«УТВЕРЖДАЮ»:
Заместитель декана биологического факультета
МГУ имени М.В.Ломоносова,
доктор биологических наук,
профессор А.М.Рубцов

«____» _____ 202__ г.

ЗАКЛЮЧЕНИЕ
кафедры биоинженерии биологического факультета Федерального государственного бюджетного образовательного учреждения высшего образования «Московский государственный университет имени М.В.Ломоносова»

Диссертация «Анализ интерактома нуклеосом в хроматине и его роль в патогенезе заболеваний» выполнена на кафедре биоинженерии биологического факультета МГУ имени М.В.Ломоносова.

В период подготовки диссертации Грибкова Анна Кирилловна обучается в очной аспирантуре биологического факультета на кафедре биоинженерии по специальности 1.5.8 – «Математическая биология, биоинформатика» с 01.10.2019 г., планируемая дата окончания обучения 30.09.2023 г.; а также работает на биологическом факультете в должности младшего научного сотрудника.

В 2019 г. окончила Федеральное государственное бюджетное образовательное учреждение высшего образования «Московский государственный университет имени М. В. Ломоносова» по специальности «Биология».

Свидетельство об окончании аспирантуры, подтверждающее сдачу кандидатских экзаменов, будет выдано в 2023 г.

Научный руководитель – д.ф.-м.н., профессор РАН, чл.-корр. РАН Шайтан Алексей Константинович, доцент кафедры биоинженерии Биологического факультета Федерального государственного бюджетного образовательного

учреждения высшего образования «Московский государственный университет имени М. В. Ломоносова».

По итогам обсуждения принято следующее заключение.

Работа выполнена на хорошем методическом уровне. Личный вклад автора является определяющим во всех проведенных исследованиях. Результаты работы были представлены в виде 5 устных и стендовых докладов на международных и российских конференциях. Автор разработала классификацию белков хроматина, с помощью которой провела анализ обогащения белков хроматина клеток человека, классифицировала структуры комплексов нуклеосом с негистоновыми белками и проанализировала дифференциально экспрессированные гены в пациентах с множественной миеломой. Результаты расширяют и дополняют современные представления о белках хроматина и могут способствовать разработке новых лекарственных средств, взаимодействующих с белками хроматина человека. Результаты также будут способствовать поиску биомаркеров онкологических заболеваний среди белков хроматина человека.

6. Текст диссертации соответствует установленным правилам научного цитирования, библиографические ссылки оформлены корректно.

7. Диссертационное исследование по своему содержанию соответствует заявленной специальности 1.5.8 – «Математическая биология, биоинформатика».

8. Основные идеи и положения работы изложены в 6 научных работах автора общим объемом 4.5 п.л., в том числе 4 публикациях (объемом 4.42 п.л.) в рецензируемых научных изданиях, рекомендованных для защиты в диссертационном совете МГУ по специальности.

9. В своих научных трудах соискатель сравнила широкоиспользуемые базы данных белков на предмет качественного и количественного состава ядра, обнаружила особенности белков, выделяемых при экспериментальной очистке белков хроматина.

Диссертация «Анализ интерактома нуклеосом в хроматине и его роль в патогенезе заболеваний» Грибковой Анны Кирилловны по всем пунктам соответствует требованиям, установленным в соответствии с Федеральным законом "О науке и государственной научно-технической политике". Представленная диссертация рекомендуется к защите на соискание ученой степени кандидата физико-математических наук по научной специальности 1.5.8 – «Математическая биология, биоинформатика».

Заключение принято на заседании кафедры биоинженерии биологического факультета МГУ имени М.В.Ломоносова. Присутствовало на заседании 18 чел. Результаты голосования: «за» - 18 чел., «против» - 0 чел., «воздержалось» - 0 чел., протокол № 9 от «17» августа 2023 г.

Зам. заведующего кафедрой биоинженерии
Биологического факультета
МГУ имени М.В.Ломоносова
Д.ф.м.н., профессор К.В. Шайтан

МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
имени М.В. ЛОМОНОСОВА
БИОЛОГИЧЕСКИЙ ФАКУЛЬТЕТ

На правах рукописи

Грибкова Анна Кирилловна

Анализ интерактома нуклеосом в хроматине и его роль в патогенезе заболеваний

1.5.8 - «Математическая биология, биоинформатика»

ДИССЕРТАЦИЯ
на соискание ученой степени
кандидата физико-математических наук

Научный руководитель:
д.ф.-м.н. Шайтан Алексей Константинович

Москва – 2023

Оглавление

Введение

Актуальность диссертационной работы

Хроматин, комплекс нуклеиновых кислот и белков в ядре клеток эукариот, основной участник важнейших молекулярно-биологических процессов, таких как транскрипция, репликация и репарация ДНК.

Первые попытки описать физико-химические свойства хроматина предпринимались еще в начале 20-го века, уже во второй половине 20-го века были охарактеризованы некоторые классы негистоновых белков хроматина. На сегодняшний день известно о нескольких тысячах белков в ядре клеток человека, но детали их функционирования остаются недостаточно изученными.

Развитие экспериментальных методик выделения протеома ядра и хроматина с последующим масс-спектрометрическим анализом позволило описать белковый состав некоторых клеточных линий, тканей человека и ядер клеток модельных организмов. Вместе с тем предпринимаются усилия по функциональной классификации белков хроматина, как полуавтоматизированными подходами в Gene Ontology, так и путем создания специализированных баз данных отдельных классов белков. Однако полного понимания разнообразия состава хроматома, в том числе в разных тканях и на разных временах жизни клеток, на сегодняшний день отсутствует.

Не смотря на то, что для белков хроматина, отвечающих за реализацию наследственного материала, известны драйверные мутации и нарушения экспрессии, приводящие к онкологическим нарушения, лишь малая часть белков хроматина является мишенями для ингибирования лекарственными препаратами. Это связано с недостаточной изученностью репертуара функционирования белков хроматина и их взаимодействий на структурном уровне.

Цель диссертационной работы

Целью диссертационной работы является исследование функционального разнообразия белков ядра клеток человека, их физико-химических свойств, взаимодействий с нуклеосомами и их роли в онкологических заболеваниях.

Задачи исследования

1. Провести сравнительный анализ систем классификации и охвата источников информации (базы данных, эксперименты) о локализации белков ядра и хроматина, экспериментально полученных хроматомов.

2. Построить схему эмпирической классификации белков хроматина, добавить в категории белки из баз данных и литературы, оценить с помощью нее представленность белков хроматина в ядре клеток.
3. Выявить особенности физико-химических свойств и доменной организации белков хроматина.
4. Описать разнообразие структур комплексов нуклеосом, провести структурный анализ комплексов нуклеосом человека с негистоновыми белками ядерного протеома.
5. Идентифицировать рекуррентные мутации и нарушения экспрессии белков ядерного протеома в образцах пациентов с онкологическими заболеваниями, провести структурную интерпретацию мутаций белков на интерфейсе взаимодействий нуклеосом с негистоновыми белками хроматина.

Положения, выносимые на защиту

1. Среди общепринятых источников информации о локализации белков отсутствует консенсус в определении структур ядра и хроматина и их белковом составе.
2. Методики выделения хроматина недостаточно чувствительны и специфичны, а белковый состав хроматомов в среднем только на 62% совпадает с составом белков содержимого клеточного ядра, аннотированных в базах UniProt или HPA.
3. Разработанная эмпирическая классификация белков хроматина человека позволяет оценивать разнообразие и обогащение категорий белков хроматина в наборах белков для исследований.
4. Выявлены качественные и количественные особенности физико-химических свойств и доменной организации белков хроматина человека, по сравнению с белками цитоплазмы.

Научная новизна работы

В работе предлагается классификация белков внутреннего содержимого ядра, которая, в отличии от Gene Ontology, содержит сравнительно небольшое количество терминов. В то же время предложенные термины описывают преобладающее количество процессов в ядре клетки, и, в терминах теории графов, имеет малую глубину дерева, что облегчает работу и интерпретацию результатов. С помощью разработанной классификации произведена оценка соотношений массовых долей гистоновых белков относительно разных наборов негистоновых белков, с учетом множественной локализации и мультифункциональности белков.

В работе впервые описано качественное разнообразие структур нуклеосом по части их гистоновых и негистоновых компонентов.

Теоретическая и практическая значимость работы

Настоящая работа описывает качественный и количественный состав хроматома человека, онкологические нарушения в нем и разнообразие структур комплексов нуклеосом с негистоновыми белками. Результаты расширяют и дополняют современные представления о белках хроматина и косвенно могут способствовать разработке лекарственных средств, взаимодействующих с белками хроматина человека, а также служить предпосылкой для поиска биомаркеров прогрессии онкологических заболеваний среди белков хроматина человека.

Степень достоверности и апробация результатов

Результаты данной работы были представлены в виде 5 устных и стендовых докладов на международных и российских конференциях:

1. EMBL: Chromatin and Epigenetics, Гейдельберг, Германия, 15-18 мая
2. International Symposium on Chromatin Architecture: Structure and Function 2023, Online, Япония, 24-25 января 2023
3. ИТИС(б) — Информационные технологии и системы. Биоинформатика., Огниково, Моск. обл., Россия, 18 февраля - 21 марта 2022
4. 66th Biophysical Society Annual Meeting, Online, Сан-Франциско, США, 19-23 февраля 2022
5. The 44th FEBS Congress, Krakow, Poland, July 6-11., Краков, Польша, 6-11 июля 2019

Публикации

По материалам диссертации опубликовано 4 статьи в рецензируемых журналах, индексируемых в наукометрических базах данных Web of Science, Scopus и RSCI (РИНЦ):

1. Seal R. L., Denny P., Bruford E.A., **Gribkova A. K.,** Landsman D., Marzluff W.F., McAndrews M., Panchenko A.R., Shaytan A.K., Talbert P. A standardized nomenclature for mammalian histone genes // Epigenetics & Chromatin. 2022. Т. 15. № 1. С. 34. IF 4.7
2. Armeev G. A.*, **Gribkova A. K.***, Shaytan A. K. Nucleosomes and their complexes in the cryoEM era: Trends and limitations // Front. Mol. Biosci. 2022. Т. 9. С. 1070489. IF 5.5
3. Espiritu D.*, **Gribkova A.K.***, Shubhangi G., Shaytan A.K., Panchenko A.R. Molecular Mechanisms of Oncogenesis through the Lens of Nucleosomes and Histones // J. Phys. Chem. B. 2021. IF: 2,9
4. Armeev G. A., **Gribkova A.K.**, Pospelova I., Komarova G.A., Shaytan A.K. Linking chromatin composition and structural dynamics at the nucleosome level // Current Opinion in Structural Biology. 2019. Т. 56. С. 46–55. IF: 6,9

Личный вклад автора

Автор провела весь описанный в тексте биоинформатический анализ биологических данных, сбор необходимых данных и написание программных кодов для анализа. Автору принадлежит идея, реализация и разработка программной библиотеки для применения классификации белков внутреннего содержимого ядра и хроматина.

Глава 1. Обзор литературы

В ядре клеток эукариот молекула ДНК обвивает кор белков гистонов, образуя нуклеосомы - первый уровень компактизации ДНК. Нуклеосомы, с одной стороны, препятствуют взаимодействию ДНК с узнающими ее белками, с другой стороны, способствуют протеканию биологических процессов через взаимодействия функциональных белков хроматина с гистонами и/или нуклеосомной ДНК. Далее будут рассмотрены особенности белков гистонов, строение нуклеосомы и функциональные классы белков хроматина.

1.1. Гистоны

1.1.1. Разнообразие генов и белков гистонов, варианты гистонов и канонические изоформы

Исторически, первый протокол выделения ядер из клеток из гнойных повязок раненных предложил Фредерик Мишер в 1870-ых. Также он охарактеризовал соотношение химических элементов в нуклине, подчеркнув наличие фосфора, и выделил протамины - основные белки в сперматозоидах, которые связываются с кислотой нуклином. В 1979-ом году Вальтер Флемминг с помощью окрашивания клеток анилиновыми красителями выявил вещество в ядрах, которое назвал хроматином. В 1884 Альбрехт Коссель, впоследствии получивший Нобелевскую премию по физиологии или медицине, открыл гистоны, которые по свойствам напоминали протамины Мишера. Коссель считал, что гистоны содержатся только в некоторых тканях живых организмов. 1900-ых первыми электрофоретическими методами было показано, что компонент клеточного ядра нуклеопротеин мигрирует к аноду, а изолированные гистоны движутся к катоду и, следовательно, имеют положительный заряд [1]. Настоящий прогресс в исследовании гистонов был достигнут только после разработки методов хроматографического фракционирования и гель-электрофореза в конце 1950-х годов. К 1965-му году, благодаря серии работ Э.В. Джонса по разработке методов фракционирования и идентификации гистонов, стали известны основные типы гистонов и их свойства [2].

Гистоны делятся на коровые (типы H2A, H2B, H3, H4), образующие нуклеосому, **Рисунок 1А**, и линкерные (H1 и H5 у птиц), способствующие образованию хроматосом посредством связывания с нуклеосомой в местах "входа и выхода" ДНК. Коровые гистоны состоят из глобулярного домена (три альфа-спирали, соединенных короткими петлями), который фланкирован неупорядоченными фрагментами - хвостами гистонов (N-концы у всех типов гистонов и C-концы у H2A и H2B), **Рисунок 2**. Как глобулярный домен, так и хвосты гистонов подвержены пост-трансляционным модификациям, которые будут описаны в **Разделе 1.1.3.**

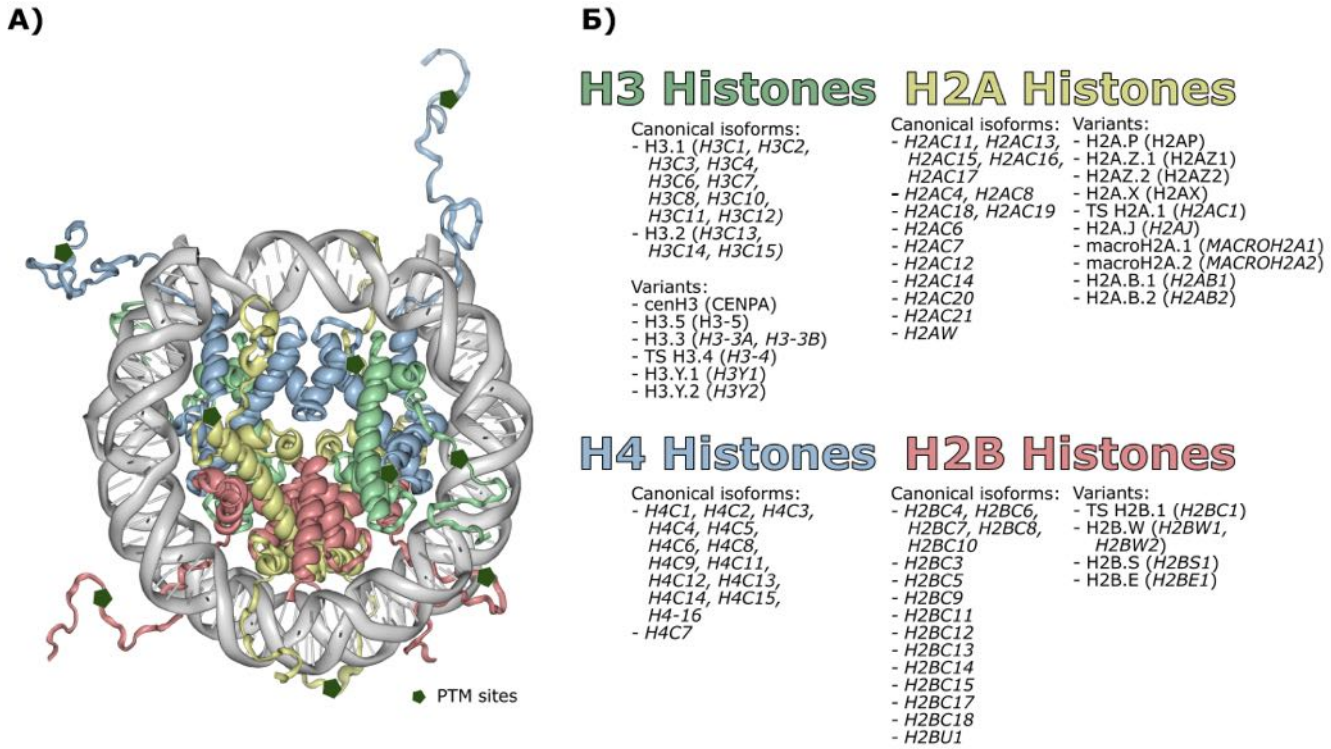


Рисунок 1. Первый уровень компактизации ДНК. А) Структура коровой частицы нуклеосомы (на основе PDB ID: 1KX5). Б) Список гистоновых генов и белков человека, классифицированных по типам, вариантам и изоформам гистонов. Каждый тип гистонов (H3, H4, H2A, H2B) содержит канонические изоформы гистонов и варианты гистонов. Названия генов выделены курсивом и сгруппированы по названиям их белковых продуктов. Адаптировано из [3].

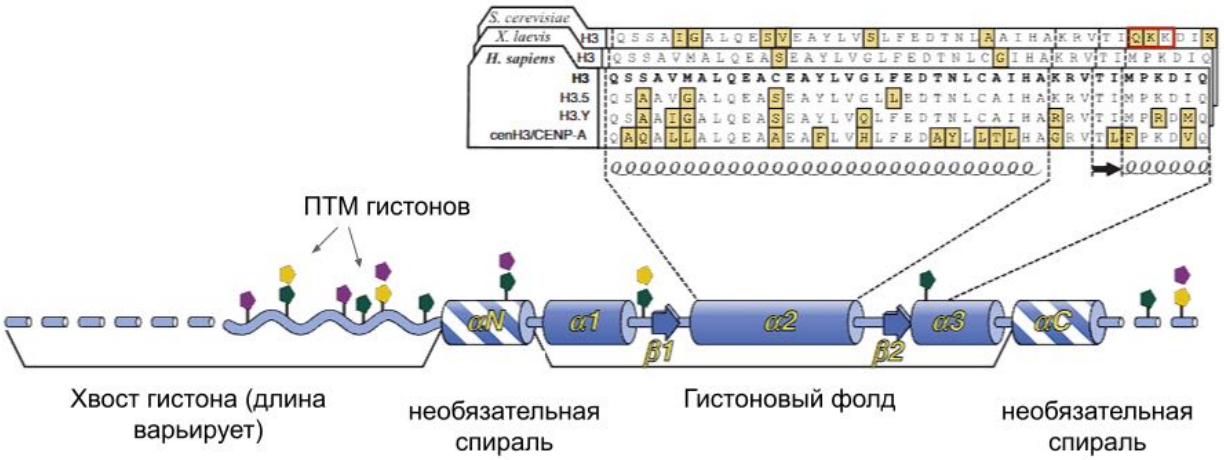


Рисунок 2. Обобщенная схема строения корового гистона. Приведено множественное выравнивание α2 спирали канонических гистонов *Xenopus laevis, S. cerevisiae, Homo sapiens* и вариантных гистонов человека (H3.5, H3.Y, cenH3/CENP-A). Отличия последовательностей от канонического гистона H3 человека, выделенные желтым, демонстрируют большую вариабельность гистонов на уровне вариантов относительно межвидовых различий. Специфический для грибов мотив QKK, обеспечивающий снижение стабильности нуклеосом, выделен красной рамкой. Адаптировано из [4].

Гистоны присутствуют во всех видах живых организмов среди эукариот (долгое время считалось, что гистонов нет у динофлагеллят, но последние исследования подтверждают их наличие [5]), в некоторых видах архей [6], вирусов [7], а в 2023 году их обнаружили даже в некоторых видах бактерий [8] (однако, компактизация ДНК в этих бактериях происходит не через образование канонической структуры нуклеосомы).

Гены гистонов делятся на зависимые от репликации и независимые, также называемые каноническими и вариантными, соответственно [9]. Гены канонических гистонов кластеризованы, у млекопитающих они находятся в четырех кластерах. Кластеры с наибольшим количеством генов находятся на шестой хромосоме (более 60-ти генов) и на первой (10-12 генов). Гены канонических гистонов имеют следующие особенности: не содержит интронов, у ряда генов соответствующую мРНК фланкирует шпилечная структура вместо поли-А-хвоста, экспрессия генов происходит в S-фазу клеточного цикла. Транскрипция и процессинг мРНК канонических гистонов происходит в тельцах гистоновых локусов, немембранных органеллах ядра.

В геноме человека 130 генов кодируют гистонов. Некоторые канонические гены гистонов с различающимися последовательностями нуклеотидов кодируют идентичные белки, что приводит к делению канонических генов гистонов на изоформы. В настоящее время известно 10 изоформ для канонических гистонов H2A, 12 - для H2B, два - для H3 и два - для H4. Все эти изоформы отличаются друг от друга несколькими аминокислотами - от двух (для H3) до восьми (для H2A), за исключением продукта гена H4C7, который отличается от основной изоформы H4 по 19 аминокислотным остаткам, **Рисунок 1Б**.

Независимые от репликации гистоны (вариантные) характерны для гистонов типа H2A, H2B и H3. Они экспрессируются на протяжении всего клеточного цикла, гены содержат интроны, а мРНК фланкирует поли-А-хвост. Для транскрипции генов вариантных гистонов не требуются тельца гистоновых локусов, поэтому они не находятся в кластерах генов гистонов. Замена канонических гистонов на вариантные проходит в ряде клеточных процессов, в частности, при клеточном делении, транскрипции, репарации ДНК и ремоделировании хроматина [10]. Варианты с небольшими изменениями канонической последовательности гистонов важны для протекания ряда биологических процессов, например среди вариантов

гистонов H#: H3.3 важен для пластичности нейронов [11], H3.5 - для сперматогенеза [12], а H3.Y изменяет регуляцию клеточного цикла [13] в ответ на внешние стимулы, **Рисунок 2**. Варианты гистонов облегчают (например, H3.3, H2A.B), репрессируют (например, macroH2A) или регулируют транскрипцию (например, H2A.Z); участвуют в репарации ДНК (например, H2A.X), функционировании центромеры (cenH3/CENP-A) и других процессах [14]. Изменение структурных свойств нуклеосомы за счет встраивания вариантов гистонов будет рассмотрено в **Разделе 1.2.1**.

Информация о последовательностях гистоновых белков, классифицированных по типам и вариантам гистонов, собрана в HistoneDB 2.0 [15]. Также HistoneDB 2.0 предоставляет структурную аннотацию последовательностей и инструменты для сравнения и анализа гистоновых белков.

1.1.2. Номенклатура генов гистонов

В данном разделе используются материалы из [16].

Открытие многочисленных вариантов гистонов в эпоху геномики привело к необходимости стандартизации их наименований. Вопросы номенклатуры генов гистонов обсуждались на семинаре EMBO по гистоновым вариантам в 2011 году [17]. В принятой тогда системе классификации в названии генов гистонов указывался геномный кластер, например, для гена гистона человека HIST1H2AA: HIST1 в начале гена означал нахождение гена в первом по количеству генов кластере. Далее следовал тип гистона (например, H2A) и уникальный буквенный идентификатор (например, A). Такая запись длиннее общепринятых наименований генов (что неудобно, по отзывам клинических сообществ) и могла быть неправильно интерпретирована неспециалистами. С другой стороны, схожая организация генов гистонов в кластеры сохраняется внутри видов млекопитающих, но различается внутри группы позвоночных животных (например, в геноме курицы только один кластер канонических гистонов) - что затрудняет перенос названия генов между видами живых организмов. В свою очередь, номенклатура гистоновых вариантов включала тип гистона, символ "F" и идентифицирующая вариант буква, принятая в сообществе исследователей. Например, H2AFZ для гистонового варианта H2A.Z. Таким образом, для описания генов семейства гистонов использовали 2 независимых номенклатуры, которые не всегда однозначно переносились на другие виды живых организмов, что и послужило началом работы над новой номенклатурой генов гистонов комитетами по номенклатуре генов человека и мыши (HUGO Gene Nomenclature Committee - HGNC, Mouse Genomic Nomenclature Committee - MGNC) с привлечением экспертизы сообщества исследователей.

В результате совместной работы комитета по номенклатуре генов человека (HGNC) и позвоночных животных (Vertebrate Gene Nomenclature Committee, VGNC) с привлечением экспертизы исследователей был опубликован новый подход к номенклатуре генов гистонов млекопитающих [16]. В новой номенклатуре сначала указывается тип гистона. Для канонических гистонов далее следует символ "C" и порядковый номер. Благодаря тому, что порядок генов в кластерах канонических генов сохраняется, кураторы комитета по номенклатуре позвоночных животных вручную присвоили генам гистонов шимпанзе, макаки-резус, собаки, кошки, свиньи, лошади и крупного рогатого скота те же названия, что и у ортологичных генов гистонов человека. Нумерация генов в кластерах не отражает порядок генов для названий среди курируемых видов живых организмов, что позволяет добавлять новые названия генов гистонов, отсутствующие у человека (например, гены H4C19 и H2BC25, которые присутствуют в геномах крупного рогатого скота, лошади и собаки и отсутствуют в геномах человека и мыши). Для обозначения варианта гистона после типа гистона указывается название варианта. В названии белка он отделяется точкой, а в названии гена принято не использовать разделитель между типом гистона и указанием варианта. При необходимости отделять числовые записи применяется дефис в систематике человека и других позвоночных и символ 'f' в систематике генов мыши (там дефис зарезервирован для специального обозначения аллелей).

Для канонических генов гистонов характерно наличие псевдогенов. Изучение организации кластеров генов гистонов у млекопитающих позволило переименовать псевдогены гистонов человека и мыши в соответствии с их белок-кодирующими ортологами. В обновленной номенклатуре для идентификации псевдогенов используют символ "P" в конце названия гена (например, H2AC10P).

Более детальные особенности номенклатуры генов гистонов человека и ряда других видов позвоночных описаны в [16]. Интерактивная таблица, включающая идентификаторы генов, транскриптов и белков гистонов человека доступна в базе HitoneDB 2.0 [15] по адресу: https://histdb.intbio.org/human/.

1.1.3. Посттрансляционные модификации гистонов

Одним из механизмов эпигенетической регуляции является добавление посттрансляционных модификаций (ПТМ) на белки гистоны. Наиболее часто встречаемые ПТМ - метилирование, фосфорилирование, ацетилирование и убиквитинилирование; эффект метки зависит от типа метки, позиции аминокислотного остатка гистона и часто от других меток на этой или соседних нуклеосомах [18]. Добавляют, считывают и стирают эти метки специализированные белки, однако не для всех меток известна вся белковая машинерия. Далее будут описаны некоторые широко распространенные и наиболее исследованные ПТМ гистонов.

К наиболее известным меткам, связанным с транскрипцией, относится триметилирование лизина 4 гистона 3 (H3K4me3). Гистоны в промоторах большинства активных генов эукариот обогащены этой ПТМ, значение обогащения которой достигает максимума в районе сайта старта транскрипции [19]. С одной стороны, H3K4me3 обеспечивает рекрутирование транскрипционной машинерии, что способствует транскрипции [15]. Однако данные функциональных экспериментов в многочисленных модельных системах свидетельствуют о том, что H3K4me3 не требуется для большинства транскрипций [21]. Напротив, локальное распространение метки H3K4me3 умеренно активирует экспрессию генов в строго контекстно-зависимой манере [22]. Таким образом, точная роль удивительно консервативного обогащения H3K4me3, наблюдаемого на большинстве активных промоторов, остается неясной.

Метками активных энхансеров являются H3K4me1 и ацетилирование гистона H3K27 (H3K27ac). H3K36me3 тесно коррелирует с активно транскрибируемыми областями благодаря рекрутированию H3K36-метилтрансферазы SETD2 РНК-полимеразой II в процессе элонгации. H3K27me3 и моноубиквитилирование гистона H2A (H2Aub) являются связанными характеристиками факультативного гетерохроматина и обусловлены различной активностью репрессивных комплексов Polycomb PRC2 и PRC1, соответственно. H3K9me3 - классическая метка конститутивного гетерохроматина, обогащена на транскрипционно молчащих участках генома. H3K9me3 связывается с HP1, что способствует компактизации хроматина за счет рекрутирования других функциональных белков хроматина (метилтрансфераз и деацетилаз) и, возможно, разделения жидких фаз.

Посттрансляционные модификации гистонов приводят к изменениям в структуре хроматина. Модификации, в частности изменяющие заряд аминокислот, могут приводить к изменению подвижности нуклеосом за счет изменения контактов между ДНК и гистонами; могут влиять и на формирование структур более высокого порядка за счет изменения контактов между нуклеосомами или могут быть прочитаны специальными эффекторными белками, которые в свою очередь запускают каскад реакций с привлечением других функциональных белков хроматина [23]. Концепция гистонового кода, разработанная Дэвидом Аллисом в 2000-ых годах предполагала, что каждая метка узнается своим доменом и несет определенный смысл [24]. Позже было показано, что одну метку могут узнавать несколько доменов, у ряда доменов нет определенного аминокислотного контекста узнавания, что привело к появлению концепции мультивалентности - то есть кооперативности поверхности домена или доменов в узнавании нескольких связанных меток гистонов [23]. Первое систематическое описание ковстречаемости ПТМ гистонов на одной нуклеосоме в цис и транс-положениях гистонов или на разных нуклеосомах, а также систематический поиск потенциальных мультивалетных

белков-считывателей ПТМ гистонов были проведены в [23], что послужило неким заделом для дальнейшего обсуждения нуклеосомного кода.

1.1.4. Онкогистоны

В данном разделе используются материалы из [3].

Мутации в генах гистонов могут приводить к нарушениям в жизнедеятельности клеток. Впервые рекуррентные мутации гистонов были обнаружены в образцах пациентов с педиатрической глиомой высокой степени тяжести (pHGG) около десяти лет назад [25,26]. В течение последующих нескольких лет были зарегистрированы многочисленные онкологические мутации в гистонах среди различных типов рака [27–31]. По последним данным, мутации гистонов могут встречаться не менее чем у ~5% онкологических больных [31]. Далее будут рассмотрены наиболее изученные и часто встречаемые в различных типах онкологий мутации в гистонах.

Замены K27M (лизина на метионин в 27-ой позиции) в гистоне H3 являются одними из первых и наиболее хорошо задокументированных гистоновых мутаций у онкологических больных [25,26]. Наличие мутаций K27M в одном гене, кодирующем канонический гистон H3.1 или гистоновый вариант H3.3, ассоциируется с глобальным снижением ди- и триметилирования (me2 и me3, соответственно) в остатках H3K27 [32,33]. Такое глобальное снижение связывают с устранением боковой цепи лизина, необходимой для метилирования H3K27 (цис-ингибирование), и инактивацией поликомб репрессорного комплекса 2 (PRC2) (транс-ингибирование). Как правило, PRC2 метилирует H3K27 за счет метилтрансферазной активности своего домена EZH2 [34]. Гидрофобные взаимодействия между замещенной боковой цепью метионина H3 и некоторыми ароматическими остатками EZH2 приводят к ингибированию активности PRC2, препятствуя метилированию H3K27 на соседних нуклеосомах [32], что также было подтверждено in vivo [35].

H3K36M - еще один пример драйверной мутации в ключевом сайте ПТМ гистона H3. Эта мутация была обнаружена в генах канонических гистонах H3 и варианте H3 при хондробластоме и плоскоклеточной карциноме головы и шеи [36,37]. Мутация приводит к глобальному снижению метилирования (ди-, три-) в этом сайте. Мутантные H3K36M мышиные модели демонстрируют тяжелую анемию с остановкой эритропоэза и быстрой летальностью, что подтверждает пагубное влияние этой мутации на развитие и пролиферацию клеток [38]. Интересно, что аминокислотные замены в H3G34 (как в канонических вариантах H3, так и в H3.3) также могут приводить к снижению уровня H3K36me3 в цис-положении, предположительно за счет введения громоздкой боковой цепи, которая не позволяет SET-домену метилтрансферазы SETD2 связываться с N-концевым хвостом H3 и метилировать H3K36

[39,40]. В свою очередь, отсутствие H3K36me3 в этих клетках может препятствовать связыванию MutSα (гетеродимера MSH2-MSH6) с N-концевым хвостом H3, что может привести к нарушению репарации несоответствий оснований и повышению частоты мутаций, наблюдаемому в клетках, мутантных по H3G34 [39]. Эффекты мутаций H3G34 могут распространяться и на изменение профилей метилирования ДНК за счет нарушения связывания ДНК-метилтрансферазы de novo DNTM3a/b через ее PWWP-домен с H3K36me3 [41].

Помимо прямого воздействия на сайты PTM, миссенс- и нонсенс-мутации гистонов способны нарушать четвертичную структуру гистонового октамера и структуру хроматина более высокого порядка. В качестве примера можно привести функциональные эффекты мутаций E76K в H2B, которые, по данным литературы, являются наиболее часто встречающимися мутациями в каноническом гистоновом гене во всех типах рака [42]. Было установлено, что эти мутации происходят одновременно с дестабилизацией нуклеосомы, предположительно за счет нарушения взаимодействий на границе H2B-H4 [42,43]. Это может приводить к релаксации структуры хроматина более высокого порядка, что способствует доступу транскрипционных факторов к множеству генов, которые обычно репрессированы [42,43]. Более того, в клетках с мутациями H2BE76K наблюдается повышенная экспрессия генов, связанных с дифференцировкой, пролиферацией, миграцией, апоптозом и клеточной сигнализацией. И наоборот, эти клетки демонстрируют пониженную экспрессию генов, связанных с биосинтезом клеток, транспортом митохондриальных мембран, гомеостазом глюкозы, ответом на факторы роста и адгезию [42]. Подобно мутациям, нарушающим структуру нуклеосом, раковые мутации в кислотном участке гистонового октамера могут непосредственно влиять на взаимодействие с другими белками хроматина [31], что, в свою очередь, может изменять структуру хроматина более высокого порядка и изменять доступность ДНК для факторов транскрипции [44].

Мутации в гистонах также могут изменять силу взаимодействия гистонов с ДНК. Например, мутации G53D в H2B были обнаружены примерно в 5% случаев протоковой аденокарциномы поджелудочной железы [45]. Остаток G53 в H2B расположен в непосредственной близости от ДНК в структуре нуклеосомы [45]. Анализы по силовому растягиванию нуклеосом при помощи оптического пинцета показывают, что мутации G53D могут ослаблять взаимодействие между гистоновым октамером и нуклеосомной ДНК [45]. Так, барьер для частичного разворачивания ДНК для нуклеосомы дикого типа составил 17,81 кДж/моль по сравнению с 9,66 кДж/моль для мутанта G53D. Кроме того, эксперименты in vitro на клетках рака поджелудочной железы показали, что H2B G53D может повышать эффективность прохождения нуклеосом РНК-полимеразой II, что, в свою очередь, может приводить к аберрантной экспрессии генов, которые обычно подавляются связыванием

гистонов [45]. Среди экспериментально охарактеризованных мутаций H2A.Z R80C, как показали исследования in vitro, снижает стабильность нуклеосом, что согласуется с ее расположением в важном сайте связывания с ДНК, где боковая цепь аргинина проникает в малую бороздку ДНК. Интересно, что аналогичные мутации также найдены и в вариантах TS H2A.1 и H2A.P.

Отдельно стоит отметить мутации в канонических изоформах гистонов. С появлением методов анализа экспрессии генов стало понятно, что канонические изоформы гистонов дифференциально экспрессируются в различных физиологических состояниях [46]. Эффекты этой дифференциальной экспрессии могут быть соотнесены со стабильностью нуклеосом. Например, замена M51L в изоформах, кодируемых генами H2AC18 и H2AC12, расположена на границе димеров H2A-H2B и снижает температуру плавления H2A-H2B примерно на 3 °C [47]. Кроме того, при экзогенной сверхэкспрессии он способствует пролиферации клеток контекстно-зависимым образом, по-видимому, за счет разрыхления структуры хроматина. Замены между каноническими изоформами играют роль в стабильности нуклеосом, взаимодействии с ДНК и регуляции доступности сайтов ПТМ [48].

Первое наблюдение, свидетельствующее о нарушении экспрессии изоформ гистонов при раке, было получено для генов H2A при хроническом лимфоцитарном лейкозе. Было показано, что в B-клетках пациентов у гена H2AC6 повышена экспрессия, а у генов H2AC4 и H2AC8 экспрессия понижена [49]. Более того, уровень экспрессии этих генов изменяется в зависимости от степени агрессивности раковых клеток. Также было обнаружено, что соотношение экспрессии изоформ H2AC6 и H2AC12 различается в нормальных и опухолевых тканях печени [47]. Аналогично, ген H2AC6 был сверхэкспрессирован в клетках эстроген-рецептор положительного рака молочной железы, где он выполняет роль медиатора эстрадиол-зависимой транскрипции онкогенов BCL2 и C-MYC [50]. Таким образом, H2AC6 может выступать в роли мастер-регулятора онкогенов. Аналогично, изоформа, которую экспрессирует ген H2AC20, является одним из основных регуляторов EGF-сигналинга при раке молочной железы. Экспрессия H2AC20 стимулируется EGF, что создает петлю положительной обратной связи, способствующую EGF-индуцированной пролиферации, росту и эпителиально-мезенхимальному переходу клеток [51]. Гистон типа H4 считается одним из наиболее консервативных гистоновых белков, так не включает гистоновые варианты. Было показано, что одна из двух его канонических изоформ, кодируемая геном H4C7, локализуется в ядрышках и усиливает транскрипцию рДНК. В раке молочной железы уровень экспрессии H4C7 коррелирует со стадией прогрессии рака, а в моделях опухолевых ксенотрансплантатов нокаут H4C7 приводит к снижению опухолеобразования [52].

Причины различной экспрессии изоформ гистонов могут быть связаны с дифференциальным метилированием ДНК соответствующих генов. Например, гиперметилирование гена H2AC8 (с меньшим уровнем экспрессии по сравнению с H2AC6 в раковых опухолях, о чем говорилось выше) коррелирует с плохим прогнозом при гепатоцеллюлярной карциноме [53], а ген H2BC11 был амплифицирован и гипометилирован в образцах метастазов рака молочной железы в мозг [54]. Появляются данные и о том, что экспрессия изоформ гистонов вносит свой вклад в терапевтический ответ: экспрессия мРНК псевдогена H2BC20P коррелирует с устойчивостью к паклитакселу в клетках трижды негативного рака молочной железы [55], уровень экспрессии H2BC21 изменяется в ответ на эпигенетическую терапию [56], а экспрессия H2BC13 снижена в клеточных линиях рака поджелудочной железы, устойчивых к гемцитабину [57].

1.2. Нуклеосома - структурная единица первого уровня компактизации хроматина

Молекула ДНК в эукариотах на первом уровне компактизации обматывает кор белков гистонов, **рисунок 1А**. При сборке нуклеосомы димеры H3 и H4 формируют тетрамер, с которым соединяются два димера H2A-H2B. Нуклеосомная ДНК формирует левозакрученную спираль, которая ~1.65 раз обвивает нуклеосомный кор. На более высоких уровнях компактизации в месте "входа и выхода" ДНК с нуклеосомой соединяется гистон H1 (или H5 у птиц).

Нуклеосомная ДНК содержит 14 сайтов связывания с гистонами, расположенных в малой бороздке. Преимущественно, эти контакты представлены водородными связями между гистонами и сахарофосфатным остовом нуклеиновой кислоты. Кроме водородных связей, взаимодействие ДНК и белков проявляется в гидрофобных взаимодействиях и солевых мостиках. В гистоновом октамере выделяют участок, называемый «acidic patch», который состоит из 8-ми отрицательно заряженных аминокислотных остатков, расположенных на гистонах H2A (E56, E61, E64, D90, E91, E92) H2B (E102, E110). С кислотным лоскутом взаимодействуют: хвост гистона H4 соседней нуклеосомы при формировании хроматиновых фибрилл; различные функциональные классы белков хроматина посредством "аргининовых якорей" [58]; некоторые белки вирусов (например, LANA герпесвируса, ассоциированного с саркомой Капоши; IE1 цитомегаловируса человека) [59].

1.2.1. Влияние вариантов гистонов на структуру и динамику хроматина

Раздел написан по материалам [4].

Варианты гистонов могут изменять структурные свойства нуклеосом, приводить к обертыванию меньшего или большего количества пар оснований ДНК, изменять стабильность

нуклеосом [60]. Например, cenH3 (предыдущее название CENP-A) (вариант, играющий ключевую роль в формировании центромер во время клеточного деления) имеет более короткую спираль aN, по сравнению с каноническим H3. Эта особенность обусловливает высокую гибкость ДНК-линкеров в нуклеосомах, содержащих cenH3, и изменяет связывание гистонов с линкерной ДНК [61]. Кроме того, cenH3 влияет на отворачивание ДНК от кора нуклеосомы и способствует петлеобразованию [62].

В последние годы было показано, что небольшие вариации последовательностей вариантов гистонов могут оказывать влияние на стабильность и динамику нуклеосом. Например, специфический для H3.5 остаток лейцина 103 дестабилизирует нуклеосому, что важно для процесса транскрипции в клетках семенников человека [12]. Как показано в работах [12,63], метионин 124 в H3.Y способствует стабильной ассоциации тетрамера H3.Y-H4 с ДНК. Другой специфический для H3.Y остаток - лизин 42 - играет ключевую роль изгибании линкерной ДНК. Более того, вариации в канонических изоформах гистонов человека, также могут функционально влиять на стабильность нуклеосом. Вариации M51L и K99R в гене гистона H2AC12 (относительно других канонических генов гистонов H2A) приводят к стабилизации нуклеосом и изменяют пролиферацию клеток [47]. Аналогичным образом в настоящее время описаны эффекты небольших вариаций между каноническими последовательностями у разных видов. Специфический для грибов мотив гистона H3 QKK **Рисунок 2**, расположенный на оси диады нуклеосом, ухудшает сборку октамеров в клетках дрожжей [64].

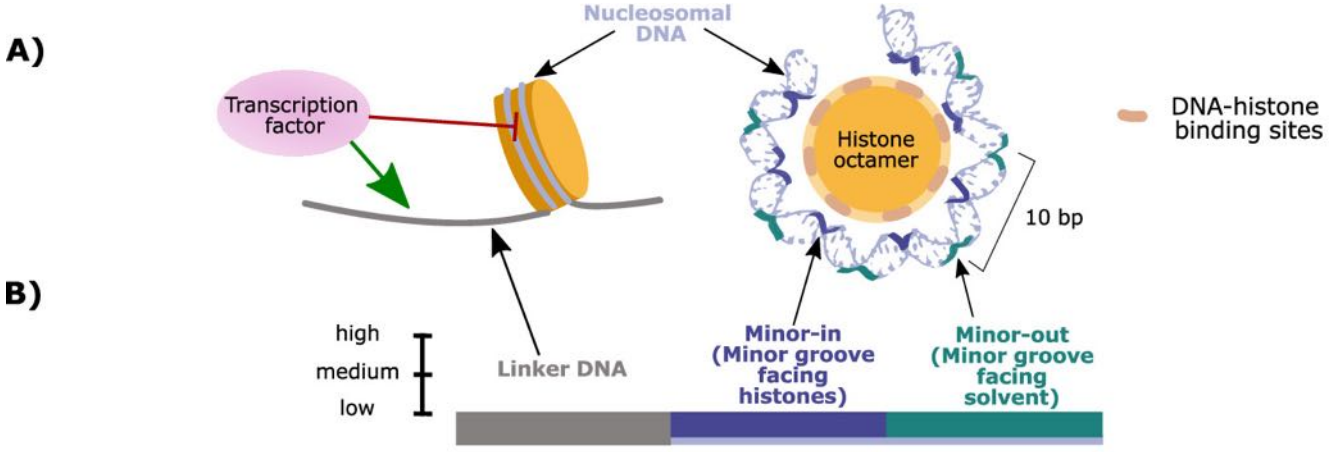
1.2.2. Влияние доступности ДНК в нуклеосомах на мутагенез

Раздел написан по материалам [3].

Оборачивание ДНК вокруг нуклеосомного кора уменьшает доступность ДНК и стерически затрудняет взаимодействия с функциональными белками хроматина [65]. Исследования *in vitro* [66] и на уровне генома [65] показывают, что нуклеосомы препятствуют связыванию транскрипционных факторов (ТФ) с ДНК, **Рисунок 3А**. Однако особенности структуры нуклеосом могут модулировать эти взаимодействия, в частности, для узнавания нуклеосомной ДНК пионерными факторами транскрипции.

Геномные исследования клеток рака молочной железы показали, что детали позиционирования нуклеосом имеют решающее значение для взаимодействия рецепторов прогестерона с их сайтами связывания [67]. Аналогичным образом, перераспределение нуклеосом широко распространено на ранних стадиях прогрессии аденокарциномы легкого и толстой кишки и потенцирует связывание факторов хроматина. Было высказано предположение, что изменение распределения нуклеосом по геному может служить ранним маркером этих

онкологических заболеваний [68]. Было показано *in vitro*, что разворачивание ДНК облегчает связывание транскрипционных факторов [66], что свидетельствует о возникновении нуклеосомно-опосредованной кооперативности между связыванием нескольких ТФ одновременно [69]. Хотя описанное явление и не единственный механизм онкогенеза, вполне вероятно, что измененные состояния позиционирования нуклеосом в раковых клетках нарушают кооперативность взаимодействия ТФ, которые важны для прогрессии рака. В частности, анализ экспрессии ТФ в различных типах рака позволяет предположить, что потеря кооперации между различными транскрипционными факторами является общей чертой раковых клеток [70].



	UV-light (CPD, 6-4PP)	Smoking (BPDE-dG)	ROS (8-oxo-G)	Spontaneous deamination 5meC
Mutation type	C>T	C>A	T>C|G	C>T
DNA damage rate			?	
DNA repair rate	NER	NER	BER	MMR ? ?
Mutation rate			?	

Рисунок 3. Влияние доступности ДНК в нуклеосомах на мутагенез и связывание белков с ДНК. А) Схема кора нуклеосомы (оранжевый цвет), показывающая расположение линкерной и нуклеосомной ДНК (слева), показана периодичность (10-11 п.н.) во взаимодействии нуклеосомной ДНК с гистоновым октамером (справа).
Б) Диаграмма, качественно связывающая положение ДНК в нуклеосоме, вероятность повреждения ДНК, репарацию и результирующую частоту мутаций. Рассмотрены четыре типа мутагенных процессов/агентов: ультрафиолетовое облучение, приводящее к образованию циклобутан-пиримидиновых димеров ДНК (CPD) и (6-4)пиримидин-пиримидоновых фотопродуктов (6-4PP); курение табака, приводящее к осаждению аддуктов бензо[а]пирен-7,8-диол-9,10-эпоксида на гуанинах (BPDE-dG); активные формы кислорода (ROS) повреждают гуаниновые основания с образованием 8-гидроксидезоксигуанозина (8-oxo-G); спонтанное деаминирование 5-метилцитозинов (5mC) в контексте CpG. На диаграммах показаны качественные относительные показатели (низкий, средний, высокий) повреждения, репарации и накопления мутаций в различных участках нуклеосомной и линкерной ДНК. Системы репарации ДНК, участвующие в восстановлении соответствующих повреждений, сокращенно обозначены как: эксцизионная репарация нуклеотидов (NER), эксцизионная репарация оснований (BER) и репарация ошибочно спаренных нуклеотидов (MMR). Адаптировано из [3].

Мутагенез ДНК - еще один ключевой процесс, на который влияет изменение доступности и динамики ДНК в нуклеосомах. Мутации в клетке возникают в результате взаимодействия процессов повреждения и репарации ДНК. Доступ белков механизмов репарации ДНК может быть ограничен присутствием нуклеосом, что может приводить к более высокой эффективности репарации в линкерной ДНК по сравнению с нуклеосомной [71], **Рисунок 3А**. Эти эффекты проявляются в основном для систем репарации одноцепочечных повреждений ДНК, таких как эксцизионная репарация оснований (BER), эксцизионная репарация нуклеотидов (NER) и репарация ошибочно спаренных нуклеотидов (MMR) [72]. Однако периодичность вращения спирали нуклеосомной ДНК относительно поверхности гистонового октамера по-разному влияет на распознавание сайтов повреждения ДНК в различных позициях. Из экспериментов по футпринтингу ДНК известно, что цепи ДНК демонстрируют характерную периодичность в 10-11 п.н. с точки зрения склонности к расщеплению ДНК: меньшая интенсивность расщепления наблюдается в местах контакта цепи ДНК с гистоновым октамером и более выраженная интенсивность расщепления - в обращенных к растворителю участках ДНК [73]. Однако большинство экспериментальных данных о мутационных процессах объединены по обеим цепям двойной спирали ДНК. В этом случае участки вдоль двойной спирали нуклеосомной ДНК лучше различать по положению малой бороздки (внутри или снаружи, minor-in и minor-out, соответственно), **Рисунок 3А**.

Эксперименты на дрожжах и клеточных линиях фибробластов человека, в которых сравнивали возникновение мутаций ДНК сразу после и через несколько часов после воздействия мутагена, показали, что BER и NER проявляют меньшую эффективность в minor-in положении [74,75], **Рисунок 3Б**. Вероятно, процесс репарации ошибочно спаренных

нуклеотидов протекает по аналогичной схеме, что требует дальнейшей экспериментальной проверки.

Однако, на скорость мутаций влияет и скорость процессов повреждения ДНК.

Наиболее распространенный мутационный процесс связан с деаминированием 5-метилцитозина (5meC), приводящим к переходу цитозина в тимин (C > T). Для протекании этой реакции необходимо, чтобы, благодаря подвижности ДНК, молекулы воды получили доступ к месту реакции. Таким образом, уменьшение подвижности ДНК в нуклеосоме защищает ДНК от мутаций данного типа [76]. Несмотря на снижение репарационного потенциала ДНК в коре нуклеосомы, мутации C > T наблюдаются чаще в линкерных сегментах ДНК, по сравнению с ДНК кора нуклеосомы [77], **Рисунок 3Б**.

Образование мутаций, ассоциированных с курением табака, которые возникают в результате образования бензо[a]пирен-диол-эпоксид-производного гуанинового аддукта (BPDE-dG), ингибирует оборачивание нуклеосомной ДНК вокруг октамера гистонов [78].

Считается, что белки защищают ДНК от повреждения активными формами кислорода (АФК, ROS), однако при взаимодействии гистонов с ROS образуются белковые (пероксильные) радикалы, которые в свою очередь могут окислять связанную ДНК [79]. В результате анализа сигнатур мутаций было показано, что повреждение ROS приводит к большему количеству мутаций в положении minor-in по сравнению с положением minor-out, что связано с изменением эффективности BER-репарации [80], **Рисунок 3Б**. В случае другого типа мутагена, УФ-света, было показано различие в скорости повреждения ДНК в различных участках нуклеосомной ДНК, где образование циклобутан-пиримидиновых димеров (ЦПД) и (6-4) пиримидин-пиримидоновых фотопродуктов (6-4ПП) было более вероятным в участках нуклеосомной ДНК в положении minor-out, по сравнению с положением minor-in или линкерной ДНК [80]. Это, в свою очередь, приводит к увеличенной частоте мутаций нуклеосомной ДНК в положении minor-out, по сравнению с ДНК в положении minor-in или линкерной ДНК [74], **Рисунок 3Б**. Описанные факты однозначно указывают на то, что трансляционное и ротационное позиционирование нуклеосом влияет на процессы образования соматических мутаций в ходе опухолеобразования и прогрессии рака [75,81].

2.3. Белки хроматина и подходы к их классификации

Не смотря на то, что первые белки хроматина (гистоны и протамины) были описаны еще в конце 19-го века, на протяжении последующего столетия развитие биохимических методов не позволяло разделить, очистить и охарактеризовать другие функциональные группы белков хроматина.

Одними из первых негистоновыми белками хроматина были описаны белки группы HMG (High Mobility Group), которые были выявленные за счет небольших размеров и относительно высокой подвижности в полиакриламидном геле [81]. В 70-80-ых годах прошлого века также были описаны структурные белки, остающиеся в комплексе с ДНК даже при промывании сильными солями [82], и белки транскрипции - РНК полимеразы и факторы транскрипции [83,84]. Среди основных функциональных классов белков хроматина можно выделить следующие. Гистоновые шапероны обеспечивают транспорт гистонов, их хранение, формирование и разборку нуклеосом. Ремоделеры хроматина обеспечивают перемещение нуклеосом вдоль ДНК за счет энергии АТФ. Их разделяют на следующие классы: SWI/SNF, ISWI, CHD и INO80 (по типу АТФазной субъединицы). Другую крупную группу образуют белки, наносящие и удаляющие ПТМ, а также белки, отрезающие гистоновые хвосты. Считывание ПТМ гистонов приводит к привлечению других белков, в том числе транскрипционных факторов, транскрипционной и репликационной машинерии. Нанесение и считывания модификаций ДНК осуществляют отдельные группы белков.

Первые попытки классифицировать белки ядра были предприняты ван Хольдом в 1989 году [82]. Он разделил белки ядра на 3 основные группы - гистоновые и негистоновые белки хромосомные белки, нехромосомные белки (белки ядерной оболочки, ламины и др.). В свою очередь негистоновые хромосомные белки делились на 5 групп - ферменты, связывающиеся с хроматином; белки, регулирующие транскрипцию; белки-рецепторы и белки-гормоны; белки группы HMG и белки хромосомного каркаса (в понимании тех времен). Однако, ван Хольде подчеркивал, что четкого определение негистоновым хромосомным белкам дать невозможно, эта фракция зависит от способа и химических условий выделения хроматина. Тем не менее, существует значительное количество негистоновых белков, которые остаются связанными даже при выделении хроматина радикальными методами. Ван Хольде также вывел соотношение компонентов хроматина по массе, на основе анализа литературы, согласно которому масса ДНК и гистонов соотносится как 1:0.99, а масса ДНК и негистоновых белков как 1:0.3-0.8 [83].

На сегодняшний день термин негистоновые белки хроматина подвергается критике [84]: если понимать его в широком смысле, то он должен включать все белки ядра, кроме гистонов - то есть структурные, РНП, все ферменты ядра (ДНК- и РНК-полимеразы, ДНК-топоизомеразы, ферменты, осуществляющие ремоделирование хроматина, ферменты репарации ДНК, ферменты сплайсинга), все белки, участвующие в регуляции транскрипции и репликации и т. д. А в узком смысле: только белки, формирующие хроматиновую фибриллу (HMG, HP1, Polycomb, белок MENT, MeCP2).

Среди хроматиновых или эпигенетических баз можно выделить: Epifactors [85], которая включает в белки, задействованные в эпигенетических процессах, их субстраты и комплексы, а

также длинные некодирующие РНК и гистоны, всего 902 белка. Ряд баз данных FACER [86], CRdb [87], Cistrome [88], базы нарушений хроматиновой регуляции в различных патологиях, в частности при онкологических нарушениях: CR2Cancer [89]. Среди более частных баз - базы данных гистонов и гистоновых вариантов, в частности HistoneDB 2.0 [15], HISTome2 (также включает модификаторы гистонов) [90]. Базы транскрипционных факторов TRRUST [91], HOCOMOCO [92], база ремоделеров хроматина семейства SWI/SNF Infobase [93], база данных белков, ассоциированных с метильными и ацетильными метками: WERAM [94] и др.

Самым крупным ресурсом с функциональной аннотацией генов/белков является Gene Ontology (GO) [95]. GO представляет собой ациклический граф, в котором узлы - это термины, а ребра - отношения между ними. Наиболее часто встречаемые отношения - это (is a); часть (part of, has part); регулирует, отрицательно регулирует и положительно регулирует. GO состоит из трех частей - биологические процессы (BP), молекулярные функции (MF) и клеточная локализация (CC). Для хроматина в части CC есть одноименный термин хроматин (GO:0000785), в который входят термины как отдельных комплексов, например, FACT (GO:0035101), ASTRA (GO:00702209), HDA1 (GO:0070823) и др., так и термины функций, например, активность РНК полимераз, репрессия транскрипции (GO:0106250). В части BP есть термин организации хроматина (GO:0006325). Интересно, что термины хроматин, организация хроматина и термины функциональных белков хроматина (например, модификаторы гистонов) не связаны ребрами в графе GO. К другим ограничениям GO можно отнести [96]:

- Ограниченный набор генов (белков), для которых есть экспериментальная проверка (на январь 2015 у 57% белков человека аннотация была подтверждена экспериментально).
- Смещение в сторону активно исследуемых генов и белков.
- Систематические ошибки, накапливающиеся при распространении информации в различных типах свидетельств и между версиями.
- Общий и стандартизированный подход к классификации, при котором теряются детали о более частных и варьируемых группах белков.

Таким образом, для описания белков ядра и хроматина есть или очень специализированные ресурсы и базы данных, или очень общий и не учитывающий специфику хроматина GO.

Глава 2. Материалы и методы исследования

2.1. Источники информации о локализации белков (UniProt, HPA, OpenCell)

Накопление экспериментальных данных о локализации белков, в частности, полученных методами скрининга генных ловушек и иммунофлуоресцентного анализа, в конце 20-го века привело к созданию ряда баз данных о локализации. В начале 2000-ых к популярным базам данных локализации белков относились, в частности, The Nuclear Protein Database (NPD) [97], LOCATE [98], база ChromDB с фокусом на белках растений [99], LocDB [100]. В 2004-ом году был основан всеобъемлющий ресурс о белках UniProt, в котором также есть раздел о клеточной локализации, который стал одним из основных ресурсов информации о белках [101].

Словарь клеточной локализации UniProt имеет древовидную структуру и включает 561 термин. Каждому термину соотнесен код свидетельства (ECO), то есть источник информации о локализации белка. Наибольшее количество терминов локализации подтверждается экспериментальными данными (ECO:0000269). Среди кодов есть схожесть последовательностей (ECO:0000250), заключение куратора (ECO:0000305), заявление автора без прослеживаемой поддержки (ECO:0000303), однако у ряда тегов локализации отсутствуют какие-либо коды свидетельств.

Более поздние протеомные инициативы, например, Human Protein Atlas (HPA) [102] и OpenCell [103], предоставляют непосредственно экспериментальные данные о локализации белков.

Авторы HPA провели иммунофлуоресцентный анализ с дальнейшим анализом изображений конфокальной микроскопии для 13 тысяч белков человека с использованием 35 клеточных линий. Словарь терминов локализации HPA включает 30 терминов для 13 клеточных органелл. В HPA также есть 4 кода свидетельства локализации, основанные на количестве связавшихся антител, совпадений с литературными данными и уровне экспрессии белка.

Иммунофлуоресцентный анализ проводят на фиксированных клетках, поэтому авторы OpenCell предложили новую методику для определения клеточной локализации белков в живых клетках. Схема эксперимента в OpenCell следующая - в ген белка интереса с помощью системы CRISPR-Cas встраивают фрагмент mNG11 флуоресцентного белка split-mNeonGreen$_2$. Такой эксперимент удалось провести только для 1600 белков, которые, однако, находятся среди 50% наиболее представленных белков человека. В качестве свидетельства достоверности локализации белка авторы используют 3 оценки - выраженная локализация, менее выраженная локализация и слабые паттерны, близкие к пределу обнаружения.

2.2. Протеомные подходы для исследования белков хроматина

Описание белкового состава хроматина, то есть хроматома, способствует идентификации белков, связанных с регуляцией генов, а также их последующему изучению в контексте различных заболеваний и разработки лекарственных препаратов. Сравнительный анализ хроматомов может способствовать выявлению ассоциация между состоянием клеток и белковым устройством хроматина.

2.2.1. Экспериментальные техники выделения белков хроматина

Перед проведением протеомных экспериментов необходимо выделить белки хроматина из клеток. Однако эта процедура осложняется физическими свойствами хроматина: хроматин имеет диффузный размер и плотность, что затрудняет его отделение от других органелл; хроматин не отделен мембраной и подвержен загрязнению белками других компартментов; один из способов взаимодействия белков с ДНК - электростатический, то есть хроматин - заряженный полимер, нерастворимый в солях с низкой ионной силой. Свойства хроматина явились предпосылкой для разработки различных методик для его выделения. Обобщенная схема выделения хроматина включает в себя следующие этапы: лизис клеток, центрифугирование, выделение фракции ядер (осадок), пермеабилизация ядерной мембраны, промывки в солях, буферах, центрифугирование, выделение фракции хроматина (осадок).

В начале 20-го века развивались протоколы выделения ядер и хроматина по большей части для исследования роли ДНК и белков в механизмах передачи наследственной информации. Предлагались разные протоколы экстракции нуклеогистонов из лизированных клеток, в частности, использование 1M NACl, что приводило к диссоциации гистонов и других белков с ДНК [104]. Несмотря на то, что часть белков снова связывалась с ДНК при уменьшении концентрации соли до 0.14M, большая часть информации была утеряна. Важной вехой является подбор более "мягких" условий для выделения хроматина [105,106] в 1950-1960-ых годах, который предполагал уменьшение концентрации соли в растворе до 0.075M NaCl, добавление ингибиторов ферментов и ЭДТА для хелатирования двухвалентных ионов, активирующих нуклеазы. В тоже время развитие методов хроматографического фракционирования и гель-электрофореза, в том числе двумерного, привело к появлению первых оценок качественного состава белков хроматина.

На сегодняшний день развитие методов экстракции хроматина позволяют выделять как белки хроматина в целом, так и отдельные фракции, например транскрипционно-активный и неактивный хроматин, новосинтезированный хроматин, белки, связанные с конкретными геномными локусами и др.. Далее будут рассмотрены методики выделения всех белков хроматина для последующего анализа методами протеомики.

В первых работах (начало 2000-ых) фракцию белков хроматина получали общей экстракцией (далее некоторые выжимки из протокола, полный протокол в оригинальной публикации [107]): после промывки клеток в стандартном фосфатном буфере (PBS) и буфере с солью низкой ионной силы (10 mM KCl, 1.5 mM $MgCl_2$), добавляли Triton X-100 для пермеабилизации клеток, фракцию ядер (осадок) получали центрифугированием, осадок промывали в буферах с 10 mM KCl, ЭДТА и ингибиторами протеаз, далее центрифугирование и растворение осадка (хроматин) в буферах с SDS. Фракционирование клеток позволяет анализировать фракции белков ядра и цитоплазмы, что важно в контексте исследования белков с множественной локализацией, которые при определенных условиях могут функционировать в другом клеточном компартменте. Однако, в результате использования методов фракционирования не удается избежать загрязнения образца хроматина цитоплазматическими белками [108].

Существует методика солевой экстракции хроматина, в которой клетки сначала ресуспендируют в гипотоническом лизирующем буфере (например, 10 mM KCl), а полученный после осадок ядер ресуспендируют в буфере с высокой концентрацией соли (например, 420 mM KCl) с дальнейшим снижением концентрации соли путем x-кратного разбавления [109].

Выделение белков хроматина, как полной фракции, так и отдельных фракций эу- и гетерохроматина, возможно с обработкой образца микрококковой нуклеазой (МНКаза) в различных количествах и в различные моменты времени [109–111]. Метод разделения солями (chromatin-enriching salt separation, CHESS) также позволяет разделять фракции белков хроматина последовательным изменением концентрации соли: 150 mM NaCl для белков нуклеоплазмы, 250 mM NaCl для белков эухроматина, 600 mM NaCl для белков гетерохроматина [112].

Для уменьшения загрязнения цитоплазматическими белками в протоколы выделения хроматина был добавлен этап фиксации белков и ДНК формальдегидом. Одним из таких протоколов является обогащение хроматина для протеомики (Chromatin Enrichment for Proteomics, ChEP) [113], где клетки до выделения ядер фиксируют формальдегидом, а осадок ядер после центрифугирования обрабатывают 4% ионным детергентом SDS и 8M мочевиной. Авторы провели 35 экспериментов в различных биохимических и биологических условиях для определения вероятностной оценки функционирования белка в хроматине. Данная методика применялась для выделения белков хроматина из клеток и клеточных линий разных организмов, в частности, человека [108,114], курицы [115], возбудителя малярии [116], мыши [117] и др. Однако также было показано, что данная методика приводит к загрязнению фракции белков хроматина белками митохондрий [118].

Другой разновидностью методики экстракции белков хроматина с фиксацией белков и ДНК является метод обогащения на основе плотности для МС-анализа (density-based enrichment for MS analysis of chromatin, DEMAC) [119], где после выделения ядер центрифугированием проводят фиксацию ДНК и белков, обработку ультразвуком. Далее разделяют фракции свободных белков, нуклеиновых кислот и комплексов ДНК-белок путем ультрацентрифугирования в градиенте CsCl на протяжении 48 часов. С помощью разработанного метода авторы описали изменения белкового состава хроматина, происходящие в ходе клеточного цикла и выявили степень сохранения белков-регуляторов в ходе митоза.

В 2021-ом году был предложен метод Hi-MS [120], где подготовка белков хроматина для масс-спектрометрии напоминает протокол Hi-C: хроматин обрабатывают формальдегидом, расщепляют эндонуклеазой рестрикции *HaeIII*, к липким конца ДНК пришивают биотин, после обработки ультразвуком вытягивают комплексы ДНК с белками магнитными шариками со стрептавидином. Авторы подчеркивают, что в отличии от ChEP и DEMAC данная процедура является относительно щадящей и позволяет сохранить большее количество белков, ассоциированных с хроматином. Основной целью исследования являлась оценка способности белков хроматина разделять жидкостные фазы, поэтому авторы включили этап обработки клеток 1,6-гександиолом для получения соответсвтующей количественной оценки (1,6-гександиол нарушает гидрофобные взаимодействия в конденсатах, приводя к их диссоциации).

2.2.2. Методы протеомики для исследования белкового состава хроматина

Первые эксперименты по описанию разнообразия белкового состава хроматина были проведены в 1970-1980-ых годах. В 1972-ом году благодаря разделению белков хроматина с помощью ионообменной хроматографии была получена первая количественная оценка разнообразия белков хроматина: 10-15 белков могут составлять порядка 70% негистоновых белков хроматина [121]. Несколькими годами позже методом двумерного гель-электрофореза было получено порядка 450 белков хроматина [122] и порядка 1200 белков методом изоэлектрического фокусирования [123], однако в полученном наборе белков могли быть и разные изоформы одних белков, и модифицированные белки и продукты деградации.

После 2000-ых годов подавляющее большинство исследований белков хроматина проводится масс-спектрометрическим анализом (МС), в котором измеряется отношение массы к заряду фрагмента молекулы. Восходящий (bottom-up) дизайн предполагает в общем случае: разделение белков хроматина и количественный анализ одномерным или двумерным гель-электрофорезом, расщепление белков трипсином до пептидов (расщепляет карбоксильные связи лизина и аргинина) [124], разделение пептидов ионообменной хроматографией и

непосредственно анализ методом МС или методом жидкостной хроматографии и тандемной масс-спектрометрии (LC–MS/MS).

К разновидностям МС относится тандемная МС, в которой после первого масс-анализатора ионы анализируемых пептидов фрагментируются путем соударения с молекулами инертного газа или лазера и анализируются во втором масс-анализаторе.

2.2.3. Оценка количественного состава белков в клетке

Для сравнительного анализа двух наборов белков применяют мечение изотопами. Например, в методике изотопных меток (Isotope Coded Affinity Tag, ICAT) [125] цистеины белков модифицируют меткой, состоящей из: тиол-реагирующей группы (для присоединения к цистеину), изотопно-меченого линкера и биотина для последующего выделения модифицированных пептидов. Белки в двух группах сравнения обрабатываются легкой (^{1}H) или тяжелой (восемь ^{2}H), смешивают и анализируют вместе. В методике стабильного изотопного мечения аминокислотами в культуре клеток (Stable Isotope Labeling by Amino acids in cell Culture, SILAC) [126] в питательную среду, на которой выращивают одну из сравниваемых культур клеток, вводят аминокислоты с мечеными атомами, например, гидрохлорид аргинина с ^{13}C6, ^{15}N4 и гидрохлорид лизина с ^{13}C6 и ^{15}N2. Далее культуры клеток также смешивают и анализируют вместе. Как и в методе ICAT, количественная оценка белка достигается сравнением интенсивностей пептидов с легкими и тяжелыми аминокислотными остатками.

Кроме исследования качественного состава белков и их относительного количества между образцами, методы МС позволяют также оценить количество белка в абсолютных величинах: молярную концентрацию или количество копий на клетку. Для этого используют различные подходы, например: методы соотнесения суммарного сигнала МС с визуализацией структур в клетке методом криоэлектронной томографии [127]; метод абсолютного количественного определения на основе интенсивности (iBAQ), в котором сумму интенсивностей пиков всех пептидов, соответствующих определенному белку, делят на число теоретически наблюдаемых пептидов [128]; или с использованием эталона в виде эндогенных белков, количественно определяемых по точно охарактеризованным меченным изотапами пептидам [129]. В определении абсолютного количества белков относительно данных по известному референсу ограничивающим фактором является точное измерение концентрации референсных белков.

В работе [130] было продемонстрировано выполнение следующего соотношения, **формула 1:**

$$\frac{\text{сигнал МС белка}}{\text{общий МС сигнал}} \approx \frac{\text{масса белка}}{\text{общая масса белков}} \quad (1)$$

Концептуально, общая масса белков может быть и количеством белка в данном объеме, и одним граммом). В работе [131] было предположено, что, **формула 2**:

$$\frac{\text{масса гистонов}}{\text{общая масса белков}} \approx \frac{\text{сигнал МС гистонов}}{\text{общий МС сигнал}} \approx \frac{\text{масса ДНК}}{\text{масса белков клетки}}, \quad (2)$$

основанием для которой является оценка отношения массы гистонов к массе ДНК как 1:1 [83]. Такое предположение не требует дополнительного подсчета клеток в образце и определения концентрации референсных белков или пептидов; масса ДНК в диплоидной клетке человека составляет 6.5 пг. Оценка авторов массы белков клетки по 4 клеточным линиям различается в 1.24 ± 0.29 раза по сравнению с оценкой, сделанной методом подсчета клеток. Метод оценки количества копий белков на клетку через отношение МС интенсивностей гистонов назвали «протеомной линейкой» (proteomic ruler). Также авторы продемонстрировали, что использование стандартной библиотеки пептидов для идентификации белков, по сравнению с библиотекой, в которой есть пептиды гистонов с различными комбинациями ПТМ, пренебрежимо ухудшает результат (изменяется только относительное количество гистона H3 на 5-10%). Таким образом, количество копий белка на клетку можно оценить следующим образом, **формула 3:**

$$\text{Кол} - \text{во копий белка на клетку} = \frac{\text{МС сигнал белка}}{\text{общий МС сигнал}} \times \frac{N_A}{M} \times \text{масса белков клетки} =$$

$$= \text{МС сигнал белка} \times \frac{N_A}{M} \times \frac{\text{Масса ДНК}}{\text{МС сигнал гистонов}}, \quad (3)$$

где: N_A - постоянная Авогадро, M - молярная масса белка.

2.2.4. База PAXdb - унифицированные данные о представленности белков в клетке

Накопление и унификация информации о представленности белков в клетках - одна из целей базы PAXdb [132]. В базе уделяется внимание одинаковому процессингу протеомных данных для сравнения и сопоставления информации отдельных экспериментов. В качестве единицы измерения представленности белков используется метрика parts per million (ppm). ppm - описывает представленность каждого белка относительно всего экспрессируемого протеома и. Метрика не зависит от размера клетки, и кроме того, ее определение позволяет сравнивать произвольные внеклеточные структуры, объемы или разведения. Метрика сопоставим между образцами тканей и клеточных культур, а также между различными модельными организмами с существенно отличающимися размерами клеток и структурой тканей.

В случае биохимического, биофизического или количественного МС анализа без использования меток значение ppm рассчитывают напрямую, путем пересчета предоставленных авторами оценок по их суммарному значению. Если данные значений спектров МС, то подсчет производится следующим образом. Сначала оценивают каждый ожидаемый пептид в белке по

его предполагаемой вероятности обнаружения в зависимости от его длины (было показано, что эта вероятность в настоящее время относительно одинакова для различных организмов и масс-спектрометров [133]). Затем рассчитывают фактическое покрытие пептидами каждого белка (неоднозначные пептиды учитываются дробно для каждого совпадающего белка) и нормируют это количество на ожидаемое покрытие пептидами белка. Наконец, все спектральные показатели организма суммируются и представленность нормируется на эту сумму.

Авторами PAXdb также был разработана оценка качества данных. Предпосылкой для нее ее является предположение, что у функционально связанных белков должны быть похожие уровни экспрессии. Для таких пар белков (с оценкой взаимодействия >0.9 по базе STRING) рассчитываются соотношения представленности. Чем ближе медиана таких соотношений к единице, тем лучше согласованность данных в наборе. Чтобы преобразовать этот показатель в понятную и непротиворечивую оценку, авторы также вычисляют медиану для того же набора данных после перестановки значений представленности; это делается несколько сотен раз, и фактическая медиана сравнивается с распределением рандомизированных медиан по Z-оценке. Такая метрика будет являться оценкой согласованности взаимодействий. Далее на основе индивидуальных оценок согласованности взаимодействий наборов данных для каждого организма рассчитывается интегрированный набор, который соответствует средневзвешенным значениям представленности белков из индивидуальных наборов. В интегрированном наборе данных белкам, которые не были обнаружены, присваивается нулевое значение. При расчете средневзвешенной величины решение о том, какой вес придать каждому набору данных, принимается вручную (для некоторых наборов данных он может быть и нулевым). Сначала набору данных, получившему наилучшую оценку, присваивается вес 1.0, затем для второго лучшего набора данных выбирается вес, максимизирующий оценку для полученной взвешенной комбинации. Эта процедура повторяется до тех пор, пока добавление еще одного набора данных не перестанет увеличивать общую оценку интегрированного набора данных. Иногда добавление какого-либо набора данных не повышает общий балл, но позволяет получить дополнительные белки и тем самым увеличить общий охват. В этом случае он включается, если его качество признано приемлемым. В целом назначение весовых коэффициентов, по заявлениям авторов, является в некоторой степени произвольным.

Интегрированный набор представленности белков человека в PAXdb рассчитан по 175 датасетам разных тканей и клеточных линий. Он включает в себя 19566 белков, что на 99% покрывает белки референсного протеома человека UniProt.

2.3. База NucleosomeDB - коллекция структур нуклеосом и их комплексов с негистоновыми белками хроматина

Первая кристаллическая структура нуклеосомы с высоким разрешением (2.8 Å) была получена в 1992 году [134] методом рентгеноструктурного анализа. С тех пор количество структур нуклеосом и их комплексов с негистоновыми белками увеличилось, на что, в том числе, оказало влияние развитие методов крио-электронной микроскопии.

База данных и веб-ресурс NucleosomeDB (https://nucldb.intbio.org) разработаны для сбора и анализа структур нуклеосом и их комплексов с негистоновыми белками [135]. NucleosomeDB позволяет исследователям искать, изучать и сравнивать нуклеосомы между собой, несмотря на различия в составе и особенности их представления. В веб-ресурсе реализован анализ проекций α-спиралей гистонов, геометрических параметров динуклеотидов ДНК и контактов белок-белок и белок-ДНК.

Глава 3. Сравнительный анализ состава ядерного протеома и хроматома человека на основе различных экспериментов и баз данных

3.1. Сравнительный анализ онтологий локализации белков и их наполнения между UniProt, HPA, OpenCell

Было проведено сравнение онтологий локализации ядерных белков и их состава из базы данных UniProt и из протеомных консорциумов HPA и OpenCell. Далее будут описаны критерии отбора белков для анализа.

Из UniProt был загружен референсный протеом человека (human proteome UP000005640, release 2022_2, reviewed:yes) - 20354 записи, 20225 - уникальных идентификаторов белков, 20272 уникальных идентификаторов генов. Онтология белков ядра включает следующие термины: ядрышко, ядерная оболочка, нуклеоплазма, ядерный матрикс, ядерные тельца, **Рисунок 4**, **слева**. Хромосома в UniProt - отдельная структура, не входящая в состав ядра, к ней относятся центромера, кинетохор и теломера. Для анализа мы использовали белки, принадлежащие ядру, хромосоме или их подструктурам. В UniProt для 14149 белков референсного протеома человека в записи о локализации присутствовал источник информации. Записи о локализации белков ядра и хромосомы из UniProt содержат следующие коды достоверности: ECO:0000269 - наличие экспериментального свидетельства, ECO:0000250 - решение куратора, основанное на схожести последовательностей белков, ECO:0000305 - решение куратора, ECO:0000255 - решение куратора, основанное на предсказании модели о сходстве последовательностей мотивов или доменов базы данных InterPro, ECO:0000303 - заявление автора без дополнительной поддержки, **Рисунок 5А**. 1082 пары белок (ядра или хромосомы) - клеточная локализация не содержат источника информации о локализации, они не вошли в дальнейший анализ.

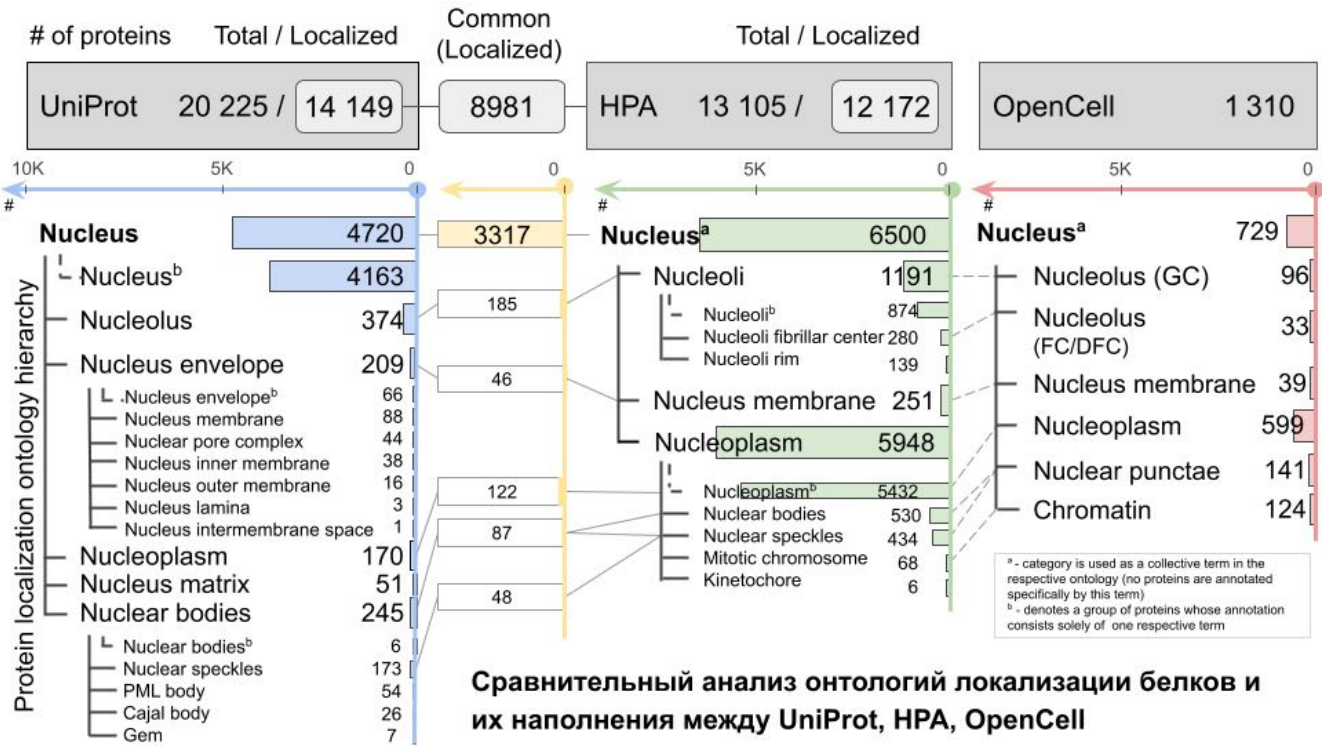


Рисунок 4. Сравнительный анализ онтологий локализации белков и их наполнения между UniProt, HPA и OpenCell. Количество белков в источниках представлено в виде: общее (Total на схеме) - количество белков человека в соответствующем источнике, Localized - количество белков с данными о локализации, удовлетворяющим определенным условиям (подробнее в тексте). Для каждого источника изображена онтология (слева от названий категорий) и количество белков в категориях (справа от категорий). Желтые столбики - общее количество белков в одноименных категориях в UniProt и HPA. [a] - собирательная категория, отсутствующая в оригинальном источнике, [b] - группа белков, аннотация которых включает только этот термин.

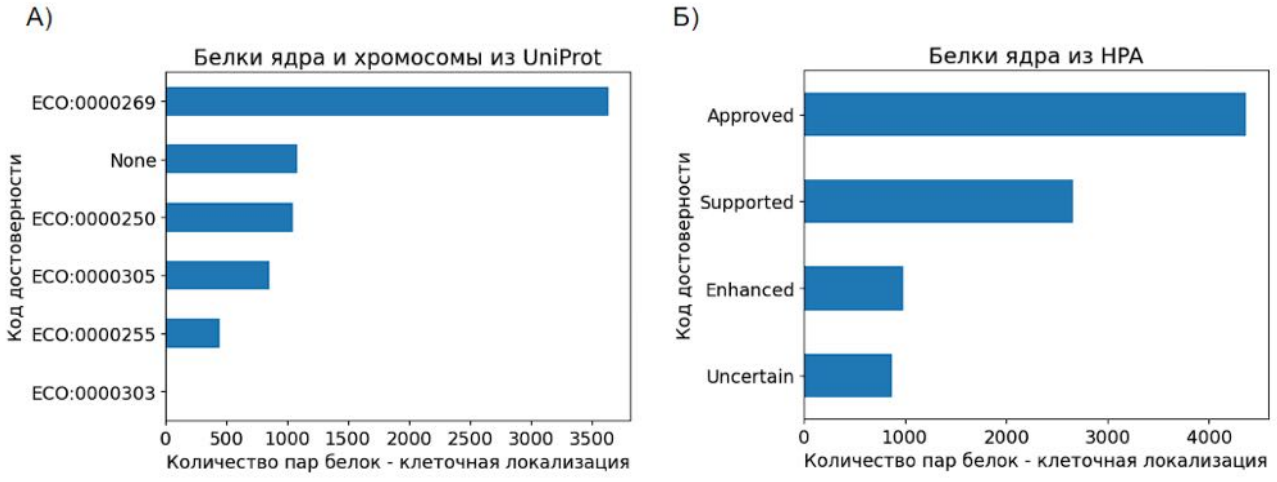


Рисунок 5. Типы источников с подтверждением информации о локализации белков ядра в А) UniProt, Б) HPA.

Информация о клеточной локализации белков из второго ресурса HPA была загружена из HPA версии 22.0. Всего в HPA проанализировано 13105 белков. Онтология категорий ядра изображена на **Рисунке 4, в центре**. В HPA ядерные белки делятся на три группы: ядрышко, ядерная мембрана и наиболее многочисленная группа нуклеоплазма, включающая также белки митотической хромосомы. Коды достоверности в системе HPA включают в себя следующие: Enhanced - одно или более антитело прошло валидацию (согласно пяти "столпам" валидации антител [136]) и нет противоречащей информации в литературе; Supported - не было валидации используемого антитела, но аннотированная локализация описана в литературе; Approved - локализация белка частично согласуется с данными из литературы или не была описана ранее; Uncertain - картина окрашивания антителами противоречит экспериментальным данным или экспрессия гена не обнаружена на уровне РНК. Количество пар белок - клеточная локализация с кодами достоверности представлены на **Рисунке 5Б**. Для дальнейшего анализа были выбраны белки с клеточной локализацией с тегами Enhanced, Supported и Approved.

Третьим проанализированным источником информации о локализации белков был OpenCell. Используя редактирование генов и конфокальную микроскопию на живых клетках, авторам удалось проанализировать всего 1310 генов. В связи с небольшим количеством информации, дополнительная фильтрация данных по кодам достоверности не проводилась. В OpenCell 729 белков являются ядерными, онтология включает в себя следующие термины: гранулярный и фибриллярный компоненты ядрышка, ядерная мембрана, нуклеоплазма, ядерные тельца и хроматин, **Рисунок 4, справа**. Результаты анализа пересечения категорий белков ядра из двух крупнейших источников информации показали различия как в онтологии, так и в количественном наполнении. Пересечение наборов ядерных белков из UniProt, HPA, OpenCell изображено на **Рисунке 6А**. Наборы ядерных белков между UniProt и HPA пересекаются только на 46%.

Было выявлено отсутствие консенсуса в информации о локализации ядерных белков, в частности:

1. В HPA ядрышко состоит из подструктур и включает в 3 раза больше белков, чем UniProt, а пересечение составляет меньше 50% белков UniProt (z-test, p-value: 4.3e-132).
2. Мембранные белки пересекаются менее чем на четверть от размера соответствующих наборов данных.
3. Большинство белков ядра HPA относятся к нуклеоплазме (5949), в то время как в UniProt всего 170 белков в нуклеоплазме. Это можно объяснить особенностью методов - конфокальная микроскопия позволяет локализовать белки в цитоплазме,

в то время как анализ литературы из UniProt относит большинство белков к категории Ядро без дальнейшего уточнения (4163).

Было выявлено также отсутствие консенсуса в плане определения структур ядра: хромосома не часть ядра в UniProt, в HPA митотическая хромосома - это часть нуклеоплазмы, в OpenCell хроматин - одна из структур ядра.

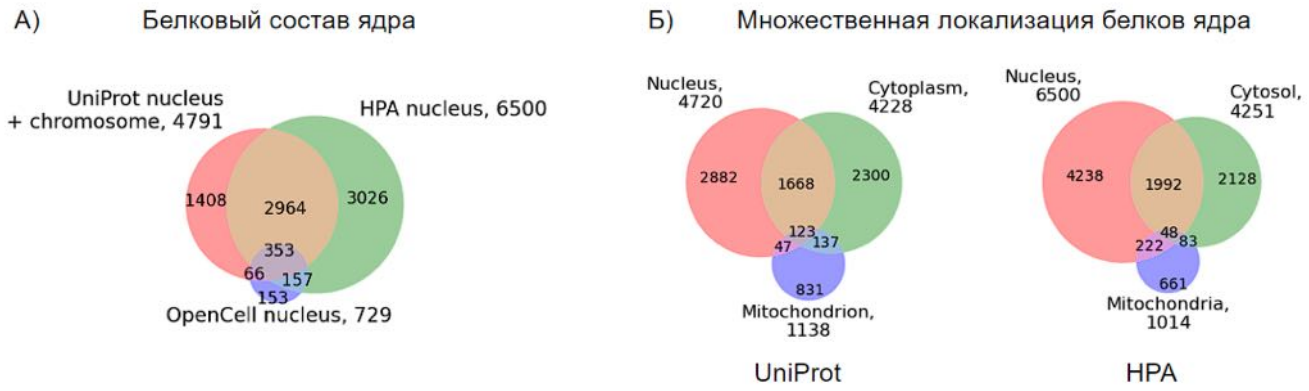


Рисунок 6. А) Диаграмма Венна белков ядра из трех ресурсов (UniProt, HPA, OpenCell). Б) Множественная локализация белков ядра из UniProt и HPA: 2 наиболее многочисленных компартмента с ядерными белками - это митохондрии и цитоплазма.

Причинами таких расхождений в определении структур ядра могут быть следующие:

- Хроматин - это динамическая органелла (fuzzy organelle) [108], качественный и количественный состав которой зависит от: типа ткани и ее возраста [137], стадии клеточного цикла [138], протекающих биологических процессов (репликация, экспрессия генов, репарация ДНК и др.), от экспериментального метода экстракции белков [82,108].
- Множественная локализация белков ядра: среди всех ядерных белков 34 и 39% белков в UniProt и HPA, соответственно, встречаются также в цитоплазме и в митохондрии, **Рисунок 6Б**.

Но если тип эксперимента и выбор клеточных линий в определении локализации белков в HPA и OpenCell еще может влиять на белковый состав исследуемых структур, то информация из базы данных UniProt, собранная по данным разных источников (литература, предсказательные модели и др.), казалось бы, должна быть более "устойчивой".

3.2. Количественный анализ экспериментально-полученных хроматомов

Мы провели поиск наборов данных белков хроматина в базе PubMed по ключевым словам: chromatome, experimental chromatin proteins, nuclear proteins. Хроматомы с делением на наборы данных в историческом контексте и наборы, полученные после 2000-го года приведены

в **Приложении А**. Не смотря на то, что проведенный анализ белков хроматина затрагивает только белки человека, в **Приложении А** приведены также некоторые хроматомы из других организмов, наборы из исследований, посвященных составу хроматина на разных стадиях клеточного цикла и в тканях разного возраста.

Для сравнительного анализа были использованы все найденные публично доступные наборов данных белков хроматина человека, **Рисунок 7**. В хроматоме ChEP были выбраны белки, с вероятностной оценкой принадлежности к белкам хроматина выше, чем 0.5 [108], далее kustatscher_2011. В хроматоме DEMAC были белки были выбраны по следующим критериям: было идентифицировано как минимум 2 пептида во всех трех репликах как минимум одной клеточной стадии, были выбраны канонические изоформы для всех белков [119], далее ginno_2018. В хроматоме Hi-MS были использованы данные, полученные в эксперименте с расщепление белков перед МС в растворе, а не в геле (авторы статьи выбрали данное условие, так как в результате было идентифицировано больше белков) [120], далее shi_2021. В статье Torrente et al., 2011 [109] сравнивали 3 метода экстракции хроматина: общую экстракцию, солевую экстракцию и расщепление МНКазой, для анализа были выбраны все белки из трех идентифицированных наборов, далее torrente_2011; в качестве идентификаторов белков были приведены GI accession, после перевода идентификаторов в Uniprot Entry осталось 1527 белков из референсного протеома человека, 653 - не из референсного, они были отфильтрованы. Также был загружен набор белков хроматина, полученный с использованием МНКазы [139], далее dutta_2011. Наборы хроматина получены из экспериментов на разных клеточных линиях, в частности, torrente_2011 - HeLa S3; dutta_2014 HEK293F; kustatscher_2014 - HepG2, HeLa, MCF-7; ginno_2018 - T98G (мультиформная глиобластома), shi_2021 - K562.

Описанные методики выделения белков хроматина влияют на идентифицируемый белковый состав, вымывая часть белков вследствии солевых промывок и других экспериментальных этапов. Поэтому мы проанализировали также представленность белков в клеточном ядре из работы [140], где оценили (методом «протеомной линейки») количество белков HeLa в различных органеллах клеток, разделяемых центрифугированием, в том числе количество белков ядра без дополнительных этапов очистки этой фракции.

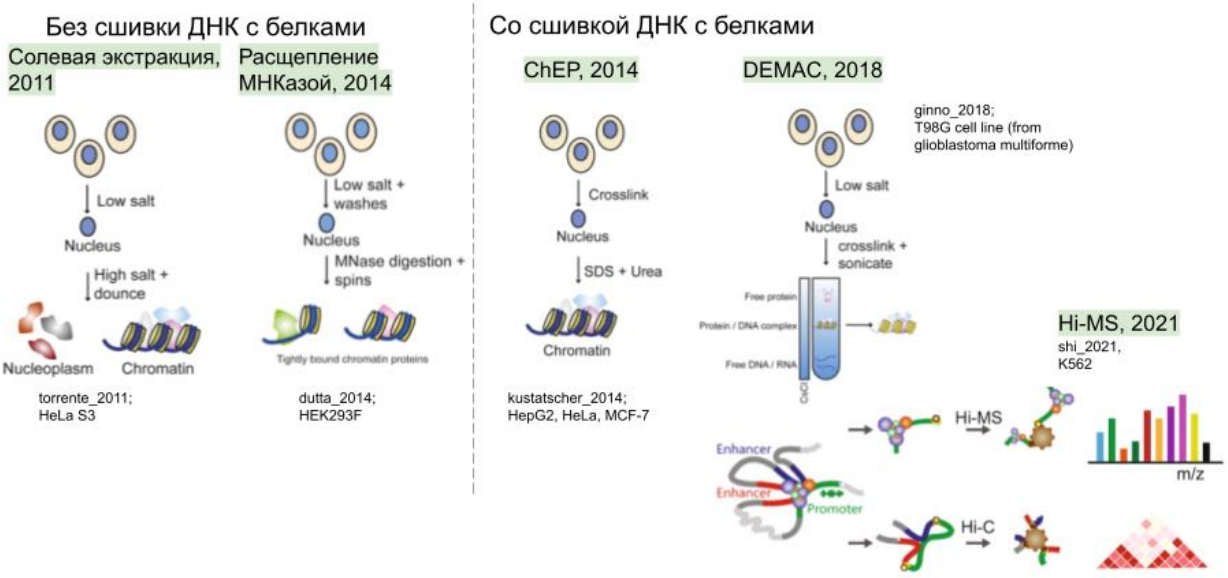


Рисунок 7. Экспериментальные методы экстракции хроматина и наборы данных для сравнительного анализа. Методики поделены на две группыы на основании присутствия или отсутствия этапа фиксации ДНК с белками формальдегидом. Адаптировано из [120,141].

В результате анализа литературы, было найдено 5 наборов данных белков хроматина человека, полученные разными методами выделения белков хроматина. На **Рисунке 8** представлены наборы данных белков хроматина человека с указанием методов очистки хроматина, количество белков в них, а также количество общих белков в наборе белков хроматина и наборе белков ядра из баз данных локализации UniProt или HPA. Размеры наборов белков хроматина варьируются от 482 (Dutta et al., 2014), до 3184 (Shi et al., 2021). Можно отметить, что наименьшее количество белков хроматина в наборах данных, полученных с расщеплением МНКазой (Dutta et al., 2014; Torrente et al. 2011), а наибольшее - в хроматомах, полученных с этапом сшивания ДНК с белками формальдегидом (Shi et al., 2021; Ginno et al., 2018; Kustatscher et al., 2014). Наибольшее присутствие белков ядра из локализационных баз данных обнаружено в (Kustatscher et al., 2014), 88% от набора, что можно объяснить использованием авторами вероятностной оценки для белков хроматина, а также формирование набора данных с отсечкой по вероятностному значению > 0.5. Неожиданно высокое пересечение множеств белков обнаружено между набором белков ядра без дополнительных очисток (Itzhak et al., 2011) и белками ядра их локализационных баз данных, 85%.

В среднем, белковый состав хроматомов на 62% пересекается с ядерными белками из UniProt или HPA, что отражает одну из проблем выбора методики очистки белков хроматина - протокол должен быть достаточно чувствительным и специфичным.

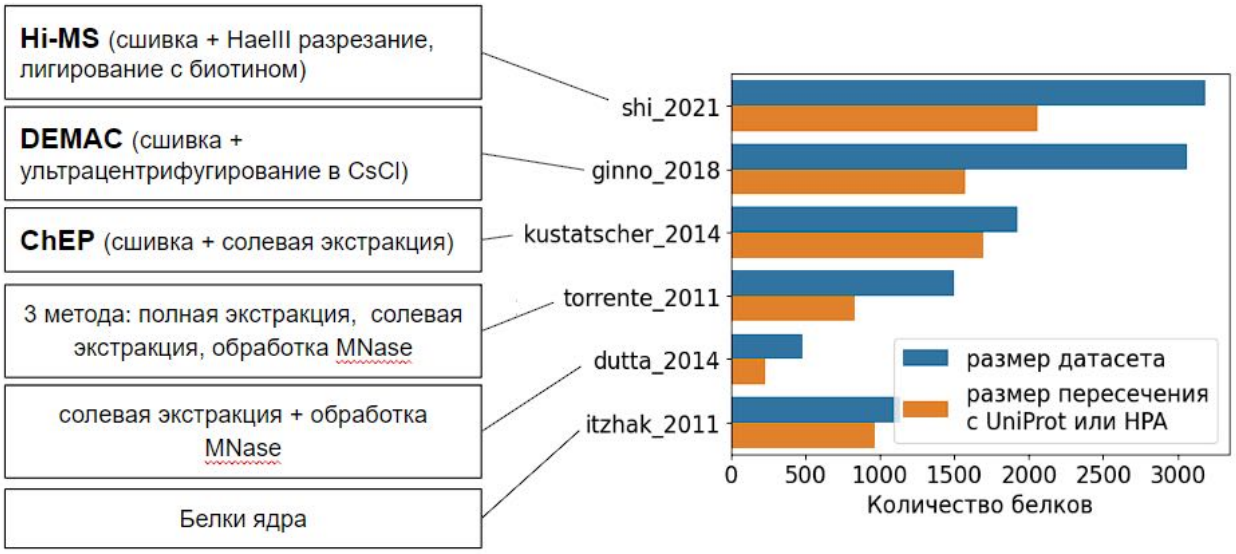


Рисунок 8. Количество белков в хроматомах человека, полученных с использованием разных методик очистки хроматина. Слева - указание особенностей методик, справа - столбчатая диаграмма с количеством белков в хроматомах и их пересечением с ядерными белками из UniProt или HPA.

3.3. Построение эмпирической классификации белков хроматина и ее наполнение

Для более детального анализа качественного состава хроматомов человека нужна функциональная классификация белков хроматина. К существующим системам можно отнести генную онтологию GO и ряд баз данных белков хроматина. В GO нельзя автоматизировано подобрать отсечку для глубины дерева тегов, на которой термины с одной стороны, не очень общие, с другой - не очень специализированные. Существующие базы белков хроматина и эпигенетических регуляторов (CR Cistrome 2013, FACER 2018, CR2Cancer 2018, CRdb 2023, EpiFactors 2023) к белкам хроматина относят следующие категории - гистоны, белки, задействованные в нанесении, считывании и стирании модификаций на гистоны и ДНК, шапероны гистонов, ремоделеры хроматина. В EpiFactors набор категорий также включает следующие: транскрипционные факторы, модификаторы РНК, белки группы Polycomb и белки скаффолда (Scaffold protein), которые по большей части включают в себя белки, участвующие в процессинге РНК. Таким образом, существующие системы для классификации белков хроматина содержат или слишком много разноуровневых терминов, или наоборот, довольно ограниченный набор. Это послужило предпосылкой к разработке эмпирической классификации белков хроматина.

Мы разделили белки ядра на белки ядерной оболочки, ядрышка и нуклеоплазмы/хроматина. Далее белки нуклеоплазмы делятся на гистоны и негистоновые белки. Внутри негистоновых белков деление осуществляется на классы по функциям, процессам, геномной локализации и свойствам (здесь белки HMG, обладающие высокой

электрофоретической подвижностью). Полный список классов перечислен на **Рисунке 9** и в **Приложение Б**. Как видно из **Рисунка 9**, разработанная классификация включает все основные категории белков из баз данных белков хроматина, а также содержит категории белков хроматина, отсутствующие в GO и базах данных (белки HMG; белки, удаляющие "хвост" гистонов; белки скаффолда, то есть хромосомного каркаса; пионерные транскрипционные факторы).

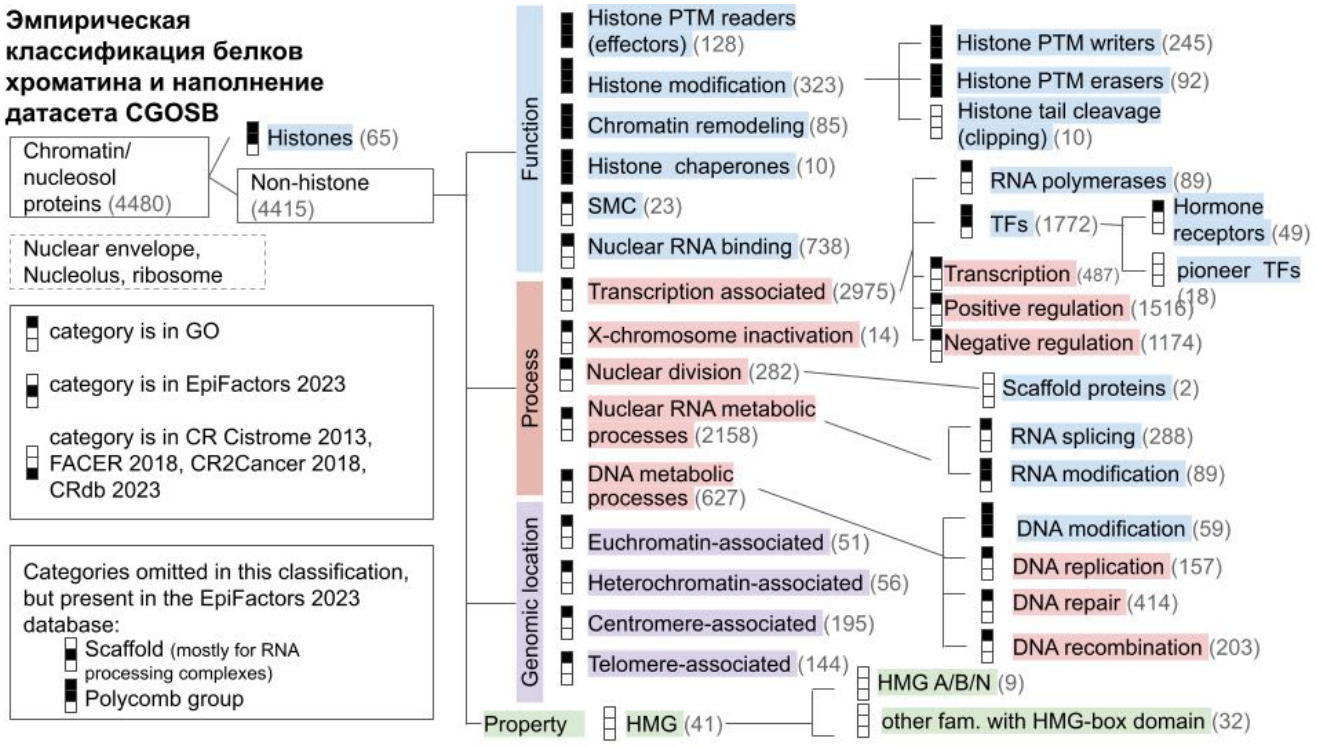


Рисунок 9. Разработанная эмпирическая классификация белков хроматина и нуклеоплазмы. Изображена древовидная структура терминов с указанием количества белков в каждой категории. Негистоновые белки хроматина также имеют принадлежность к одному из классов: функции, процессы, геномной локализация и свойства. Квадратики слева от терминов символизируют наличие термина в системе GO, в базе EpiFactors и в ряде баз данных белков хроматина (CR Cistrome 2013, FACER 2018, CR2Cancer 2018, CRdb 2023, EpiFactors 2023).

Для наполнения терминов из классификации использовали систему GO (соотнесение терминов классификации и GO проводилось в ручном режиме), специализированные базы данных и для некоторых категорий - информацию из литературы. Итоговый набор белков нуклеоплазмы и хроматина получил название Chromatin Gene Ontology Saturated Based (CGOSB). Полный список источников информации для наполнения категорий приведен в **Приложении Б**. Стоит отметить, что в GO нет отдельного термина для гистонов, они относятся к термину structural constituent of chromatin, в котором также есть 2 негистоновых гена: HMGA1 (HMG), LMNTD2 (компонент ламины). Также были проведены и другие дополнительные шаги по очистке наборов белков от нерелевантных белков, которые могли быть загружены с GO:

1. из всех категорий белков были удалены белки, участвующие в транскрипции в митохондриях (GO:0006390);
2. из категорий RNA binding, RNA metabolic process, RNA modification, RNA splicing были удалены не ядерные белки по аннотации UniProt или HPA, что привело к сокращению состава категорий на 535 белков, 447, 63 и 21, соответственно;
3. белки категорий ядрышко, рибосома и гистоны находятся только в этих категориях и были удалены из всех остальных.

3.4. Количественный состав ядерного протеома и хроматома человека: оценка массы гистонов

Для оценки количественного состава ядерного протеома мы использовали разработанную эмпирическую классификацию белков хроматина и нуклеоплазмы, наборы ядерных белков из UniProt и HPA и данные о представленности белков в живых клетках из базы PAXdb [132]. Из PAXdb версии 4.2 был загружен интегрированный набор данных о представленности белков человека (*Homo Sapiens* - Whole organism, Integrated), со средневзвешенными усредненными значениями представленности по 175-ти наборам протеомов из различных тканей и клеточных линий, полученных методом масс-спектрометрии. Единица измерения - parts per million (ppm), количество молекул белка, нормированное на миллион. К анализируемым категориям белков была добавлена категория с белками, отсутствующими в классификации, но являющимися белками содержимого ядра по аннотации UniProt или HPA (на рисунках - Other nuclear (UniProt | HPA)). Так как в разработанной классификации белок может быть в нескольких категориях, то перед анализом количественного состава было проведено устранение пересечения белкового состава функциональных категорий CGOSB. Алгоритм описан **формулой 4:**

$$S_i^{new} = S_i - S_i \cap \left(\bigcup_{j<i}^{N} S_j \right), \quad (4)$$

где: S_i - набор белков в категории, S_i^{new} - новый набор белков категории без пересечения с другими категориями, и состоит из следующих шагов:

1. Получение отсортированного списка категорий белков по медиане представленности белков.
2. Последовательное устранение пересечений между множествами, то есть из категорий с меньшим медианным значением представленности белков удаляются белки категорий с бОльшим медианным значением представленности. Обобщающие категории, подавляющее число белков которых принадлежат дочерним категориям (процесс

метаболизма ДНК, транскрипция и модификаторы гистонов), в данной процедуре не участвовали.

3. Пересортировка категории на основе обновленной медианы представленности белков.

На **Рисунке 10А** изображена представленность белков в категориях до и после устранения пересечений.

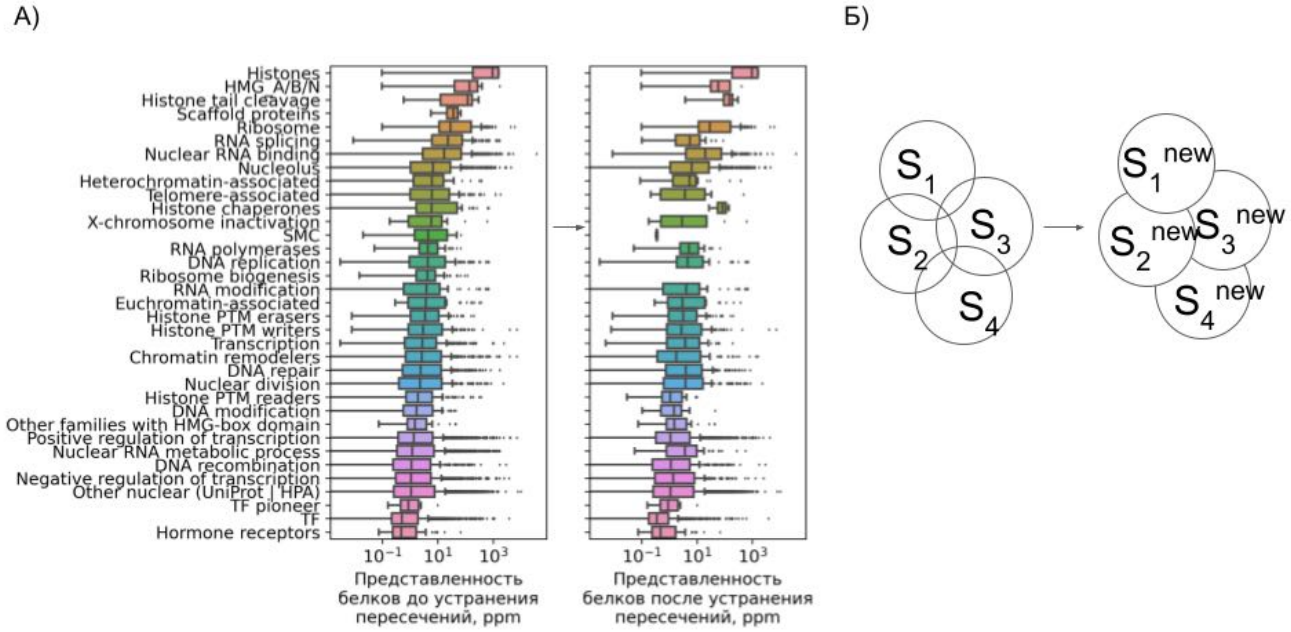


Рисунок 10. Процедура устранения пересечения белкового состава в категориях CGOSB. А) Боксплот с представленностью белков в категориях до и после процедуры, ось абсцисс в логарифмических координатах. Б) Схематическое описание процедуры устранения пересечений в категориях CGOSB.

Далее мы оценили общую представленность белков содержимого ядра, результаты представлены на **Рисунке 10**. Относительно белков содержимого ядра (из набора CGOSB или из аннотации UniProt или HPA), наибольшее медианное значение представленности белков у гистонов и белков группы HMG, что не противоречит литературным данным, рисунок X. Присутствие категории белков, отщепляющих хвост гистонов, можно объяснить мультифункциональной природой белков. Так, у некоторых широко представленных металлопротеиназ и катепсинов в определенных условиях были выявлены функции отщепления хвоста гистонов. Переходе от медианы представленности белков категорий к массовой доли осуществлялся согласно **формуле 5**:

$$M_{cat} = \sum_i A_i \times M_i \quad (5)$$

где: M_{cat} - массовая доля белков категории с учетом их представленности, A_i - представленность белка, M_i - молекулярная масса белка. Полученная оценка для категорий

изображена на рисунке Х, Г. Наибольшей массой обладают следующие категории ядерных белков: белки, связывающиеся с РНК (24%); белки ядрышка (23%); другие ядерные, отсутствующие в наборе CGOSB (21%); белки, активирующие транскрипцию (9%); белки, наносящие ПТМ гистонов (5%), гистоны (4%).

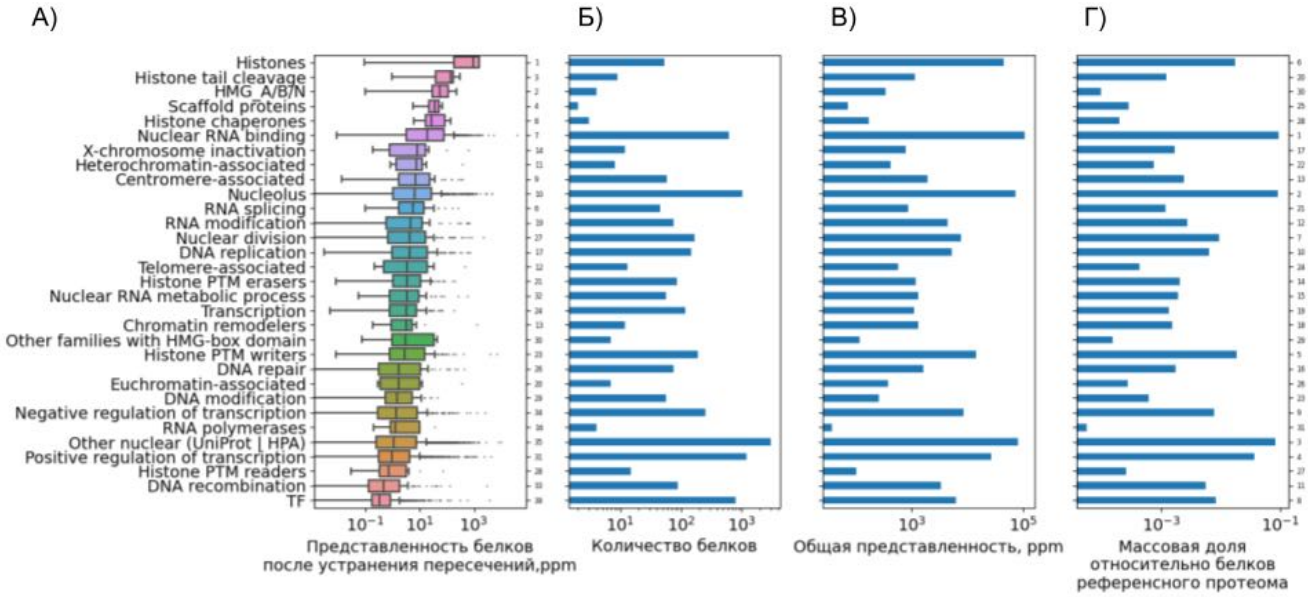


Рисунок 11. Представленность белков содержимого ядра в клетке человека по данным обобщенного набора данных из PAXDB (детали в тексте). Слева изображены категории белков ядра из разработанной классификации после устранения пересечений белкового состава общие для всех панелей, ось абсцисс на всех панелях в логарифмических координатах.
А) представленность белков в ppm (protein per million), категории отсортированы по убыванию медианного значения, индексы справа указывают порядковый номер категории в аналогичном списке до устранения пересечений белкового состава категорий.
Б) Количество белков в категориях.
В) Общая представленность белков каждой из категорий.
Г) массовая доля белков относительно референсного протеома, индексы справа указывают на порядковый номер в отсортированном по убыванию списке соответствующих значений.

Нами были предприняты попытки оценить массовую долю гистонов в клеточном ядре и в хроматине. Широко цитируемая оценка массовой доли гистонов осуществлена Ван Хольде [83], где соотношение массы ДНК : гистоны : негистоновые белки хроматина составляет 1 : 0.99 : 0.3-0.8.

Для получения оценки массы белков хроматина, мы провели описанные выше расчеты не учитывая следующие категории: РНК-связывающие белки (белки ядрышка; ядерные белки, связывающие РНК, осуществляющие метаболизм РНК, модификации РНК, сплайсинг РНК); другие ядерные белки, отсутствующие в CGOSB; белки, участвующие в делении клеточного ядра; собирательные категории, белки которых практически полностью распределены по категориям нижележащих уровней: белки, ассоциированные с транскрипцией, метаболизм ДНК, модификациями гистонов, рибосомами. В результате массовая доля гистонов составила

11% и 19%, **Рисунок 12**, если учитывать белки, которые присутствуют также в группах РНК-связывающих белков и нет, соответственно.

Массовая доля белков гистонов при этом увеличилась с 4% (относительно белков содержимого ядра) до 18% (относительно белков хроматина). Полученная оценка 18% - является оценкой "снизу", так как в рассмотрении участвовали все белки, относящиеся к категориям классификации CGOSB. Для получения оценки "сверху" мы проанализировали те белки, которые по аннотации UniProt или HPA относятся только к белкам ядра. Массовая доля гистонов относительно белков хроматина с единственной ядерной локализацией составила 44%, **Рисунок 12**.

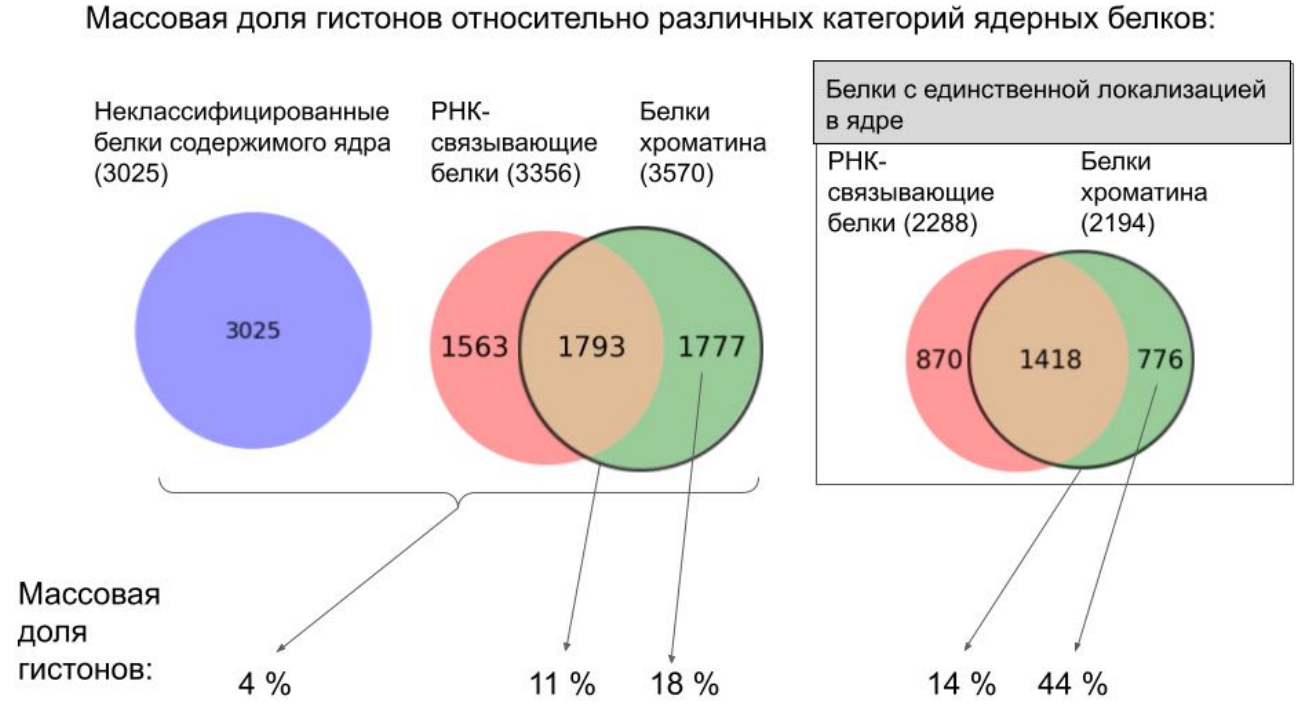


Рисунок 12. Массовая доля гистонов (рассчитанная по **формуле 5**) относительно различных категорий ядерных белков, указанных внизу рисунка.

В качестве дополнительного анализа мы провели оценку массовой доли гистонов в экспериментально-полученных наборах белков хроматина человека, в которых были опубликованы значения суммарной интенсивности сигнала MC для белков (Shi et. al., 2021; Ginno et al., 2018; Kustatscher et al., 2014).

Массовая доля гистонов относительно других белков набора составила 0.57% (Kustatscher et al., 2014), 1.25% (Ginno et al., 2018) и 41% (Shi et al., 2021). При пересчете массовой доли гистонов относительно белков хроматина, то с исключением категорий РНК-связывающих белков и неклассифицированных белков ядра, массовая доля гистонов изменяется до: 0.71% (Kustatscher et al., 2014), 1.4% (Ginno et al., 2018) и 45 (Shi et al., 2021). Наибольшая массовая доля гистонов, 45%, получена в наборе (Shi et. al., 2021), где авторы

использовали метод выделения белков хроматина схожий с тем, который используют в экспериментах Hi-C. Методы экстракции хроматина в трех наборах хоть и содержат модификации, но в общем предполагают солевую экстракцию, то есть промывку образца в растворах солей с низкой и высокой концентрацией, что по всей видимости приводит к отмывке и доли гистоновых белков.

Для нивелирования фактора солевой промывки мы проанализировали долю гистонов по массе в наборе белков ядра (Itzhak et al., 2011). Несмотря на то, что в эксперименте анализировали ядро клеток без дополнительных этапов очистки, доля массовая доля гистонов относительно белков хроматина составила 9%.

Таким образом, с помощью разработанной классификации белков содержимого ядра и интегрированного набора представленности белков из базы PAXdb мы оценили массовую долю категорий ядерных белков, а также получили оценку массовой доли гистонов относительно белков хроматина: она составляет 19 - 41%. Полученное значение отличается от оценки Ван Хольде [83] в 1.5 раза в меньшую сторону и является предметом дальнейших дискуссий. К ограничениям проведенного вычисления можно отнести:

1 - наличие шума в классификации белков по системе GO, которая являлась одним из источников информации для наполнения набора CGOSB;

2 - эксперименты масс-спектрометрического анализа количества белков, депонированные в PAXdb, проводили на целых клетках, без очистки ядерных белков и белков хроматина, а чувствительность методики могла не позволила выявить наименее представленные белки.

С другой стороны, оценки количества гистонов за счет разделения белков на электрофорезе, усредненные в [83], предполагают сравнение 2 групп белков - гистоны являются однородной группой по размеру и заряду и образуют одну полосу, в то время как негистоновые белки хроматина представляют собой разнородную группу белков, которые не концентрируются на гель-электрофорезе в одну полосу, затрудняя тем самым оценку их количества.

3.5. Физико-химический анализ белков хроматина

Мы провели сравнение физико-химических свойств белков хроматина из разных источников с белками цитоплазмы по аннотации UniProt или HPA. В качестве анализируемых признаков были рассмотрены: фракции отдельных аминокислот и сгруппированных по свойствам, заряд, молекулярная масса, доля неупорядоченных регионов. Сравнение проводилось с использованием U-критерия Манна-Уитни, так как данные не распределены нормально согласно тесту Шапиро-Уилка. В качестве поправки на множественное сравнение была выбрана поправка Бонферрони, значимыми считались результаты с уровнем p-value < 0.00014. Результаты сравнения экспериментальных наборов хроматина, данных из UniProt, HPA,

OpenCell, а также 3 дополнительных наборов данных: histone_ppi - белков-партнеров гистонов (загруженных из баз белок-белковых взаимодействий, полученных в высокопроизводительных экспериментах, STRING [142] версия 11, IntAct [143] версия 4.2.16, BioGRID [144] версия 4.3, и депонированных в базе HistonePPIDB, доступной по веб-адресу https://intbio.org/histoneppidb/), nlsdb_prediction - белки человека, содержащие сигналы ядерной локализации из базы NLSDB [145], deeploc_acc_nucl - результаты предсказания нейросетевой модели Deeploc 2.0 [146], обученной на аннотации клеточной локализации белков по UniProt. Результаты проведенных статистических тестов сравнения медианных значений признаков в парах набор белков ядра или хроматина и белков цитоплазмы приведены на **Рисунке 13**. Медианное значение части признаков статистически значимо отличается во всех наборах белков хроматина или ядра по сравнению с цитоплазмой (доля положительно заряженных аминокислот; доли серина, пролина и лейцина; значение изоэлектрической точки; заряд белков). Медианное значение доли тирозина статистически значимо не различается ни в одной паре сравнения набор белков хроматина или ядра и белков цитоплазмы.

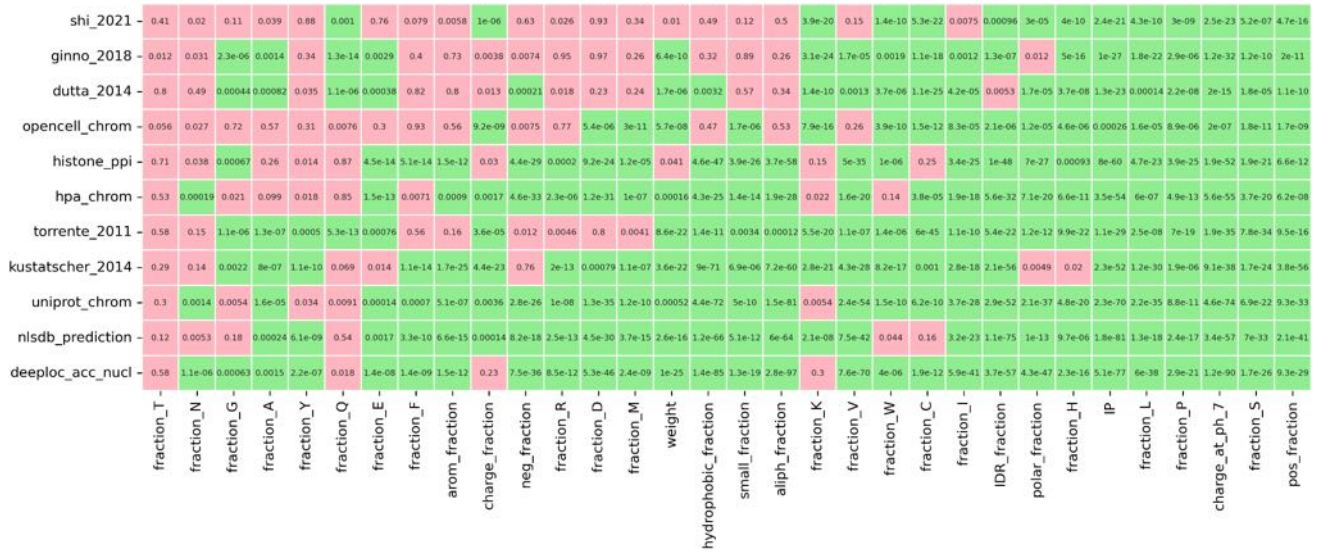


Рисунок 13. Результаты сравнения медианных значений физико-химических свойств белков хроматина и ядерных белков из различных источников (эксперименты, базы данных, протеомные инициативы, предсказания) относительно белков цитоплазмы (по аннотации UniProt или HPA). Физико-химические свойства представлены в столбцах, наборы данных хроматина или ядра в строках, на пересечении - результаты U-критерия Манна-Уитни с поправкой Бонферрони, зеленые ячейки - статистически значимые отличия, красные - статистически незначимые отличия.

Для наборов белков хроматина, которые статистически значимо отличаются от белков цитоплазмы, мы оценили разницу в медианных значениях признака, **Рисунок 14**. Медианное значение доли ароматических и отрицательно-заряженных аминокислот во всех наборах хроматина ниже, чем соответствующее значение белков цитоплазмы, а медианное значение доли положительно-заряженных аминокислот выше. Для физико-химических признаков, у

которых медианное значение в белках хроматина меньше, чем в белках цитоплазмы, различие медианных значений с белками цитоплазмы в среднем составляет 0.4%. Для признаков с большим медианным значением в белках хроматина по сравнению с белками цитоплазмы, в среднем различие составляет 0.6%. Медианное значение доли аминокислот (метионин, гистидин, цистеин, аспарагиновая кислота, валин, аланин, изолейцин, пролин, серин) в белках ядра и хроматина отличается как в меньшую, так и в большую сторону относительно белков цитоплазмы.

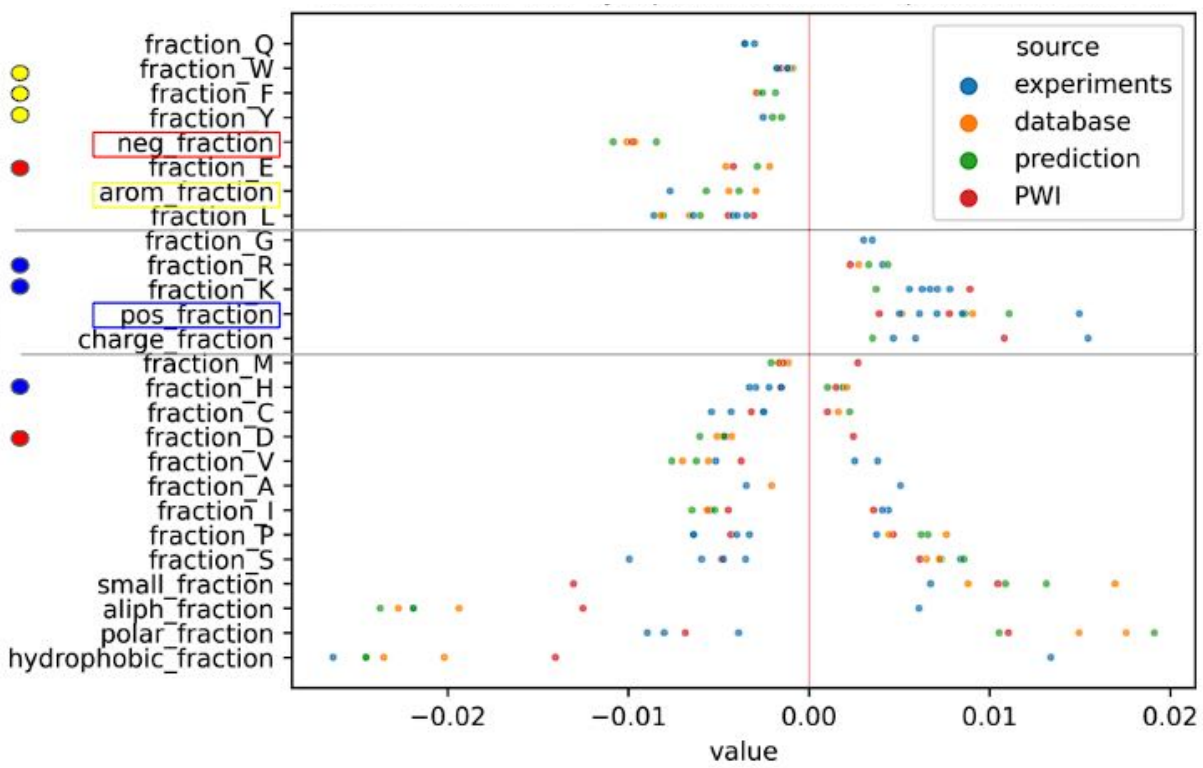


Рисунок 14. Различие в медианных значениях физико-химических признаков белков хроматина и ядра относительно соответствующего значения белков цитоплазмы (вертикальная красная линия). Наборы белков хроматина сгруппированы по типу источника информации (эксперименты, базы данных, предсказания, протеомные инициативы). Рамками и кругами слева выделены группы аминокислот и отдельные аминокислоты, входящие в состав группы, соответственно; отрицательно заряженные выделены красным цветом, положительно заряженные синим, ароматические желтым.

3.6. Особенности распределения заряда в белках хроматина

В ядрах клеток находится отрицательно-заряженная молекула ДНК, с которой связываются положительно-заряженные белки гистоны. Анализ общего заряда белков ядра и хроматина показал статистически значимую разницу между медианами значений заряда по сравнению с белками цитоплазмы, **Рисунок 15А**. Далее мы проанализировали заряд белков в ядре с учетом их представленности, **Рисунок 15Б**. На диаграмме представленности белков ядра с разбивкой на заряды заметен пик (заряд=15) с коровыми гистонами, наиболее представленными положительно-заряженными ядерными белками, **Рисунок 15Б.** Линкерные

гистоны отличаются от коровых по строению и свойствам, на **Рисунке 15Б** они расположены в правой части графика (заряд=50-60).

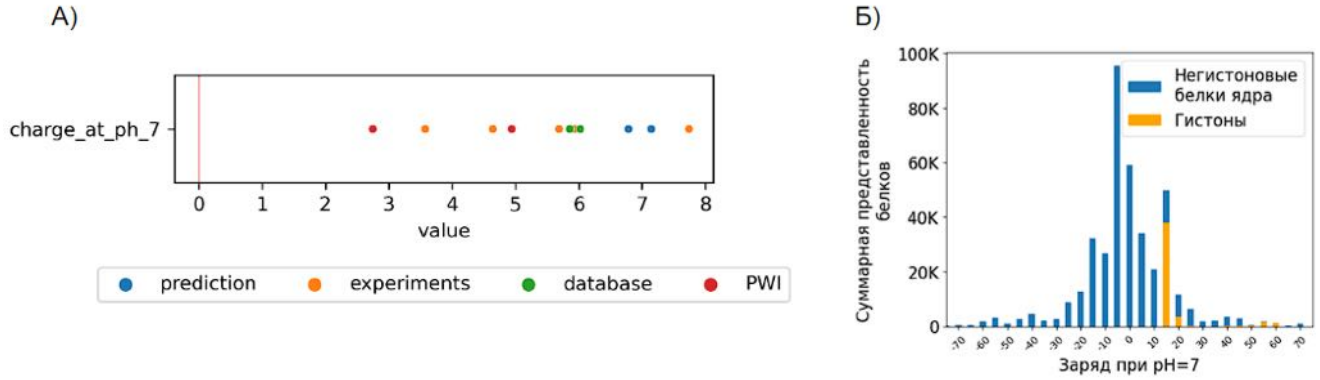


Рисунок 15. А) Различие в медианных значениях заряда белков хроматина и ядра относительно соответствующего значения для белков цитоплазмы (красная линия). Наборы белков хроматина сгруппированы по типу источника информации (эксперименты, базы данных, предсказания, протеомные инициативы). Б) Диаграмма представленности заряда ядерных белков, ось абсцисс обрезана по значениям (-75, 75).

Детальный анализ распределения положительно и отрицательно заряженных белков внутреннего содержимого ядра по разработанной классификации показал, что наибольший положительный заряд в ядре несут белки категорий: гистоны, ядерные РНК-связывающие белки и белки ядрышка, **Рисунок 16**. Интересно, что только один из гистоновых белков заряжен отрицательно - это гистоновый вариант H2A.P, относящийся к группе коротких гистонов (117 а.о.). При этом отрицательно заряженные белки с учетом их представленности преобладают в категориях ядерных РНК-связывающих, белках ядрышка, неклассифицированных белках ядра, и ряде категорий белков хроматина: белках, регулирующих транскрипцию, наносящих и стирающих ПТМ гистонов, осуществляющих репликацию ДНК и др., **Рисунок 16**.

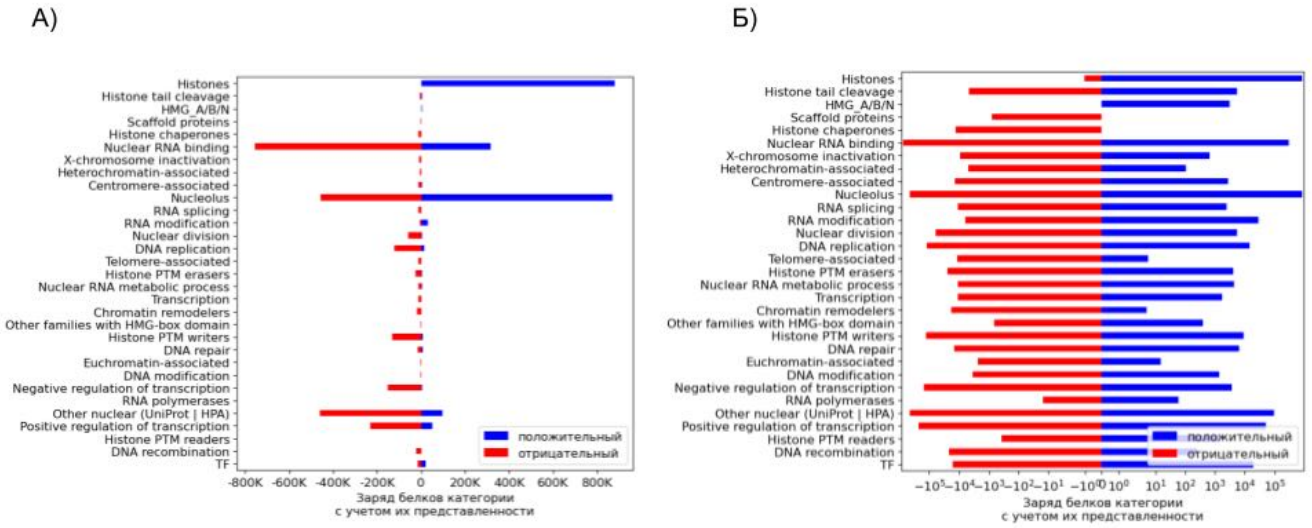


Рисунок 16. Заряд категорий белков ядра с учетом их представленности. А) Представленность положительно и отрицательно заряженных белков с разбивкой на категории из разработанной классификации. Порядок категорий такой же, как на рисунках **Раздела 3.4** после устранения пересечения белков между категориями. Б) тот же график в симметричных логарифмических координатах.

В результате проведенного сравнительного анализа аминокислотного состава белков хроматина относительно белков цитоплазмы мы выявили следующие статистически значимые отличия. В среднем по наборам белков хроматина разница медианного значения доли аминокислот в белках хроматина относительно белков цитоплазмы составляет: −1% для доли отрицательно заряженных аминокислот, −0.5% для доли ароматических аминокислот и +0.8% для доли положительно заряженных аминокислот. Различия в заряде белков можно объяснить природой взаимодействия белков, нейтрализацией отрицательного заряда ДНК. Различия в составе ароматических аминокислот в белках хроматина и цитоплазмы может быть связано с их ролью в узнавании нуклеиновых кислот: было показано, что в сайтах связывания нуклеиновых кислот в белках увеличено количество ароматических аминокислот по сравнению с ожидаемым распределением [147]. При этом нельзя говорить о том, что белки хроматина (кроме гистонов) несут сильный положительный заряд. Белки в большинстве категорий хроматина заряжены отрицательно.

3.7. Доменная архитектура белков хроматина

Специфичные для белков хроматина функции определяются разнообразием белковых доменов. Для анализа доменной архитектуры белков была использована аннотация доменов из Pfam [148]. Для более детального анализа были выбраны белки, которые относятся к следующим функциональным категориям хроматина: наносящие, считывающие и удаляющие

ПТМ гистонов, шапероны гистонов и ремоделеры хроматина. С основными функциональным категориями белков хроматина были соотнесены домены по аннотации Pfam, используя поиск ключевых слов регулярными выражениями в описаниях доменов, а также данные из литературы, **Таблица 1**.

Таблица 1. Домены по системе Pfam основных функциональных классов белков хроматина.

Функция	**Домены (идентификаторы Pfam)**
Метилирование гистонов	PRMT5_TIM, PRMT5, PRMT5_C, SET, Pre-SET
Удаление метильных меток с гистонов	JmjC, JHD, JmjN
Считывание метильных меток с гистонов	ADD_ATRX, ADD_DNMT3, Ank_2, Ank_4, Ank, Ank_5, Ank_3, BAH, Chromo, PWWP, TTD, Tudor-knot, Tudor_2, 53-BP1_Tudor, TUDOR, WD40, zf-CW, PHD, PHD_3, PHD_2, PHD_4, RAG2_PHD, zf-HC5HC2H_2, MBT
Ацетилирование гистонов	HAT_KAT, zf-MYST, Acetyltransf_1, Acetyltransf_13, Hat1_N, EPL1
Стирание ацетильных меток с гистонов	HDAC4_Gln, Hist_deacetyl, SIR2
Считывание ацетильных меток с гистонов	Bromodomain, PHD
Шапероны	ASF1_hist_chap, CHZ,
Хроматиновые ремоделлеры	Helicase_C, SNF2-rel_dom
Узнавание метилированной ДНК	MBD, MBDa
Метилирование ДНК	DNA_methylase

Белковые семейства, домены и повторы, присутствующие в более чем шести белков обозначенных выше категорий, были сгруппированы по функциям; **рисунок 17**.

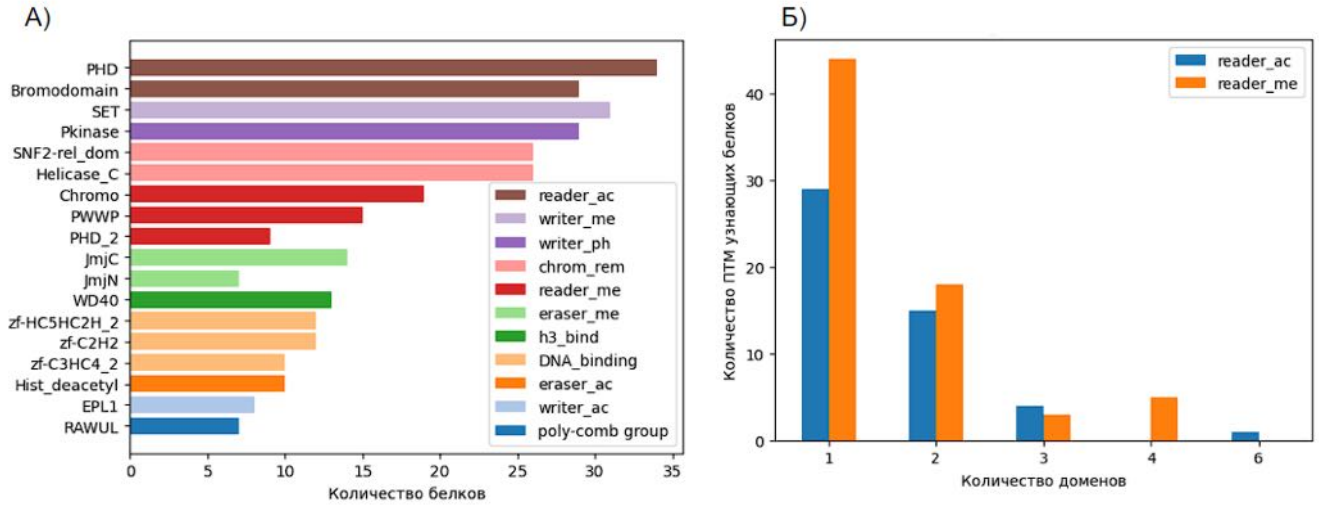


Рисунок 17. А) Белковые семейства, домены и повторы по аннотации Pfam некоторых групп белков хроматина (подробнее в тексте), сгруппированные по функциям, присутствующие в более шести белков. Б) Количество доменов одного типа, узнающих метильные или ацетильные метки гистонов, в белках, узнающих ПТМ гистонов согласно разработанной классификации.

Мы проанализировали ко-встречаемость доменов разных функциональных групп в одних белках, **Рисунок 18**, и ко-встречаемость доменов в белках, узнающих ПТМ гистонов, **Рисунок 19**.

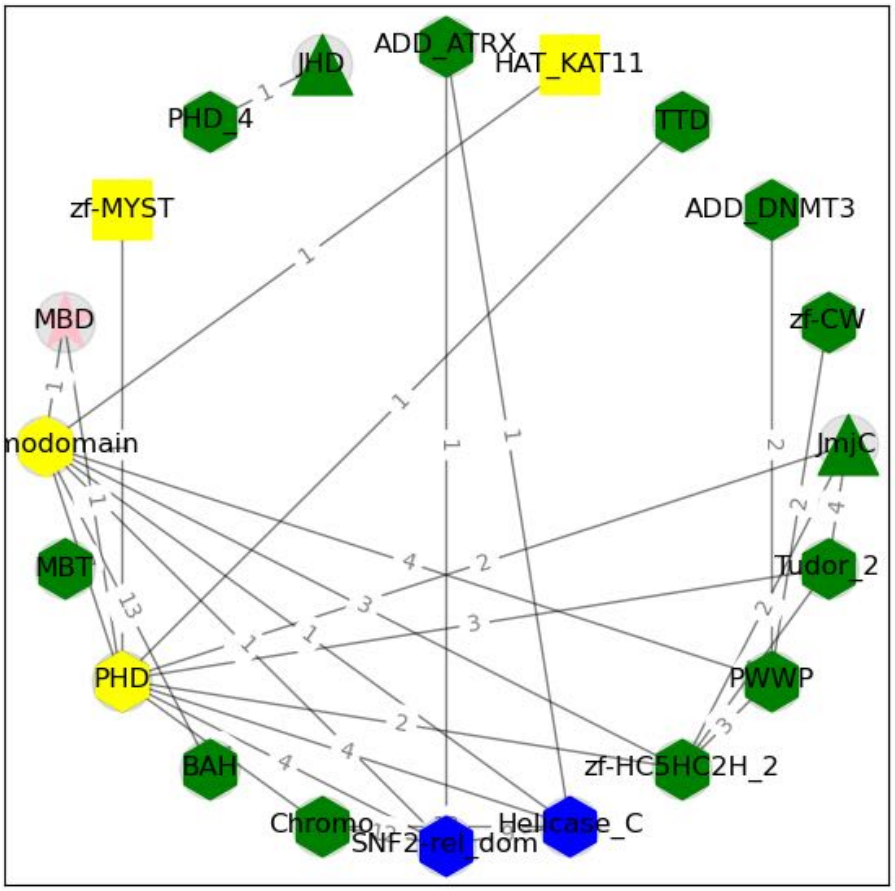


Рисунок 18. Ко-встречаемость доменов, узнающих ПТМ гистонов и доменов других функциональных категорий хроматина (подробнее в тексте). Легенда: зеленый - белки, связанные с метилирование гистонов, желтый - с ацетилированием, синий - ремоделеры хроматина, розовая звезда - узнавание метилированной ДНК; шестиугольники - узнавание метки или ремоделер хроматина, треугольник - удаление метки, квадрат - нанесение метки. Число связей между доменами - количество белков функциональных групп хроматина (подробнее в тексте), в которых присутствует пара доменов.

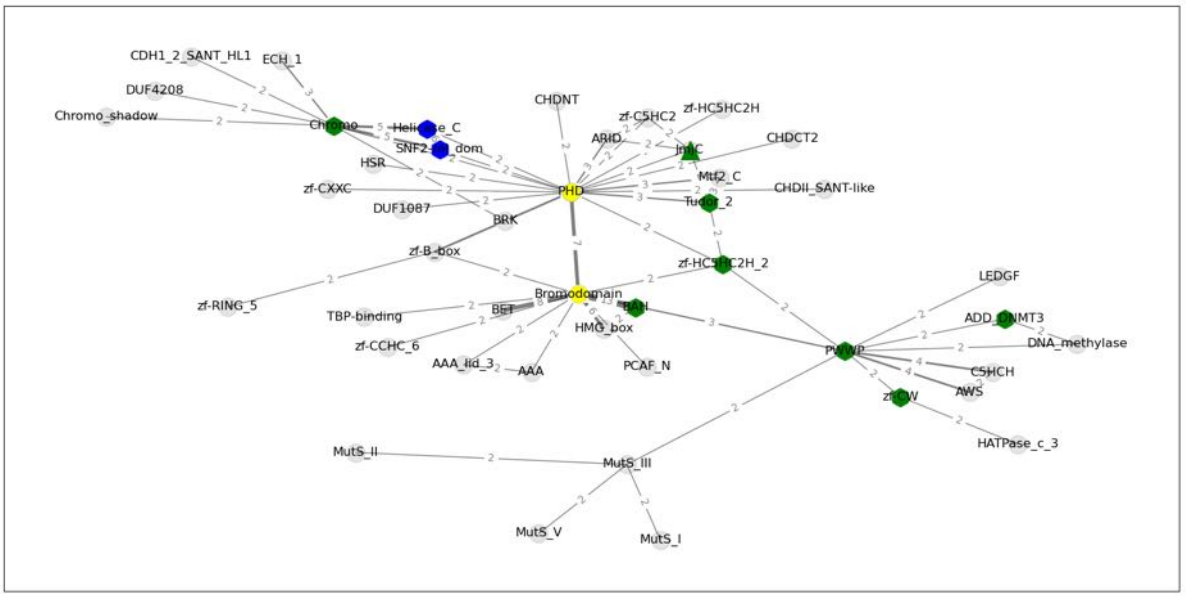


Рисунок 19. Доменная организация белков, узнающих ПТМ гистонов, из разработанной классификации. Ковстречаемость одного домена в белках не показана. Вес связей - количество белков с данной комбинацией доменов. Цветовая схема аналогична **Рисунку 18**.

Полученные результаты, **Рисунок 18** и **Рисунок 19**, свидетельствуют о наличии мультидоменных белков хроматина, которые за счет доменной архитектуры способны осуществлять несколько функций. Результаты уточняют и дополняют описанную в литературе ковстречаемость доменов, узнающих ПТМ гистонов, в белках человеческого протеома [23].

Не менее важной чертой белков являются неупорядоченные регионы. В наборах белков внутреннего содержимого ядра медианное значение доли неупорядоченных регионов для двух третей наборов выше, чем медианное значение в белках цитоплазмы, **Рисунок 20**. Однако различий в распределении зарядов в неупорядоченных концевых фрагментов белков хроматина по сравнению с белками цитоплазмы не было выявлено, **Рисунок 21**.

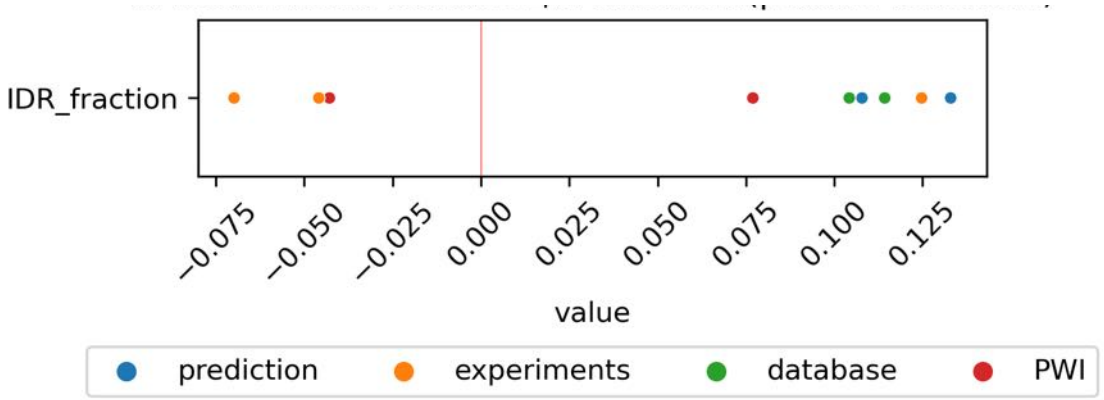


Рисунок 20. Различие доли неупорядоченных регионов белков хроматина и ядра относительно соответствующего значения белков цитоплазмы (красная линия). Наборы белков хроматина сгруппированы по типу источника информации (эксперименты, базы данных, предсказания, протеомные инициативы).

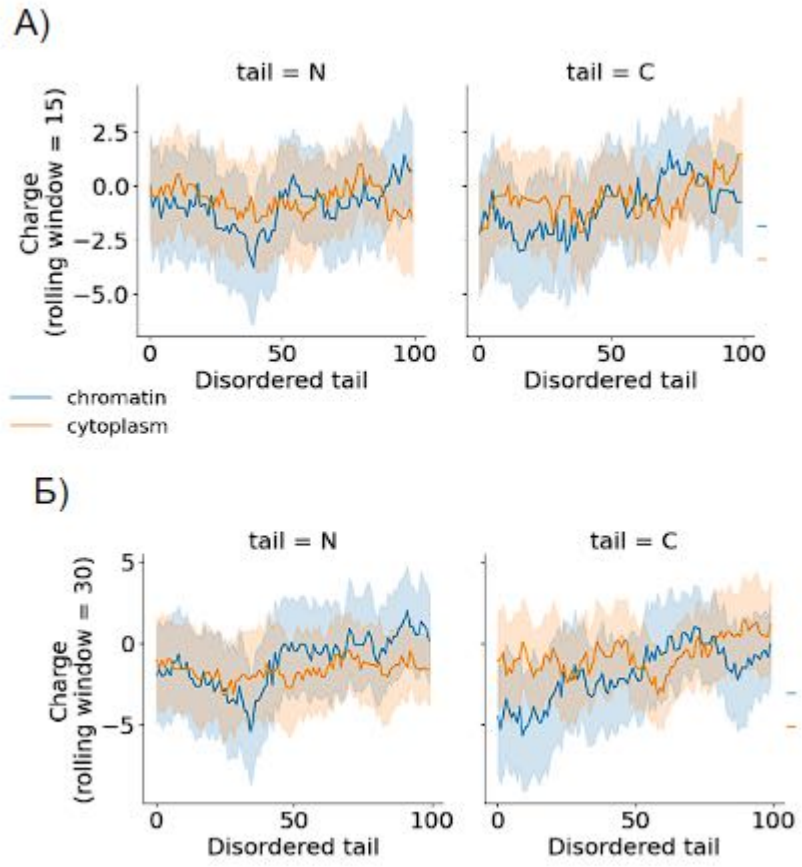


Рисунок 21. Распределение зарядов в неупорядоченных концевых фрагментах белков хроматина и цитоплазмы, для скользящего окна размером А) 15 а.о., Б) 30 а.о. N-конец белков слева, С-конец справа.

Глава 4. Анализ разнообразия структур нуклеосом и их комплексов с негистоновыми белками хроматина

4.1. Представленность вариантов гистонов в структурах нуклеосом

Мы провели качественный и количественный анализ гистоновых и негистоновых белков структур нуклеосом и их комплексов, депонированных в базе NucleosomeDB (по состоянию на 20 сентября 2022-года). Источником информации о гистоновых вариантах была база HistoneDB 2.0 [15], классификация негистоновых белков хроматина проводилась с помощью разработанной эмпирической классификации белков хроматина (описанной в **Разделе 3.3**) и ручной курации.

По состоянию на 20 сентября 2022 года в базе была депонирована 441 структура, из которых в 193 структурах были негистоновые белковые партнеры. В результате анализа представленности гистонов в структурах было выявлено, что гистон H3 - самый представленный, в структурах содержатся следующие варианты: cenH3, H3.3, TS H3.4, H3.5, H3.6, H3.Y. (то есть все основные, наиболее изученные варианты). Варианты гистона H2A представлены следующие: H2A.1, H2A.B, H2A.X, H2A.Z и macroH2A. Однако macroH2A присутствует в неполной форме (107 из 370 аминокислотных остатков, отсутствует макро-домен). Структуры нуклеосом с вариантами H2A.L, H2A.P, H2A.W - отсутствуют в PDB. Последний упомянутый вариант H2A.W - специфичен для растений и содержит уникальный мотив SPKK на С-конце, влияние которого на структуру нуклеосом еще предстоит охарактеризовать. Варианты гистонов H2B описываются одним вариантом - H2B.1, в то время как другие варианты гистонов, участвующие в спермиогенезе (H2B.W, sperm H2B и subH2B), отсутствуют.

Как уже было упомянуто, для гистонов характерна бОльшая изменчивость на уровне гистоновых вариантов одного вида, по сравнению с изменчивостью последовательностей одного варианта между видами живых организмов. Однако даже последовательности канонических гистонов различаются в разных видах живых организмов. Большинство структур нуклеосом содержат гистоны человека и модельных организмов: *Homo sapiens* (219 структур), *Xenopus laevis* (172), *Saccharomyces cerevisiae* (12), *Drosophila melanogaster* (8), *Mus musculus* (7). Однако есть и нетипичные структуры: структура нуклеосомы с гистонами метилотрофных дрожжей *Komagataella pastoris*, которых используют при наработке белка (PDB ID: 7WLR) [149]; нуклеосома патогенного одноклеточного эукариотического паразита *Giardia lamblia* (PDB ID: 7D69) [150] и нуклеосома человека с включением гистона H3 внутриклеточного паразита *Leishmania* (PDB ID: 6KXV) [151]. Гистоны архей присутствуют в одной структуре так называемой "гипернуклеосомы" из *Methanothermus fervidus* (PDB ID: 5T5K) [152], в которой

ДНК обернута вокруг "бесконечного" кора гистонов. В 2021 были разрешены структуры с гистонами вирусов группы *Marseilleviridae* (гигантские вирусы, заражающие амеб), которые кодируют свои гистоны (PDB IDs: 7LV8, 7LV9, 7N8N) [153,154]. Гистоны *Marseilleviridae* присутствуют в дублетах (H4 сливается с H3, а H2B сливается с H2A) и образуют частицы, похожие на эукариотические "канонические" нуклеосомы, хотя и менее стабильные, содержащие только 120 п.о. ДНК.

Количество структур хроматосомы, то есть нуклеосомы с линкерным гистоном H1 в месте входа и выхода ДНК, составляет 33, 20 из которых содержат вариант гистона H1.4 человека. Также структуры нуклеосом содержат варианты человека H1.0 и H1.10, курицы H1.5 и по одной структуре с вариантами H1.0 и H1.8 лягушки (*Xenopus*). Не разрешены структуры с линкерными гистонами H1: scH1, TS H1.6, TS H1.7 и TS H1.9.

4.2. Качественный и количественный анализ белков-партнеров нуклеосомы по структурным данным

Регуляция хроматина осуществляется в том числе за счет взаимодействия негистоновых белков с гистонами и ДНК нуклеосомы. В базе NucleosomeDB 193 структуры содержат негистоновые белки. Первые шесть структур были разрешены методом рентгено-структурного анализа (РСА) в период с 2006 по 2013 год. В тех работах негистоновые белки партнеры были относительно небольшими пептидами и белками - фрагмент белка LANA герпесвируса саркомы Капоши (PDB ID: 1ZLA), фактора конденсации хромосом (RCC1, PDB ID: 3MVD), регуляторный белок SIR3 (PDB IDs: 3TU4, 4JJN, 4KUD, 4LD9). С 2016 года количество комплексов резко увеличилось, большинство из них получены с помощью крио-электронной микроскопии (крио-ЭМ) с относительно низким разрешением. Наибольшее количество структур комплексов нуклеосом с негистоновыми белками были получены в 2019 и 2020 годах - 60 и 38 структур, соответственно (против 21 и 29 структур одиночных нуклеосом). Молекулярная масса взаимодействующих с нуклеосомой небелковых компонентов в структурах варьирует от 0,4 до 687 кДа, медианное значение 53 кДа. Самым тяжелым комплексом является комплекс гистоновой ацетилтрансферазы NuA4 (PDB ID: 7VVZ).

Количество структурных негистоновых белков, взаимодействующих с нуклеосомой составляет 215. Большинство белков принадлежит к белкам *Saccharomyces cerevisiae* (96 белков в 54 структурах) и белкам человека (78 белков в 102 структурах). Интересно, что только 15 нуклеосом содержат гистоны *Saccharomyces*. Половина негистоновых белков-партнеров (94 из 215) присутствует только в одной структуре, а 89 - в двух и более структурах. Как известно, не все белковые последовательности разрешаются в структурах, к ним относятся неупорядоченные регионы, однако длина разрешенных белковых фрагментов увеличивается с годами, **Рисунок**

22А.

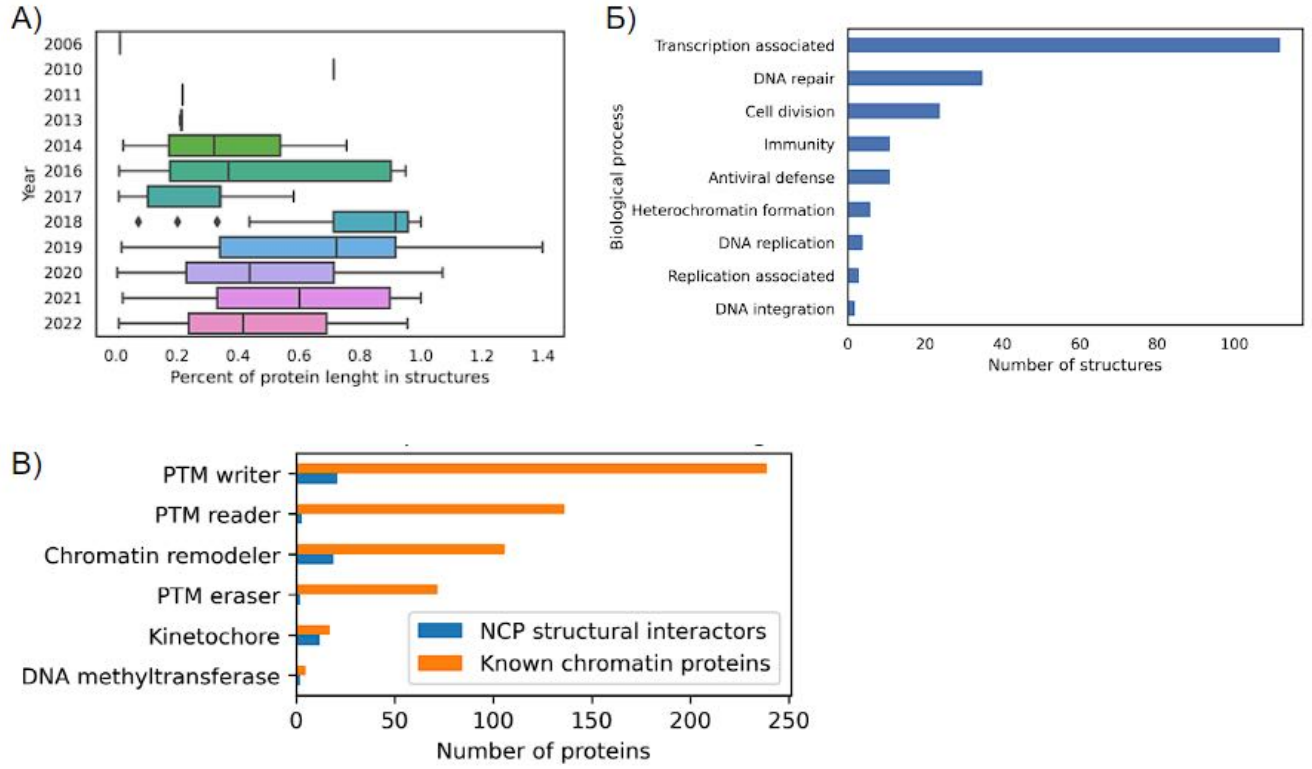


Рисунок 22. А) Изменение разрешенной в структурах комплексов нуклеосом с негистонвоыми белками длины белков. Б) Количество структур комплексов нуклеосом с негистоновыми белками, ассоциированными с биологическими процессами. В) Сравнение количества известных белков хроматина некоторых функциональных категорий и белков из структур комплексов нуклеосом. По [155].

Согласно разработанной эмпирической классификации белков хроматина, большинство структур нуклеосом с негистоновыми партнерами ассоциированы со следующим биологическим процессам: транскрипция (112 структур), репарация ДНК (34) и деление клеток (24), **Рисунок 22Б**. К наиболее представленным в структурах относятся белки следующих функциональных групп: наносящие ПТМ гистонов (56 структур), ремоделеры хроматина (40), транскрипционные факторы(19), РНК-полимеразы (17) и комплексы с циклической GMP-AMP синтазой (гуанозин монофосфат-аденозин монофосфат синтазой) (cGAS) (11), **Рисунок 23.** Не смотря на 17 структур нуклеосомы с РНК полимеразой, даже в структурах с гистонами человека, белки-партнеры из *Saccharomyces cerevisiae* или *Komagataella phaffii*.

Также мы проанализировали представленность белков хроматина человека различных функциональных классов в структурах комплексов нуклеосом, **Рисунок 22Б**. Наиболее представленным в структурах классом являются компоненты кинетохора (70% белков данного класса есть в структурах нуклеосом). В других классах (ремоделирование хроматина, нанесение, считывание и удаление ПТМ гистонов) в структурах представлено от 2 до 18 % белков.

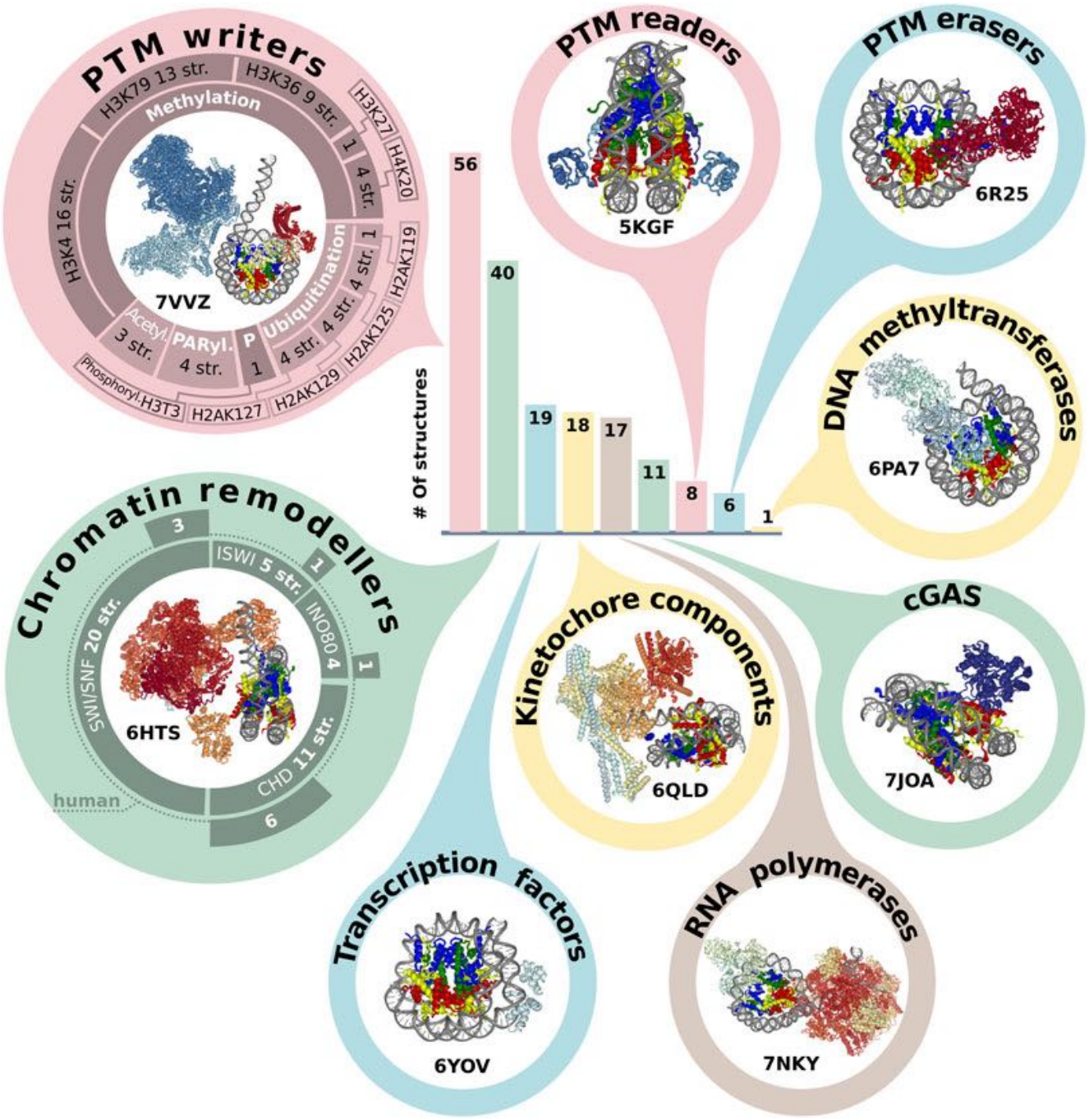


Рисунок 23. Разнообразие негистоновых белков в структурах комплексов нуклеосом. Изображено количество структур, ассоциированных с молекулярными функциями. Структуры представителей класса изображены кружками с PDB-идентификаторами. На панеле вокруг белков, наносящих ПТМ гистонов - разбиение структур по типу и сайту метки (метилирование, убиквитинирование, ацетилирование, PARилирование и фосфорилирование). На панеле вокруг ремоделеров хроматина представлено количество структур с делением на семейства, которые дополнительно подразделяются на комплексы белков человека и других видов живых организмов. В инфографику не включены следующие категории: шапероны гистонов (три структуры), обмен гистонов (две структуры), интеграция ДНК вируса (две структуры). По [155].

Одним из уровней регуляции эпигенетики являются посттрансляционные модификации гистонов, которые могут изменять структурную динамику нуклеосом и регулировать доступность ДНК [156–158]. С помощью разработанного подхода классификации негистоновых

белков, белкам из классов наносящих, считывающих и удаляющих ПТМ гистонов были дополнительно присвоены теги по типу модификации (метилирование, ацетилирование и др.) и аминокислотной позиции гистона. Мы проанализировали структуры комплексов нуклеосом на предмет ассоциации с ПТМ гистонов и выявили, что одна треть комплексов связана с метилированием и одна седьмая - с убиквитинированием, **Рисунок 23**. По аннотации HISTome 2 [90] наиболее изученные сайты метилирования у человека: H3K4, H3K9, H3K27, H3K36, H3K79, H4K20, H3R8, H4R3. Наибольшее количество структур связано с метилированием H3K4 (16 структур) преимущественно комплексом COMPASS, эта метка указывает на активные промоторы генов. Метилирование H3K79 - один из маркеров регуляции транскрипции и ответа на повреждение ДНК. Белки, ассоциированные с этой меткой, присутствуют в 13 структурах, а белки, ассоциированные с меткой активного хроматина H3K36 - в 9 структурах.

На момент анализа 8 из 30 человеческих лизин-метилтрансфераз гистонов были представлены в структурах комплексов нуклеосом, а структуры комплексов нуклеосом с человеческими метилтрансферазами для сайтов H3K9 и H4R3 отсутствовали. Первые структуры с ацетилтрансферазами появились только в 2022 году (PDB: 7W9V, 7VVZ, 7VVU). Менее 10 структур связаны с записью и считыванием ПТМ следующих сайтов: фосфорилирование H3T3; метилирование H3K27, H4K20; деметилирование H3K4, H3K9 и убиквитинирование H2AK15, H2A119, H2AK127, H2AK125, H2AK129, H2BK120, **Рисунок 23**.

Другим ключевым функциональным классом белков хроматина является класс ремоделеров хроматина, которые изменяют состав и/или расположение нуклеосом, затрачивая энергию АТФ. Революция разрешения в крио-электронной микроскопии способствовала появлению в 2017 году первых структур комплексов нуклеосом с хроматиновыми ремоделерами. Ремоделеры хроматина разделены на четыре семейства, наиболее представленным в комплексах с нуклеосомами является семейство SWI/SNF (29 структур), далее следуют CHD (13), INO80 (6) и ISWI (5). Депонированные в NucleosomeDB структуры охватывают все известные семейства ремоделеров хроматина, и даже, несмотря на низкое разрешение, позволяют получать механистические представления о взаимодействии белков с ДНК.

4.2. Структуры комплексов нуклеосом с патологическими изменениями

Для выявления структур нуклеосом с известными онкологическими мутациями был применен поиск по ключевым словам в текстовой информации о структуре нуклеосомы. В результате были найдены 23 структуры со следующими мутациями в гистонах: замены лизина на глутамин (KQ) в гистон-фолдовых доменах H3 и H4 [159], онкогистоновые мутации H3.3K36M (нуклеосомы дрожжей и человека) [160], мутации в аминокислотных остатках H3 и

H4, расположенных на интерфейсах взаимодействия белок-ДНК вблизи нуклеосомной диады [161].

Для выявления среди негистоновых белков онкогенов и генов-супрессоров опухолей была использована аннотация генов OncoKB [162]. Так как только 78 из 215 белковых партнеров нуклеосом в структурах относятся к белкам человека, для увеличения числа белков мы рассматривали также белки-ортологи человека по аннотации OrthoDB [163]. В результате 24 из 105 генов человека аннотированы в OncoKB [162]: пять онкогенов (EZH2, NSD2, SOX2, DOT1L, SMARCE1) и 18 генов-супрессоров опухолей (ARID1A, BRCA1, DNMT3A, EP300, EZH2, PBRM1, SETD2, SMARCA4, SMARCB1, ARID2, BARD1, KMT2A, KMT2C, SUZ12, PARP1, DNMT3B, SMARCE1, TP53BP1).

Разработка новых соединений для эпигенетических терапий требует детальной структурной информации об интерфейсе взаимодействия между нуклеосомой и белком-партнером. Из структурных белков-партнеров нуклеосомы только 4 негистоновых белка одобрены в качестве мишеней для эпигенетических препаратов: Поли [ADP-рибоза] полимераза 1 (P09874), поли [ADP-рибоза] полимераза 2 (Q9UGN5), ДНК (цитозин-5)-метилтрансфераза 3A (Q9Y6K1), гистон-лизин N-метилтрансфераза EZH2 (Q15910). В клинических испытаниях находится одна молекула, нацеленная на лизин-специфичную деметилазу гистонов 1A (O60341), по данным базы данных биологически активных молекул с лекарственно-подобными свойствами ChEMBL [164].

Глава 5. Биоинформатический и структурный анализ геномных и транскриптомных данных опухолей с точки зрения организации хроматина на нуклеосомном уровне

5.1. Нарушения экспрессии гистонов и белков хроматина по данным TCGA

Мы проанализировали предобработанные данные экспрессии гистонов и белков хроматина, для которых выявлены взаимодействия с нуклеосомами на уровне структур, в образцах 15 типов онкологических заболеваний консорциума The Cancer Genome Atlas (TCGA) и в образцах соответствующих тканей здоровых доноров консорциума The Genotype-Tissue Expression (GTEx) [165], загруженных с портала UCSC Xena [166]. Предобработка данных экспрессии включала выравнивание прочтений на референсный транскриптом методом STAR и подсчет каунтов методом RNA-Seq by Expectation-Maximization (RSEM), что позволило сравнить наборы данных из TCGA и GTEx между собой. Общее количество образцов составило 19131, из них 11521 образцов пациентов с онкологическими заболеваниями. Для анализа были выбраны когорты с количеством образцов более 50. Количественные и качественные характеристики проанализированных наборов данных представлены в **Таблице 2**.

Таблица 2. Проанализированные датасеты экспрессии генов RNA-seq: опухолевые данные из TCGA, референсные данные соответствующих тканей без опухоли из консорциума GTEX; для каждого датасета приведено количество секвенированных образцов.

№	Тип заболевания	TCGA аббре-виатура	датасет TCGA	кол-во обр. в TCGA	датасет GTEx	кол-во обр. в GTEx
1	Острая миелоидная лейкемия	LAML	Acute Myeloid Leukemia	173	Blood	337
2	Глиома головного мозга 2 степени	LGG	Brain Lower Grade Glioma	523	Brain	1141
3	Инвазивная карцинома молочной железы	BRCA	Breast invasive carcinoma	1099	Breast	179
4	Аденокарцинома толстой кишки	COAD	Colon adenocarcinoma	290	Colon	308
5	Рак пищевода	ESCA	Esophageal carcinoma	182	Esophagus	653
6	Мультиформная глиобластома	GBM	Glioblastoma multiforme	166	Brain	1141
7	Гепатоцеллюлярная карцинома	LIHC	Liver hepatocellular carcinoma	371	Liver	110
8	Серозная цистаденокарцинома яичника	OV	Ovarian serous cystadenocarcinoma	427	Ovary	88

9	Аденокарцинома поджелудочной железы	PAAD	Pancreatic adenocarcinoma	179	Pancreas	167
10	Аденокарцинома предстательной железы	PRAD	Prostate adenocarcinoma	496	Prostate	100
11	Аденокарцинома желудка	STAD	Stomach adenocarcinoma	414	Stomach	174
12	Опухоли тестикулярных эмбриональных клеток	TGCT	Testicular Germ Cell Tumors	154	Testis	165
13	Карцинома щитовидной железы	THCA	Thyroid carcinoma	512	Thyroid	279
14	Карциносаркома матки	UCS	Uterine Carcinosarcoma	57	Uterus	78
15	Рак шейки матки	UCEC	Uterine Corpus Endometrial Carcinoma	181	Uterus	78

Для 15-ти онкологических заболевания были оценены дифференциально экспрессирующиеся гены с помощью библиотеки DESeq2. Значимым считалось изменение экспрессии $|\log2FoldChange| > 1$ при скорректированном p-value < 0,01. Был получен список генов гистонов и белков из структуры комплексов с нуклеосомами, для которых обнаружены статистически значимые отклонения в уровнях экспрессии в образцах раковых опухолей по сравнению с нормальными тканями. Визуальное отображение в виде тепловой карты для всех типов опухолей приведено на **Рисунке 24А**.

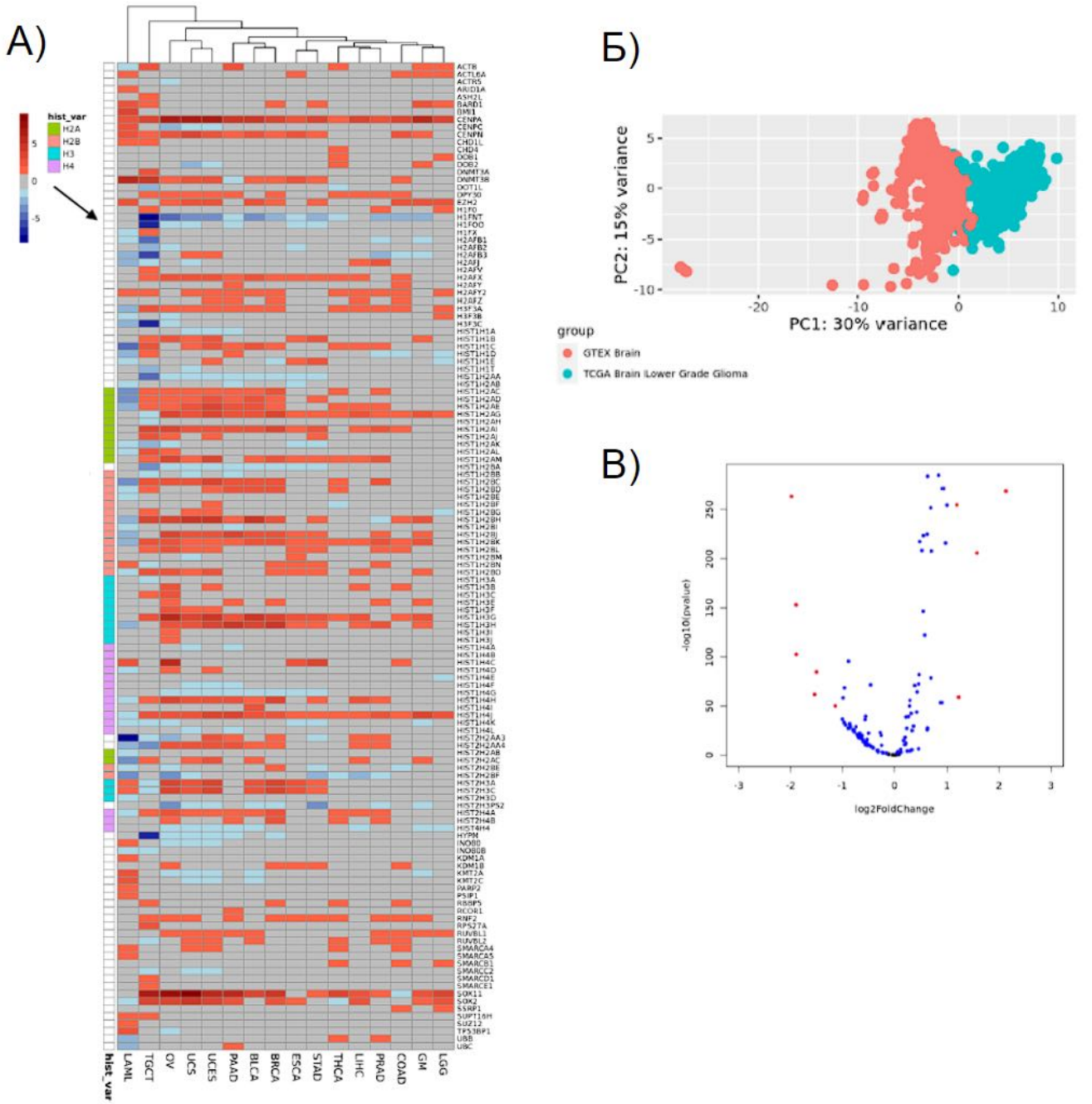


Рисунок 24. Анализ дифференциальной экспрессии генов гистонов и белков-партнеров (из структурного интерактома) по данным секвенирования РНК пациентов с онкологическими заболеваниями. Для каждого типа опухоли сравнение проводилось относительно соответствующих здоровых тканей из консорциума GTEx (приведены в **Таблице 2**).

А) Тепловая карта дифференциальной экспрессии генов для всех проанализированных типов онкологических заболеваний, единицы значений - log2FoldChange, цветом выделены значения > 1 (красным), и <-1 (синим) при скорректированном p-value < 0.01, серым - статистически незначимые значения или log2FoldChange в диапазоне -1 - 1. Гены канонических гистонов проаннотированы на панели слева. По оси абсцисс аббревиатуры онкологических заболеваний по номенклатуре TCGA (расшифровка в **Таблице 2**).

Б) График PCA первых двух компонент экспрессии генов гистонов и их партнеров в размеченных когортах на примере TCGA Brain Lower Grade Glioma и GTEx Brain. Видно, что образцы в когортах хорошо скоррелированы между собой, а когорты хорошо разделимы.

B) Volcano диаграмма дифференциальной экспрессии генов в TCGA Brain Lower Grade Glioma, синим выделены гены при padj < 0.01, красным при |log2FC|>1 и p-adj < 0.01.

Можно отметить различные паттерны экспрессии генов гистонов в различных типах онкологических заболеваний. Так, у центромерного гистона CENPA во всех наборах опухолевых данных повышена экспрессия, **Рисунок 24А**. В то же время, несмотря на высокую консервативность последовательностей канонических генов гистонов, их профили экспрессии довольно сильно различаются, **Рисунок 24А**. Например, у канонических изоформы гистона H2B: H2BC13 (HIST1H2BL) и H2BC17 (HIST1H2BO) повышена экспрессия в PRAD, COAD и в GM и LIHC, соответственно. Среди генов со статистически значимой повышенной экспрессией в THCA, TGCT, COAD, LAML представлены: ремоделеры хроматиновых (nBAF, SWI/SNF); белки, наносящие ПТМ (метилтрансферазы: Set1C/COMPASS, MLL3/4, ацетилтрансфераза: NuA4). В BLCA, PRAD, UCS и UCES повышена экспрессия у комплексов ремоделлеров SWR1, INO80.

Гены с пониженной экспрессией входят в состав комплексов: в TGCT, OV - INO80, в BLCA в комплекс, осуществляющий метилирование H3K4. Метка H3K4me кодирует активные промотеры и привлекает хроматиновые ремоделеры семейства ISWI. Пониженная экспрессия перечисленных выше генов в TGCT, OV и BLCA может приводить к более компактном состоянию хроматина и уменьшению транскрипции ряда генов.

5.2. Анализ дифференциальной экспрессии гистонов и других белков хроматина в образцах пациентов с множественной миеломой

Для выполнения исследования были использованы необработанные прочтения секвенирования РНК клеток CD138+ аспирата костного мозга пациентов с множественной миеломой (ММ). Выходящее за рамки диссертационной работы секвенирование тотальной РНК (очищенной от рРНК) было проведено на оборудовании Illumina HiSeq 3000 для одноконцевого секвенирования, длина прочтений 50 п.н., порядка 30 млн. «сырых» прочтений на образец. Проверка качества данных проводилась с помощью программы Illumina SAV. Де-мультиплексирование проводили с помощью программного обеспечения Illumina Bcl2fastq2 v 2.17.

Образцы аспирата костного мозга были получены от 58-ми пациентов (возраст от 29 лет до 79, 35 мужчин и 27 женщин). В качестве контроля были образцы 11-ти здоровых доноров (возраст от 24 лет до 41, 5 мужчин и 6 женщин), ранее опубликованных в рамках атласа секвенирования РНК здоровых тканей [167].

Нами был проведен контроль качества «сырых» прочтений был проведен с помощью FastQC [168]. FASTQ файлы были обработаны программной библиотекой Salmon [169] в

режиме mapping-based для количественного определения числа транскриптов с использованием индекса из последовательностей транскриптов (GENCODE release 32). Матрица экспрессии в пересчете на гены была построена с помощью библиотеки tximeta [170]. Идентификаторы генов из Ensembl конвертировали в символы генов HGNC с помощью библиотеки biomaRt [171]. Дифференциальную экспрессию оценивали с помощью пакета DESeq2 [172]. Дифференциально экспрессированными считались гены с |log2FoldChange| >2,0 и скорректированным p-value < 0,05.

Количество выровненных на гены прочтений (далее каунты) в образцах ММ и доноров варьировало от 19106 до 18683993 (оценка Salmon). Образцы (один от донора и шесть от больных ММ), в которых было получено менее 2,2 млн прочтений, были исключены из дальнейшего анализа. К таблице каунтов была применено преобразование, стабилизирующее дисперсию (variance stabilizing transformations, VST), далее расчитано евклидово расстояние между образцами и проведена иерархическая кластеризация образцов. Отфильтрованные образцы объединялись в обособленный кластер. Таким образом, для дальнейшего анализа дифференциальной экспрессии генов было отобрано 62 образца (52 из ММ и 10 от доноров).

Проведенный анализ дифференциальной экспрессии генов выявил 629 генов с повышенной и 472 гена с повышенной экспрессией в образцах пациентов относительно контрольных образцов. При проведении для найденных генов обогащения терминов Gene Ontology были выявлены термины, связанные с функционированием хроматина, поэтому был проведен анализ обогащения генов по разработанной эмпирической классификации белков хроматина. Среди категорий, обогащенных в дифференциально экспрессируемых генах были выявлены, в частности: гистоны; рецепторы гормонов; белки HMG; белки, связанные с эухроматином; белки транскрипции и ее регуляция, **Рисунок 25А**.Интересно, что среди генов с пониженной экспрессией присутствовали следующие канонические гены гистонов: семь гистонов H3 (H3C2, H3C3, H3C7, H3C8, H3C10, H3C11, H3C14), шесть H2A (H2AC7, H2AC13, H2AC14, H2AC15, H2AC16, H2AC17), пять гистонов H2B (H2BC6, H2BC9, H2BC13, H2BC14, H2BC17) и четыре канонических гистона H4 (H4C1, H4C6, H4C9, H4C13), **Рисунок 25Б**. Среди вариантных гистонов с пониженной экспрессией - MACROH2A2 и гены H1-0, H1-1, H1-5. На уровне белков изменения в содержании гистонов, вероятно, не столь значительны, поскольку гены гистонов могут кодировать одинаковые белки, **Рисунок 25Б**.

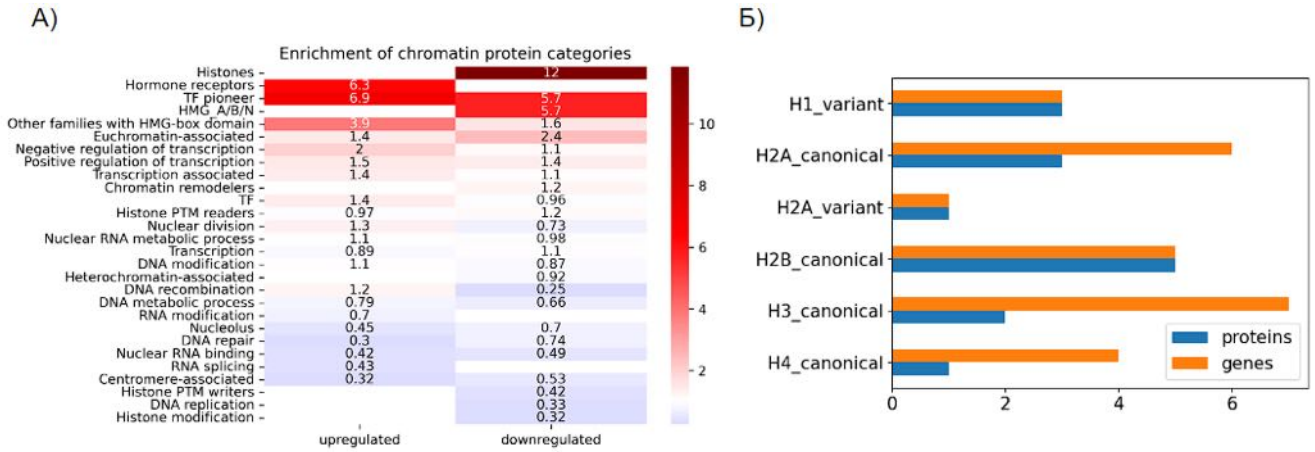


Рисунок 25. Анализ дифференциально экспрессированных генов образцов пациентов с ММ. А) Обогащение терминами разработанной эмпирической классификации белков хроматина среди дифференциально экспрессированных генов в образцах пациентов с ММ. Б) Количество генов и белков гистонов с пониженной экспрессией в образцах пациентов с ММ с классификацией по типу гистонов и каноничности.

5.3. Онкомутации в гистонах по данным TCGA

Были загружены мутации в образцах пациентов с онкологическими заболеваниями по данным консорциума TCGA с помощью портала cBioPortal [173]. Реккурентными считались мутации, обнаруженные в девяти образцах разных пациентов для канонических гистонов и в трех - для вариантов гистонов. Мутации в белках канонических гистонов анализировали совместно для каждого типа гистонов. Таким образом, рекуррентные мутации гистонов были выявлены в 4723 образцах, **Рисунок 26**.

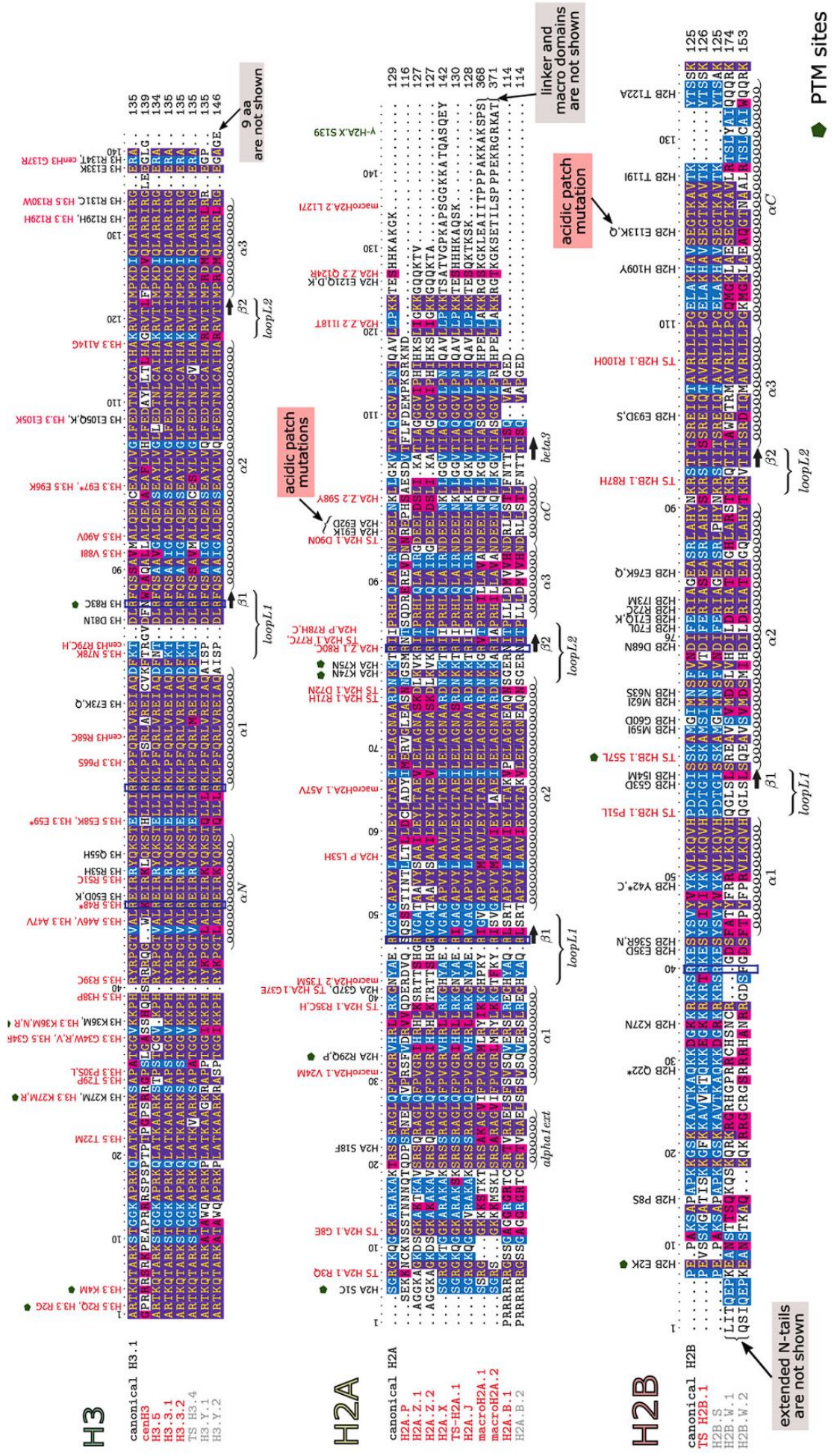

H3
H2A
H2B
canonical H3.1
cenH3
H3.5
H3.3.1
H3.3.2
TS H3.4
H3.Y.1
H3.Y.2
canonical H2A
H2A.P
H2A.Z.1
H2A.Z.2
H2A.X
TS-H2A.1
H2A.J
macroH2A.1
macroH2A.2
H2A.B.1
H2A.B.2
canonical H2B
TS H2B.1
H2B.S
H2B.W.1
H2B.W.2
9 aa are not shown
linker and macro domains are not shown
acidic patch mutations
acidic patch mutation
extended N-tails are not shown
PTM sites

Рисунок 26. Онкологические мутации в гистонах человека. Множественное выравнивание последовательностей гистонов типов H2A, H2B и H3, окрашивание по сходству последовательностей. Структурные элементы гистонов аннотированы под выравниванием; синими рамками выделены аргинины, проникающие в малую бороздку ДНК. Рекуррентные онкомутации для канонических гистонов и гистоновых вариантов аннотированы над выравниваниями черным и красным цветом, соответственно. Визуализация множественного выравнивания последовательностей выполнена с помощью TexShade [174]. По [3].

В гистоне H3.3 было выявлено 12 реккурентных мутаций, в H3.5 - 15. H3.5 - специфичный для семенников вариант, встраивание которого в нуклеосомы в районе сайта старта транскрипции приводит к их меньшей стабильности по сравнению с каноническими [12]. Однако эффекты мутаций остаются неизученными. TS H2A.1 - менее изученный вариант гистона, сходный с каноническим и специфичный для для семенников, ооцитов и зигот млекопитающих. Для TS H2A.1 наш анализ выявил 7 мутаций. Мутация TS H2A.1 G37E находится в той же позиции, что и другая онкомутация G37D в каноническом гистоне H2A; обе эти мутации вводят отрицательно заряженные аминокислоты в конец α-1 спирали гистонового фолда в H2A. Интересно, что в этой позиции также наблюдаются замены, изменяющие заряд G>R в H2A.Z и G>E в H2A.P - дивергентном варианте, лишенном C-концевого хвоста, который присутствует у плацентарных млекопитающих (eutheria). Среди вариантов H2B мутации, по результатам проведенного анализа, присутствуют только в TS H2B.1 (4 мутации), который, возможно, вместе с TS H2A.1 участвует в перепрограммировании клеток [175]; в то время как в вариант H2B.W, участвующем в сперматогенезе, не было выявлено рекуррентных онкомутаций.

5.3. Структурная интерпретация мутаций белков, взаимодействующих с нуклеосомой

Для проведения структурной интерпретации мутаций, находящихся на поверхности взаимодействий негистоновых белков с гистонами нуклеосомы, были загружены:

- структуры комплексов нуклеосом из базы NucleosomeDB;
- мутации образцов пациентов с онкологическими заболеваниями по данным консорциума TCGA с помощью портала cBioPortal.

В структурах 72 структурах комплексов нуклеосом было выявлено 54 негистоновых белка человека, в которых, по данным TCGA, 15352 онкомутации. Далее путем выравнивания последовательности белка из структуры на полную последовательность из базы UniProt, было проведено картирование мутаций на аминокислотные остатки в структурах. Количество мутаций негистоновых белков, которые находятся в разрешенных фрагментах структур составило 12696. Для 10762 мутаций референсный аминокислотный остаток совпадал с таковым в структуре PDB, а для 1934 отличается, что может свидетельствовать, во-первых, о

полиморфизмах или врожденных мутациях в клетках крови пациента, которые являлись референсом, а во-вторых, о мутациях белков в процессе подготовки к кристаллизации. Взаимодействие между негистоновыми и гистоновыми белками в структурах комплексов выполнялось с помощью библиотеки MDAnalysis и определялось расстоянием между аминокислотными остатками - не более 4,5 Å. Для 5132 мутаций были найдены контакты между гистоновыми и негистоновыми белками (на уровне белковых цепей), а 236 мутаций находились непосредственно в сайтах контактов. Онкомутации были отфильтрованы по расположению на поверхности взаимодействия белков и соответствии референса последовательности белка в UniProt и PDB.

Для пар гистоновых и негистоновых белков из 56-ти структур, содержащих отобранные 4926 неуникальные онкомутации, был проведен следующий структурный анализ. Поверхность взаимодействия белков была оптимизирована в полуэмпирическом силовом поле FoldX [176]. Для каждой мутации рассчитывалось изменение энергии связывания негистоновых белков с нуклеосомой (ΔΔG) по формуле 6:

$$\Delta\Delta G = \Delta G(mutant) - \Delta G(WT), \quad (6)$$

где WT - комплекс без мутации. Описываемая в литературе погрешность FoldX составляет 0.5 ккал/моль, поэтому были выбраны следующие пороговые значения для выявления стабилизирующих (ΔΔG < -1) и дестабилизирующих (ΔΔG > 1) мутаций.

Гистограмма распределения значений ΔΔG для проанализированных мутаций изображена на **Рисунке 27**. Рассчитанные значения ΔΔG находятся в диапазоне от -8.5 ккал/моль (мутация G655R в Histone-lysine N-methyltransferase EZH2) до +2.5 ккал/моль (мутация G137R в histone H3K79 methyltransferase, **Рисунок 27Г**). Так как для анализа использовались все доступные в PDB структуры комплексов нуклеосом, эффект ряда мутаций был проанализирован на более чем одной структуре. Таким образом, было проанализировано 1552 уникальных мутаций, 1472 из которых являются нейтральными, 51 стабилизирующая и 36 дестабилизирующих.

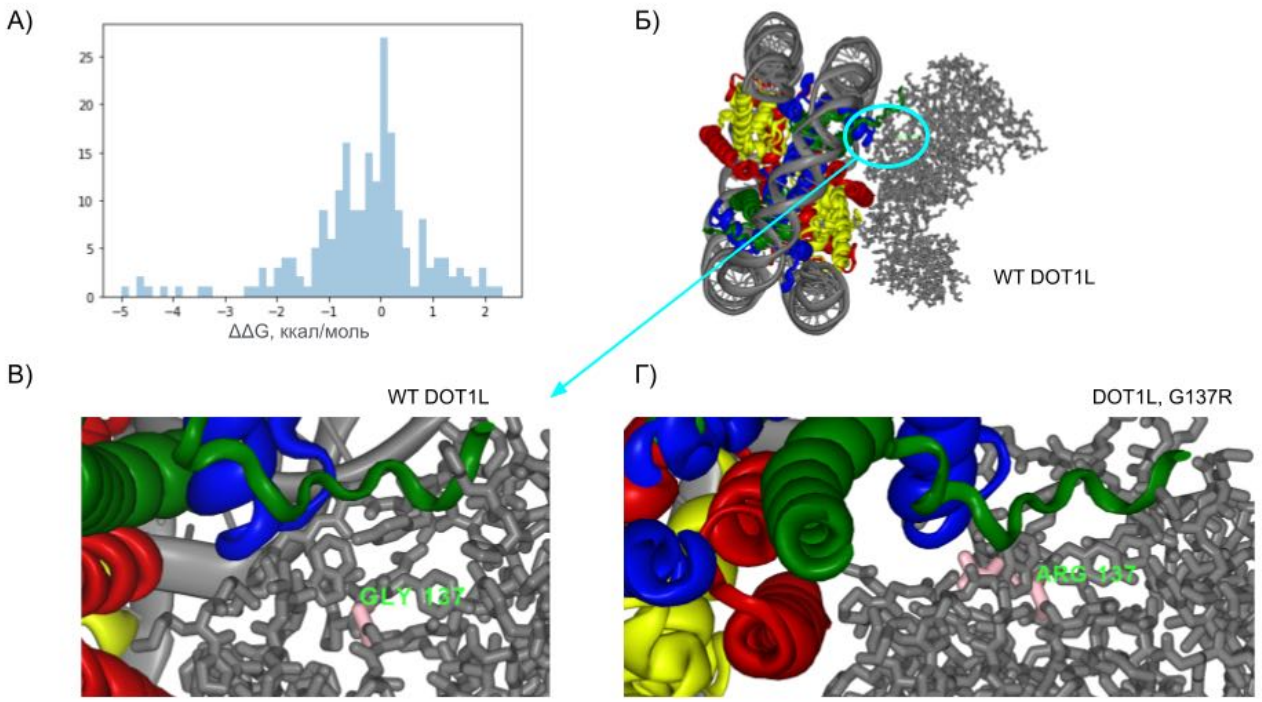


Рисунок 27. А) Гистограмма распределения значений изменения энергии связывания (ΔΔG, ккал/моль) при онкомутациях в негистоновых белках.
Б,В,Г) Структура нуклеосомы с WT H3K79-специфичной гистоновой метилтрансферазой DOT1L и с мутацией (G137R) (PDB ID: 6J99); голубым овалом выделен сайт мутации; DOT1L G137R, которая приводит к максимальному из проанализированных мутаций увеличению энергии связывания DOT1L с нуклеосомой на 2.5 ккал/моль.

Из стабилизирующих или дестабилизирующих мутаций 37 находятся в сайтах контактов негистоновых белков хроматина с нуклеосомой, из которых 3 встречаются в более чем 2-ух пациентах (EZH2 F667L, DOT1L W305C, RNF2 R98I). Интересны мутации, не находящиеся в сайтах контактов, но приводящие к увеличению или уменьшению энергии связывания. Среди 9 мутаций, найденных в образцах более чем 2 пациентов: 5 дестабилизирующих (CHD4 R877W, R1105W, R1105Q; DOT1L S132F; EZH2 T586I) и 4 стабилизирующих (CGAS R339H; CHD4 R813C, R877Q; DOT1L SK132F).

Эффект ряда мутаций рассчитывался на нескольких структурах. Для 7 мутаций эффект различается в зависимости от выбранных структуры: CENPC P527S, CGAS P257L, CGAS W330C, DOT1L: G137R, P122R, S132F, W305C (например, W305C дестабилизирующая мутация в структуре с PDB ID: 6NQA и стабилизирующая в структурах с PDB ID: 6JM9, 6J99).

Для обнаруженных онкомутаций были предприняты попытки объяснить эффект мутаций. Одна из найденных мутация в гистновой метилтрансферазе EZH2 F667L находится в ароматической полости, узнающей субстрат (позицию гистона H3K27). Соответственно, мутация приводит к более сильному связыванию метилтрансферазы с субстратом, что может замедлять активность PRC2 комплекса (в состав которого входит EZH2), способствующего

образованию гетерохроматина [177]. Метки, нанесенные комплексом PRC2 узнают белки комплекса PRC1. Одна из субъединиц PRC1 RNF2 убиквитинилирует гистоны H2A по K119. Вторая найденная нами стабилизирующая мутация находится в белке RNF2, она также может потенциально приводить к нарушению формирования гетерохроматина и дальнейшему аберрантному увеличению экспрессии генов.

Эффект мутации DOT1L (DOT1L - метилтрансфераза H3K79, метка H3K79me2/3 - метка активных энхансеров [178] W305C в не так хорошо исследован, и в литературе нет описания влияния мутации на функционирование белка, однако, мутация встречается в базах данных онкологических нарушений. Учитывая противоположный эффект мутации в разных структурах было бы интересно проверить эффект мутации экспериментально.

Среди рекуррентных мутаций не в сайтах контактов, но на взаимодействующих цепях негистоновых белков с гистонами нуклеосомы также обнаружены те, механизм действий которых подробно не описан в литературе, они могут служить кандидатами для экспериментальной проверки.

Заключение

Диссертационная работа посвящена комплексному и разностороннему исследованию белков хроматина человека. Описано разнообразие генов и белков гистонов человека, негистоновых белков хроматина человека и структур их комплексов с нуклеосомами. Проведено сравнение содержания источников информации о локализации белков и их функциональной аннотации, составлена и наполнена эмпирическая классификация белков хроматина. Выявлены особенности физико-химических свойств и доменной архитектуры белков хроматина человека. Проведен биоинформатический анализ геномных и транскриптомных данных опухолей с точки зрения выявления нарушения организации хроматина на уровне нуклеосом.

Выводы

1. Белковый состав ядра по базам данных UniProt или HPA в среднем на 62% пересекается с составом экспериментальных хроматомов, лучшее пересечение с составом белков ядерной фракции без дополнительной очистки.
2. Вычислена массовая доля гистонов относительно белков хроматина: она может составлять от 19 до 41%.
3. В белках хроматина меньше доля ароматических и отрицательно-заряженных аминокислот и больше доля положительно-заряженных аминокислот, по сравнению с белками цитоплазмы.
4. Описано разнообразие гистоновых и негистоновых белков в структурах комплексов нуклеосом.
5. Выявлено влияние онкомутаций в негистоновых белках хроматина на стабильность комплексов с нуклеосомами.

Благодарности

Автор выражает благодарность

научному руководителю д.ф.-м.н. А.К. Шайтану за постановку задач, обсуждение результатов и помощь в выполнении работы;

к.ф.-м.н. Г.А. Армееву за предоставленные данные о структурах нуклеосом и их комплексов, за обсуждение результатов;

к.б.н. А.М. Сергеевой за предоставление прочтений секвенирования образцов РНК пациентов с множественной миеломой.

Работа выполнена при поддержке гранта Президента РФ МД-1131.2022.1.4.

Список сокращений

TCGA - The Genomic Data Commons

HGNC - HUGO Gene Nomenclature Committee

ПТМ - посттрансляционные модификации

ТФ - транскрипционные факторы

BER - эксцизионная репарация оснований

NER - эксцизионная репарация нуклеотидов

MMR - репарация ошибочно спаренных нуклеотидов

5meC - 5-метилцитозин

АФК, ROS - активные формы кислорода

ЦПД - циклобутан-пиримидиновые димеры

6-4ПП - (6-4) пиримидин-пиримидоновые фотопродукты

HMG - High Mobility Group

GO - Gene Ontology

BP - биологические процессы

MF - молекулярные функции

CC - и клеточная локализация

ECO - код свидетельства

ChEP - обогащение хроматина для протеомики, Chromatin Enrichment for Proteomics

DEMAC - обогащения на основе плотности для МС-анализа, density-based enrichment for MS analysis of chromatin

МС - масс-спектрометрический анализ

ppm - parts per million

РСА - рентгеноструктурный анализ

крио-ЭМ - криоэлектронная микроскопия

ММ - множественная миелома

Список литературы

1. Huiskamp. W. Ueber die Elektrolyse der Salze des Nucleohistons und Histons. // Hoppe-Seyler´s Zeitschrift für physiologische Chemie. 1901. Vol. 34, № 1. P. 32–54.
2. Butler J.A.V., Johns E.W., Phillips D.M.P. Recent investigations on histones and their functions // Progress in Biophysics and Molecular Biology. 1968. Vol. 18. P. 209–244.
3. Espiritu D. et al. Molecular Mechanisms of Oncogenesis through the Lens of Nucleosomes and Histones // J. Phys. Chem. B. 2021.
4. Armeev G.A. et al. Linking chromatin composition and structural dynamics at the nucleosome level // Current Opinion in Structural Biology. 2019. Vol. 56. P. 46–55.
5. Marinov G.K., Lynch M. Diversity and Divergence of Dinoflagellate Histone Proteins // G3: Genes|Genomes|Genetics. Oxford University Press, 2016. Vol. 6, № 2. P. 397.
6. Henneman B. et al. Structure and function of archaeal histones // PLOS Genetics. Public Library of Science, 2018. Vol. 14, № 9. P. e1007582.
7. Talbert P.B., Armache K.-J., Henikoff S. Viral histones: pickpocket's prize or primordial progenitor? // Epigenetics & Chromatin. 2022. Vol. 15, № 1. P. 21.
8. Hocher A. et al. Histone-organized chromatin in bacteria. bioRxiv, 2023. P. 2023.01.26.525422.
9. Marzluff W.F., Wagner E.J., Duronio R.J. Metabolism and regulation of canonical histone mRNAs: life without a poly(A) tail // Nat Rev Genet. 2008. Vol. 9, № 11. P. 843–854.
10. Talbert P.B., Henikoff S. Histone variants at a glance // Journal of Cell Science. 2021. Vol. 134, № 6. P. jcs244749.
11. Maze I. et al. Critical Role of Histone Turnover in Neuronal Transcription and Plasticity // Neuron. 2015. Vol. 87, № 1. P. 77–94.
12. Urahama T. et al. Histone H3.5 forms an unstable nucleosome and accumulates around transcription start sites in human testis // Epigenetics and Chromatin. BioMed Central, 2016. Vol. 9, № 1. P. 1–16.
13. Wiedemann S.M. et al. Identification and characterization of two novel primate-specific histone H3 variants, H3.X and H3.Y // Journal of Cell Biology. 2010. Vol. 190, № 5. P. 777–791.
14. Martire S., Banaszynski L.A. The roles of histone variants in fine-tuning chromatin organization and function: 9 // Nature Reviews Molecular Cell Biology. Nature Publishing Group, 2020. Vol. 21, № 9. P. 522–541.
15. Draizen E.J. et al. HistoneDB 2.0: a histone database with variants—an integrated resource to explore histones and their variants // Database. 2016. Vol. 2016. P. baw014.
16. Seal R.L. et al. A standardized nomenclature for mammalian histone genes // Epigenetics & Chromatin. 2022. Vol. 15, № 1. P. 34.
17. Talbert P.B. et al. A unified phylogeny-based nomenclature for histone variants // Epigenetics & Chromatin. 2012. Vol. 5, № 1. P. 7.
18. Millán-Zambrano G. et al. Histone post-translational modifications — cause and consequence of genome function // Nat Rev Genet. 2022.
19. Talbert P.B., Meers M.P., Henikoff S. Old cogs, new tricks: the evolution of gene expression in a chromatin context // Nat. Rev. Genet. 2019. Vol. 20, № 5. P. 283–297.
20. Vermeulen M. et al. Selective Anchoring of TFIID to Nucleosomes by Trimethylation of Histone H3 Lysine 4 // Cell. Elsevier, 2007. Vol. 131, № 1. P. 58–69.
21. Henikoff S., Shilatifard A. Histone modification: cause or cog? // Trends in Genetics. Elsevier, 2011. Vol. 27, № 10. P. 389–396.
22. Cano-Rodriguez D. et al. Writing of H3K4Me3 overcomes epigenetic silencing in a sustained but context-dependent manner: 1 // Nat Commun. Nature Publishing Group, 2016. Vol. 7, № 1. P. 12284.
23. Ruthenburg A.J. et al. Multivalent engagement of chromatin modifications by linked binding modules // Nat. Rev. Mol. Cell Biol. 2007/11/27 ed. 2007. Vol. 8, № 12. P. 983–994.
24. Jenuwein T., Allis C.D. Translating the Histone Code // Science. American Association for the Advancement of Science, 2001. Vol. 293, № 5532. P. 1074–1080.

25. Wu G. et al. Somatic histone H3 alterations in pediatric diffuse intrinsic pontine gliomas and non-brainstem glioblastomas // Nat Genet. 2012. Vol. 44, № 3. P. 251–253.
26. Schwartzentruber J. et al. Driver mutations in histone H3.3 and chromatin remodelling genes in paediatric glioblastoma // Nature. 2012. Vol. 482, № 7384. P. 226–231.
27. Behjati S. et al. Distinct H3F3A and H3F3B driver mutations define chondroblastoma and giant cell tumor of bone: 12 // Nature Genetics. Nature Publishing Group, 2013. Vol. 45, № 12. P. 1479–1482.
28. Zhao S. et al. Mutational landscape of uterine and ovarian carcinosarcomas implicates histone genes in epithelial-mesenchymal transition // Proc Natl Acad Sci U S A. 2016. Vol. 113, № 43. P. 12238–12243.
29. Okosun J. et al. Integrated genomic analysis identifies recurrent mutations and evolution patterns driving the initiation and progression of follicular lymphoma // Nat Genet. 2014. Vol. 46, № 2. P. 176–181.
30. Papillon-Cavanagh S. et al. Impaired H3K36 methylation defines a subset of head and neck squamous cell carcinomas // Nat Genet. 2017. Vol. 49, № 2. P. 180–185.
31. Nacev B.A. et al. The expanding landscape of “oncohistone” mutations in human cancers // Nature. 2019. Vol. 567, № 7749. P. 473–478.
32. Lewis P.W. et al. Inhibition of PRC2 activity by a gain-of-function H3 mutation found in pediatric glioblastoma // Science. 2013. Vol. 340, № 6134. P. 857–861.
33. Chan K.-M. et al. The histone H3.3K27M mutation in pediatric glioma reprograms H3K27 methylation and gene expression // Genes Dev. 2013. Vol. 27, № 9. P. 985–990.
34. Kuzmichev A. et al. Histone methyltransferase activity associated with a human multiprotein complex containing the Enhancer of Zeste protein // Genes Dev. 2002. Vol. 16, № 22. P. 2893–2905.
35. Jain S.U. et al. Histone H3.3 G34 mutations promote aberrant PRC2 activity and drive tumor progression // PNAS. National Academy of Sciences, 2020.
36. Lu C. et al. Histone H3K36 mutations promote sarcomagenesis through altered histone methylation landscape // Science. 2016. Vol. 352, № 6287. P. 844–849.
37. Fang D. et al. The histone H3.3K36M mutation reprograms the epigenome of chondroblastomas // Science. 2016. Vol. 352, № 6291. P. 1344–1348.
38. Brumbaugh J. et al. Inducible histone K-to-M mutations are dynamic tools to probe the physiological role of site-specific histone methylation in vitro and in vivo: 11 // Nature Cell Biology. Nature Publishing Group, 2019. Vol. 21, № 11. P. 1449–1461.
39. Fang J. et al. Cancer-driving H3G34V/R/D mutations block H3K36 methylation and H3K36me3-MutSα interaction // Proc Natl Acad Sci U S A. 2018. Vol. 115, № 38. P. 9598–9603.
40. Shi L. et al. Histone H3.3 G34 Mutations Alter Histone H3K36 and H3K27 Methylation In Cis // J Mol Biol. 2018. Vol. 430, № 11. P. 1562–1565.
41. Lutsik P. et al. Globally altered epigenetic landscape and delayed osteogenic differentiation in H3.3-G34W-mutant giant cell tumor of bone // Nat Commun. 2020. Vol. 11, № 1. P. 5414.
42. Bennett R.L. et al. A Mutation in Histone H2B Represents A New Class Of Oncogenic Driver // Cancer Discov. 2019. Vol. 9, № 10. P. 1438–1451.
43. Arimura Y. et al. Cancer-associated mutations of histones H2B, H3.1 and H2A.Z.1 affect the structure and stability of the nucleosome // Nucleic Acids Res. 2018. Vol. 46, № 19. P. 10007–10018.
44. Kalashnikova A.A. et al. The role of the nucleosome acidic patch in modulating higher order chromatin structure // J R Soc Interface. 2013. Vol. 10, № 82. P. 20121022.
45. Wan Y.C.E. et al. Cancer-associated histone mutation H2BG53D disrupts DNA-histone octamer interaction and promotes oncogenic phenotypes // Signal Transduct Target Ther. 2020. Vol. 5, № 1. P. 27.
46. Khare S.P. et al. Overexpression of histone variant H2A.1 and cellular transformation are related in N-nitrosodiethylamine-induced sequential hepatocarcinogenesis // Exp Biol Med (Maywood). 2011. Vol. 236, № 1. P. 30–35.

47. Bhattacharya S. et al. Histone isoform H2A1H promotes attainment of distinct physiological states by altering chromatin dynamics // Epigenetics and Chromatin. BioMed Central, 2017. Vol. 10, № 1. P. 1–19.
48. Singh R. et al. Corrigenda: Replication-dependent histone isoforms: a new source of complexity in chromatin structure and function // Nucleic Acids Res. 2018. Vol. 46, № 18. P. 9893–9894.
49. Singh R. et al. Proteomic profiling identifies specific histone species associated with leukemic and cancer cells // Clin Proteom. 2015. Vol. 12, № 1. P. 22.
50. Su C.-H. et al. An H2A histone isotype regulates estrogen receptor target genes by mediating enhancer-promoter-3′-UTR interactions in breast cancer cells // Nucleic Acids Res. 2014. Vol. 42, № 5. P. 3073–3088.
51. Monteiro F.L. et al. The histone H2A isoform Hist2h2ac is a novel regulator of proliferation and epithelial-mesenchymal transition in mammary epithelial and in breast cancer cells // Cancer Lett. 2017. Vol. 396. P. 42–52.
52. Long H. et al. H2A.Z facilitates licensing and activation of early replication origins: 7791 // Nature. Nature Publishing Group, 2020. Vol. 577, № 7791. P. 576–581.
53. Jung N. et al. Pharmacological unmasking microarray approach-based discovery of novel DNA methylation markers for hepatocellular carcinoma // J Korean Med Sci. 2012. Vol. 27, № 6. P. 594–604.
54. Salhia B. et al. Integrated Genomic and Epigenomic Analysis of Breast Cancer Brain Metastasis // PLoS One. 2014. Vol. 9, № 1.
55. Liu Y.-R. et al. Comprehensive Transcriptome Profiling Reveals Multigene Signatures in Triple-Negative Breast Cancer // Clin Cancer Res. 2016. Vol. 22, № 7. P. 1653–1662.
56. Berenguer-Daizé C. et al. OTX015 (MK-8628), a novel BET inhibitor, displays in vitro and in vivo antitumor effects alone and in combination with conventional therapies in glioblastoma models // Int J Cancer. 2016. Vol. 139, № 9. P. 2047–2055.
57. Zhou M. et al. Genomic analysis of drug resistant pancreatic cancer cell line by combining long non-coding RNA and mRNA expression profling // Int J Clin Exp Pathol. 2015. Vol. 8, № 1. P. 38–52.
58. Skrajna A. et al. Comprehensive nucleosome interactome screen establishes fundamental principles of nucleosome binding // Nucleic Acids Research. 2020. Vol. 48, № 17. P. 9415–9432.
59. Lagadec F., Parissi V., Lesbats P. Targeting the Nucleosome Acidic Patch by Viral Proteins: Two Birds with One Stone? // mBio. American Society for Microbiology, 2022. Vol. 13, № 2. P. e01733-21.
60. Kurumizaka H., Kujirai T., Takizawa Y. Contributions of Histone Variants in Nucleosome Structure and Function // Journal of Molecular Biology. 2021. Vol. 433, № 6. P. 166678.
61. Roulland Y. et al. The Flexible Ends of CENP-A Nucleosome Are Required for Mitotic Fidelity // Molecular Cell. Elsevier Inc., 2016. Vol. 63, № 4. P. 674–685.
62. Stumme-Diers M.P. et al. Nanoscale dynamics of centromere nucleosomes and the critical roles of CENP-A // Nucleic Acids Research. Oxford University Press, 2018. Vol. 46, № 1. P. 94–103.
63. Kujirai T. et al. Identification of the amino acid residues responsible for stable nucleosome formation by histone H3.Y // Nucleus. Taylor & Francis, 2017. Vol. 8, № 3. P. 1–10.
64. Leung A. et al. Unique yeast histone sequences influence octamer and nucleosome stability // FEBS Letters. 2016. Vol. 590. P. 2629–2638.
65. Widom J. Role of DNA sequence in nucleosome stability and dynamics // Quarterly reviews of biophysics. 2002/02/13 ed. 2001. Vol. 34, № 3. P. 269–324.
66. Polach K.J., Widom J. Mechanism of Protein Access to Specific DNA Sequences in Chromatin: A Dynamic Equilibrium Model for Gene Regulation // Journal of Molecular Biology. 1995. Vol. 254, № 2. P. 130–149.
67. Ballaré C. et al. Nucleosome-driven transcription factor binding and gene regulation // Mol Cell. 2013. Vol. 49, № 1. P. 67–79.
68. Druliner B.R. et al. Comprehensive nucleosome mapping of the human genome in cancer progression // Oncotarget. 2015. Vol. 7, № 12. P. 13429–13445.

69. Mirny L.A. Nucleosome-mediated cooperativity between transcription factors // Proceedings of the National Academy of Sciences of the United States of America. 2010/12/15 ed. 2010. Vol. 107, № 52. P. 22534–22539.
70. Wang J. et al. Disrupted cooperation between transcription factors across diverse cancer types // BMC Genomics. 2016. Vol. 17.
71. Yazdi P.G. et al. Increasing Nucleosome Occupancy Is Correlated with an Increasing Mutation Rate so Long as DNA Repair Machinery Is Intact // PLOS ONE / ed. Imhof A. 2015. Vol. 10, № 8. P. e0136574.
72. Rodriguez Y., Smerdon M.J. The Structural Location of DNA Lesions in Nucleosome Core Particles Determines Accessibility by Base Excision Repair Enzymes // J. Biol. Chem. 2013. Vol. 288, № 19. P. 13863–13875.
73. Shaytan A.K. et al. Hydroxyl-radical footprinting combined with molecular modeling identifies unique features of DNA conformation and nucleosome positioning // Nucleic Acids Research. Oxford University Press, 2017. Vol. 45, № 16. P. 9229–9243.
74. Mao P. et al. Chromosomal landscape of UV damage formation and repair at single-nucleotide resolution // Proc Natl Acad Sci U S A. 2016. Vol. 113, № 32. P. 9057–9062.
75. Pich O. et al. Somatic and Germline Mutation Periodicity Follow the Orientation of the DNA Minor Groove around Nucleosomes // Cell. 2018. Vol. 175, № 4. P. 1074-1087.e18.
76. Hauer M.H., Gasser S.M. Chromatin and nucleosome dynamics in DNA damage and repair // Genes Dev. 2017. Vol. 31, № 22. P. 2204–2221.
77. Prendergast J.G.D., Semple C.A.M. Widespread signatures of recent selection linked to nucleosome positioning in the human lineage // Genome Res. 2011. Vol. 21, № 11. P. 1777–1787.
78. Gräslund A., Jernström B. DNA–carcinogen interaction: covalent DNA-adducts of benzo(a)pyrene 7, 8-dihydrodiol 9, 10-epoxides studied by biochemical and biophysical techniques // Quart. Rev. Biophys. 1989. Vol. 22, № 1. P. 1–37.
79. Zhou C., Greenberg M.M. DNA damage by histone radicals in nucleosome core particles // Journal of the American Chemical Society. American Chemical Society, 2014. Vol. 136, № 18. P. 6562–6565.
80. Gonzalez-Perez A., Sabarinathan R., Lopez-Bigas N. Local Determinants of the Mutational Landscape of the Human Genome // Cell. 2019. Vol. 177, № 1. P. 101–114.
81. Struhl K., Segal E. Determinants of nucleosome positioning // Nat. Struct. Mol. Biol. 2013. Vol. 20, № 3. P. 267–273.
82. van Holde K.E. The Proteins of Chromatin. II. Nonhistone Chromosomal Proteins // Chromatin / ed. van Holde K.E. New York, NY: Springer, 1989. P. 181–218.
83. van Holde K.E. The First Hundred Years // Chromatin / ed. van Holde K.E. New York, NY: Springer, 1989. P. 1–15.
84. Разин С.В., Быстрицкий А.А. Хроматин: упакованный геном. 4 (электронное). Москва: БИНОМ. Лаборатория знаний, 2020. 191 p.
85. Marakulina D. et al. EpiFactors 2022: expansion and enhancement of a curated database of human epigenetic factors and complexes // Nucleic Acids Research. 2023. Vol. 51, № D1. P. D564–D570.
86. Lu J. et al. FACER: comprehensive molecular and functional characterization of epigenetic chromatin regulators // Nucleic Acids Res. 2018. Vol. 46, № 19. P. 10019–10033.
87. Zhang Y. et al. CRdb: a comprehensive resource for deciphering chromatin regulators in human // Nucleic Acids Research. 2023. Vol. 51, № D1. P. D88–D100.
88. Zheng R. et al. Cistrome Data Browser: expanded datasets and new tools for gene regulatory analysis // Nucleic Acids Research. 2019. Vol. 47, № D1. P. D729–D735.
89. Ru B. et al. CR2Cancer: a database for chromatin regulators in human cancer // Nucleic Acids Research. 2018. Vol. 46, № D1. P. D918–D924.
90. Shah S.G. et al. HISTome2: a database of histone proteins, modifiers for multiple organisms and epidrugs // Epigenetics & Chromatin. 2020. Vol. 13, № 1. P. 31.
91. Han H. et al. TRRUST v2: an expanded reference database of human and mouse transcriptional

regulatory interactions // Nucleic Acids Res. 2018. Vol. 46, № D1. P. D380–D386.
92. Kulakovskiy I.V. et al. HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis // Nucleic Acids Research. 2018. Vol. 46, № D1. P. D252–D259.
93. Mani U. et al. SWI/SNF Infobase-An exclusive information portal for SWI/SNF remodeling complex subunits // PLoS ONE. Public Library of Science, 2017. Vol. 12, № 9. P. e0184445.
94. Xu Y. et al. WERAM: a database of writers, erasers and readers of histone acetylation and methylation in eukaryotes // Nucleic Acids Res. 2017. Vol. 45, № Database issue. P. D264–D270.
95. Gene Ontology Consortium. Gene Ontology Consortium: going forward // Nucleic Acids Res. 2015. Vol. 43, № Database issue. P. D1049-1056.
96. Tomczak A. et al. Interpretation of biological experiments changes with evolution of the Gene Ontology and its annotations: 1 // Sci Rep. Nature Publishing Group, 2018. Vol. 8, № 1. P. 5115.
97. Dellaire G., Farrall R., Bickmore W.A. The Nuclear Protein Database (NPD): sub-nuclear localisation and functional annotation of the nuclear proteome // Nucleic Acids Res. 2003. Vol. 31, № 1. P. 328–330.
98. Sprenger J. et al. LOCATE: a mammalian protein subcellular localization database // Nucleic Acids Res. 2008. Vol. 36, № Database issue. P. D230-233.
99. Gendler K., Paulsen T., Napoli C. ChromDB: the chromatin database // Nucleic Acids Res. 2008. Vol. 36, № Database issue. P. D298-302.
100. Rastogi S., Rost B. LocDB: experimental annotations of localization for Homo sapiens and Arabidopsis thaliana // Nucleic Acids Res. 2011. Vol. 39, № Database issue. P. D230-234.
101. The UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2023 // Nucleic Acids Research. 2022. P. gkac1052.
102. Thul P.J. et al. A subcellular map of the human proteome // Science. American Association for the Advancement of Science, 2017. Vol. 356, № 6340. P. eaal3321.
103. Cho N.H. et al. OpenCell: Endogenous tagging for the cartography of human cellular organization // Science. American Association for the Advancement of Science, 2022. Vol. 375, № 6585.
104. Mirsky A.E., Ris H. ISOLATED CHROMOSOMES // Journal of General Physiology. 1947. Vol. 31, № 1. P. 1–6.
105. R.F. Steiner. PHYSICO-CHEMICAL STUDIES ON THE COMPONENTS OF THYMUS CELL NUCLE. 1952.
106. Zubay G., Doty P. The isolation and properties of deoxyribonucleoprotein particles containing single nucleic acid molecules // Journal of Molecular Biology. 1959. Vol. 1, № 1. P. 1-IN1.
107. Shiio Y. et al. Quantitative proteomic analysis of chromatin-associated factors // J. Am. Soc. Mass Spectrom. American Society for Mass Spectrometry. Published by the American Chemical Society. All rights reserved., 2003. Vol. 14, № 7. P. 696–703.
108. Kustatscher G. et al. Proteomics of a fuzzy organelle: interphase chromatin // EMBO J. 2014. Vol. 33, № 6. P. 648–664.
109. Torrente M.P. et al. Proteomic Interrogation of Human Chromatin // PLOS ONE. Public Library of Science, 2011. Vol. 6, № 9. P. e24747.
110. Dutta B. et al. Profiling of the Chromatin-associated Proteome Identifies HP1BP3 as a Novel Regulator of Cell Cycle Progression // Mol Cell Proteomics. 2014. Vol. 13, № 9. P. 2183–2197.
111. Alajem A. et al. Differential Association of Chromatin Proteins Identifies BAF60a/SMARCD1 as a Regulator of Embryonic Stem Cell Differentiation // Cell Reports. 2015. Vol. 10, № 12. P. 2019–2031.
112. Federation A.J. et al. Highly Parallel Quantification and Compartment Localization of Transcription Factors and Nuclear Proteins // Cell Reports. 2020. Vol. 30, № 8. P. 2463-2471.e5.
113. Kustatscher G. et al. Chromatin enrichment for proteomics: 9 // Nat Protoc. Nature Publishing Group, 2014. Vol. 9, № 9. P. 2090–2099.
114. Kito Y. et al. Cell cycle–dependent localization of the proteasome to chromatin: 1 // Sci Rep. Nature Publishing Group, 2020. Vol. 10, № 1. P. 5801.

115. Ohta S. et al. The Protein Composition of Mitotic Chromosomes Determined Using Multiclassifier Combinatorial Proteomics // Cell. 2010. Vol. 142, № 5. P. 810–821.
116. Batugedara G. et al. The chromatin bound proteome of the human malaria parasite // Microb Genom. 2020. Vol. 6, № 2. P. e000327.
117. van Mierlo G., Wester R.A., Marks H. A Mass Spectrometry Survey of Chromatin-Associated Proteins in Pluripotency and Early Lineage Commitment // PROTEOMICS. 2019. Vol. 19, № 14. P. 1900047.
118. Kustatscher G., Grabowski P., Rappsilber J. Multiclassifier combinatorial proteomics of organelle shadows at the example of mitochondria in chromatin data // Proteomics. 2016. Vol. 16, № 3. P. 393–401.
119. Ginno P.A. et al. Cell cycle-resolved chromatin proteomics reveals the extent of mitotic preservation of the genomic regulatory landscape: 1 // Nat Commun. Nature Publishing Group, 2018. Vol. 9, № 1. P. 4048.
120. Shi M. et al. Quantifying the phase separation property of chromatin-associated proteins under physiological conditions using an anti-1,6-hexanediol index // Genome Biology. 2021. Vol. 22, № 1. P. 229.
121. Elgin S.C.R., Bonner J. Partial fractionation and chemical characterization of the major nonhistone chromosomal proteins // Biochemistry. 1972. Vol. 11, № 5. P. 772–781.
122. Peterson J.L., McConkey E.H. Non-histone chromosomal proteins from HeLa cells. A survey by high resolution, two-dimensional electrophoresis. // Journal of Biological Chemistry. 1976. Vol. 251, № 2. P. 548–554.
123. Fleischer-Lambropoulos H., Pollow K. Comparison of nonhistone chromosomal proteins from neuronal and glial chromatin by isoelectric focussing and microdisc electrophoresis // Biochemical and Biophysical Research Communications. 1978. Vol. 80, № 4. P. 773–780.
124. Gundry R.L. et al. Preparation of Proteins and Peptides for Mass Spectrometry Analysis in a Bottom-Up Proteomics Workflow // Curr Protoc Mol Biol. 2009. Vol. CHAPTER. P. Unit10.25.
125. Gygi S.P. et al. Quantitative analysis of complex protein mixtures using isotope-coded affinity tags: 10 // Nat Biotechnol. Nature Publishing Group, 1999. Vol. 17, № 10. P. 994–999.
126. Ong S.-E. et al. Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics // Mol Cell Proteomics. 2002. Vol. 1, № 5. P. 376–386.
127. Malmström J. et al. Proteome-wide cellular protein concentrations of the human pathogen Leptospira interrogans: 7256 // Nature. Nature Publishing Group, 2009. Vol. 460, № 7256. P. 762–765.
128. Schwanhäusser B. et al. Global quantification of mammalian gene expression control: 7347 // Nature. Nature Publishing Group, 2011. Vol. 473, № 7347. P. 337–342.
129. Beck M. et al. The quantitative proteome of a human cell line // Molecular Systems Biology. John Wiley & Sons, Ltd, 2011. Vol. 7, № 1. P. 549.
130. Wiśniewski J.R. et al. Extensive quantitative remodeling of the proteome between normal colon tissue and adenocarcinoma // Molecular Systems Biology. John Wiley & Sons, Ltd, 2012. Vol. 8, № 1. P. 611.
131. Wiśniewski J.R. et al. A “Proteomic Ruler” for Protein Copy Number and Concentration Estimation without Spike-in Standards* // Molecular & Cellular Proteomics. 2014. Vol. 13, № 12. P. 3497–3506.
132. Wang M. et al. Version 4.0 of PaxDb: Protein abundance data, integrated across model organisms, tissues, and cell‐lines // Proteomics. 2015. Vol. 15, № 18. P. 3163–3168.
133. Weiss M. et al. Shotgun proteomics data from multiple organisms reveals remarkable quantitative conservation of the eukaryotic core proteome // Proteomics. 2010. Vol. 10, № 6. P. 1297–1306.
134. Luger K. et al. Crystal structure of the nucleosome core particle at 2.8 A resolution: 6648 // Nature. 1997/09/26 ed. Macmillan Magazines Ltd., 1997. Vol. 389, № 6648. P. 251–260.
135. Armeev G.A., Gribkova A.K., Shaytan A.K. NucleosomeDB - a database of 3D nucleosome structures and their complexes with comparative analysis toolkit. bioRxiv, 2023. P.

2023.04.17.537230.
136. Uhlen M. et al. A proposal for validation of antibodies // Nat Methods. 2016. Vol. 13, № 10. P. 823–827.
137. Oliviero G. et al. Distinct and diverse chromatin proteomes of ageing mouse organs reveal protein signatures that correlate with physiological functions // eLife / ed. Denzel M.S., Kaeberlein M. eLife Sciences Publications, Ltd, 2022. Vol. 11. P. e73524.
138. Ginno P.A. et al. Cell cycle-resolved chromatin proteomics reveals the extent of mitotic preservation of the genomic regulatory landscape: 1 // Nat Commun. Nature Publishing Group, 2018. Vol. 9, № 1. P. 4048.
139. Dutta B. et al. Elucidating the temporal dynamics of chromatin-associated protein release upon DNA digestion by quantitative proteomic approach // Journal of Proteomics. 2012. Vol. 75, № 17. P. 5493–5506.
140. Itzhak D.N. et al. Global, quantitative and dynamic mapping of protein subcellular localization // eLife / ed. Hegde R.S. eLife Sciences Publications, Ltd, 2016. Vol. 5. P. e16950.
141. van Mierlo G., Vermeulen M. Chromatin Proteomics to Study Epigenetics — Challenges and Opportunities // Molecular & Cellular Proteomics. 2021. Vol. 20. P. 100056.
142. Szklarczyk D. et al. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets // Nucleic Acids Research. 2019. Vol. 47, № D1. P. D607–D613.
143. Orchard S. et al. The MIntAct project--IntAct as a common curation platform for 11 molecular interaction databases. // Nucleic Acids Res. 2014. Vol. 42, № Database issue. P. D358-63.
144. Oughtred R. et al. The BioGRID interaction database: 2019 update // Nucleic Acids Res. 2019. Vol. 47, № Database issue. P. D529–D541.
145. Bernhofer M. et al. NLSdb—major update for database of nuclear localization signals and nuclear export signals // Nucleic Acids Research. 2018. Vol. 46, № D1. P. D503–D508.
146. Thumuluri V. et al. DeepLoc 2.0: multi-label subcellular localization prediction using protein language models // Nucleic Acids Research. 2022. Vol. 50, № W1. P. W228–W234.
147. Baker C.M., Grant G.H. Role of aromatic amino acids in protein–nucleic acid recognition // Biopolymers. 2007. Vol. 85, № 5–6. P. 456–470.
148. El-Gebali S. et al. The Pfam protein families database in 2019 // Nucleic Acids Res. 2019. Vol. 47, № D1. P. D427–D432.
149. Fukushima Y. et al. Structural and biochemical analyses of the nucleosome containing Komagataella pastoris histones // The Journal of Biochemistry. 2022. Vol. 172, № 2. P. 79–88.
150. Sato S. et al. Cryo-EM structure of the nucleosome core particle containing *Giardia lamblia* histones // Nucleic Acids Research. 2021. Vol. 49, № 15. P. 8934–8946.
151. Dacher M. et al. Incorporation and influence of Leishmania histone H3 in chromatin // Nucleic Acids Research. 2019. Vol. 47, № 22. P. 11637–11648.
152. Mattiroli F. et al. Structure of Histone-based Chromatin in Archaea // Science. 2017. Vol. 357, № 6351. P. 609–612.
153. Liu Y. et al. Virus-encoded histone doublets are essential and form nucleosome-like structures // Cell. Elsevier, 2021. Vol. 0, № 0.
154. Valencia-Sánchez M.I. et al. The structure of a virus-encoded nucleosome // Nat Struct Mol Biol. 2021. Vol. 28, № 5. P. 413–417.
155. Armeev G.A., Gribkova A.K., Shaytan A.K. Nucleosomes and their complexes in the cryoEM era: Trends and limitations // Front. Mol. Biosci. 2022. Vol. 9. P. 1070489.
156. Bernier M. et al. Linker histone H1 and H3K56 acetylation are antagonistic regulators of nucleosome dynamics // Nat Commun. 2015. Vol. 6, № 1. P. 10152.
157. Brehove M. et al. Histone Core Phosphorylation Regulates DNA Accessibility* // Journal of Biological Chemistry. 2015. Vol. 290, № 37. P. 22612–22621.
158. Kim J., Lee J., Lee T.-H. Lysine Acetylation Facilitates Spontaneous DNA Dynamics in the Nucleosome // J. Phys. Chem. B. 2015. Vol. 119, № 48. P. 15001–15005.
159. Iwasaki W. et al. Comprehensive Structural Analysis of Mutant Nucleosomes Containing Lysine

to Glutamine (KQ) Substitutions in the H3 and H4 Histone-Fold Domains // Biochemistry. American Chemical Society, 2011. Vol. 50, № 36. P. 7822–7832.
160. Liu Y. et al. Cryo-EM structure of SETD2/Set2 methyltransferase bound to a nucleosome containing oncohistone mutations: 1 // Cell Discov. Nature Publishing Group, 2021. Vol. 7, № 1. P. 1–12.
161. Muthurajan U.M. et al. Crystal structures of histone Sin mutant nucleosomes reveal altered protein–DNA interactions // The EMBO Journal. John Wiley & Sons, Ltd, 2004. Vol. 23, № 2. P. 260–271.
162. Chakravarty D. et al. OncoKB: A Precision Oncology Knowledge Base // JCO Precision Oncology. 2017. № 1. P. 1–16.
163. Zdobnov E.M. et al. OrthoDB in 2020: evolutionary and functional annotations of orthologs // Nucleic Acids Research. 2021. Vol. 49, № D1. P. D389–D393.
164. Mendez D. et al. ChEMBL: towards direct deposition of bioassay data // Nucleic Acids Research. 2019. Vol. 47, № D1. P. D930–D940.
165. Carithers L.J. et al. A Novel Approach to High-Quality Postmortem Tissue Procurement: The GTEx Project // Biopreserv Biobank. 2015. Vol. 13, № 5. P. 311–319.
166. Goldman M.J. et al. Visualizing and interpreting cancer genomics data via the Xena platform // Nat Biotechnol. 2020. Vol. 38, № 6. P. 675–678.
167. Suntsova M. et al. Atlas of RNA sequencing profiles for normal human tissues // Sci Data. 2019. Vol. 6. P. 36.
168. Sheng Q. et al. Multi-perspective quality control of Illumina RNA sequencing data analysis // Brief Funct Genomics. 2017. Vol. 16, № 4. P. 194–204.
169. Patro R. et al. Salmon provides fast and bias-aware quantification of transcript expression: 4 // Nat Methods. Nature Publishing Group, 2017. Vol. 14, № 4. P. 417–419.
170. Love M.I. et al. Tximeta: Reference sequence checksums for provenance identification in RNA-seq // PLOS Computational Biology. Public Library of Science, 2020. Vol. 16, № 2. P. e1007664.
171. Durinck S. et al. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt: 8 // Nat Protoc. Nature Publishing Group, 2009. Vol. 4, № 8. P. 1184–1191.
172. Love M.I., Huber W., Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2 // Genome Biology. 2014. Vol. 15, № 12. P. 550.
173. Gao J. et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal // Sci Signal. 2013. Vol. 6, № 269. P. pl1.
174. Beitz E. TEXshade: shading and labeling of multiple sequence alignments using LATEX2 epsilon // Bioinformatics (Oxford, England). 2000/06/08 ed. 2000. Vol. 16, № 2. P. 135–139.
175. Shinagawa T. et al. Histone Variants Enriched in Oocytes Enhance Reprogramming to Induced Pluripotent Stem Cells // Cell Stem Cell. 2014. Vol. 14, № 2. P. 217–227.
176. Buß O., Rudat J., Ochsenreither K. FoldX as Protein Engineering Tool: Better Than Random Based Approaches? // Computational and Structural Biotechnology Journal. 2018. Vol. 16. P. 25–33.
177. Chammas P., Mocavini I., Di Croce L. Engaging chromatin: PRC2 structure meets function: 3 // British Journal of Cancer. Nature Publishing Group, 2020. Vol. 122, № 3. P. 315–328.
178. Godfrey L. et al. DOT1L inhibition reveals a distinct subset of enhancers dependent on H3K79 methylation: 1 // Nature Communications. Nature Publishing Group, 2019. Vol. 10, № 1. P. 2803.
179. Uchiyama S. et al. Proteome Analysis of Human Metaphase Chromosomes* // Journal of Biological Chemistry. 2005. Vol. 280, № 17. P. 16994–17004.
180. Kurisaki A. et al. Chromatin-related proteins in pluripotent mouse embryonic stem cells are downregulated after removal of leukemia inhibitory factor // Biochemical and Biophysical Research Communications. 2005. Vol. 335, № 3. P. 667–675.
181. Barthéléry M. et al. 2-D DIGE identification of differentially expressed heterogeneous nuclear ribonucleoproteins and transcription factors during neural differentiation of human embryonic

stem cells // Prot. Clin. Appl. 2009. Vol. 3, № 4. P. 505–514.
182. Zhang P. et al. An Overview of Chromatin-Regulating Proteins in Cells // Curr Protein Pept Sci. 2016. Vol. 17, № 5. P. 401–410.
183. Burgess R.J., Zhang Z. Histone chaperones in nucleosome assembly and human disease // Nat. Struct. Mol. Biol. 2013. Vol. 20, № 1. P. 14–22.
184. Musselman C.A. et al. Perceiving the epigenetic landscape through histone readers // Nature Structural & Molecular Biology. 2012. Vol. 19, № 12. P. 1218–1227.
185. Liu H. et al. Clipping of arginine-methylated histone tails by JMJD5 and JMJD7 // Proceedings of the National Academy of Sciences. Proceedings of the National Academy of Sciences, 2017. Vol. 114, № 37. P. E7717–E7726.
186. Azad G.K. et al. Modifying Chromatin by Histone Tail Clipping // Journal of Molecular Biology. 2018. Vol. 430, № 18, Part B. P. 3051–3067.
187. Shin Y. et al. MMP-9 drives the melanomagenic transcription program through histone H3 tail proteolysis: 4 // Oncogene. Nature Publishing Group, 2022. Vol. 41, № 4. P. 560–570.
188. Duncan E.M. et al. Cathepsin L Proteolytically Processes Histone H3 During Mouse Embryonic Stem Cell Differentiation // Cell. Elsevier, 2008. Vol. 135, № 2. P. 284–294.
189. Khalkhali-Ellis Z. et al. Cleavage of Histone 3 by Cathepsin D in the Involuting Mammary Gland // PLoS ONE / ed. Burchell J.M. 2014. Vol. 9, № 7. P. e103230.
190. Ali M.A.M. et al. Matrix metalloproteinase-2 mediates ribosomal RNA transcription by cleaving nucleolar histones // The FEBS Journal. 2021. Vol. 288, № 23. P. 6736–6751.
191. Cheung P. et al. Repression of CTSG, ELANE and PRTN3-mediated histone H3 proteolytic cleavage promotes monocyte-to-macrophage differentiation: 6 // Nat Immunol. Nature Publishing Group, 2021. Vol. 22, № 6. P. 711–722.
192. Poonperm R. et al. Chromosome Scaffold is a Double-Stranded Assembly of Scaffold Proteins: 1 // Sci Rep. Nature Publishing Group, 2015. Vol. 5, № 1. P. 11916.
193. Mayran A., Drouin J. Pioneer transcription factors shape the epigenetic landscape // J. Biol. Chem. Elsevier, 2018. Vol. 293, № 36. P. 13795–13804.

Приложение

Приложение А.

Список наборов белков хроматина из экспериментальных источников с указанием метода исследования, метода выделения белков хроматина и количеством белков, отсортированный по дате публикации.

Экспериментально полученные хроматомы				
№	**Год**	**Тип клеток**	**# белков**	**Ссылка**
Первые исследования				
1	1972	liver cells	10-15	[121]
2	1976	HeLa cells	450	[122]
3	1978	Cerebral nuclei from Wistar rats	1200	[123]
Данные с 2000-ого года				
4	2003	human B lymphocytes cell line, P493-6	282	[107]
5	2005	human cell line (BALL-1)	209	[179]
6	2005	mouse embryonic stem cells (D3)	51	[180]
7	2009	hESC and hESC-derived neurospheres	1521	[181]
8	2011	human, HeLa S3	1900	[109]
9	2012	rat (8 weeks old), liver cells	694	[139]
10	2014	293F cells	481	[110]
11	2014	human cell lines HepG2, HeLa, MCF-7	7635	[108]
12	2015	Mouse R1 embryonic stem cells, neural progenitor cells	150	[111]
13	2018	human T98G cell line (from glioblastoma multiforme)	3065	[119]
14	2021	human K562 cell line	3185	[120]
15	2022	tissues of male C57BL/6J mice	863	[137]

Приложение Б.

Набор терминов из разработанной эмпирической классификации белков хроматина, источники наполнения терминов, с делением на Gene Ontology, базы данных и литературу.

Термины классификации	**Термины из GO**	**Базы данных и литература**
Centromere-associated	CC GO:0000775 chromosome, centromeric region	
Chromatin remodelers	GO:0140658 ATP-dependent chromatin remodeler activity	SWI/SNF Infobase, [182]
DNA modification	BP GO:0006304 DNA modification	
DNA recombination	BP GO:0006310 DNA recombination	
DNA repair	BP GO:0006281 DNA repair	
DNA replication	BP GO:0006260 DNA replication	
Euchromatin	CC GO:0000791 euchromatin	
Heterochromatin	CC GO:0000792 heterochromatin	
Histone chaperones	MF GO:0140713 histone chaperone activity	[183]
Histone modification	BP GO:0016570 Histone modification	
Histone PTM erasers	GO:0160009,GO:0016578,GO:0004407,GO:0016575,GO:0016577,GO:0032452	WERAM, Histome2
Histone PTM readers	MF GO:0140566 histone reader activity, MF GO:0035064 methylated histone binding, MF GO:0106153 phosphorylated histone binding, MF GO:0061649 ubiquitin modification-dependent histone binding, MF GO:0070577 lysine-acetylated histone binding	WERAM, [184]
Histone PTM writers	GO:0140068,GO:0140069,GO:0035173,GO:0120295,GO:0004402,GO:0120297,GO:0106078,GO:0042054,GO:0106229,GO:0140789,GO:0000412,GO:0016574,GO:0016573,GO:0016571,GO:0140852,GO:0061922,GO:0120301	WERAM, Histome2
Histone tail cleavage	-	[185,186], (review), [187] (MMP9),[188] (CTSL),[189] (CTSD),[190] (MMP2),[191] (CTSG, ELANE, PRTN3)
Histones	MF GO:0030527 structural constituent of chromatin	HistoneDB 2.0
HMG	-	
HMG_A/B/N	-	Genes
Hormone receptors	MF GO:0004879 nuclear receptor activity	
Nuclear division (meiotic/mitotic)	BP GO:0140013 meiotic nuclear division, BP GO:0140014 mitotic nuclear division	
Other families with HMG-box domain	-	HOCOMOCO, The Human Transcription Factors
Ribosome biogenesis	BP GO:0042254 ribosome biogenesis	
RNA metabolic process	GO:0016070	
RNA modification	GO:0009451	

RNA polymerases	MF GO:0097747 RNA polymerase activity, CC GO:0030880 RNA polymerase complex	
RNA splicing	BP GO:0008380 RNA splicing	
Scaffold	-	[192]
SMC	CC GO:0008278 cohesin complex, CC GO:0000796 condensin complex, CC GO:0030915 Smc5-Smc6 complex	
Telomere-associated	BP GO:0034397 telomere localization, BP GO:0032200 telomere organization, CC GO:0000781 chromosome, telomeric region	
TF	MF GO:0003700 DNA-binding transcription factor activity	TRRUST v2, HOCOMOCO, The Human Transcription Factors
TF pioneer	-	[193]
Transcription associated	BP GO:0006351 DNA-templated transcription	
X-chromosome inactivation	BP GO:0007549 dosage compensation	
Ribosome	GO:0005840	
Nuclear RBP	GO:0035770, GO:1990904	
RNA binding	GO:0003723	
Nucleolus	GO:0005730	
Positive regulation of transcription	GO:0045893 Positive regulation of DNA-templated transcription	
Negative regulation of transcription	GO:0045892 Negative regulation of DNA-templated transcription	