующих генов для *Arabidopsis thaliana*). Наиболее изучены в сравнении с другими модельными объектами представители царства бактерий.

8. Исследование образцов клеток ткани печени человека и клеточной линии HepG2 показало, что наибольшая специфичность (90 %) в сравнении с транскриптомом того же образца достигается при использовании направленного масс-спектрометрического анализа методом $ММР_{мс}$, при этом чувствительность метода составляет только 40 %.

СПИСОК ПУБЛИКАЦИЙ ПО ТЕМЕ ДИССЕРТАЦИИ

1. Арчаков, А. И., Крохин, Н. В., Лисица, А. В., **Пономаренко, Е. А.**, Старовойтов, А. В. Биодекриптомика макромолекул // Информатизация и связь, — 2010. — № 2. — С. 65–67.

2. Евдокимов, П. А., **Пономаренко, Е. А.**, Аверчук, В. В., Кудрявцев, А. М. Проект BioKnol: социальная сеть для молекулярных биологов // Интеграл. — 2012. — № 3 (65). — С. 16–17.

3. Киселева, Я. Ю., Птицын, К. Г., Тихонова, О. В., Радько, С. П., Курбатов, Л. К., Вахрушев, И. В., Згода, В. Г., **Пономаренко, Е. А.**, Лисица, А. В., Арчаков, А. И. ПЦР-анализ абсолютного числа копий транскриптов хромосомы 18 человека в клетках печени и линии HepG2 — Биомедицинская химия. — 2017. — Том 63(2). — С.147–153.

4. Лисица, А. В., **Пономаренко, Е. А.**, Лохов, П. Г., Арчаков, А. И., Постгеномная медицина: альтернатива биомаркерам // Вестник Российской академии медицинских наук. — 2016. — № 71 (3). — С. 255–260.

5. **Пономаренко, Е. А.**, Згода, В. Г., Копылов, А. Т., Поверенная, Е. В., Ильгисонис, Е. В., Лисица, А. В., Арчаков, А. И., Россия в международном проекте «Протеом человека»: первые итоги и перспективы // Биомедицинская химия. — 2015. — № 61 (2). — С. 169–175.

6. **Пономаренко, Е. А.**, Лисица, А. В., Ильгисонис, Е. В., Арчаков, А. И. Создание семантических сетей белков с использованием Pubmed/MEDLINE // Молекулярная биология. — 2010. — № 44 (1). — С. 152–161.

6а. **Ponomarenko, E. A.**, Lisitsa, A. V., Ilgisonis, E. V., Archakov, A. I. Construction of Protein Semantic Networks Using PubMed/MEDLINE // Molecular Biology. — 2010. — V. 44(1). — P. 140–149.

7. **Пономаренко, Е. А.**, Лисица, А. В., Карузина, И. И., Мирошниченко, Ю. В. Автоматизированное аннотирование функциональных свойств белков надсемейства цитохромов Р450 // Аллергия, астма и клиническая иммунология. — 2003. — № 7(8). — С. 95–99.

8. **Пономаренко, Е. А.**, Лисица, А. В., Петрак, И., Мошковский, С. А., Арчаков, А. И. Выявление дифференциально-экспрессирующихся белков с использованием автоматического мета-анализа протеомных публикаций // Биомедицинская химия. — 2009. — № 55(1). — С. 5–14.

8а. **Ponomarenko, E. A.**, Lisitsa, A. V., Petrak, J., Moshkovskii, S. A., Archakov, A. I. Identification of Differentially Expressed Proteins Using Automated Meta-Analysis of Proteomics-Related Articles //Biochemistry (Moscow) Supplement Series B: Biomedical Chemistry. — 2009. — № 3(1). — С. 10–16.

9. **Пономаренко, Е. А.**, Ильгисонис, Е. В., Лисица, А. В. Технологии знаний в протеомике // Биоорганическая химия. — 2011. — № 37(2). — С. 190–198.

9a. **Ponomarenko, E. A.**, Lisitsa, A. V., Ilgisonis, E. V. Knowledge-based Technologies in Proteomics // Russian Journal of Bioorganic Chemistry. — 2011. — V. 37(2). — P. 168–175.

10. Ромашова, Ю. А., **Пономаренко, Е. А.**, Евдокимов, П. А., Щербаков, А. Ю., Лисица, А. В., Арчаков, А. И. База знаний как инструмент для мониторинга постгеномных медико-биологических исследований // Вестник Российской академии медицинских наук. — 2011. — № 8. — С. 20–24.

11. Федорова, А. А., Ильгисонис, Е. В., Бурнышова, К. М., Мирошниченко, Ю. В., **Пономаренко, Е. А.** Мета-анализ публикаций в области молекулярных механизмов развития болезни Альцгеймера // Интеграл. — 2012. — № 3(65). — С. 24–26.

12. Чернобровкин, А. Л., Трифонова, О. П., Петушкова, Н. А., **Пономаренко, Е. А.**, Лисица, А. В. Выбор допустимой погрешности определения массы пептида при идентификации белков методом пептидного картирования // Биоорганическая химия. — 2011. — № 37(1). — С. 132–136.

12a. Chernobrovkin, A. L., Petushkova, N. A., **Ponomarenko, E. A.**, Lisitsa, A. V., Trifonova, O. P. Selection of the Peptide Mass Tolerance Value for Protein Identification with Peptide Mass Fingerprinting // Russian Journal of Bioorganic Chemistry. — 2011. —V. 37(1). — P. 119–122.

13. Швец, В. И., Корнюшко, В. Ф., Угольникова, О. А., Лисица, А. В., **Пономаренко, Е. А.** Информационная поддержка систем формирования предметной области проблемно-ориентированных баз данных по белкам и биологически активным соединениям с использованием ассоциативного библиометрического анализа // Интеграл. — 2011. — № 1. — С. 24–26.

14. Archakov, A., Lisitsa, A., **Ponomarenko, E.**, Zgoda, V., Recent Advances in Proteomic Profiling of Human Blood: Clinical Scope // Expert Review of Proteomics. — 2015. — № 12 (2). — P. 111–113.

15. Archakov, A., Aseev, A., Bykov, V., Grigoriev, A., Govorun, V., Ivanov, V., Khlunov, A., Lisitsa, A., Mazurenko, S., Makarov, A., **Ponomarenko, E.**, Sagdeev, R., Skryabin, K. Gene-centric View on the Human Proteome Project: the Example of the Russian Roadmap for Chromosome 18 // Proteomics. — 2011. — V. 11(10). — P. 1853–1856.

16. Archakov, A., Zgoda, V., Kopylov, A., Naryzhny, S., Chernobrovkin, A., **Ponomarenko, E.**, Lisitsa, A. Chromosome-centric Approach to Overcoming Bottlenecks in the Human Proteome Project // Expert Review of Proteomics. — 2012. — V. 9(6). — P. 667–676.

17. Evdokimov, P., Kudryavtsev, A., Ilgisonis, E., **Ponomarenko, E.**, Lisitsa, A. Use of Scientific Social Networking to Improve the Research Strategies of PubMed Readers // BMC Research Notes. — 2016. — V. 9(113). — P. 1–6.

18. Horvatovich, P., Lundberg, E. K., Chen, Y. J., Sung, T. Y., He, F., Nice, E. C., Goode, R. J., Yu, S., Ranganathan, S., Baker, M. S., Domont, G. B., Velasquez, E., Li, D., Liu, S., Wang, Q., He, Q. Y., Menon, R., Guan, Y., Corrales, F. J., Segura, V., Casal, J. I., Pascual-

Montano, A., Albar, J. P., Fuentes, M., Gonzalez-Gonzalez, M., Diez, P., Ibarrola, N., Degano, R. M., Mohammed, Y., Borchers, C. H., Urbani, A., Soggiu, A., Yamamoto, T., Salekdeh, G. H., Archakov, A., **Ponomarenko, E.**, Lisitsa, A., Lichti, C. F., Mostovenko, E., Kroes, R. A., Rezeli, M., Végvári, Á., Fehniger, T. E., Bischoff, R., Vizcaíno, J. A., Deutsch, E. W., Lane, L., Nilsson, C. L., Marko-Varga, G., Omenn, G. S., Jeong, S. K., Lim, J. S., Paik, Y. K., Hancock, W. S. Quest for Missing Proteins: Update 2015 on Chromosome-Centric Human Proteome Project // Journal of Proteome Research. — 2015. — V. 14(9). — P. 3415–3431.

19. Lisitsa, A. V., Poverennaya, E. V., **Ponomarenko, E. A.**, Archakov, A. I. The Width of the Human Plasma Proteome Compared With a Cancer Cell Line and Bacteria // Biomolecular Research & Therapeutics. — 2015. — V. 4. — P. 132.

20. Lisitsa, A. V., **Ponomarenko, E. A.**, Kiseleva, O. I., Poverennaya, E. V., Archakov, A. I. Molar Concentration Welcomes Avogadro in Postgenomic Analytics // Biochemistry & Analytical Biochemistry. — 2015. — V. 4. — P. 216–220.

21. Lisitsa, A., Moshkovskii, S., Chernobrovkin, A., **Ponomarenko, E.**, Archakov, A. Profiling Proteoforms: Promising Follow-up of Proteomics for Biomarker Discovery // Expert Review of Proteomics. — 2014. — V. 11(1). — P.121–129.

22. Kopylov, A. T., Ilgisonis, E. V., Tikhonova, O. V., Moysa, A. A., Zavialova, M. G., Novikova, S. E., Lisitsa, A. V., **Ponomarenko, E. A.**, Moshkovskii, S. A., Markin, A. A., Grigoriev, A. I., Zgoda, V. G., Archakov, A. I. Targeted Quantitative Screening of Chromosome 18 Encoded Proteome in Plasma Samples of Astronaut Candidates // Journal of Proteome Research. — 2016. — V. 15(11). — P. 4039–4046.

23. Krasnov, G. S., Dmitriev, A. A., Kudryavtseva, A. V., Shargunov, A. V., Karpov, D. S., Uroshlev, L. A., Melnikova, N. V., Blinov, V. M., Poverennaya, E. V., Archakov, A. I., Lisitsa, A. V., **Ponomarenko, E. A.** PPLine: An Automated Pipeline for SNP, SAP, and Splice Variant Detection in the Context of Proteogenomics // Journal of Proteome Research. — 2015. — V. 14(9). — P. 3729–3737.

24. Naryzhny, S. N., Zgoda, V. G., Maynskova, M. A., Novikova, S. E., Ronzhina, N. L., Vakhrushev, I. V., Khryapova, E. V., Lisitsa, A. V., Tikhonova, O. V., **Ponomarenko, E. A.**, Archakov, A. I. 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. — № 37 (2). — P. 302–309.

25. Naryzhny, S. N., Lisitsa, A. V., Zgoda, V. G., **Ponomarenko, E. A.**, Archakov, A. I. 2DE-based Approach for Estimation of Number of Protein Species in a Cell // Electrophoresis. — 2014. — V. 35(6). — P. 895–900.

26. **Ponomarenko, E. A.**, Kopylov, A. T., Lisitsa, A. V., Radko, S. P., Kiseleva, Y. Y., Kurbatov, L. K., Ptitsyn, K. G., Tikhonova, O. V., Moisa, A. A., Novikova, S. E., Poverennaya, E. V., Ilgisonis, E. V., Filimonov, A. D., Bogolubova, N. A., Averchuk, V. V., Karalkin, P. A., Vakhrushev. I. V., Yarygin, K. N., Moshkovskii, S. A., Zgoda, V. G., Sokolov, A. S., Mazur, A. M., Prokhortchouck, E. B., Skryabin, K. G., Ilina, E. N., Kostrjukova, E. S.,

Alexeev, D. G., Tyakht, A. V., Gorbachev, A. Y., Govorun, V. M., Archakov, A. I. Chromosome 18 Transcriptoproteome of Liver Tissue and HepG2 Cells and Targeted Proteome Mapping in Depleted Plasma: Update 2013 // Journal of Proteome Research. — 2014. — V. 13(1). — P. 183–190.

27. **Ponomarenko, E. A.**, Poverennaya, E. V., Ilgisonis, E. V., Pyatnitskiy, M. A., Kopylov, A. T., Zgoda, V. G., Lisitsa, A. V., Archakov, A. I. The Size of the Human Proteome: The Width and Depth // International Journal of Analytical Chemistry. — 2016. — V. 2016. — P. 1–6.

28. **Ponomarenko, E.**, Lisitsa, A., Archakov, A., Baranova, A., Albar, J. P. The Chromosome-Centric Human Proteome Project at FEBS Congress // Proteomics. — 2014. — V. 14(2-3). — P. 147–152.

29. **Ponomarenko, E.**, Poverennaya, E., Pyatnitskiy, M., Lisitsa, A., Moshkovskii, S., Ilgisonis, E., Chernobrovkin, A., Archakov, A. Comparative Ranking of Human Chromosomes Based on Post-Genomic Data // Omics: a Journal of Integrative Biology. — 2012. — V. 16(11). — P. 604–611.

30. Poverennaya, E. V., Kopylov, A. T., **Ponomarenko, E. A.**, Ilgisonis, E. V., Zgoda, V. G., Tikhonova, O. V., Novikova, S. E., Farafonova, T. E., Kiseleva, Y. Y., Radko, S. P., Vakhrushev, I. V., Yarygin, K. N., Moshkovskii, S. A., Kiseleva, O. I., Lisitsa, A. V., Sokolov, A. S., Mazur, A. M., Prokhortchouck, E. B., Skryabin, K. G., Kostrjukova, E. S., Tyakht, A. V., Gorbachev, A. Y., Ilina, E. N., Govorun, V. M., Archakov, A. I. State of the Art on Chromosome 18-centric HPP in 2016: Transcriptome and Proteome Profiling of Liver Tissue and HepG2 Cells // Journal of Proteome Research. — 2016. — V. 15(11). — P. 4030–4038.

31. Poverennaya, E. V., Bogolubova, N. A., Bylko, N. N., **Ponomarenko, E. A.**, Lisitsa, A. V., Archakov, A. I. Gene-centric Content Management system // Biochimica et Biophysica Acta - Proteins and Proteomics. — 2014. — № 1844(1 PtA). — P. 77–81.

32. Poverennaya, E. V., Bogolubova, N. A., **Ponomarenko, E. A.**, Lisitsa, A. V., Archakov, A. I. GenoCMS - The Content Management System for genes and proteins // Journal of Proteomics & Bioinformatics. — 2013. — V. 6(8). — P. 176–182.

33. Shargunov, A. V., Krasnov, G. S., **Ponomarenko, E. A.**, Lisitsa, A. V., Shurdov, M. A., Zverev, V. V., Archakov, A. I., Blinov, V. M. Tissue-specific Alternative Splicing Analysis Reveals the Diversity of Chromosome 18 Transcriptome // Journal of Proteome Research. — 2014. — V. 13(1). — P. 173–182.

34. Zgoda, V. G., Kopylov, A. T., Tikhonova, O. V., Moisa, A. A., Pyndyk, N. V., Farafonova, T. E., Novikova, S. E., Lisitsa, A. V., **Ponomarenko, E. A.**, Poverennaya, E. V., Radko, S. P., Khmeleva, S. A., Kurbatov, L. K., Filimonov, A. D., Bogolyubova, N. A., Ilgisonis, E. V., Chernobrovkin, A. L., Ivanov, A. S., Medvedev, A. E., Mezentsev, Y. V., Moshkovskii, S. A., Naryzhny, S. N., Ilina, E. N., Kostrjukova, E. S., Alexeev, D. G., Tyakht, A. V., Govorun, V. M., Archakov, A. I. Chromosome 18 Transcriptome Profiling and Targeted Proteome Mapping in Depleted Plasma, Liver Tissue and HepG2 cells // Journal of Proteome Research. — 2013. — V. 12(1). — P. 123–134.

Материалы трудов конференций

35. Арчаков, А. И., Лисица, А. В., **Пономаренко, Е. А.** Биоинформационный взгляд на проект «Протеом человека» // Сборник трудов Международной научно-практической конференции «Биотехнологии и качество жизни». — Москва, 2014.

36. Иванов, Н. А., Лисица, А. В., **Пономаренко, Е. А.**, Арчаков, А. И. Тематический анализ резюме научных публикаций в области цитохромов Р450 // Сборник материалов Сессии ИВТН. — Москва, 2003. — С. 28-29.

37. Лисица, А. В., Мирошниченко, Ю. В., **Пономаренко, Е. А.** База знаний по цитохромам Р450 // Сборник научных трудов X Российского национального конгресса «Человек и лекарство». — 2003. — С. 730.

38. Лисица, А. В., Поверенная, Е. В., Боголюбова, Н. А., **Пономаренко, Е. А.** Создание базы знаний по 18-й хромосоме человека с использованием геноцентричного принципа // Сборник тезисов докладов научного форума «Наука и общество». — Санкт-Петербург, 2012.

39. Поверенная, Е. В., Лисица, А. В., **Пономаренко, Е. А.** Создание базы знаний по 18-й хромосоме человека с использование геноцентричного принципа // Тезисы II Международной научно-практической конференции «Постгеномные методы анализа в биологии, лабораторной и клинической медицине: геномика, протеомика, биоинформатика». — Новосибирск, 2011. — С. 125.

40. Поверенная, Е. В., Чернобровкин, А. Л., **Пономаренко, Е. А.**, Лисица, А. В. Выявление белок-белковых взаимодействий путем анализа масс-спектрометрических данных // Сборник трудов XX Российского национального конгресса «Человек и лекарство». — Москва, 2013. — С. 408.

41. Поверенная, Е. В., **Пономаренко, Е. А.**, Пятницкий, М. А., Лисица, А. В., Мошковский, С. А., Ильгисонис, Е. В., Чернобровкин, А. Л., Арчаков, А. И. Сравнительный анализ хромосом человека на основе постгеномных данных // Сборник тезисов докладов научной конференции ФГБУ «ИБМХ» РАМН. — Москва, 2013. — С. 17.

42. Поверенная, Е. В., Лисица, А. В., **Пономаренко, Е. А.** Gene-Centric Content Management System: база знаний по белкам 18-й хромосомы человека // Сборник трудов конференции «Человек и лекарство». — 2012. — С. 555.

43. **Пономаренко, Е. А.**, Лисица, А. В., Копылов, А. Т., Згода, В. Г., Арчаков, А. И. Перспективы исследования протеома человека. // Конгресс «Мир биотехнологий». — Москва, 2017.

44. **Пономаренко, Е. А.**, Лисица, А. В., Гусев, С. А. База знаний по цитохромам Р450 // Материалы международной школы-конференции молодых ученых «Системная биология и биоинженерия». — МАКС Пресс : Москва, 2005. — С. 50.

45. **Пономаренко, Е. А.**, Лисица, А. В., Карузина И.И., Гусев С.А. База знаний по цитохромам Р450 // Сборник материалов Сессии ИВТН. — Москва, 2006. — С. 32.

46. **Пономаренко, Е. А.**, Лисица, А. В., Арчаков, А. И. Лингвистические методы поиска взаимосвязанных белков // Сборник трудов конференции «Человек и лекарство». — 2008. — С. 523.

47. **Пономаренко, Е. А.**, Лисица, А. В., Арчаков, А. И. Лингвистические методы поиска взаимосвязанных белков // Сборник трудов конференции «Математика. Компьютер. Образование». — 2009. — С. 68.

48. Archakov, A. I., **Ponomarenko, E. A.**, Poverennaya, E. V., Lisitsa, A. V., Kopylov, A. T., Zgoda, V. G. State of the Chromosome 18-centric HPP in 2016: Transcriptome and Proteome Profiling of Liver Tissue and HepG2 cells // In: Proceedings HUPO. — Taipei, 2016.

49. Archakov, A. I., **Ponomarenko, E. A.**, Poverennaya, E. V., Lisitsa, A. V., Kopylov, A. T., Zgoda, V. G. The First Master Proteome of Single Chromosome: Example of Human Chromosome 18 // In: Proceedings HUPO. — Vancouver, 2015. — P. 120.

50. Archakov, A. I., **Ponomarenko, E. A.**, Poverennaya, E. V., Lisitsa, A. V., Kopylov, A. T., Zgoda, V. G. Size of Human Proteome: Example of Chr 18 // In: Proceedings the HUPO 13th Annual World Congress. — Madrid, 2014.

51. Archakov, A. I., **Ponomarenko, E. A.**, Poverennaya, E. V., Lisitsa, A. V., Kopylov, A. T., Zgoda, V. G. Size of Master Protein Proteome Expressed by Single Chromosome in Different Tissues and Cells // In: Proceedings the Joint 7th AOHUPO Congress and 9th International Symposium of the Protein Society of Thailand. — Bangkok, 2014.

52. Kopylov, A. T., Zgoda, V. G., Lisitsa, A. V., **Ponomarenko, E. A.**, Poverennaya, E. V., Ilgisonis, E. V., Moisa, A. A., Filimonov, A. D., Archakov, A. I. Hidden Proteome: Multiplex Quantitation of low- and Ultralow-copy Number Proteins in HepG2 cells // In: Proceedings the FEBS Journal. — Saint Petersburg, 2013. — V. 280. — P. 637.

53. Lisitsa, A. V., Poverennaya, E. V., **Ponomarenko, E. A.**, Archakov, A. I. Comprehensive Characterization of a Liver Tissue and HepG2 Cells Transctiproproteome for Human Chromosome 18 // In: Proceedings the Joint 7th AOHUPO Congress and 9th International Symposium of the Protein Society of Thailand. — Bangkok, 2014.

54. Lisitsa, A. V., Poverennaya, E. V., **Ponomarenko, E. A.**, Archakov, A. I. Consolidating Chr 18 Data Using Knowledgebase of Protein and Transcript Annotations // In: Proceedings the HUPO 13th Annual World Congress. — Madrid, 2014.

55. Lisitsa, A. V., Poverennaya, E. V., Bogolubova, N. A., **Ponomarenko, E. A.** Consolidating Chr18 Data Using Knowledge Base of Protein and Transcript Features // In: Proceedings the Proteomic Forum 2013. — Berlin, 2013. — P. 143.

56. Lisitsa, A. V., Poverennaya, E. V., Bogolubova, N. A., Bylko, N. N., **Ponomarenko, E. A.**, Archakov, A. I. Consolidating Chr 18 Data Using Knowledgebase of Protein and Transcript Annotations // In: Proceedings the FEBS Journal. — Saint Petersburg, 2013. — V. 280. — P. 635.

57. Lisitsa, A., Poverennaya, E., Bogolubova, N., **Ponomarenko, E.** Knowledgebase for Single Chromosome Proteins // In: Proceedings 6th Congress AOHUPO. — Beijing, 2012. — P. 69.

58. Lisitsa, A., Poverennaya, E., Bogolyubova, N., **Ponomarenko, E.** Gene-Centric Knowledgebase for a Single Chromosome // In: Proceedings European Proteomics Association 2012 Scientific Congress Hosted by The British Society for Proteome Research. — Glasgow, 2012.

59. Lisitsa, A. V., Poverennaya, E. V., Bogolyubova, N. A., **Ponomarenko, E. A.** Gene-Centric Knowledgebase for a Single Chromosome on the Web // In: Proceedings the HUPO 11th Annual World Congress. — Boston, 2012. — P. 99.

60. Lisitsa, A., Poverennaya, E., Filimonov, A., **Ponomarenko, E.** Creation of Knowledgebase Using Gene-Centric Content Management System (CMS) // In: Proceedings 5th Central and Eastern European Proteomic Conference. — Prague, 2011.

61. Lisitsa, A. V, **Ponomarenko, E. A.**, Karuzina, I. I., Ivanov, A. S., Archakov, A. I. Balance Sheet for Cytochrome P450 Knowledgebase // In: Proceedings 13-th International Conference on Cytochromes P450. — Prague, 2003. — P. 67–73.

62. Lisitsa, A. V., **Ponomarenko, E. A.**, Gusev, S. A., Kuznetsova, G. P., Karuzina, I. I., Lewi, P., Archakov, A. I. Cytochrome P450 Knowledgebase: Structure and Functionality // In: Proceedings 14th International Conference on Cytochromes P450: Biophysics and Bioinformatics. — Dallas, USA, 2005. — P. 29–34.

63. **Ponomarenko, E. A.** Gene-centric Knowledgebase as a Tool for Estimating Protein Species Number // In: Proceedings HUPO. — Vancouver, 2015. — P. 120.

64. **Ponomarenko, E. A.**, Krasnov, G. S., Poverennaya, E. V., Kiseleva, O. I., Lisitsa, A. V., Archakov, A. I. Estimation of Protein Species Number Based on Consolidation of Postgenomic Data in Gene-centric Knowledgebase // In: Proceedings EUPA. — Milano, 2015. — P. 49

65. **Ponomarenko, E. A.**, Naryzhny, S. N., Poverennaya, E. V., Pyatnitskii, M. A., Lisitsa, A. V., Archakov, A. I. Estimation of Protein Species Number for Mammalian, Bacteria, Insecta and Yeast // In: Proceedings the FEBS Journal. — Saint Petersburg, 2013. — V. 280. — P. 635.

66. **Ponomarenko, E.**, Naryzhny, S., Poverennaya, E., Pyatnitskii, M., Lisitsa, A., Archakov, A. Estimation of Protein Species Number for Mammalian, Bacteria, Insecta and Yeast // In: Proceedings the HUPO 12th Annual World Congress. — Yokohama, 2013. — P. 23.

67. **Ponomarenko, E.**, Pyatnitskiy, M., Poverennaya, E., Lisitsa, A. Comparative Analysis of Human Chromosomes Based on Post-genomic Data // In: Proceedings 6th Congress AOHUPO. — Beijing, 2012. — P. 322.

68. **Ponomarenko, E. A.**, Lisitsa, A. V., Archakov, A. I. Text Mining Tools in Analysis of High-Throughput Data // Материалы конференции CMTPI. — 2007. — P. 135.

69. **Ponomarenko, E. A.**, Lisitsa, A. V., Archakov, A. I. Searching for Related Proteins Using Textomic Approach // Сборник трудов конференции HUPO. — 2007. — P. 103.

70. **Ponomarenko, E. A.**, Lisitsa, A. V., Petrak, J., Moshkovskii, S. A., Archakov, A. I. Textomics Tools for Automatically Update the Hit-parade of Repeatedly Identified Proteins // Сборник материалов международной конференции GPBNM. — 2008. — P. 36.

71. **Ponomarenko, E. A.**, Lisitsa, A. V., Petrak, J., Moshkovskii, S. A., Archakov, A. I. Automated meta-analysis Confirms the Hit-parade of Repeatedly Identified Proteins // Сборник материалов международной конференции HUPO. — 2008. — P. 1669.

72. Poverennaya, E., **Ponomarenko, E.**, Lisitsa, A. Chromosome-centric View of Human Protein-protein Interactions Based on In Silico Analysis // In: Proceedings the Proteomic Forum 2013. — Berlin, 2013. — P. 138.

73. Poverennaya, E. V., Chernobrovkin, A. L., **Ponomarenko, E. A.**, Lisitsa, A. V. Chromosome-centered Interactome of Human Chromosome 18 by Analysis of GPMDB Datasets // In: Proceedings the FEBS Journal.— Saint Petersburg, 2013. — V. 280. — P. 636.

74. Poverennaya, E. V., Chernobrovkin, A. L., **Ponomarenko, E. A.**, Lisitsa, A. V. Obtaining Interactome Map by Sifting the Collection of MS-Data // In: Proceedings the HUPO 12th Annual World . Yokohama, 2013. — P. 186.

75. Poverennaya, E., Bogolubova, N., Lisitsa, A., **Ponomarenko, E.** Gene-Centric Knowledgebase on the Web // In: Proceedings 8th International Conference of Bioinformatics of Genome Regulation and Structure/systems Biology. — Novosibirsk, 2012.

76. **Пономаренко, Е. А.**, Худоклинова, Ю. Ю., Мирошниченко, Ю. В., Мишень для определения концентрационной чувствительности лазерных времяпролетных масс-спектрометрометров с матричным типом ионизации биообразцов. — 2009. — Патент на полезную модель. — Рег. номер: RU 00090073 U1.