

МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ  
имени М.В. ЛОМОНОСОВА  
БИОЛОГИЧЕСКИЙ ФАКУЛЬТЕТ

*На правах рукописи*

**Грибкова Анна Кирилловна**

**Классификация белков хроматина человека на основе  
функциональных, физико-химических и структурных свойств методами  
биоинформатики и вычислительной биологии**

Специальность 1.5.8. Математическая биология, биоинформатика

**ДИССЕРТАЦИЯ**  
на соискание ученой степени  
кандидата физико-математических наук

Научный руководитель:  
доктор физико-математических наук,  
член-корреспондент РАН  
Шайтан Алексей Константинович

Москва – 2026

## Оглавление

|                                                                                                                                         |           |
|-----------------------------------------------------------------------------------------------------------------------------------------|-----------|
| <b>Оглавление .....</b>                                                                                                                 | <b>2</b>  |
| <b>Введение .....</b>                                                                                                                   | <b>5</b>  |
| Актуальность темы исследования и степень ее разработанности.....                                                                        | 5         |
| Цель и задачи диссертационной работы .....                                                                                              | 6         |
| Научная новизна работы.....                                                                                                             | 7         |
| Теоретическая и практическая значимость работы .....                                                                                    | 8         |
| Методология и методы исследования .....                                                                                                 | 8         |
| Положения, выносимые на защиту.....                                                                                                     | 10        |
| Публикации.....                                                                                                                         | 11        |
| Степень достоверности и апробация результатов .....                                                                                     | 12        |
| Личный вклад автора .....                                                                                                               | 13        |
| <b>Глава 1. Современные представления о структурно-функциональной организации белков хроматина и подходах к их систематизации .....</b> | <b>14</b> |
| 1.1. Хроматин как иерархическая многокомпонентная система .....                                                                         | 14        |
| 1.2. Структурно-функциональные особенности белков хроматина.....                                                                        | 18        |
| 1.2.1. Гистоновые белки хроматина .....                                                                                                 | 18        |
| 1.2.1.1. Номенклатура генов гистонов.....                                                                                               | 19        |
| 1.2.1.2. Посттрансляционные модификации гистонов .....                                                                                  | 20        |
| 1.2.1.3. Онкомутации гистонов .....                                                                                                     | 22        |
| 1.2.2. Негистоновые белки хроматина .....                                                                                               | 23        |
| 1.2.2.1. Суперсемейство HMG.....                                                                                                        | 24        |
| 1.2.2.2. Шапероны гистонов .....                                                                                                        | 24        |
| 1.2.2.3. Комплексы ремоделирования хроматина .....                                                                                      | 26        |
| 1.2.2.4. Белки, считывающие модификации гистонов, и их домены .....                                                                     | 27        |
| 1.3. Внутренне неупорядоченные регионы белков и методы их предсказания .....                                                            | 29        |
| 1.3.1. Физико-химические свойства ВНР.....                                                                                              | 29        |
| 1.3.2. Методы исследования и предсказания ВНР .....                                                                                     | 31        |
| 1.4. Подходы к аннотации доменов в белках и инструменты структурного сопоставления...33                                                 |           |
| 1.5. Подходы к классификации белков хроматина и основные источники информации .....                                                     | 35        |
| 1.5.1. Подходы к аннотации клеточной локализации белков ядра и хроматина .....                                                          | 36        |
| 1.5.2. Подходы к функциональной классификации белков хроматина .....                                                                    | 39        |
| 1.5.3. Протеомные подходы к анализу белков хроматина .....                                                                              | 43        |
| 1.5.3.1. Обогащение фракции белков хроматина .....                                                                                      | 43        |
| 1.5.3.2. Общие принципы протеомных исследований.....                                                                                    | 48        |
| 1.5.3.3. Количественный анализ белков хроматина.....                                                                                    | 49        |
| 1.5.3.4. База данных PaxDb.....                                                                                                         | 51        |
| <b>Глава 2. Материалы и методы исследования.....</b>                                                                                    | <b>52</b> |
| 2.1. Сбор, обработка и анализ данных о репертуаре белков хроматина и ядра клетки человека.....                                          | 52        |
| 2.1.1. Данные о субклеточной локализации белков и методы их анализа.....                                                                | 52        |
| 2.1.2. Данные о функциях белков хроматина .....                                                                                         | 53        |
| 2.1.3. Данные протеомных исследований хроматина и методы их анализа.....                                                                | 54        |

|                                                                                                                                         |            |
|-----------------------------------------------------------------------------------------------------------------------------------------|------------|
| 2.1.4. Другие источники информации .....                                                                                                | 57         |
| 2.2. Разработка функционально-ориентированной классификации белков хроматина человека SimChrom .....                                    | 57         |
| 2.3. Построение эталонных наборов белков по локализации .....                                                                           | 60         |
| 2.4. Подготовка протеомных данных для количественного анализа белков хроматина .....                                                    | 63         |
| 2.4.1. Данные из базы данных PaxDb .....                                                                                                | 63         |
| 2.4.2. Данные о представленности белков из протеомных исследований .....                                                                | 64         |
| 2.5. Методы анализа физико-химических свойств белков .....                                                                              | 64         |
| 2.6. Методы структурной характеристики и анализа доменного состава белков .....                                                         | 66         |
| <b>Глава 3. Критический анализ разнородных источников информации о репертуаре белков хроматина и ядра клетки человека .....</b>         | <b>69</b>  |
| 3.1. Особенности аннотирования хроматин-ассоциированных белков в базе знаний Gene Ontology .....                                        | 69         |
| 3.2. Сравнительный анализ данных о клеточной локализации белков .....                                                                   | 71         |
| 3.2.1. Количество и доля ядерных белков в протеоме человека .....                                                                       | 72         |
| 3.2.2. Различия в иерархиях терминов субъядерной локализации белков .....                                                               | 72         |
| 3.2.3. Сравнительный анализ аннотаций субъядерной локализации белков в UniProt, HPA и OpenCell .....                                    | 74         |
| 3.2.4. Множественная локализация белков ядра .....                                                                                      | 77         |
| 3.3. Сравнительный анализ протеомных данных о составе хроматина .....                                                                   | 81         |
| 3.3.1. Анализ белков, идентифицированных в протеомных экспериментах .....                                                               | 81         |
| 3.3.2. Анализ белков из протеомных исследований хроматина, не имеющих аннотаций о локализации в ядре .....                              | 84         |
| 3.3.3. Анализ не идентифицированных в протеомных экспериментах белков хроматина .....                                                   | 87         |
| 3.4. Сравнение репертуара белков хроматина по данным разнородных источников информации .....                                            | 89         |
| 3.5. Обсуждение результатов .....                                                                                                       | 90         |
| <b>Глава 4. Эмпирическая функционально-ориентированная классификация белков хроматина SimChrom и эталонные наборы белков ядра .....</b> | <b>93</b>  |
| 4.1. Эмпирическая функционально-ориентированная классификация белков хроматина SimChrom .....                                           | 93         |
| 4.2. Эталонные наборы белков ядра и цитоплазмы .....                                                                                    | 99         |
| 4.3. Верификация набора белков хроматина человека SimChrom .....                                                                        | 100        |
| 4.4. База данных SimChrom .....                                                                                                         | 103        |
| 4.5. Обсуждение результатов .....                                                                                                       | 105        |
| <b>Глава 5. Структурно-функциональные свойства белков хроматина человека .....</b>                                                      | <b>107</b> |
| 5.1. Количественная представленность белков хроматина в клетках .....                                                                   | 108        |
| 5.1.1. Особенности распределения белков протеома и хроматома по представленности .....                                                  | 108        |
| 5.1.2. Анализ представленности конститутивных белков хроматина .....                                                                    | 111        |
| 5.1.3. Представленность белков функциональных категорий хроматина .....                                                                 | 112        |
| 5.1.4. Анализ представленности белков по данным МС-исследований .....                                                                   | 122        |
| 5.2. Физико-химические свойства белков хроматина .....                                                                                  | 124        |
| 5.2.1. Длина белков хроматина .....                                                                                                     | 125        |
| 5.2.2. Внутренне неупорядоченные регионы белков хроматина .....                                                                         | 127        |

|                                                                                |            |
|--------------------------------------------------------------------------------|------------|
| 5.2.3. Заряд белков хроматина.....                                             | 129        |
| 5.2.4. Особенности аминокислотного состава белков хроматина .....              | 132        |
| 5.3. Структурная организация белков хроматина .....                            | 140        |
| 5.3.1. Структурная характеристика белков хроматина.....                        | 141        |
| 5.3.2. Структурные домены белков хроматина .....                               | 143        |
| 5.3.3. Разнообразие функциональных доменов в белках хроматина .....            | 147        |
| 5.4. Мультивалентные взаимодействия белков хроматина.....                      | 153        |
| 5.4.1. Определение доменов, вовлеченных в мультивалентные взаимодействия ..... | 153        |
| 5.4.2. Парные сочетания функциональных доменов в белках хроматина .....        | 154        |
| 5.4.3. Сочетания функциональных доменов в белках хроматина .....               | 158        |
| <b>Заключение.....</b>                                                         | <b>163</b> |
| <b>Благодарности.....</b>                                                      | <b>166</b> |
| <b>Список сокращений.....</b>                                                  | <b>167</b> |
| <b>Список литературы.....</b>                                                  | <b>169</b> |
| <b>Приложение А .....</b>                                                      | <b>182</b> |
| <b>Приложение Б .....</b>                                                      | <b>188</b> |

## Введение

### **Актуальность темы исследования и степень ее разработанности**

В ядре эукариотических клеток геномная ДНК организована в виде хроматина — комплекса, состоящего из ДНК, РНК и белков [1,2]. Суммарная длина молекул ДНК в ядре одной клетки человека достигает примерно двух метров при диаметре ядра около десяти микрометров, что требует высокоорганизованной упаковки генетического материала.

Белки хроматина не только участвуют в упаковке ДНК в ядре клетки, но и играют ключевую роль в регуляции её функциональной активности [3,4]. Несколько тысяч белков, ассоциированных с хроматином, обеспечивают реализацию основных молекулярно-биологических процессов, включая транскрипцию, репликацию и репарацию ДНК [5].

Функциональное состояние хроматина связано с физико-химическими свойствами ассоциированных с ним белков, их количеством, наличием в них структурных доменов и неупорядоченных регионов, которые формируют динамическую сеть специфических и неспецифических многокомпонентных взаимодействий между молекулами.

Первые работы по систематизации белков ядра клетки были выполнены К. Е. ван Хольдом в 1980-х годах [1]. Он разделил их на гистоновые, негистоновые хромосомные и нехромосомные, а также продемонстрировал существенную зависимость состава выделяемой фракции белков хроматина от метода пробоподготовки.

В современных исследованиях принадлежность белка к хроматину определяют с помощью различных методов. Среди них — исследования методами масс-спектрометрии очищенных фракций белков хроматина [6], анализ субклеточной локализации белка с использованием флуоресцентного мечения и конфокальной микроскопии [7,8], а также анализ функций белка по данным из специализированных баз данных и литературы [9,10]. Однако получение согласованной информации о репертуаре белков хроматина затруднено по ряду

причин: ограниченные чувствительность и разрешение методов анализа, функциональная и пространственная компартментализация хроматина, зависимость состава белков от фазы клеточного цикла и типа клеток, а также динамическое перемещение белков между ядром и цитоплазмой в процессе жизнедеятельности клетки. Таким образом, накопленные данные о функциях, качественном и количественном разнообразии белков хроматина остаются разрозненными.

Несмотря на то что современные представления о пространственной организации генома и структуре отдельных ДНК-белковых комплексов в последнее время значительно расширились благодаря развитию методов эпигеномики, 3D-геномики, структурной биологии и биофизики [11–13], имеющихся данных о белках недостаточно для формализации описания функционирования и пространственно-временной организации хроматина.

Таким образом, предметом диссертационной работы является разработка системного подхода, интегрирующего гетерогенную информацию о функциональных, физико-химических и структурных свойствах белков хроматина в единую классификацию. Это необходимо для формирования целостного представления о белковом составе хроматина и его роли в организации и регуляции генома человека.

### **Цель и задачи диссертационной работы**

**Целью** диссертационной работы является разработка эмпирической функционально-ориентированной классификации белков хроматина человека на основе интеграции данных об их функциях, субклеточной локализации, физико-химических свойствах и доменной организации.

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

- 1) Провести критический мета-анализ разнородных источников информации о функциях, локализации и свойствах белков хроматина и ядра клетки человека.
- 2) Разработать на основе интеграции гетерогенных данных о составе и свойствах белков хроматина человека эмпирическую функционально-

ориентированную классификацию SimChrom, а также эталонные наборы белков ядра.

- 3) Применить разработанную классификацию белков хроматина SimChrom для оценки количественной представленности белков различных функциональных категорий (по данным масс-спектрометрических исследований) и выявления функционально значимых физико-химических свойств этих белков.
- 4) Охарактеризовать организацию белков хроматина человека на уровне структурных и функциональных доменов. Определить функционально значимые сочетания доменов, участвующих в мультивалентных взаимодействиях.

### **Научная новизна работы**

Разработанная эмпирическая функционально-ориентированная классификация белков хроматина человека SimChrom разделяет 3045 белков на 39 категорий. Каждый белок в классификации характеризуется независимыми параметрами: (i) субклеточной локализацией, (ii) функциональной аннотацией, (iii) особенностями строения (физико-химическими свойствами и доменным составом). Предложенная классификация обеспечивает системное сопоставление независимых параметров в рамках единой системы.

Применение данной классификации позволяет устанавливать функционально значимые особенности строения белков хроматина, в частности определять параметры их неупорядоченных регионов и распределение в них заряженных аминокислотных остатков, выявлять характерные сочетания функциональных доменов в белках, обуславливающие мультивалентные взаимодействия в хроматине, а также идентифицировать новые структурные домены, которые отсутствуют в экспериментальных базах данных.

Таким образом, разработанная классификация не только систематизирует информацию о белках хроматина человека, но и позволяет проводить интегративный анализ накопленных гетерогенных данных, а именно на основе

данных о субклеточной локализации, физико-химических свойствах и доменном составе белка уточнять и прогнозировать его функциональную роль в хроматине.

### **Теоретическая и практическая значимость работы**

Теоретическая значимость работы заключается в систематизации и интеграции разнородных данных о качественном и количественном составе белков хроматина человека, накопленных в научной литературе. Полученные результаты расширяют представление о белковом составе хроматина человека, а также углубляют понимание молекулярных механизмов функционирования и регуляции генома.

Практическая значимость работы определяется следующим:

- 1) Интеграция гетерогенных данных о белках хроматина в единую функционально-ориентированную классификацию, доступную в сети Интернет (<https://simchrom.intbio.org/>), позволяет проводить более детальную функциональную аннотацию белков хроматина человека на основе сопоставления различных аспектов их свойств, а также уточнять или прогнозировать функции отдельных белков.
- 2) Разработанные эталонные наборы белков ядра и цитоплазмы клетки с различными критериями достоверности и согласованности данных применимы в качестве обучающих наборов в задачах предсказания субклеточной локализации белков методами машинного обучения.

### **Методология и методы исследования**

В ходе исследования разработана оригинальная методология комплексного анализа структурно-функциональных свойств белков ядра и хроматина клетки человека. Методология объединяет три взаимосвязанных этапа: (i) системный мета-анализ существующих способов определения принадлежности белков к хроматину и ядру, (ii) разработку эмпирической функционально-ориентированной классификации белков хроматина и (iii) применение классификации для выявления структурно-функциональных особенностей организации белков хроматина человека (**Рисунок 1**). Итеративный характер предложенной методологии

заключается в следующем: предварительные варианты классификации использовались для оценки качества источников информации, что, в свою очередь, позволило уточнить состав категорий классификации. Такой подход обеспечил внутреннюю согласованность разработанной классификации и эталонных наборов белков.



**Рисунок 1** — Краткая схема методологии исследования, состоящая из трех этапов

Исследование выполнено с использованием современных методов биоинформатики, вычислительной биологии и статистического анализа. В работе особое внимание уделено приведению гетерогенных данных, извлекаемых из специализированных баз данных и литературы, к сопоставимому виду. Для оценки согласованности данных использованы метрики, включая коэффициент сходства Жаккара и гипергеометрический тест.

Выявление статистически значимых различий физико-химических свойств белков хроматина и цитоплазмы выполнено с применением U-критерия Манна–

Уитни с поправкой Бонферрони. Визуализация различий между указанными группами белков по аминокислотному составу осуществлена методом нелинейного снижения размерности Uniform Manifold Approximation and Projection (UMAP). Предсказание трехмерных структур белков хроматина и соотнесение доменов белков с их структурными аналогами выполнены с использованием методов машинного обучения (методы AlphaFold2, FoldSeek). Для анализа сочетания доменов в белках хроматина разработан подход, состоящий из функциональной классификации доменов и расчета условной вероятности нахождения в белке домена одной функциональной группы при наличии домена другой.

### **Положения, выносимые на защиту**

- 1) Разработка эмпирической функционально-ориентированной классификации белков хроматина человека возможна при условии итеративной интеграции гетерогенных данных об их составе и свойствах (локализация, функции, физико-химические свойства, доменная организация). Предложенная классификация позволяет проводить системный анализ структурно-функциональных особенностей организации белков хроматина, выявлять новые закономерности, уточнять известные функции и прогнозировать свойства неохарактеризованных белков.
- 2) Выявленные характерные признаки строения белков хроматина (высокая доля неупорядоченных регионов, специфическое распределение заряженных аминокислотных остатков) лежат в основе их структурных и регуляторных функций, а также способствуют участию этих белков в многокомпонентных взаимодействиях.
- 3) Мультидоменная организация белков хроматина с преобладанием комбинаций доменов, узнающих модифицированные гистоны и ДНК, способствует регуляции эпигенетических процессов.
- 4) На основе разработанной классификации белков хроматина SimChrom с применением методов искусственного интеллекта возможно выявление ранее неохарактеризованных структурных доменов в белках хроматина человека.

## Публикации

Основные публикации Грибковой Анны Кирилловны по теме диссертационной работы в рецензируемых научных изданиях, рекомендованных для защиты в диссертационном совете МГУ по специальности 1.5.8. Математическая биология, биоинформатика (физико-математические науки), в конце приведен вклад автора и объем публикации в печатных листах (п.л.):

- 1) **Gribkova A. K.**, Armeev G. A., Kirpichnikov M. P., Shaytan A. K. (Re)defining the human chromatome: an integrated meta-analysis of localization, function, abundance, physical properties and domain composition of chromatin proteins // Nucleic Acids Research. — 2026. — Vol. 54. — № 2. — gkaf1489. DOI: 10.1093/nar/gkaf1489. Импакт-фактор 13,1 (JIF). 3 из 3,8 п.л.
- 2) Хасанова У. Н., **Грибкова А. К.**, Романова Т. А., Шайтан А. К., Армеев Г. А. Анализ белок-белковых взаимодействий пионерных транскрипционных факторов человека на основе предсказаний структур их комплексов // Биофизика. — 2025. — Т. 70 — № 5. — С. 837–853. EDN: XGEIMO. Импакт-фактор 0,593 (РИНЦ). 0,35 из 0,93 п.л.
- 3) Бигильдеев А. Е., Алексеев В. И., **Грибкова А. К.**, Тимохин Г. С., Комарова Г. А., Шайтан А. К. Роль изменений в структуре и динамике хроматина при COVID-19 // Генетика. — 2024. — Т. 60. — № 1. — С. 16–41. EDN: XIFEVZ. Импакт-фактор 0,673 (РИНЦ). 0,2 из 1,07 п.л.
- 4) Armeev G. A., **Gribkova A. K.**, Shaytan A. K. Nucleosomes and their complexes in the cryoEM era: Trends and limitations // Frontiers in Molecular Biosciences. — 2022. — Vol. 9. — P.1070489. EDN: BOMNID. Импакт-фактор 4,0 (JIF). 0,2 из 0,57 п.л.
- 5) 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 and Chromatin. — 2022. — Vol. 15. — P. 34. EDN: BLBOIK. Импакт-фактор 3,5 (JIF). 0,2 из 1,19 п.л.

- 6) Espiritu D., **Gribkova A. K.**, Shubhangi G., Shaytan A. K., Panchenko A. R. Molecular Mechanisms of Oncogenesis through the Lens of Nucleosomes and Histones // Journal of Physical Chemistry B. — 2021 — Vol. 125. — № 16. — Pp. 3963-3976. EDN: KANSFD. Импакт-фактор 2,9 (JIF). 0,45 из 1,17 п.л.
- 7) 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. — Vol. 56. — Pp. 46–55. EDN: YPOJYJ. Импакт-фактор 7 (JIF). 0,2 из 0,79 п.л.

### **Степень достоверности и апробация результатов**

Основные результаты по теме диссертации опубликованы в 7 статьях, общий объем публикаций составляет 9,52 п.л.

Изложенные результаты работы были представлены в виде устных и стендовых докладов на всероссийских и международных конференциях: «Инженерия живых систем — достижения и перспективы: к 25-летию кафедры биоинженерии биологического факультета МГУ» (Москва, 2025 г.), «Current Trends and Achievements in Life Science» (Шэньчжэнь, Китай, 2025 г.), «Генетические технологии в биомедицине» (Москва, 2025 г.), Moscow Conference on Computational Molecular Biology (Москва, 2025 г.), Russian international conference on cryo-electron microscopy (Москва, 2025 г.), «Генетическая модификация и анализ генома клеток» (Москва, 2024 г.), EMBL Conference: Chromatin and epigenetics (Гейдельберг, Германия, 2023 г.), International Symposium on Chromatin Architecture: Structure and Function (Япония, онлайн, 2023 г.), «Информационные технологии и системы. Биоинформатика. ИТИС(б)» (Огниково, 2022 г.), «XXIX Международная конференция Математика. Компьютер. Образование» (Дубна, онлайн, 2022 г.), «FEBS Congress» (Краков, Польша, 2019 г.).

### **Личный вклад автора**

Личный вклад в работах соискателя, подготовленных им в соавторстве с другими исследователями, заключается в следующем: в работе (Gribkova et al., 2026) соискателем проведен анализ источников информации о белках хроматина

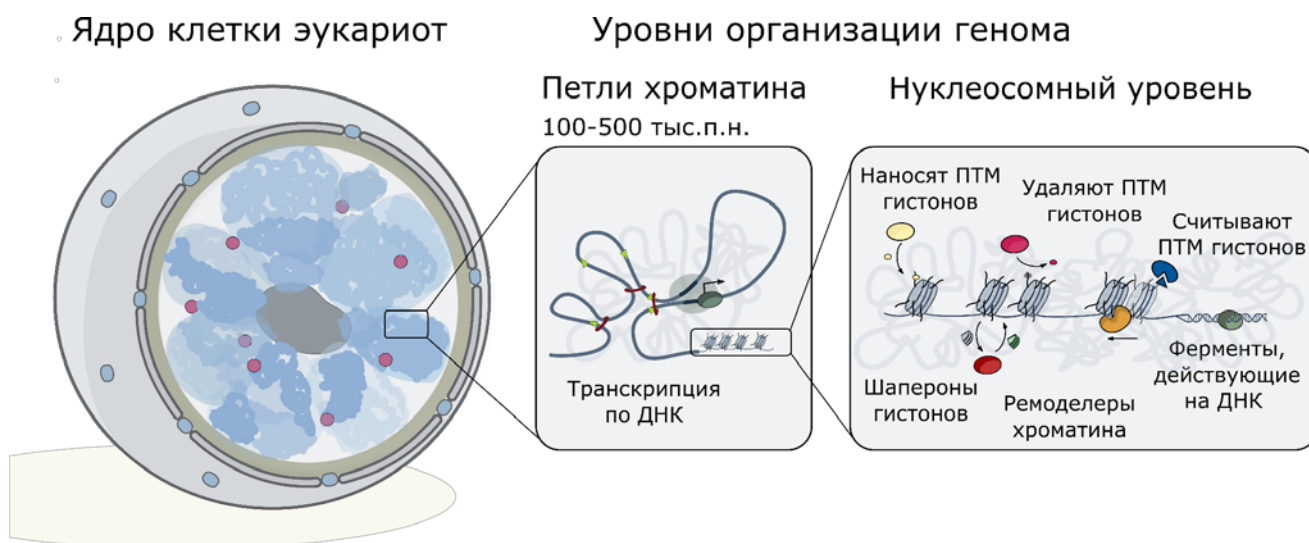
(результаты представлены в Главе 3), разработаны эмпирическая функционально-ориентированная классификация белков хроматина человека SimChrom, эталонные наборы белков хроматина, ядра и цитоплазмы, база данных белков хроматина SimChrom (результаты представлены в Главе 4). Техническая реализация интерактивной веб-версии базы данных SimChrom выполнена Г. А. Армеевым. Соискателем проведена характеристика белков хроматина человека по физико-химическим свойствам и структурной организации (результаты представлены в Главе 5).

В работе (Armeev et al., 2022) соискателем выполнена классификация негистоновых белковых партнеров в комплексах нуклеосом (результаты представлены в Главе 4). В работе (Armeev et al., 2019) соискателем проведен литературный обзор влияния вариантов гистонов на динамику нуклеосом; в работе (Espiritu et al., 2021) проведен анализ онкологических мутаций в гистонах; в работе (Seal et al., 2022) проведено внедрение новой номенклатуры генов гистонов человека в базу данных HistoneDB 2.0; в работе (Хасанова и др., 2025) проведена функциональная классификация белков хроматина из предсказанных комплексов с пионерными транскрипционными факторами; в работе (Бигильдеев и др., 2024) проведен обзор механизмов влияния COVID-19 на профили доступности хроматина в иммунных клетках (результаты представлены в Главе 1).

## Глава 1. Современные представления о структурно-функциональной организации белков хроматина и подходах к их систематизации

### 1.1. Хроматин как иерархическая многокомпонентная система

В ядре эукариотической клетки генетический материал организован в виде хроматина — комплекса геномной ДНК, РНК и белков [1,2] (**Рисунок 2**). С хроматином связаны процессы реализации и хранения наследственной информации — транскрипция, репликация и репарация ДНК. Ключевую роль в организации и регуляции этих процессов играют белки хроматина, совокупность которых обозначают термином хроматом [5,14].



**Рисунок 2** — Строение ядра клетки с деталями организации хроматина на уровне доменов хроматина и на нуклеосомном уровне, показаны функции негистоновых белков хроматина.

В основе взаимодействий макромолекул в хроматине лежат различные физико-химические механизмы, среди которых разделение жидких фаз [15,16], диффузия в условиях макромолекулярного скопления (краудинга) [17] и мультивалентные кооперативные взаимодействия [18]. Важную роль также играют физико-химические и структурные свойства белков хроматина, их

посттрансляционные модификации, доменная архитектура и параметры внутренне неупорядоченных регионов (ВНР, англ. intrinsically disordered regions, IDR).

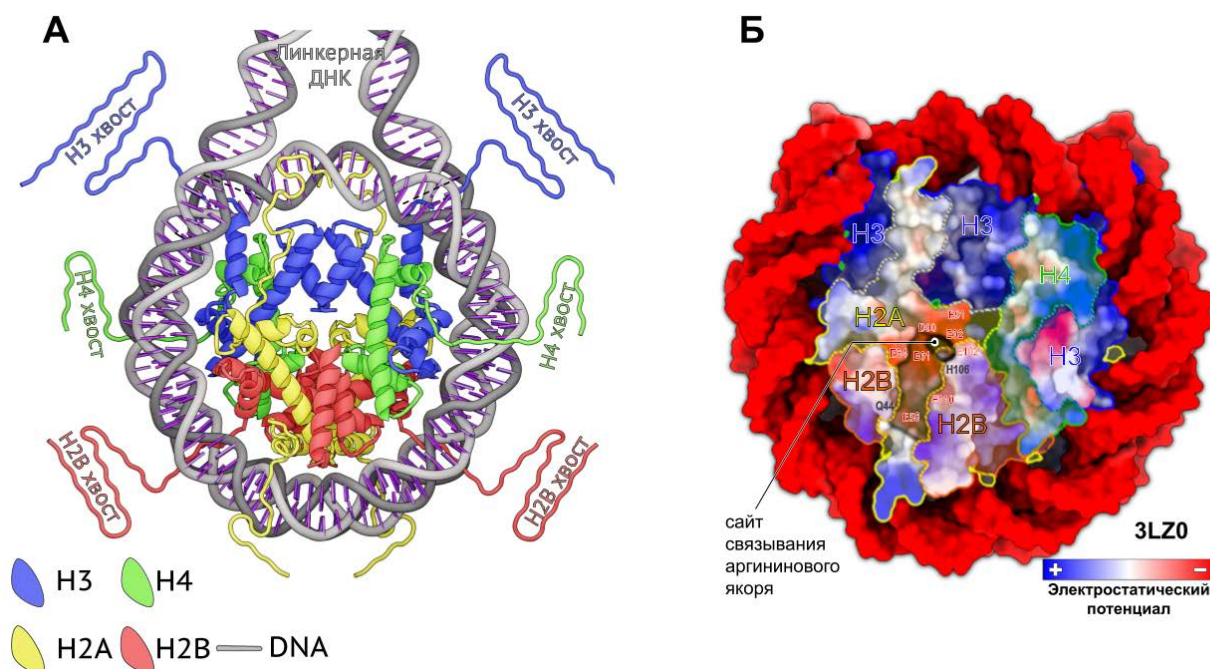
Иерархическая структура хроматина ассоциирована с функциональной активностью генома [19]. На высоком уровне пространственной организации хроматина интерфазные хромосомы занимают в ядре клетки определенные области пространства — хромосомные территории [20] и взаимодействуют с различными компонентами клеточного ядра, включая ламину ядерной оболочки и ядрышко [21,22]. Между хромосомными территориями, а также внутри них, располагаются обедненные хроматином участки — интерхроматиновый компартмент, обеспечивающий транспорт молекул [23]. Транскрипция происходит преимущественно в транскрипционных фабриках на поверхности хроматина в перихроматиновом слое, что способствует транспорту предшественников РНК и новосинтезированных молекул РНК [24].

Белки хроматина вместе с молекулами ДНК и РНК участвуют в организации и регуляции безмембранных ядерных конденсатов интерхроматинового компартмента [25,26], к которым относятся, например, ядрышко (место синтеза и процессинга рибосомных РНК, а также сборки рибосом), тельца Кахаля (структуры, в которых происходит модификация малых ядерных и малых ядрышковых РНК) и тельца гистоновых локусов (места процессинга 3'-концов мРНК гистонов).

Для описания уровней организации хроматина, различающихся степенью конденсации и транскрипционной активностью, используется деление на компартменты А и В, соответствующие эухроматину и гетерохроматину. Развитие методов анализа пространственной организации генома, основанных на фиксации конформации хромосом (англ. chromosome conformation capture) [27], позволило предложить модель организации хроматина в виде топологически ассоциированных доменов (ТАДов), для которых характерна высокая вероятность геномных контактов. Размер ТАДов в геноме человека составляет порядка 100–1000 тысяч пар нуклеотидов.

Сближение пространственно удаленных геномных локусов осуществляется посредством формирования петель хроматина, экструзию которых обеспечивают когезиновые комплексы семейства белков структурного поддержания хромосом (англ. structural maintenance of chromosomes, SMC-комплексы) [28] (**Рисунок 2**). В частности, взаимодействие промотора и энхансера, необходимое для регуляции экспрессии генов, реализуется за счет образования соответствующих петель хроматина [29].

Базовыми структурными единицами хроматина являются нуклеосомные коровые частицы, которые состоят из ДНК (145–147 пар нуклеотидов), обернутой вокруг октамера гистоновых белков [30] (**Рисунок 3А**). Этот октамер состоит из тетрамера гистонов  $(H3-H4)_2$  и двух димеров  $H2A-H2B$ . Нуклеосомная коровая частица вместе с линкерным фрагментом ДНК образует нуклеосому. Линкерный гистон  $H1/H5$  взаимодействует с входящими и выходящими из нуклеосомы участками ДНК, образуя хроматосому.



**Рисунок 3** — Нуклеосомная коровая частица. (А) Структура нуклеосомной коровой частицы, адаптировано из [31]. (Б) Электростатический потенциал поверхности нуклеосомной коровой частицы, адаптировано из [32].

Нуклеосомная коровая частица представляет собой высоко консервативную структуру, лежащую в основе упаковки генома эукариот. Взаимодействие белков хроматина с нуклеосомами способствует структурным и динамическим перестройкам последних, что необходимо для реализации различных генетических процессов [31].

Развитие методов крио-электронной микроскопии привело к значительному прогрессу в изучении функциональной организации хроматина на нуклеосомном уровне. Согласно данным базы данных NucleosomeDB (дата обращения май 2026 г.) в Protein Data Bank (PDB) депонировано порядка 600 структур комплексов нуклеосомы с негистоновыми белками. В большинстве структур белки хроматина взаимодействуют с отрицательно заряженной поверхностью нуклеосомы, называемой кислотный лоскут (англ. acidic patch) (**Рисунок 3Б**). Ключевую роль в связывании белков хроматина с кислотным лоскутом играет остаток аргинина (так называемый аргининовый якорь) [32,33].

С нуклеосомной ДНК способны связываться пионерные транскрипционные факторы, которые облегчают доступ к ДНК другим белкам, в частности, ремоделерам хроматина и белкам, модифицирующим гистоны [34].

Промоторные и энхансерные области активных генов часто представлены участками ДНК, свободными от нуклеосом, что обеспечивает доступ транскрипционной машинерии и других регуляторных белков [35]. Большинство промоторов локализовано в CpG-островках — регионах длиной более 1000 п.о. с высоким содержанием динуклеотидов CpG [36]. Наиболее распространенной модификацией ДНК является метилирование цитозина в составе CpG. Как правило, метилирование CpG-островков коррелирует с репрессией транскрипции [37] и снижает аффинность связывания транскрипционных факторов [38]. В свою очередь, связывание транскрипционных факторов с мотивами ДНК в CpG-островках может препятствовать метилированию этих участков.

Другим важным классом молекул в ядре клетки являются некодирующие РНК (нкРНК), которые участвуют в ремоделировании хроматина и регуляции

экспрессии генов посредством взаимодействий с белками хроматина, геномными локусами и другими РНК [39].

Таким образом, хроматин представляет собой иерархическую, многокомпонентную, пространственно и функционально неоднородную систему, состояние которой определяется сложной сетью взаимодействий между ДНК, РНК и белками.

## **1.2. Структурно-функциональные особенности белков хроматина**

### **1.2.1. Гистоновые белки хроматина**

Кроме компактизации генетического материала, гистоны играют активную роль в регуляции молекулярно-биологических процессов. Это реализуется посредством замены канонических гистонов в нуклеосомах на вариантные, а также посредством посттрансляционных модификаций гистонов, обеспечивающих рекрутирование белков-эффекторов.

В процессе клеточного деления, наряду с репликацией ДНК, синтезируется четыре типа коровых гистонов (H2A, H2B, H3, H4). У многоклеточных животных эти белки кодируются набором зависимых от репликации генов гистонов (также называемых каноническими или кластерными). Транскрипты этих генов не содержат интронов и вместо полиаденилированного хвоста фланкированы структурой шпилька [40]. Транскрипция и процессинг мРНК канонических гистонов локализованы в тельцах гистоновых локусов.

В отличие от канонических гистонов, экспрессия гистоновых вариантов не зависит от репликации. Гены вариантных гистонов не кластеризованы, их мРНК полиаденилирована и содержит интроны. Показано, что замена канонических гистонов на вариантные изменяет структурные и динамические свойства нуклеосом, в частности стабильность нуклеосомы, сайты взаимодействия с другими белками хроматина и сайты посттрансляционных модификаций гистонов [41]. Гистоновые варианты регулируют активность генома и участвуют в

различных процессах, например, транскрипции (H2A.Z, macroH2A, H2A.B, H3.3) и репарации ДНК (H2A.X), формировании центромер (cenH3) [42], а также в регуляции тканеспецифичных процессов, таких как сперматогенез, в котором задействованы короткие гистоновые варианты — H2A.B, H2A.L, H2A.P [43].

#### **1.2.1.1. Номенклатура генов гистонов**

В геномах млекопитающих канонические гены гистонов объединены в четыре кластера. В геноме человека идентифицировано порядка 130 генов гистонов, более 60 из которых расположены в кластере на шестой хромосоме. Среди канонических гистонов одного типа можно выделить изоформы, кодирующие идентичные белки. Для гистона H2A описано 10 изоформ, для H2B — 10, гистоны H3 и H4 представлены двумя изоформами каждая. Указанные изоформы различаются ограниченным числом замен: от двух (в случае гистонов H3) до восьми (в случае H2A). Исключением является белковый продукт гена H4C7, отличающийся от основной изоформы гистона H4 по 19 остаткам. Показано, что даже единичные замены между каноническими изоформами могут влиять на стабильность нуклеосом, взаимодействие с ДНК и регуляцию доступности сайтов модификации гистонов [44].

Гистоны были открыты еще в XIX веке [45], однако систематическая характеристика их типов и свойств была проведена только в 1960-х годах с развитием биохимических методов фракционирования белков [46]. В связи с накоплением данных о многообразии гистонов и открытии новых вариантов, в 2011 году была принята номенклатура гистоновых белков, основанная на филогенетической информации [47]. Это, в свою очередь, потребовало пересмотра номенклатуры генов гистонов, поскольку предыдущая система не была применима ко всем видам позвоночных и не соответствовала номенклатуре белков [47].

В 2022 году комитеты по номенклатуре генов человека и мыши (Human Gene Nomenclature Committee и Mouse Gene Nomenclature Committee соответственно) предложили пересмотренную номенклатуру генов гистонов млекопитающих [48]. Ключевым принципом новой системы является учет консервативности расположения генов в геномах без привязки к отдельным кластерам, а также отказ

от отражения тканеспецифичного характера экспрессии генов в названии. Такой подход обеспечивает унификацию названий генов гистонов у разных видов позвоночных животных.

Информация о последовательностях гистоновых белков, систематизированных по типам и вариантам гистонов, а также инструменты для их анализа доступны в базе данных HistoneDB 2.0 [49].

В контексте биоинформатических исследований принципиально важно отслеживать обновления номенклатуры генов и белков для корректной интеграции данных. Использование гетерогенных идентификаторов может приводить к потере информации и нарушению воспроизводимости результатов.

#### **1.2.1.2. Посттрансляционные модификации гистонов**

В состав коровых гистонов (H2A, H2B, H3, H4) входит глобулярный домен, состоящий из трех  $\alpha$ -спиралей, соединенных короткими петлями L1 и L2. Этот домен, называемый гистоновым фолдом, фланкирован внутренне неупорядоченными фрагментами — хвостами гистонов (N-концы у всех типов гистонов и C-концы у H2A и H2B). Посттрансляционные модификации гистоновых белков играют важную роль в эпигенетической регуляции.

К наиболее распространенным типам модификаций гистонов относятся метилирование, ацетилирование, фосфорилирование, поли(АДФ)-рибозилирование и убиквитинилирование [3,50]. Известно порядка 30 гистоновых меток, число которых продолжает расти. В последние годы были описаны редкие модификации гистонов — серотонилирование и допаминилирование, которые участвуют в регуляции специфических процессов в головном мозге [51,52].

Эффект модификации определяется ее типом и положением модифицированного остатка гистона. Отдельные модификации могут влиять на добавление, удаление или распознавание соседних меток [3]. Посттрансляционные модификации гистонов изменяют структурное и функциональное состояние хроматина посредством нескольких механизмов. Изменение заряда аминокислотных остатков (например, при ацетилировании остатков лизина) влияет на стабильность и подвижность нуклеосом [53], межнуклеосомные контакты [54] и

свойства жидкофазных конденсатов хроматина [55]. Такие эффекты наиболее характерны для модификаций коровой части нуклеосомы. В свою очередь, модификации хвостов гистонов служат платформой для рекрутирования эффекторных белков хроматина, специфически распознающих эти модификации [56].

Развитие методов иммунопреципитации хроматина с последующим секвенированием (ChIP-Seq) позволило определить распределение модифицированных гистонов вдоль генома и связать их с функциональным состоянием хроматина [57]. Так, метка H3K4me3 (триметилирование остатка лизина-4 гистона H3) обогащена в активных промоторах генов эукариот; метки H3K4me1 и H3K27ac (монометилирование остатка лизина-4 гистона H3, ацетилирование остатка лизина-27 гистона H3) служат маркерами активных энхансеров; расположение меток H3K36me3 коррелирует с активно транскрибируемыми областями благодаря рекрутированию метилтрансферазы гистонов с помощью РНК-полимеразы II в процессе элонгации [58]. В районах факультативного гетерохроматина обогащены метки H3K27me3 и моноубиквитинилирование гистона H2A, которые наносят комплексы группы Polycomb — PRC1 и PRC2 (polycomb repressive complex 1 и 2). В свою очередь конститутивный гетерохроматин обогащен метками H3K9me3, которые специфически узнает белок HP1 (heterochromatin protein 1), способствуя расширению области гетерохроматина [59].

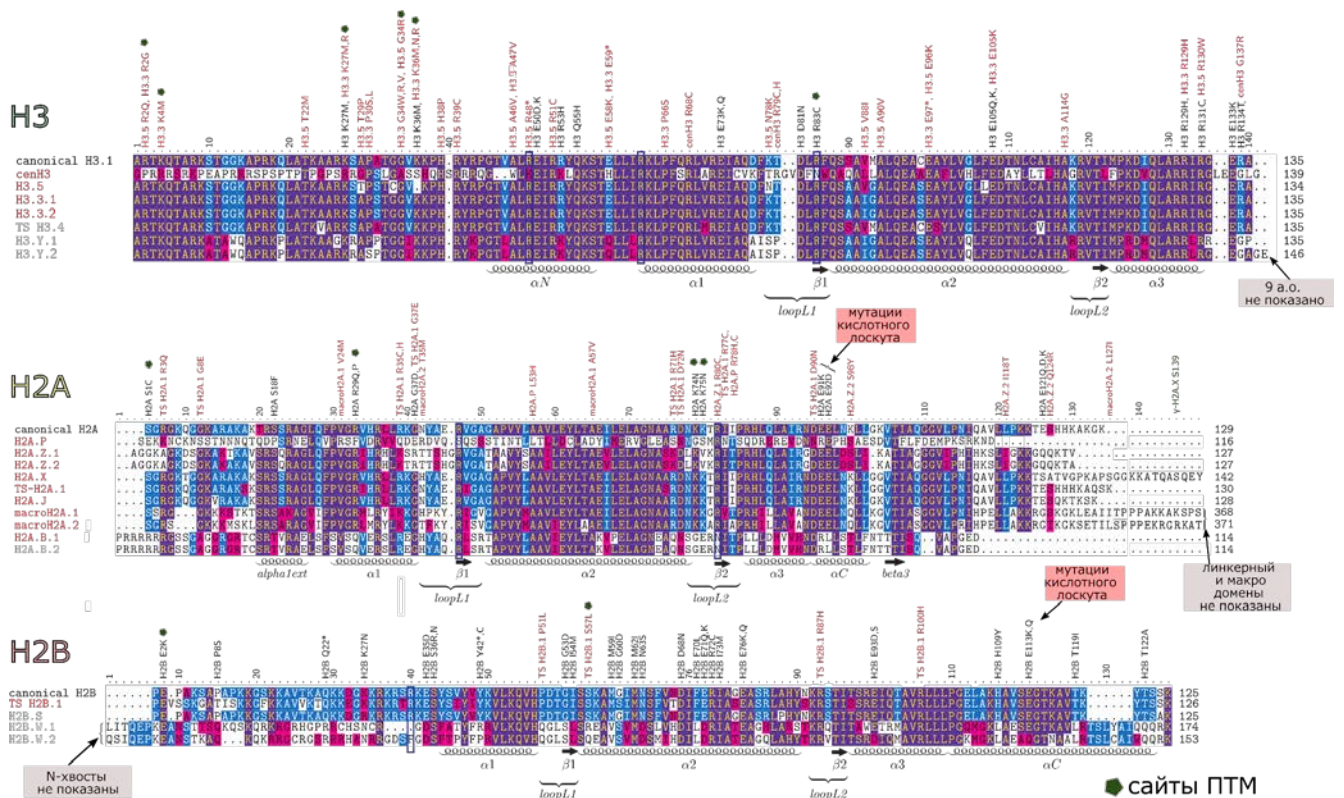
Исследование посттрансляционных модификаций гистонов привело к формированию в 2000-х годах концепции гистонового кода [56,60]. Согласно этой концепции, комбинации модификаций гистонов определяют функциональное состояние соответствующих участков генома [24]. Впоследствии эта концепция была расширена до концепции гистонового контекста, учитывающей взаимосвязь и соотношение различных модификаций. На основе данных об эпигенетических профилях клеток, регионы генома разбивают на непересекающиеся отрезки, так называемые «цвета хроматина», соответствующие различным эпигенетическим/функциональным состояниям [61,62].

Короткие линейные мотивы гистоновых белков, содержащие сайты ПТМ, могут подвергаться мимикрии со стороны белков патогенов, что позволяет последним модулировать эпигенетическое состояние клеток хозяина и влиять на противовирусный ответ [63–65].

### 1.2.1.3. Онкомутации гистонов

Определенные позиции белков гистонов являются сайтами онкогенных аминокислотных замен (онкомутаций). Эти замены влияют на различные аспекты строения и функционирования хроматина, в числе которых изменение профилей посттрансляционных модификаций гистонов, дестабилизация нуклеосомы за счет нарушения взаимодействий между гистонами [71,72], а также нарушение связывания негистоновых белков с нуклеосомой [66,67] (**Рисунок 4**). Известны онкомутации гистонов, которые способствуют релаксации структуры хроматина более высокого порядка, что облегчает доступ транскрипционных факторов к генам, которые в норме репрессированы [71,73].

В образцах пациентов ряда онкологических заболеваний (педиатрические глиомы высокой степени злокачественности, хондробластомы и плоскоклеточные карциномы головы и шеи) были идентифицированы мутации H3K36M и H3K27M (замены остатков лизина на метионин), приводящие к нарушению паттерна метилирования гистонов [68,69]. Показано, что отдельные мутации способствуют изменению профилей метилирования соседних остатков в гистонах и ДНК. В частности, выявлена ассоциация замены H3G34 со снижением уровня метилирования H3K36, а также снижением уровня метилирования ДНК за счет нарушения связывания *de novo* ДНК-метилтрансферазы с модификацией H3K36me3 [70].



**Рисунок 4** — Онкогенные мутации в белках гистонах. Показано множественное выравнивание гистонов H3, H2A, H2B человека, отмечены элементы вторичной структуры гистонов, сайты онкомутаций и модификаций гистонов, адаптировано из [74].

### 1.2.2. Негистоновые белки хроматина

Благодаря развитию методов крио-электронной микроскопии значительно расширились представления о молекулярных механизмах взаимодействий белков с ДНК и функционировании крупных макромолекулярных комплексов хроматина [78]. В данном разделе рассмотрены отдельные репрезентативные классы негистоновых белков хроматина с акцентом на их взаимодействия с гистоновыми белками или нуклеосомой или на физико-химические и структурные особенности, связанные с их функциональной ролью.

#### 1.2.2.1. Суперсемейство HMG

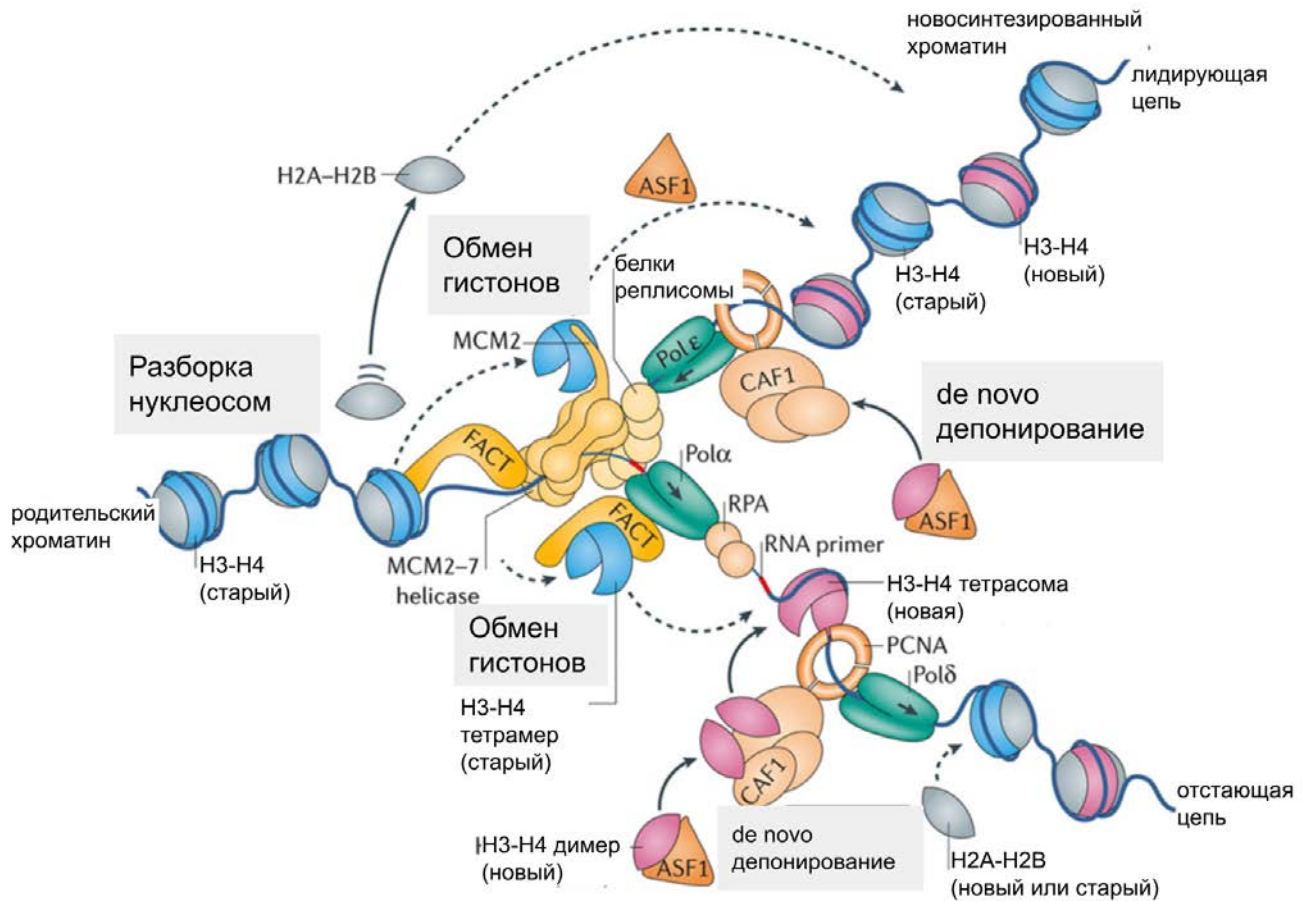
Одними из первых охарактеризованных негистоновых белков хроматина были белки группы HMG (High Mobility Group), которые отличались небольшими

размерами и относительно высокой электрофоретической подвижностью в полиакриламидном геле [79]. В настоящее время суперсемейство HMG подразделяют на три семейства: HMGA (белки содержат домен AT-hook); HMGB (белки содержат ДНК-связывающий мотив HMG-box) и HMGN (белки содержат домен, связывающийся с нуклеосомой). Белки HMG относятся к архитектурным белкам хроматина, которые модулируют его структуру посредством конформационных изменений ДНК и облегчают связывание других регуляторных белков.

#### **1.2.2.2. Шапероны гистонов**

Поддержание целостности хроматина невозможно без динамической сборки и разборки нуклеосом, которые происходят в ходе транскрипции, репликации и репарации ДНК. Ключевую роль в этих процессах играют шапероны гистонов, которые транспортируют гистоны из цитоплазмы в ядро, депонируют их и обеспечивают правильную сборку канонических и вариантных нуклеосом.

Функционирование шаперонов не зависит от энергии АТФ. Шапероны связывают димеры и тетрамеры гистонов, экранируя их положительный заряд и поверхности взаимодействия с ДНК и другими гистонами. Кроме того, они удерживают гистоны в ненативных конформациях [80]. Многие шапероны образуют функциональные сети, обеспечивающие передачу димеров и тетрамеров гистонов между различными комплексами посредством электростатических взаимодействий белковых доменов [81] (**Рисунок 5**).



**Рисунок 5** — Распределение гистонов на хроматине при репликации ДНК. Показано встраивание новосинтезированных гистонов (новых) и «родительских» (старых) гистонов в нуклеосомы на отстающей и лидирующей цепи, которое осуществляют комплексы FACT, Asf1, CAF1. Адаптировано из [82].

Различные шапероны демонстрируют специфичность к отдельным вариантам гистонов. Например, ASF1 (anti-silencing function 1) связывает новосинтезированные димеры H3–H4 через Ig-подобный домен, взаимодействуя с интерфейсом тетрамеризации гистонов ( $\alpha 2$ – $\alpha 3$  спирали гистона H3) [83], и облегчает перенос гистонов к комплексу CAF-1 (chromatin assembly factor-1) для их репликационно-зависимого встраивания в хроматин [84] (**Рисунок 5**). Шапероны HJURP (holliday junction recognition protein) и DAXX (death domain-associated protein 6) взаимодействуют с вариантными димерами гистонов (cenH3 и H3.3 соответственно), распознавая  $\alpha 2$ -спираль H3 через домен HBD (histone-binding domain) [85]. Шапероны семейства NAP (nucleosome assembly protein) осуществляют перенос димера H2A-H2B [86] и способны вытеснять его из

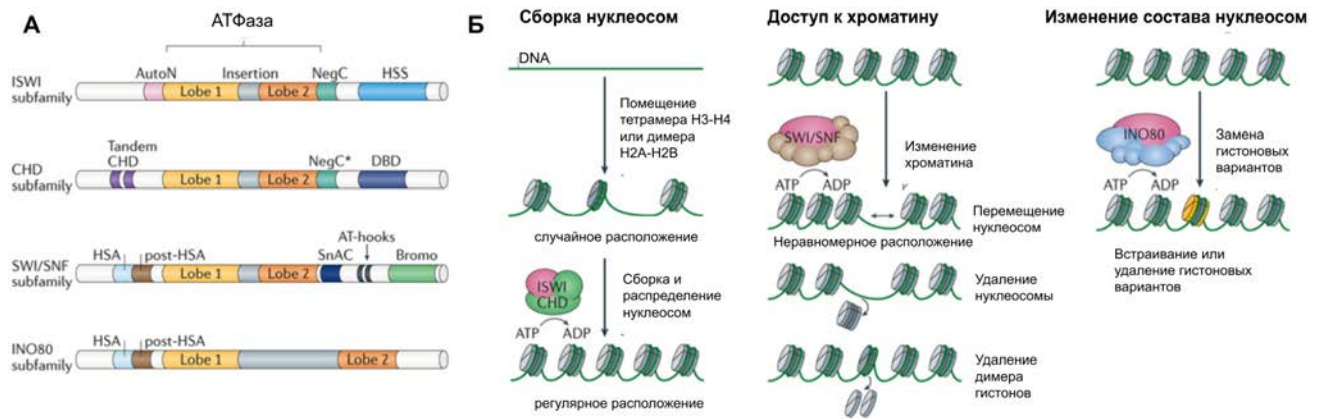
нуклеосом за счет гибкого С-концевого фрагмента [87]. Многофункциональный белок NPM1 из семейства шаперонов NPM (nucleophosmin) связывает несколько гистонов за счет образования пентамерных или декамерных комплексов с помощью N-концевого домена [88].

Комплекс FACT (facilitates chromatin transcription) облегчает прохождение РНК-полимеразы через хроматин путем обратимого разворачивания нуклеосомы [89]. Он состоит из белков SPT16 (suppressor of ty 16) и SSRP1 (structure specific recognition protein 1). Пластичность комплекса способствует формированию мультикомпонентных взаимодействий: N-концевой домен белка SPT16 взаимодействует с димером H3-H4, домен HMG белка SSRP1 — с ДНК, внутренне неупорядоченный домен белка SSRP1 необходим для временного вытеснения димера H2A-H2B [90]. Показано, что FACT может стабилизировать временно раскрытую конформацию нуклеосомы без полной диссоциации гистонов, однако молекулярный механизм этого процесса остается предметом активного изучения.

### 1.2.2.3. Комплексы ремоделирования хроматина

Комплексы ремоделирования хроматина используют энергию АТФ для перемещения нуклеосом вдоль ДНК, что приводит к изменению позиционирования нуклеосом и изменению доступности нуклеосомной ДНК для взаимодействия с другими белками [91] (**Рисунок 6**).

В состав комплексов ремоделирования входит белок с АТФазной активностью, по типу которого выделяют семейства SWI/SNF (SWItch/Sucrose Non-Fermentable), ISWI (Imitation SWItch), CHD (Chromodomain Helicase DNA-binding) и INO80 (**Рисунок 6А**). Комплексы ремоделирования различаются по функциональной специализации: комплексы ISWI и CHD преимущественно участвуют в сборке и позиционировании нуклеосом, тогда как комплексы SWI/SNF обеспечивают увеличение доступности хроматина для регуляторных белков (**Рисунок 6Б**). В состав комплексов ремоделирования могут также входить белки, взаимодействующие с посттрансляционными модификациями гистонов [92].



**Рисунок 6** — Комплексы ремоделирования хроматина. (А) Доменная организация АТФазных субъединиц, определяющих принадлежность комплексов к различным семействам ремоделеров. (Б) Основные функции ремоделеров: сборка нуклеосом, регуляция доступа к хроматину и замена гистоновых вариантов в нуклеосоме. АТФазные субъединицы показана розовым цветом; дополнительные субъединицы обозначены зелёным (ISWI и CHD), коричневым (SWI/SNF) и синим (INO80). Адаптировано из [93].

#### 1.2.2.4. Белки, считывающие модификации гистонов, и их домены

Вместе с концепцией гистонowego кода было установлено существование специализированных белков, обеспечивающих нанесение, распознавание и удаление посттрансляционных модификаций гистонов. Эти процессы лежат в основе регуляции функциональной активности хроматина [94].

Белки, распознающие модификации гистонов (белки-эффекторы), способны изменять свойства хроматина посредством взаимодействия с модифицированными нуклеосомами, связывания нескольких нуклеосом и рекрутирования белков, участвующих в ремоделировании хроматина [95]. Домены белков, связывающиеся с модификациями гистонов, нередко обладают перекрывающейся субстратной специфичностью [96] (**Рисунок 7А**).

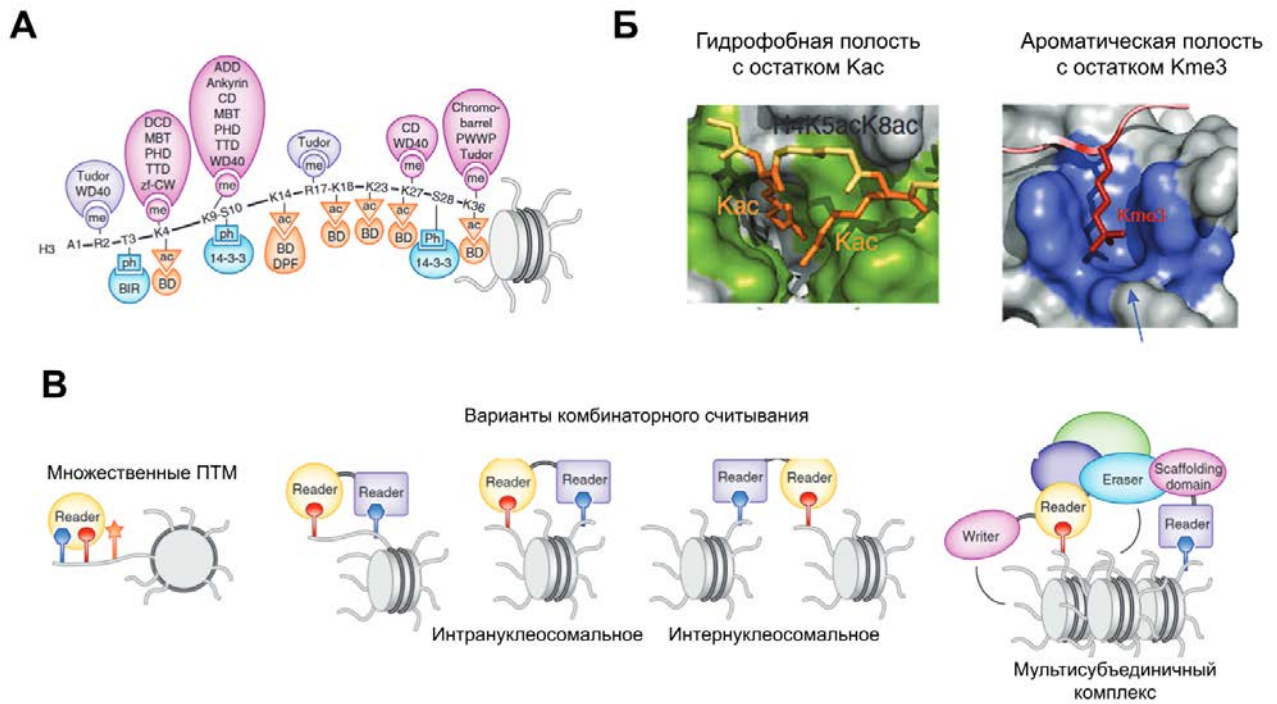
Распознавание метилированных остатков лизина и аргинина гистонов осуществляется различными классами доменов, включая PHD (plant homeodomain), chromodomain, WD40 (tryptophan-aspartate 40 repeat-containing), Tudor, Double Tudor, MBT (malignant brain tumor), ankyrin repeats, BAH (bromo-adjacent homology), CW (cysteine-tryptophan) и PWWP (proline-tryptophan-tryptophan-proline) [96]. Для большинства из них характерно формирование ароматической полости (англ.

aromatic cage), которая стабилизирует метилированный остаток за счет  $\pi$ -катионных взаимодействий и сил Ван-дер-Ваальса [97] (**Рисунок 7Б**). Аминокислотный состав и размер полости обуславливают избирательность в узнавании моно-, ди- и триметилированных производных лизина.

Бромодомены (Bromodomain) являются основными доменами, распознающими ацетилированные остатки лизина посредством гидрофобной полости, содержащей консервативный остаток аспарагина [98] (**Рисунок 7Б**). Для ряда бромодоменов характерна низкая специфичность в отношении аминокислотного контекста, окружающего модифицированный остаток. Они способны связывать несколько различных ацетилированных остатков гистонов, проявляя при этом кооперативный эффект [99].

В дополнение к концепции гистонового кода была предложена концепция мультивалентности. Она подразумевает кооперативное связывание нескольких субстратов агентом, содержащим более одной дискретной поверхности взаимодействия [18]. В хроматине такие взаимодействия реализуются мультидоменными белками и стабильными мультисубъединичными белковыми комплексами, способными одновременно распознавать и связывать различные модификации гистонов, ДНК или несколько нуклеосом (**Рисунок 7В**). Это обеспечивает комбинаторное считывание эпигенетических состояний и значительно увеличивает потенциал регуляторных взаимодействий в хроматине [18].

Таким образом, функциональные свойства белков хроматина определяются сочетанием их структурной организации, физико-химических характеристик, способности к распознаванию посттрансляционных модификаций и участию в мультикомпонентных взаимодействиях. Разнообразие доменной архитектуры, наличие внутренне неупорядоченных регионов и функциональная специализация белков формируют основу структурной и функциональной неоднородности хроматина. Учет этих особенностей необходим для систематического анализа белков хроматина и разработки интегративных подходов к их классификации.



**Рисунок 7** — Узнавание модификаций гистонов белковыми доменами. (А) Примеры доменов, распознающих модификации на N-хвосте гистона H3. (Б) Структуры гидрофобной и ароматической полостей, узнающих ацетилированные и метилированные остатки лизина соответственно. (В) Примеры комбинаторного считывания модификаций гистонов. Показано интрануклеосомальное и интернуклеосомальное узнавание множественных модификаций, а также пример узнавания модификаций мультисубъединичным комплексом. Адаптировано из [96].

### 1.3. Внутренне неупорядоченные регионы белков и методы их предсказания

#### 1.3.1. Физико-химические свойства ВНР

Внутренне неупорядоченные регионы (ВНР) белков характеризуются отсутствием стабильной пространственной структуры и конформационной пластичностью [100] (**Рисунок 8**). Длина ВНР может варьировать в диапазоне от 5-10 а.о. до 1000 а.о. ВНР вовлечены в различные клеточные процессы, в том числе в передачу клеточных сигналов, клеточное деление, транскрипцию и трансляцию [101], а также являются движущей силой жидкофазного разделения [102].

Аминокислотный состав и паттерны расположения остатков в ВНР определяют их способность формировать различные типы взаимодействий с белками и нуклеиновыми кислотами. ВНР способны участвовать как в динамических взаимодействиях между неупорядоченными регионами белков, так и во взаимодействиях с упорядоченными белковыми доменами и нуклеиновыми кислотами [103]. Заряженные и ароматические остатки ВНР обеспечивают формирование электростатических и  $\pi$ - $\pi$ -взаимодействий (**Рисунок 8А**), тогда как алифатические и полярные остатки участвуют в гидрофобных взаимодействиях и

в образовании водородных связей. Остатки пролина, вследствие стерических ограничений, препятствуют формированию стабильных вторичных структур [100].

Свойства аминокислотной последовательности ВНР связаны с функцией белка и его субклеточной локализацией [104] (**Рисунок 8Б**). Выявлено, что ВНР белков ядрышка обогащены повторами остатков лизина, а ВНР ядерных спеклов — остатками аргинина [104]. Для белков, регулирующих транскрипцию, характерны глутамин-богатые ВНР, тогда как в белках, участвующих в организации хроматина, часто встречаются чередующиеся положительно и отрицательно заряженные участки [104]. Показано, что на специфичность связывания транскрипционных факторов могут влиять слабые взаимодействия их ВНР с геномом [105].

Внутренне неупорядоченные регионы лежат в основе жидкофазного разделения и формирования динамических биомолекулярных конденсатов, обеспечивающих пространственную компартментализацию молекулярно-биологических процессов и регуляцию функционального состояния хроматина [106].

### 1.3.2. Методы исследования и предсказания ВНР

Экспериментальные исследования свойств ВНР современными биофизическими методами (спектроскопия, ядерный магнитный резонанс, Ферстеровский резонансный перенос энергии, малое угловое рассеяние) позволяют получить информацию о конформационном ансамбле ВНР (совокупности динамических конформаций), их локальных свойствах, термодинамике и кинетике взаимодействий с другими молекулами [107] (**Рисунок 8В**). Одновременно активно развиваются базы данных с экспериментально подтвержденными и предсказанными аннотациями неупорядоченных регионов в белках (DisProt, MobiDB, ELM и другие) [108–110], а также базы данных белков, разделяющих жидкие фазы с мета-данными экспериментальных условий (LLPSDB [111], PhaSepDB [112], PhaSePro [113]).

За последнее десятилетие разработано большое количество вычислительных инструментов для идентификации и характеристики функционально значимых неупорядоченных регионов белков [114]. Они учитывают физико-химические свойства и паттерны расположения аминокислотных остатков в белках, энергию взаимодействий остатков, эволюционную консервативность, а также контекстно-зависимые факторы внешней среды (**Рисунок 8В**). Наибольшее распространение получили методы машинного обучения, позволяющие по аминокислотной последовательности белков предсказать наличие в них неупорядоченных регионов и функциональных мотивов, а также способность белков разделять жидкие фазы и взаимодействовать с упорядоченными доменами [115,116].

Одним из широко используемых инструментов для предсказания неупорядоченных регионов в белках является IUPred (IUPred2, IUPred3), в котором реализована энергетическая оценка внутримолекулярных взаимодействий на основе статистического потенциала, полученного из анализа глобулярных белков [117]. Развитие методов предсказания структуры белков позволило аннотировать неупорядоченные регионы. Было показано, что анализ структур, предсказанных AlphaFold2, с учетом значений показателя уверенности модели (pLDDT, англ. Predicted Local Distance Difference Test), а также площади поверхности, доступной растворителю, обеспечивает высокую точность такой аннотации [118]. В связи с этим в данной работе для аннотации ВНР белков использовали подход, основанный на анализе структур белков, предсказанных AlphaFold2.

#### **1.4. Подходы к аннотации доменов в белках и инструменты структурного сопоставления**

Домены белков определяют как эволюционно консервативные блоки, обладающие сходством на уровне аминокислотных последовательностей, пространственных структур или выполняемых функций [119].

Наиболее широко используемая коллекция функциональных доменов представлена в базе данных Pfam (Protein Families) [120], которая является частью

консорциума InterPro. Pfam предоставляет информацию о доменных моделях, идентифицируемых с помощью скрытых марковских моделей (англ. Hidden Markov Models, HMM) на основе множественных выравниваний аминокислотных последовательностей белков [120]. Каждая доменная модель представлена в Pfam соответствующим профилем HMM, где для каждой позиции множественного выравнивания определены состояния модели и вероятности переходов между состояниями совпадения, вставки и делеции. Профили HMM используются для поиска в базах данных последовательностей белков с использованием программы HMMER. В Pfam выделяют два уровня достоверности данных: домены категории Pfam-A являются наиболее достоверными, информацию о них курируют вручную; домены категории Pfam-B генерируют в автоматическом режиме. С развитием методов искусственного интеллекта появились аннотации категории Pfam-N, полученные с помощью различных нейросетевых архитектур. Использование Pfam-N привело к аннотации порядка 22 миллионов аминокислотных последовательностей белков, среди которых более 1400 последовательностей белков человека, которые не были аннотированы ранее классическими методами поиска по последовательности [120].

Чувствительность методов аннотации и/или идентификации доменов снижается с уменьшением уровня идентичности аминокислотных последовательностей [121]. Структурные методы анализа позволяют преодолеть это ограничение, так как пространственная структура белков эволюционно более консервативна.

Подход, реализованный в базе данных CATH (Class, Architecture, Topology, Homology), ориентирован на классификацию структурных доменов посредством структурного выравнивания экспериментально полученных структур, депонированных в базе данных PDB [122]. Иерархическая классификация CATH включает уровни описания доменов на основе элементов вторичной структуры (класс), общей пространственной укладки вторичных структур без учета и с учетом связности между ними (архитектура и топология соответственно), предположительного наличия общего эволюционного предка (гомологичное

суперсемейство). Для определения гомологичных суперсемейств в CATH был реализован метод динамического программирования SSAP (англ. Secondary Structure Alignment Program), в котором сравниваются локальные геометрические окружения атомов C $\alpha$ . Для аннотации доменов в CATH используют широкий спектр алгоритмов на основе сравнения структур, последовательностей, а также применяют языковые модели [123].

Прогресс в предсказании структур белков, связанный с разработкой алгоритмов AlphaFold, привел к появлению новых подходов в характеристике структур и обнаружению новых структурных доменов. Аннотации доменов, идентифицированных в предсказанных структурах белков, доступны в базе данных TED (The Encyclopedia of Domains) [124]. В TED идентифицировано более 365 миллионов отдельных структурных доменов в более чем 1 миллионе видов живых организмов. Для предсказания границ структурных доменов и получения консенсусных определений в TED используют алгоритмы Merizo, Chainsaw и UniDoc. Сравнение идентифицированных доменов TED с доменами из CATH показало высокий уровень согласованности границ доменов (77 %) [124]. Таким образом, база данных TED значительно расширяет набор потенциальных структурных доменов белков.

Развитие алгоритмов предсказания структур и накопление данных привели к появлению быстрых высокопроизводительных алгоритмов структурного выравнивания запроса с коллекциями (поиска структуры в коллекции). Примером является алгоритм FoldSeek [125], который позволяет находить удаленных структурных гомологов в базах данных PDB, CATH, а также среди моделей, предсказанных методами AlphaFold и ESMFold. Поиск осуществляется на основе третичной структуры белка-запроса, которая преобразуется в одномерную последовательность в соответствии со «структурным алфавитом». Авторами Foldseek был предложен структурный алфавит “3Di”, описывающий геометрические параметры взаимодействия остатков в пространстве и позволяющий выполнять поиск преобразованной одномерной последовательности среди предварительно преобразованных структур из баз данных, применяя

адаптированный алгоритм MMseqs2 [126]. Для повышения производительности MMseqs2 использует поиск по коротким участкам, что позволяет отфильтровать нерелевантные записи перед финальным выравниванием алгоритмом Смита — Ватермана.

Таким образом, развитие методов машинного обучения способствовало увеличению количества аннотаций доменов, повышению точности их классификации и ускорению сопоставления с экспериментально полученными и предсказанными структурами. В тоже время границы доменов могут различаться в зависимости от используемого подхода.

### **1.5. Подходы к классификации белков хроматина и основные источники информации**

Белки хроматина выполняют ключевые функции в реализации молекулярно-биологических процессов при нормальном функционировании клетки. Кроме того, они вовлечены в модулирование структуры хроматина при патологиях, включая онкологические заболевания [127] и вирусные инфекции [128]. Поэтому крайне важно иметь полное и систематизированное представление о репертуаре этих белков.

Долгое время определение репертуара хроматин-ассоциированных белков носило операционный характер: результаты экспериментов существенно зависели от выбранного метода экспериментальной очистки белков хроматина и последующего способа их идентификации [1].

Строгое терминологическое определение репертуара белков хроматина до сих пор остается предметом дискуссий. В узком смысле к негистоновым белкам хроматина относят преимущественно структурные белки — например, семейство HMG или компоненты SMC-комплексов. Однако такого набора недостаточно даже для описания многообразия процессов, которые происходят на нуклеосомном уровне организации хроматина. Более того, методами микроскопии было показано,

что многие белки, ассоциированные с хроматином, взаимодействуют с ним транзистентно [129].

В настоящее время не существует единого терминологического и методологического определения репертуара белков хроматина. Принадлежность белка к хроматину определяют по его субъядерной локализации и функции или по результатам протеомного анализа обогащенных фракций белков. При этом отсутствует единая система, позволяющая интегрировать и совместно анализировать различные параметры белков хроматина.

### 1.5.1. Подходы к аннотации клеточной локализации белков ядра и хроматина

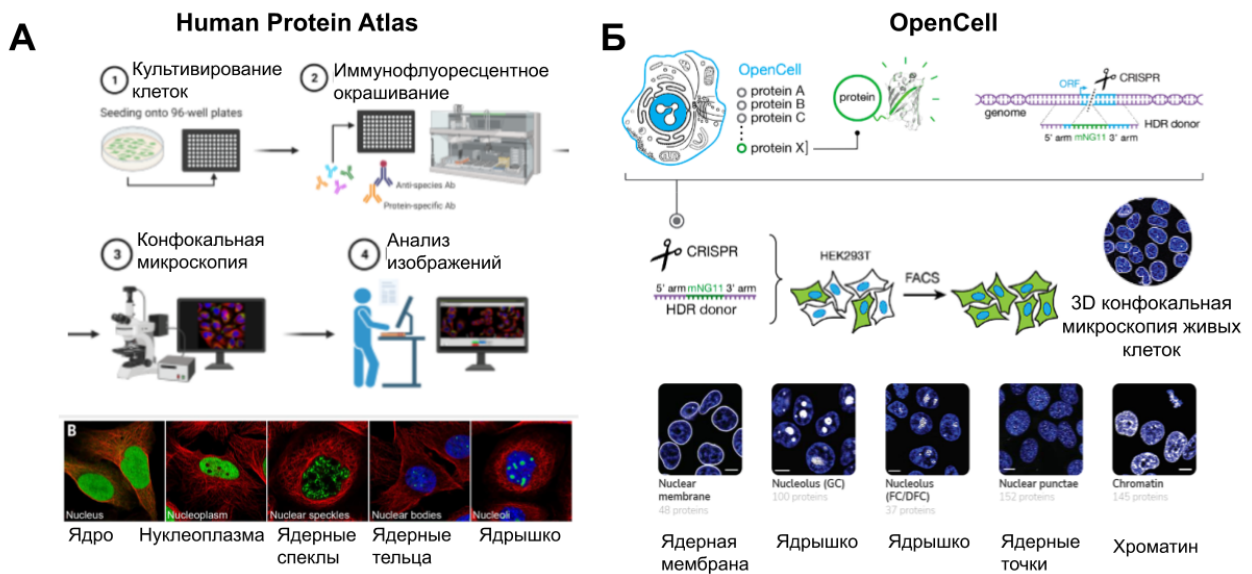
Накопление экспериментальных данных о локализации белков, в частности полученных методами скрининга генных ловушек (англ. gene trap) [130] и иммунофлуоресцентного анализа, привело к созданию специализированных баз данных. В начале 2000-х годов были разработаны первые ресурсы по аннотации локализации белков ядра клетки: The Nuclear Protein Database (NPD) [131] — база данных субъядерной локализации и функциональной аннотации белков ядра; LOCATE [132] — база данных субклеточной локализации белков млекопитающих; ChromDB [133] — база данных белков, ассоциированных с хроматином растений; LocDB [134] — база данных экспериментально подтвержденной локализации белков *Homo sapiens* и *Arabidopsis thaliana*. В 2004 году был основан UniProt — репозиторий, интегрирующий информацию о белках из литературных источников, включая сведения о субклеточной локализации [135].

На настоящий момент наиболее полными и/или достоверными источниками информации о субклеточной локализации белков являются UniProt [135], Human Protein Atlas (HPA) [7] и OpenCell [8]. В UniProt представлен референсный протеом человека, включающий 20225 белков (версия 2022\_2), из которых для 14107 белков доступны аннотации субклеточной локализации. В HPA данные о локализации представлены для 13105 белков, в OpenCell — для 1312. Каждый из ресурсов

использует собственную иерархическую систему (онтологию) для описания субклеточной локализации белков и систему оценки достоверности информации.

Онтология UniProt содержит порядка 550 терминов для описания субклеточных компартментов. Термины локализации присваиваются белкам с указанием источника информации согласно онтологии Evidence Code Ontology (ECO). Наибольшее количество аннотаций получено на основе экспериментальных данных из литературных источников (ECO:0000269), переноса информации о локализации белков на основе сходства последовательностей (ECO:0000250), заключений кураторов (ECO:0000305) и утверждений авторов без дополнительных подтверждающих данных (ECO:0000303). В отдельных случаях термины локализации могут быть присвоены без указания источника информации (например, в случае гистонов).

Более поздние протеомные инициативы, такие как HPA и OpenCell, присваивают белкам аннотации о локализации на основе экспериментальных данных. В HPA локализация белков была определена с использованием иммунофлуоресцентного окрашивания фиксированных клеток и анализа изображений, полученных с помощью конфокальной микроскопии (**Рисунок 9А**). Эксперименты были проведены для ~13 тысяч белков человека на 35 клеточных линиях. Онтология для описания субклеточной локализации белков состоит из 30 терминов, распределенных по 13 органеллам клетки. Для оценки достоверности информации используют четыре уровня, различающиеся количеством связавшихся с белком антител, согласованностью информации с литературными данными и уровнем экспрессии белка.



**Рисунок 9** — Методы определения субклеточной локализации белка. (А, Б) Подходы, реализованные в HPA и OpenCell соответственно, адаптировано из [7,8].

В рамках OpenCell использовали метод определения клеточной локализации белков в живых клетках, основанный на встраивании в ген белка-мишени фрагмента mNG11 флуоресцентного белка split-mNeonGreen2 с помощью системы CRISPR/Cas (**Рисунок 9Б**). Экспрессия в клетке комплементарного фрагмента (mNG1–10) приводит к восстановлению флуоресценции при корректной сборке сплит-системы, что позволяет визуализировать эндогенную локализацию белка без существенного нарушения его функции. Таким методом было охарактеризовано около 1600 белков с преимущественно высоким уровнем экспрессии. Для описания компартментов клетки в OpenCell используют онтологию из 15 терминов, 6 из которых относятся к ядру клетки. OpenCell является одним из немногих ресурсов, в котором интерфазный хроматин выделен в качестве отдельного термина локализации. Достоверность информации о локализации белка оценивается по трем уровням: выраженная локализация, менее выраженная локализация и слабый сигнал, близкий к пределу обнаружения.

### 1.5.2. Подходы к функциональной классификации белков хроматина

Первая систематическая классификация белков ядра была предложена ван Хольде в 1989 году [136]. На основе анализа литературных данных он разделил ядерные белки на три основные класса: гистоновые и негистоновые хромосомные белки, нехромосомные белки (белки ядерной оболочки, ламины и другие). Кроме того, ван Хольде определил следующие массовые соотношения компонентов хроматина: отношение массы ДНК к массе гистонов составляет 1:0,99, тогда как отношение массы ДНК к массе негистоновых белков находится в диапазоне 1:0,3-0,8 [136]. В соответствии с представлениями того периода негистоновые хромосомные белки были разделены на следующие функциональные категории: ферменты, ассоциированные с хроматином; белки, участвующие в регуляции транскрипции; белки-рецепторы и белки-гормоны; белки группы HMG; а также белки «каркаса» хромосом.

В настоящее время ресурсы, предоставляющие информацию о функциональной аннотации белков хроматина, можно разделить на три типа: онтология генов Gene Ontology (GO), базы данных белков, регулирующих эпигенетические процессы (EpiFactors и другие) и более специализированные базы данных или литературные источники, содержащие информацию об определенных классах белков. Эти источники и краткая информация о них представлены в **Таблице 1.**

**Таблица 1** — Источники информации о функциональных классах белков хроматина

| Ресурс                                                                                                                               | Количество белков | Категории белков, количество белков в них                                                                                                                             |
|--------------------------------------------------------------------------------------------------------------------------------------|-------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Gene Ontology                                                                                                                        | ~20000            | хроматин (2019), организация хроматина (825), ремоделирование хроматина (646), транскрипция РНК полимеразой II (445), ядерная ДНК репликация (27)                     |
| Базы данных белков, регулирующих эпигенетические процессы (CR Cistrome, CR2Cancer, CRdb, FACER, EpiFactors)                          | 400-850           | в EpiFactors: гистоны и протамины (65), наносят/считывают/удаляют ПТМ гистонов (145,80,58), модификаторы ДНК и РНК, ремоделеры хроматина (80), шапероны гистонов (23) |
| Специализированные базы данных (HistoneDB 2.0, SWI/SNF Infobase, The human transcription factors, TRRUST, HOCOMOCO, WERAM, HISTome2) | 30-1500           | гистоны (~60), ремоделеры хроматина семейства SWI/SNF (~30), ПТМ машинерия гистонов (~100), транскрипционные факторы (~1500)                                          |

Наиболее полным на сегодняшний день ресурсом по аннотации генов и их белковых продуктов является Gene Ontology [9,137]. Ресурс предоставляет информацию о генах (белках) в соответствии с их молекулярными функциями, участием в биологических процессах и локализацией в компартментах клетки. Онтология терминов Gene Ontology представляет собой ориентированный ациклический граф — сложную взаимосвязанную систему, включающую порядка 40 000 терминов (узлов графа), различающихся по уровням детализации. Термины связаны гетерогенными типами связей (например, «А» является частью «Б», «А» регулирует «Б», «А» встречается в «Б»). Для подавляющего большинства белков референсного протеома человека (97 %) в Gene Ontology доступны аннотации. Однако несмотря на кажущуюся полноту данных, в литературе отмечают следующие недостатки системы [138]:

- 1) ограниченный набор генов (белков), аннотации которых подтверждены экспериментально: по состоянию на январь 2015 доля таких белков составляла 57 % от протеома человека;
- 2) предвзятость аннотаций — хорошо изученные белки аннотированы более детально;
- 3) систематические ошибки, накапливающиеся при переносе информации между видами живых организмов, версиями GO, а также из-за задержек в обновлении аннотаций;
- 4) избыточная генерализация: стандартизированный подход к классификации приводит к потере деталей о более частных и вариабельных группах белков.

Более того, информация о белках хроматина в Gene Ontology представлена в большом количестве терминов, что затрудняет анализ соответствующих аннотаций.

К числу более специфичных для белков хроматина ресурсов относится база данных EpiFactors [10], в которой представлены данные о функциях эпигенетических регуляторов (белков, регулирующих эпигенетические процессы) на основе интеллектуального анализа литературных источников.

В EpiFactors содержится информация о ~100 гистоновых и протаминовых белках, а также о ~800 негистоновых белках. В числе последних — белки, которые изменяют структуру хроматина (АТФ-зависимые ремоделеры хроматина; белки, наносящие, считывающие и удаляющие метки гистонов; шапероны гистонов), а также модификаторы ДНК/РНК, влияющие на экспрессию генов, но не через посттранскрипционную модификацию РНК в рамках ее созревания. В базе данных также представлены функциональные комплексы (например, белки семейства Polycomb) и кофакторы таких комплексов. Кроме того, в EpiFactors представлена информация о длинных некодирующих РНК, вовлеченных в эпигенетическую регуляцию.

Определение и классификация эпигенетических регуляторов из EpiFactors используется в ряде других баз данных, включая CR Cistrome 2013 [139], FACER 2018 [140], CR2Cancer 2018 [141], CRdb 2023 [142]. Однако основное внимание в

этих базах данных уделено другим аспектам, в частности взаимодействиям белков с геномом и их роли в канцерогенезе.

Несмотря на полноту сведений об эпигенетических регуляторах, база данных EpiFactors не охватывает ряд других функциональных категорий белков хроматина, к которым относятся транскрипционные факторы, субъединицы РНК-полимераз, архитектурные белки группы HMG, а также белки, участвующие в репликации и репарации ДНК.

Более специализированные базы данных предоставляют информацию об ограниченном числе функциональных категорий белков хроматина, к ним относятся: гистоны; ремоделлеры хроматина; белки, наносящие, считывающие или удаляющие модификации гистонов; транскрипционные факторы. Так, база данных HistoneDB [49] содержит курируемую информацию о генах и белках гистонов; The Human Transcription Factors [143] и каталог транскрипционных факторов Gene Ontology [144] — информацию о ~1500 транскрипционных факторах; HISTome2 [145] и WERAM [146] — о модификаторах гистонов; SWI/SNF Infobase [147] — о ремоделерах хроматина семейства SWI/SNF.

Для некоторых функциональных категорий белков хроматина информация о репертуаре белков доступна только в литературных источниках или в виде групп генов на веб-ресурсе HUGO Gene Nomenclature Committee (HGNC) и не может быть загружена из ресурсов GO или EpiFactors. К таким категориям относятся, например, белки, отщепляющие хвосты гистонов [148]; шапероны гистонов [82]; пионерные транскрипционные факторы [34]; белки HMG [149]; ядерные РНК-связывающие белки [150,151]; субъединицы SMC-комплексов [152].

Таким образом, несмотря на представленность в литературе источников информации о функциональном разнообразии белков хроматина, до сих пор отсутствует специализированный ресурс, в котором была бы представлена полная и непротиворечивая классификация функциональных категорий белков хроматина человека.

### **1.5.3. Протеомные подходы к анализу белков хроматина**

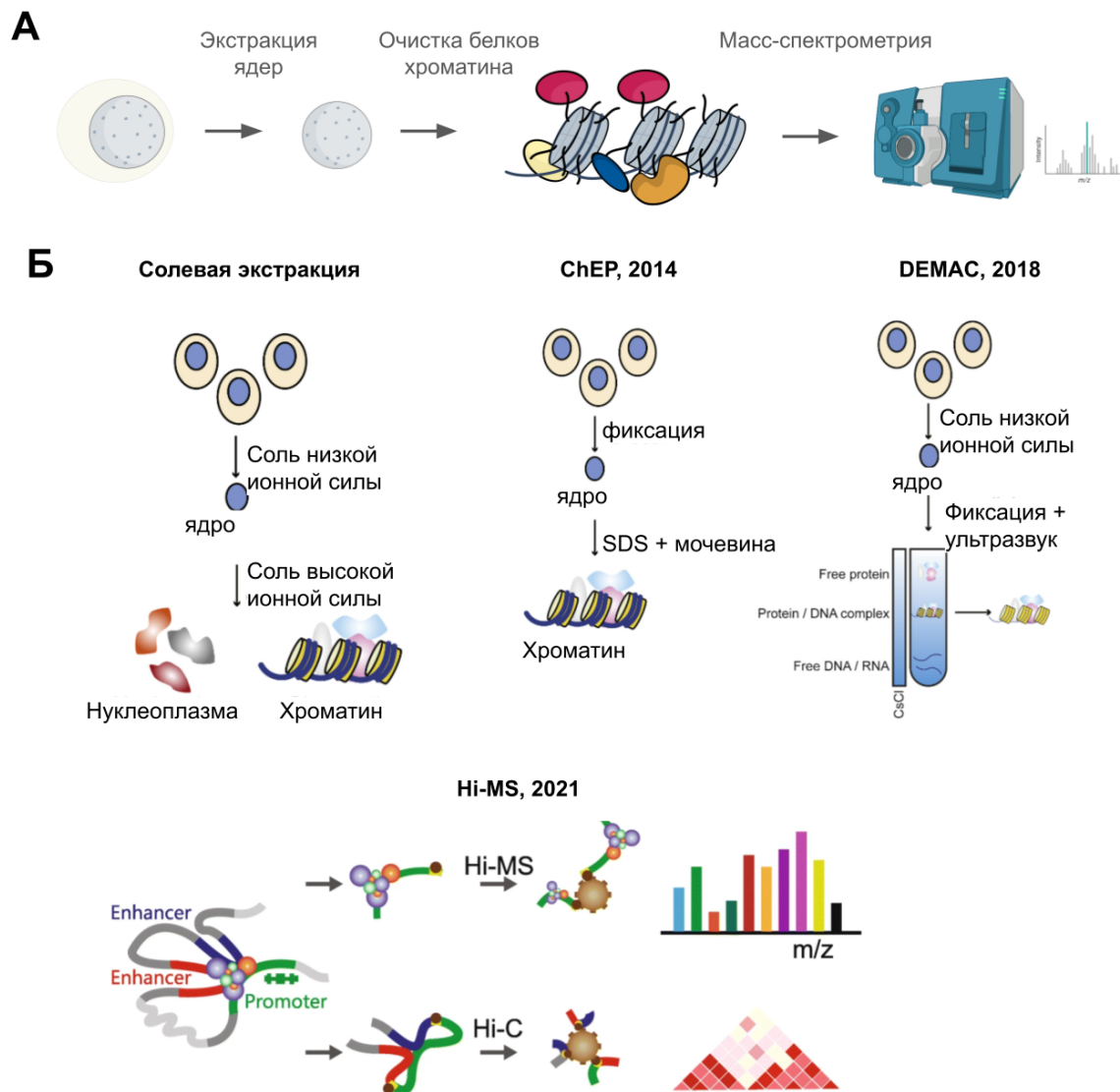
Для определения качественного и количественного состава белков в образце используют протеомные подходы, основанные на методах масс-спектрометрии (МС).

Ключевым этапом протеомного анализа белков хроматина является получение препаратов очищенного хроматина или фракции ядерных белков. Эта задача осложнена физико-химическими свойствами хроматина, который представляет собой заряженный полимер с динамичными взаимодействиями компонентов, нерастворимый в буферах с низкой ионной силой. Кроме того, диффузные размер и плотность хроматина, а также отсутствие мембранной оболочки затрудняют его отделение от других органелл клетки и способствуют контаминации препарата белками из других субклеточных компартментов [6,136].

К другим трудностям исследования хроматина протеомными методами относятся ограниченная идентификация мало представленных белков, зависимость белкового состава хроматина от типа клеток и стадии клеточного цикла, а также сложность интерпретации интенсивностей сигналов для белков с высокой степенью сходства аминокислотных последовательностей, например гистонов [153].

#### **1.5.3.1. Обогащение фракции белков хроматина**

Обобщенная схема выделения хроматина включает в себя следующие этапы: лизис клеток, центрифугирование, осаждение фракции ядер, пермеабилзация ядерной мембраны, промывки в солях и буферах, центрифугирование, осаждение фракции хроматина (**Рисунок 10**).



**Рисунок 10** — Подходы к очистке белков хроматина для протеомных исследований. (А) Общая схема выделения белков хроматина из клеток. (Б) Схематичное изображение протоколов некоторых методов, в частности солевой экстракции, ChEP, DEMAC, Hi-MS. Адаптировано из [6,154].

Протоколы выделения ядер и хроматина развивались с начала XX века в контексте изучения роли ДНК и белков в механизмах передачи наследственной информации. В контексте использования протоколов экстракции «нуклеогистонов» из разрушенных клеток было отмечено, что использование высокой концентрации соли (1 М NaCl) приводит к обратимой диссоциации гистонов и других белков от ДНК [155]. Это наблюдение послужило предпосылкой

для разработки более мягких условий выделения хроматина в 1950-1960-х годах [156,157], которые предполагали уменьшение концентрации соли в растворе до 0,075 М, добавление ингибиторов ферментов и этилендиаминтетрауксусную кислоту (ЭДТА) для хелатирования двухвалентных ионов, активирующих нуклеазы. В то же время развитие методов хроматографического фракционирования и двумерного гель-электрофореза привело к появлению первых оценок качественного состава белков хроматина.

В работах начала 2000-х годов фракцию белков хроматина получали методом общей экстракции: после промывки клеток в стандартном фосфатном буфере и буфере с солью низкой ионной силы (10 мМ KCl, 1,5 мМ MgCl<sub>2</sub>), добавляли Triton X-100 для пермеабилзации клеток, фракцию ядер осаждали центрифугированием и промывали в буферах, содержащих 10 мМ KCl, ЭДТА и ингибиторы протеаз; затем проводили центрифугирование и растворение осадка (хроматина) в буферах с додецилсульфатом натрия [158].

При проведении солевой экстракции хроматина клетки ресуспендируют в гипотоническом лизирующем буфере (например, 10 мМ KCl), полученный после осаждения ядер осадок ресуспендируют в буфере с высокой концентрацией соли (например, 420 мМ KCl) с последующим снижением ионной силы путем многократного разбавления [14] (**Рисунок 10Б**).

Для снижения контаминации цитоплазматическими белками в протоколы выделения хроматина был добавлен этап фиксации белков и ДНК формальдегидом. Примером такого подхода является метод обогащения хроматина для протеомики (англ. Chromatin Enrichment for Proteomics, ChEP) [159]. При применении данного протокола клетки обрабатывают формальдегидом, проводят дифференциальную экстракцию в денатурирующих условиях с додецилсульфатом натрия и мочевиной, обрабатывают РНКазой А и центрифугируют для осаждения хроматина (**Рисунок 10Б**). Авторы протокола ChEP [160] применили классификатор Multiclassifier Combinatorial Proteomics (МССР). Этот метод объединяет результаты количественного протеомного анализа, полученные с помощью стабильного изотопного мечения аминокислотами в культуре клеток (англ. Stable Isotope

Labeling by Amino acids in Cell culture, SILAC) в экспериментах с различными биологическими и биохимическими условиями. Затем с помощью модели машинного обучения на основе случайного леса (англ. random forest) авторы интегрировали результаты классификатора и литературные данные, что позволило для каждого идентифицированного белка рассчитать вероятность его ассоциации с интерфазным хроматином. Методика ChEP широко применяется для выделения белков хроматина из клеточных линий различных видов живых организмов — человека [160,161], курицы [162], возбудителя малярии [163], мыши [164].

Другой разновидностью методики экстракции белков хроматина с фиксацией белков и ДНК является метод обогащения на основе плотности для МС-анализа (англ. density-based enrichment for MS analysis of chromatin, DEMAC) [165]. В DEMAC после осаждения ядер центрифугированием проводят фиксацию ДНК и белков, затем обработку ультразвуком. Далее фракции свободных белков, нуклеиновых кислот и комплексов ДНК–белок разделяют с помощью ультрацентрифугирования в градиенте CsCl в течение 48 часов (**Рисунок 10Б**). С использованием разработанного метода авторы охарактеризовали изменение белкового состава хроматина в ходе клеточного цикла и определили степень сохранения регуляторных белков при митозе.

В 2021 году был предложен метод Hi-MS [154], в котором подготовка белков хроматина для масс-спектрометрии схожа с протоколом Hi-C: хроматин обрабатывают формальдегидом, геномную ДНК расщепляют эндонуклеазой рестрикции NaeIII (по сайтам GGCC), к концам фрагментов ДНК пришивают биотин, а после обработки ультразвуком вытягивают комплексы ДНК с белками с помощью магнитных шариков со стрептавидином (**Рисунок 10Б**). Авторы протокола отмечают, что в отличие от ChEP и DEMAC данная процедура является относительно щадящей и позволяет сохранить наибольшее количество белков, ассоциированных с хроматином. Кроме разработки метода Hi-MS, авторы исследования определили способность белков хроматина разделять жидкие фазы, рассчитав оценку чувствительности белков к 1,6-гександиолу, который нарушает гидрофобные взаимодействия в конденсатах и приводит к их диссоциации.

В настоящее время предложены подходы как для очистки общей массы белков хроматина (англ. total chromatin), так и его отдельных фракций, в частности транскрипционно-активного и неактивного хроматина, новосинтезированного хроматина, а также белков, связанных с конкретными геномными локусами.

Для разделения хроматина на фракции используют обработку образцов различными концентрациями микрококковой нуклеазы (МНКазы) [14]. Другой подход, метод разделения солями (англ. chromatin-enriching salt separation, CHES), основан на последовательном изменении концентрации соли: 150 мМ NaCl — для выделения белков нуклеоплазмы, 250 мМ NaCl — для белков эухроматина, 600 мМ NaCl — для белков гетерохроматина [166].

Для выделения белков новосинтезированного хроматина используют метод захват вновь синтезированного хроматина (англ. nascent chromatin capture, NCC) и выделение белков на вновь синтезированной ДНК (англ. isolation of proteins on nascent DNA, iPOND) [167]. В методе NCC клетки метят биотин-dUTP во время репликации ДНК, после чего через определенные промежутки времени фиксируют формальдегидом для получения новосинтезированного или «зрелого» хроматина. В методе iPOND клетки сначала фиксируют формальдегидом, затем обрабатывают 5-этинил-2'-дезоксинуридином (EdU). Далее с помощью реакции клик-химии присоединяют биотин-азид для выделения целевой фракции белков, связанных с новосинтезированной ДНК.

Описанное разнообразие подходов к выделению белков хроматина может быть систематизировано следующим образом.

- 1) Методы, основанные на грубой дифференциальной экстракции белков детергентом или солями. Из-за заряда компонентов хроматина такие подходы могут приводить к существенной контаминации препарата белками цитоплазмы.
- 2) Методы, включающие этап сшивания молекул формальдегидом перед экстракцией хроматина (ChEP, DEMAC, Hi-MS, NCC, iPOND). Это позволяет уменьшить контаминацию нецелевыми белками.
- 3) Метод, включающий стадию ультрацентрифугирования (DEMAC).

- 4) Метод, основанный на адаптации протокола Hi-C (Hi-MS). В нем обогащение ДНК-белковых комплексов происходит за счет экстракции биотинилированных фрагментов ДНК. Это позволяет сохранять белки хроматина с транзитным взаимодействием с ДНК.
- 5) Методы с использованием MNКаза позволяют получить белки, устойчиво связанные с нуклеосомной ДНК.
- 6) Методы, предполагающие использование дополнительных этапов обработки данных для более точного отнесения белков к хроматину или ядру клетки, например, использование пороговых значений интенсивностей сигнала (Itzhak et al., 2016) или методов машинного обучения (Kustatscher et al., 2014).

### **1.5.3.2. Общие принципы протеомных исследований**

Благодаря развитию методов ионизации молекул масс-спектрометрия стала основным методом определения первичной структуры пептидов начиная с 2000-х годов. В масс-спектрометрии измеряют отношение массы ионизированного фрагмента к его заряду ( $m/z$ ) [168]. Дизайн эксперимента по типу «снизу вверх» обычно включает следующие этапы: очистка целевой фракции белков, денатурация, расщепление белков трипсином до пептидов (трипсин гидролизует пептидные связи после остатков лизина и аргинина), масс-спектрометрический анализ полученной смеси пептидов и сопоставление масс протеолитических пептидов с теоретическими [169]. Метод жидкостной хроматографии в сочетании с tandemной масс-спектрометрией (ЖХ-МС/МС) предполагает разделение смеси пептидов с помощью жидкостной хроматографии, отбор ионов в первом масс-анализаторе, их последующую фрагментацию и анализ во втором масс-анализаторе.

### **1.5.3.3. Количественный анализ белков хроматина**

Первые эксперименты по оценке потенциального количества уникальных белков хроматина без их идентификации были проведены в 1970-1980-х годах. Полученные оценки составили: 10-15 белков (метод ионообменной хроматографии) [170], 450 белков (метод двумерного геля-электрофореза) [171] и

1200 белков, среди которых могли присутствовать различные изоформы и протеоформы одних и тех же белков (метод изоэлектрического фокусирования) [172].

Современные методы количественной протеомики позволяют определять как абсолютное, так и относительное содержание целевых белков. Количество вещества в пробе пропорционально площади пика на масс-хроматограмме (далее МС-сигнал), построенной для иона этого вещества.

Для сравнительного анализа двух наборов белков используют изотопное мечение или внутренние стандарты. В методике изотопного мечения (англ. Isotope Coded Affinity Tag, ICAT) [173] остатки цистеинов белков модифицируют меткой, состоящей из тиол-реагирующей группы для присоединения к цистеину, изотопно-меченого линкера и биотина для последующего выделения модифицированных пептидов. В лизаты двух сравниваемых групп вводят химические метки, различающиеся по числу замен водорода на дейтерий, затем смешивают их и анализируют совместно.

В методике SILAC [174] в питательную среду для выращивания одной из сравниваемых культур клеток добавляют аминокислоты, меченные стабильными изотопами (например,  $^{13}\text{C}_6$ ). Далее обе культуры клеток смешивают и анализируют вместе. Как и в методе ICAT, количественную оценку белка проводят, сравнивая интенсивности сигналов пептидов с легкими и тяжелыми изотопными метками.

Методы МС позволяют также определять количество белка в абсолютных величинах: молярную концентрацию или количество копий белка на клетку. Для этого используют различные подходы: метод абсолютного количественного определения на основе интенсивности (англ. Intensity-Based Absolute Quantification, iBAQ), в котором сумму интенсивностей пиков всех пептидов, соответствующих определенному белку, делят на число теоретически наблюдаемых пептидов [175]; подход с эндогенными белками-стандартами, которые количественно определяют по точно охарактеризованным изотопно-меченым пептидам [176]. При определении абсолютного количества белков относительно известного эталона ограничивающим фактором является точное измерение его концентрации.

В работе [177] было продемонстрировано выполнение следующего соотношения (**формула 1**)

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

В последствии в работе [178] на основании соотношения массы гистонов к массе ДНК как 1:1 [179], было предложено следующее соотношение (**формула 2**)

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

Применение такого соотношения не требует дополнительного подсчета клеток в образце и определения концентрации эталонных белков или пептидов. На основании данных о массе ДНК в диплоидной клетке человека (6,5 пг) были рассчитаны массы белков в четырех клеточных линиях; полученные результаты различались в  $1,24 \pm 0,29$  раза по сравнению с результатами, полученными с подсчетом клеток. Такой метод расчета количества копий белков на клетку с использованием данных об МС-сигнале гистонов получил название «протеомная линейка» (англ. proteomic ruler). В итоге, количество копий белка на клетку можно оценить следующим образом (**формула 3**)

$$\begin{aligned} \frac{\text{МС сигнал белка}}{\text{общий МС сигнал}} \times \frac{N_A}{M} \times \text{масса белков клетки} = \\ = \text{МС сигнал белка} \times \frac{N_A}{M} \times \frac{\text{Масса ДНК}}{\text{МС сигнал гистонов}}, \end{aligned} \quad (3)$$

где  $N_A$  — постоянная Авогадро,  $M$  — молекулярная масса белка.

Также было показано, что использование стандартной библиотеки пептидов для идентификации белков, по сравнению с библиотекой, в которой есть пептиды

гистонов с различными комбинациями ПТМ, ухудшает результат: относительное количество гистона H3 изменяется на 5-10 % [178].

#### **1.5.3.4. База данных PaxDb**

Различие подходов к количественному определению белков, а также накопление экспериментальных данных послужили стимулом для разработки базы данных PaxDb [180] — ресурса, предназначенного для интеграции информации о количественном содержании белков (их представленности, англ. abundance) на уровне организма и отдельных тканей. В версии PaxDb v6.0 собраны данные о представленности белков для 392 видов живых организмов, полученные из 1639 экспериментальных наборов данных, что делает PaxDb одним из наиболее крупных ресурсов в области сравнительной протеомики.

В качестве единицы измерения представленности белков в PaxDb используется метрика ppm (от англ. parts per million — частей на миллион), которая характеризует относительное количество каждого белка в протеоме. Такой подход позволяет сравнивать между собой различные биологические образцы. Для оценки согласованности данных в наборе используют сведения о взаимодействии белков из базы данных STRING: считается, что функционально связанные белки должны иметь схожие уровни экспрессии.

## **Глава 2. Материалы и методы исследования**

### **2.1. Сбор, обработка и анализ данных о репертуаре белков хроматина и ядра клетки человека**

Для проведения биоинформатического анализа необходимо было привести идентификаторы генов и белков к сопоставимому виду. В рамках работы все полученные записи о белках были сопоставлены с набором идентификаторов белков референсного протеома человека (UniProt proteome ID UP000005640, версия 2022\_2), представленного в базе знаний UniProt Knowledgebase (UniProtKB/Swiss-Prot). Набор содержит записи о 20272 генах, соответствующих 20225 уникальным идентификаторам канонических изоформ белков человека.

#### **2.1.1. Данные о субклеточной локализации белков и методы их анализа**

Аннотации клеточной локализации белков были загружены из UniProtKB [135] (версия 2022\_2, раздел субклеточной локализации), Human Protein Atlas [7] (версия 22, раздел субклеточной локализации, дата обращения 9 июня 2022 г.) и OpenCell [8] (<https://opencell.czbiohub.org/>, дата обращения 10 июня 2022 г.). Для повышения качества анализа аннотации субклеточной локализации были отфильтрованы в соответствии с критериями достоверности, специфичными для каждой базы данных. Так, из UniProt были отобраны аннотации с указанием источника информации согласно онтологии ECO, из HPA и OpenCell — с уровнем достоверности выше самого низкого.

Для анализа множественной локализации белков аннотации их локализации были сгруппированы в следующие обобщенные категории согласно определениям, используемым в HPA: ядро, цитоплазма, эндомембранная система, другие белки (хромосомные, секреторные и внеклеточные белки согласно классификации UniProt).

Для оценки обогащения или недостаточной представленности белков в одной базе данных по сравнению с другой был проведен анализ обогащения терминами локализации с использованием гипергеометрического теста (точный критерий Фишера) с порогом значимости  $p$ -значение  $<0,05$ .

Для оценки согласованности аннотаций субъядерной локализации между базами данных UniProt (набор белков А) и НРА (набор белков Б) использовали **формулы 4-11**

$$\text{ИП} = |A \cap B|; \quad (4)$$

$$\text{ЛП} = |B \cap \neg A|; \quad (5)$$

$$\text{ЛО} = |A \cap \neg B|; \quad (6)$$

$$\text{Объединение} = |A \cup B|; \quad (7)$$

$$\text{коэффициент сходства Жаккара} = \frac{\text{ИП}}{\text{Объединение}}; \quad (8)$$

$$\text{точность (Precision)} = \frac{\text{ИП}}{\text{ИП} + \text{ЛП}}; \quad (9)$$

$$\text{полнота (Recall)} = \frac{\text{ИП}}{\text{ИП} + \text{ЛО}}; \quad (10)$$

$$\text{F1 – мера} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}; \quad (11)$$

где ИП — истинно положительные, ЛП — ложноположительные, ИО — истинно отрицательные, ЛО — ложноотрицательные.

### 2.1.2. Данные о функциях белков хроматина

Данные о функциях белков (для сравнительного анализа источников информации и аннотации белков в разрабатываемой классификации SimChrom, **раздел 2.2**) были загружены из базы знаний Gene Ontology через базу данных

QuickGo [42] (набор аннотаций от 6 марта 2025 г.) с обращением по REST API. Для каждого интересующего термина из GO отбирали белки, ассоциированные с этим термином, а также с его дочерними терминами, связанными отношениями "is\_a", "part\_of" и "occurs\_in".

Информация о белках, регулирующих эпигенетические процессы, была получена из базы данных EpiFactors (версия 2.1 от 10 сентября 2024 г.) [10]. ДНК-связывающие транскрипционные факторы были отобраны согласно курируемой коллекции (<https://www.ebi.ac.uk/QuickGO/targetset/dbTF>, дата обращения 1 сентября 2024 г.) [144].

### **2.1.3. Данные протеомных исследований хроматина и методы их анализа**

Для отбора масс-спектрометрических исследований белков хроматина был проведен поиск в базе данных PubMed с использованием ключевых слов: «chromatome», «chromatin proteins», «experimentally obtained chromatin proteins», «nuclear proteins» и «nuclear proteome», а также в статьях, ссылающихся на основные методы обогащения белков хроматина (ChEP и другие) с последующей фильтрацией исследований, в которых доступны результаты идентификации белков хроматина в клеточных линиях человека.

Идентификаторы белков из исследований были сопоставлены с идентификаторами белков референсного протеома человека согласно базе знаний UniProtKB (версия 2022\_2). Идентификаторы сплайс-изоформ белков были приведены к форме записи канонических изоформ. В случае отсутствия совпадений для идентификаторов из экспериментов дополнительно использовали идентификаторы генов (при наличии), а также учитывали различные версии номенклатур генов и белков.

Информация об отобранных для анализа результатах протеомных экспериментов (8 наборов данных) с указанием исследуемой фракции хроматина, клеточной линии и количества идентифицированных белков приведена в **Таблице 2**.

**Таблица 2** — Протеомные исследования хроматина, используемые для сравнительного анализа репертуара белков

| Исследование             | Исследуемая фракция хроматина                                                                     | Тип клеток, тканей                         | Количество белков, идентифицированных авторами                                                                                                    | Количество отобранных белков |
|--------------------------|---------------------------------------------------------------------------------------------------|--------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------|
| Torrente et al., 2011    | Общий хроматин                                                                                    | HeLa S3                                    | 1038 (тотальная экстракция), 1388 (высокосолева), 949 (обработка МНКазой); 751 (частичная обработка МНКазой); 1912 (все идентифицированные белки) | 1501                         |
| Kustatscher et al., 2014 | Общий интерфазный хроматин                                                                        | HeLa, MCF-7, HepG2, HEK293, U2OS, DT40     | 1980 (с вероятностью принадлежности к интерфазному хроматину > 0,5); 7635 (все идентифицированные белки)                                          | 1956                         |
| Alabert et al., 2014     | Новосинтезированный vs «зрелый»                                                                   | HeLa S3                                    | 426 (новосинтезированный); 3995 (все идентифицированные белки)                                                                                    | 3861                         |
| Itzhak et al., 2016      | Общий ядерный протеом                                                                             | HeLa                                       | 1133 (ядерные); 672 (ядерно-цитоплазматические); 8710 (всего)                                                                                     | 1092                         |
| Ginno et al., 2018       | Общий хроматин в разные фазы клеточного цикла G1, S, M)                                           | T98G (мультиформная глиобластома человека) | 3065 (хроматом); 6242 (общий протеом)                                                                                                             | 3051                         |
| Shi et al., 2021         | Хроматин промоторных регионов                                                                     | K562                                       | 3228                                                                                                                                              | 2848                         |
| Ugur et al., 2023        | Общий хроматин                                                                                    | ESCs (H9)                                  | 2487                                                                                                                                              | 1730                         |
| Alvarez et al., 2023     | Новосинтезированный хроматин, хроматин в разные фазы клеточного цикла (G2/M, ранняя и поздняя G1) | HeLa S3                                    | 1454 (идентифицированы во всех временных точки во всех репликах); 5770 (все идентифицированные белки)                                             | 1478 (2894 всего)            |
|                          |                                                                                                   | TIG-3 фибробласты                          | 2351 (идентифицированы в $\geq 4$ из 5 реплик)                                                                                                    | 2397 (2894 всего)            |

Ниже приведены детали отбора белков хроматина по результатам масс-спектрометрических исследований в указанные наборы данных.

Набор (Kustatscher et al., 2014) [160]: отобраны белки с вероятностью принадлежности к интерфазному хроматину выше 0,5.

Набор (Alabert et al., 2014) [167]: исключены записи без идентификаторов белков, генов или значения интенсивности МС-сигнала; для уточнения сопоставления идентификаторов белков использовали идентификаторы генов. Дополнительный отбор белков на основании обогащения белков во фракции новосинтезированного хроматина или вероятности принадлежности к интерфазному хроматину не проводился.

Набор (Ginno et al., 2018) [167]: отобраны белки, идентифицированные по крайней мере двумя уникальными пептидами и имеющие согласованный сигнал во

всех трех репликах хотя бы для одной стадии клеточного цикла (G1, S или M). Несмотря на добавление в набор данных белков митотического хроматина, идентифицированный белковый состав на разных стадиях клеточного цикла был практически идентичен, и исключение белков M-фазы не привело бы к значительному изменению набора данных.

Набор (Shi et al., 2021) [181]: отобраны белки, ассоциированные с хроматином, выделенным в нативных условиях с использованием протокола Hi-MS. Это соответствовало условиям эксперимента с использованием обработки препарата эндонуклеазой рестрикции HaeIII и отсутствием обработки 1,6-гександиолом («условие 2»). Были отобраны только белки с ненулевыми значениями iBAQ во всех трех репликах.

Набор (Torrente et al., 2011) [14]: объединены списки хроматин-ассоциированных белков, полученных тремя методами экстракции (тотальная экстракция хроматина, высокосолева экстракция и экстракция с обработкой MNKазой). Идентификаторы генов (GenInfo Identifier) были преобразованы в идентификаторы UniProt.

Набора (Alvarez et al., 2023) [182]: был сформирован объединенный список белков, полученных методами NCC и iPOND.

Набора (Ugur et al., 2023) [183]: отобраны белки с ненулевыми значениями логарифма интенсивности сигнала ( $\log_2$  raw intensity) во всех трех репликах.

Набора белков ядра (Itzhak et al., 2016) [184]: отобраны белки, аннотированные как «преимущественно ядерные», то есть те, у которых доля интенсивности MC-сигнала в ядерной фракции превышала 85 % от суммарной интенсивности сигнала по всем исследованным субклеточным фракциям, согласно примененному авторами классификатору (Global classifier 2).

#### 2.1.4. Другие источники информации

К белкам — продуктам генов домашнего хозяйства (далее — конститутивным белкам) — были отнесены те, которые, согласно ресурсу НРА

(версия 23), были обнаружены во всех проанализированных тканях методом РНК-секвенирования (всего 8899 белков).

Стабильные мультисубъединичные комплексы были загружены из базы данных Complex Portal (дата обращения 7 января 2025 г.) [47]. Для анализа совместной встречаемости доменов были отобраны комплексы, в состав которых входили исключительно белки хроматина.

## **2.2. Разработка функционально-ориентированной классификации белков хроматина человека SimChrom**

Разработка эмпирической функционально-ориентированная классификация белков хроматина SimChrom включала в себя формирование иерархической структуры категорий (онтологию) и определение соответствующего набора белков.

В связи с тем, что на основании существующих данных не представлялось возможным провести строгое разграничение между белками хроматина и нуклеоплазмы, в основу классификации была положена обобщенная категория белков хроматина/нуклеоплазмы.

Для учета различных свойств белков хроматина были выделены четыре аспекта классификации — молекулярная функция, биологический процесс, физико-химические свойства и геномная локализация. Первые два аспекта соответствуют подходу, используемому в классификации Gene Ontology. Основанием для выделения аспекта физико-химических свойств послужило семейство HMG-белков, характеризующихся высокой электрофоретической подвижностью. С учетом пространственно-функциональной компартментализации генома в аспекте геномной локализации были выделены категории, соответствующие эу- и гетерохроматину, а также центромерным и теломерным регионам хромосом.

Формирование списка категорий классификации основывалось на общепринятых представлениях о функциональных классах белков хроматина [136]. В первую очередь, белки хроматина/нуклеоплазмы были разделены на

гистоновые и негистоновые. Категории белков, регулирующих эпигенетические процессы, были добавлены на основании информации из базы данных EpiFactors (шапероны гистонов, ремоделеры хроматина, белки машинерии посттрансляционных модификаций гистонов, а также модификаторы РНК и ДНК). Далее были добавлены категории, описывающие основные молекулярно-биологические процессы, связанные с функционированием хроматина (транскрипция, репарация и рекомбинация ДНК). На основе анализа литературы были добавлены другие функционально-значимые категории белков, например, SMC-комплексы, транскрипционные факторы, белки, осуществляющие отщепление хвостов гистонов. Категория РНК-связывающих белков была включена в классификацию с учетом данных HUGO Gene Nomenclature Committee о широком распространении в клетке длинных некодирующих РНК, вовлеченных в эпигенетическую регуляцию [185], а также большого числа белков, связывающих РНК, идентифицированных в МС-исследованиях хроматина.

Верификацию структуры классификации проводили путем сравнительного анализа с терминами GO, содержащими ключевые слова, связанные с хроматин-ассоциированными процессами, а также путем анализа терминов GO, обогащенных для белков, идентифицированных методом масс-спектрометрии в очищенных фракциях хроматина.

После формирования структуры классификации, в категории SimChrom были добавлены белки из базы знаний GO (набор аннотаций от 6 марта 2025 г.) на основе вручную определенных соответствий между категориями SimChrom и терминами GO, представленными в **Таблице А1 (Приложение А)**. Далее наборы белков в категориях были дополнены белками из специализированных баз данных, обзорных статей и экспериментальных исследований. Так, гистоновые белки были добавлены из HistoneDB 2.0 [49], белки, наносящие ПТМ гистонов — из Histome2 [145]; шапероны гистонов из [82]; белки, отщепляющие хвосты гистонов из [148,186]; пионерные ТФ из [34].

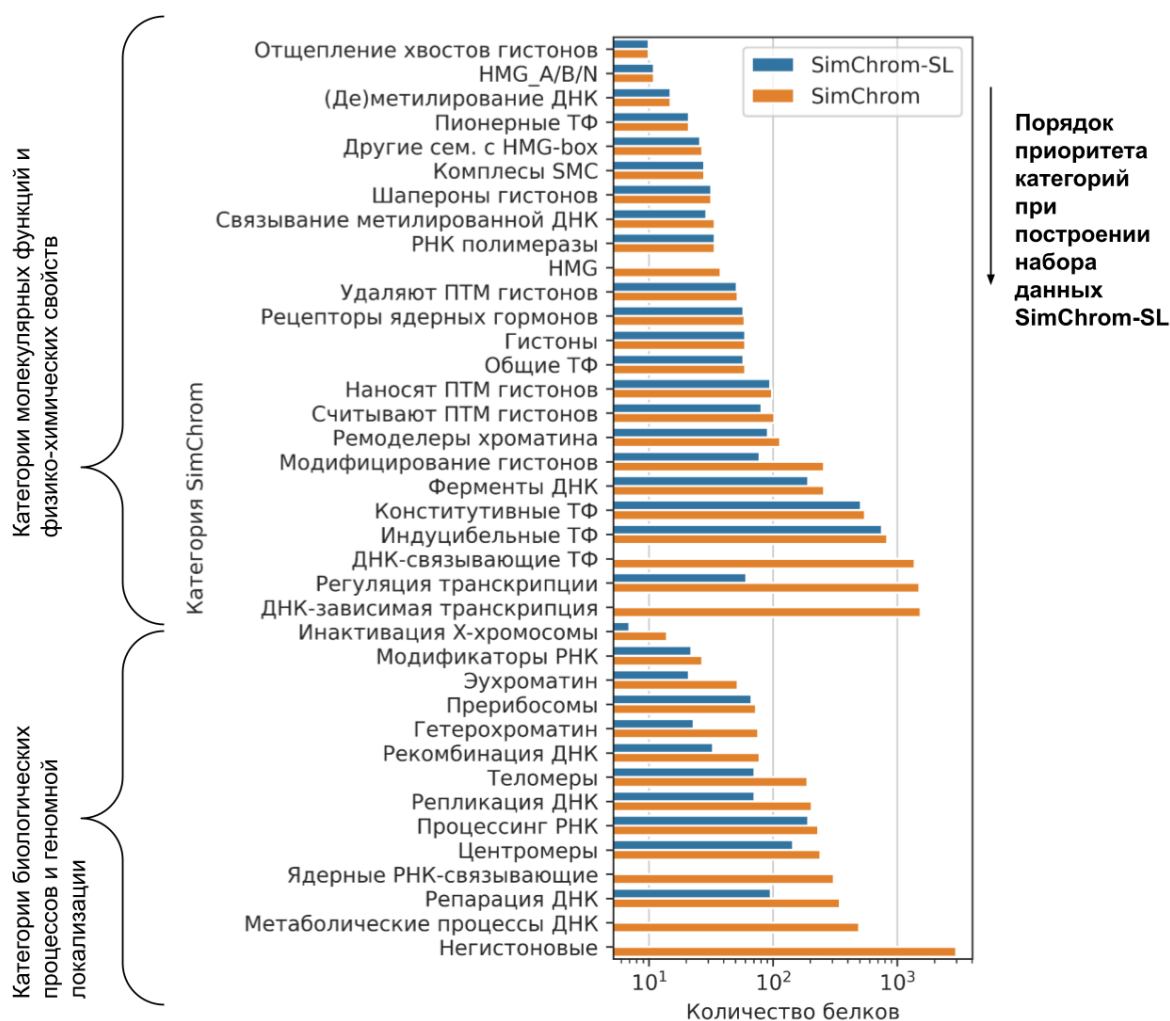
Для категории «ядерные РНК-связывающие белки» была проведена дополнительная фильтрация на основании данных о подтвержденной

экспериментально локализации в ядре клетки [150,151]. Гистоновые белки были исключены из категорий негистоновых белков.

Процесс разработки классификации был итеративным: с помощью предварительных вариантов классификации проводили оценку качества источников информации о репертуаре белков хроматина и ядра клетки человека, что, в свою очередь, позволило уточнить состав категорий классификации. Такой подход обеспечил самосогласованность разработанной классификации и эталонных наборов белков.

Разработку классификации SimChrom-SL (single labeled), в которой каждый белок отнесен только к одной категории, проводили в соответствии с заданным порядком приоритетов категорий (**Рисунок 11**). Приоритетными являлись категории с меньшим количеством белков и описывающие молекулярные функции или физико-химические свойства, за ними следовали категории с большим количеством белков и описывающие биологические процессы и геномную локализацию.

Для оценки качества собранного набора белков хроматина SimChrom был проведен анализ обогащения терминами GO для белков, присутствующих в SimChrom, но отсутствующих в разработанных наборах белков ядра NULOC\_CS и NULOC\_JT\_NECF, и наоборот, присутствующих в NULOC\_CS и отсутствующих в SimChrom. Обогащение терминами GO было проведено с помощью g:Profiler [187] (порог значимости  $p$ -значение  $< 0,05$  с поправкой Бонферрони на множественное тестирование). Для интерпретации были отобраны только драйверные GO-термины (согласно определению g:Profiler).



**Рисунок 11** — Приоритет категорий классификации при отборе белков для классификации SimChrom-SL.

### 2.3. Разработка эталонных наборов белков по локализации

Разработка эталонных наборов белков, локализованных в ядре или цитоплазме клетки, осуществлялась на основании комбинаций аннотаций субклеточной локализации из баз данных UniProt, HPA и OpenCell.

При формировании наборов белков учитывали следующие критерии: локализацию белков, согласованность информации между источниками, уникальность локализации в клеточном компартменте и достоверность аннотаций.

Для обозначения сформированных наборов использовали следующие буквенные сокращения:

- 1) локализация белков (ядерные — NULOC, неядерные — NON\_NULOC, цитоплазматические — CYTLOC);
- 2) согласованность информации между источниками (белки с согласованной локализацией в UniProt и HPA — CS, consensus; белки, аннотированные хотя бы в одном из источников — JT, joint);
- 3) уникальность локализации в клеточном компартменте (белки с аннотацией локализации только в соответствующем клеточном компартменте — UL, unique localization);
- 4) достоверность информации (наборы белков без учета критериев достоверности информации были помечены суффиксом NECF, no evidence code filtration). В качестве критериев достоверности использовали наличие источника информации согласно онтологии ECO для аннотаций UniProt; уровни достоверности Enhanced, Supported или Approved для HPA; а также второй и третий уровни достоверности для OpenCell.

Полный список сформированных наборов белков представлен в **Таблице 3**.

**Таблица 3** — Эталонные наборы белков ядра и цитоплазмы, сформированные на основе данных UniProt, HPA и OpenCell с учетом согласованности и достоверности аннотаций

| Название набора данных | Определение набора данных                                                                                                                                                                                                                    | Количество белков |
|------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------|
| NULOC_CS               | Белки, аннотированные как ядерные в UniProt (указан источник информации) и HPA (критерии достоверности: Enhanced, Supported, Approved), исключая белки, помеченные только как неядерные в OpenCell (критерии достоверности: уровень 2 или 3) | 3296              |
| NULOC_CS_NECF          | Белки, аннотированные как ядерные в UniProt и HPA, исключая белки, помеченные только как неядерные в OpenCell                                                                                                                                | 3988              |
| NULOC_CS_UL            | Белки, аннотированные только как ядерные в UniProt (указан источник информации) и HPA (крит. дост.: Enhanced, Supported, Approved), исключая белки, помеченные только как неядерные в OpenCell (крит. дост.: уровень 2 или 3)                | 1322              |
| NULOC_CS_UL_NECF       | Белки, аннотированные только как ядерные в UniProt и HPA, исключая белки, помеченные только как неядерные в OpenCell                                                                                                                         | 1322              |
| NULOC_JT               | Белки, аннотированные как ядерные хотя бы в одном из ресурсов: UniProt (указан источник информации), HPA (крит. дост.: Enhanced, Supported, Approved) или OpenCell (крит. дост.: уровень 2 или 3)                                            | 8048              |
| NULOC_JT_NECF          | Белки, аннотированные как ядерные хотя бы в одном из ресурсов: UniProt, HPA или OpenCell                                                                                                                                                     | 8912              |
| NULOC_JT_UL            | Белки, аннотированные только как ядерные хотя бы в одном из ресурсов: UniProt (указан источник информации), HPA (крит. дост.: Enhanced, Supported, Approved) или OpenCell (крит. дост.: уровень 2 или 3)                                     | 4292              |
| NULOC_JT_UL_NECF       | Белки, аннотированные только как ядерные хотя бы в одном из ресурсов: UniProt или HPA, исключая белки, помеченные только как неядерные в OpenCell                                                                                            | 4292              |
| NON_NULOC_CS           | Белки, аннотированные как неядерные одновременно в UniProt (указан источник информации) и HPA (крит. дост.: Enhanced, Supported, Approved)                                                                                                   | 3674              |
| NON_NULOC_JT           | Белки, аннотированные как неядерные хотя бы в одном из ресурсов: UniProt (указан источник информации), HPA (крит. дост.: Enhanced, Supported, Approved) или OpenCell (крит. дост.: уровень 2 или 3)                                          | 11479             |
| CYTLOC_CS_UL           | Белки с аннотацией только в цитоплазме одновременно в UniProt (указан источник информации) и HPA (крит. дост.: Enhanced, Supported, Approved)                                                                                                | 2026              |

## 2.4. Подготовка протеомных данных для количественного анализа белков хроматина

### 2.4.1. Данные из базы данных PaxDb

Для анализа количественной представленности (относительного количественного содержания) белков протеома человека использовали два набора данных из базы данных PaxDb (версия 4.2) [180]:

- 1) набор «H. sapiens — Whole organism (Integrated)» с данными представленности для 99 % протеома, далее PaxDb\_INT;
- 2) набор «Whole organism, SC (PeptideAtlas, Aug 2014)» с данными для 84 % протеома, но с наибольшей согласованностью данных (на основе информации о белок-белковых взаимодействиях), далее PaxDb\_PA.

На основе значений представленности белков, выраженных в единицах parts per million (ppm), отражающих относительную долю белка в протеоме, белки протеома человека были разделены на две группы: низкопредставленные (<1 ppm) и высокопредставленные (>1 ppm). Для белков, кодируемых несколькими генами, например, гистонов, значения представленности, приведенные для генов, суммировали.

Суммарная представленность всех белков категории рассчитывалась как сумма представленностей отдельных белков. Массовую долю белков категории в хроматоме рассчитывали по **формуле 12**

$$W_{\text{категории}} = \frac{\sum_{i \in \text{категории}} A_i \times M_i}{\sum_{j \in \text{хроматом}} A_j \times M_j}, \quad (12)$$

где  $A$  — представленность белка (ppm),  $M$  — молекулярная масса белка.

Для визуализации распределений белков по представленности были построены гистограммы логарифмированных значений представленности (ширина столбца 0,15) с последующим сглаживанием гауссовым ядром.

### 2.4.2. Данные о представленности белков из протеомных исследований

Для сравнительного анализа относительной представленности белков различных категорий хроматина также использовали данные из оригинальных исследований хроматомов (значения интенсивностей MS-сигналов или значения iBAQ). Если одно значение представленности соответствовало нескольким белкам, то оно равномерно распределялось между ними. Для белков с несколькими записями в рамках одного исследования брали медианное значение представленности. Для анализа результатов исследований (Kustatscher et al., 2014; Alabert et al., 2014) представленность белка определяли как сумму MS-сигналов, для (Ginno et al., 2018) — как медиану интенсивности MS-сигнала по шести репликам эксперимента (по три реплики на каждую фазу клеточного цикла G1 и S) на основе набора данных «Chromatome Reporters». Для анализа (Shi et al., 2021) использовали медиану значений iBAQ по трем репликам; для анализа (Itzhak et al., 2016) — рассчитанное число копий на клетку (англ. estimated copy number per cell), для изоформ одного белка брали медианное значение; для анализа (Ugur et al., 2023) — медианное значение логарифма интенсивности сигнала по трем репликам.

## 2.5. Методы анализа физико-химических свойств белков

Для аннотации внутренне неупорядоченных регионов использовали структуры белков, предсказанные с помощью AlphaFold2, а также рассчитанную на их основе площадь поверхности, доступную растворителю (по данным [118]). Регион белка относили к внутренне неупорядоченному, если он содержал не менее четырех последовательно расположенных остатков со значениями площади поверхности, доступной растворителю, выше 0,55, рассчитанными со скользящим окном шириной 20 а.о. [118]. Неупорядоченные регионы считали разделенными при наличии между ними упорядоченного фрагмента длиной более четырех остатков, что соответствует минимальной длине  $\alpha$ -спирали. В качестве неупорядоченных хвостов белков рассматривали N- или C-концевые неупорядоченные регионы.

Заряд белка определяли как положительный ( $\geq 1$ ), отрицательный ( $\leq -1$ ) или нейтральный ( $-1 < \text{заряд} < 1$ ) в зависимости от количества положительно (лизины и аргинины) и отрицательно (аспартаты и глутаматы) заряженных остатков.

Суммарный средний заряд остатков хвостов анализируемой группы белков рассчитывали как сумму зарядов остатков, усредненных по каждой позиции (для хвостов длиннее 80 остатков анализировали только первые 80 позиций). Профиль заряда для хвостов той же группы белков строили по усредненным значениям заряда в каждой позиции с последующим сглаживанием скользящим окном размером 10 а.о.

В рамках анализа заряда белков использовали три метрики: (i) заряд (положительный, отрицательный, нейтральный) отдельных белков; (ii) заряд (положительный, отрицательный, нейтральный) белковых молекул с учетом данных о представленности белков согласно набору данных PaxDb\_PA; (iii) суммарный заряд белковых молекул с учетом величины заряда и представленности белковых молекул.

Для визуализации различий в аминокислотном составе белков хроматина и цитоплазмы использовали нелинейный метод снижения размерности Uniform Manifold Approximation and Projection (UMAP) с параметрами по умолчанию и с предварительной стандартизацией относительного количества аминокислотных остатков в белках (метод StandardScaler, библиотека scikit-learn).

Для сравнения групп белков по физико-химическим признакам (медианам значений относительного количества аминокислотных остатков) использовали критерий Манна-Уитни с поправкой Бонферрони, уровень значимости  $p$ -значение  $< 0,05$ . Обогащение определяли как статистически значимое повышенное относительное содержание аминокислотного остатка в белках одной группы по сравнению с другой; коэффициент обогащения рассчитывали как отношение соответствующих медианных долей.

## **2.6. Методы структурной характеристики и анализа доменного состава белков**

Соотнесение аминокислотных остатков последовательностей белков и экспериментально разрешенных структур соответствующих белков, депонированных в базе данных PDB, проводили с использованием PDB API (дата обращения 15 января 2025 г.).

Аннотации структурных доменов CATH использовали согласно данным базы данных CATH v4.3.0 [188].

В качестве доменных моделей Pfam использовали регионы семейства Pfam-A (версия Pfam 37.0); отбирали только следующие типы регионов: domain, family, repeat, coiled-coil. Доля доменных моделей Pfam, представленных в PDB, была получена путем анализа данных из базы данных InterPro. Данные об аннотации функциональных доменов белков из других базах данных (например, InterPro, PANTHER) были получены с помощью InterPro API (база данных InterPro версия 103.0); отбирали только следующие типы доменов: domain, homologous\_superfamily, family, repeat, coiled\_coil. Аннотации TED-доменов были загружены с использованием TED API v1 [124].

Аннотации доменов из разных источников (TED, CATH и Pfam) считали перекрывающимися, если длина перекрытия превышала половину длины наиболее короткой аннотации.

Для поиска структурных соответствий TED-доменов и доменных моделей Pfam в базах данных PDB («PDB100 20240101») или CATH («CATH50 4.3.0») использовали Foldseek Search [125] с модулем easy-search, опцией «exhaustive-search» и остальными параметрами по умолчанию. Критериями совпадения запрашиваемой структуры и мишени считали вероятность (оцененную вероятность для запроса и мишени быть гомологичными, т.е., например, относиться к одному суперсемейству базы данных SCOPe) выше 0,9 и покрытие последовательности запроса последовательностью мишенью выше 0,5.

Для каждого TED-домена было идентифицировано наилучшее структурное соответствие в PDB (с наибольшим значением оценки вероятности по FoldSeek). TED-домены разделяли на группы на основе:

- 1) уровня идентичности последовательности с наилучшим структурным соответствием из PDB;
- 2) таксономической принадлежности структурного соответствия («Homo sapiens», «Mammalia», «Other Vertebrata», «Protostomia», «Viridiplantae», «Fungi», «Archaea», «Bacteria», «Viruses», «Other», синтетические конструкции или неклассифицированные последовательности «N/A»).

Дополнительное соотнесение TED-доменов с доменами базы данных CATH проводили путем поиска по аминокислотной последовательностям с использованием аннотаций CATH/GENE3D.

Аналогичным образом с помощью FoldSeek выполнили поиск структурных соответствий в PDB для доменных моделей Pfam. Для каждой доменной модели была рассчитана медиана идентичности аминокислотных последовательностей наилучших соответствий входящих в ее состав доменов.

Неохарактеризованные структурные домены (потенциально представляющие новые типы структурных гомологичных суперсемейств) были определены как TED-домены, которые не удалось сопоставить ни с одной известной структурой белка или гомологичным суперсемейством белковых структур (с помощью поиска FoldSeek в PDB и CATH). Для этих доменов функциональные и другие аннотации были загружены из InterPro, при этом рассматривали только соответствия, в которых аннотации совпадали более чем на половину длины аннотации TED-домена. Предсказание терминов Gene Ontology для этих структурных доменов выполнили с использованием графовой сверточной нейронной сети DeepFri [189] на основе аминокислотной последовательности доменов (модель CNN) и структуры (модель GCN).

Доменное разнообразие рассчитывали как количество различных доменных моделей Pfam в белках одной категории SimChrom-SL; относительное доменное

разнообразие — как отношение доменного разнообразия к количеству белков в соответствующей категории SimChrom-SL.

Из доменных моделей Pfam были отобраны те, которые присутствовали в нескольких копиях или в комбинации с другими доменами по крайней мере в трех белках, регулирующих эпигенетические процессы (то есть относящиеся к следующим категориям SimChrom-SL: шапероны гистонов; белки, наносящие/считывающие/удаляющие ПТМ гистонов; ремоделеры хроматина; белки, наносящие и удаляющие модификации ДНК; белки, модифицирующие РНК). Далее эти домены были вручную классифицированы по функциям на основе анализа литературы. Для классификации доменов были использованы следующие группы: метилирование/ацетилирование/фосфорилирование гистонов; ремоделирование хроматина; связывание с гистонами; связывание с ДНК; связывание с РНК; метилирование ДНК; димеризация/олигомеризация белков; белок-белковые взаимодействия. Внутри групп, ассоциированных с ПТМ, было введено дополнительное деление по подгруппам: наносящие, считывающие или удаляющие ПТМ гистонов.

Совместное присутствие доменных моделей в белке (наличие домена А при присутствии другого домена Б) оценивали с помощью условной вероятности (формула 13)

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad (13)$$

### **Глава 3. Критический анализ разнородных источников информации о репертуаре белков хроматина и ядра клетки человека**

Неоднозначность определения репертуара белков хроматина и существенный разброс в оценках их количества явились предпосылками для проведения сравнительного анализа и критического осмысления концептуально различающихся источников информации. В анализ также были добавлены белки ядра клетки, так как данные о клеточной локализации являются независимым способом валидации белков интерфазного хроматина.

В данной главе представлены результаты комплексного сравнительного мета-анализа онтологий классификаций и репертуара белков хроматина и ядра клетки человека. В результате были определены преимущества и ограничения различных подходов к получению наборов белков ядра и хроматина.

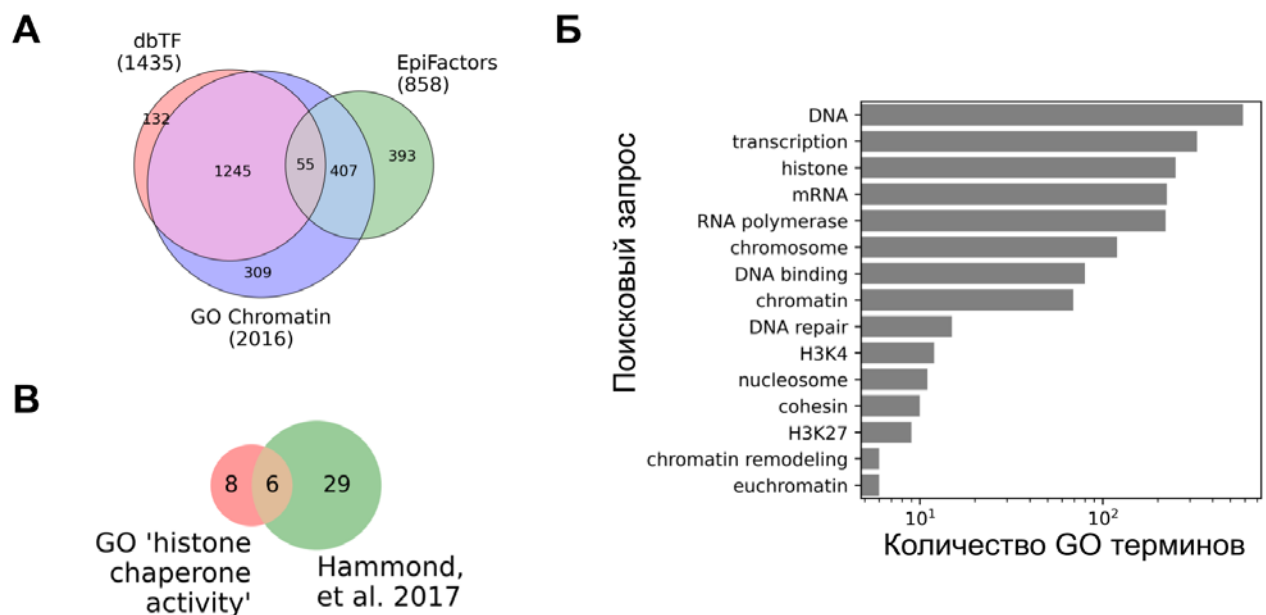
#### **3.1. Особенности аннотирования белков хроматина в базе знаний Gene Ontology**

Анализ структуры онтологии базы знаний Gene Ontology, а также соответствующего набора белков, ассоциированных с хроматином человека, привел к выявлению ряда ограничений системы. В частности, отсутствуют группы белков, относящиеся к отдельным функциональным или исторически сложившимся категориям. Это препятствует формированию репрезентативного списка белков хроматина.

В базе знаний Gene Ontology термин «хроматин» относится к аспекту клеточной локализации и определяется как «упорядоченный и организованный комплекс ДНК, белков и иногда РНК, образующий хромосому». К данному термину отнесено 2019 белков человека.

Установлено, что ряд функционально значимых терминов, которые должны быть связаны с термином «хроматин», в иерархии терминов Gene Ontology с ним не связаны. К ним относятся «связывание с нуклеосомной ДНК», «активность

ДНК-связывающего транскрипционного фактора», «активность моно-метилтрансферазы гистонов по положению H3K27», «связывание с гистонами», «ремоделирование хроматина» и другие. Вследствие этого более 500 белков человека с функциями, характерными для белков хроматина (в частности, белки, регулирующие эпигенетические процессы согласно базе данных EpiFactors и транскрипционные факторы из базы данных dbTF), отсутствуют в термине «хроматин» Gene Ontology (**Рисунок 12А**).



**Рисунок 12** — Особенности аннотации Gene Ontology по части белков хроматина человека. **(А)** Диаграмма Венна пересечения списков белков хроматина: (i) белки, регулирующие эпигенетические процессы (из базы данных EpiFactors), (ii) транскрипционные факторы (из базы данных dbTF), (iii) белки, отнесенные к термину «хроматин» (Gene Ontology). **(Б)** Количество терминов Gene Ontology, в названии которых присутствуют слова, ассоциированные с хроматином. **(В)** Диаграмма Венна пересечения списков белков, относящихся к шаперонам гистонов, согласно термину Gene Ontology «активность шаперонов гистонов» и литературному источнику (Hammond et al., 2017).

Кроме того, в Gene Ontology представлено более 100 терминов, содержащих хроматин-ассоциированные ключевые слова ( «ДНК», «транскрипция», «гистон», «РНК-полимераза» и другие), что делает затруднительным ручной отбор этих терминов (**Рисунок 12Б**).

Иерархический принцип организации Gene Ontology приводит к присутствию аннотаций белков в терминах более высокого порядка. Вследствие этого некоторые хроматин-ассоциированные термины содержат белки дочерних терминов иной функциональной принадлежности. Например, термин «транскрипция по матрице ДНК» включает в себя «митохондриальную транскрипцию», а термин «экспрессия генов» объединяет функционально разнородные процессы, такие как «созревание белка» и «трансляция».

Выявлено отсутствие отдельных терминов Gene Ontology для описания некоторых исторически сложившихся групп белков хроматина, например, белков HMG; белков, отщепляющих хвосты гистонов; а также общих и пионерных транскрипционных факторов. Кроме того, в GO отсутствует отдельный термин для гистонов — они, наряду с негистоновыми белками HMGA1 и LMNTD2, входят в состав термина «структурная составляющая хроматина».

В то же время, вследствие того что Gene Ontology не является специализированной базой знаний для описания белков хроматина, она содержит устаревшие термины (например, «ядерный матрикс») и включает в себя не все аннотации генов, представленные в современной литературе. В частности, только 6 из 35 белков гистоновых шаперонов, перечисленных в работе [82], входят в состав одноименного термина Gene Ontology (**Рисунок 12В**).

### **3.2. Сравнительный анализ данных о клеточной локализации белков**

Был проведен анализ онтологий и соответствующих репертуаров белков ядра клетки по данным из баз данных UniProt, HPA и OpenCell. Списки белков с учетом критериев достоверности информации из UniProt и HPA совпадают только на 60-70 % (8981 белок). Суммарно три ресурса предоставляют информацию о 17335 белках человека (86 % референсного протеома).

### 3.2.1. Количество и доля ядерных белков в протеоме человека

По данным UniProt, HPA и OpenCell количество белков ядра клетки человека, отобранных с учетом критериев достоверности информации, составляет 4717, 6500 и 640 соответственно, что соответствует 33 %, 53 % и 49 % от числа белков в этих ресурсах (**Рисунок 13**). Объединение данных из трех ресурсов позволило получить «оценку сверху» для количества ядерных белков, которая составляет ~8000 белков (46 % от числа белков с клеточной локализацией согласно хотя бы одному ресурсу).

Таким образом, согласно современным базам данных о локализации белков, примерно половина белков протеома человека относится к ядру клетки. В зависимости от выбранной строгости критериев достоверности информации оценка доли белков ядра в протеоме человека составляет 35-65 %.

### 3.2.2. Различия в иерархиях терминов субъядерной локализации белков

Онтологии терминов, используемые для описания локализации ядерных белков в клетке, различаются между ресурсами UniProt, HPA и OpenCell. В HPA и OpenCell представлено 9 и 6 терминов для описания структур ядра соответственно. В UniProt, несмотря на наибольшее разнообразие терминов субъядерной локализации в онтологии (более 15), только для 19 % ядерных белков (882 белка) указана морфологическая подструктура ядра (**Рисунок 13**). В тоже время в онтологиях HPA и OpenCell отсутствует обобщающий термин «ядро», из-за чего подавляющее большинство ядерных белков в HPA относится к нуклеоплазме (92 %), и только для 33 % белков доступны аннотации о локализации в других структурах ядра.



**Рисунок 13** — Информация о репертуаре белков ядра согласно ресурсам UniProt, HPA и OpenCell. Для каждого из ресурсов показаны онтологии субъядерной локализации белков ядра (слева) и количество соответствующих белков протеома человека (справа от названия каждой категории). Количество общих белков в одноименных категориях в UniProt и HPA выделено желтым цветом. На схеме белок может относиться к нескольким категориям. В верхней части схемы указано общее количество белков в базах данных, а также количество белков с учетом применения критериев достоверности информации.

Интерактивная версия рисунка доступна по адресу:

<https://simchrom.intbio.org/#localization>.

Общими в исследуемых онтологиях являются следующие ядерные компартменты: «оболочка ядра или ядерная мембрана», «ядрышко», «нуклеоплазма», «ядерные тельца/спеклы/точки» (**Рисунок 13**). Вместе с тем одноименные термины из разных ресурсов могут находиться в онтологиях на разных уровнях иерархии, а также быть представленными в качестве

самостоятельных или собирательных терминов (например, «нуклеоплазма» в НРА — это и самостоятельный термин, и собирательное название для группы терминов).

Термин «хроматин» явно присутствует только в онтологии OpenCell. В связи с отсутствием соответствующего термина в UniProt и НРА, ожидается, что белки хроматина в этих ресурсах должны относиться к терминам «ядро» и «нуклеоплазма» соответственно. В UniProt присутствует отдельный термин «хромосома», не входящий в состав «ядра» и унаследованный из Gene Ontology. При этом 84 % белков, относящихся к хромосоме по UniProt, имеют также аннотации о локализации в ядре клетки. В онтологии НРА присутствует термин «митотическая хромосома», который является частью «нуклеоплазмы».

Вышеупомянутые различия в онтологиях систем классификации ядерных белков отражают: (1) разные подходы к классификации, (2) наследование сторонних онтологий (например, GO в случае UniProt), (3) эволюцию представлений об организации ядра (например, онтология UniProt содержит устаревший термин «ядерный матрикс»), (4) предел по разрешающей способности экспериментальных методов (в случае НРА и OpenCell).

Приведенные наблюдения подчеркивают сложность организации живых клеток и, как следствие, сложность описания клеточных компартментов в виде иерархической онтологии.

### **3.2.3. Сравнительный анализ аннотаций субъядерной локализации белков в UniProt, НРА и OpenCell**

Наибольшей по количеству белков группой ядерных белков в НРА и OpenCell является «нуклеоплазма», 5948 и 487 белков соответственно. В UniProt к нуклеоплазме отнесены только 170 белков (еще 245 белков отнесены к «ядерным тельцам» — к компартменту, который не является частью нуклеоплазмы согласно данным UniProt, но относится к ней по онтологии НРА). Из 3288 белков с согласованной аннотацией о локализации в ядре по данным из UniProt и НРА для 95 % белков с аннотацией о локализации в нуклеоплазме по данным НРА

отсутствует детальная информация о локализации в субъядерной структуре по данным UniProt.

Второй по числу белков категорией как в UniProt, так и в HPA является категория «ядрышко» (372 и 1191 белок соответственно). Однако это не верно для OpenCell, предположительно, из-за разрешающей способности используемого метода анализа.

Для сравнения репертуара белков между онтологиями в категориях, различающихся по составу входящих в них терминов, были разработаны обобщенные термины (категории) субъядерной локализации: «нуклеоплазма», «ядерная оболочка», «ядрышко», «ядерные тельца». Так, согласованность между наборами белков обобщенных субъядерных категорий из UniProt и HPA, рассчитанная с использованием коэффициента сходства Жаккара или метрики F1 оказалась низкой (**Таблица 4**). Среди белков ядерной мембраны/оболочки только 46 белков имеют согласованную аннотацию между UniProt и HPA, в то время как аннотации 368 белков не совпадают.

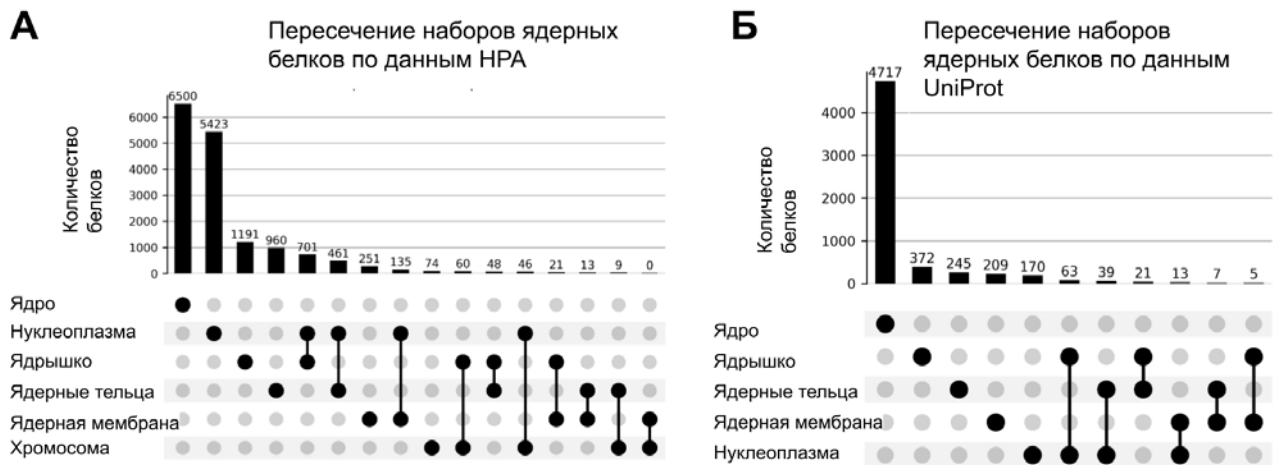
Таким образом, терминологическое определение субъядерных локализаций, а также представленные в них белки существенно различаются между ресурсами UniProt и HPA.

Другим аспектом анализа была одновременная локализация белков в нескольких ядерных компартментах. Результаты приведены на графиках Upset (**Рисунок 14**). По данным из HPA, для большинства белков с локализацией в ядрышке (59 %) характерна локализация в нуклеоплазме, что объясняется отсутствием мембраны у ядрышка и диффузией белков между этими компартментами. По данным UniProt только 17 % белков ядрышка локализованы в нуклеоплазме, что является следствием недостаточной аннотации белков категории «ядрышко» в данном ресурсе.

**Таблица 4** — Метрики оценки сходства наборов белков из обобщенных категорий субъядерной локализации по данным UniProt и HPA

|                               | Белки с достоверной локализацией | «Ядро» | «Ядрышко» | «Ядерная оболочка» | «Ядерные тельца» | «Нуклеоплазма» |
|-------------------------------|----------------------------------|--------|-----------|--------------------|------------------|----------------|
| UniProt (A)                   | 14147                            | 4717   | 373       | 209                | 245              | 170            |
| HPA (B)                       | 12172                            | 6500   | 1191      | 251                | 960              | 5423           |
| Истинно положительные (A И B) | 8981                             | 3288   | 185       | 46                 | 87               | 122            |
| Объединение (A ИЛИ B)         | 17340                            | 7932   | 1380      | 414                | 1118             | 5471           |
| Коэффициент сходства Жаккара  | 0,52                             | 0,41   | 0,13      | 0,11               | 0,08             | 0,02           |
| Ложно-положительные (B, НЕ A) | 3191                             | 3212   | 1006      | 205                | 873              | 5301           |
| Ложно-отрицательные (A, НЕ B) | 5168                             | 1432   | 189       | 163                | 158              | 48             |
| Точность ИП / (ИП + ЛП)       | 0,74                             | 0,51   | 0,16      | 0,18               | 0,09             | 0,02           |
| Полнота ИП / (ИП + ЛО)        | 0,63                             | 0,7    | 0,49      | 0,22               | 0,36             | 0,72           |
| F1-мера                       | 0,68                             | 0,59   | 0,24      | 0,2                | 0,14             | 0,04           |

Аналогичная ситуация наблюдалась при сравнении множественной субъядерной локализации белков ядерных телец. Так, 48 % и 16 % белков ядерных телец по данным HPA и UniProt соответственно имеют аннотации о локализации в нуклеоплазме. Другим наблюдением является наличие у 54 % белков ядерной мембраны по данным HPA второй локализации в нуклеоплазме, в то время как для соответствующих белков из UniProt этот показатель составляет 6 %. Это может отражать временную локализацию белков нуклеоплазмы в области ядерной мембраны посредством взаимодействия с белками, содержащими трансмембранные домены.



**Рисунок 14** — Сравнительный анализ наборов белков по их субъядерной локализации: (А) согласно данным HPA; (Б) согласно данным UniProt. Используются обобщенные термины субъядерной локализации.

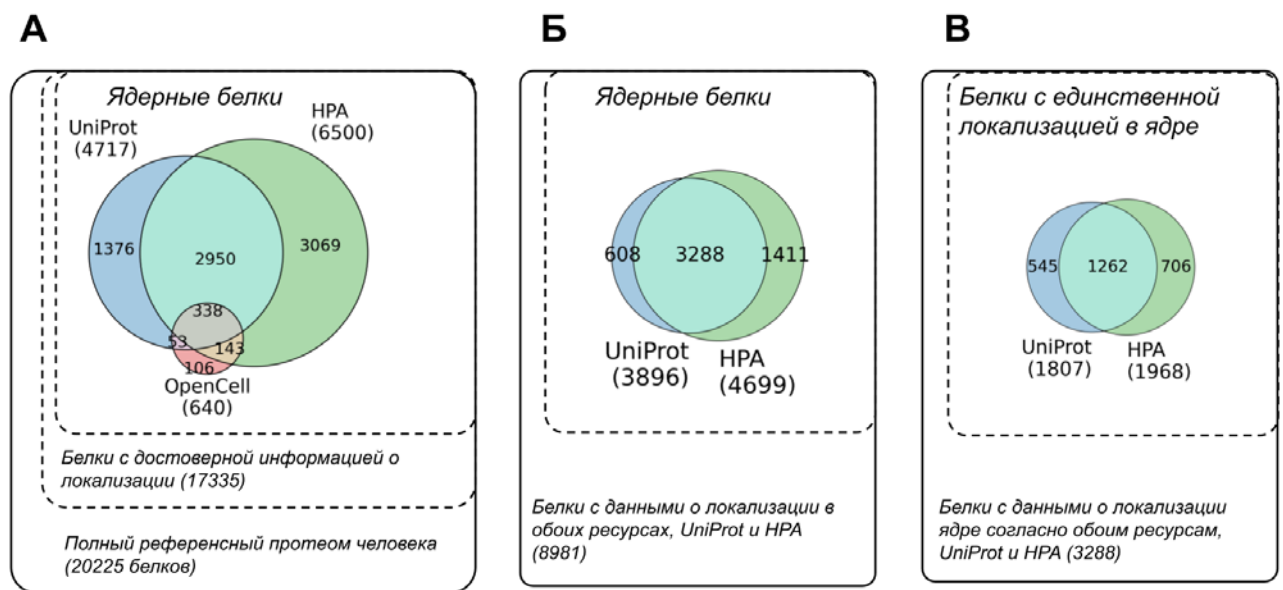
### 3.2.4. Множественная локализация белков ядра

Ресурсы UniProt и HPA отличаются количеством аннотаций о локализации белков: среднее число аннотаций в UniProt — одна, в HPA — две. Из чего можно предположить наличие более подробной аннотации субклеточной локализации белков в HPA. Для дальнейшего анализа термины клеточной локализации из UniProt и HPA были сгруппированы в три категории по принадлежности к ядру, цитоплазме или эндомембранной системе.

Количество белков с аннотациями о ядерной локализации согласно одновременно двум ресурсам, UniProt и HPA, составляет 3288 белков, что соответствует 51 % ядерных белков по данным HPA и 70 % — по данным UniProt (**Рисунок 15А**).

Вследствие того, что количество белков с наличием достоверных аннотаций о какой-либо локализации в клетке в обоих ресурсах составляет 8981 белок, для оценки согласованности аннотаций между UniProt и HPA были сделаны расчеты с учетом общих для этих ресурсов белков. Так, из ~9000 белков ~3300 аннотированы как ядерные одновременно в UniProt и HPA, в то время как оставшиеся 2019 белков имеют аннотации о ядерной локализации только в одном из ресурсов (~600 — в

UniProt и ~1400 — в HPA) (**Рисунок 15Б**). Из этих 2019 белков: (i) 1670 (83 %) белков имеют совпадающие аннотации о локализации в более крупных клеточных структурах/компартаментах (цитоплазма, эндомембранная система); (ii) 1254 белка локализованы по крайней мере в одной общей более крупной структуре клетки; (iii) 765 белков не содержат общих сведений о других (неядерных) локализациях. Коэффициент сходства Жаккара для этих наборов белков составляет 0,62 (3288/5307).

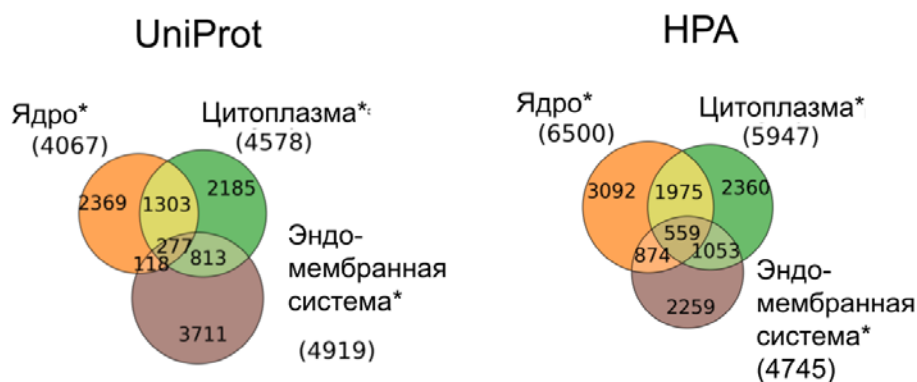


**Рисунок 15** — Согласованность аннотаций UniProt и HPA о ядерной локализации белков. (**А**) Диаграмма Венна для списков белков ядра по данным UniProt и HPA, составленным из набора белков с достоверными сведениями о локализации по крайней мере в одном из ресурсов. (**Б**) Диаграмма Венна для белков ядра, имеющих достоверные сведения о локализации в обоих ресурсах (UniProt и HPA). (**В**) Диаграмма Венна для белков с единственной локализацией в ядре и имеющих достоверные сведения о локализации в обоих ресурсах (UniProt и HPA).

При отборе белков с достоверными аннотациями клеточной локализации в двух ресурсах одновременно (UniProt и HPA) консенсус относительно единственной локализации в ядре клетки был достигнут только для половины белков (**Рисунок 15В**). Так, на каждые пять белков, отнесенных исключительно к ядру по данным одного из ресурсов, два белка в альтернативном ресурсе имеют аннотации о неядерной локализации.

Полученные данные отражают недостаточную специфичность информации о локализации белков в базах данных. Это связано с наличием частичной информации о других (неядерных) локализациях, а также с возможным присутствием ложноположительных аннотаций, возникающих из-за особенностей экспериментальных методов или некорректного назначения термина локализации. Устранение влияния подобных ошибок представляет собой нетривиальную задачу.

По данным UniProt и HPA половина белков ядра локализованы исключительно в ядре, ~40 % ядерных белков также локализованы в цитоплазме, 12–22 % — в эндомембранной системе (**Рисунок 16**).



**Рисунок 16** — Множественная локализация белков ядра в обобщенных компартментах клетки (цитоплазма и эндомембранная система).

С целью выявления причин несогласованности аннотаций был проведен перекрестный анализ белков, ядерная локализация которых была указана только в одном из ресурсов (154 белка по данным UniProt и 611 белков по данным HPA). Большинство таких белков в UniProt относится к ядерным белкам без более детальной аннотации, а в HPA — преимущественно к белкам нуклеоплазмы или ядрышка. Среди неядерных категорий в альтернативном ресурсе в обоих случаях одной из наиболее многочисленных по числу белков была «цитоплазма», однако, согласно результатам гипергеометрического теста, обогащение данного термина не было статистически значимым.

Для белков, аннотированных как ядерные только в UniProt, в HPA статистически значимым обогащенным термином (p-значение <0,05) с

наибольшим числом белков был термин «везикулы» (коэффициент обогащения — 1,3). Для белков с ядерной локализацией только по данным НРА, в UniProt статистически значимое обогащение было выявлено для следующих терминов с наибольшим количеством белков: «секретируемые» и «внеклеточный матрикс», коэффициенты обогащения — 1,6 для обоих. Термины «хромосома» и «кинетохор» также были обогащены (коэффициенты обогащения 6,1 и 2,7 соответственно), но включали менее 20 белков.

Отмеченные выше случаи множественной локализации белков ядра в других клеточных компартментах (цитоплазма, эндомембранная система) могут отражать функционально значимое перемещение белков между компартментами клетки. В частности, некоторые транскрипционные факторы (например, NF-κB, STAT, p53, TAF7, YAP/TAZ) регулируют клеточные процессы посредством перемещения между цитоплазмой и ядром [190–192]. Более того, даже для некоторых гистоновых белков, в частности гистонов типа H2B, было отмечено перемещение в цитоплазму в ответ на клеточный стресс для выполнения неканоничных функций [193,194]. С другой стороны, учитывая синтез ядерных белков в цитоплазме и последующий импорт в ядро клетки, ошибочные аннотации множественной локализации белков ядра могут являться следствием экспериментальных артефактов или неоптимального выбора пороговых соотношений сигнала и шума.

Выявленные расхождения в аннотациях ядерной локализации белков могут быть объяснены несколькими причинами. Во-первых, неполной аннотацией множественной локализации белков: из ~2000 белков ядра, аннотированных только в одном из ресурсов, около 60 % имеют совпадающие аннотации о локализации в клеточных компартментах, отличных от ядра. Во-вторых, наличием неточностей в аннотациях локализации: в НРА выявлена тенденция аннотации ядерных белков как везикулярных, что может объяснить до 30 % случаев несовпадающих аннотаций, тогда как в UniProt отмечена тенденция аннотации ядерных белков как секретируемых или относящихся к внеклеточному матриксу, что может объяснить до 15 % таких случаев.

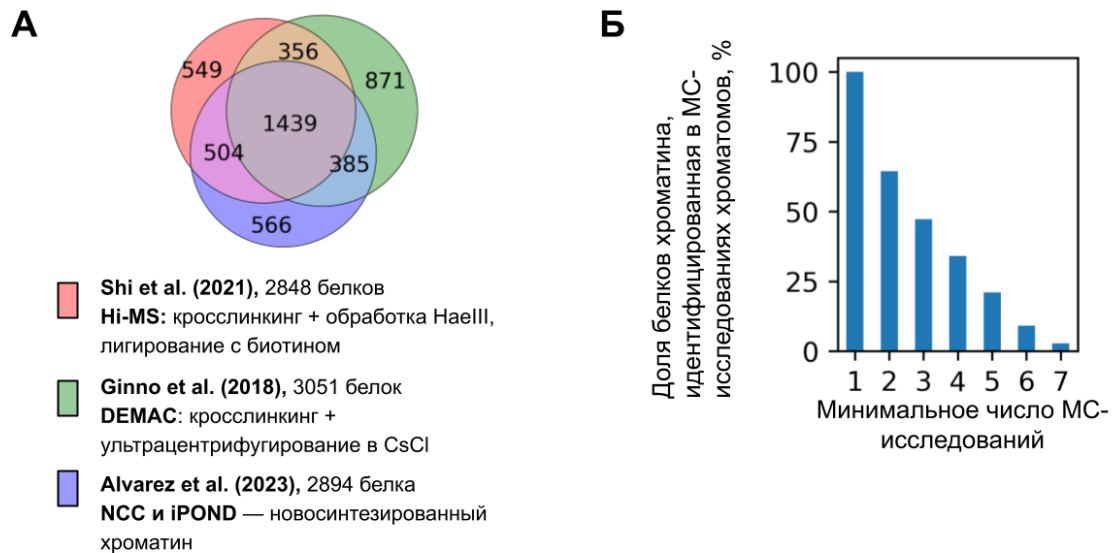
### 3.3. Сравнительный анализ протеомных данных о составе хроматина

Другим источником информации о белковом составе хроматина являются результаты масс-спектрометрической идентификации белков, полученные при анализе очищенных фракций хроматина или ядерного экстракта (нуклеома) [6,195]. Для сравнительного анализа были отобраны публично доступные результаты масс-спектрометрических исследований хроматомов, опубликованные после 2010-го года (**Таблица 2**). Из них пять исследований посвящены хроматомам интерфазного хроматина, два — новосинтезированному хроматину и одно — ядерному нуклеому.

Для описания результатов в этом разделе будут использованы разработанные референсные наборы белков хроматина и ядра клеток человека, которые подробно описаны в **Главе 4**. Кратко, SimChrom — список 3045 белков хроматина человека, NULOC\_JT\_NECF — наиболее полный список 8912 потенциальных белков ядра по аннотации UniProt или HPA без учета критериев достоверности информации; NULOC\_CS — наиболее достоверный список 3296 белков с достоверной аннотацией в ядре одновременно в UniProt и HPA.

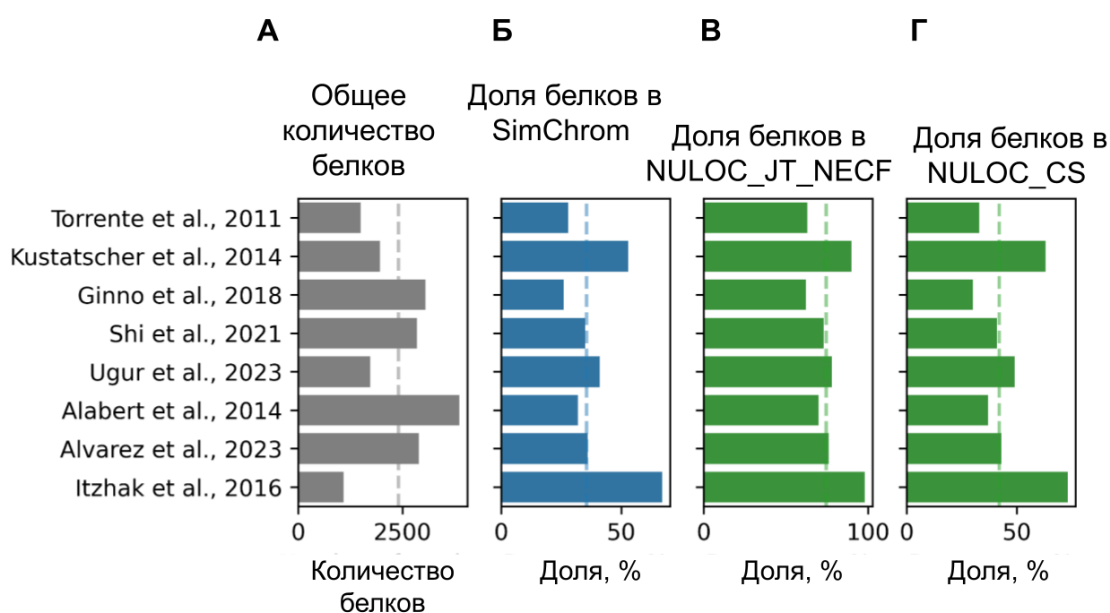
#### 3.3.1. Анализ белков, идентифицированных в протеомных экспериментах

Количество идентифицированных в МС-исследованиях белков хроматина человека варьирует от 1,5 до 3 тысяч. Эти различия обусловлены особенностями пробоподготовки фракций белков хроматина и чувствительностью масс-спектрометрических детекторов. В результате попарного сравнения наборов белков хроматина сопоставимого размера (~3000 белков) показано, что они совпадают не более чем на 68% (**Рисунок 17А**). При этом количество белков, идентифицированных во всех семи проанализированных исследованиях хроматина, составило 179 белков, что свидетельствует о выраженной зависимости получаемых результатов от биологического материала (клеточной линии) и техники экстракции хроматина (**Рисунок 17Б**).



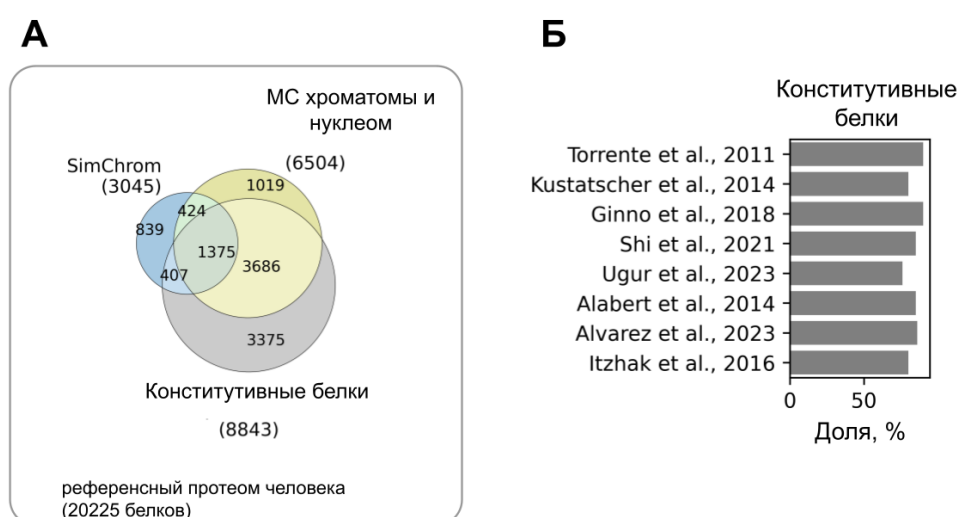
**Рисунок 17** — Сравнение результатов протеомных исследований хроматина. (А) Диаграмма Венна для списков белков сопоставимого размера, идентифицированных в протеомных исследованиях (Shi et al., 2021; Ginno et al., 2018; Alvarez et al., 2023). (Б) Кумулятивная доля белков, идентифицированных как минимум в  $x$  МС-исследованиях хроматина, где  $x$  принимает значения от 1 до 7; за 100 % принимается совокупность белков, выявленных хотя бы в одном МС-исследовании.

Наиболее крупными по числу идентифицированных белков хроматина являются следующие наборы белков: Shi et al., 2021; Ginno et al., 2018; Alabert et al., 2014; Alvarez et al., 2023. По размеру они сопоставимы с эталонным набором белков хроматина человека SimChrom (~3000 белков), однако доля общих белков в них невысока (25-35 %) (**Рисунок 18А,Б**). То же верно и для набора белков хроматина (Torrente et al., 2011), содержащего вдвое меньшее число белков. Более высокая согласованность экспериментальных данных с эталонным набором белков хроматина SimChrom (50-60 %) выявлена для наборов (Itzhak et al., 2016) и (Kustatscher et al., 2014). В этих исследованиях применяли дополнительные методы отбора преимущественно ядерных и хроматиновых белков с использованием классификаторов машинного обучения и литературных данных (**Рисунок 18Б**).



**Рисунок 18** — Анализ хроматомов и нуклеома из МС-исследований. **(А)** Общее количество уникальных белков, идентифицированных в исследованиях. **(Б, В, Г)** Пересечение белковых наборов с белками хроматина SimChrom (Б), белками ядра согласно NULOC\_JT\_NECF (В) и NULOC\_CS (Г).

В результате анализа представленности конститутивных белков в МС-исследованиях хроматина выявлено, что к ним относится 80 % идентифицированных белков, в то время как ожидаемая доля таких белков составляет 62 % (**Рисунок 19**).

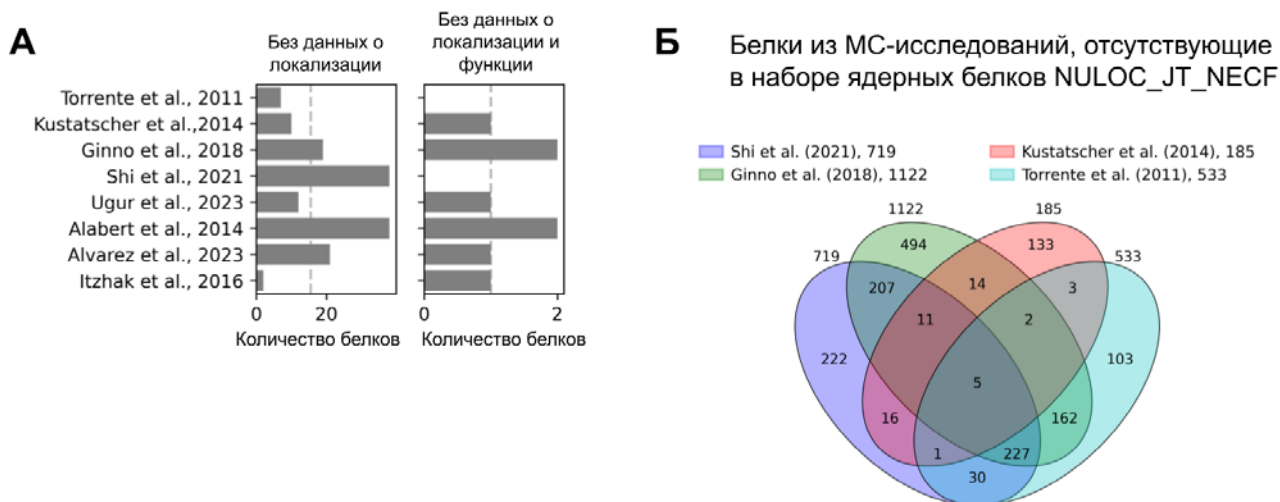


**Рисунок 19** — Конститутивные белки в хроматомах, идентифицированных методами протеомики. **(А)** Диаграмма Венна пересечений белков хроматина из SimChrom, белков хроматина из МС-исследований и конститутивных белков согласно НРА. **(Б)** Доля конститутивных белков в хроматомах из МС-исследований.

### **3.3.2. Анализ белков из протеомных исследований хроматина, не имеющих аннотаций о локализации в ядре**

Выявлено, что от 10 до 38 % белков, идентифицированных в МС-исследованиях хроматина, отсутствуют в наборе белков NULOC\_JT\_NECF, то есть не имеют аннотации о ядерной локализации ни в одной из баз данных (UniProt или HPA) (**Рисунок 18В**). Это может свидетельствовать как о потенциальной контаминации образцов белками из других клеточных компартментов (цитоплазма, эндомембранная система), так и о идентификации потенциально новых белков ядра, не выявленных в других исследованиях.

Общее число белков, отсутствующих в наборе NULOC\_JT\_NECF, составило 2232. При этом только для 94 из них в базах данных UniProt и HPA отсутствовали какие-либо сведения о клеточной локализации (**Рисунок 20А**). Из 2232 белков только пять были идентифицированы в четырех наиболее многочисленных по количеству белков исследованиях интерфазного хроматина (**Рисунок 20Б**). К ним относятся FLNB (Filamin B, actin-binding protein), GMPS (Guanine Monophosphate Synthase), CHERP (Calcium Homeostasis Endoplasmic Reticulum Protein), ILKAP (ILK Associated Serine/Threonine Phosphatase) и PLEC (Plectin). Несмотря на то что в UniProt и HPA эти белки аннотированы преимущественно как цитоплазматические, в литературе присутствуют экспериментальные свидетельства об их потенциальной локализации в клеточном ядре [196].



**Рисунок 20** — Анализ белков хроматина, идентифицированных методами протеомики, для которых отсутствуют сведения о локализации в ядре клетки по данным UniProt и HPA. **(А)** Белки из МС-исследований без аннотаций о локализации в клетке по данным UniProt и HPA (слева) и без аннотаций в GO (справа). **(Б)** Диаграмма Венна пересечения списков белков из МС-исследований хроматина, которые отсутствуют в наборе белков ядра NULOC\_JT\_NECF.

Оценка контаминации фракций белков хроматина нецелевыми белками затруднена из-за отсутствия эталонных наборов белков других компартментов и неполной аннотации функциональных белков ядра. Тем не менее, для 2232 белков хроматина из МС-исследований, отсутствующих в наборе NULOC\_JT\_NECF, в результате анализа обогащения терминами GO выявлены категории, связанные с неядерными органеллами клетки, клеточным метаболизмом, а также трансляцией и созреванием белков. В частности, 2025 белков (90 %) относились к термину «цитоплазма», 115 белков — к «транспорт в везикулах аппарата Гольджи». При этом не выявлено обогащения терминами, связанными с митохондриями. Для 30 белков выявлено обогащение терминами GO, связанными с трансляцией («трансляция», «активность факторов инициации трансляции»), в то же время в литературе есть свидетельства функциональной активности некоторых белков клеточного аппарата трансляции в ядре клетки [197]. Аналогично, аппарат Гольджи связан с ядром через эндоплазматический ретикулум и систему везикулярного транспорта.

Дополнительно для анализа клеточной локализации по данным литературы из 195 белков, выявленных в пяти из семи МС-исследований хроматина и отсутствующих в наборе NULOC\_JT\_NECF, случайным образом было отобрано 20 белков. Для пяти из них в литературе были представлены экспериментальные свидетельства локализации в клеточном ядре: CALR (Calreticulin) [198], PDIA4 (Protein disulfide-isomerase A4) [199], ABCF2 (ATP-binding cassette sub-family F member 2) [200], SEC23B (Protein transport protein Sec23B) [201], EIF3D (Eukaryotic translation initiation factor 3 subunit D) [202]. Полученный результат свидетельствует о прогностической силе протеомных подходов для выявления новых белков хроматина.

Таким образом, протеомные исследования являются независимым источником информации о белковом составе хроматина. При этом до 35 % белков, идентифицированных в протеомных исследованиях, не имеют аннотаций о локализации в ядре клетки по данным UniProt и HPA. Вместе с тем до 25 % белков без аннотаций о ядерной локализации, идентифицированных по крайней мере в пяти из семи МС-исследований, могут быть локализованы в ядре клетки на основании литературных данных. Проведенный анализ подтвердил литературные данные о загрязнении фракций белков хроматина белками цитоплазмы. Было показано, что дополнительные этапы обработки данных (фильтрация белков на основе сравнения интенсивностей МС-сигналов или использование классификатора машинного обучения, обученного, в том числе, на данных из литературы) существенно увеличивают долю белков с аннотациями локализации в ядре или функциями, ассоциированными с хроматином. В то же время, избыточная фильтрация данных может приводить к снижению вероятности идентификации не описанных ранее белков хроматина.

### 3.3.3. Анализ белков хроматина, не идентифицированных в протеомных экспериментах

Был проведен сравнительный анализ белков из МС-исследований хроматина и белков из категорий классификации SimChrom. С одной стороны, такой анализ позволил выявить ограничения протеомных методов в идентификации белков хроматина, а с другой — являлся проверкой наличия основных категорий, ассоциированных с хроматином, в онтологии SimChrom.

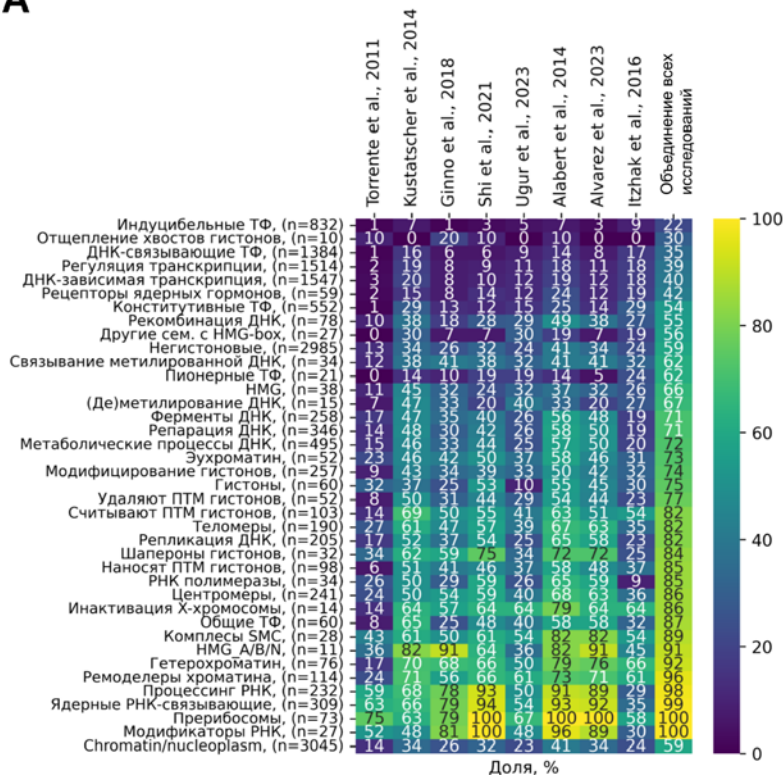
Из набора SimChrom 1246 белков (41 %) не были идентифицированы в протеомных исследованиях хроматина (**Рисунок 21**). Большинство из них (839 белков, 67 %) относятся к индуцибельным белками по аннотации НРА, то есть их уровень экспрессии мог отсутствовать или быть низким в исследуемых клеточных линиях. Наибольшее число неидентифицированных белков (1148) приходится на категорию «ДНК-связывающие транскрипционные факторы», из которых только 394 белка являются конститутивными.

Во всех МС-исследованиях хроматина доля транскрипционных факторов была низкой: доля конститутивных ТФ не превышала 30 % от эталонного списка белков, а индуцибельных — 10 % (**Рисунок 21А**). При объединении всех идентифицированных в МС белков, доля конститутивных и индуцибельных ТФ составила 54 % и 22 % соответственно (**Рисунок 21А**). Полученные результаты можно объяснить низким уровнем экспрессии ТФ, а также их слабым или транзистентным характером взаимодействия с геномом, в результате чего такие белки могут вымываться из целевой фракции на этапах пробоподготовки хроматина.

К другим категориям белков хроматина, состав которых в МС-исследованиях варьировал, предположительно, в зависимости от способа пробоподготовки, относятся субъединицы РНК-полимераз и гистоны. Так, в работе (Ginno et al., 2018) идентифицированная доля белков из одноименных категорий SimChrom составила 29 % и 25 % соответственно; тогда как в работе (Shi et al., 2021) — 59 % и 53 % соответственно (**Рисунок 21А**).

Гистоны (60 белков человека) относятся к группе белков, наиболее консервативных по аминокислотной последовательности и наиболее представленных по числу копий в ядре клетки. Поэтому был проведен более детальный анализ идентифицированных в МС-исследованиях хроматина гистонов. Закономерно, тканеспецифичные варианты гистонов (например, из тканей половой системы) не были обнаружены ни в одном из МС-исследований. Однако и канонические гистоны, которые экспрессируются во всех типах клеток, были идентифицированы не в полном составе, особенно гистоны типа H2B (Рисунок 21Б).

А



Б



**Рисунок 21** — Анализ белков хроматина из набора SimChrom, идентифицированных в МС-исследованиях хроматина. (А) Доля белков из категорий SimChrom, выявленная в МС-исследованиях хроматина. (Б) Количество генов гистонов, белковые продукты которых были выявлены в МС-исследованиях хроматина.

Так, из 14 канонических изоформ гистонов типа H2B в работе (Shi et al., 2021) было идентифицировано шесть белков (соответствующих продуктам генов *H2BC1*, *H2BC3*, *H2BC4*, *H2BC13*, *H2BC18*, *H2BC26*), тогда как в работе (Itzhak et al., 2016)

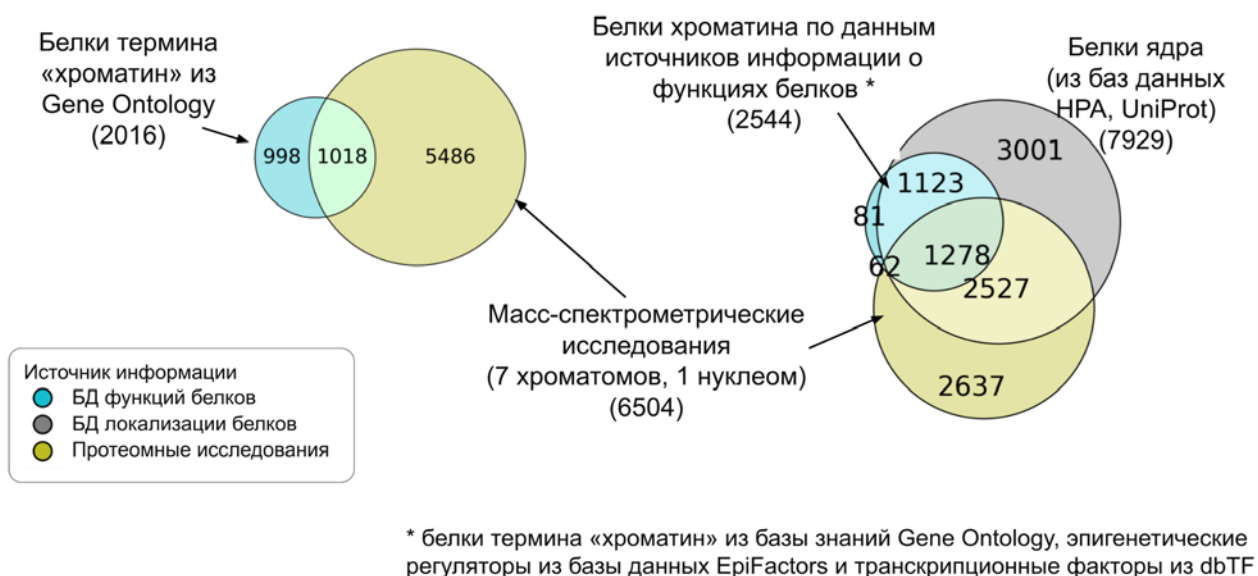
— четыре белка (соответствующие продуктам генов *H2BC4*, *H2BC11*, *H2BC12*, *H2BC13*) (**Рисунок 21Б**). Несмотря на то, что канонические гистоны одного типа, предположительно, коэкспрессируются в клетке, расхождения в идентификации продуктов отдельных генов могли быть обусловлены как различиями в чувствительности МС-методов, так и вариациями экспрессии гистонов в различных клеточных линиях.

Таким образом, выявлены категории белков хроматина с низким уровнем обнаружения в МС-исследованиях. Это может быть связано с их низкой экспрессией, транзистентным характером взаимодействия с другими компонентами хроматина, а также вымыванием слабосвязанных белков при пробоподготовке.

### **3.4. Сравнение репертуара белков хроматина по данным разнородных источников информации**

Для более комплексного понимания преимуществ и ограничений описанных выше источников информации о белках ядра и хроматина (баз данных о локализации белков, баз данных о функциях белков и протеомных исследований) был проведен сравнительный анализ всех аннотированных в них белков.

Выявлена частичная согласованность наборов белков между разнородными источниками (**Рисунок 22**). Так, для 500 белков (25 %) из категории GO «хроматин» отсутствовали сведения о ядерной локализации согласно данным НРА, а 998 белков (50 %) не были идентифицированы в МС-исследованиях хроматина (**Рисунок 22А**). Примечательно, что 2699 белков (42 %), идентифицированных в МС-исследованиях хроматина и ядра, не имеют сведений о локализации в ядре ни в UniProt, ни в НРА, а для 254 белков и вовсе отсутствуют сведения о клеточной локализации (**Рисунок 22Б**).



**Рисунок 22** — Сравнение репертуара белков хроматина и ядра по данным различных источников информации. Слева: пересечение белков категории «хроматин» (Gene Ontology) с белками, идентифицированными в масс-спектрометрических исследованиях хроматина и ядра клетки. Справа: пересечение трех групп белков: (i) белки категории «хроматин» (Gene Ontology), белки, регулирующие эпигенетические процессы (EpiFactors), транскрипционные факторы (dbTF); (ii) белки с локализацией в ядре (UniProt и HPA); (iii) белки, идентифицированные в масс-спектрометрических исследованиях хроматина и ядра клетки.

Таким образом, в результате сравнительного анализа разнородных источников информации выявлена ограниченная согласованность данных о репертуаре белков хроматина и ядра клетки человека.

### 3.5. Обсуждение результатов

В данной работе впервые проведен интегративный мета-анализ репертуара белков хроматина и ядра клетки человека по данным разнородных источников информации, включающих функциональные аннотации, данные клеточной локализации и результаты протеомных исследований. Полученные результаты свидетельствуют о том, что различные источники информации о белках хроматина и ядра клетки человека предоставляют частично пересекающиеся, но не полностью согласованные наборы белков. Эти различия обусловлены особенностями

используемых экспериментальных подходов, принципов аннотации и критериев отнесения белков к функциональным или локализационным категориям.

Было выявлено, что существующие функциональные аннотации белков хроматина существенно различаются по степени детализации категорий и принципам организации онтологий. В результате существующих специализированных ресурсов недостаточно для описания процессов, протекающих на нуклеосомном уровне организации хроматина, а более универсальные ресурсы содержат избыточное количество разрозненных терминов.

В свою очередь, ресурсы о клеточной локализации белков позволяют оценить распределение белков между клеточными и субъядерными компартментами, а также определить верхнюю границу числа белков хроматина и нижнюю границу количества ядерных белков. Однако их использование также сопряжено с рядом ограничений, к которым относятся частичная аннотация протеома человека, сложности интерпретации множественной локализации белков и различия в иерархиях субъядерных компартментов.

Протеомные исследования обогащенных фракций хроматина предоставляют дополнительную информацию о составе хроматома и позволяют идентифицировать ранее не аннотированные другими методами белки. Вместе с тем такие подходы зависят от используемых протоколов выделения хроматина, чувствительности масс-спектрометрического анализа и биологических особенностей исследуемых клеток.

Дополнительную сложность для исследований представляет изменчивый белковый состав хроматина, который зависит от функционального состояния хроматина, типа клеток, стадии клеточного цикла [160]. Многие белки способны одновременно участвовать в нескольких биологических процессах, перемещаться между компартментами клетки и формировать транзистентные взаимодействия с геномом. В результате использование единственного критерия — функции, локализации или идентификации протеомными методами — оказывается недостаточным для описания репертуара белков хроматина. Таким образом, полученные результаты подчеркивают необходимость объединения разнородных

данных о белках хроматина и ядра клетки человека в рамках единой функционально-ориентированной классификационной системы.

## **Глава 4. Эмпирическая функционально-ориентированная классификация белков хроматина SimChrom и эталонные наборы белков ядра**

С учетом преимуществ и недостатков существующих источников информации о белках хроматина, для получения целостного представления об их репертуаре была разработана иерархическая структура классификации (онтология) белков хроматина и соответствующий набор белков (непосредственно классификация). Основным требованием к классификации было сочетание простоты использования и полноты данных. Кроме того, были разработаны эталонные наборы белков ядра и цитоплазмы клеток человека с учетом различной степени согласованности данных и достоверности информации.

### **4.1. Эмпирическая функционально-ориентированная классификация белков хроматина SimChrom**

Согласно проведенному литературному анализу, к белкам хроматина в широком смысле можно отнести все белки, способные напрямую или опосредованно взаимодействовать с геномной ДНК и участвовать в процессах, связанных с организацией и функционированием хроматина, за исключением белков, прочно связанных с ядерной оболочкой. Так как на основании доступных данных невозможно явно разделить белки хроматина и нуклеоплазмы, разработанная структура классификации SimChrom (онтология категорий) включает в себя категории обеих групп.

В основе классификации SimChrom лежит иерархическая онтология, которая объединяет категории белков по четырем аспектам: (1) молекулярная функция; (2) биологический процесс; (3) геномная локализация, которая косвенно связана с выполняемыми функциями; (4) физико-химические свойства. Использование нескольких аспектов позволяет формализовать разнородные характеристики белков и учесть их мультифункциональность.

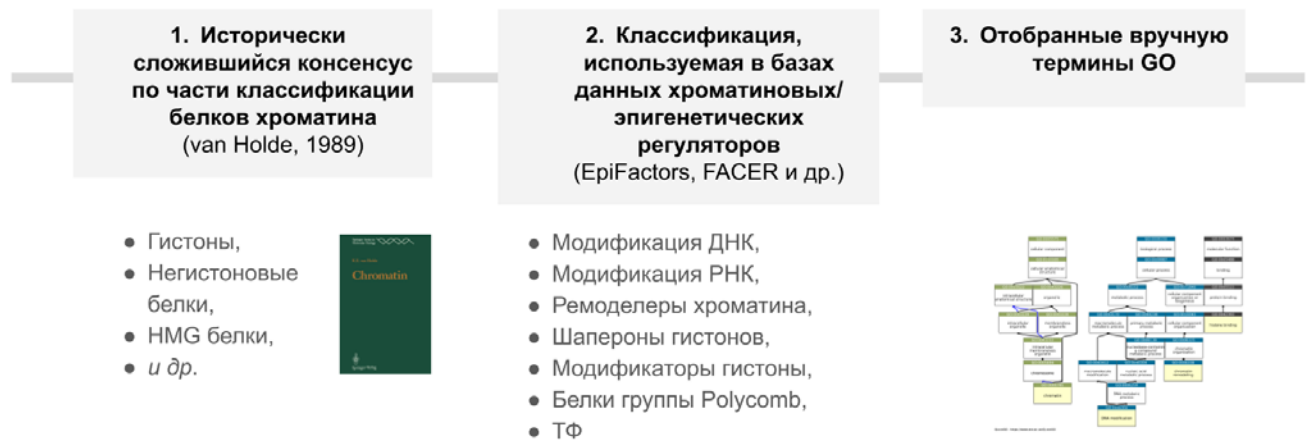
В классификацию были включены традиционно выделяемые группы белков (гистоновые, негистоновые, белки HMG [136]); белки, регулирующие эпигенетические процессы (согласно базе данных EpiFactors), а также категории белков, ассоциированные с хроматином из литературы и базы знаний Gene Ontology (поиск по ключевым словам ДНК, гистон, нуклеосома, хроматин и другие). Анализ белков, идентифицированных методами масс-спектрометрии, а также данные о важной регуляторной роли некодирующих РНК [185] послужили основанием для добавления РНК-связывающих белков в классификацию.

В категории классификации были добавлены белки из терминов базы знаний Gene Ontology на основе вручную установленных соответствий между категориями SimChrom и терминами GO, перечисленными в **Таблице A1 (Приложение А)**, а также белки из специализированных баз данных (The Human Transcription Factors, HISTome2, HistoneDB) и научной литературы (**Рисунок 23**). Дополнительно была проведена ручная проверка и фильтрация данных, в том числе путем сравнительного анализа с результатами МС-исследований и данными о локализации белков в клетке.

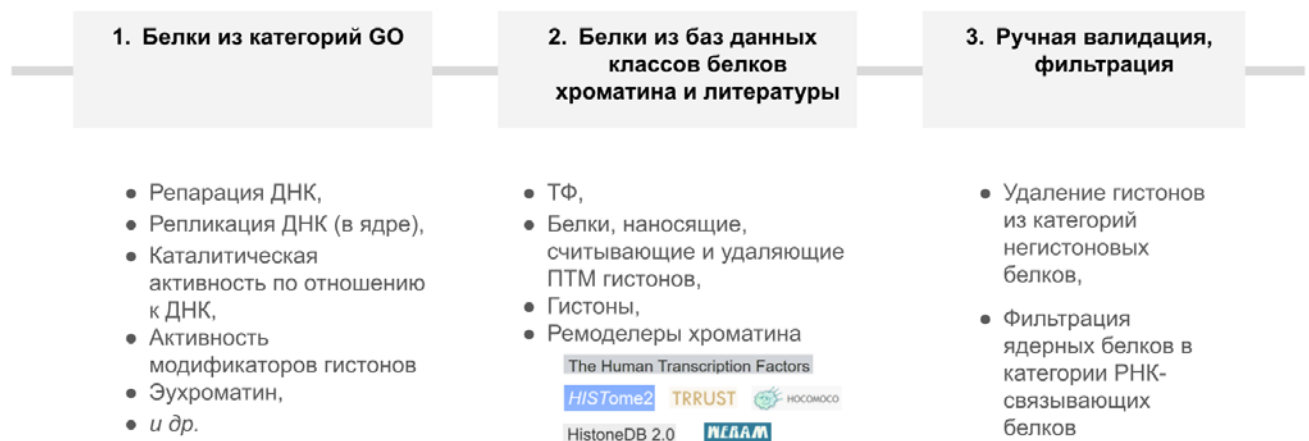
В классификации SimChrom белок может быть отнесен к нескольким категориям, что обусловлено разнородностью аспектов категорий и мультифункциональностью белков. Для проведения последующего анализа на основе описанной выше онтологии была разработана классификация SimChrom-SL, в которой каждый белок отнесен только к одной категории. Выбор категории для каждого белка осуществлялся с учетом специфичности категорий (их свойств и количества белков в них).

В итоге разработанная классификация состоит из 39 категорий (гистоновые белки и 36 категорий негистоновых белков), к которым отнесено 3045 белков хроматина человека. Некоторые категории классификации подразделены на подкатегории, максимальная глубина иерархии не превышает пяти уровней (**Рисунок 24**).

**А** Для создания онтологии **SimChrom** были использованы:



**Б** Для создания набора белков **SimChrom** были использованы:



**Рисунок 23** — Схема разработки эмпирической функционально-ориентированной классификации белков хроматина SimChrom, (А) формирование онтологии классификации, (Б) формирование эталонного набора белков хроматина человека.



**Рисунок 24** — Разработанная эмпирическая функционально-ориентированная классификация белков хроматина SimChrom. Для каждой категории справа приведено количество белков, слева обозначена принадлежность к аспекту (молекулярная функция, биологический процесс, геномное положение); пиктограммы указывают на наличие или отсутствие аналогичных категорий в онтологиях других баз данных (Gene Ontology, EpiFactors и другие). Интерактивная версия рисунка доступна по адресу: <https://simchrom.intbio.org/#classification/>.

Наиболее крупной по количеству белков категорией классификации SimChrom является «ДНК-зависимая транскрипция» (1547 белков). В ее состав входят: субъединицы РНК-полимераз I, II и III типов (34 белка), а также 1514 белков, участвующих в регуляции транскрипции. Среди них:

- «Общие ТФ» (60 белков) — неспецифические транскрипционные факторы, необходимые для сборки преинициаторного комплекса на промоторах и рекрутирования РНК-полимеразы;
- «ДНК-связывающие ТФ» (1384 белка) — специфичные к последовательности ТФ и активирующие или репрессирующие транскрипцию; белки данной группы были дополнительно разделены на конститутивные (552) и индуцибельные (832);

- «Пионерные ТФ» (21 белок) — группа белков, способных связываться с нуклеосомной ДНК и рекрутировать другие белки для изменения функционального состояния хроматина;
- «Рецепторы ядерных гормонов» (59 белков) – группа лиганд-активируемых транскрипционных факторов, среди которых рецепторы стероидных и тиреоидных гормонов, андрогенов и эстрогенов (**Рисунок 24**).

К категориям белков, регулирующих эпигенетические процессы, отнесены:

- «Ремоделеры хроматина» (114 белков) — компоненты комплексов ремоделирования хроматина, которые осуществляют АТФ-зависимые перестройки хроматина (семейства SWI/SNF, ISWI, CHD и INO80);
- «Шапероны гистонов» (32 белка) — белки, транспортирующие и депонирующие гистоны, а также обеспечивающие их обмен в нуклеосомах без затрат энергии АТФ;
- «Модификаторы гистонов» (257 белков) — белки, катализирующие нанесение ПТМ гистонов (сукцинирование, допаминилирование, ацетилирование, серотонилирование, пропионалирование, фосфорилирование, метилирование, цитрунирование, бутирилирование, кротонилирование, лактилирование, глутарилирование и убиквитинилирование) (98); гидролазы, удаляющие ПТМ гистонов (52), а также разнородные белки, неканонической функцией которых является отщепление хвостов гистонов (10);
- «Связывание ПТМ гистонов» (103 белка) — белки-эффекторы, узнающие преимущественно метилированные, ацетилированные, фосфорилированные, убиквитинированные гистоны посредством доменов Chromodomain, Tudor, MBT, PHD и других;
- Ферменты метилирования и деметилирования ДНК (15 белков) – группа, к которой отнесены *de novo* ДНК-метилтрансфераза (DNMT3); метилтрансферазы, поддерживающие уровень метилирования при репликации (DNMT1), диоксигеназы семейства TET (TET1, TET2 и TET3) окисляющие метилцитозин;

- «Связывание метилированной ДНК» (34 белка) — группа, к которой отнесены, например, белки с доменом MBD (methyl DNA binding domain) и цинковыми пальцами (Kaiso-подобные белки и другие);
- «Модификаторы РНК» (27 белков) — белки, катализирующие нанесение и удаление химических модификаций (метилирование, ацетилирование) мРНК, тРНК, рРНК, малые ядерные РНК.

В состав категории «Ферменты, действующие на ДНК» (258 белков) вместе с категорией «(Де)метилирование ДНК» входят белки, участвующие в репликации и репарации (ДНК-полимеразы, ДНК-лигазы, ДНК-нуклеазы, ДНК-фотолиазы, апуриновые/апиримидиновые эндонуклеазы, фосфодиэстеразы, ДНК-гликозилазы); а также ДНК-топоизомеразы, осуществляющие топологические изменения ДНК.

В классификации SimChrom отдельно выделены категории, описывающие «Метаболические процессы ДНК» (495 белков): «Репарация ДНК» (346 белков), «Репликация ДНК» (205 белков) и «Рекомбинация ДНК» (78 белков) (**Рисунок 24**). К другим категориям биологических процессов относится «Инактивация X-хромосомы» (14 белков), в которой белки, участвующие в дозовой компенсации.

На основании физико-химических свойств в отдельную группу были выделены белки суперсемейства HMG (семейства A, B, N), характеризующиеся высокой электрофоретической подвижностью. Вместе с другими белками, большинство из которых относится к транскрипционным факторам и содержит домен HMG\_box, они формируют сборную категорию HMG (38 белков).

В состав категории «SMC-комплексы» (28 белков) входят АТФ-зависимые комплексы, осуществляющие экструдию ДНК, включая когезин, осуществляющий когезию сестринских хроматид (гетеродимер SMC1-SMC3), конденсин, участвующий в конденсации хроматина (гетеродимер SMC2-SMC4), а также комплексы с гетеродимером SMC5-SMC6, вовлеченные в репарацию ДНК.

К сборной группе архитектурных белков хроматина были отнесены белки суперсемейства HMG, белки SMC-комплексов, линкерные гистоны H1 и транскрипционный фактор CTCF.

В классификации SimChrom категория «Ядерные РНК-связывающие белки» объединяет белки, ассоциированные с прерибосомами (73 белка), а также белки, участвующие в метаболизме РНК: в созревании — категория «Процессинг РНК» (205 белков) и в химических модификациях — категория «Модификация РНК» (27 белков) (**Рисунок 24**). В данную группу были включены только белки с экспериментально подтвержденными функцией и ядерной локализацией [150,151].

На основе геномного положения выделены категории «Центромерные» (241 белок), «Теломерные» (190 белков), «Гетерохроматин» (76), «Эухроматин» (52). Однако состав этих категорий требует дополнительного уточнения на основе экспериментальных данных.

Стоит отметить, что классификация включает в себя не только категории, представленные в существующих ресурсах (Gene Ontology, EpiFactors), но и категории, отсутствующие в других базах данных (например, пионерные ТФ; белки, отщепляющие хвосты гистонов; белки SMC-комплексов) (**Рисунок 24**).

Таким образом, SimChrom представляет собой эмпирическую функционально-ориентированную классификацию белков хроматина человека, объединяющую сведения о функциях и участии белков в процессах, связанных с организацией и функционированием хроматина.

## 4.2. Эталонные наборы белков ядра и цитоплазмы

Из-за несогласованности данных о локализации ядерных белков в различных ресурсах, выявленной в **Главе 3**, были сформированы эталонные наборы белков ядра и цитоплазмы. Эти наборы различаются степенью согласованности данных, достоверностью информации и учетом множественной локализации белков. Список наборов белков и их определения приведены в **Таблице 3 (раздел 2.3)**. Так, NULOC\_CS — набор из 3296 ядерных белков, согласно аннотациям с высоким уровнем достоверности информации из UniProt и HPA и отсутствием противоречий с данными из OpenCell; NULOC\_JT\_NECF — наиболее полный набор из 8912 ядерных белков, включающий все белки, аннотированные как ядерные хотя бы в

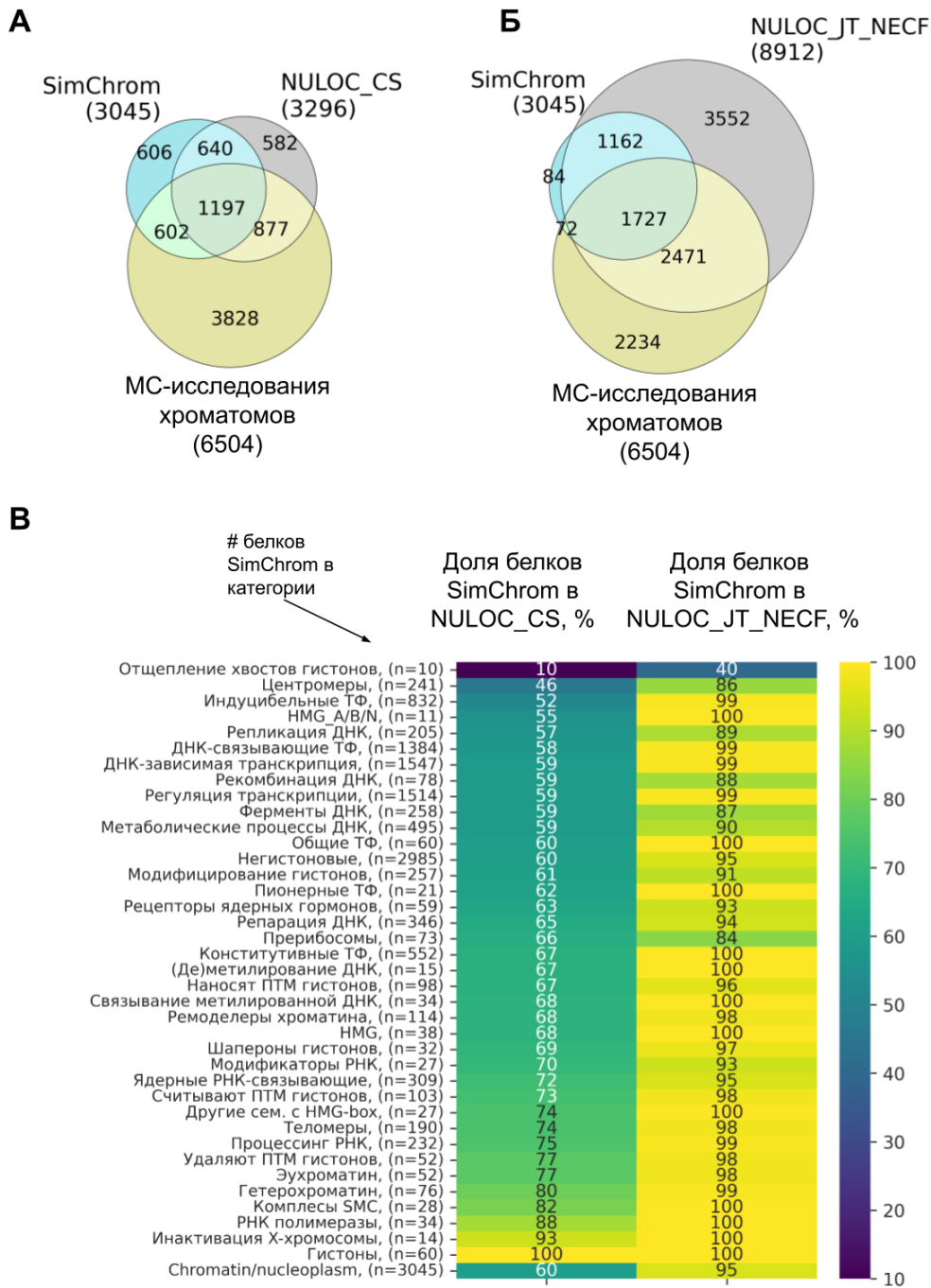
одном из ресурсов (UniProt, HPA или OpenCell), независимо от уровня достоверности информации.

#### **4.3. Верификация набора белков хроматина человека SimChrom**

Для оценки качества набора белков хроматина человека SimChrom был проведен его сравнительный анализ с наборами ядерных белков NULOC\_CS и NULOC\_JT\_NECF, а также с белками хроматина, идентифицированными в МС-исследованиях.

Из 3045 белков набора SimChrom для 1837 доступна достоверная информация о ядерной локализации (входят в набор NULOC\_CS), причем 1197 белков также идентифицированы хотя бы в одном МС-исследовании хроматина (**Рисунок 25А**). Эти 1197 белков можно рассматривать как наиболее достоверный набор белков хроматина человека.

Выявлено, что 1208 белков набора SimChrom отсутствуют в наборе NULOC\_CS (**Рисунок 25А**). Учитывая низкую долю белков референсного протеома человека с достоверными аннотациями одновременно в UniProt и HPA (44 %), полученный результат является ожидаемым.



**Рисунок 25** — Сравнение репертуара белков набора SimChrom с наборами ядерных белков NULOC\_CS и NULOC\_JT\_NECF. (А, Б) Диаграмма Венна пересечения белков SimChrom, белков, идентифицированных в MS-исследованиях хроматомов и белков из набора NULOC\_CS (А) и NULOC\_JT\_NECF (Б). (В) Доля белков категорий SimChrom в наборах NULOC\_CS и NULOC\_JT\_NECF.

В качестве дополнительной независимой проверки принадлежности белков SimChrom к ядерному компартменту клетки был проведен анализ обогащения терминами GO для 1208 белков SimChrom, отсутствующих в NULOC\_CS, с

дальнейшим отбором терминов, не ассоциированных с ядром клетки. Выявлено сравнительно небольшое число белков, входящих в состав обогащенных терминов, не связанных с ядром. Большинство из них относились к белкам, ассоциированным с центромерой: 12 из 156 белков данной категории по GO относились к мультивезукулярным тельцам и комплексу динактина.

С другой стороны, выявлено 1459 ядерных белков из NULOC\_CS, отсутствующих в SimChrom; из них 877 белков идентифицированы в МС-исследованиях хроматомов. Эти белки требуют дальнейшего анализа для уточнения их возможной принадлежности к хроматину (**Рисунок 25А**). В результате анализа обогащения терминами GO этих белков не были выявлены функционально значимые категории хроматина, которые отсутствовали бы в онтологии SimChrom, что дополнительно свидетельствует о ее полноте.

Подавляющее число белков хроматина SimChrom, как и ожидалось, присутствовало в наборе белков ядра NULOC\_JT\_NECF, составленном с наименее строгими критериями (**Рисунок 25Б**). Большинство из 156 белков SimChrom, отсутствующих в NULOC\_JT\_NECF, относятся к категориям «Ферменты, действующие на ДНК» и «Белки, ассоциированные с центромерой». При этом для 32 белков отсутствуют сведения о клеточной локализации, а 27 белков (большинство из которых — центромерные белки SimChrom) аннотированы как белки хромосомы согласно данным UniProt. Среди оставшихся 124 белков путем ручного анализа выявлены белки как с функциями в клеточном ядре (ацетил- и метилтрансферазы гистонов), так и потенциально неядерные белки (рибосомальные белки, различные киназы и белки, участвующие в процессинге митохондриальной ДНК).

Наибольшая доля белков без аннотации о ядерной локализации относится к категории «Отщепление хвостов гистонов»: только 4 из 10 белков этой категории присутствуют в наборе NULOC\_JT\_NECF (**Рисунок 25Б**). Это согласуется с тем фактом, что для многих ферментов, осуществляющих отщепление хвостов гистонов (в частности, металлопротеиназ, катепсинов, эластаз нейтрофилов),

данная функция не является основной и реализуется только в определенных условиях или на определенных стадиях развития клетки [148].

Валидацию классификации SimChrom провели путем ее применения для функциональной аннотации негистоновых белков хроматина, представленных в структурах нуклеосом и их комплексов из базы данных NucleosomeDB [203]. Для описания порядка 600 структур (по состоянию на май 2026 г.) в текущую версию классификации были дополнительно включены следующие категории: белки комплекса кинетохор, cGAS (cyclic GMP-AMP Synthase, синтаза циклического ГМФ-АМФ), а также, с учетом неспецифичности базы данных по отношению к белкам человека, интегразы вирусов и компоненты системы CRISPR/Cas.

Таким образом, несмотря на то что ряд категорий разработанной эмпирической функционально-ориентированной классификации SimChrom требует дальнейшего более тщательного уточнения состава белков и фильтрации ложноположительных аннотаций с использованием данных о локализации белков и доменном составе, классификация SimChrom охватывает наиболее полный репертуар белков хроматина, что обосновывает её использование в биоинформатических исследованиях.

#### 4.4. База данных SimChrom

Для систематизации информации о белках хроматина, полученной в ходе данного исследования, а также поддержки дальнейших исследований белков хроматина была разработана база данных SimChrom. Записи в базе данных являются белок-центричными (на основе референсного протеома человека согласно определению UniProtKB, версия 2022\_2).

Для каждого белка протеома человека в базе представлены следующие сведения:

- идентификатор гена (основной и альтернативные HGNC ID с гиперссылкой на страницу гена на ресурсе HGNC);

- идентификатор белка (UniProt ID с гиперссылкой на страницу белка на сайте UniProtKB/Swiss-Prot);
- название белка (UniProtKB Protein name);
- аннотации субклеточной локализации из баз данных UniProt, HPA и OpenCell (как с фильтрацией по критериям достоверности информации, так и без нее);
- категории белка согласно эмпирической функционально-ориентированной классификации SimChrom, а также единственная категория по классификации SimChrom-SL;
- представленность белка (на основе данных набора PaxDb\_PA);
- протеомные исследования хроматина, в которых этот белок был идентифицирован;
- состав доменов согласно аннотации Pfam;
- длина белка (количество аминокислотных остатков);
- мультисубъединичные комплексы в состав которых входит белок (по данным базы данных ComplexPortal).

База данных реализована в виде интерактивного веб-ресурса (<https://simchrom.intbio.org/>), представляющего собой serverless-приложение на платформе GitHub Pages.

Раздел "Localization" (локализация) содержит интерактивный графический интерфейс для сравнения онтологий и репертуара ядерных белков по данным UniProt, HPA, OpenCell, позволяющий с использованием логических операций пересекать и объединять наборы белков из различных терминов и баз данных.

Раздел "SimChrom classification" содержит эмпирическую функционально ориентированную классификацию с интерактивными инструментами отбора белков.

Раздел "Domain composition" (состав доменов) позволяет выполнять отбор белков на основе функциональных аннотаций (согласно SimChrom-SL) и состава доменов согласно аннотации Pfam (Pfam-домены).

Раздел "Domain co-occurrence" (парная ко-встречаемость доменов) содержит инструменты отбора белков с парными комбинациями Pfam-доменов, размеченных

по функциональным группам. Для пар доменных моделей Pfam рассчитаны условные вероятности их ко-встречаемости в белках хроматина и указано количество соответствующих белков.

Раздел "Download reference datasets" (загрузка референсных наборов белков) позволяет скачать как эталонные наборы белков ядра, разработанные в данном исследовании, так и унифицированные списки идентификаторов белков UniProt, идентифицированных в протеомных исследованиях хроматина.

Раздел "Novel structural domains" содержит интерактивную таблицу с характеристиками TED-доменов, присутствующих в белках хроматина, но не имеющих структурных соответствий в базах данных PDB и CATH, а также функциональных аннотаций в ресурсе InterPro.

#### **4.5. Обсуждение результатов**

В связи с отсутствием устоявшегося терминологического определения репертуара белков хроматина нами была разработана функционально-ориентированная классификация SimChrom с применением эмпирического подхода. Это позволило на основе анализа данных из баз данных, литературы, протеомных исследований хроматина и исследований субклеточной локализации белков объединить основные классы белков хроматина с учетом их молекулярных функций, биологических процессов, геномной локализации и физико-химических свойств. Разработанный подход позволил учесть как основные структурные компоненты хроматина, так и динамически рекрутируемые белки, участвующие в регуляции геномных процессов.

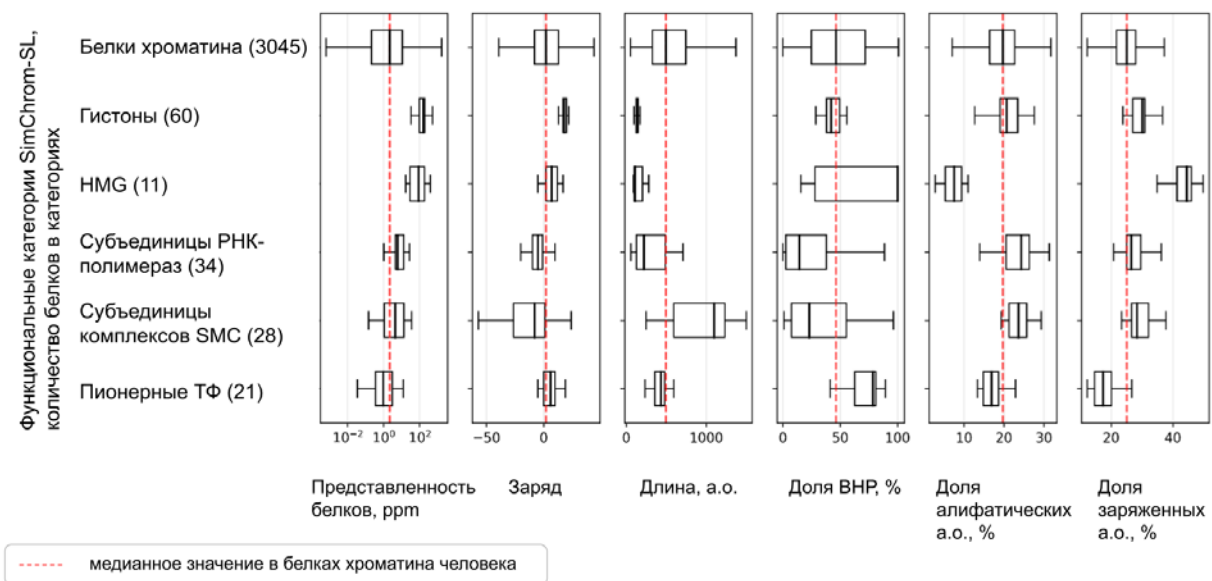
Основным источником аннотаций белков для включения в классификацию служила база знаний Gene Ontology. Поэтому даже при тщательном отборе сопоставляемых категорий между Gene Ontology и SimChrom остается возможность включения ложноположительных аннотаций вследствие систематических ошибок и неполноты данных Gene Ontology. Кроме того, для ряда категорий, особенно основанных на геномной локализации белков, необходимо

дальнейшее уточнение состава белков на основе экспериментальных данных. Несмотря на то, что немембранные органеллы ядра не были выделены в отдельные категории SimChrom, белки этих органелл могли быть включены в классификацию на основании их молекулярных функций или участия в соответствующих биологических процессах.

Таким образом, разработанная классификация SimChrom создает основу для интегративного анализ белков хроматина человека и исследования их функций, структурных особенностей и физико-химических свойств.

## Глава 5. Структурно-функциональные свойства белков хроматина человека

В рамках разработанной эмпирической функционально-ориентированной классификации белков хроматина человека SimChrom каждому белку сопоставлены данные о его субклеточной локализации, представленности в клетке (относительное количество молекул), физико-химических свойствах (длина, доля внутренне неупорядоченных регионов, заряд, аминокислотный состав) и доменной организации. Сравнение этих параметров между категориями классификации позволило выявить структурно-функциональные паттерны, характерные для категорий белков хроматина, что демонстрирует эвристический потенциал классификации (**Рисунок 26**).



**Рисунок 26** — Свойства белков хроматина из категорий классификации SimChrom-SL (на примере некоторых свойств и функциональных групп). Длина «усов» соответствует 1,5 межквартильному размаху.

Например, пионерные ТФ, в отличие от других белков хроматина, характеризуются меньшей длиной (медиана 321 а.о. против 496 а.о.) и повышенной долей внутренне неупорядоченных регионов (медиана 73 % против 47 %) (**Рисунок 26**), что объясняет их способность связываться с плотно упакованным хроматином

и низкую детектируемость в масс-спектрометрических экспериментах (не более 25 % белков категории). Архитектурные белки хроматина из суперсемейства HMG обладают небольшой длиной (медиана 109 а.о.) и являются наиболее неупорядоченными (медиана 99 %) (**Рисунок 26**), что обуславливает их гибкость для взаимодействия с ДНК и другими белками. В свою очередь, сложные макромолекулярные машины, например, субъединицы РНК-полимеразы, практически полностью лишены неупорядоченных регионов (**Рисунок 26**).

## **5.1. Количественная представленность белков хроматина в клетках**

### **5.1.1. Особенности распределения белков протеома и хроматома по представленности**

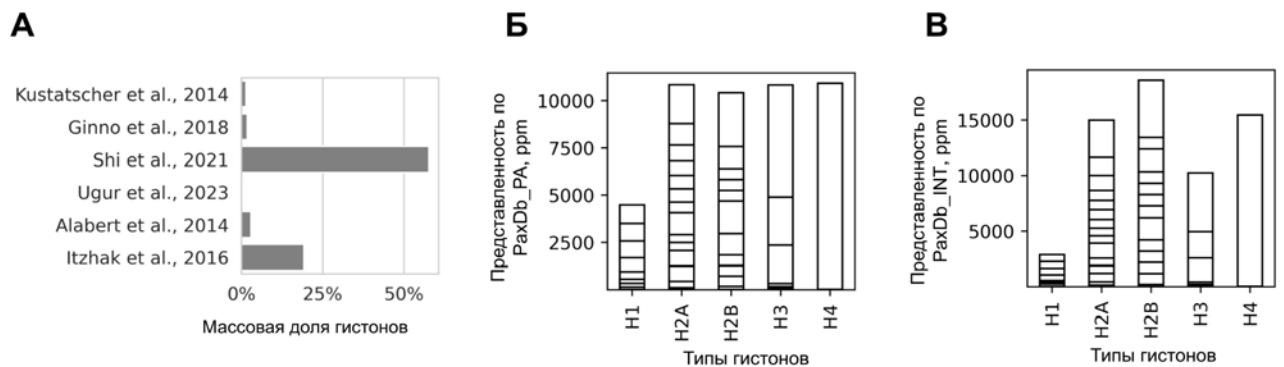
Функционирование хроматина зависит не только от качественного разнообразия входящих в него белков, но и от количества отдельных молекул в ядре клетки (далее представленности) и их стехиометрических соотношений.

В результате анализа масс-спектрометрических данных из восьми исследований (семь хроматомов и один нуклеом) была выявлена высокая вариабельность массовой доли гистонов в образцах (0,1-58 %), рассчитанная по интенсивностям МС-сигналов или значений iBAQ (**Рисунок 27А**). Это наблюдение отражает как биологические различия между исследуемыми образцами, так и различия в методиках пробоподготовки и алгоритмах обработки данных.

Таким образом, для точного установления качественного и количественного состава белковой фракции хроматина методами протеомики требуются дальнейшие исследования. Более того, метод «proteomic ruler» [178], который преобразует интенсивности МС-сигналов в число копий белка на основе информации о количестве гистонов, по всей видимости, малоприменим к препаратам хроматина.

В связи с указанными методическими разногласиями, в качестве меры относительного количества белковых молекул в данной работе использовали их

представленность, выраженную в относительной доле белка в протеоме (parts per million, ppm) по данным из базы данных PaxDb [180]. Анализ представленности белков проводили с использованием наборов данных PaxDb\_INT (данные для 99 % протеома) и PaxDb\_PA (с данными для 84 % протеома, но с наибольшей согласованностью данных). На основании значений представленности белки были разделены на низкопредставленные (менее 1 ppm) и высокопредставленные (более 1 ppm).

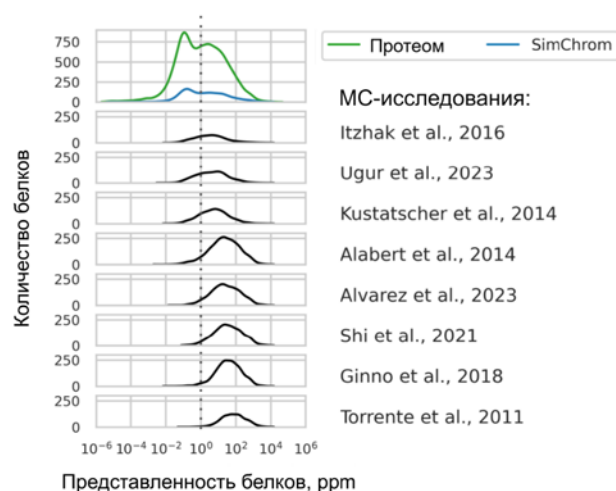


**Рисунок 27** — Анализ представленности гистонов в хроматине и в клетках человека. (А) Массовая доля гистонов по данным протеомных экспериментов. (Б, В) Представленность гистонов различных типов по данным PaxDb\_PA (Б) и PaxDb\_INT (В).

Набор данных PaxDb\_INT содержит на 2741 белок больше, чем набор PaxDb\_PA, причем ~700 из них относятся к белкам ядра или хроматина с низким уровнем представленности. В то же время только в наборе PaxDb\_PA выявлено ожидаемое эквимоллярное соотношение гистонов разных типов (H2A, H2B, H3, H4) (**Рисунок 27Б,В**). Поэтому для анализа распределения представленности белков хроматина был выбран набор PaxDb\_INT, а для более детального исследования относительного количества белков различных функциональных категорий хроматина — набор PaxDb\_PA.

Распределение белков человека по представленности (набор данных PaxDb\_INT) является бимодальным: около половины из них относятся к низкопредставленным белкам (**Рисунок 28**). Распределения белков хроматина и ядра (наборы SimChrom, NULOC\_CS и NULOC\_JT), построенные по значениям представленности из набора PaxDb\_INT, воспроизводят бимодальную форму

распределения белков протеома (**Рисунок 28**). В каждом из этих наборов половина белков также относится к низкопредставленным (40 %, 44 % и 48 % для NULOC\_CS, NULOC\_JT и SimChrom, соответственно), что дополнительно свидетельствует о сбалансированности эталонных наборов по представленности входящих в них белков.



**Рисунок 28** — Распределение белков по количественной представленности (по данным на уровне организма из набора PaxDb\_INT). Показаны наборы: протеом человека, хроматом (SimChrom) и хроматомы по данным масс-спектрометрических исследований. Значения представленности логарифмированы, гистограмма построена с шириной столбца 0,15, выполнено сглаживание гауссовым ядром.

В отличие от бимодальных распределений эталонных наборов, распределения белков, идентифицированных в МС-исследованиях хроматина, по значениям представленности из набора PaxDb\_INT являются унимодальными, с максимумами, смещёнными в область более высоких значений (**Рисунок 28**). Наибольшее смещение в область высоких значений представленности отмечено для хроматомов, пробоподготовка которых включала этап фиксации молекул формальдегидом (Alabert et al., 2023; Shi et al., 2021; Ginno et al., 2018; Torrente et al., 2011).

Доля низкопредставленных белков в хроматомах из МС-исследований составляет от 1 до 27 % (**Рисунок 28**). Наибольшая доля низкопредставленных белков выявлена в результатах исследований, включающих дополнительный этап биоинформатической фильтрации или обработки данных (Kustatscher et al., 2014;

Ugur et al., 2023; Itzhak et al., 2016) (**Рисунок 28**). Полученные результаты согласуются с ранее отмеченной недостаточной идентификацией низкопредставленных белков методами масс-спектрометрии.

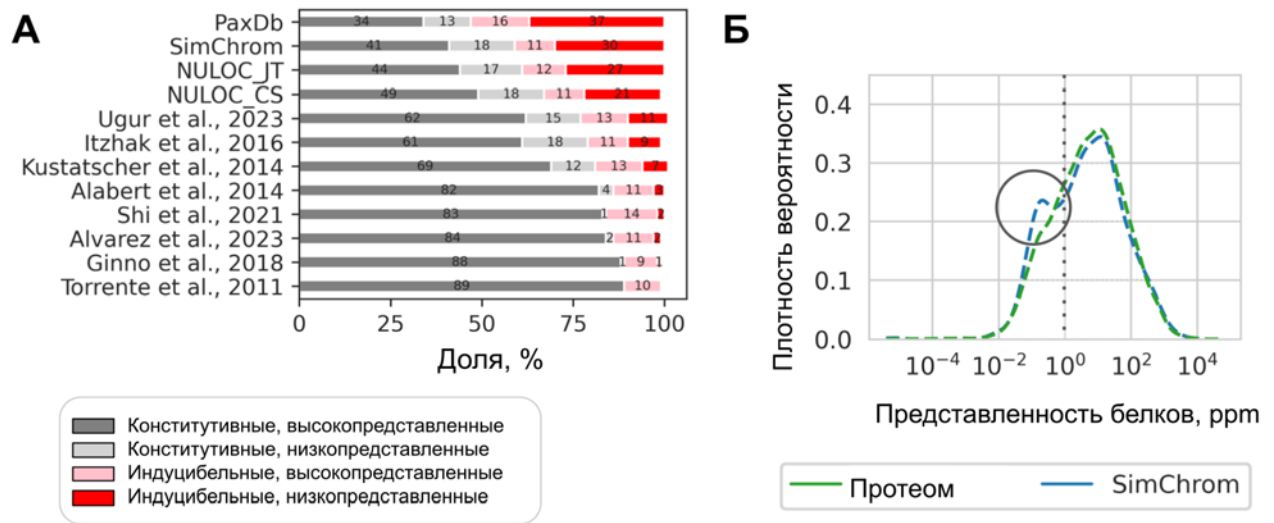
### 5.1.2. Анализ представленности конститутивных белков хроматина

В результате анализа распределения конститутивных и индуцибельных белков по представленности показано, что хроматом человека (набор белков SimChrom) обогащен конститутивными белками: их доля в хроматоме составляет 60 %, в протеоме — 40 %. Исключение из рассмотрения транскрипционных факторов, к которым относится 45 % белков хроматина, приводит к увеличению доли конститутивных белков хроматина до 74 %. Кроме того, конститутивные белки чаще являются высокопредставленными: в протеоме их доля составляет 34 % против 13 % низкопредставленных, в хроматоме — 41 % против 18 % (**Рисунок 29А**).

Обогащение хроматина конститутивными белками связано с увеличением их доли в группе низкопредставленных белков: наблюдаемое значение составило 38 %, тогда как ожидаемое — 32 % (**Рисунок 29Б**). Ожидаемое значение рассчитано по формуле:  $25 \% \times 60 \% / 47 \% = 32 \%$ , где 25 % — доля конститутивных белков среди низкопредставленных белков протеома, 60 % — доля конститутивных белков хроматома, 47 % — доля конститутивных белков протеома. Большинство конститутивных низкопредставленных белков хроматина (64 %) относится к ДНК-связывающим транскрипционным факторам. Этот результат подчеркивает важную регуляторную роль белков с низким уровнем экспрессии в функционировании хроматина. В свою очередь ожидаемая доля конститутивных высокопредставленных белков хроматина должна была составить 86 %, тогда как наблюдаемая составила 80 %.

Хроматомы из протеомных исследований, напротив, значительно обогащены конститутивными белками с высокой представленностью, что, вероятно, является

следствием недостаточной детекции белков с низкой экспрессией (**Рисунки 28, 29А**).



**Рисунок 29** — Представленность конститутивных белков. (**А**) Доля белков протеома, хроматомов (набор SimChrom и результаты МС-исследований) и нуклеомов (наборы NULOC\_JT, NULOC\_CS, (Itzhak et al., 2016)) с делением по типу экспрессии (конститутивные или индуцибельные) и уровню представленности (высокая или низкая). (**Б**) Распределение представленности конститутивных белков протеома и хроматина (набор SimChrom).

### 5.1.3. Представленность белков функциональных категорий хроматина

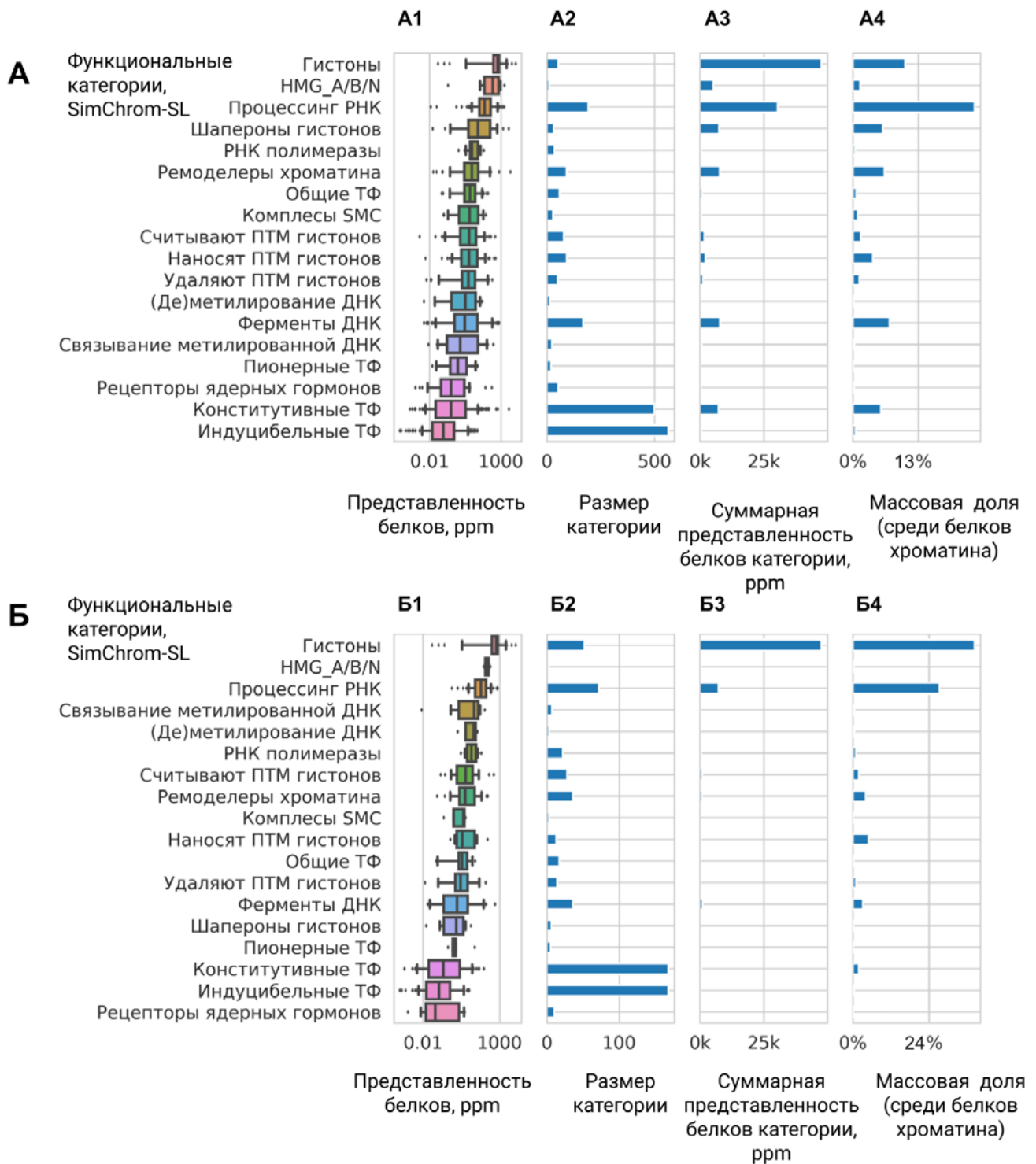
Функциональные категории белков хроматина классификации SimChrom охарактеризованы по трем параметрам: медианная представленность белков в категории, суммарная представленность белков категории и массовая доля категории в хроматоме (**Рисунок 30**). Используемые значения представленности белков (набор данных PaxDb\_PA) рассчитаны на уровне целого организма вне зависимости от внутриклеточной локализации белка. Поэтому аналогичный анализ был проведен также для набора белков хроматина с единственной локализацией в ядре (согласно набору NULOC\_CS\_UL, 802 белка).

Медианные значения представленности белков различных функциональных категорий хроматина различаются более чем на четыре порядка (от 0,09 до

570 ppm), что отражает широкий диапазон концентраций, необходимых для реализации их функций (**Рисунок 30А1**).

Несмотря на небольшое качественное разнообразие гистонов (60 уникальных белков), их суммарная представленность выше, чем у белков других категорий хроматина (**Рисунок 30А3**). Этот результат сохраняется и при анализе белков с единственной локализацией в ядре клетки (**Рисунок 30Б3**). Массовая доля гистонов в хроматоме составила 10 % при рассмотрении всех белков набора SimChrom и 38 % — при анализе белков с единственной локализацией в ядре. Полученные значения обусловлены относительно небольшим размером гистонов (~15 кДа) и дополнительной локализацией большинства белков хроматина вне ядра клетки.

Массовая доля ядерных РНК-связывающих белков в хроматоме составила 30 % и 32 % при анализе белков хроматина с единственной локализацией в ядре (**Рисунок 30А4,Б4**). При этом их доля в протеоме составила 4,8 %, что свидетельствует о распространенности процессов, связанных с метаболизмом РНК и регуляцией транскрипции в клеточном ядре.



**Рисунок 30** — Представленность белков хроматина. **(А)** Представленность белков функциональных категорий хроматина (классификация SimChrom-SL, данные набора RaxDb\_PA). Показано: А1 — распределение белков хроматина по представленности, границы «ящика» соответствуют 25 и 75 перцентилям, «усы» — 5-95 перцентилям; А2 — количество уникальных белков в категории, А3 — суммарная представленность белков категории; А4 — массовая доля белков категории в хроматоме. Категории SimChrom отсортированы по убыванию медианного значения представленности. **(Б)** То же самое для белков хроматина SimChrom с единственной локализацией в ядре клетки.

Половина белков SimChrom (1345 уникальных белков) относится к ДНК-связывающим ТФ, однако их вклад в массовую долю белков хроматина составляет 6,3 %. Несмотря на относительно низкую экспрессию, эти белки важны для регуляции клеточных процессов, таких как реакция на стимулы окружающей среды, дифференцировка тканей и развитие организма [204].

К другим функционально значимым категориям белков хроматина с высоким медианным значением представленности относятся: белки суперсемейства HMG; белки, отщепляющие хвосты гистонов; шапероны гистонов; субъединицы РНК-полимераз; субъединицы комплексов ремоделирования хроматина (**Рисунок 30A1**).

По медиане представленности белки суперсемейства HMG уступают только гистонам (**Рисунок 30A1**). Численные соотношения белков семейств HMG и нуклеосом составляют 1:8 для представителей семейства HMGA, 1:2 — для HMGB и 1:3 — для HMGN. Однако большинство белков HMG также локализуется в неядерном компартменте клетки, что приводит к завышению оценки их представленности в хроматине. Белки, отщепляющие хвосты гистонов, выполняют основные функции вне ядра клетки, что также искажает оценку их представленности в хроматине.

Среди белков хроматина наиболее представленным по относительному количеству молекул является гистон H4 (~11000 ppm). Этот белок кодируется семейством генов гистонов H4 (за исключением гена *H4C7*). Далее представленность белков будет охарактеризована с использованием значения в ppm (набор PaxDb\_PA) и в процентах относительно представленности гистона H4.

Суммарная представленность гистонов каждого типа (H2A, H2B, H3, H4) сопоставима (~10400-10900 ppm), что соответствует их ожидаемому эквимольному соотношению в клетке. Для линкерного гистона H1 суммарная представленность составила 4500 ppm, что соответствует соотношению одна молекула гистона H1 на одну нуклеосому. Согласно проведенному анализу, наиболее распространенными вариантными гистонами являются: H3.3 (2530 ppm, 23 % от представленности гистона H4), H2A.X (714 ppm, 6,5 %), H2A.Z (1140 ppm,

10 %), H2A.W (423 ppm, 3,8 %) (**Рисунок 31**). К наименее распространенным гистоновым вариантам (менее 1 ppm) относятся H2A.B и H1.7 (**Рисунок 31**).

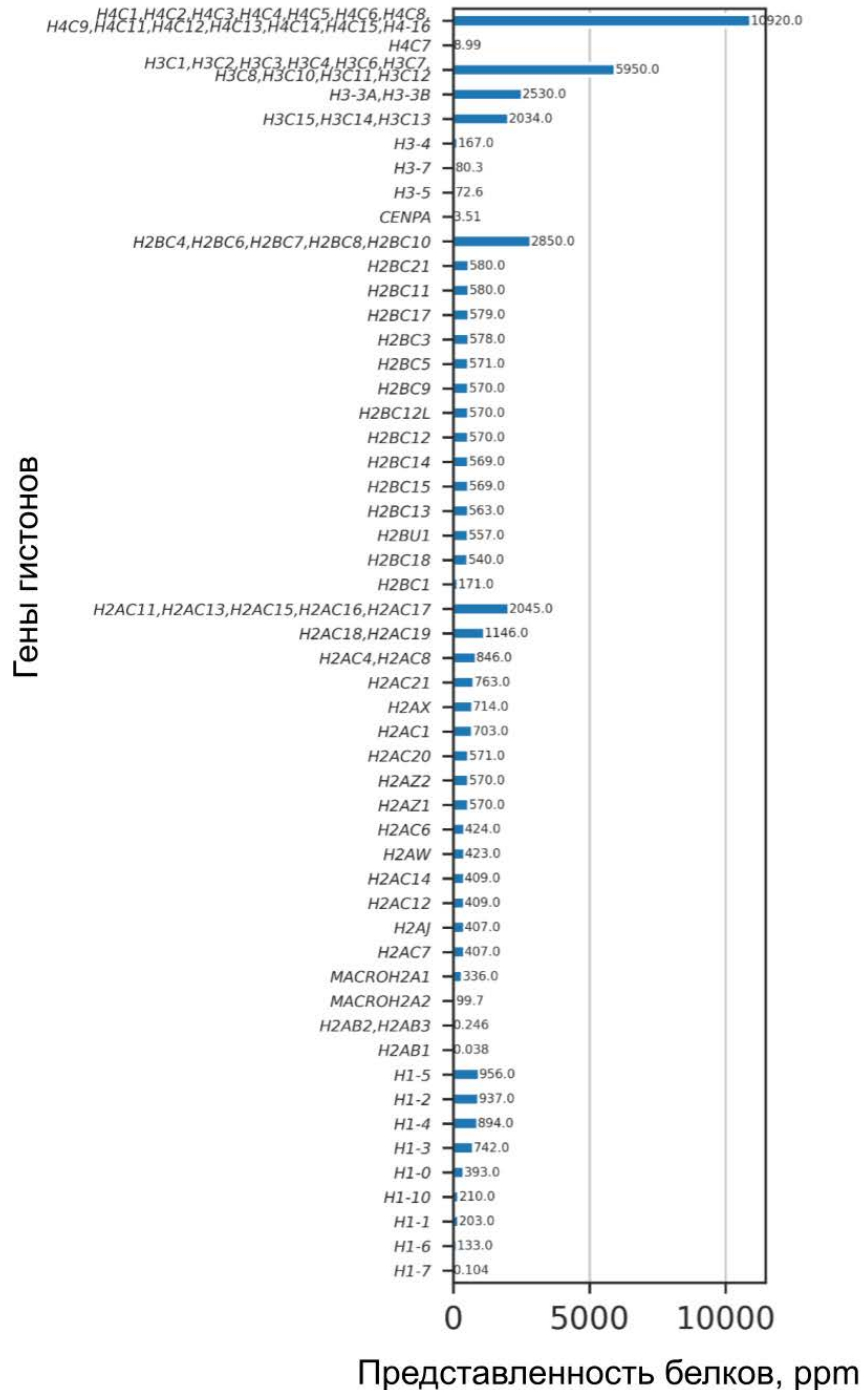
Среди гистоновых шаперонов наиболее представлены: NPM1 (специфичен для димера H3-H4, 32 % от представленности гистона H4); NCL (специфичен для димера H2A-H2B, 12 % от H4); ANP32E (специфичен для димера H2A.Z-H2B, 2 % от гистона H4) (**Рисунок 32**). Количество шаперона гистонов FACT, состоящего из двух субъединиц SSRP1 и SUPT16H, составляет 1 % от гистона H4, что соответствует одному комплексу FACT на ~50 нуклеосом.

Среди субъединиц РНК-полимераз наиболее представлена субъединица POLR2E — общая для РНК-полимераз I, II и III (0,58 % от представленности H4 или 1 белок на ~90 нуклеосом). Представленность субъединиц, эксклюзивных для РНК-полимеразы II (например, POLR2B, POLR2C, POLR2D), находится в диапазоне 0,05-0,3 % от представленности H4, что соответствует одному комплексу на 10 генов. Оценка получена с учетом средней длины гена человека (~24 тысяч пар нуклеотидов) и среднего расстояния между центрами двух соседних нуклеосом (~200 нуклеотидов).

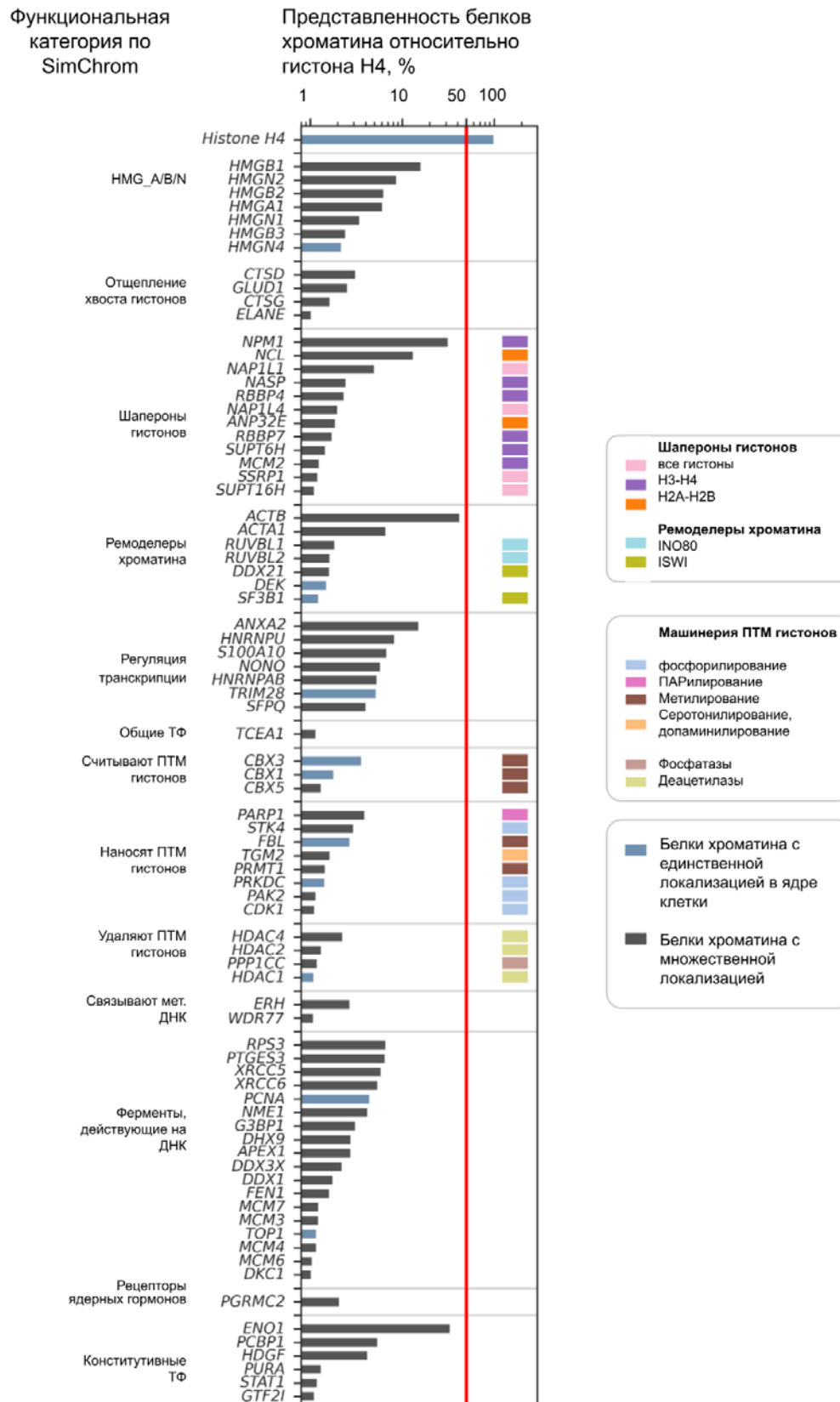
Среди компонентов комплексов ремоделирования хроматина наиболее представлены изоформы актина (ACTB, ACTA1) и актиноподобный белок ACTL6A (**Рисунок 32**). Несмотря на то что актин действительно входит в состав ремоделеров семейства SWI/SNF [205], его высокая представленность обусловлена в первую очередь белками, участвующими в формировании цитоскелета в цитоплазме.

Представленность белков RUVBL1 и RUVBL2 (семейство AAA+ АТФаз, англ. ATPases Associated with diverse cellular Activities), составляет 1,6–1,9 % от представленности гистона H4 (**Рисунок 32**). Эти белки формируют гетерогексамерные кольца в ремоделерах хроматина семейства INO80 [205], однако эта функция не является для них основной [206]. Аналогично, белок UCHL5, представленность которого составила 0,6 % от представленности H4, является необязательным компонентом комплексов ремоделирования хроматина семейства INO80. Основная АТФаза ремоделеров этого семейства (ген *INO80*),

имеет значительно меньшую представленность — 0,006 % от представленности H4, что соответствует одному комплексу данного семейства на ~8680 нуклеосом.



**Рисунок 31** — Представленность гистоновых белков согласно набору данных PaxDb\_PA.



**Рисунок 32** — Наиболее представленные белки хроматина в клетках человека, относящиеся к категориям молекулярных функций или физико-химических свойств согласно SimChrom. Показаны белки с относительной представленностью более 1 % от представленности гистона H4.

Представленность АТФазных субъединиц других семейств ремоделеров хроматина составила: SMARCA5 (семейство ISWI) — 0,7 % от представленности гистона H4; вспомогательные субъединицы DDX21, SF3B1, MYBBP1A — 0,6-1,6 %; CHD4 (семейство CHD) — 0,6 %; SMARCA4/SMARCA2 (семейство SWI/SNF) — 0,2 %; вспомогательные субъединицы SMARCC2, SMARCC1, SMARCE — 0,6 %. На основании этих значений были получены следующие соотношения: один комплекс ISWI приходится примерно на 70 нуклеосом, CHD — на 80, SWI/SNF — на 200–300 нуклеосом.

Среди белков, участвующих в формировании комплексов SMC, наиболее представлены SMC3 и SMC1A (0,8 % от представленности гистона H4). Совместно с белком RAD21 (0,46 % от представленности H4) они формируют комплекс митотического когезина, который, как недавно было обнаружено, не только обеспечивает сцепление сестринских хроматид, но и осуществляет экструзию петель ДНК для поддержания топологически ассоциированных доменов и трехмерной организации генома [207]. Таким образом, один комплекс SMC1A-SMC3 приходится на ~60 нуклеосом.

Среди белков, распознающих посттрансляционные модификации гистонов, наиболее представлены HP1 $\alpha$ , HP1 $\beta$  и HP1 $\gamma$ , которые связываются с метилированным остатком 9-го лизина гистона H3 (H3K9me3) и способствуют формированию гетерохроматина. Их представленность составила: HP1 $\gamma$  (ген *CBX3*) — 3,65 % от представленности гистона H4, HP1 $\beta$  (ген *CBX1*) — 1,82 %, HP1 $\alpha$  (ген *CBX5*) — 1,33 % (**Рисунок 32**).

Среди белков, наносящих ПТМ на белки (в том числе гистоны), наиболее распространены (**Рисунок 32**):

- PARP1 (3,97 % от представленности гистона H4) — ПАРилирует, то есть наносит поли(АДФ)-рибозные цепочки на белки-мишени, включая гистоны;
- киназы STK4 (2,99 %), PRKDC (1,46 %), PAK2 (1,17 %), CDK1 (1,13 %);
- FBL (2,73 %) — метилирует РНК и остаток глутамина-104 гистона H2A (H2AQ104) [208];
- PRMT1 (1,5 %) — метилирует остаток аргинина-3 гистона H4 (H4R3);

- PRMT5 (0,7 %) — метилирует гистоны H2A, остаток аргинина-3 гистона H4 (H4R3) и остаток аргинина-8 гистона H3 (H3R8);
- NAA50 (0,87 %) — ацетирует гистон H4;
- TGM2 (1,67 %) — осуществляет серотонилирование и дофаминилирование гистона H3.

К ключевым эпигенетическим регуляторам относятся комплексы поликомб: polycomb repressive complex 1 (PRC1) осуществляет моноубиквитинирование остатка лизина-119 гистона H2A (H2AK119), а polycomb repressive complex 2 (PRC2) триметирует остаток лизина-27 гистона H3 (H3K27). Количество субъединиц комплекса PRC1, а именно каталитической субъединицы RNF2 и рекрутирующих субъединиц CBX2/4/6/7/8, составляет 0,02–0,18 % от представленности гистона H4, что соответствует одному комплексу PRC1 на 275 нуклеосом. В свою очередь, количество субъединиц комплекса PRC2 оказалось в два раза меньше. Представленность основных субъединиц его минимального каталитически активного комплекса (EZH2, EED и SUZ12) составила 0,08-0,09 %, что соответствует одному комплексу PRC2 на 550 нуклеосом.

Среди наиболее представленных белков, удаляющих ПТМ гистонов, были выявлены следующие:

- HDAC4, HDAC2 и HDAC1 (2,28 %, 1,34 %, 1,1 % соответственно) — деацетилазы, неспецифичные по отношению к гистоновым белкам;
- MTA2 (0,6 %) — деацетилаза, ключевой компонент комплекса ремоделирования Nucleosome Remodeling Deacetylase (NuRD)
- PPP1CC, PPPCA, PPP2CB, PPP5C (0,6–1,2 %) — фосфатазы.

Наиболее распространенной деметилазой гистонов оказался белок JMJD6 (0,18 %), деметилирующий гистоны по положениям H3R2 и H4R3 (**Рисунок 32**).

К наименее представленным в клетках человека белкам хроматина относятся ДНК-связывающие ТФ (**Рисунок 30**). Представленность индуцибельных ДНК-связывающих ТФ находится в диапазоне от 0,00008 до 23,9 ppm, что свидетельствует о минимальной экспрессии белков данной группы при усреднении

значений на уровне организма. Для большинства конститутивных ТФ также была выявлена сравнительно низкая представленность (медианное значение — 0,31 ppm). Вместе с тем отдельные белки данной группы демонстрируют значительно более высокие уровни представленности (**Рисунок 32**). Так, среди белков с единственной локализацией в ядре наибольшая представленность отмечена у PURB — 91,8 ppm (0,84 % от представленности H4). Среди белков с множественной локализацией, включая ядерную, максимальная представленность отмечена у ENO1 — 3667 ppm (33,6 % от представленности H4).

Согласно данным К.Е. ван Хольде, масса гистоновых белков примерно соотносится с массой ДНК в клетке, тогда как доля негистоновых хромосомных белков варьирует от 0,3 до 0,8 г на 1 г ДНК [1]. Такие оценки могут существенно различаться в зависимости от экспериментального подхода и его разрешающей способности [209]. Насколько эти оценки согласуются с определенным репертуаром белков хроматина из данной работы (набор белков SimChrom), требует дальнейшего исследования. Тем не менее полученные нами результаты согласуются с общепринятым представлением о том, что гистоны преобладают в ядре как по числу молекул, так, вероятно, и по массе.

В то же время белки, участвующие в процессинге РНК (факторы сплайсинга; белки, связывающие пре-мРНК и другие), сопоставимы с гистонами по массовой доле за счет большей длины аминокислотных последовательностей белков. Клеточное ядро играет ключевую роль не только в хранении и реализации наследственной информации, но и в производстве и созревании молекул РНК. Более того, наиболее распространенные белки ядра — гистоновые шапероны нуклеофозмин (NPM1) и нуклеолин (NCL) — также участвуют в организации ядрышка и биогенезе рибосом. Полученные результаты также согласуются с ранее полученными данными о высокой представленности коротких белков суперсемейства HMG, которые неспецифически взаимодействуют с ДНК или нуклеосомой для модуляции архитектуры и динамики хроматина [210].

#### 5.1.4. Анализ представленности белков по данным МС-исследований

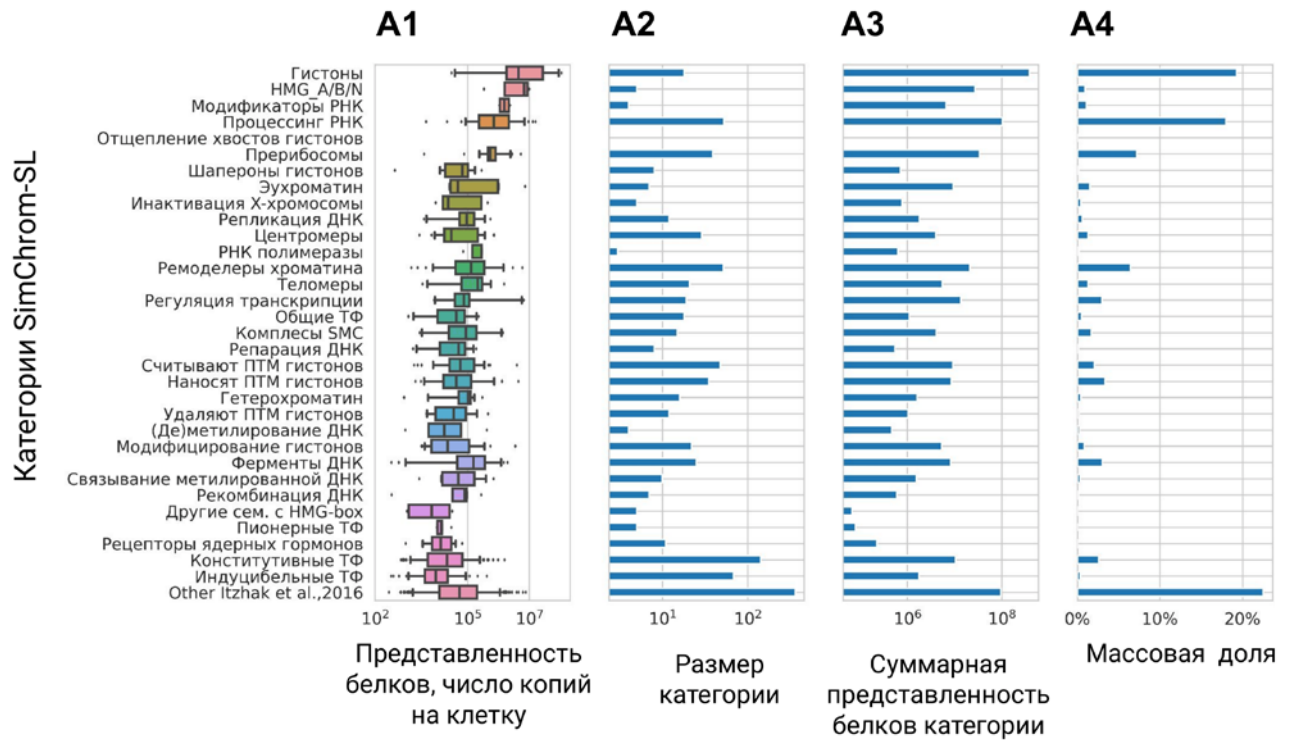
Аналогичный анализ был проведен для белков, идентифицированных в масс-спектрометрических исследованиях хроматомов и нуклеома. Белки были сгруппированы по категориям хроматина согласно классификации SimChrom-SL, а их относительная представленность и массовая доля были рассчитаны на основе исходных данных.

Массовая доля белков хроматина из набора SimChrom в МС-исследованиях составила от 33 до 81 %: по данным (Ginno et al., 2018) — 33 %; (Ugur et al., 2023) — 43 %; (Alabert et al., 2014) — 65 %; (Itzhak et al., 2016) — 77 %; (Shi et al., 2021) — 78 %; (Kustatscher et al., 2014) — 81 %. При этом в некоторых исследованиях (Ginno et al., 2018, Ugur et al., 2023) белки хроматина из набора SimChrom не преобладали по массовой доле среди других идентифицированных белков.

В МС-исследованиях (Shi et al., 2021) и (Alabert et al., 2014) преобладание массовой доли белков хроматина SimChrom было обусловлено высокой массовой долей одной из категорий: в наборе (Shi et al., 2021) — гистонов, в наборе (Alabert et al., 2014) — ферментов, действующих на ДНК.

В работе (Itzhak et al., 2016) авторы отбирали преимущественно ядерные белки, поэтому распределение белков из этого набора по категориям SimChrom-SL, выраженное в количестве копий белковых молекул на клетку (copy number per cell), было наиболее сбалансированным (**Рисунок 33**). Массовая доля идентифицированных белков в этом исследовании составила: «Ядерные РНК-связывающие белки» — 26 %, «Гистоны» — 19 %, «Ремоделеры хроматина» — 6,4 %, «Белки, наносящие ПТМ гистонов» — 3,4 % (**Рисунок 33**).

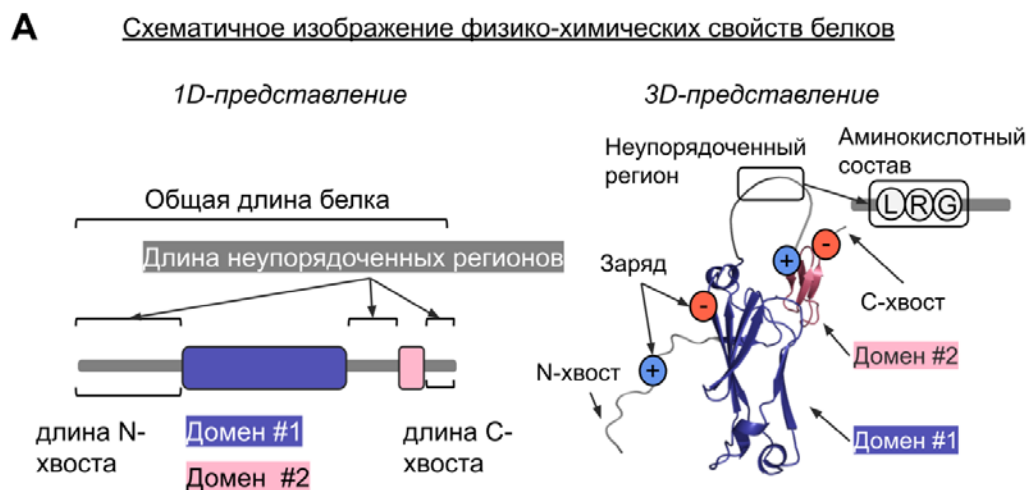
Таким образом, в МС-исследованиях хроматомов и нуклеома выявлена высокая вариабельность оценок представленности молекул белков хроматина.



**Рисунок 33** — Представленность идентифицированных в МС-исследовании (Itzhak et al., 2016) белков хроматина по классификации SimChrom-SL, выраженная в количестве копий белков на клетку. Показано: A1 — распределение идентифицированных белков по представленности, границы «ящика» соответствуют 25 и 75 перцентилям, «усы» — 5-95 перцентилям; A2 — количество уникальных белков в категории, A3 — суммарная представленность белков категории; A4 — массовая доля белков категории среди всех идентифицированных белков.

## 5.2. Физико-химические свойства белков хроматина

Анализ физико-химических свойств белков хроматина проводили на основе их первичной структуры (аминокислотной последовательности) (**Рисунок 34**). Для выявления специфических особенностей строения белков хроматина было проведено сравнение их свойств со свойствами белков с локализацией в цитоплазме клетки (набор белков CYTLOC\_CS\_UL) или с белками с единственной локализацией в ядре клетки (набор NULOC\_CS\_UL).



**Б** Ключевые различия между белками хроматина и цитоплазмы

↑ - в среднем выше, ↓ - в среднем ниже

| Свойство белков                | Белки хроматина     |
|--------------------------------|---------------------|
| Общая длина                    | ↑                   |
| Длина N-хвоста                 | ↑                   |
| Длина C-хвоста                 | ↑                   |
| Суммарный заряд                | более положительный |
| Заряд N-хвоста                 | отрицательный       |
| Длина неупорядоченных регионов | ↑                   |
| Доля неупорядоченности         | ↑                   |
| Доля остатков C,S,P,H,K        | ↑                   |
| Доля гидрофобных остатков      | ↓                   |

**Рисунок 34** — Анализ физико-химических свойств белков хроматина. (**А**) Схематичное изображение проанализированных физико-химических свойств на последовательности и структуре белка. (**Б**) Краткое изложение результатов раздела.

### 5.2.1. Длина белков хроматина

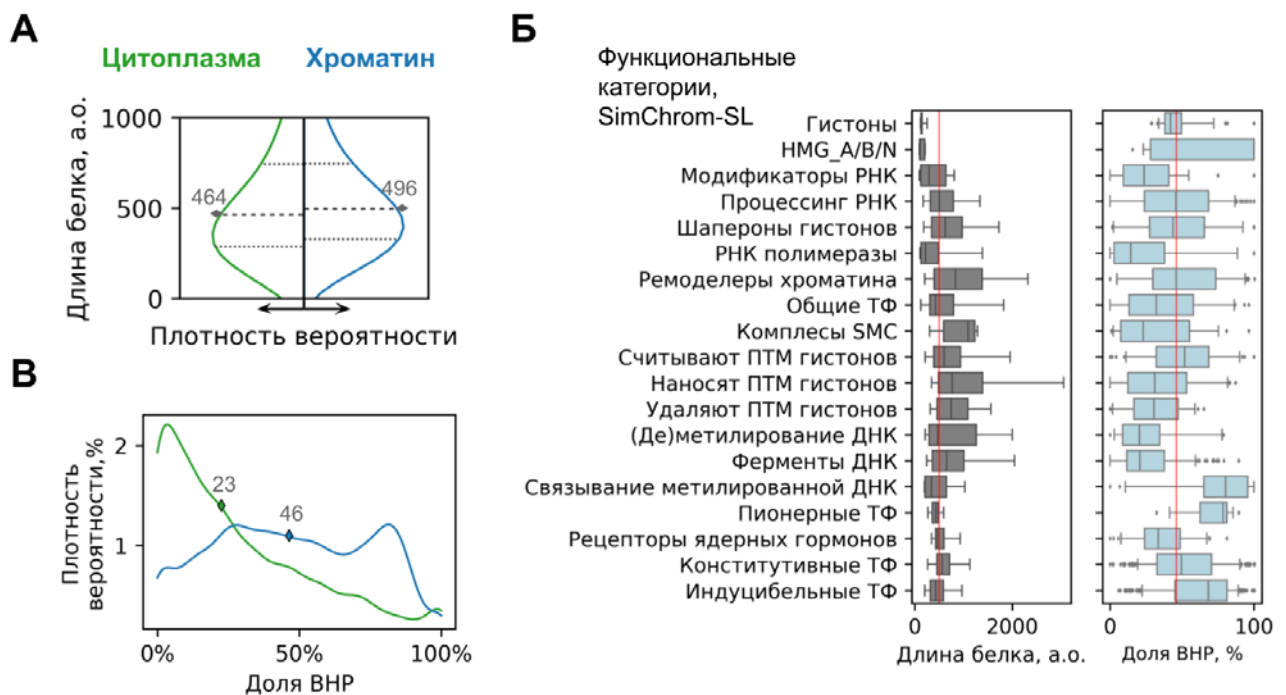
В качестве показателя длины белка анализировали число его аминокислотных остатков (а.о.). Медианные значения длин белков хроматина (набор SimChrom) и ядра клетки (набор NULOC\_CS\_UL) составили 496 а.о. и 500 а.о., что на 7 % превышает медианную длину белков цитоплазмы (набор CYTLOC\_CS\_UL) — 464 а.о. Выявлены различия в форме распределений белков хроматина и цитоплазмы по длине аминокислотных последовательностей: в области значений длины до 800 а.о. распределение белков хроматина смещено в сторону более высоких значений (**Рисунок 35А**). Доля коротких белков (менее 200 а.о.) среди белков хроматина в 1,6 раз ниже, чем среди белков цитоплазмы (8 и 13 % соответственно), что свидетельствует о меньшей распространенности коротких белков в составе хроматина.

При более детальном анализе категорий белков хроматина были отмечены следующие различия в длинах белков (**Рисунок 35Б**). Самыми короткими являются гистоны и белки НМГ (медианные значения 130 а.о. и 109 а.о. соответственно), а также субъединицы комплексов РНК-полимеразы (POLR2K — 58 а.о., POLR2L — 67 а.о., при медианном значении длины белков категории 220 а.о.).

Наибольшие медианные значения длины были выявлены для следующих категорий белков хроматина: «SMC-комплексы» (1096 а.о.), «Белки, наносящие ПТМ гистонов» (766 а.о.), «Белки, стирающие ПТМ гистонов» (747 а.о.), «Ферменты, действующие на ДНК» (655 а.о.).

К категориям белков хроматина с наибольшей вариабельностью длины аминокислотной последовательности (на основании разницы между 5-м и 95-м перцентилями распределения длины белков категории) относятся следующие категории: «Белки, считывающие ПТМ гистонов» (1741 а.о.), «Общие ТФ» (1746 а.о.), «Шапероны гистонов» (1768 а.о.), «Ферменты, действующие на ДНК» (1810 а.о.), «Белки, осуществляющие (де)метиляцию ДНК» (1830 а.о.), «Белки, ассоциированные с центромерами» (2013 а.о.), «Ремоделеры хроматина» (2314 а.о.), «Белки, наносящие ПТМ гистонов» (2988 а.о.) (**Рисунок 35Б**).

К категориям белков хроматина с наименьшей вариабельностью длины белков и количеством белков, превышающим 15, были отнесены следующие: «Гистоны» (медиана 144 а.о.), «Пионерные ТФ» (321 а.о.), «Ядерные рецепторы гормонов» (598 а.о.), «Индукцибельные ДНК-связывающие ТФ» (766 а.о.), «Белки, участвующие в модификации РНК» (768 а.о.), «Белки, узнающие метилирование ДНК» (856 а.о.), «Конститутивные ДНК-связывающие ТФ» (874 а.о.), «Другие семейства белков с доменом HMG-box» (943 а.о.) (**Рисунок 35Б**).



**Рисунок 35** — Сравнение белков хроматина и цитоплазмы по длине аминокислотной последовательности и доле внутренне неупорядоченных регионов. (**А, В**) Распределения белков хроматина и цитоплазмы по длине (**А**) и доле ВНР (**В**); показаны медианные значения. (**Б**) Длина и доля ВНР в белках хроматина с разделением по функциональным категориям согласно классификации SimChrom-SL; границы «ящика» соответствуют 25 и 75 перцентилям, «усы» — 5-95 перцентилям, красной линией показано медианное значение для белков хроматина.

### 5.2.2. Внутренне неупорядоченные регионы белков хроматина

В качестве метрики для аннотация внутренне неупорядоченных регионов (ВНР) использовали площадь поверхности, доступной растворителю (solvent-

accessible surface area, SASA), рассчитанную по структурам белков, предсказанных с помощью AlphaFold 2 и усредненную по 20 а.о. [118]. Регионы белков, не относящиеся к ВНР, считались внутренне упорядоченными регионами (ВУР).

Распределения белков хроматина и цитоплазмы по доле неупорядоченных регионов в белках хроматина и цитоплазмы значительно различаются (**Рисунок 35В**). Медианное значение доли ВНР в белках хроматина и ядра составило 46 % и 51 % соответственно, для белков цитоплазмы — 23 % (**Рисунок 35В**). Таким образом, у половины белков хроматина большая часть последовательности относится к внутренне неупорядоченным регионам, тогда как для белков с единственной локализацией в цитоплазме это верно лишь для четверти белков.

Выявлено, что увеличение доли ВНР в белках хроматина обусловлено большей длиной отдельных ВНР и большим количеством таких регионов в белках. Медианные значения длин отдельных ВНР в белках хроматина составили ~70 а.о., в белках цитоплазмы — ~50 а.о. Доля белков с количеством ВНР от двух до четырех была значительно выше среди белков хроматина и ядра по сравнению с белками цитоплазмы. Так, в состав 38 % и 41 % белков ядра и хроматина соответственно входило более двух ВНР, в то время как доля таких белков цитоплазмы составила 27 %. И наоборот, доля белков без ВНР или с одним ВНР была выше среди белков цитоплазмы (20 %), по сравнению с белками ядра (18 %).

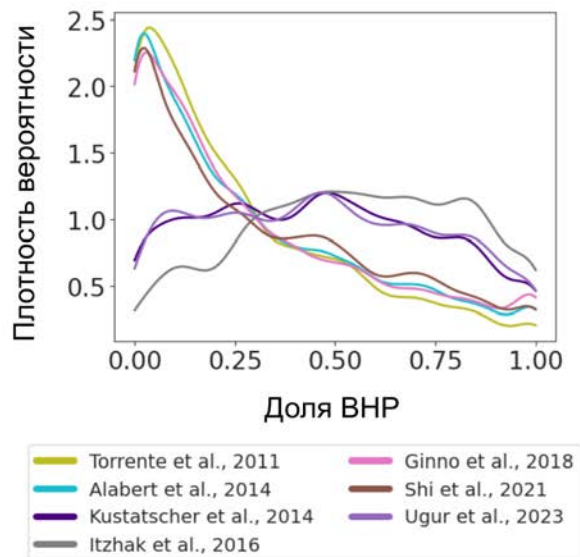
Был проведен детальный анализ доли ВНР в белках из категорий классификации SimChrom-SL. Это позволило выявить категории белков хроматина с наименее и наиболее упорядоченными белками (**Рисунок 35Б**). К категориям с медианным значением доли ВНР меньшим, чем для белков цитоплазмы, относятся: субъединицы РНК-полимераз (доля ВНР 0 %), «Белки, отщепляющие хвосты гистонов» (3 %), «Белки, ассоциированные с прерибосомой» (10 %), «Ферменты, действующие на ДНК» (16 %), «Белки, модифицирующие РНК» (16 %), «Белки, осуществляющие (де)метиляцию ДНК» (17 %), субъединицы «SMC-комплексов» (18 %), «Белки, ассоциированные с репарацией ДНК» (19 %).

Наибольшее медианное значение доли ВНР в белках хроматина было отмечено для следующих категорий белков: «HMG A/B/N» (99 %, при этом 7 из 11 белков были полностью неупорядоченными), «Другие семейства белков с доменом HMG-box» (74 %), «Пионерные ТФ» (73 %), «Белки, узнающие метилирование ДНК» (62 %).

Дополнительный пик в распределении белков хроматина по доле ВНР в области высоких значений (~80 %) обусловлен транскрипционными факторами, прежде всего индуцибельными (в них медианное значение доли ВНР 68 %). Среди них 453 белка с долей ВНР, превышающей 50 % длины аминокислотной последовательности. Форма распределения белков хроматина без учета ДНК-связывающих ТФ по доле ВНР в большей степени совпадает с формой распределения белков цитоплазмы.

Аналогичный анализ был проведен для наборов белков хроматина, идентифицированных методами масс-спектрометрии (**Рисунок 36**). Для хроматомов из работ (Kustatscher et al., 2014), (Itzhak et al., 2016) и (Ugur et al., 2023) форма распределения белков по доле ВНР соотносится с формой распределения для эталонных наборов белков хроматина (SimChrom) и ядра (NULOC\_CS\_UL) (**Рисунок 36**). Это связано с наличием в этих наборах белков с низкой представленностью, в частности, индуцибельных транскрипционных факторов. Форма распределений для наборов белков из других исследований соответствует форме распределения для белков цитоплазмы (CYTLOC\_CS\_UL).

Разбиение белков хроматина по функциональным категориям SimChrom позволило определить, что индуцибельные ДНК-связывающие ТФ (453 белка с долей ВНР более 50 %) вносят наибольший вклад в увеличение доли неупорядоченности белков хроматина.



**Рисунок 36** — Распределение белков хроматина из МС-исследований хроматомов по доле неупорядоченных регионов.

Таким образом, в белках хроматина увеличены длина и количество отдельных ВНР, что важно для функционирования белков: ВНР необходимы для нахождения целевых последовательностей ДНК [211] и регуляции транскрипции за счет разделения жидких фаз [212].

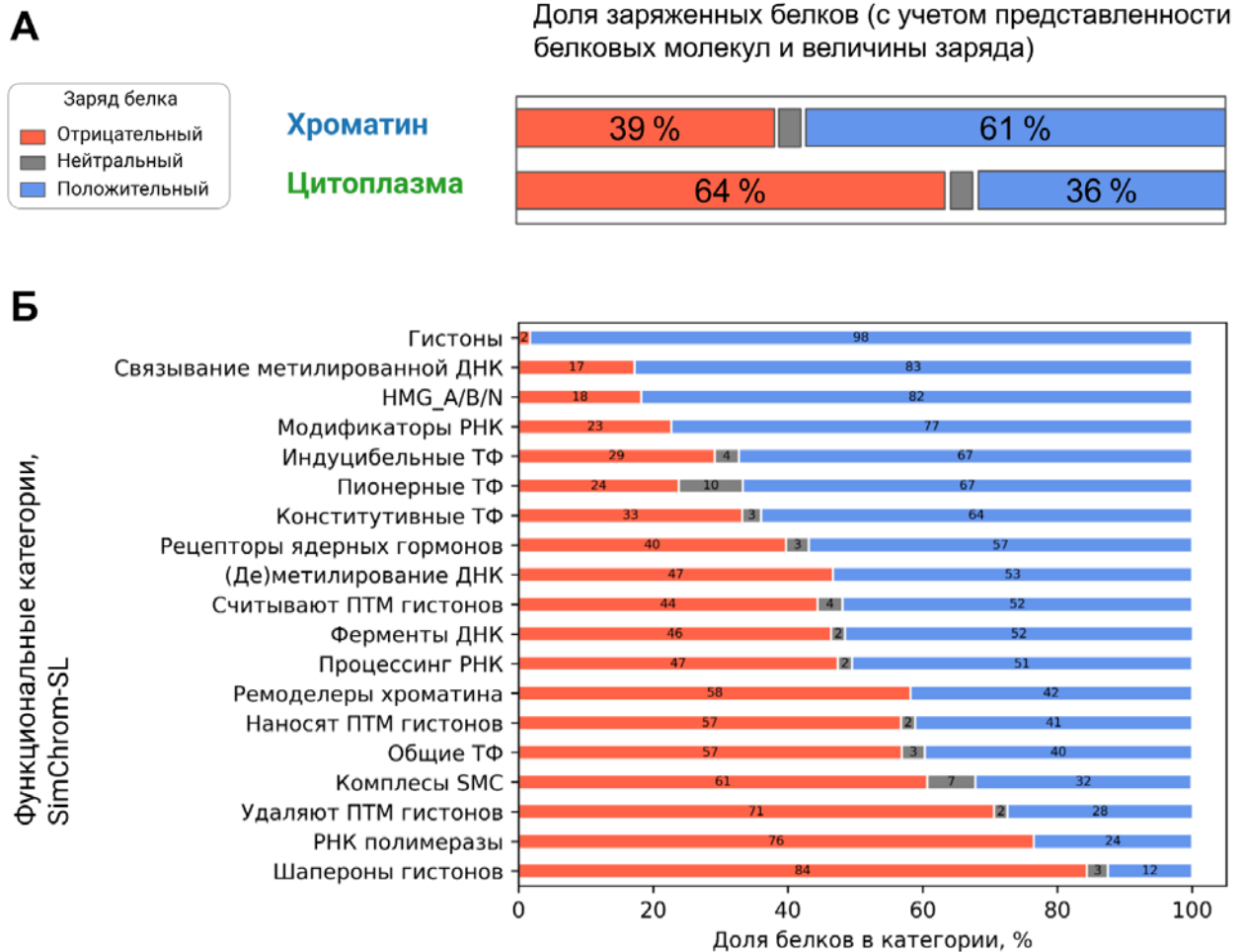
### 5.2.3. Заряд белков хроматина

Белки хроматина в среднем характеризуются положительным зарядом для компенсации отрицательного заряда молекулы ДНК. Был проведен сравнительный анализ распределения заряда белков хроматина (набор SimChrom), белков с единственной локализацией в ядре (NULOC\_CS\_UL) или цитоплазме (CYTLOC\_CS\_UL) (**Рисунок 37**).

Для анализа распределения белков по знаку заряда применены три подхода.

В рамках первого подхода рассматривали заряд отдельных белков (положительный, отрицательный или нейтральный). Так, доля положительно заряженных белков хроматина и ядра составила 54 и 58 % соответственно; в то время как доля положительно заряженных белков цитоплазмы — 39 %.

Наибольшая доля отрицательно заряженных белков в хроматине выявлена в следующих категориях: «Шапероны гистонов» (84 % белков категории), «РНК-полимеразы» (76 %), «Белки, стирающие ПТМ гистонов» (71 %) (**Рисунок 37Б**). Это объясняется ассоциацией отрицательно заряженных белков не с отрицательно заряженной молекулой ДНК, а с положительно заряженными гистонами.



**Рисунок 37** — Заряд белков хроматина. (А) Распределение белков хроматина (набор SimChrom) и цитоплазмы (набор CYTLOC\_CS\_UL) по знаку заряда с учетом представленности белковых молекул и величины заряда. (Б) Распределение белков функциональных категорий хроматина (классификация SimChrom-SL) по знаку заряда.

Однако такой подход не учитывает ни представленность белков, ни величину их заряда. Поэтому в рамках второго подхода были рассчитаны доли заряженных белковых молекул (с учетом представленности каждого белка согласно набору данных PaxDb\_PA). В рамках третьего подхода был рассчитан суммарный заряд

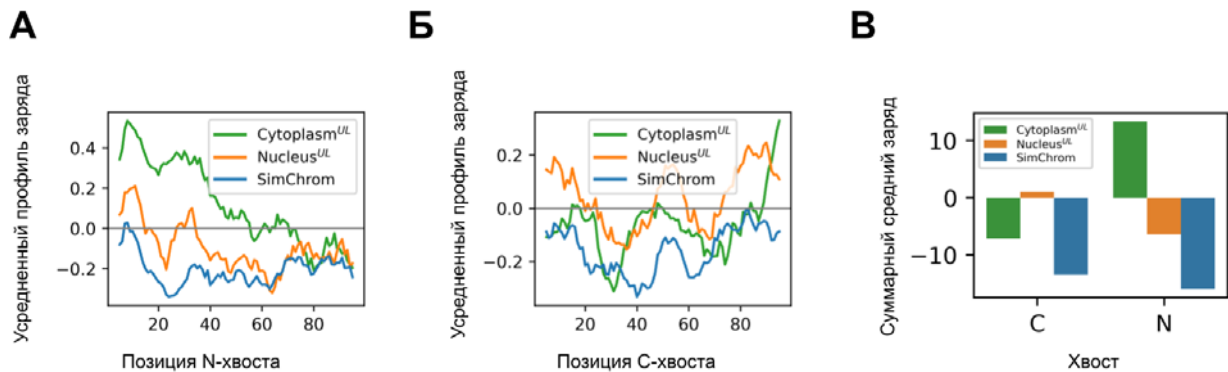
белковых молекул с учетом как величины заряда отдельных молекул, так и их представленности (по PaxDb\_PA) (**Рисунок 37А**).

Существенное влияние учета представленности белковых молекул и величины заряда наблюдалось только для набора белков с единственной локализацией в клеточном ядре: доля положительно заряженных белковых молекул составила 82 %, а доля суммарного положительного заряда (с учетом величины заряда и представленности белков) — 89 %. Это связано с высокой представленностью и значительным положительным зарядом гистоновых белков, входящих в набор NULOC\_CS\_UL. В то же время для набора белков хроматина SimChrom распределения положительно заряженных белков, белковых молекул и суммарного заряда отличались незначительно: 54, 60 и 61 % соответственно.

В результате анализа распределения белков хроматина по суммарному заряду с разбиением на функциональные категории были выявлены категории со значительным по величине отрицательным зарядом: «Процессинг РНК», «Ремоделирование хроматина», «Шапероны гистонов» и «Ферменты, действующие на ДНК». Однако в составе этих категорий присутствуют белки с двойной локализацией в цитоплазме и ядре клетки. Поэтому в результате аналогичного анализа для белков хроматина с единственной локализацией в ядре величина суммарного отрицательного заряда в белках отмеченных категорий была ничтожно малой по сравнению с величиной положительного заряда молекул гистонов.

Были выявлены особенности распределения заряда в N- и C-концевых фрагментах (хвостах) белков. Медианные значения длин N-/C-концевых хвостов составили 51/47, 72/84 и 67/79 а.о. для белков из наборов CYTLOC\_CS\_UL, NULOC\_CS\_UL и SimChrom соответственно. Различия в знаке заряда усредненных профилей заряда между белками цитоплазмы и белками ядра/хроматина наблюдались только для N-концевых хвостов (**Рисунок 38А,Б**). Суммарный усредненный заряд N-концевых хвостов составил –16 для белков хроматина, –6 для ядерных и +13 для белков цитоплазмы (**Рисунок 38В**). Было выявлено, что наиболее положительно заряженные хвосты находятся в гистонах, а наиболее

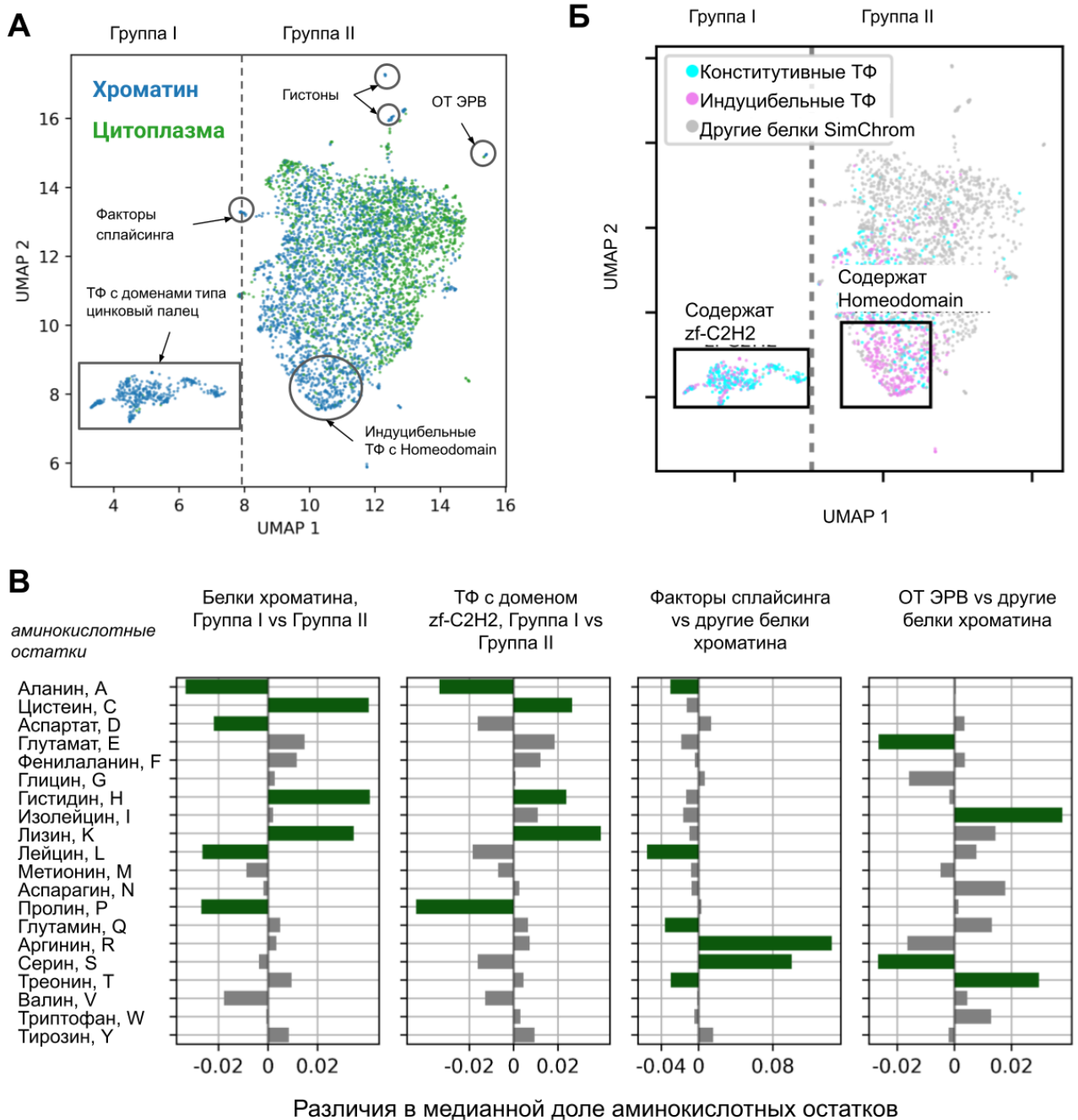
отрицательные — в гистоновых шаперонах и белках HMG. Примечательно, что хвосты факторов транскрипции в среднем заряжены отрицательно, несмотря на общий положительный заряд белков данной группы.



**Рисунок 38** — Анализ распределения заряда в N- и C-концевых хвостах белков хроматина (набор белков SimChrom), ядра (набор NULOC\_CS\_UL) и цитоплазмы (CYTLOC\_CS\_UL). (А) Усредненный профиль заряда N-концевых хвостов белков. (Б) Усредненный профиль заряда C-концевых хвостов белков. (В) Суммарный усредненный (для каждого положения а.о.) заряд N- и C-концевых хвостов белков.

#### 5.2.4. Особенности аминокислотного состава белков хроматина

Для более полной характеристики различий между белками хроматина (набор SimChrom) и цитоплазмы (набор CYTLOC\_CS\_UL) был проведен сравнительный анализ аминокислотного состава белков. С использованием метода нелинейного снижения размерности UMAP были выявлены общие закономерности относительного содержания аминокислотных остатков в белках хроматина и цитоплазмы (**Рисунок 39**).



**Рисунок 39** — Сравнительный анализ аминокислотного состава белков хроматина и цитоплазмы. (А) UМАР-проекция белков хроматина и цитоплазмы; подписаны группы белков с отличающимися свойствами, в частности, гистоны типов Н1, Н2А, Н2В и Н3, а также факторы сплайсинга и обратные транскриптазы эндогенных ретровирусов (ОТ ЭРВ). (Б) Те же проекции для белков хроматина, с указанием конститутивных и индуцибельных ТФ с характерными доменами. (В) Сравнительный анализ различий в аминокислотном составе между группами белков, зеленым отмечены статистически значимые различия ( $p$ -значение  $< 0,05$ ).

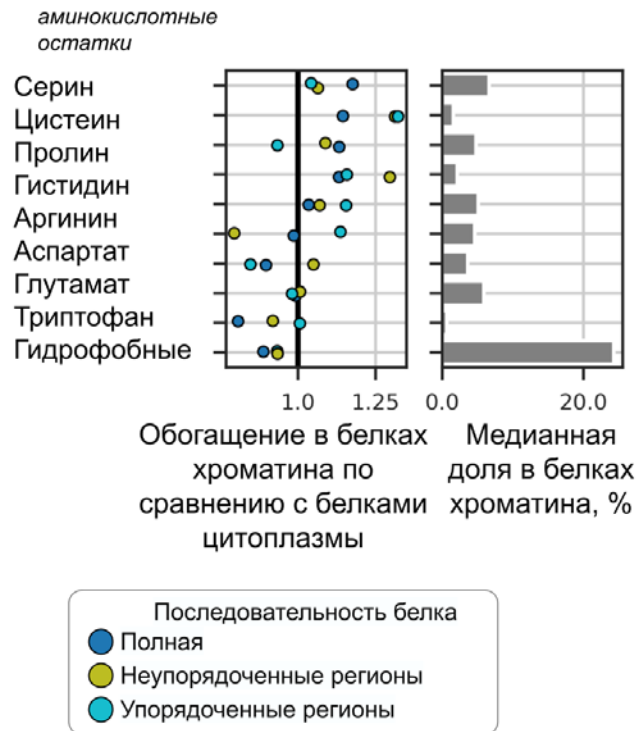
На двумерной UMAP-проекции, где каждая точка соответствует белку, а ее координаты зависят от аминокислотного состава, наблюдается значительное перекрывание облаков точек, соответствующих наборам белков хроматина и цитоплазмы, однако центры распределений смещены относительно друг друга (**Рисунок 39А**). Это свидетельствует о различиях между группами белков по совокупности их физико-химических свойств.

На UMAP-проекциях определены две группы белков, представляющие собой совокупности точек, компактно расположенных в пределах одной области пространства (**Рисунок 39А**). Большинство белков Группы I, 415 из 422, относятся к ДНК-связывающим ТФ, содержащим домен типа цинковый палец (zf-C2H2) (**Рисунок 39А,Б**). Из них 240 белков относятся к конститутивным белкам, 175 — к индуцибельным. Белки Группы I отличаются от белков Группы II повышенным количеством доменов zf-C2H2, медианные значения 10 и 3 соответственно, которое также влияет на статистически значимое увеличение доли остатков цистеина и гистидина в белках Группы I по сравнению с белками Группы II (**Рисунок 39В**). В свою очередь в Группе II сосредоточены ДНК-связывающие ТФ, содержащие домен Homeodomain (**Рисунок 39А,Б**).

Среди отличающихся по аминокислотному составу групп белков выявлены гистоны (обогащены положительно заряженными а.о.); факторы сплайсинга, обогащенные остатками серина и аргинина (англ. serine/arginine-rich splicing factors, SRSFs); обратные транскриптазы эндогенных ретровирусов (англ. reverse transcriptases of endogenous retroviruses, ERVs), обогащенные остатками изолейцина и треонина (**Рисунок 39А,В**).

Для описания различий в аминокислотном составе белков хроматина и цитоплазмы использовали следующие понятия: обогащение — повышенная относительная доля остатка в белках одной группы по сравнению с белками другой группы (расчет проводили с использованием критерия Манна-Уитни с поправкой Бонферрони на множественное тестирование); коэффициент обогащения — отношение медианной доли остатка в белках одной группы к медианной доле этого остатка в белках другой группы.

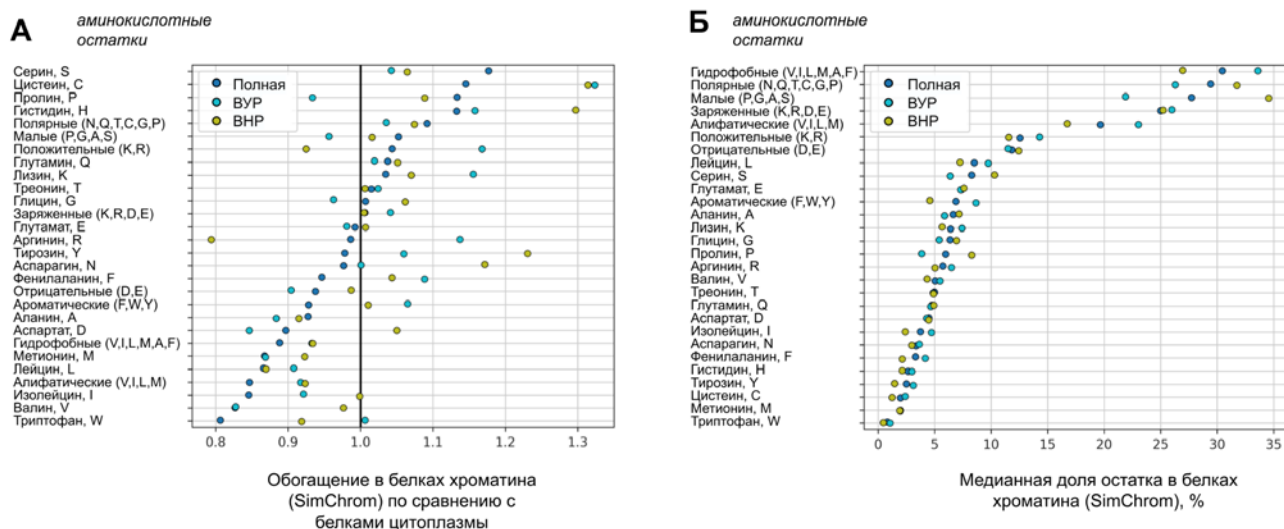
Белки хроматина по сравнению с белками цитоплазмы наиболее статистически значимо обогащены остатками серина, цистеина, пролина и гистидина. Соответствующие коэффициенты обогащения находятся в диапазоне 1,13-1,18 (Рисунок 40).



**Рисунок 40** — Различия в аминокислотном составе белков хроматина по сравнению с белками цитоплазмы. Слева показаны остатки, по которым белки хроматина обогащены или обеднены по сравнению с белками цитоплазмы, с отдельным рассмотрением полной последовательности белков, ВУР и ВНР; справа приведены медианные доли аминокислотных остатков в белках хроматина.

Приведенный выше анализ позволил охарактеризовать белки хроматина в целом. Однако белки, относящиеся к разным функциональным категориям, могут обладать специфическими физико-химическими свойствами. В зависимости от количества белков в составе категории такие особенности могут оставаться невыявленными при рассмотрении совокупности белков. В связи с этим был проведен сравнительный анализ относительного содержания аминокислотных остатков в белках хроматина, относящихся к различным функциональным категориям согласно классификации SimChrom, по сравнению с белками цитоплазмы, а также с разбиением аминокислотных последовательностей белков на внутренне упорядоченные и неупорядоченные регионы (ВУР и ВНР).

соответственно). Обобщенные результаты приведены на **Рисунках 40 и 41**, полные результаты приведены на **Рисунках Б1, Б2, Б3 (Приложение Б)**.



**Рисунок 41** — Аминокислотные остатки, содержание которых повышено или снижено в белках хроматина по сравнению с белками цитоплазмы.

Приведены результаты анализа полных аминокислотных последовательностей, ВНР и ВУР.

Белки хроматина по сравнению с белками цитоплазмы обогащены полярными (N, Q, T, C, G, P), малыми (P, G, A, S) и положительно заряженными (K, R) остатками (**Рисунок 41А**).

Остатки серина и пролина наиболее распространены в белках: медианная доля в белках хроматина — 7-8 %, в белках цитоплазмы — 5-6 % (**Рисунок 41**). Вместе с тем медианные доли остатков серина различаются между ВНР и ВУР: в белках хроматина — 10,3 и 6,3 %, в белках цитоплазмы — 9,6 и 6,1 % соответственно. Из этого следует, что в большей степени повышенное содержание остатков серина в белках хроматина связано с большей неупорядоченностью этих белков (медиана доли неупорядоченных регионов 46 % против 23 % в белках цитоплазмы). Неупорядоченные регионы, в свою очередь, характеризуются более высоким относительным содержанием остатков серина по сравнению с упорядоченными регионами.

Обогащение остатками серина ВНР белков хроматина по сравнению с ВНР белков цитоплазмы отмечено для белков практически всех функциональных категорий SimChrom, за исключением белков HMG и субъединиц РНК-полимераз (**Рисунок Б2**).

В ВУР белков хроматина различных категорий по сравнению с ВУР белков цитоплазмы отмечено как увеличение, так и уменьшение доли остатка серина, а вклад в общее обогащение серином достигается в основном за счет количественного преобладания белков, относящихся к конститутивным транскрипционным факторам (**Рисунок Б3**).

Обогащение остатками пролина белков хроматина также обусловлено их повышенным содержанием в ВНР по сравнению с ВНР белков цитоплазмы и, в большей степени, повышенным содержанием ВНР в белках хроматина (**Рисунок 41, Рисунок Б2**). Коэффициент обогащения остатка пролина в ВНР по сравнению с ВУР составил 1,8 и 2,1 для белков цитоплазмы и хроматина соответственно. При этом в ВУР общее обогащение остатками пролина в белках хроматина по сравнению с белками цитоплазмы не было статистически значимым.

В некоторых малочисленных функциональных категориях белков хроматина были отмечены значительные отклонения в содержании остатка пролина в ВУР, а именно повышенная доля в белках HMG и пониженная доля в гистонах (**Рисунок Б3**).

Обогащение остатками пролина в ВНР наблюдалось во многих категориях белков хроматина (**Рисунок Б2**). Наибольшее обогащение остатками пролина в ВНР было отмечено в категориях индуцибельных и пионерных ТФ (коэффициенты обогащения 1,37 и 1,55 соответственно). В то же время в ВНР конститутивных ТФ содержание остатков пролина было заметно ниже (коэффициент обогащения 0,92). Медианные доли пролина в конститутивных и индуцибельных ТФ составили 7 и 10,4 % соответственно.

Относительное содержание остатков цистеина и гистидина в белках хроматина невысокое, медианные значения составляют 1-2,5 % (**Рисунок 41**). Статистически значимое обогащение белков хроматина этими остатками по

сравнению с белками цитоплазмы полностью обусловлено наличием в транскрипционных факторах ДНК-связывающих доменов типа цинковый палец (zf-C2H2) (**Рисунок Б1**). Исключение этой группы белков из анализа привело к исчезновению выявленной закономерности.

При сравнении доли положительно заряженных остатков в белках хроматина по сравнению с белками цитоплазмы выявлено статистически значимое умеренное обогащение остатками лизина (коэффициент обогащения 1,03), но не аргинина. При этом как ВНР, так и ВУР белков хроматина статистически значимо обогащены остатками лизина (коэфф. обогащ. 1,07 и 1,6 соответственно), в то время как содержание остатка аргинина статистически значимо повышено в ВНР и снижено в ВУР (коэфф. обогащ. 0,79 и 1,14 соответственно) (**Рисунки Б1, Б2, Б3**).

Статистически значимые изменения в содержании отрицательно заряженных остатков в белках хроматина выявлены только для аспартата: содержание остатка снижено в полных последовательностях (коэфф. обогащ. 0,9), повышено в ВУР (коэфф. обогащ. 1,05) и снижено в ВНР (коэфф. обогащ. 0,8) (**Рисунок 40**). Содержание остатка глутамата статистически не значимо снижено в белках хроматина (коэфф. обогащ. 0,9).

Таким образом, более высокий суммарный положительный заряд белков хроматина по сравнению с белками цитоплазмы (61 % против 36 %) (**Рисунок 37**) обусловлен преимущественно повышенным содержанием остатков лизина, пониженным содержанием остатков аргинина в ВУР и пониженным содержанием остатков аспартата в ВНР.

Другим выявленным отличием ВНР белков хроматина по сравнению с ВНР белков цитоплазмы является значительное обогащение этих регионов остатками тирозина и аспарагина (коэфф. обогащ. 1,23 и 1,17 соответственно), при этом основной вклад в обогащение вносят транскрипционные факторы (**Рисунок Б2**).

Содержание остатков триптофана и гидрофобных/алифатических аминокислот, таких как валин, изолейцин, лейцин и метионин, снижено в белках хроматина по сравнению с белками цитоплазмы. Триптофан относится к наименее распространенным в белках остаткам (относительная доля 1 %) (**Рисунок 41Б**) и

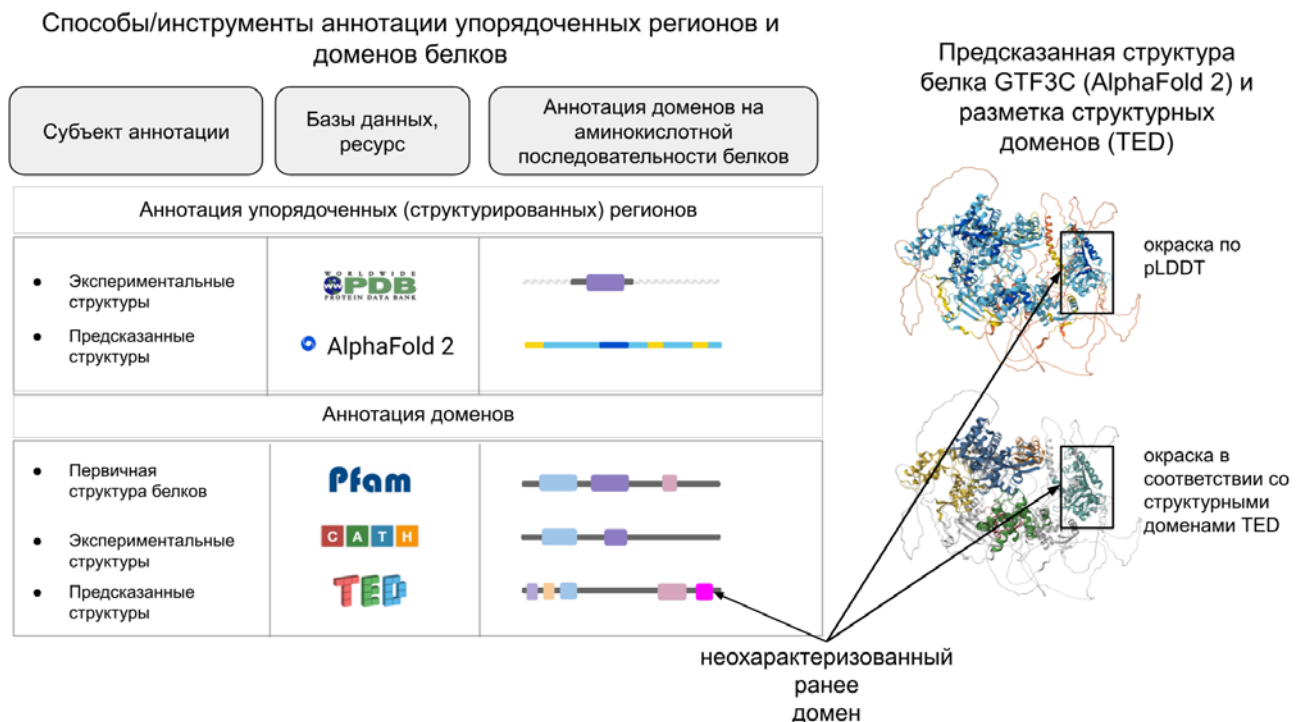
полностью отсутствует в белках некоторых категорий по SimChrom — в гистонах и белках семейства HMG. Однако в некоторых категориях белков хроматина, в частности, в пионерных ТФ; белках, участвующих в (де)метиловании ДНК и нанесении/считывании ПТМ гистонов, наблюдается обогащение остатками триптофана в ВУР соответствующих групп белков по сравнению с белками цитоплазмы (**Рисунок Б1В**). При этом относительная доля гидрофобных/алифатических остатков в ВНР белков хроматина снижена по сравнению с их долей в ВУР этих же белков, что в совокупности с увеличенной долей ВНР в белках хроматина и обуславливает этот эффект (**Рисунок 41Б**).

Приведенные выше особенности распределения аминокислотных остатков верны и для белков с единственной локализацией в ядре (набор белков NULOC\_CS\_UL, **Рисунки Б1, Б2, Б3**), за исключением выраженного обогащения остатками цистеина. Это можно объяснить тем, что большинство транскрипционных факторов, содержащих домен zf-C2H2, согласно аннотациям баз данных UniProt и HPA, характеризуются двойной локализацией (ядро и цитоплазма) и, следовательно, не были включены в набор белков с единственной ядерной локализацией NULOC\_CS\_UL. Действительно, среди конститутивных и индуцибельных транскрипционных факторов только 33 % и 29 % соответственно относились к белкам с единственной локализацией в ядре.

Таким образом, в результате проведенного сравнительного анализа были выявлены следующие особенности аминокислотного состава белков хроматина человека по сравнению с белками цитоплазмы. Обогащение остатками цистеина и гистидина обусловлено присутствием в транскрипционных факторах ДНК-связывающих доменов zf-C2H2. Обогащение остатками серина и пролина связано в основном с увеличенной долей неупорядоченных регионов в этих белках. Обогащение остатками лизина и обеднение остатками аспартата обуславливает высокий суммарный положительный заряд белков хроматина человека.

### 5.3. Структурная организация белков хроматина

Для характеристики структурной организации белков хроматина человека был проведен систематический анализ аннотаций упорядоченных (структурированных) регионов в белках (на основе результатов из **раздела 5.2.2**) и состава доменов — функциональных доменных моделей Pfam (на основании аннотации аминокислотных последовательностей белков согласно базе данных Pfam), а также структурных доменов (согласно аннотации структур белков из базы данных CATH, далее CATH-домены, и аннотации предсказанных структурных доменов из базы данных TED, далее TED-домены) (**Рисунок 42**).

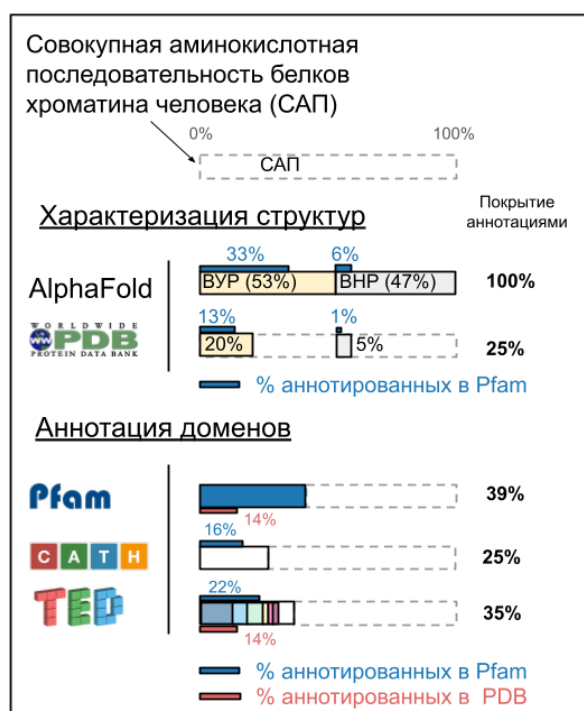


**Рисунок 42** — Источники информации об аннотации упорядоченных регионов в белках (на основе анализа структур, предсказанных методом AlphaFold2, и экспериментально полученных структур белков, депонированных в PDB), а также аннотации доменов на основе аминокислотных последовательностей белков (Pfam-домены), экспериментальных структур (CATH-домены) и предсказанных структур белков (TED-домены). Справа показан пример выявленного в исследовании ранее неохарактеризованного структурного домена в белке GTF3C. Структура белка, предсказанная AlphaFold2, окрашена в соответствии с метрикой уверенности модели pLDDT и разметкой структурных доменов TED.

### 5.3.1. Структурная характеристика белков хроматина

Прежде всего мы охарактеризовали белки хроматина (набор белков SimChrom) в соответствии с аннотациями упорядоченных регионов и доменов согласно различным ресурсам. Анализ был проведен по объединенной (конкатенированной) совокупной аминокислотной последовательности белков хроматина человека (далее САП), принимаемой за 100 %.

Доля упорядоченных регионов в белках хроматина человека, согласно анализу структур, предсказанных методом AlphaFold2, составляет 53 % САП. При этом только 35 % САП отнесено к предсказанным структурным доменам (TED-домены), а экспериментальные структуры по данным PDB доступны только 25 % САП (**Рисунок 43**). Таким образом, консервативная оценка доли САП, доступной для структурных исследований методами структурной биологии и биофизики, составляет 35 %.



**Рисунок 43** — Анализ аннотаций совокупной аминокислотной последовательности белков хроматина человека по экспериментально полученным структурам (база данных PDB), упорядоченным и неупорядоченным регионам белков согласно предсказаниям AlphaFold2, а также по аннотациям доменов — структурных (CATH, TED) и функциональных (Pfam). Для всех аннотаций отражена соответствующая доля САП.

Различия в аннотациях упорядоченных регионов при сравнении предсказанных структур (AlphaFold2) и разметки TED-доменов, предположительно, обусловлены следующими факторами:

1) склонностью AlphaFold2 к предсказанию длинных одиночных спиралей [213], которые могли быть не соотнесены со структурными доменами согласно алгоритмам, реализованным в TED;

2) неспособностью алгоритмов TED в ряде случаев выявить повторяющиеся фрагменты, содержащие визуально идентифицируемые элементы вторичной структуры в составе больших многодоменных белков;

3) в данной работе к ВУР или ВНР были отнесены регионы белков длиной не менее 4 а.о., тогда как медианное значение длины домена в белках человека согласно аннотации TED составляет 108 а.о.

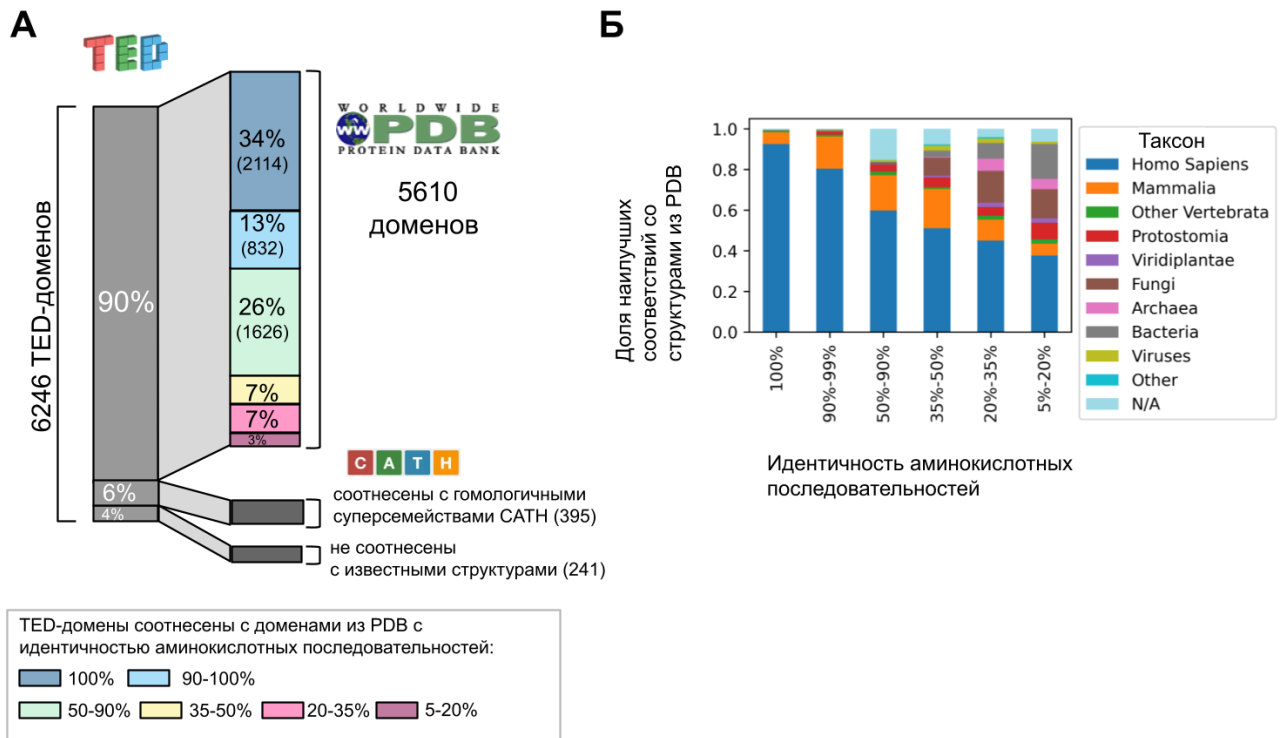
Также следует учитывать, что аннотации TED-доменов получены для одиночных белков без учета мультидоменных белковых комплексов. Примером описанных выше различий служат 9 % САП, для которой отсутствует аннотация структурных TED-доменов, однако соответствующие остатки разрешены в базе данных PDB (**Рисунок 43**).

В некоторой степени аннотации доменов по базам данных Pfam, CATH и TED являются взаимодополняющими: к 39 % САП, аннотированной функциональными Pfam-доменами, для 13 % САП доступны аннотации TED-доменов и еще для 3 % САП — CATH-домены. Таким образом, аннотации функциональных и структурных доменов доступны для 55 % от аминокислотной последовательностей белков хроматина человека.

### **5.3.2. Структурные домены белков хроматина**

Был проведен детальный анализ аннотации САП структурными TED-доменами (**Рисунок 44**). Для идентифицированных в белках хроматина человека 6246 отдельных TED-доменов было выполнено соотнесение их структур со структурами из базы данных PDB и с гомологичными суперсемействами доменов

(согласно классификации CATH) методом FoldSeek с последующим отбором наилучшего соответствия. Домены, для которых не удалось установить такого соотнесения, были проанализированы с целью возможного отнесения к новым структурным суперсемействам, а также оценки их потенциальной функциональной роли.



**Рисунок 44** — Анализ структурных TED-доменов белков хроматина человека. (А) Найденные структурные соответствия для TED-доменов в базах данных PDB и CATH с различной степенью идентичности аминокислотных последовательностей. (Б) Таксономическое распределение организмов — источников структурных соответствий для TED-доменов по данным базы данных PDB.

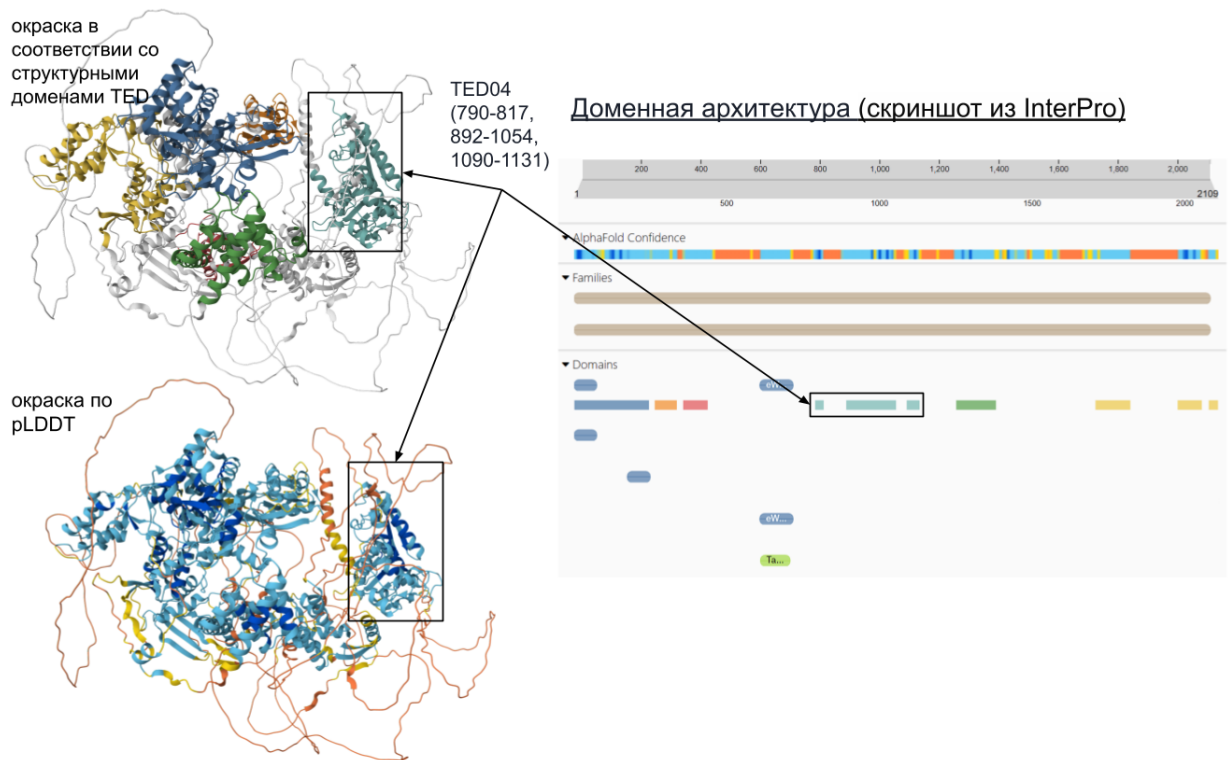
Среди 6246 TED-доменов 34 % имели точные соответствия со структурами из PDB (со 100 % идентичностью аминокислотных последовательностей); 56 % соответствовали структурам из PDB с меньшей степенью идентичности последовательностей (от 99 до 5 %) (Рисунок 44А). Большинство соотнесенных гомологичных доменов представляли собой паралогичные последовательности белков человека (Рисунок 44Б). Среди сопоставленных доменов с идентичностью аминокислотных последовательностей в диапазоне 35-50 % доля белков человека

составила 51 %, доля белков других млекопитающих — 20 %. Среди доменов с более низкой идентичностью последовательностей (менее 35 %) структурные соответствия были представлены в основном структурами белков грибов, первичноротых и бактерий (**Рисунок 44Б**).

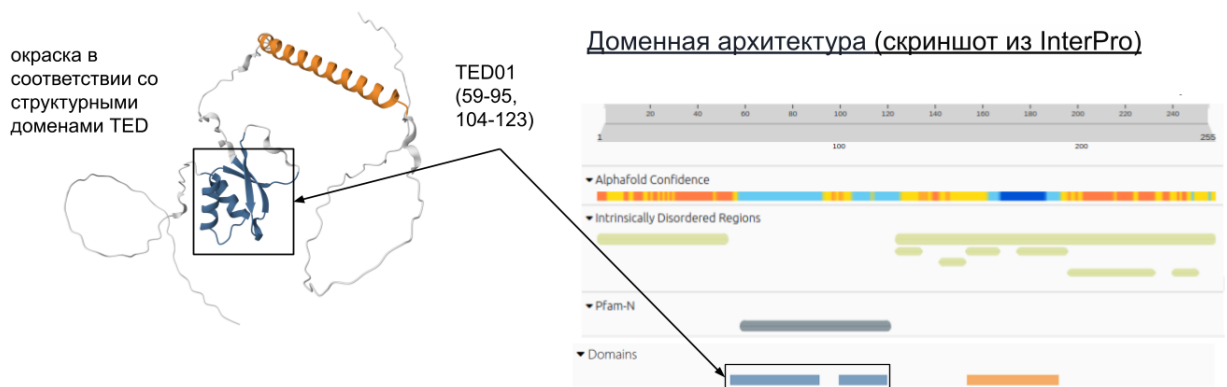
Для 395 TED-доменов (6 %) без структурных соответствий в PDB были получены структурные соответствия с гомологичными суперсемействами из базы данных CATH. Таким образом, учет информации о потенциальных структурных вариациях в представителях гомологичных суперсемейств CATH совместно с предсказанными AlphaFold2 структурами позволил выявить наиболее отдаленных структурно охарактеризованных гомологов.

Остальные 241 TED-домен (4 %) не удалось соотнести со структурами в PDB или со структурными суперсемействами по классификации CATH. Эти домены потенциально могут являться новыми типами гомологичных суперсемейств. Список таких доменов доступен на странице веб-сервиса SimChrom ([https://simchrom.intbio.org/#novel\\_structural\\_domains](https://simchrom.intbio.org/#novel_structural_domains)). Примерно половина из этих доменов (123 домена) имеет аннотации по Pfam или другим базам данных, представленным в ресурсе InterPro. В свою очередь для 118 доменов, относящихся к 106 белкам хроматина человека, на момент анализа отсутствовали какие-либо аннотации. Они являются перспективными объектами для исследований методами структурной биологии и биофизики. Среди таких белков GTF3C1 (General Transcription Factor III Subunit 1), в состав которого входит ранее не аннотированный структурный домен длиной 233 а.о. (**Рисунок 45А**), а также глобулярный домен линкерного гистона H1.7 (**Рисунок 45Б**).

**А** Общий фактор транскрипции (gene *GTF3C1*, protein Q12789)



**Б** Линкерный гистон H1.7 (gene *H1-7*, protein Q75WM6)



**Рисунок 45** — Примеры идентифицированных структурных доменов в белках хроматина человека: **(А)** общий фактор транскрипции GTF3C1, **(Б)** линкерный гистон H1.7. Слева показаны предсказанные методом AlphaFold2 структуры белков, окрашенные в соответствии со степенью уверенности модели (pLDDT) и разметкой TED-доменов. Справа приведены скриншоты из базы данных InterPro, рамкой выделены структурные TED-домены.

Гистон H1.7 (предыдущий идентификатор HANP1/H1T2) специфичен для тканей семенников. В предсказанной структуре этого гистона глобулярный домен

отличается по топологии от доменов других гистонов типа H1: часть глобулярного домена, называемая «крыло», состоит из двух  $\beta$ -листов, а не из трех. Несмотря на значительное количество исследований, посвященных структуре линкерных гистонов [214], данный факт в полной мере не обсуждался в научной литературе.

Из-за аминокислотных различий между последовательностями гистона H1.7 и другими линкерными гистонами ранее этот домен не удавалось идентифицировать классическими методами анализа последовательностей белков, реализованных в базе данных Pfam. Это удалось преодолеть с развитием принципиально новых подходов к аннотированию доменов, основанных на методах глубокого обучения (таких как Pfam-N) (**Рисунок 45Б**).

Для 118 TED-доменов, не имеющих структурных соответствий (в PDB и CATH) и каких-либо аннотаций (в InterPro), были предсказаны термины Gene Ontology молекулярных функций и биологических процессов с помощью графовой сверточной нейронной сети DeepFRI на основе аминокислотных последовательностей и предсказанных структур.

Предсказанные GO-термины носили довольно общий характер. Наиболее распространенными GO-терминами молекулярных функций были: связывание с ионами, связывание с органическими циклическими соединениями, связывание с гетероциклическими соединениями, связывание с белками, связывание с катионами, связывание с ионами металлов, связывание с нуклеиновыми кислотами. Наибольшее количество доменов было аннотировано следующими GO-терминами биологических процессов: процесс метаболизма органических веществ, первичный метаболический процесс, клеточный метаболический процесс, процесс метаболизма азотистых соединений, процесс метаболизма макромолекул, процесс метаболизма азоторганических соединений, регуляция клеточных процессов, процесс метаболизма клеточных макромолекул, процесс метаболизма клеточных азотистых соединений, реакция клеток на раздражители.

Для девяти из 118 TED-доменов с помощью модели DeepFRI не удалось предсказать термины GO. Два из них присутствовали в белковых продуктах

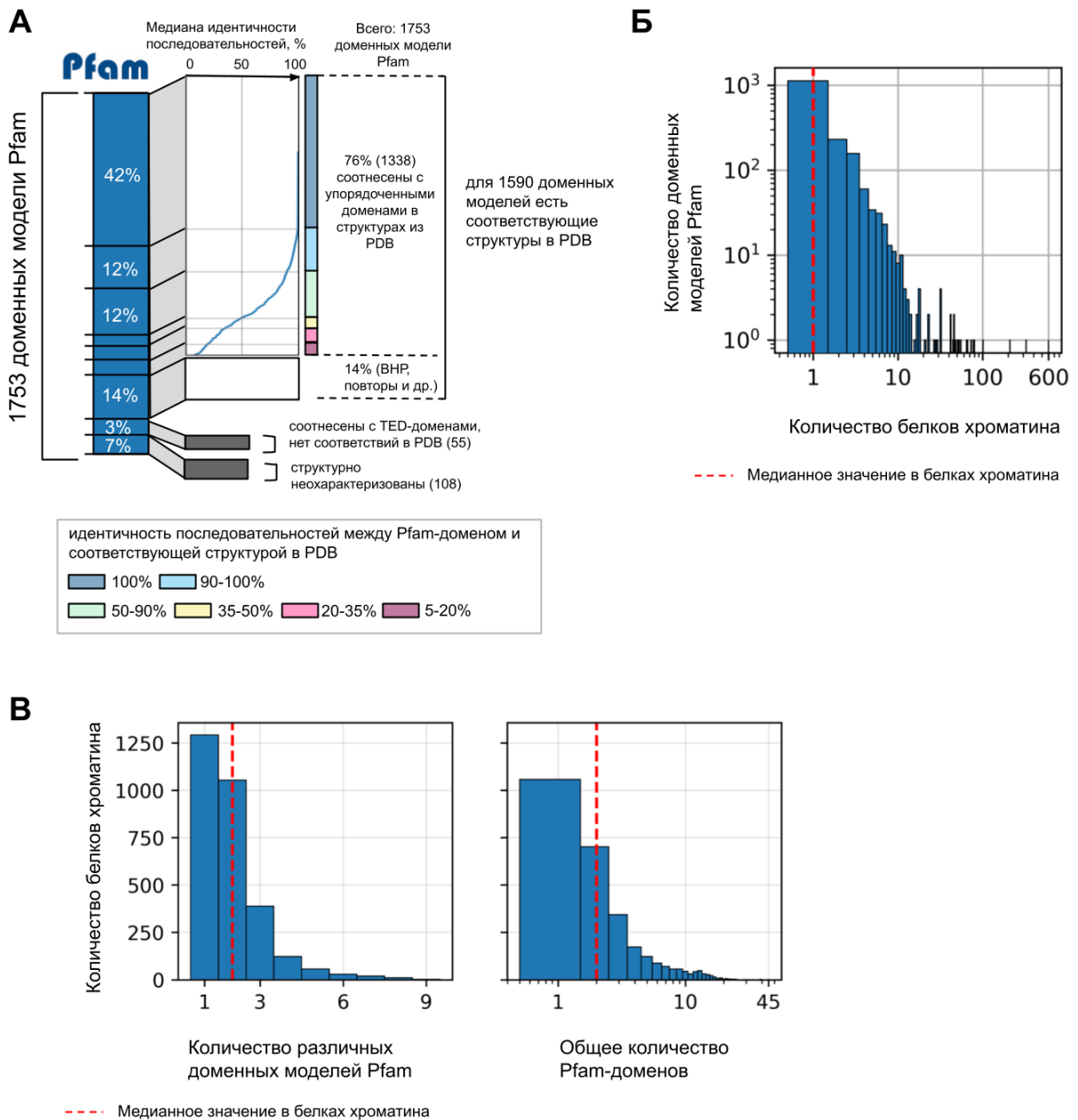
представителей семейства генов регуляторного фактора X (англ. regulatory factor X family), кодируемых генами *RFX1*, *RFX5*.

Таким образом, применение методов искусственного интеллекта для предсказания структур расширяет возможности аннотации функциональных и структурных доменов, выявления наиболее отдаленных структурных гомологов, а также углубляет представление о структурной организации белков хроматина, в частности в задачах идентификации ранее не охарактеризованных структурных доменов.

### 5.3.3. Разнообразие функциональных доменов в белках хроматина

Для исследования доменной архитектуры и разнообразия доменов в белках хроматина человека (набор белков SimChrom) были выбраны аннотации доменных моделей из базы данных Pfam (объединяющие сходные эволюционно родственные домены). В белках хроматина человека идентифицировано 11147 отдельных Pfam-домов, соответствующих 1753 доменным моделям. Для 70 белков хроматина аннотации доменов по Pfam отсутствовали.

Для оценки структурной охарактеризованности доменных моделей Pfam проведено их структурное соотнесение с доменами из баз PDB и TED методом FoldSeek (**Рисунок 46А**). Для каждой доменной модели рассчитана медиана идентичности аминокислотных последовательностей с наилучшими структурными соответствиями из PDB. В результате выявлено, что для 76 % доменных моделей Pfam по крайней мере один индивидуальный домен может быть сопоставлен со структурой из PDB (**Рисунок 46А**). Полностью охарактеризованными (100 % идентичность последовательностей Pfam-домов и структуры из PDB) признаны 42 % доменных моделей. Для 34 % доменных моделей характерна частичная охарактеризованность — медианная идентичность Pfam-домов и структуры из PDB находится в диапазоне от 50 % до 99 % (**Рисунок 46А**).



**Рисунок 46** — Структурная характеристика аннотаций доменных моделей по Pfam в белках хроматина человека. **(А)** Анализ структурных соответствий для доменных моделей Pfam в базах данных PDB и TED. **(Б)** Гистограмма отображает количество доменных моделей Pfam (ось Y), точное количество которых входит в состав белков хроматина (ось X). **(В)** Распределение белков хроматина по количеству различных доменных моделей Pfam (слева) и по общему количеству Pfam-доменов (справа).

Для 252 доменных моделей Pfam (14 %) не были обнаружены соответствия структурам PDB методом FoldSeek, однако такие домены были идентифицированы в PDB методами поиска по последовательности (**Рисунок 46А**). Эти домены преимущественно соответствовали ВНР, гибким доменам с ВНР, повторам,

коротким ДНК-связывающим мотивам и мотивами типа «спиральная катушка» (англ. coiled-coil); к последним относились 34 доменные модели Pfam.

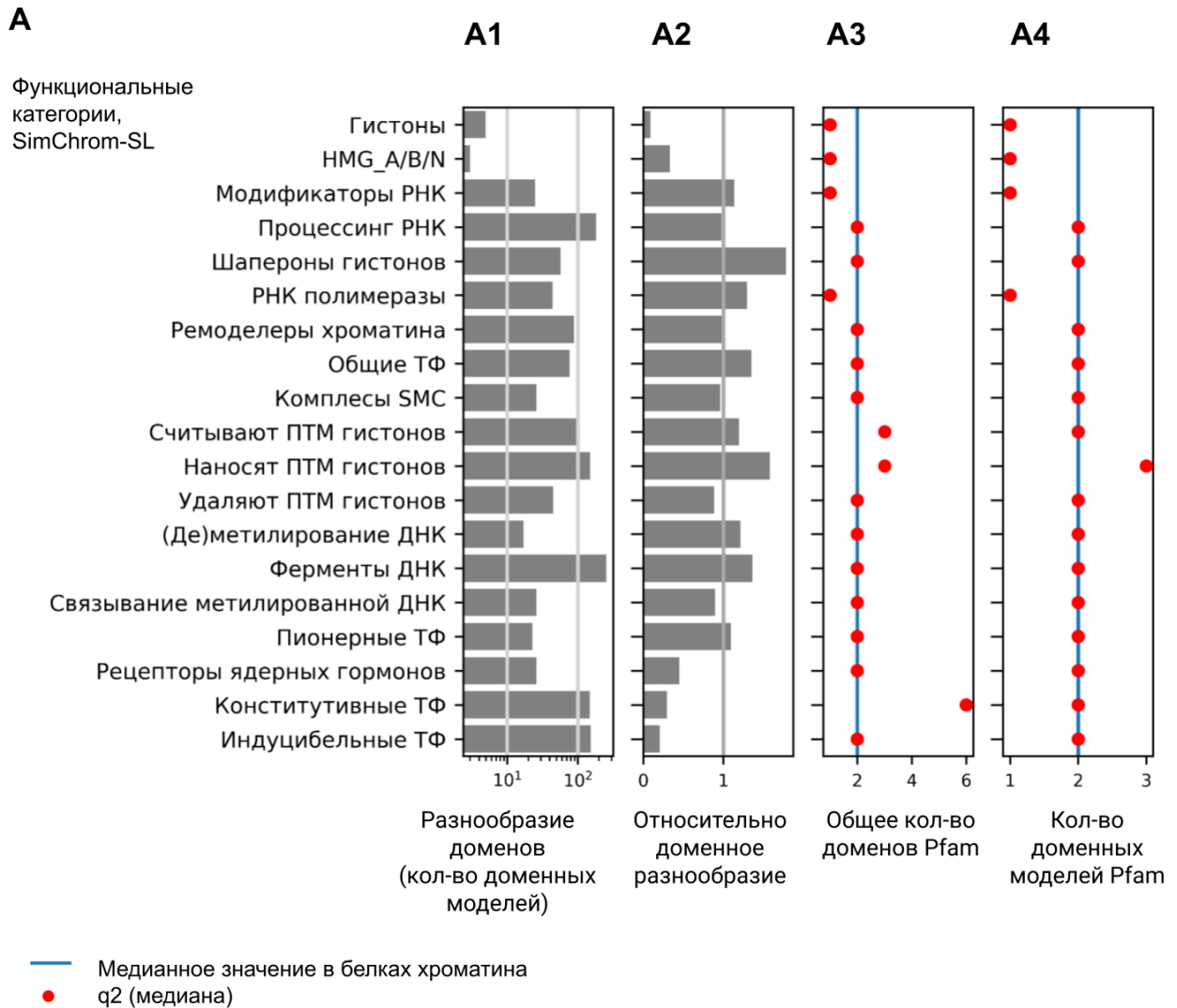
Для 55 доменных моделей Pfam (3 %) не выявлено структурных соответствий в PDB, однако они были сопоставлены с предсказанными структурными доменами из базы данных TED (**Рисунок 46А**). Эти домены представляют собой перспективные объекты для валидации и дальнейшего исследования методами структурной биологии и биофизики. К ним относятся домены, потенциально связанные с ремоделированием хроматина (SANTA, zf-C3Hc3H), нанесением ПТМ гистонов (DUF7030, COMPASS-Shg1), а также различные типы цинковых пальцев (zf\_CCCH\_4, zf-LITAF-like, zf-WIZ, SWIM).

Наконец, для 108 доменных моделей Pfam (7 %) отсутствовала структурная информация, которая могла бы быть получена путем поиска структурных доменов в базах данных PDB и TED по аминокислотной последовательности.

Далее был проведен анализ количества Pfam-доменов и доменных моделей в белках хроматина. Выявлено, что в общем случае в белке хроматина находится два Pfam-домена, относящихся к разным доменным моделям. Распределение белков хроматина по числу доменных моделей Pfam приведено на **Рисунке 46Б**. При этом большинство доменных моделей Pfam присутствует только в одном белке хроматина (**Рисунок 46В**), за исключением, например, доменов типа цинковый палец (zf-C2H2).

Для белков, вовлеченных в регуляцию эпигенетических процессов (ремоделеры хроматина, шапероны гистонов, белки ПТМ машинерии гистонов), отмечено увеличенное число доменных моделей Pfam. Так, 16 % таких белков содержат три доменные модели Pfam, 10 % — четыре, 7 % — пять, и 10 % — от шести до четырнадцати.

Далее был проведен анализ разнообразия Pfam-доменов и доменных моделей в белках различных функциональных категорий SimChrom-SL (**Рисунок 47**).



**Рисунок 47** — Разнообразие Pfam-доменов в функциональных категориях белков хроматина, где A1 — суммарное количество доменных моделей Pfam в белках категории, A2 — относительное доменное разнообразие; A3 — медианное значение количества доменов в белках категории; A4 — медианное значение количества различных доменных моделей Pfam в белках категорий хроматина.

Ожидается, что наиболее крупные по числу белков категории SimChrom-SL (более 100 белков) содержали наибольшее число различных доменных моделей, (**Рисунок 47A1**). В связи с этим для каждой категории белков хроматина была рассчитана оценка относительного доменного разнообразия, определяемая как число различных доменных моделей, приходящееся на один белок (**Рисунок 47A2**). Значение относительного доменного разнообразия для совокупности белков хроматина составляет один. Наиболее низкое доменное разнообразие выявлено в следующих категориях:

- 1) транскрипционные факторы — функциональная вариабельность белков в значительной степени обусловлена различными комбинациями доменов типа цинковый палец, описываемых ограниченным числом моделей Pfam;
- 2) гистоны — функциональная вариабельность белков часто определяется небольшими изменениями аминокислотной последовательности;
- 3) белки HMG — малочисленная группа, включающая девять белков и три модели Pfam.

Медианное значение количества Pfam-доменов в белках хроматина человека составляет два, что согласуется с полученными ранее оценками на основе анализа предсказанных структур (**Раздел 5.2.2, Рисунок 47А3**). В некоторых категориях белков (транскрипционные факторы, ПТМ машинерия гистонов) выявлено увеличенное число Pfam-доменов. Наличие дополнительных доменов у белков, наносящих и считывающих ПТМ гистонов, вероятно, эволюционно обусловлено функциональной необходимостью мультивалентного взаимодействия с различными компонентами хроматина.

К категориям хроматина, в составе которых преимущественно однодоменные белки, относятся: «Белки, ассоциированные с центромерами», «Репарация ДНК», «Регуляция транскрипции», «РНК-полимеразы» (при этом в белках данной категории может быть до 42 доменов), а также белки категорий «Рекомбинация ДНК», «Модификации РНК», «Гистоны» и белки «HMG A/B/N».

Наибольшее медианное значение количества Pfam-доменов, три, выявлено в белках, наносящих ПТМ гистонов (**Рисунок 47А4**). Тем не менее многие другие белки хроматина содержат до 8-9 различных доменных моделей. Такие белки относятся к категориям «Ферменты, действующие на ДНК», «Белки, наносящие ПТМ гистонов», «Шапероны гистонов», «Ремоделеры хроматина», «Транскрипционные факторы». В совокупности 118 белков хроматина содержат пять и более различных доменных моделей (**Рисунок 46В**). Это подчёркивает мультивалентность взаимодействий белков хроматина, учитывая, что многие из них дополнительно формируют белок-белковые комплексы, увеличивая тем самым потенциал взаимодействий.

Для некоторых категорий белков хроматина выявлено наличие общих доменных моделей (**Рисунок 48**). В частности, такие модели были обнаружены:

- 1) в транскрипционных факторах и белках, связывающих метилированную ДНК (домены zf-C2H2, Homeodomain, KRAB);
- 2) в категориях белков, осуществляющих эпигенетическую регуляцию, включая машинерию ПТМ гистонов и ремоделеры хроматина (домены Chromo, Bromodomain, PHD).

## **5.4. Мультивалентные взаимодействия белков хроматина**

### **5.4.1. Определение доменов, вовлеченных в мультивалентные взаимодействия**

В связи с выявленными особенностями многодоменного состава белков хроматина (см. **Раздел 5.3**), был проведен анализ доменов, вовлеченных в мультивалентные взаимодействия (далее ВМВВ-домены). Для ограничения размера анализируемой выборки доменов были отобраны только те, которые присутствуют в нескольких копиях или в комбинации с другими доменами как минимум в трех белках из группы белков, регулирующих эпигенетические процессы (шапероны гистонов, ремоделеры хроматина, машинерия ПТМ гистонов, модификаторы ДНК/РНК). Таким критериям удовлетворяли 94 доменные модели Pfam, из которых 59 были вручную классифицированы по функциям на основе анализа литературы. Для классификации доменов использовали следующие группы:

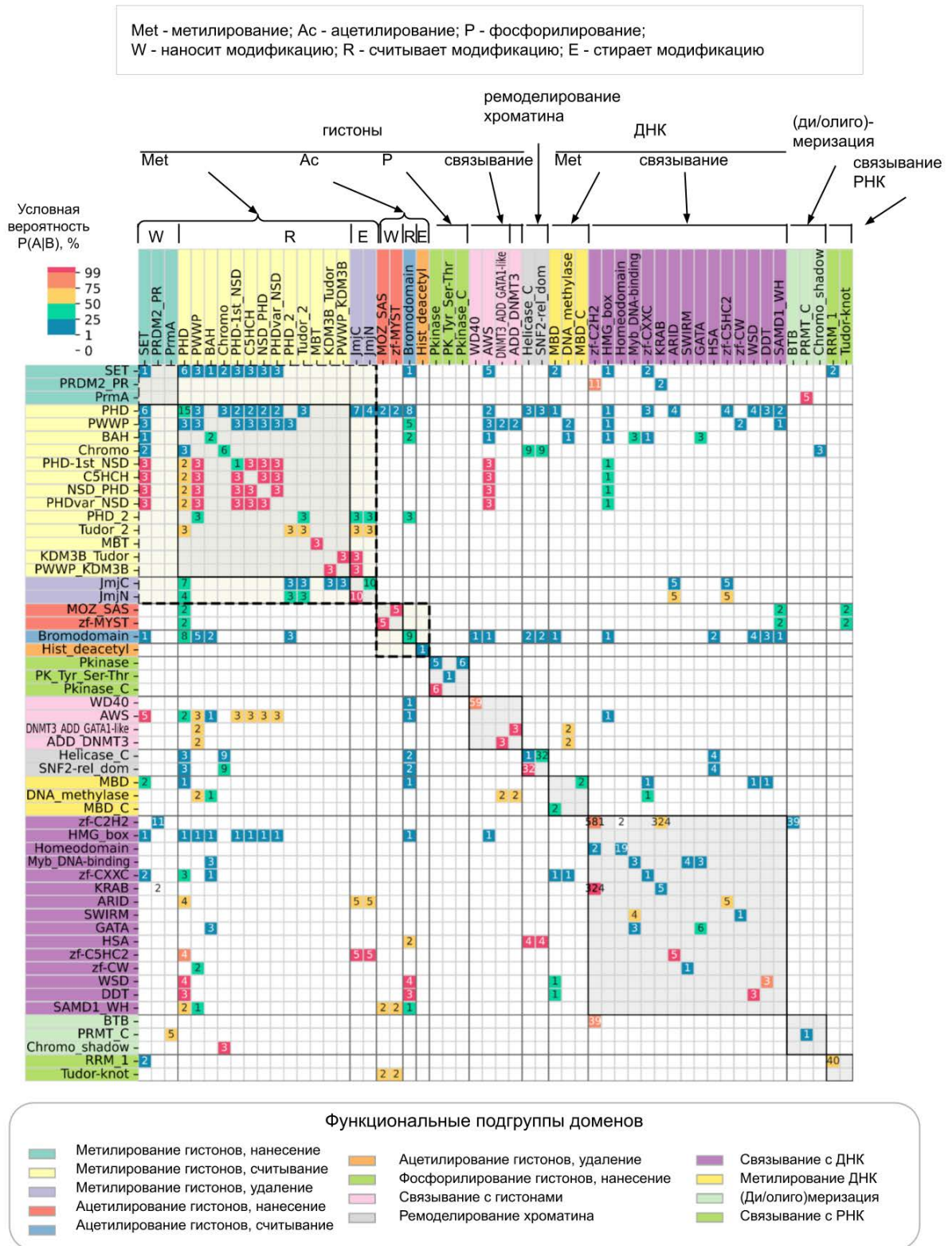
- метилирование/ацетилирование/фосфорилирование гистонов;
- ремоделирование хроматина;
- связывание с гистонами;
- связывание с ДНК;
- связывание с РНК;
- метилирование ДНК;
- димеризация/олигомеризация белков;
- белок-белковые взаимодействия.

Внутри групп доменов, ассоциированных с ПТМ гистонов, было введено дополнительное деление на подгруппы: наносящие, считывающие или удаляющие модификации. Список доменов и их подробная классификация представлены на **Рисунке 49**. Выделенное подмножество доменов считается репрезентативным для иллюстрации концепции мультивалентных взаимодействий белков хроматина, так как отобранные доменные модели в достаточной мере охарактеризованы и их функции известны.

Отобранные ВМВВ-домены (59 доменных модели) присутствуют в 992 белках хроматина (без учета транскрипционных факторов — в 306 белках). Количество белков, содержащих два и более таких домена, составляет 851 (262 белка без учета транскрипционных факторов). Наиболее распространенными доменами в белках хроматина (то есть присутствующие в наибольшем количестве белков), являются домены, ассоциированные с метилированием и ацетилизацией гистонов, связыванием ДНК, фосфорилированием белков (в частности гистонов) и ремоделированием хроматина. Наибольшее разнообразие доменных моделей выявлено среди групп доменов, ассоциированных с метилированием гистонов и связыванием ДНК, что позволяет предположить, что эти механизмы наиболее широко представлены в хроматине и играют важную роль в его регуляции (**Рисунок 49**).

#### 5.4.2. Парные сочетания функциональных доменов в белках хроматина

Далее был проведен анализ парных сочетаний доменов (их ко-встречаемость) в составе белка. Ко-встречаемость двух доменных моделей в белке оценивали как условную вероятность присутствия в белке одной доменной модели (А) при наличии второй (В) (**Рисунок 49**).



**Рисунок 49** — Парные сочетания доменов (доменных моделей по аннотации Pfam) в белках хроматина человека. Показаны домены, присутствующие в белках, регулирующих эпигенетические процессы. Домены сгруппированы по функциям.

Значения в ячейках отражают количество белков хроматина, в которых данные домены сочетаются, цвет ячейки указывает на условную вероятность присутствия в белке хроматина домена А (в столбце) при наличии домена В (в строке).

Интерактивная версия рисунка: [https://simchrom.intbio.org/#domain\\_co-occurrence](https://simchrom.intbio.org/#domain_co-occurrence).

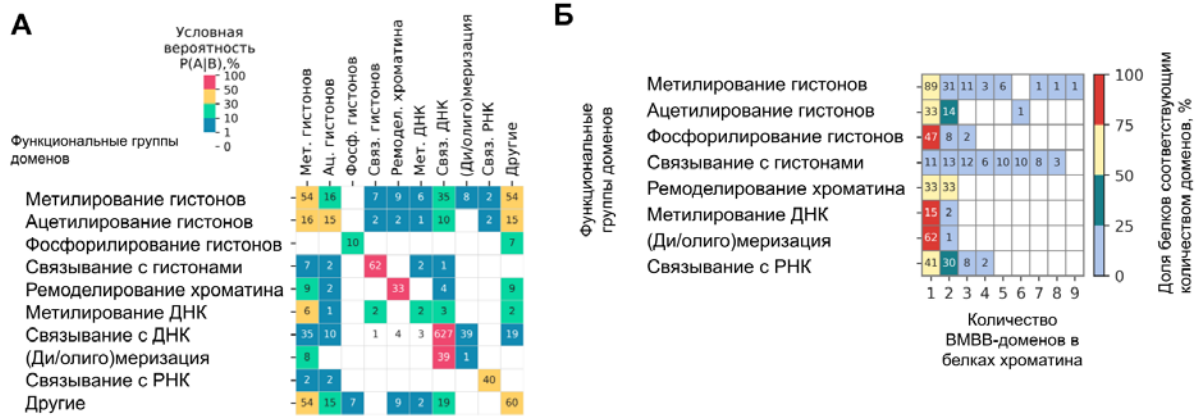
Для 49 пар BMBB-доменов выявлена 100 % ко-встречаемость, из которых в 18 случаях (9 пар доменов) ко-встречаемость является «симметричной», то есть  $P(A|B) = P(B|A)$ . Для некоторых пар такая «обязательная» ко-встречаемость объясняется тем, что домены посредством структурных взаимодействий образуют структуры более высокого порядка (выявлено в результате визуального анализа моделей белков, предсказанных AlphaFold2). В частности, к таким парам относятся MOZ\_SAS и zf-MYST, ADD\_DNMT3 и DNMT3\_ADD\_GATA1-like. В других случаях причиной «обязательной» ко-встречаемости может быть общая функциональность: большинство таких доменов осуществляют считывание метильных меток гистонов и находятся в метилтрансферазах гистонов (KDM3B\_Tudor и PWWP\_KDM3B; C5HCH, NSD\_PHD, PHD-1st\_NSD и PHDvar\_NSD).

Среди BMBB-доменов, которые присутствуют в белках хроматина совместно с большим количеством других доменных моделей, выявлены следующие: PHD (45 совместно присутствующих доменных моделей), Bromodomain (38), SET (40) и PWWP (28), а также АТФазные субъединицы ремоделеров хроматина Helicase\_C (33) и SNF2-rel\_dom (31) (**Рисунок 49**).

Ряд доменов присутствует в белках хроматина в нескольких копиях (диагональные элементы на **Рисунке 49**). Особенно это характерно для доменов MBT, WD40 и цинковых пальцев (zf-C2H2), которые представляют собой короткие повторы, необходимые для формирования функциональной структуры. Домены PHD и Bromodomain также представлены в нескольких копиях в 15 и 8 белках хроматина соответственно.

Более общее представление о стратегии мультивалентных взаимодействий в белках хроматина было получено путем анализа сочетаний доменов из разных функциональных групп (**Рисунок 50А**). Выявлено, что домены, ассоциированные с метилированием гистонов (наносят, считывают или удаляют ПТМ), как правило, совместно присутствуют в белках хроматина: условная вероятность наличия другого домена из этой функциональной группы составляет 38 %. Вместе с тем, домены, ассоциированные с метилированием гистонов, совместно присутствуют в

белках хроматина с доменами практически из всех других функциональных групп за исключением доменов, ассоциированных с фосфорилированием белков (и в частности гистонов) (**Рисунок 50А**). В белках хроматина может присутствовать до 9 доменов, ассоциированных с метилированием гистонов (**Рисунок 50Б**).



**Рисунок 50** — Парные сочетания функциональных групп доменов (доменных моделей по аннотации Pfam) в белках хроматина человека. (**А**) Парные сочетания доменов, объединенных в функциональные группы, в белках хроматина. Значения в ячейках отражают количество соответствующих белков хроматина; цвет ячеек отражает условную вероятность присутствия в белке домена одной функциональной группы А (в столбце) при наличии домена из другой В (в строке). (**Б**) Количество доменов из одной функциональной группы в белках хроматина. Значения в ячейках отражают количество белков хроматина, в состав которых входят BMBB-домены данной функциональной группы; цвет ячеек отражает долю белков с соответствующим количеством доменов.

Для доменов, ассоциированных с ацетилированием и фосфорилированием гистонов, условные вероятности присутствия в белке другого домена из этой же функциональной группы составили 31 % и 18 % соответственно.

Выявлены следующие закономерности сочетания доменов из различных функциональных групп в белках хроматина. Домены, ассоциированные с метилированием гистонов, присутствуют в белках хроматина в сочетании с большим количеством других BMBB-доменов, среди которых домены из групп: связывание с ДНК (35 белков хроматина), ацетилирование гистонов (16), ремоделирование хроматина (9), (ди/олиго)меризация (8), связывание с гистонами

(7), метилирование ДНК (6), связывание с РНК (2) (**Рисунок 50А**). Домены, ассоциированные с ацетилизацией гистонов, отмечены в комбинациях с доменами тех же функциональных групп, за исключением доменов (ди/олиго)меризации и встречаются в меньшем количестве белков хроматина.

Количество белков с комбинациями ДНК-связывающих доменов с доменами других функциональностей является невысоким по сравнению с количеством белков, главным образом транскрипционных факторов, в которых присутствуют исключительно ДНК-связывающие домены (**Рисунок 50А**).

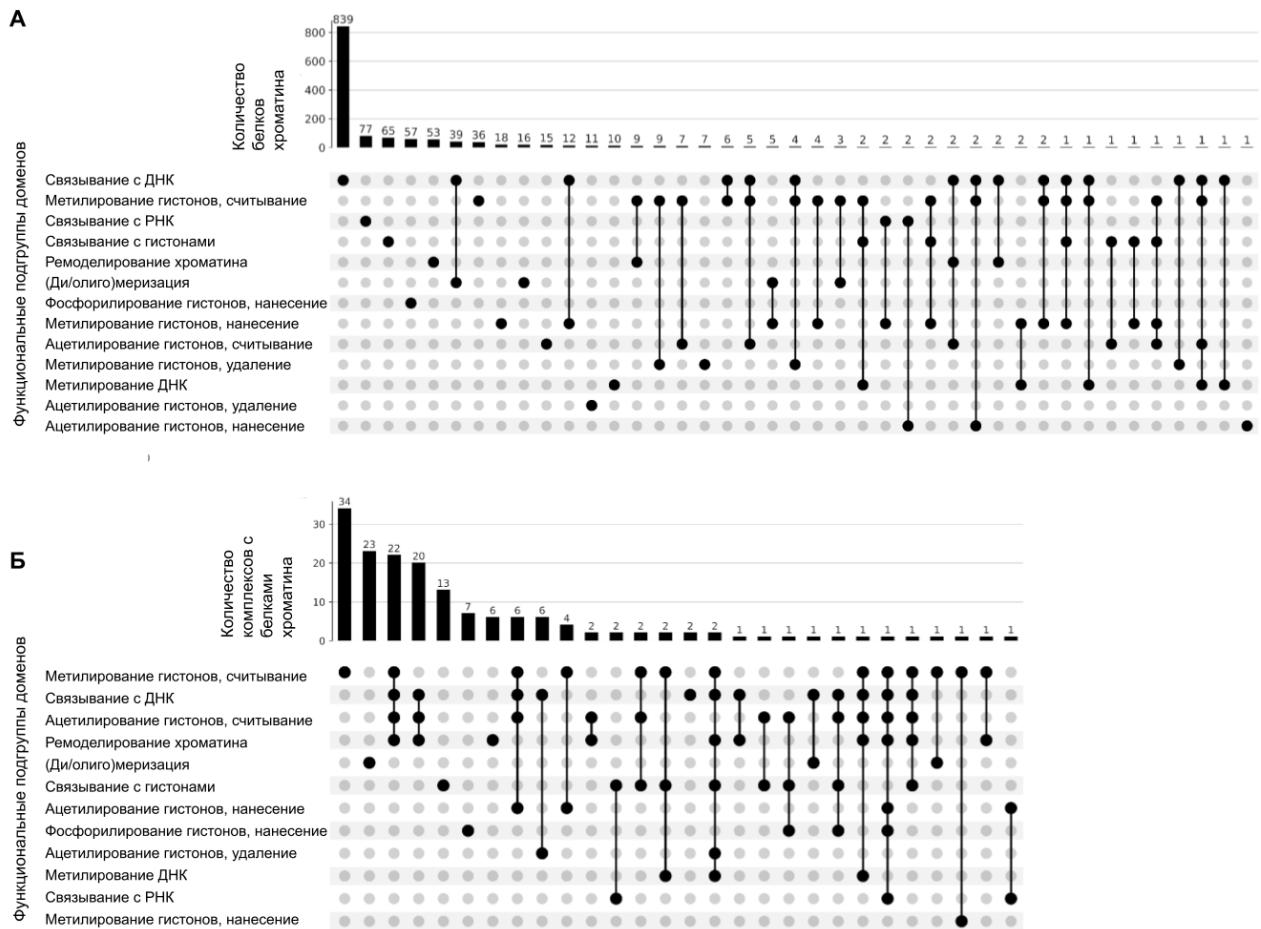
Домены, представляющие собой АТФазные субъединицы ремоделеров хроматина, выявлены в комбинациях с доменами из групп: метилирование гистонов (9 белков), ДНК-связывающие домены (4), ацетилизацию гистонов (2) (**Рисунок 50А**). При этом домены, ассоциированные с ПТМ гистонов, представлены только доменами, узнающими метилированные (PHD, Chromo) или ацетилированные (Bromodomain) гистоны.

Примечательно, что в проведенном анализе не были выявлены случаи ко-встречаемости доменов фосфорилирования белков (в частности, гистонов) с доменами других функциональных групп (**Рисунок 50А**). Это может свидетельствовать об эволюционной стратегии, в рамках которой домены, ассоциированные с метилированием и ацетилизацией гистонов, эволюционировали согласованно в составе различных сочетаний, обеспечивающих тонкую регуляцию экспрессии генов на эпигенетическом уровне, в то время как фосфорилирование белков оставалось более общим и универсальным механизмом, задействованным в широком спектре клеточных сигнальных путей.

#### 5.4.3. Сочетания функциональных доменов в белках хроматина

Для формирования более целостного представления о стратегиях мультивалентного взаимодействия в белках хроматина в дополнение к анализу парной ко-встречаемости доменов, был проведен анализ всех возможных

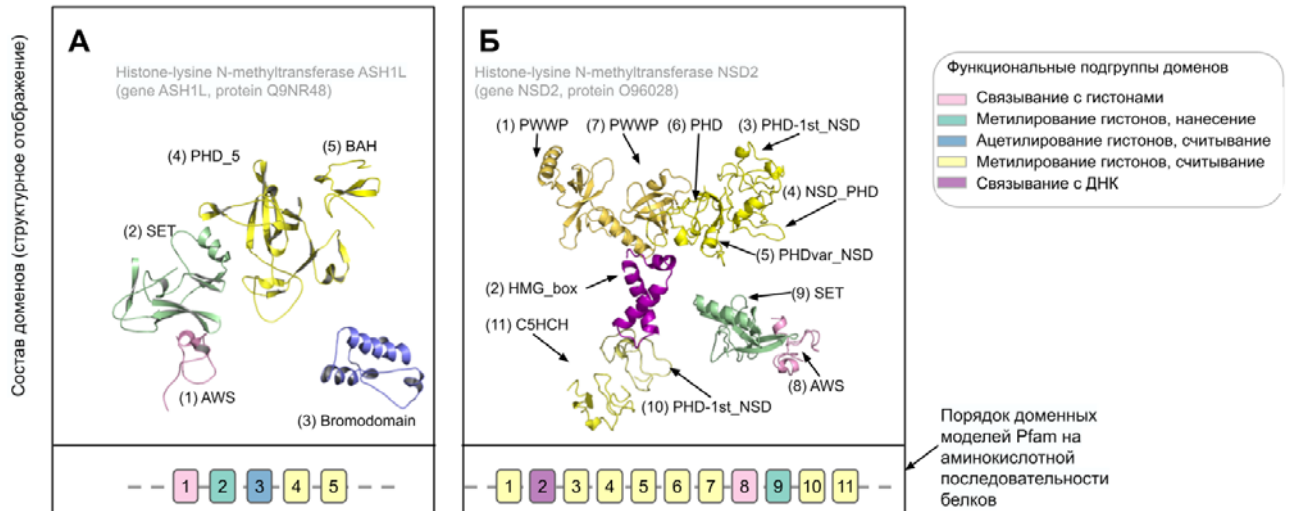
комбинаций совместно присутствующих доменов, а также аналогичный анализ для мультисубъединичных комплексов белков хроматина (**Рисунок 51**).



**Рисунок 51** — Комбинации доменов (доменных моделей по аннотации Pfam), относящихся к различным функциональным подгруппам, в отдельных белках и в мультисубъединичных комплексах белков хроматина (А и Б соответственно).

Выявлено, что в составе белков хроматина может быть до 4 доменов из разных функциональных подгрупп (**Рисунок 51**). К таким белкам относятся модификаторы гистонов. Например, в составе метилтрансферазы гистонов ASH1L идентифицировано 5 доменов из 4 функциональных подгрупп, обеспечивающих считывание и нанесение метильных меток гистонов (PHD\_5, BAN и SET соответственно), считывание ацетилирования гистонов (Bromodomain), а также связывание с гистонами (AWS) (**Рисунок 52А**). Примером белка хроматина с наибольшим числом доменов (11) служит метилтрансфераза гистонов NSD2, в

состав которой входят домены, участвующие в считывании метильных меток гистонов (PWWP, PHD-1st\_NSD, NSD\_PHD, PHDvar\_NSD), их нанесении (SET), связывании с гистонами (AWS) и связывании с ДНК (HMG\_box) (**Рисунок 52Б**).

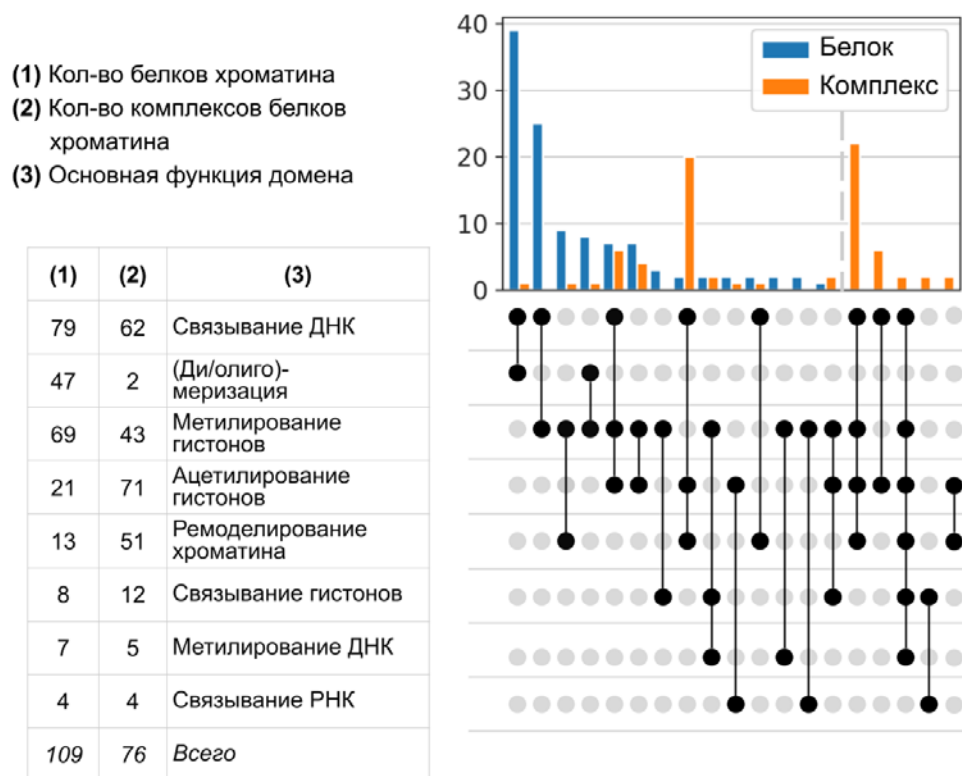


**Рисунок 52** — Примеры доменных архитектур в белках хроматина, содержащих наибольшее количество функциональных доменов, где **А** — метилтрансфераза гистонов ASHL1, **Б** — метилтрансфераза гистонов NSD2. Для каждого белка показаны трехмерные структуры доменов (из предсказанных с помощью AlphaFold2 структур), окрашенные в соответствии с основной функцией. Внизу показан порядок расположения доменов на уровне аминокислотной последовательности.

Далее результаты будут приведены с использованием функциональных групп доменов (**Рисунок 53**). В отдельных белках хроматина выявлены сочетания доменов, относящиеся к трем различным функциональным группам (**Рисунок 53**). В частности, домены, связывающие ДНК, совместно присутствовали с доменами ди-/олигомеризации; метилирования гистонов; метилирования и ацетилирования гистонов; ацетилирования гистонов и ремоделирования хроматина (**Рисунок 53**).

Для проведения аналогичного анализа мультисубъединичных белковых комплексов по данным базы данных ComplexPortal было отобрано 513 комплексов, в которых все белки, входящие в комплекс, относились к набору белков хроматина человека (SimChrom). В составе 181 из отобранных комплексов присутствовали белки хроматина, содержащие BMBV-домены. При этом в 101 комплексе было выявлено более одного BMBV-домена, а в составе 80 комплексов встречались

ВМВВ-домены из разных функциональных групп. Большинство комплексов последней категории (53) представляли собой различные варианты комплексов ремоделеров хроматина. Остальные относились к комплексам, осуществляющим ацетилирование и деацетилирование гистонов (13 и 4 комплекса соответственно), а также метилирование ДНК (2 комплекса).



**Рисунок 53** — Варианты сочетания доменов различных функциональных групп в отдельных белках и в мультисубъединичных белковых комплексах хроматина.

Наибольшее количество комплексов (22) состояли из доменов четырех функциональных групп, которые были ассоциированы со связыванием ДНК, метилированием и ацетилированием гистонов, а также ремоделированием хроматина (**Рисунок 53**). В двух комплексах, участвующих в ремоделировании хроматина (MBD2 или MBD3/NuRD — nucleosome remodeling and deacetylase complex), было выявлено наличие доменов из шести функциональных групп: перечисленные выше, а также ассоциированных с метилированием ДНК и связыванием с гистонами (**Рисунок 53**). Примечательно, что в комплексах белков

домены, ассоциированные с ацетилированием гистонов, были чаще представлены в комбинациях с другими функциональными доменами, чем в отдельных белках хроматина. Количество таких комплексов и белков составило 71 и 21 соответственно (**Рисунок 53**). Однако данное наблюдение может быть следствием неполноты сведений об известных белковых комплексах хроматина, а также об их вариабельности.

Таким образом, объединение белков хроматина в мультисубъединичные комплексы значительно расширяет репертуар сочетающихся доменов, участвующих в мультивалентных взаимодействиях в хроматине.

## Заключение

В диссертационной работе разработана эмпирическая функционально-ориентированная классификация белков хроматина человека SimChrom, полученная на основе интеграции гетерогенных данных об их репертуаре (из экспериментов, баз данных и литературы) и свойствах (по результатам вычислительных расчетов). Разработанная классификация позволила объединить в рамках единой функционально-ориентированной системы следующие сведения о белках: функциональную аннотацию, субклеточную локализацию, физико-химические свойства, доменный состав и количественную представленность в клетке.

Применение классификации позволило охарактеризовать функциональные категории белков хроматина, выявить статистически значимые различия в их физико-химических свойствах, а также определить характерные функциональные домены и их сочетания, ассоциированные с мультивалентными взаимодействиями в хроматине. В белках хроматина выявлено более 200 структурных доменов, ранее не охарактеризованных экспериментально; они являются потенциальными мишенями для дальнейших исследований методами структурной биологии и биофизики.

Полученные результаты демонстрируют, что разработанная классификация может служить инструментом обнаружения скрытых структурно-функциональных закономерностей белков хроматина и обладает эвристической ценностью. Она позволяет сопоставлять характерные признаки белков с их функциями, что дает возможность уточнять функции известных белков хроматина или предсказывать наиболее вероятные функции для ранее неохарактеризованных белков.

Для поддержки дальнейших исследований в этой области разработана интерактивная база данных SimChrom (<https://simchrom.intbio.org/>), в которой собраны инструменты для сопоставления аннотаций о локализации и функциях

белков хроматина, исследования доменной организации и парных сочетаний доменов в составе белков хроматина.

Таким образом, полученные результаты формируют целостное представление о составе, свойствах и принципах организации белков хроматина человека, расширяют понимание молекулярных механизмов регуляции генома, а также закладывают основу для дальнейшего прогнозирования и уточнения функций белков хроматина и построения количественных моделей его пространственно-временной организации.

### **Основные выводы диссертации:**

1. Проведена систематизация и мета-анализ разнородных источников информации о белковом составе хроматина и ядра клетки человека. Определены причины несогласованности данных между различными источниками информации, которые включают различия в подходах к аннотации множественной локализации белков, а также методические трудности детекции в составе фракции хроматина белков с низкой экспрессией или транзистентным характером взаимодействия.
2. Разработана эмпирическая функционально-ориентированная классификация белков хроматина человека SimChrom (39 категорий, 3045 белков), интегрирующая сведения о функциях, субклеточной локализации, физико-химических свойствах и доменной организации белков. Разработаны эталонные наборы белков ядра и цитоплазмы клетки человека с различным уровнем достоверности и согласованности данных между источниками информации.
3. На основе анализа масс-спектрометрических данных установлено, что доминирующими по массе компонентами хроматина являются гистоны и белки, участвующие в процессинге РНК: их массовая доля составляет около 40–60 % от общей массы всех белков хроматина.
4. Установлены функционально значимые физико-химические и структурные свойства белков хроматина человека.

- 4.1. Показано, что белки хроматина по сравнению с белками цитоплазмы являются менее структурированными (медианные значения доли неупорядоченных регионов 46 % и 23 % соответственно), что обусловлено увеличением числа и длины неупорядоченных регионов (медианы 70 и 50 а.о.). Высокая неупорядоченность характерна для пионерных транскрипционных факторов и архитектурных белков семейства HMG, тогда как высокой упорядоченностью отличаются субъединицы РНК-полимераз.
- 4.2. Выявлены особенности белков хроматина, обуславливающие их электростатические взаимодействия с другими белками и ДНК. В хроматине преобладают положительно заряженные белки (61 %), обогащенные аминокислотными остатками лизина, аргинина (в упорядоченных регионах) и обедненные остатками аспартата (в неупорядоченных регионах).
- 4.3. В составе белков хроматина выявлены сочетания функциональных доменов, способствующие мультивалентным взаимодействиям. Среди них преобладают комбинации доменов, ассоциированных с метилированием и ацетилированием гистонов, а также со связыванием ДНК. Это свидетельствует о кооперативной эволюции систем метилирования и ацетилирования гистонов для эпигенетической регуляции экспрессии генов.

## Благодарности

Автор выражает искреннюю признательность научному руководителю Алексею Константиновичу Шайтану за постановку задач, обсуждение результатов, внимание к деталям и поддержку на всех этапах исследования. Автор благодарит Григория Алексеевича Армеева за техническую реализацию базы данных SimChrom, предоставление данных о структурах нуклеосом и их комплексов из базы данных NucleosomeDB, техническое сопровождение и обсуждение результатов.

Автор выражает благодарность А.Л. Сивкиной, Д.К. Малининой, Н.С. Герасимовой, А.В. Любителю, С.В. Ульянову и А.А. Гаврилову за обсуждение категорий классификации SimChrom, которое способствовало ее развитию и улучшению.

Работа поддержана грантами Российского научного фонда № 25-14-00046 (разработка классификации SimChrom, структурная характеристика белков) и № 23-74-10012 (анализ физико-химических свойств белков хроматина), а также выполнена в рамках соглашения с Министерством науки и высшего образования Российской Федерации № 075-15-2024-539 от 24.04.2024 по теме «Эпигенетика как основа для разработки новых стратегий лечения болезней» (анализ мультидоменных белков хроматина).

## Список сокращений

а.о. — аминокислотный остаток

ВМВВ-домены — домены, вовлеченные в мультивалентные взаимодействия

ВНР — внутренне неупорядоченный регион белка (англ. intrinsically disordered region, IDR)

ВУР — внутренне упорядоченный регион белка

ИО — истинно отрицательный

ИП — истинно положительный

ЛП — ложноположительный

МНКза — микрококковая нуклеаза

МС — масс-спектрометрия

ОТ ЭРВ — обратная транскриптаза эндогенных ретровирусов (англ. reverse transcriptase of endogenous retroviruses, ERVs)

ПТМ — посттрансляционная модификация белка

САП — объединённая (конкатенированная) совокупная аминокислотная последовательность белков хроматина человека

ТФ — транскрипционный фактор

ЭДТА — этилендиаминтетрауксусная кислота

ChEP (Chromatin Enrichment for Proteomics) — метод обогащения хроматина для протеомики

CHESS (chromatin-enriching salt separation) — метод разделения фракций белков хроматина солями

CS (consensus) — консенсусный набор белков, аннотации о локализации которых согласованы в UniProt и HPA

CYTLOC — набор белков цитоплазмы

DEMAC (density-based enrichment for MS analysis of chromatin) — метод обогащения на основе плотности для МС-анализа

ECO (Evidence Code Ontology) — онтология источников информации

GO — Gene Ontology

HMG — белки семейства High mobility Group

HPA — Human Protein Atlas

ICAT (Isotope Coded Affinity Tag) — метод изотопного мечения в количественной протеомике

iBAQ (Intensity-Based Absolute Quantification) — метод абсолютного количественного определения на основе интенсивности сигналов

iPOND (isolation of proteins on nascent DNA) — метод выделения белков на вновь синтезированной ДНК

ISWI (Imitation SWItch)— семейство ремоделеров хроматина

JT (joint) — объединенный набор белков с необязательной согласованностью информации между UniProt и HPA)

NCC (nascent chromatin capture)— метод захвата вновь синтезированного хроматина

NECF (no evidence code filtration) — набор белков без фильтрации по критериям достоверности

NULOC — набор белков ядра

PaxDb\_INT — набор данных из базы данных PaxDb о представленности белков («H. sapiens — Whole organism (Integrated)»)

PaxDb\_PA — набор данных из базы данных PaxDb о представленности белков («Whole organism, SC (PeptideAtlas, Aug 2014)»)

ppm (parts per million) — частей на миллион

PRC1 — комплекс polycomb repressive complex 1

PRC2 — комплекс polycomb repressive complex 2

SILAC (Stable Isotope Labeling by Amino acids in Cell culture) — стабильное изотопное мечение аминокислотами в культуре клеток

SimChrom-SL — классификация SimChrom, в которой каждый белок отнесен только к одной категории (single labeled — с одной категорией)

SMC (structural maintenance of chromosomes) — семейство белков структурного поддержания хромосом

SWI/SNF (SWItch/Sucrose Non-Fermentable) — семейство ремоделеров хроматина

UL (unique localization) — набор белков с единственной локализацией в определенном клеточном компартменте

UMAP (Uniform Manifold Approximation and Projection) — метод снижения размерности

zf-C2H2 — домен типа цинковый палец

## Список литературы

1. Van Holde K. E. Chromatin. Springer-Verlag, 1989.
2. Bernstein E., Allis C. D. RNA meets chromatin // *Genes Dev. Cold Spring Harbor Lab*, 2005. T. 19, № 14. C. 1635–1655.
3. Millán-Zambrano G., Burton A., Bannister A. J., Schneider R. Histone post-translational modifications — cause and consequence of genome function // *Nat Rev Genet*. 2022.
4. Berger S. L. The complex language of chromatin regulation during transcription // *Nature*. Nature Publishing Group, 2007. T. 447, № 7143. C. 407–412.
5. Imhof A., Bonaldi T. “Chromatomics” the analysis of the chromatome // *Mol. BioSyst. The Royal Society of Chemistry*, 2005. T. 1, № 2. C. 112–116.
6. van Mierlo G., Vermeulen M. Chromatin Proteomics to Study Epigenetics — Challenges and Opportunities // *Molecular & Cellular Proteomics*. 2021. T. 20. C. 100056.
7. Thul P. J., Åkesson L., Wiking M., Mahdessian D., Geladaki A., Ait Blal H., Alm T., Asplund A., Björk L., Breckels L. M., Bäckström A., Danielsson F., Fagerberg L., Fall J., Gatto L., Gnann C., Hober S., Hjelmare M., Johansson F., Lee S., Lindskog C., Mulder J., Mulvey C. M., Nilsson P., Oksvold P., Rockberg J., Schutten R., Schwenk J. M., Sivertsson Å., Sjöstedt E., Skogs M., Stadler C., Sullivan D. P., Tegel H., Winsnes C., Zhang C., Zwahlen M., Mardinoglu A., Pontén F., von Feilitzen K., Lilley K. S., Uhlén M., Lundberg E. A subcellular map of the human proteome // *Science. American Association for the Advancement of Science*, 2017. T. 356, № 6340. C. eaal3321.
8. Cho N. H., Cheveralls K. C., Brunner A.-D., Kim K., Michaelis A. C., Raghavan P., Kobayashi H., Savy L., Li J. Y., Canaj H., Kim J. Y. S., Stewart E. M., Gnann C., McCarthy F., Cabrera J. P., Brunetti R. M., Chhun B. B., Dingle G., Hein M. Y., Huang B., Mehta S. B., Weissman J. S., Gómez-Sjöberg R., Itzhak D. N., Royer L. A., Mann M., Leonetti M. D. OpenCell: Endogenous tagging for the cartography of human cellular organization // *Science. American Association for the Advancement of Science*, 2022. T. 375, № 6585. C. eabi6983.
9. Gene Ontology Consortium и др. The Gene Ontology knowledgebase in 2023 // *Genetics*. 2023. T. 224, № 1. C. iyad031.
10. Marakulina D., Vorontsov I. E., Kulakovskiy I. V., Lennartsson A., Drabløs F., Medvedeva Y. A. EpiFactors 2022: expansion and enhancement of a curated database of human epigenetic factors and complexes // *Nucleic Acids Research*. 2023. T. 51, № D1. C. D564–D570.
11. Armeev G. A., Gribkova A. K., Shaytan A. K. Nucleosomes and their complexes in the cryoEM era: Trends and limitations // *Front. Mol. Biosci*. 2022. T. 9. C. 1070489.
12. Pletenev I. A., Bazarevich M., Zagirova D. R., Kononkova A. D., Cherkasov A. V., Efimova O. I., Tiukacheva E. A., Morozov K. V., Ulianov K. A., Komkov D., Tvorogova A. V., Golimbet V. E., Kondratyev N. V., Razin S. V., Khaitovich P., Ulianov S. V., Khrameeva E. E. Extensive long-range polycomb interactions and weak compartmentalization are hallmarks of human neuronal 3D genome // *Nucleic Acids Research*. 2024. T. 52, № 11. C. 6234–6252.
13. Dekker C., Haering C. H., Peters J.-M., Rowland B. D. How do molecular motors fold the genome? // *Science*. 2023. T. 382, № 6671. C. 646–648.
14. Torrente M. P., Zee B. M., Young N. L., Baliban R. C., LeRoy G., Floudas C. A., Hake S. B., Garcia B. A. Proteomic Interrogation of Human Chromatin // *PLOS ONE. Public Library of Science*, 2011. T. 6, № 9. C. e24747.
15. Ulianov S. V., Velichko A. K., Magnitov M. D., Luzhin A. V., Golov A. K., Ovsiyannikova N., Kireev I. I., Gavrikov A. S., Mishin A. S., Garaev A. K., Tyakht A. V., Gavrilov A. A., Kantidze O. L., Razin S. V. Suppression of liquid–liquid phase separation by 1,6-hexanediol partially compromises the 3D genome organization in living cells // *Nucleic Acids Research*. 2021. T. 49, № 18. C. 10524–10541.
16. Rippe K. Liquid–Liquid Phase Separation in Chromatin // *Cold Spring Harb Perspect Biol. Cold Spring Harbor Lab*, 2022. T. 14, № 2. C. a040683.
17. Kanada R., Terakawa T., Kenzaki H., Takada S. Nucleosome Crowding in Chromatin Slows the Diffusion but Can Promote Target Search of Proteins // *Biophys J*. 2019. T. 116, № 12. C. 2285–2295.

18. Ruthenburg A. J., Li H., Patel D. J., David Allis C. Multivalent engagement of chromatin modifications by linked binding modules // *Nat Rev Mol Cell Biol.* Nature Publishing Group, 2007. T. 8, № 12. C. 983–994.
19. Cremer T., Cremer M., Cremer C. The 4D Nucleome: Genome Compartmentalization in an Evolutionary Context // *Biochemistry (Mosc).* 2018. T. 83, № 4. C. 313–325.
20. Cremer T., Cremer M. Chromosome Territories // *Cold Spring Harb Perspect Biol.* 2010. T. 2, № 3.
21. van Steensel B., Belmont A. S. Lamina-associated domains: links with chromosome architecture, heterochromatin and gene repression // *Cell.* 2017. T. 169, № 5. C. 780–791.
22. Gupta S., Bersaglieri C., Bär D., Raingeval M., Schaab L., Santoro R. The nucleolar granular component mediates genome-nucleolus interactions and establishes their repressive chromatin states // *Mol Cell.* 2025. T. 85, № 11. C. 2165–2175.e6.
23. Cremer T., Cremer M., Hübner B., Silahtaroglu A., Hendzel M., Lanctôt C., Strickfaden H., Cremer C. The Interchromatin Compartment Participates in the Structural and Functional Organization of the Cell Nucleus // *BioEssays.* 2020. T. 42, № 2. C. 1900132.
24. Masiello I., Siciliani S., Biggiogera M. Perichromatin region: a moveable feast // *Histochem Cell Biol.* 2018. T. 150, № 3. C. 227–233.
25. Sabari B. R., Dall’Agnese A., Young R. A. Biomolecular Condensates in the Nucleus // *Trends in Biochemical Sciences.* 2020. T. 45, № 11. C. 961–977.
26. Razin S. V., Gavrilov A. A. The Role of Liquid-Liquid Phase Separation in the Compartmentalization of Cell Nucleus and Spatial Genome Organization // *Biochemistry (Mosc).* 2020. T. 85, № 6. C. 643–650.
27. Akgol Oksuz B., Yang L., Abraham S., Venev S. V., Krietenstein N., Parsi K. M., Ozadam H., Oomen M. E., Nand A., Mao H., Genga R. M. J., Maehr R., Rando O. J., Mirny L. A., Gibcus J. H., Dekker J. Systematic evaluation of chromosome conformation capture assays // *Nat Methods.* 2021. T. 18, № 9. C. 1046–1055.
28. Fudenberg G., Imakaev M., Lu C., Goloborodko A., Abdennur N., Mirny L. A. Formation of Chromosomal Domains by Loop Extrusion // *Cell Rep.* 2016. T. 15, № 9. C. 2038–2049.
29. Schoenfelder S., Fraser P. Long-range enhancer-promoter contacts in gene expression control // *Nat Rev Genet.* 2019. T. 20, № 8. C. 437–455.
30. Luger K. Crystal structure of the nucleosome core particle at 2.8 Å resolution. 1997. T. 389. C. 10.
31. Fedulova A. S., Armeev G. A., Romanova T. A., Singh-Palchevskaia L., Kosarim N. A., Motorin N. A., Komarova G. A., Shaytan A. K. Molecular dynamics simulations of nucleosomes are coming of age // *WIREs Computational Molecular Science.* 2024. T. 14, № 4. C. e1728.
32. Oleinikov P. D., Fedulova A. S., Armeev G. A., Motorin N. A., Singh-Palchevskaia L., Sivkina A. L., Feskin P. G., Glukhov G. S., Afonin D. A., Komarova G. A., Kirpichnikov M. P., Studitsky V. M., Feofanov A. V., Shaytan A. K. Interactions of Nucleosomes with Acidic Patch-Binding Peptides: A Combined Structural Bioinformatics, Molecular Modeling, Fluorescence Polarization, and Single-Molecule FRET Study: 20 // *IJMS. Multidisciplinary Digital Publishing Institute*, 2023. T. 24, № 20. C. 15194.
33. McGinty R. K., Tan S. Principles of nucleosome recognition by chromatin factors and enzymes // *Current Opinion in Structural Biology.* 2021. T. 71. C. 16–26.
34. Mayran A., Drouin J. Pioneer transcription factors shape the epigenetic landscape // *J. Biol. Chem. Elsevier*, 2018. T. 293, № 36. C. 13795–13804.
35. Nizovtseva E. V., Clauvelin N., Todolli S., Polikanov Y. S., Kulaeva O. I., Wengrzynek S., Olson W. K., Studitsky V. M. Nucleosome-free DNA regions differentially affect distant communication in chromatin // *Nucleic Acids Res.* 2017. T. 45, № 6. C. 3059–3067.
36. Antequera F. Structure, function and evolution of CpG island promoters // *Cellular and Molecular Life Sciences (CMLS).* 2003.
37. Shi J., Xu J., Chen Y. E., Li J. S., Et. A. The concurrence of DNA methylation and demethylation is associated with transcription regulation // *Nature Communications.* 2021.
38. Yin Y., Morgunova E., Jolma A., Kaasinen E., Et. A. Impact of cytosine methylation on DNA binding specificities of human transcription factors // *Science.* 2017.

39. Statello L., Guo C.-J., Chen L.-L., Huarte M. Gene regulation by long non-coding RNAs and its biological functions // *Nat Rev Mol Cell Biol.* Nature Publishing Group, 2021. T. 22, № 2. C. 96–118.
40. 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. T. 9, № 11. C. 843–854.
41. Kurumizaka H., Kujirai T., Takizawa Y. Contributions of Histone Variants in Nucleosome Structure and Function // *Journal of Molecular Biology.* 2021. T. 433, № 6. C. 166678.
42. Talbert P. B., Henikoff S. Histone variants at a glance // *J. Cell Sci.* 2021. T. 134, № 6. C. jcs244749.
43. Molaro A., Young J. M., Malik H. S. Evolutionary origins and diversification of testis-specific short histone H2A variants in mammals // *Genome Res.* Cold Spring Harbor Lab, 2018. T. 28, № 4. C. 460–473.
44. Singh R., Bassett E., Chakravarti A., Parthun M. R. Corrigenda: Replication-dependent histone isoforms: a new source of complexity in chromatin structure and function // *Nucleic Acids Res.* 2018. T. 46, № 18. C. 9893–9894.
45. Kossel A. Ueber einen peptonartigen Bestandtheil des Zellkerns. // *bchm.* 1884. T. 8, № 6. C. 511–515.
46. 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. T. 18. C. 209–244.
47. Talbert P. B., Ahmad K., Almouzni G., Ausió J., Berger F., Bhalla P. L., Bonner W. M., Cande W. Z., Chadwick B. P., Chan S. W. L., Cross G. A. M., Cui L., Dimitrov S. I., Doenecke D., Eirin-López J. M., Gorovsky M. A., Hake S. B., Hamkalo B. A., Holec S., Jacobsen S. E., Kamieniarz K., Khochbin S., Ladurner A. G., Landsman D., Latham J. A., Loppin B., Malik H. S., Marzluff W. F., Pehrson J. R., Postberg J., Schneider R., Singh M. B., Smith M. M., Thompson E., Torres-Padilla M.-E., Tremethick D. J., Turner B. M., Waterborg J. H., Wollmann H., Yelagandula R., Zhu B., Henikoff S. A unified phylogeny-based nomenclature for histone variants // *Epigenetics & Chromatin.* 2012. T. 5, № 1. C. 7.
48. 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. B. A standardized nomenclature for mammalian histone genes // *Epigenetics & Chromatin.* 2022. T. 15, № 1. C. 34.
49. Draizen E. J., Shaytan A. K., Mariño-Ramírez L., Talbert P. B., Landsman D., Panchenko A. R. HistoneDB 2.0: a histone database with variants—an integrated resource to explore histones and their variants // *Database.* 2016. T. 2016. C. baw014.
50. Zhao Y., Garcia B. A. Comprehensive Catalog of Currently Documented Histone Modifications // *Cold Spring Harb Perspect Biol.* Cold Spring Harbor Lab, 2015. T. 7, № 9. C. a025064.
51. Farrelly L. A., Thompson R. E., Zhao S., Lepack A. E., Lyu Y., Bhanu N. V., Zhang B., Loh Y.-H. E., Ramakrishnan A., Vadodaria K. C., Heard K. J., Erikson G., Nakadai T., Bastle R. M., Lukasak B. J., Zebroski H., Alenina N., Bader M., Berton O., Roeder R. G., Molina H., Gage F. H., Shen L., Garcia B. A., Li H., Muir T. W., Maze I. Histone serotonylation is a permissive modification that enhances TFIID binding to H3K4me3 // *Nature.* 2019. T. 567, № 7749. C. 535–539.
52. Lepack A. E., Werner C. T., Stewart A. F., Fulton S. L., Zhong P., Farrelly L. A., Smith A. C. W., Ramakrishnan A., Lyu Y., Bastle R. M., Martin J. A., Mitra S., O'Connor R. M., Wang Z.-J., Molina H., Turecki G., Shen L., Yan Z., Calipari E. S., Dietz D. M., Kenny P. J., Maze I. Dopaminylation of histone H3 in ventral tegmental area regulates cocaine seeking // *Science.* New York, N.Y., 2020. T. 368, № 6487. C. 197–201.
53. Cosgrove M. S., Boeke J. D., Wolberger C. Regulated nucleosome mobility and the histone code // *Nat. Struct. Mol. Biol.* 2004. T. 11, № 11. C. 1037–1043.
54. Shogren-Knaak M., Ishii H., Sun J.-M., Pazin M. J., Davie J. R., Peterson C. L. Histone H4-K16 Acetylation Controls Chromatin Structure and Protein Interactions // *Science.* 2006. T. 311, № 5762. C. 844–847.
55. Gibson B. A., Doolittle L. K., Schneider M. W. G., Jensen L. E., Gamarra N., Henry L., Gerlich D. W., Redding S., Rosen M. K. Organization of Chromatin by Intrinsic and Regulated Phase Separation // *Cell.* 2019. T. 179, № 2. C. 470-484.e21.
56. Jenuwein T., Allis C. D. Translating the Histone Code // *Science.* American Association for the Advancement of Science, 2001. T. 293, № 5532. C. 1074–1080.

57. Landt S. G., Marinov G. K., Kundaje A., Kheradpour P., Pauli F., Batzoglu S., Bernstein B. E., Bickel P., Brown J. B., Cayting P., Chen Y., DeSalvo G., Epstein C., Fisher-Aylor K. I., Euskirchen G., Gerstein M., Gertz J., Hartemink A. J., Hoffman M. M., Iyer V. R., Jung Y. L., Karmakar S., Kellis M., Kharchenko P. V., Li Q., Liu T., Liu X. S., Ma L., Milosavljevic A., Myers R. M., Park P. J., Pazin M. J., Perry M. D., Raha D., Reddy T. E., Rozowsky J., Shores N., Sidow A., Slattery M., Stamatoyannopoulos J. A., Tolstorukov M. Y., White K. P., Xi S., Farnham P. J., Lieb J. D., Wold B. J., Snyder M. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia // *Genome Res.* 2012. T. 22, № 9. C. 1813–1831.
58. Walshe J. L., Ochmann M., Neef U., Dybkov O., Dienemann C., Oberthür C., Zheenbekova A., Urlaub H., Cramer P. Molecular mechanism of co-transcriptional H3K36 methylation by SETD2 // *Nat Commun.* 2025. T. 16, № 1. C. 9565.
59. Sanulli S., Gross J. D., Narlikar G. J. Biophysical Properties of HP1-Mediated Heterochromatin // *Cold Spring Harb Symp Quant Biol.* Cold Spring Harbor Laboratory Press, 2019. T. 84. C. 217–225.
60. Strahl B. D., Allis C. D. The language of covalent histone modifications // *Nature.* 2000. T. 403, № 6765. C. 41–45.
61. Ernst J., Kellis M. Chromatin state discovery and genome annotation with ChromHMM // *Nat Protoc.* 2017. T. 12, № 12. C. 2478–2492.
62. Filion G. J., van Bommel J. G., Braunschweig U., Talhout W., Kind J., Ward L. D., Brugman W., de Castro I. J., Kerkhoven R. M., Bussemaker H. J., van Steensel B. Systematic protein location mapping reveals five principal chromatin types in *Drosophila* cells // *Cell.* 2010. T. 143, № 2. C. 212–224.
63. Schaefer U., Ho J. S. Y., Prinjha R. K., Tarakhovsky A. The “histone mimicry” by pathogens // *Cold Spring Harb Symp Quant Biol.* 2013. T. 78. C. 81–90.
64. Marazzi I., Ho J. S. Y., Kim J., Manicassamy B., Dewell S., Albrecht R. A., Seibert C. W., Schaefer U., Jeffrey K. L., Prinjha R. K., Lee K., García-Sastre A., Roeder R. G., Tarakhovsky A. Suppression of the antiviral response by an influenza histone mimic: 7390 // *Nature.* Nature Publishing Group, 2012. T. 483, № 7390. C. 428–433.
65. Kee J., Thudium S., Renner D. M., Glastad K., Palozola K., Zhang Z., Li Y., Lan Y., Cesare J., Poleshko A., Kiseleva A. A., Truitt R., Cardenas-Diaz F. L., Zhang X., Xie X., Kotton D. N., Alysandratos K. D., Epstein J. A., Shi P.-Y., Yang W., Morrissey E., Garcia B. A., Berger S. L., Weiss S. R., Korb E. SARS-CoV-2 disrupts host epigenetic regulation via histone mimicry: 7931 // *Nature.* Nature Publishing Group, 2022. T. 610, № 7931. C. 381–388.
66. Nacev B. A., Feng L., Bagert J. D., Lemiesz A. E., Gao J., Soshnev A. A., Kundra R., Schultz N., Muir T. W., Allis C. D. The expanding landscape of ‘oncohistone’ mutations in human cancers // *Nature.* Nature Publishing Group, 2019. T. 567, № 7749. C. 473–478.
67. Deshmukh S., Ptack A., Krug B., Jabado N. Oncohistones: A Roadmap to Stalled Development // *FEBS J.* 2022. T. 289, № 5. C. 1315–1328.
68. Lu C., Jain S. U., Hoelper D., Bechet D., Molden R. C., Ran L., Murphy D., Venneti S., Hameed M., Pawel B. R., Wunder J. S., Dickson B. C., Lundgren S. M., Jani K. S., De Jay N., Papillon-Cavanagh S., Andrulis I. L., Sawyer S. L., Grynspan D., Turcotte R. E., Nadaf J., Fahiminiyah S., Muir T. W., Majewski J., Thompson C. B., Chi P., Garcia B. A., Allis C. D., Jabado N., Lewis P. W. Histone H3K36 mutations promote sarcomagenesis through altered histone methylation landscape // *Science.* 2016. T. 352, № 6287. C. 844–849.
69. Wu G., Broniscer A., McEachron T. A., Lu C., Paugh B. S., Becksfort J., Qu C., Ding L., Huether R., Parker M., Zhang J., Gajjar A., Dyer M. A., Mullighan C. G., Gilbertson R. J., Mardis E. R., Wilson R. K., Downing J. R., Ellison D. W., Zhang J., Baker S. J., St. Jude Children’s Research Hospital–Washington University Pediatric Cancer Genome Project. Somatic histone H3 alterations in pediatric diffuse intrinsic pontine gliomas and non-brainstem glioblastomas // *Nat Genet.* 2012. T. 44, № 3. C. 251–253.
70. Lutsik P., Baude A., Mancarella D., Öz S., Kühn A., Toth R., Hey J., Toprak U. H., Lim J., Nguyen V. H., Jiang C., Mayakonda A., Hartmann M., Rosemann F., Breuer K., Vonficht D., Grünschlager F., Lee S., Schuhmacher M. K., Kusevic D., Jauch A., Weichenhan D., Zustin J., Schlesner M., Haas S., Park

- J. H., Park Y. J., Oppermann U., Jeltsch A., Haller F., Fellenberg J., Lindroth A. M., Plass C. Globally altered epigenetic landscape and delayed osteogenic differentiation in H3.3-G34W-mutant giant cell tumor of bone // *Nat Commun.* 2020. T. 11, № 1. C. 5414.
71. Bennett R. L., Bele A., Small E. C., Will C. M., Nabet B., Oyer J. A., Huang X., Ghosh R. P., Grzybowski A. T., Yu T., Zhang Q., Riva A., Lele T. P., Schatz G. C., Kelleher N. L., Ruthenburg A. J., Liphardt J., Licht J. D. A Mutation in Histone H2B Represents A New Class Of Oncogenic Driver // *Cancer Discov.* 2019. T. 9, № 10. C. 1438–1451.
  72. Onyema A. C., DiForte C., Patel R., Poget S. F., Loverde S. M. Structural and thermodynamic impact of oncogenic mutations on the nucleosome core particle // *Biophys J.* 2025. T. 124, № 14. C. 2362–2379.
  73. Arimura Y., Ikura M., Fujita R., Noda M., Kobayashi W., Horikoshi N., Sun J., Shi L., Kusakabe M., Harata M., Ohkawa Y., Tashiro S., Kimura H., Ikura T., Kurumizaka H. Cancer-associated mutations of histones H2B, H3.1 and H2A.Z.1 affect the structure and stability of the nucleosome // *Nucleic Acids Res.* 2018. T. 46, № 19. C. 10007–10018.
  74. Espiritu D., Gribkova A. K., Gupta S., Shaytan A. K., Panchenko A. R. Molecular Mechanisms of Oncogenesis through the Lens of Nucleosomes and Histones // *J. Phys. Chem. B. American Chemical Society*, 2021. T. 125, № 16. C. 3963–3976.
  75. Khare S. P., Sharma A., Deodhar K. K., Gupta S. Overexpression of histone variant H2A.1 and cellular transformation are related in N-nitrosodiethylamine-induced sequential hepatocarcinogenesis // *Exp Biol Med (Maywood)*. 2011. T. 236, № 1. C. 30–35.
  76. Singh R., Harshman S. W., Ruppert A. S., Mortazavi A., Lucas D. M., Thomas-Ahner J. M., Clinton S. K., Byrd J. C., Freitas M. A., Parthun M. R. Proteomic profiling identifies specific histone species associated with leukemic and cancer cells // *Clin Proteom.* 2015. T. 12, № 1. C. 22.
  77. Bhattacharya S., Reddy D., Jani V., Gadewal N., Shah S., Reddy R., Bose K., Sonavane U., Joshi R., Gupta S. Histone isoform H2A1H promotes attainment of distinct physiological states by altering chromatin dynamics // *Epigenetics and Chromatin. BioMed Central*, 2017. T. 10, № 1. C. 1–19.
  78. Kurumizaka H. Structural studies of functional nucleosome complexes with transacting factors // *Proc Jpn Acad Ser B Phys Biol Sci.* 2022. T. 98, № 1. C. 1–14.
  79. Goodwin G. H., Nicolas R. H., Johns E. W. An improved large scale fractionation of high mobility group non-histone chromatin proteins // *Biochim Biophys Acta.* 1975. T. 405, № 2. C. 280–291.
  80. Hondele M., Ladurner A. G. The chaperone-histone partnership: for the greater good of histone traffic and chromatin plasticity // *Curr Opin Struct Biol.* 2011. T. 21, № 6. C. 698–708.
  81. Pardal A. J., Fernandes-Duarte F., Bowman A. J. The histone chaperoning pathway: from ribosome to nucleosome // *Essays Biochem.* 2019. T. 63, № 1. C. 29–43.
  82. Hammond C. M., Strømme C. B., Huang H., Patel D. J., Groth A. Histone chaperone networks shaping chromatin function: 3 // *Nat Rev Mol Cell Biol. Nature Publishing Group*, 2017. T. 18, № 3. C. 141–158.
  83. English C. M., Adkins M. W., Carson J. J., Churchill M. E. A., Tyler J. K. Structural Basis for the Histone Chaperone Activity of Asf1 // *Cell.* 2006. T. 127, № 3. C. 495–508.
  84. Tagami H., Ray-Gallet D., Almouzni G., Nakatani Y. Histone H3.1 and H3.3 complexes mediate nucleosome assembly pathways dependent or independent of DNA synthesis // *Cell.* 2004. T. 116, № 1. C. 51–61.
  85. Mattioli F., D’Arcy S., Luger K. The right place at the right time: chaperoning core histone variants // *EMBO Rep.* 2015. T. 16, № 11. C. 1454–1466.
  86. Fujii-Nakata T., Ishimi Y., Okuda A., Kikuchi A. Functional analysis of nucleosome assembly protein, NAP-1. The negatively charged COOH-terminal region is not necessary for the intrinsic assembly activity // *J Biol Chem.* 1992. T. 267, № 29. C. 20980–20986.
  87. Nagae F., Takada S., Terakawa T. Histone chaperone Nap1 dismantles an H2A/H2B dimer from a partially unwrapped nucleosome // *Nucleic Acids Res.* 2023. T. 51, № 11. C. 5351–5363.
  88. Box J. K., Paquet N., Adams M. N., Boucher D., Bolderson E., O’Byrne K. J., Richard D. J. Nucleophosmin: from structure and function to disease development // *BMC Mol Biol.* 2016. T. 17, № 1. C. 19.

89. Belotserkovskaya R., Oh S., Bondarenko V. A., Orphanides G., Studitsky V. M., Reinberg D. FACT facilitates transcription-dependent nucleosome alteration // *Science*. New York, N.Y., 2003. T. 301, № 5636. C. 1090–1093.
90. Volokh O., Studitsky V. M., Sokolova O. S. Beyond Chaperoning: The Multifaceted Role of FACT in Chromatin Transactions // *International Journal of Molecular Sciences*. Multidisciplinary Digital Publishing Institute, 2025. T. 26, № 11. C. 5176.
91. Hargreaves D. C., Crabtree G. R. ATP-dependent chromatin remodeling: genetics, genomics and mechanisms // *Cell Res*. 2011. T. 21, № 3. C. 396–420.
92. Petty E., Pillus L. Balancing chromatin remodeling and histone modifications in transcription // *Trends Genet*. 2013. T. 29, № 11. C. 621–629.
93. Clapier C. R., Iwasa J., Cairns B. R., Peterson C. L. Mechanisms of action and regulation of ATP-dependent chromatin-remodelling complexes // *Nature Reviews Molecular Cell Biology*. 2017.
94. Torres I. O., Fujimori D. G. Functional coupling between writers, erasers and readers of histone and DNA methylation // *Current Opinion in Structural Biology*. 2015. T. 35. C. 68–75.
95. Kouzarides T. Chromatin Modifications and Their Function // *Cell*. Elsevier, 2007. T. 128, № 4. C. 693–705.
96. Musselman C. A., Lalonde M.-E., Côté J., Kutateladze T. G. Perceiving the epigenetic landscape through histone readers // *Nature Structural & Molecular Biology*. 2012. T. 19, № 12. C. 1218–1227.
97. Vann K. R., Vishweshwaraiah Y. L., Dokholyan N. V., Kutateladze T. G. Searching for methyllysine-binding aromatic cages // *Biochem J*. 2021. T. 478, № 19. C. 3613–3619.
98. Yun M., Wu J., Workman J. L., Li B. Readers of histone modifications // *Cell Research*. 2011. T. 21, № 4. C. 564.
99. Morinière J., Rousseaux S., Steuerwald U., Soler-López M., Curtet S., Vitte A.-L., Govin J., Gaucher J., Sadoul K., Hart D. J., Krijgsveld J., Khochbin S., Müller C. W., Petosa C. Cooperative binding of two acetylation marks on a histone tail by a single bromodomain // *Nature*. 2009. T. 461, № 7264. C. 664–668.
100. Holehouse A. S., Kragelund B. B. The molecular basis for cellular function of intrinsically disordered regions // *Nat Rev Mol Cell Biol*. 2024. T. 25, № 3. C. 187–211.
101. Holehouse A. S., Kragelund B. B. The molecular basis for cellular function of intrinsically disordered protein regions // *Nat Rev Mol Cell Biol*. Nature Publishing Group, 2024. T. 25, № 3. C. 187–211.
102. Yang Y., Mu J., Wang Y., Lai L. Protein IDR linkers regulate overall protein conformation and phase separation // *CR-PHYS-SC*. Elsevier, 2025. T. 6, № 12.
103. Boeynaems S., Alberti S., Fawzi N. L., Mittag T., Polymenidou M., Rousseau F., Schymkowitz J., Shorter J., Wolozin B., Van Den Bosch L., Tompa P., Fuxreiter M. Protein Phase Separation: A New Phase in Cell Biology // *Trends Cell Biol*. 2018. T. 28, № 6. C. 420–435.
104. Ruff K. M., King M. R., Ying A. W., Liu V., Pant A., Lieberman W. E., Shinn M. K., Su X., Kadoch C., Pappu R. V. Molecular grammars of predicted intrinsically disordered regions that span the human proteome // *Cell*. Elsevier, 2026. T. 189, № 1. C. 323–342.e17.
105. Abidi A. A., Cattoglio C., Tang N. N., Fan V. B., Dailey G. M., Hay A. D., Kunamaneni P., Milkie D. E., Darzacq X., Betzig E., Tjian R., Graham T. G. W. Unstructured transcription factor interactions enable emergent specificity // *Science*. American Association for the Advancement of Science, 2026. T. 392, № 6801. C. eaeb6487.
106. Watson M., Stott K. Disordered domains in chromatin-binding proteins // *Essays Biochem*. Portland Press, 2019. T. 63, № 1. C. 147–156.
107. Maiti S., Singh A., Maji T., Saibo N. V., De S. Experimental methods to study the structure and dynamics of intrinsically disordered regions in proteins // *Curr Res Struct Biol*. 2024. T. 7. C. 100138.
108. Piovesan D., Del Conte A., Mehdiabadi M., Aspromonte M. C., Blum M., Tesei G., von Bülow S., Lindorff-Larsen K., Tosatto S. C. E. MOBIDB in 2025: integrating ensemble properties and function annotations for intrinsically disordered proteins // *Nucleic Acids Res*. 2025. T. 53, № D1. C. D495–D503.
109. Nugnes M. V., Bouhraoua K. E. A., Zoubiri M., Pancsa R., Fichó E., DisProt Consortium, Tompa P., Piovesan D., Tosatto S. C. E., Aspromonte M. C. DisProt in 2026: enhancing intrinsically disordered

- proteins accessibility, deposition, and annotation // *Nucleic Acids Res.* 2026. T. 54, № D1. C. D383–D392.
110. Kumar M., Michael S., Alvarado-Valverde J., Zeke A., Lazar T., Glavina J., Nagy-Kanta E., Donagh J. M., Kalman Z. E., Pascarelli S., Palopoli N., Dobson L., Suarez C. F., Van Roey K., Krystkowiak I., Griffin J. E., Nagpal A., Bhardwaj R., Diella F., Mészáros B., Dean K., Davey N. E., Pancsa R., Chemes L. B., Gibson T. J. ELM—the Eukaryotic Linear Motif resource—2024 update // *Nucleic Acids Res.* 2024. T. 52, № D1. C. D442–D455.
  111. Li Q., Peng X., Li Y., Tang W., Zhu J., Huang J., Qi Y., Zhang Z. LLPSDB: a database of proteins undergoing liquid–liquid phase separation in vitro // *Nucleic Acids Res. Oxford Academic*, 2020. T. 48, № D1. C. D320–D327.
  112. You K., Huang Q., Yu C., Shen B., Sevilla C., Shi M., Hermjakob H., Chen Y., Li T. PhaSepDB: a database of liquid-liquid phase separation related proteins // *Nucleic Acids Res.* 2020. T. 48, № D1. C. D354–D359.
  113. Mészáros B., Erdős G., Szabó B., Schád É., Tantos Á., Abukhairan R., Horváth T., Murvai N., Kovács O. P., Kovács M., Tosatto S. C. E., Tompa P., Dosztányi Z., Pancsa R. PhaSePro: the database of proteins driving liquid-liquid phase separation // *Nucleic Acids Res.* 2020. T. 48, № D1. C. D360–D367.
  114. Kurgan L., Hu G., Wang K., Ghadermarzi S., Zhao B., Malhis N., Erdős G., Gsponer J., Uversky V. N., Dosztányi Z. Tutorial: a guide for the selection of fast and accurate computational tools for the prediction of intrinsic disorder in proteins // *Nat Protoc. Nature Publishing Group*, 2023. T. 18, № 11. C. 3157–3172.
  115. Lotthammer J. M., Ginell G. M., Griffith D., Emenecker R. J., Holehouse A. S. Direct prediction of intrinsically disordered protein conformational properties from sequence // *Nat Methods*. 2024. T. 21, № 3. C. 465–476.
  116. Ginell G. M., Emenecker R. J., Lotthammer J. M., Keeley A. T., Plassmeyer S. P., Razo N., Usher E. T., Pelham J. F., Holehouse A. S. Sequence-based prediction of intermolecular interactions driven by disordered regions // *Science. New York, N.Y.*, 2025. T. 388, № 6749. C. eadq8381.
  117. Erdős G., Pajkos M., Dosztányi Z. IUPred3: prediction of protein disorder enhanced with unambiguous experimental annotation and visualization of evolutionary conservation // *Nucleic Acids Research*. 2021. T. 49, № W1. C. W297–W303.
  118. Akdel M., Pires D. E. V., Pardo E. P., Jänes J., Zalevsky A. O., Mészáros B., Bryant P., Good L. L., Laskowski R. A., Pozzati G., Shenoy A., Zhu W., Kundrotas P., Serra V. R., Rodrigues C. H. M., Dunham A. S., Burke D., Borkakoti N., Velankar S., Frost A., Basquin J., Lindorff-Larsen K., Bateman A., Kajava A. V., Valencia A., Ovchinnikov S., Durairaj J., Ascher D. B., Thornton J. M., Davey N. E., Stein A., Elofsson A., Croll T. I., Beltrao P. A structural biology community assessment of AlphaFold2 applications // *Nat Struct Mol Biol. Nature Publishing Group*, 2022. T. 29, № 11. C. 1056–1067.
  119. Vogel C., Bashton M., Kerrison N. D., Chothia C., Teichmann S. A. Structure, function and evolution of multidomain proteins // *Current Opinion in Structural Biology*. 2004. T. 14, № 2. C. 208–216.
  120. Paysan-Lafosse T., Andreeva A., Blum M., Chuguransky S. R., Grego T., Pinto B. L., Salazar G. A., Bileschi M. L., Llinares-López F., Meng-Papaxanthos L., Colwell L. J., Grishin N. V., Schaeffer R. D., Clementel D., Tosatto S. C. E., Sonnhammer E., Wood V., Bateman A. The Pfam protein families database: embracing AI/ML // *Nucleic Acids Res.* 2025. T. 53, № D1. C. D523–D534.
  121. Rost B. Twilight zone of protein sequence alignments // *Protein Engineering*. 1999. T. 12, № 2. C. 85–94.
  122. Sillitoe I., Bordin N., Dawson N., Waman V. P., Et. A. CATH: increased structural coverage of functional space // *Nucleic Acids Research*. 2020.
  123. Nallapareddy V., Bordin N., Sillitoe I., Heinzinger M., Littmann M., Waman V. P., Sen N., Rost B., Orengo C. CATHe: detection of remote homologues for CATH superfamilies using embeddings from protein language models // *Bioinformatics*. 2023. T. 39, № 1. C. btad029.
  124. Lau A. M., Bordin N., Kandathil S. M., Sillitoe I., Waman V. P., Wells J., Orengo C. A., Jones D. T. Exploring structural diversity across the protein universe with The Encyclopedia of Domains // *Science. American Association for the Advancement of Science*, 2024.

125. van Kempen M., Kim S. S., Tumescheit C., Mirdita M., Lee J., Gilchrist C. L. M., Söding J., Steinegger M. Fast and accurate protein structure search with Foldseek // *Nat Biotechnol.* Nature Publishing Group, 2023. C. 1–4.
126. Steinegger M., Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets // *Nat Biotechnol.* 2017. T. 35, № 11. C. 1026–1028.
127. Gryder B., Scacheri P. C., Ried T., Khan J. Chromatin Mechanisms Driving Cancer // *Cold Spring Harb Perspect Biol.* 2022. T. 14, № 3. C. a040956.
128. Locatelli M., Faure-Dupuy S. Virus hijacking of host epigenetic machinery to impair immune response // *J Virol.* T. 97, № 9. C. e00658-23.
129. Phair R. D., Scaffidi P., Elbi C., Vecerová J., Dey A., Ozato K., Brown D. T., Hager G., Bustin M., Misteli T. Global Nature of Dynamic Protein-Chromatin Interactions In Vivo: Three-Dimensional Genome Scanning and Dynamic Interaction Networks of Chromatin Proteins // *Mol Cell Biol.* 2004. T. 24, № 14. C. 6393–6402.
130. Sutherland H. G. E., Mumford G. K., Newton K., Ford L. V., Farrall R., Dellaire G., Cáceres J. F., Bickmore W. A. Large-scale identification of mammalian proteins localized to nuclear sub-compartments // *Human Molecular Genetics.* 2001. T. 10, № 18. C. 1995–2011.
131. 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. T. 31, № 1. C. 328–330.
132. Sprenger J., Lynn Fink J., Karunaratne S., Hanson K., Hamilton N. A., Teasdale R. D. LOCATE: a mammalian protein subcellular localization database // *Nucleic Acids Res.* 2008. T. 36, № Database issue. C. D230-233.
133. Gendler K., Paulsen T., Napoli C. ChromDB: the chromatin database // *Nucleic Acids Res.* 2008. T. 36, № Database issue. C. D298-302.
134. Rastogi S., Rost B. LocDB: experimental annotations of localization for Homo sapiens and Arabidopsis thaliana // *Nucleic Acids Res.* 2011. T. 39, № Database issue. C. D230-234.
135. The UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2025 // *Nucleic Acids Research.* 2025. T. 53, № D1. C. D609–D617.
136. van Holde K. E. The Proteins of Chromatin. II. Nonhistone Chromosomal Proteins // *Chromatin* / под ред. van Holde K. E. New York, NY: Springer, 1989. C. 181–218.
137. Ashburner M., Ball C. A., Blake J. A., Botstein D., Butler H., Cherry J. M., Davis A. P., Dolinski K., Dwight S. S., Eppig J. T., Harris M. A., Hill D. P., Issel-Tarver L., Kasarskis A., Lewis S., Matese J. C., Richardson J. E., Ringwald M., Rubin G. M., Sherlock G. Gene Ontology: tool for the unification of biology: 1 // *Nat Genet.* Nature Publishing Group, 2000. T. 25, № 1. C. 25–29.
138. Tomczak A., Mortensen J. M., Winnenburg R., Liu C., Alessi D. T., Swamy V., Vallania F., Lofgren S., Haynes W., Shah N. H., Musen M. A., Khatri P. Interpretation of biological experiments changes with evolution of the Gene Ontology and its annotations: 1 // *Sci Rep.* Nature Publishing Group, 2018. T. 8, № 1. C. 5115.
139. Wang Q., Huang J., Sun H., Liu J., Wang J., Wang Q., Qin Q., Mei S., Zhao C., Yang X., Liu X. S., Zhang Y. CR Cistrome: a ChIP-Seq database for chromatin regulators and histone modification linkages in human and mouse // *Nucleic Acids Res.* 2014. T. 42, № Database issue. C. D450–D458.
140. Lu J., Xu J., Li J., Pan T., Bai J., Wang L., Jin X., Lin X., Zhang Y., Li Y., Sahni N., Li X. FACER: comprehensive molecular and functional characterization of epigenetic chromatin regulators // *Nucleic Acids Res.* 2018. T. 46, № 19. C. 10019–10033.
141. Ru B., Sun J., Tong Y., Wong C. N., Chandra A., Tang A. T. S., Chow L. K. Y., Wun W. L., Levitskaya Z., Zhang J. CR2Cancer: a database for chromatin regulators in human cancer // *Nucleic Acids Research.* 2018. T. 46, № D1. C. D918–D924.
142. Zhang Y., Zhang Y., Song C., Zhao X., Ai B., Wang Y., Zhou L., Zhu J., Feng C., Xu L., Wang Q., Sun H., Fang Q., Xu X., Li E., Li C. CRdb: a comprehensive resource for deciphering chromatin regulators in human // *Nucleic Acids Research.* 2023. T. 51, № D1. C. D88–D100.
143. Lambert S. A., Jolma A., Campitelli L. F., Das P. K., Yin Y., Albu M., Chen X., Taipale J., Hughes T. R., Weirauch M. T. The Human Transcription Factors // *Cell.* 2018. T. 172, № 4. C. 650–665.

144. Lovering R. C., Gaudet P., Acencio M. L., Ignatchenko A., Jolma A., Fornes O., Kuiper M., Kulakovskiy I. V., Lægreid A., Martin M. J., Logie C. A GO catalogue of human DNA-binding transcription factors // *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*. 2021. T. 1864, № 11. C. 194765.
145. Shah S. G., Mandloi T., Kunte P., Natu A., Rashid M., Reddy D., Gadewal N., Gupta S. HISTome2: a database of histone proteins, modifiers for multiple organisms and epidrugs // *Epigenetics & Chromatin*. 2020. T. 13, № 1. C. 31.
146. Xu Y., Zhang S., Lin S., Guo Y., Deng W., Zhang Y., Xue Y. WERAM: a database of writers, erasers and readers of histone acetylation and methylation in eukaryotes // *Nucleic Acids Research*. 2017. T. 45, № D1. C. D264–D270.
147. Mani U., S A. S., Goutham R N A., Mohan S S. SWI/SNF Infobase-An exclusive information portal for SWI/SNF remodeling complex subunits // *PLoS ONE. Public Library of Science*, 2017. T. 12, № 9. C. e0184445.
148. Azad G. K., Swagatika S., Kumawat M., Kumawat R., Tomar R. S. Modifying Chromatin by Histone Tail Clipping // *Journal of Molecular Biology*. 2018. T. 430, № 18, Part B. C. 3051–3067.
149. Reeves R. High mobility group (HMG) proteins: Modulators of chromatin structure and DNA repair in mammalian cells // *DNA Repair*. 2015. T. 36. C. 122–136.
150. Sun H., Fu B., Qian X., Xu P., Qin W. Nuclear and cytoplasmic specific RNA binding proteome enrichment and its changes upon ferroptosis induction // *Nat Commun. Nature Publishing Group*, 2024. T. 15, № 1. C. 852.
151. Van Nostrand E. L., Freese P., Pratt G. A., Wang X., Wei X., Xiao R., Blue S. M., Chen J.-Y., Cody N. A. L., Dominguez D., Olson S., Sundararaman B., Zhan L., Bazile C., Bouvrette L. P. B., Bergalet J., Duff M. O., Garcia K. E., Gelboin-Burkhart C., Hochman M., Lambert N. J., Li H., McGurk M. P., Nguyen T. B., Palden T., Rabano I., Sathe S., Stanton R., Su A., Wang R., Yee B. A., Zhou B., Louie A. L., Aigner S., Fu X.-D., Lécuyer E., Burge C. B., Graveley B. R., Yeo G. W. A large-scale binding and functional map of human RNA-binding proteins // *Nature. Nature Publishing Group*, 2020. T. 583, № 7818. C. 711–719.
152. Hoencamp C., Rowland B. D. Genome control by SMC complexes // *Nat Rev Mol Cell Biol. Nature Publishing Group*, 2023. T. 24, № 9. C. 633–650.
153. El Kennani S., Adrait A., Shaytan A. K., Khochbin S., Bruley C., Panchenko A. R., Landsman D., Pflieger D., Govin J. MS\_HistoneDB, a manually curated resource for proteomic analysis of human and mouse histones // *Epigenetics & Chromatin*. 2017. T. 10, № 1.
154. Shi M., You K., Chen T., Hou C., Liang Z., Liu M., Wang J., Wei T., Qin J., Chen Y., Zhang M. Q., Li T. Quantifying the phase separation property of chromatin-associated proteins under physiological conditions using an anti-1,6-hexanediol index // *Genome Biology*. 2021. T. 22, № 1. C. 229.
155. Mirsky A. E., Ris H. ISOLATED CHROMOSOMES // *Journal of General Physiology*. 1947. T. 31, № 1. C. 1–6.
156. R.F. Steiner. PHYSICO-CHEMICAL STUDIES ON THE COMPONENTS OF THYMUS CELL NUCLE. 1952.
157. Zubay G., Doty P. The isolation and properties of deoxyribonucleoprotein particles containing single nucleic acid molecules // *Journal of Molecular Biology*. 1959. T. 1, № 1. C. 1-IN1.
158. Shiio Y., Eisenman R. N., Yi E. C., Donohoe S., Goodlett D. R., Aebersold R. Quantitative proteomic analysis of chromatin-associated factors // *J. Am. Soc. Mass Spectrom.* 2003. T. 14, № 7. C. 696–703.
159. Kustatscher G., Wills K. L. H., Furlan C., Rappsilber J. Chromatin enrichment for proteomics: 9 // *Nat Protoc. Nature Publishing Group*, 2014. T. 9, № 9. C. 2090–2099.
160. Kustatscher G., Hégarat N., Wills K. L. H., Furlan C., Bukowski-Wills J.-C., Hocheegger H., Rappsilber J. Proteomics of a fuzzy organelle: interphase chromatin // *EMBO J.* 2014. T. 33, № 6. C. 648–664.
161. Kito Y., Matsumoto M., Hatano A., Takami T., Oshikawa K., Matsumoto A., Nakayama K. I. Cell cycle-dependent localization of the proteasome to chromatin: 1 // *Sci Rep. Nature Publishing Group*, 2020. T. 10, № 1. C. 5801.
162. Ohta S., Bukowski-Wills J.-C., Sanchez-Pulido L., Alves F. de L., Wood L., Chen Z. A., Platani M., Fischer L., Hudson D. F., Ponting C. P., Fukagawa T., Earnshaw W. C., Rappsilber J. The Protein

- Composition of Mitotic Chromosomes Determined Using Multiclassifier Combinatorial Proteomics // *Cell*. 2010. T. 142, № 5. C. 810–821.
163. Batugedara G., Lu X. M., Saraf A., Sardu M. E., Cort A., Abel S., Prudhomme J., Washburn M. P., Florens L., Bunnik E. M., Le Roch K. G. The chromatin bound proteome of the human malaria parasite // *Microb Genom*. 2020. T. 6, № 2. C. e000327.
  164. van Mierlo G., Wester R. A., Marks H. A Mass Spectrometry Survey of Chromatin-Associated Proteins in Pluripotency and Early Lineage Commitment // *PROTEOMICS*. 2019. T. 19, № 14. C. 1900047.
  165. Ginno P. A., Burger L., Seebacher J., Iesmantavicius V., Schübeler D. Cell cycle-resolved chromatin proteomics reveals the extent of mitotic preservation of the genomic regulatory landscape: 1 // *Nat Commun*. Nature Publishing Group, 2018. T. 9, № 1. C. 4048.
  166. Federation A. J., Nandakumar V., Searle B. C., Stergachis A., Wang H., Pino L. K., Merrihew G., Ting Y. S., Howard N., Kutayavin T., MacCoss M. J., Stamatoyannopoulos J. A. Highly Parallel Quantification and Compartment Localization of Transcription Factors and Nuclear Proteins // *Cell Reports*. 2020. T. 30, № 8. C. 2463-2471.e5.
  167. Alabert C., Bukowski-Wills J.-C., Lee S.-B., Kustatscher G., Nakamura K., de Lima Alves F., Menard P., Mejlvang J., Rappsilber J., Groth A. Nascent chromatin capture proteomics determines chromatin dynamics during DNA replication and identifies unknown fork components // *Nat Cell Biol*. Nature Publishing Group, 2014. T. 16, № 3. C. 281–291.
  168. Glish G. L., Vachet R. W. The basics of mass spectrometry in the twenty-first century // *Nature Reviews Drug Discovery*. 2003.
  169. Gundry R. L., White M. Y., Murray C. I., Kane L. A., Fu Q., Stanley B. A., Van Eyk J. E. Preparation of Proteins and Peptides for Mass Spectrometry Analysis in a Bottom-Up Proteomics Workflow // *Curr Protoc Mol Biol*. 2009. T. CHAPTER. C. Unit10.25.
  170. Elgin S. C. R., Bonner J. Partial fractionation and chemical characterization of the major nonhistone chromosomal proteins // *Biochemistry*. 1972. T. 11, № 5. C. 772–781.
  171. 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. T. 251, № 2. C. 548–554.
  172. 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. T. 80, № 4. C. 773–780.
  173. Gygi S. P., Rist B., Gerber S. A., Turecek F., Gelb M. H., Aebersold R. Quantitative analysis of complex protein mixtures using isotope-coded affinity tags: 10 // *Nat Biotechnol*. Nature Publishing Group, 1999. T. 17, № 10. C. 994–999.
  174. Ong S.-E., Blagoev B., Kratchmarova I., Kristensen D. B., Steen H., Pandey A., Mann M. Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics // *Mol Cell Proteomics*. 2002. T. 1, № 5. C. 376–386.
  175. Schwanhäusser B., Busse D., Li N., Dittmar G., Schuchhardt J., Wolf J., Chen W., Selbach M. Global quantification of mammalian gene expression control: 7347 // *Nature*. Nature Publishing Group, 2011. T. 473, № 7347. C. 337–342.
  176. Beck M., Schmidt A., Malmstroem J., Claassen M., Ori A., Szymborska A., Herzog F., Rinner O., Ellenberg J., Aebersold R. The quantitative proteome of a human cell line // *Molecular Systems Biology*. John Wiley & Sons, Ltd, 2011. T. 7, № 1. C. 549.
  177. Wiśniewski J. R., Ostasiewicz P., Duś K., Zielińska D. F., Gnadt F., Mann M. Extensive quantitative remodeling of the proteome between normal colon tissue and adenocarcinoma // *Molecular Systems Biology*. John Wiley & Sons, Ltd, 2012. T. 8, № 1. C. 611.
  178. Wiśniewski J. R., Hein M. Y., Cox J., Mann M. A “Proteomic Ruler” for Protein Copy Number and Concentration Estimation without Spike-in Standards\* // *Molecular & Cellular Proteomics*. 2014. T. 13, № 12. C. 3497–3506.
  179. van Holde K. E. The First Hundred Years // *Chromatin* / под ред. van Holde K. E. New York, NY: Springer, 1989. C. 1–15.

180. Wang M., Herrmann C. J., Simonovic M., Szklarczyk D., von Mering C. Version 4.0 of PaxDb: Protein abundance data, integrated across model organisms, tissues, and cell-lines // *Proteomics*. 2015. T. 15, № 18. C. 3163–3168.
181. Shi M., You K., Chen T., Hou C., Liang Z., Liu M., Wang J., Wei T., Qin J., Chen Y., Zhang M. Q., Li T. Quantifying the phase separation property of chromatin-associated proteins under physiological conditions using an anti-1,6-hexanediol index // *Genome Biology*. 2021. T. 22, № 1. C. 229.
182. Alvarez V., Bandau S., Jiang H., Rios-Szwed D., Hukelmann J., Garcia-Wilson E., Wiechens N., Griesser E., Have S. T., Owen-Hughes T., Lamond A., Alabert C. Proteomic profiling reveals distinct phases to the restoration of chromatin following DNA replication // *Cell Reports*. Elsevier, 2023. T. 42, № 1.
183. Ugur E., de la Porte A., Qin W., Bultmann S., Ivanova A., Drukker M., Mann M., Wierer M., Leonhardt H. Comprehensive chromatin proteomics resolves functional phases of pluripotency and identifies changes in regulatory components // *Nucleic Acids Research*. 2023. T. 51, № 6. C. 2671–2690.
184. Itzhak D. N., Tyanova S., Cox J., Borner G. H. Global, quantitative and dynamic mapping of protein subcellular localization // *eLife* / под ред. Hegde R. S. eLife Sciences Publications, Ltd, 2016. T. 5. C. e16950.
185. Gerstberger S., Hafner M., Tuschl T. A census of human RNA-binding proteins // *Nat Rev Genet*. Nature Publishing Group, 2014. T. 15, № 12. C. 829–845.
186. Liu H., Wang C., Lee S., Deng Y., Wither M., Oh S., Ning F., Dege C., Zhang Q., Liu X., Johnson A. M., Zang J., Chen Z., Janknecht R., Hansen K., Marrack P., Li C.-Y., Kappler J. W., Hagman J., Zhang G. Clipping of arginine-methylated histone tails by JMJD5 and JMJD7 // *Proc Natl Acad Sci U S A*. 2017. T. 114, № 37. C. E7717–E7726.
187. Raudvere U., Kolberg L., Kuzmin I., Arak T., Adler P., Peterson H., Vilo J. g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update) // *Nucleic Acids Res*. 2019. T. 47, № W1. C. W191–W198.
188. Bordin N., Sillitoe I., Nallapareddy V., Rauer C., Lam S. D., Waman V. P., Sen N., Heinzinger M., Littmann M., Kim S., Velankar S., Steinegger M., Rost B., Orengo C. AlphaFold2 reveals commonalities and novelties in protein structure space for 21 model organisms // *Commun Biol*. 2023. T. 6, № 1. C. 160.
189. Gligorijević V., Renfrew P. D., Kosciolk T., Leman J. K., Berenberg D., Vatanen T., Chandler C., Taylor B. C., Fisk I. M., Vlamakis H., Xavier R. J., Knight R., Cho K., Bonneau R. Structure-based protein function prediction using graph convolutional networks: 1 // *Nat Commun*. Nature Publishing Group, 2021. T. 12, № 1. C. 3168.
190. Cartwright P., Helin K. Nucleocytoplasmic shuttling of transcription factors // *Cell Mol Life Sci*. 2000. T. 57, № 8–9. C. 1193–1206.
191. Cheng D., Semmens K., McManus E., Chen Q., Meerzaman D., Wang X., Hafner M., Lewis B. A., Takahashi H., Devaiah B. N., Geggion A., Singer D. S. The nuclear transcription factor, TAF7, is a cytoplasmic regulator of protein synthesis // *Science Advances*. American Association for the Advancement of Science, 2021.
192. Shreberk-Shaked M., Oren M. New insights into YAP/TAZ nucleo-cytoplasmic shuttling: new cancer therapeutic opportunities? // *Mol Oncol*. 2019. T. 13, № 6. C. 1335–1341.
193. Kobiyama K., Kawashima A., Jounai N., Takeshita F., Ishii K. J., Ito T., Suzuki K. Role of Extrachromosomal Histone H2B on Recognition of DNA Viruses and Cell Damage // *Front Genet*. 2013. T. 4. C. 91.
194. Zeng Z., Chen L., Luo H., Xiao H., Gao S., Zeng Y. Progress on H2B as a multifunctional protein related to pathogens // *Life Sciences*. 2024. T. 347. C. 122654.
195. Sigismondo G., Papageorgiou D. N., Krijgsveld J. Cracking chromatin with proteomics: From chromatome to histone modifications // *PROTEOMICS*. 2022. T. 22, № 15–16. C. 2100206.
196. Zhou W., Cao H., Yang X., Cong K., Wang W., Chen T., Yin H., Wu Z., Cai X., Liu T., Xiao J. Characterization of Nuclear Localization Signal in the N Terminus of Integrin-linked Kinase-associated Phosphatase (ILKAP) and Its Essential Role in the Down-regulation of RSK2 Protein Signaling \* // *Journal of Biological Chemistry*. Elsevier, 2013. T. 288, № 9. C. 6259–6271.

197. Kachaev Z. M., Ivashchenko S. D., Kozlov E. N., Lebedeva L. A., Shidlovskii Y. V. Localization and Functional Roles of Components of the Translation Apparatus in the Eukaryotic Cell Nucleus: 11 // Cells. Multidisciplinary Digital Publishing Institute, 2021. T. 10, № 11. C. 3239.
198. Pronier E., Cifani P., Merlinsky T. R., Berman K. B., Somasundara A. V. H., Rampal R. K., LaCava J., Wei K. E., Pastore F., Maag J. L. V., Park J., Koche R., Kentsis A., Levine R. L. Targeting the CALR interactome in myeloproliferative neoplasms // JCI Insight. 2018. T. 3, № 22. C. e122703.
199. Kuo T., Hsu S., Huang S., Chang C. L., Feng C., Huang M., Chen T., Yang M., Jiang S., Wen T., Yang C., Huang C., Kao S., Tsai K., Yang G., Yang W. Pdia4 regulates  $\beta$ -cell pathogenesis in diabetes: molecular mechanism and targeted therapy // EMBO Molecular Medicine. Springer Nature, 2021. T. 13, № 10. C. e11668.
200. Nishimura S., Tsuda H., Ito K., Jobo T., Yaegashi N., Inoue T., Sudo T., Berkowitz R. S., Mok S. C. Differential expression of ABCF2 protein among different histologic types of epithelial ovarian cancer and in clear cell adenocarcinomas of different organs // Human Pathology. 2007. T. 38, № 1. C. 134–139.
201. Yehia L., Liu D., Fu S., Iyer P., Eng C. Non-canonical role of wild-type SEC23B in the cellular stress response pathway // Cell Death Dis. Nature Publishing Group, 2021. T. 12, № 4. C. 1–12.
202. Gao Y., Teng J., Hong Y., Qu F., Ren J., Li L., Pan X., Chen L., Yin L., Xu D., Cui X. The oncogenic role of EIF3D is associated with increased cell cycle progression and motility in prostate cancer // Med Oncol. 2015. T. 32, № 7. C. 196.
203. 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. C. 2023.04.17.537230.
204. Palii C. G., Cheng Q., Gillespie M. A., Shannon P., Mazurczyk M., Napolitani G., Price N. D., Ranish J. A., Morrissey E., Higgs D. R., Brand M. Single-Cell Proteomics Reveal that Quantitative Changes in Co-expressed Lineage-Specific Transcription Factors Determine Cell Fate // Cell Stem Cell. 2019. T. 24, № 5. C. 812-820.e5.
205. Clapier C. R., Iwasa J., Cairns B. R., Peterson C. L. Mechanisms of action and regulation of ATP-dependent chromatin-remodelling complexes // Nat. Rev. Mol. Cell Biol. 2017. T. 18, № 7. C. 407–422.
206. Dauden M. I., López-Perrote A., Llorca O. RUVBL1–RUVBL2 AAA-ATPase: a versatile scaffold for multiple complexes and functions // Current Opinion in Structural Biology. 2021. T. 67. C. 78–85.
207. Ulianov S. V., Khrameeva E. E., Gavrillov A. A., Flyamer I. M., Kos P., Mikhaleva E. A., Penin A. A., Logacheva M. D., Imakaev M. V., Chertovich A., Gelfand M. S., Shevelyov Y. Y., Razin S. V. Active chromatin and transcription play a key role in chromosome partitioning into topologically associating domains // Genome Res. 2016. T. 26, № 1. C. 70–84.
208. Iyer-Bierhoff A., Krogh N., Tessarz P., Ruppert T., Nielsen H., Grummt I. SIRT7-Dependent Deacetylation of Fibrillarin Controls Histone H2A Methylation and rRNA Synthesis during the Cell Cycle // Cell Reports. Elsevier, 2018. T. 25, № 11. C. 2946-2954.e5.
209. Shinohara K., Toné S., Ejima T., Ohigashi T., Ito A. Quantitative Distribution of DNA, RNA, Histone and Proteins Other than Histone in Mammalian Cells, Nuclei and a Chromosome at High Resolution Observed by Scanning Transmission Soft X-Ray Microscopy (STXM) // Cells. Multidisciplinary Digital Publishing Institute, 2019. T. 8, № 2. C. 164.
210. Hock R., Furusawa T., Ueda T., Bustin M. HMG chromosomal proteins in development and disease // Trends in Cell Biology. 2007. T. 17, № 2. C. 72–79.
211. Jonas F., Navon Y., Barkai N. Intrinsically disordered regions as facilitators of the transcription factor target search // Nat Rev Genet. Nature Publishing Group, 2025. C. 1–12.
212. Chong S., Dugast-Darzacq C., Liu Z., Dong P., Dailey G. M., Cattoglio C., Heckert A., Banala S., Lavis L., Darzacq X., Tjian R. Imaging dynamic and selective low-complexity domain interactions that control gene transcription // Science. American Association for the Advancement of Science, 2018.
213. Tsaban T., Varga J. K., Avraham O., Ben-Aharon Z., Khramushin A., Schueler-Furman O. Harnessing protein folding neural networks for peptide–protein docking // Nat Commun. Nature Publishing Group, 2022. T. 13, № 1. C. 176.

214. Lyubitelev A. V., Nikitin D. V., Shaytan A. K., Studitsky V. M., Kirpichnikov M. P. Structure and functions of linker histones // *Biochemistry Moscow*. 2016. T. 81, № 3. C. 213–223.

## Приложение А

**Таблица А1** — Соответствия между категориями разработанной эмпирической функционально-ориентированной классификацией SimChrom и терминами базы знаний Gene Ontology

| Категории классификации SimChrom (русс. яз.) | Категории классификации SimChrom (англ. яз.) | Соответствующие термины из GO                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        |
|----------------------------------------------|----------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Центромерные                                 | Centromere-associated                        | CC GO:0000775 chromosome, centromeric region                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         |
| Ремоделеры хроматина                         | Chromatin remodelers                         | MF GO:0140658 ATP-dependent chromatin remodeler activity<br><br>из дочерних терминов "BP GO:0006338 chromatin remodeling" только следующие были отобраны:<br><br>CC GO:0071564 npBAF complex<br>CC GO:0140092 bBAF complex<br>CC GO:0071565 nBAF complex<br>CC GO:0035060 brahma complex<br>CC GO:0140091 mBAF complex<br>CC GO:0110016 B-WICH complex<br>CC GO:0016586 RSC-type complex<br>CC GO:0016514 SWI/SNF complex<br>CC GO:0140093 esBAF complex                                                                                                                                                                                                             |
| (Де)метилирование ДНК                        | DNA (de)methylation                          | MF GO:0035514 DNA demethylase activity<br>MF GO:0009008 DNA-methyltransferase activity<br>MF GO:0070579 5-methylcytosine dioxygenase activity                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        |
| Метаболические процессы ДНК                  | [DNA metabolic processes]                    | -                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    |
| Рекомбинация ДНК                             | DNA recombination                            | из дочерних терминов "BP GO:0006310 DNA recombination" только следующие были отобраны:<br><br>BP GO:0035825 homologous recombination<br>BP GO:0006312 mitotic recombination<br><br>из дочерних терминов "BP GO:0006310 DNA recombination" следующие были не включены:<br><br>GO:0071139 resolution of DNA recombination intermediates<br>GO:0045911 positive regulation of DNA recombination<br>GO:0042150 plasmid recombination<br>GO:0042152 RNA-mediated DNA recombination<br>GO:0000713 meiotic heteroduplex formation<br>GO:0000150 DNA strand exchange activity<br>GO:0000732 DNA strand displacement<br>GO:0051102 DNA ligation involved in DNA recombination |

|                                          |                                   |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        |
|------------------------------------------|-----------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|                                          |                                   | GO:0045910 negative regulation of DNA recombination<br>GO:0000018 regulation of DNA recombination<br>GO:0016444 somatic cell DNA recombination<br>GO:1905951 mitochondrion DNA recombination<br>GO:0000725 recombinational repair<br>GO:0045458 recombination within rDNA repeats<br>GO:1990067 intrachromosomal DNA recombination<br>GO:0042148 DNA strand invasion<br>GO:0006313 DNA transposition<br>GO:0006314 intron homing                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       |
| Репарация ДНК                            | DNA repair                        | из дочерних терминов "BP GO:0006281 DNA repair" только следующие были отобраны:<br><br>CC GO:0009380 excinuclease repair complex<br>CC GO:1990391 DNA repair complex<br>CC GO:0097196 Shu complex<br>MP GO:0000725 recombinational repair<br>MP GO:0006290 pyrimidine dimer repair<br>MP GO:0006307 DNA alkylation repair<br>MP GO:0006284 base-excision repair<br>MP GO:0006289 nucleotide-excision repair<br>MP GO:0006302 double-strand break repair<br>MP GO:0010213 non-photoreactive DNA repair<br>MP GO:0006301 postreplication repair<br>MP GO:0000012 single strand break repair<br>MP GO:0036297 interstrand cross-link repair<br>MP GO:0006298 mismatch repair<br>MP GO:0070914 UV-damage excision repair<br>MP GO:0141168 chromosomal 5-methylcytosine DNA demethylation, oxidative deamination pathway<br><br>из дочерних терминов "BP GO:0006281 DNA repair" следующие были не включены:<br><br>MP GO:0043504 mitochondrial DNA repair<br>MP GO:0000731 DNA synthesis involved in DNA repair<br>MP GO:0045739 positive regulation of DNA repair<br>MP GO:0045738 negative regulation of DNA repair<br>MP GO:0051103 DNA ligation involved in DNA repair<br>MP GO:0106300 protein-DNA covalent cross-linking repair<br>MP GO:0006282 regulation of DNA repair regulates<br>MP GO:0046787 viral DNA repair |
| Репликация ДНК                           | DNA replication                   | BP GO:0033260 nuclear DNA replication                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  |
| ДНК-связывающие транскрипционные факторы | DNA-binding transcription factors | из дочерних терминов "MF GO:0003700 DNA-binding transcription factor activity" только следующие были отобраны:<br><br>MF GO:0000981 DNA-binding transcription factor activity, RNA polymerase II-specific<br>MF GO:0098531 ligand-activated transcription factor activity<br>MF GO:0001217 DNA-binding transcription repressor activity<br>MF GO:0001216 DNA-binding transcription activator activity<br><br>из дочерних терминов "MF GO:0003700 DNA-binding                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           |

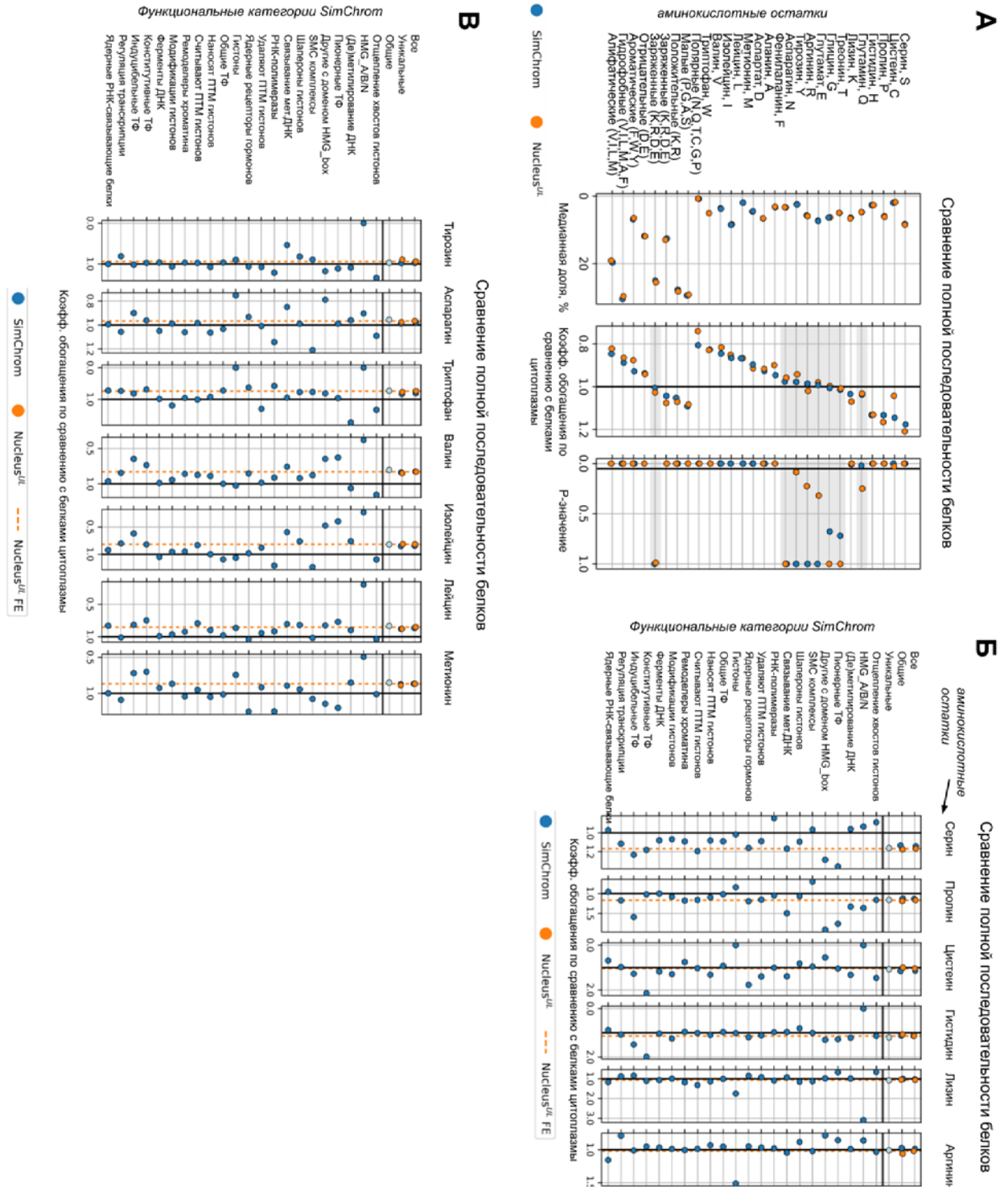
|                                |                               |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  |
|--------------------------------|-------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|                                |                               | transcription factor activity" следующие были не включены:<br><br>BP GO:0051090 regulation of DNA-binding transcription factor activity<br>BP GO:0043433 negative regulation of DNA-binding transcription factor activity<br>BP GO:0051091 positive regulation of DNA-binding transcription factor activity<br>MF GO:0016987 sigma factor activity                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               |
| Ферменты, действующие на ДНК   | DNA-acting enzymes            | из дочерних терминов "MF GO:0140097 catalytic activity, acting on DNA" были отобраны (все):<br><br>MF GO:0051575 5'-deoxyribose-5-phosphate lyase activity<br>MF GO:0140670 cohesin unloader activity<br>MF GO:0008094 ATP-dependent activity, acting on DNA<br>MF GO:0004536 DNA nuclease activity<br>MF GO:0034061 DNA polymerase activity<br>MF GO:0106334 3'-deoxyribose phosphate lyase activity<br>MF GO:0004803 transposase activity<br>MF GO:0003909 DNA ligase activity<br>MF GO:0047839 dATP(dGTP)-DNA purinetransferase activity<br>MF GO:0061775 cohesin loader activity<br>MF GO:0003906 DNA-(apurinic or apyrimidinic site) endonuclease activity<br>MF GO:0033958 DNA-deoxyinosine glycosylase activity<br>MF GO:0033959 deoxyribodipyrimidine endonucleosidase activity<br>MF GO:0070259 tyrosyl-DNA phosphodiesterase activity<br>MF GO:0003913 DNA photolyase activity<br>MF GO:0000150 DNA strand exchange activity<br>MF GO:0003912 DNA nucleotidylexotransferase activity<br>MF GO:0019104 DNA N-glycosylase activity<br>MF GO:0033821 DNA beta-glucosyltransferase activity<br>MF GO:0043336 site-specific telomere resolvase activity<br>MF GO:0090580 phosphodiesterase activity, acting on 3'-phosphoglycolate-terminated DNA strands<br>MF GO:0008907 integrase activity<br>MF GO:0003916 DNA topoisomerase activity<br>MF GO:0033820 DNA alpha-glucosyltransferase activity<br>MF GO:0009008 DNA-methyltransferase activity<br>MF GO:0035514 DNA demethylase activity |
| ДНК-зависимая транскрипция     | [DNA-templated transcription] | -                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                |
| Эухроматин                     | Euchromatin-associated        | CC GO:0000791 euchromatin                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        |
| Общие транскрипционные факторы | General TFs                   | GO:0000126 transcription factor TFIIB complex<br>GO:0000127 transcription factor TFIIC complex<br>GO:0005672 transcription factor TFIIA complex<br>GO:0005669 transcription factor TFIID complex<br>GO:0005673 transcription factor TFIIE complex<br>GO:0005674 transcription factor TFIIF complex<br>GO:0005675 transcription factor TFIIH holo complex<br>GO:0033276 transcription factor TFTC complex                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         |
| Гетерохроматин                 | Heterochromatin-              | CC GO:0000792 heterochromatin                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    |

|                                         |                                    |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  |
|-----------------------------------------|------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|                                         | associated                         |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  |
| Шапероны гистонов                       | Histone chaperones                 | -                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                |
| Модифицирование гистоны                 | Histone modification               | MF GO:0140993 histone modifying activity                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         |
| Удаление ПТМ гистонов                   | Histone PTM erasers                | MF GO:0140789 histone phosphatase activity,<br>MF GO:0140934 histone deubiquitinase activity,<br>MF GO:0032452 histone demethylase activity,<br>MF GO:0160009 histone decrotonylase activity,<br>MF GO:0004407 histone deacetylase activity                                                                                                                                                                                                                                                                                                                                                                                                                                                                      |
| Распознавание ПТМ гистонов              | Histone PTM readers                | MF GO:0035064 methylated histone binding,<br>MF GO:0106153 phosphorylated histone binding,<br>MF GO:0061649 ubiquitin modification-dependent histone binding,<br>MF GO:0070577 lysine-acetylated histone binding                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                 |
| Нанесение ПТМ гистонов                  | Histone PTM writers                | MF GO:0106078 histone succinyltransferase activity,<br>MF GO:0120297 histone dopaminytransferase activity,<br>MF GO:0004402 histone acetyltransferase activity,<br>MF GO:0120295 histone serotonyltransferase activity,<br>MF GO:0061922 histone propionyltransferase activity,<br>MF GO:0035173 histone kinase activity,<br>MF GO:0042054 histone methyltransferase activity,<br>MF GO:0140794 histone arginine deiminase activity,<br>MF GO:0140069 histone butyryltransferase activity,<br>MF GO:0140068 histone crotonyltransferase activity,<br>MF GO:0120301 histone lactyltransferase activity,<br>MF GO:0106229 histone glutaryltransferase activity,<br>MF GO:0140852 histone ubiquitin ligase activity |
| Отщепление хвостов гистонов             | Histone tail cleavage              | -                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                |
| Гистоны                                 | Histones                           | -                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                |
| [HMG]                                   | [HMG]                              | -                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                |
| HMG_A/B/N                               | HMG_A/B/N                          | -                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                |
| Конститутивные транскрипционные факторы | Housekeeping DNA-binding TFs       | -                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                |
| Индукцибельные транскрипционные факторы | Non-Housekeeping DNA-binding TFs   | -                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                |
| Связывание метилированной ДНК           | Methylated DNA binding             | MF GO:0010385 double-stranded methylated DNA binding<br>MF GO:0008327 methyl-CpG binding                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         |
| Ядерные РНК-связывающие белки           | [Nuclear RNA binding proteins]     | -                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                |
| Ядерные рецепторы гормонов              | Nuclear hormone receptors          | MF GO:0004879 nuclear receptor activity<br>MF GO:0003707 nuclear steroid receptor activity                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       |
| Другие семейства с HMG-box              | Other families with HMG-box domain | -                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                |
| Пионерные транскрипционные факторы      | Pioneer TFs                        | -                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                |

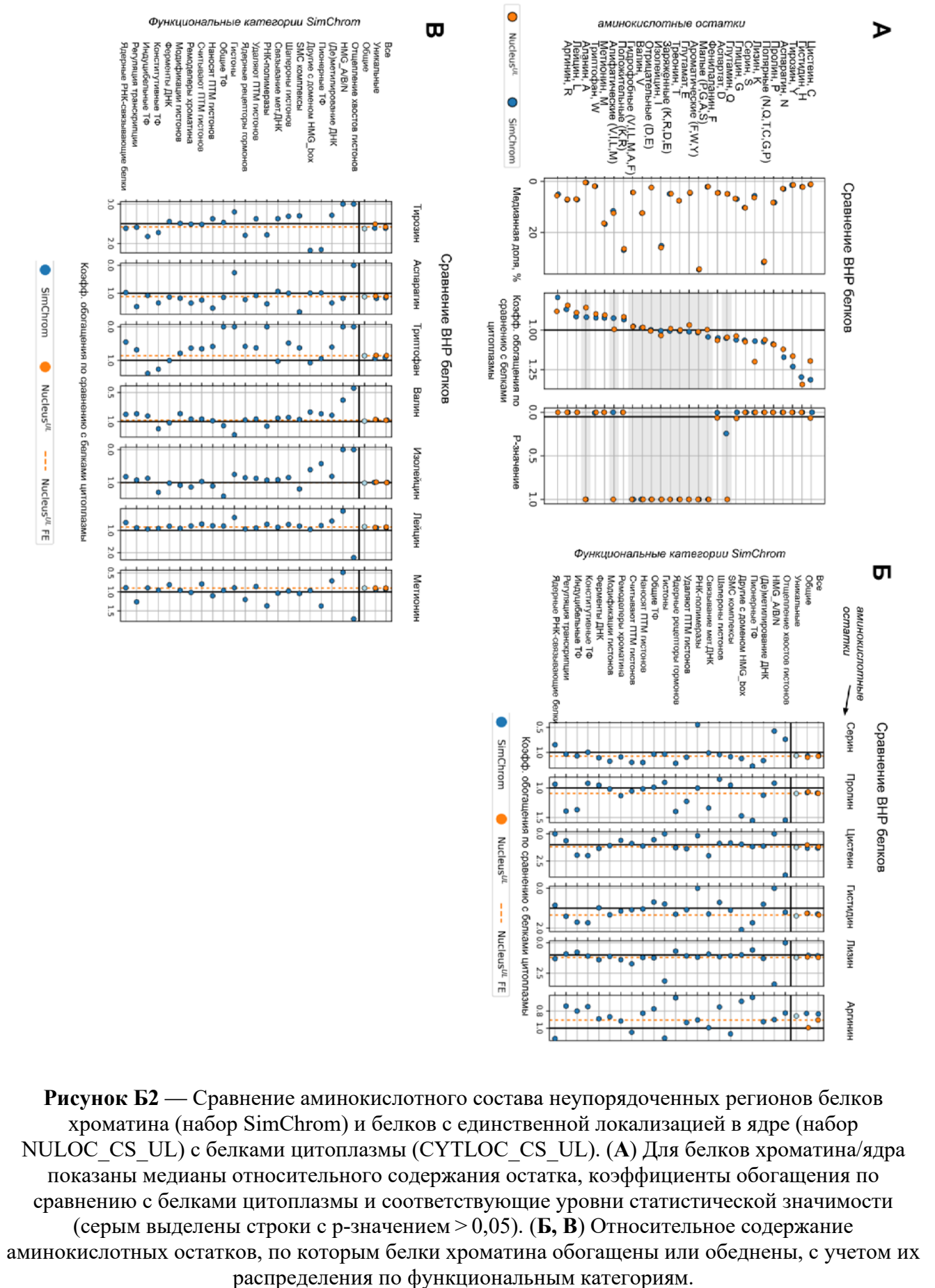
|                        |                             |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       |
|------------------------|-----------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Прерибосомы            | Preribosome-associated      | CC GO:0030684 preribosome                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             |
| Регуляция транскрипции | Regulation of transcription | CC GO:0000120 RNA polymerase I transcription regulator complex<br>CC GO:0090575 RNA polymerase II transcription regulator complex<br>CC GO:0090576 RNA polymerase III transcription regulator complex                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                 |
| Модификаторы РНК       | RNA modification            | из дочерних терминов "BP GO:0009451 RNA modification" только следующие были отобраны:<br><br>BP GO:0016556 mRNA modification<br>BP GO:0035513 oxidative RNA demethylation<br>BP GO:0040031 snRNA modification<br>BP GO:0006400 tRNA modification<br>BP GO:0000154 rRNA modification<br>BP GO:0001510 RNA methylation<br>BP GO:1990884 RNA acetylation<br><br>из дочерних терминов "BP GO:0009451 RNA modification" следующие были не включены:<br><br>BP GO:1900865 chloroplast RNA modification<br>BP GO:1900864 mitochondrial RNA modification<br>BP GO:0016553 base conversion or substitution editing<br>BP GO:0070706 RNA nucleotide deletion<br>BP GO:0070705 RNA nucleotide insertion<br>BP GO:0001522 pseudouridine synthesis |
| РНК полимеразы         | RNA polymerases             | из дочерних терминов "CC GO:0055029 nuclear DNA-directed RNA polymerase complex" только следующие были отобраны:<br><br>CC GO:0005736 RNA polymerase I complex<br>CC GO:0005665 RNA polymerase II, core complex<br>CC GO:0005666 RNA polymerase III complex                                                                                                                                                                                                                                                                                                                                                                                                                                                                           |
| Процессинг РНК         | RNA processing              | из дочерних терминов "BP GO:0006396 RNA processing" только следующие были отобраны:<br><br>CC GO:0000178 exosome (RNase complex)<br>BP GO:0031123 RNA 3'-end processing<br>BP GO:0180035 lncRNA processing<br>BP GO:0000966 RNA 5'-end processing<br>BP GO:0000379 tRNA-type intron splice site recognition and cleavage<br>BP GO:0070918 regulatory ncRNA processing<br>BP GO:0008380 RNA splicing<br>BP GO:0006397 mRNA processing<br><br>из дочерних терминов "BP GO:0006396 RNA processing" следующие были не включены:<br><br>BP GO:0000963 mitochondrial RNA processing<br>BP GO:0031425 chloroplast RNA processing<br>BP GO:0140235 RNA polyadenylation at postsynapse                                                         |
| СМC комплексы          | SMC complexes               | CC GO:0008278 cohesin complex                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         |

|                         |                           |                                                                                                                                                                                                                                                                                                                                                                                                                                    |
|-------------------------|---------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|                         |                           | CC GO:0000796 condensin complex<br>CC GO:0030915 Smc5-Smc6 complex                                                                                                                                                                                                                                                                                                                                                                 |
| Теломеры                | Telomere-associated       | BP GO:0034397 telomere localization<br>BP GO:0032200 telomere organization<br>CC GO:0000781 chromosome, telomeric region                                                                                                                                                                                                                                                                                                           |
| Инактивация X-хромосомы | X-chromosome inactivation | из дочерних терминов "BP GO:0007549 sex-chromosome dosage compensation" только следующие были отобраны:<br><br>BP GO:0042714 dosage compensation complex assembly<br>BP GO:0009048 dosage compensation by inactivation of X chromosome<br>CC GO:0046536 dosage compensation complex<br>BP GO:0009047 dosage compensation by hyperactivation of X chromosome<br>BP GO:0042464 dosage compensation by hypoactivation of X chromosome |

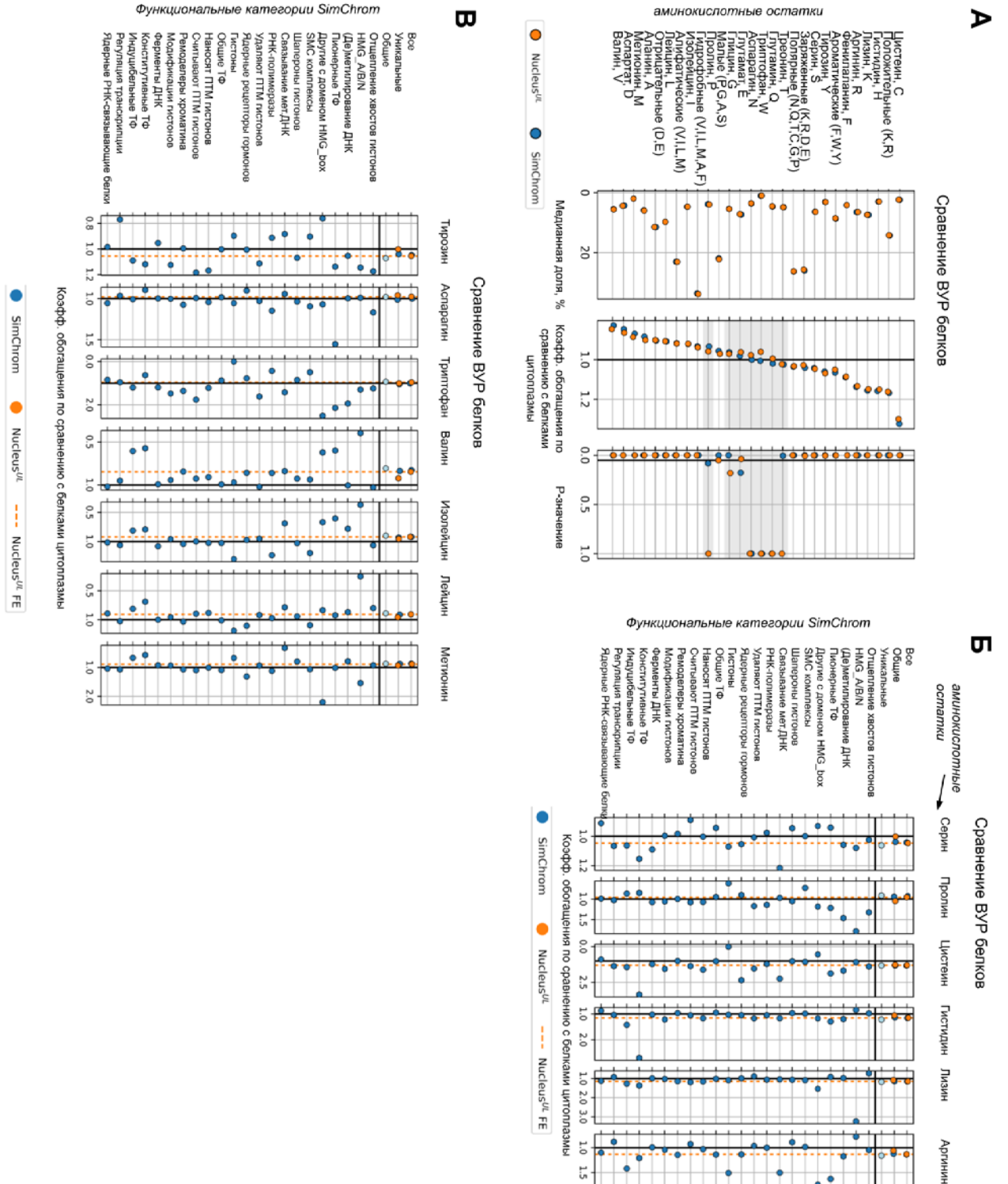
## Приложение Б



**Рисунок Б1** — Сравнение аминокислотного состава белков хроматина (набор SimChrom) и белков с единственной локализацией в ядре (набор NULOC\_CS\_UL) с белками цитоплазмы (CYTLOC\_CS\_UL) по полным последовательностям белков. **(А)** Для белков хроматина/ядра показаны медианы относительного содержания остатка, коэффициенты обогащения по сравнению с белками цитоплазмы и соответствующие уровни статистической значимости (серым выделены строки с р-значением  $> 0,05$ ). **(Б, В)** Относительное содержание аминокислотных остатков, по которым белки хроматина обогащены или обеднены, с учетом их распределения по функциональным категориям.



**Рисунок Б2** — Сравнение аминокислотного состава неупорядоченных регионов белков хроматина (набор SimChrom) и белков с единственной локализацией в ядре (набор NULOC\_CS\_UL) с белками цитоплазмы (CYTLOC\_CS\_UL). (А) Для белков хроматина/ядра показаны медианы относительного содержания остатка, коэффициенты обогащения по сравнению с белками цитоплазмы и соответствующие уровни статистической значимости (серым выделены строки с р-значением > 0,05). (Б, В) Относительное содержание аминокислотных остатков, по которым белки хроматина обогащены или обеднены, с учетом их распределения по функциональным категориям.



**Рисунок БЗ** — Сравнение аминокислотного состава внутренне упорядоченных регионов белков хроматина (набор SimChrom) и белков с единственной локализацией в ядре (набор NULOC\_CS\_UL) с белками цитоплазмы (CYTLOC\_CS\_UL). (А) Для белков хроматина/ядра показаны медианы относительного содержания остатка, коэффициенты обогащения по сравнению с белками цитоплазмы и соответствующие уровни статистической значимости (серым выделены строки с  $p$ -значением  $> 0,05$ ). (Б, В) Относительное содержание аминокислотных остатков, по которым белки хроматина обогащены или обеднены, с учетом их распределения по функциональным категориям.