МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
имени М.В.ЛОМОНОСОВА

БИОЛОГИЧЕСКИЙ ФАКУЛЬТЕТ
Кафедра биоинженерии

Грибкова Анна Кирилловна

ПОСТРОЕНИЕ И АНАЛИЗ ИНТЕРАКТОМА НУКЛЕОСОМ И БЕЛКОВ ХРОМАТИНА

CONSTRUCTION AND ANALYSIS OF AN INTERACTOME BETWEEN NUCLEOSOMES AND CHROMATIN PROTEINS

Выпускная квалификационная работа магистра

Научный руководитель:
канд. физ.-мат. наук, вед. науч. сотр.
А.К. Шайтан

Москва - 2019 г.

Оглавление

Список сокращений:

АХ/МС - аффинная хроматография/масс-спектрометрия

ББВ - белок-белковые взаимодействия

БД - база данных

ПТМ - пост-трансляционные модификации

1. Введение

Хроматин представляет собой комплекс ДНК с белками, посредством которых ДНК компактизируется. Взаимодействия между нуклеосомными белками (гистонами) и прочими белками хроматина (например, транскрипционными факторами, шаперонами, ремоделерами и т.д.) определяют доступность хроматина и обуславливают экспрессию генов. Исследование белок-белковых взаимодействий (ББВ) на нуклеосомном уровне способствует пониманию работы клеточной машинерии, поиску и идентификации фенотипических проявлений нарушений ББВ, а также аннотации белков с неизвестными функциями.

Современные базы данных (БД) ББВ содержат огромное количество информации, полученной как экспериментальными методами, так и предсказательными, а также за счет перенесения информации с генетических родственных организмов. К недостаткам высокопроизводительных экспериментальных методов можно отнести низкую специфичность взаимодействий и большое количество ложноположительных результатов. Тем не менее, гистоны обладают рядом специфических черт, которые влияют на идентификацию их взаимодействий и предполагают более аккуратный анализ.

1.1. Цель работы

Целью данной работы являлось построение нуклеосомного интерактома человека по информации из трех баз данных (STRING, BioGrid, IntAct), и его анализ на основе разработанной иерархической классификации белков хроматина.

1.2. Задачи

1. Создать обновленный список гистонов человека.

2. Загрузить и обработать информацию о белок-белковых взаимодействиях из баз данных STRING, IntAct, BioGrid. Выбрать порог вероятностной оценки для БД STRING.
3. Разработать иерархическую функциональную классификацию белков, взаимодействующих с нуклеосомой.
4. Провести качественный и количественный анализ полученного интерактома.

2. Обзор литературы

2.1. Нуклеосома

В клетках эукариот молекула ДНК занимает ограниченный объем внутриядерного пространства. На первом уровне компактизации ДНК повторяющейся элементарной единицей хроматина является нуклеосомный кор - 145-147 н.п. ДНК, обвивающих 1,65 раз гистоновый октамер. В ядро нуклеосомы входят белки 4-ех гистоновых типов: H2A, H2B, H3, H4, которые объединяются в гетеродимеры: два H3-H4 и два H2A-H2B.

Белковая часть кора имеет вид цилиндра радиусом 65 Å и высотой 60 Å. Нуклеосомный кор с линкерной ДНК длиной приблизительно 50 н.п. формирует нуклеосому (Рис. 1). Линкерный гистон H1 находится снаружи от нуклеосомы и отличается от других типов гистонов наличием глобулярного домена. Он взаимодействует с ДНК в местах входа и выхода из нуклеосомы и влияет на ее доступность. Нуклеосома с линкерным гистоном называется хроматосомой.

Нуклеосомная ДНК содержит 14 сайтов связывания с гистонами, расположенных в малой бороздке. Преимущественно, эти контакты представлены водородными связями между гистонами и сахарофосфатным остовом нуклеиновой кислоты. Помимо водородных связей, взаимодействие ДНК и белка проявляется в гидрофобных взаимодействиях и солевых мостиках. Концевые части гистонов (N-концы у всех типов гистонов и C-концы у H2A и H2B) длиной 15-30 аминокислотных остатков называют гистоновыми хвостами, выступающими за пределы нуклеосомного кора. В них сосредоточены положительно заряженные аминокислоты, преимущественно лизин и аргинин. В гистоновом октамере выделяют участок, называемый «acidic patch», который состоит из 8-ми отрицательно заряженных аминокислотных остатков, расположенных на гистонах H2A и H2B. Предполагается, что с ним взаимодействует хвост гистона H4 соседней нуклеосомы при формировании хроматиновых фибрилл, а также некоторые хроматиновые белки.

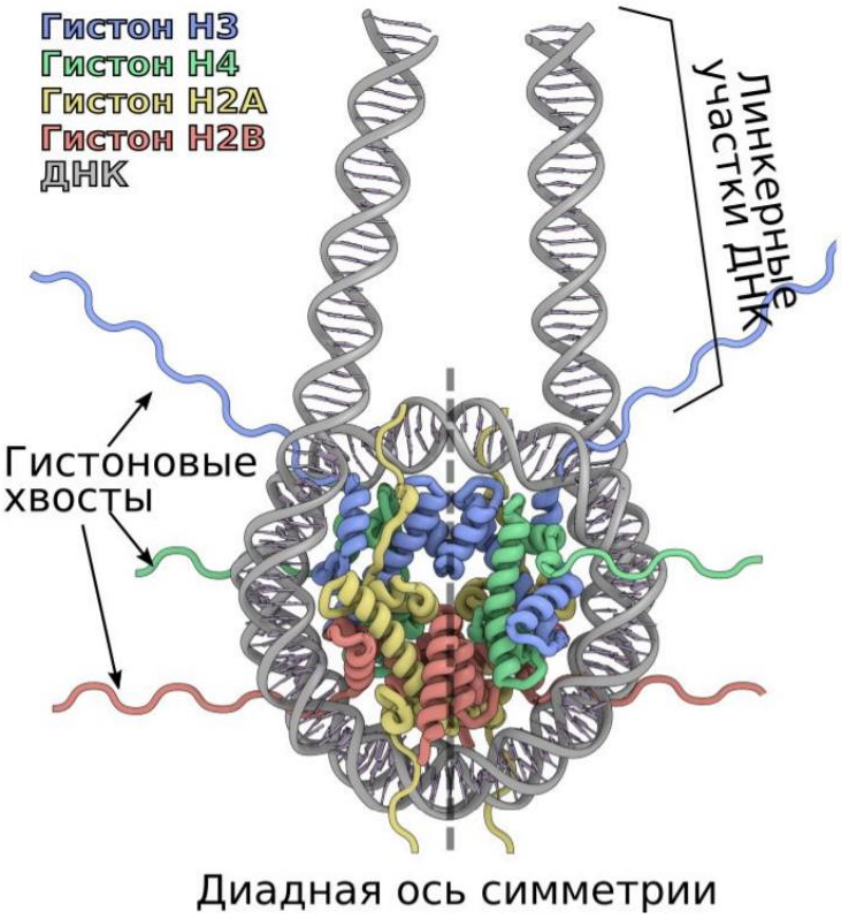

Рис. 1. Визуализированная нуклеосома, построенная по структуре с кодом 1kx5 из банка данных PDB (Davey et al. 2002). (Armeev et al., 2019)

2.2. Гистоны и их особенности

К основным ядерным белкам относят гистоны. По типу гистоны делятся на H1, H2A, H2B, H3, H4, а по классу - на канонические и вариантные. У животных канонические гистоны кластеризованы и транскрибируются только в S-фазе митоза. Отличительной особенностью канонических гистонов является отсутствие интронов в мРНК и отсутствие типичного для 3’-конца пре-мРНК поли(А)-хвоста. Его роль выполняет петлеподобная структура, регулирующая процессинг.

Паралогами по отношению к каноническим гистонам являются гистоновые варианты. Они важны для правильного функционирования многих процессов, включающих транскрипцию, хромосомную сегрегацию и репарацию

ДНК. Включение гистоновых вариантов в нуклеосому обычно не зависит от фазы клеточного цикла. Транскрипты гистоновых вариантов имеют типичное строение, то есть содержат и интроны, и полиА-хвост.

Замена канонических гистонов на вариантные обладает пространственной, ткане- и временной специфичностью и обуславливает протекание некоторых клеточных процессов. Например: замена H3.3 имеет место в регуляции транскрипции, H3.1 встраивается во время репликаци, CenH3 обуславливает специфичность центромерного региона, а H3.5 вносит вклад в конденсацию хроматина (Venkatesh and Workman, 2015). Ряд вариантных гистонов специфичен для семенников и сперматозоидов. Другой важный пример гистоновых вариантов - H2A.X, фосфорилированный гистон принимает участие в репарации двухцепочечных разрывов ДНК. А ацетилированный H2A.Z преимущественно располагается в промоторных областях генов. Даже нескольких приведенных выше примеров достаточно для понимания, насколько важную и разнообразную роль играет замена канонических гистонов на гистоновые варианты.

Следующее важное свойство гистонов - наличие пост-трансляционных модификаций, которые могут располагаться как на коровой части гистона, так и на гистоновых хвостах. Так, ацетилирование и фосфорилирование уменьшает заряд белков и, соответственно, степень взаимодействия с ДНК, приводя к изменению структуры хроматина. Другие модификации (например, метилирование, убиквитинилирование, сумоилирование и т.д.) влияют на узнавание и связывание ДНК с гистоновыми шаперонами и факторами ремоделирования хроматина. Интересно, что одни и те же гистоновые аминокислоты могут модифицироваться разными метками, что приводит к осуществлению противоположных процессов. Одни и те же метки, находящиеся на разных аминокислотах также могут приводить к совершенно противоположным результатам.

Для ряда гистонов характерно наличие сплайс-изоформ, то есть образование разных белков с одной пре-мРНК в результате альтернативного сплайсинга.

2.2.1 Биотипы генов по классификации по Ensembl

Несмотря на наличие копий гистоновых генов и сплайс-изоформ, не все транскрипты являются белок-кодирующими. Гистоновые гены могут быть следующих типов (по классификации генов и транскриптов Ensembl):

1. Белок-кодирующие гены - содержат открытую рамку считывания (ОРС).

1.1. Отдельный подтип - белки, подвергаемые нонсенс-опосредованному распаду мРНК (NMD - nonsense-mediated mRNA decay), то есть мРНК, содержащие в неправильных местах стоп-кодоны или неправильно сплайсированные. Помечаются как NMD и в тех случаях, если кодирующая последовательность транскрипта заканчивается в более чем в 50 н.п. от сайта сплайсинга и если вариант покрывает не всю кодирующую последовательность.

2. Процессируемые транскрипты - не содержат ОРС. Сюда относят длинные-некодирующие РНК, некодирующие РНК и неклассифицируемые транскрипты.

3. Псевдогены - последовательности нуклеотидов, содержащие сдвиг рамки считывания или стоп-кодоны, разрушающие ОРС.

3.1 Различают транскрибирующиеся псевдогены, то есть те, для которых гомология белков или геномная структура свидетельствуют о псевдогене, но наличие локус-специфичных транскриптов свидетельствует об экспрессии.

Среди них выделяют:

3.1.1. Процессированные псевдогены (ретропсевдогены) появились в результате встраивания в хромосомы ДНК, полученной в результате обратной транскрипции, поэтому интронов не содержат.

3.1.2. Непроцессированные (дуплицированные) появились в результате дуплицирования гена (во время гомологичной рекомбинации, например).

Потеря функциональности дуплицированного гена обычно мало заметна, потому что есть функционирующая копия. Могут содержать интроны.

3.1.3. Унитарные псевдогены - те гены, нормальным транскрипции или трансляции которых мешают различные мутации. Процесс псевдогенизации унитарных напоминает процесс непроцессированных, однако в этом случае не происходит предварительной дупликации генов. Унитарные псевдогены видоспецифичны, имеют активных ортологов среди других видов.

2.3. Негистоновые белки интерактома

Далее будет произведено краткое перечисление негистоновых белков хроматина и других функциональных групп нуклеосомного интерактома, принимающих участие в различных уровнях компактизации хроматина или способствующих протеканию каких-либо процессов в ядре. Так, транспорт гистонов, их хранение, формирование и разборку нуклеосом обеспечивают гистоновые шапероны. Хроматиновые ремоделеры обеспечивают перемещение нуклеосом вдоль ДНК за счет энергии АТФ. Выделяют следующие классы: SWI/SNF, ISWI, CHD и INO80 (по типу АТФазной субъединицы). Другую крупную группу образуют белки, наносящие и удаляющие ПТМ, а также белки, отрезающие гистоновые хвосты. Модификациям подвергается и ДНК. Считывание ПТМ приводит к привлечению других белков, в том числе транскрипционных факторов, транскрипционной и репликационной машинерии. К структурным белкам относятся белки ядерной оболочки (внутренней и наружней ядерной мембраны, ядерного порового комплекса), ядерной ламины и матрикса, а также белки теломерных и центромерной областей хромосомы. Белки внутренней ядерной мембраны не только способствуют контактам между ядерной оболочкой, хроматином и ламиной, но и способны оказывать влияние на экспрессию генов (Hetzer 2010).

Появление методов исследования пространственной организации ядра, как, например, 3C (chromosome conformation capture - захват конформации

хромосом) (Dekker et al. 2002) и Hi-C (Lieberman-Aiden et al. 2009), способствовало пониманию функционирования архитектурных белков хроматина. Таким образом, к группе архитектурных белков хроматина относят, во-первых, белки, способствующие образованию гетерохроматина: HP1, белки группы Polycomb, MENT, MeCP2, Sir. Другой не менее важной группой архитектурных белков, регулирующих динамику нуклеосомных фибрилл, являются HMG-белки. Выделяют нескольких групп: HMGA - содержат мотив AT-hook; HMGB - связываются с малой бороздкой ДНК, что приводит к ее резкому изгибу; HMGN - связываются с нуклеосомами.

EpiFactors (http://epifactors.autosome.ru/) (Medvedeva et al. 2015) - курируемая БД эпигенетических регуляторов, содержит 815 белков, в том числе 92 гистона и 3 протамина (в ядрах сперматозоидов они замещают гистоны), и 69 белковых комплексов. Для 789 генов представлены значения экспрессий как в различных нормальных тканях, так и в раковых (458 первичных культур клеток человека, 255 раковых клеточных линий и 134 человеческих посмертных образцов тканей).

В БД разработана аннотационная схема для эпигенетических регуляторов, включающая в себя: гистоны и протамины; белки, наносящие/стирающие/считывающие ПТМ, хроматиновые ремоделеры и гистоновые шапероны, белки, модифицирующие ДНК или РНК (но участвующие не в процессинге РНК), кофакторы в эпигенетических комплексах; белки скаффолда, транскрипционные факторы и белки группы polycomb. Источники информации - литературные данные, запросы по ключевым словам в Uniprot, привлечение интеллектуальный анализа текстов. Также приведены данные доменного обогащения как для всей БД, так и для отдельных функциональных категорий. БД существует в виде несвязанных таблиц гистонов и протаминов, эпигенетических регуляторов и данных о белковых комплексах, то есть содержит справочную информацию (идентификаторы, функциональная категория и домены) для каждого гена, но не содержит данных о взаимодействиях между гистонами и эпигенетическими регуляторами.

2.4. Специфические домены и мотивы

Доменом белка называют элемент третичной структуры, который функционирует, пространственно укладывается и эволюционирует независимо от других частей белка. Белки являются комбинацией различных функциональных доменов, что и обуславливает их функции. На данный момент база данных доменов Pfam (El-Gebali et al. 2019) насчитывает 17929 доменных семейств. Наиболее показательна доменная организация в структурах белков, считывающих ПТМ.

Здесь и далее ПТМ будут обозначаться следующим образом: тип гистона, аминокислотный остаток, метка и ее количество, например, метильная группа на 4-ом лизине гистона H3 - H3K4me. Так, отсутствие метки H3K4me0 узнают домены: PHD, WD40 (приводит к дальнейшему ацетилированию гистонов) и ADD (приводит к метилированию ДНК). H3K4me узнается доменами Chromo (находится на ремоделере хроматина), PHD (в зависимости от белка и других доменов, функции широко варьируются), Tudor (деметилирование или ацетилирование гистонов), MBT, Zf-CW (Yun et al. 2011).

Последнее время список белков, узнающих ПТМ на гистонах стремительно растет, также были обнаружены примеры комбинаторного узнавания меток. На активность считывания ПТМ могут влиять соседние ПТМ. Одним из примеров является ингибирование связывания хромодомена на HP1 (Heterochromatin protein1) с H3K9me3 при наличии метки H3S10ph, что происходит во время митоза. Другой пример, H3T3ph (ПТМ - фосфат) препятствует связыванию домена DCD на CHD1 с H3K4me3 (Flanagan et al. 2005). H3K4me препятствует связыванию H3-специфического домена PHD fingers и ADD, в то время как H4K8ac увеличивает связывание бромодомена с H4K5ac (Morinière et al. 2009).

Выделяют цис-комбинаторное прочтение меток, то есть находящихся на одном гистоновом хвосте, например, каждый бромодомен на TAF1 узнает по одному ацетилированному лизину на одном и том же диацетилированном

гистоне H4 (Jacobson et al. 2000). Совместное узнавание разных меток разными белками было показано для PHD-бромодоменовых кассет белков TRIM24 и TRIM33. Так, домен PHD finger белка TRIM33 связывается с H3K9me3, а бромодомен с H3K18ac на одном и том же гистоне H3 (Xi et al. 2011).

Комбинаторное считывание меток эффектором и каталитическим гистон-связывающим доменом было показано для гистоновых деметилаз PHF8 и KIAA1718, где связывание домена PHD finger с H3K4me3 в дважды метилированных гистонах H3K4me3K9me2 и H3K4me3K27me2 способствует связыванию каталитических доменов с K9me2 и K27me2, соответственно (Horton et al. 2010).

Еще более сложными оказываются узнавания ПТМ на гистонах множеством белков, находящихся в хроматиновых комплексах. Такие механизмы способствуют очень специфическим взаимодействиям. Более того, многие ядерные комплексы являются динамичными по отношению к субъединичному составу, что является мощным механизмом регулирования функций хроматина (Musselman et al. 2012).

Узнавание ПТМ может происходить также на разных гистоновых хвостах или даже на разных нуклеосомах. Такие события транс-считывания уникальны и известно лишь несколько примеров. Так, субъединица BPTF комплекса хроматин-ремоделинга NURF, содержит домены PHD finger, специфичный к H3K4me3, и бромодомен, связывающийся с H4K16ac, соединенные жестким альфа-спиральным линкером. Данное событие впервые было показано с помощью молекулярного моделирования (Ruthenburg et al. 2007), а затем подтверждено экспериментально на клеточных линиях (Ruthenburg et al. 2011).

Соответственно, интересно перейти от взаимодействий между структурами (выраженных в названии генов), к взаимодействиям на уровне доменов, что не только поможет идентифицировать функции белков, полученных в высокопроизводительных экспериментах, но и определить функциональные взаимоотношения между взаимодействующими белками.

Множество исследований демонстрируют, что доменно-доменные взаимодействия (DDI) из разных экспериментов сильнее согласуются друг с другом, чем их соответствующие ББВ (Memišević, Wallqvist, and Reifman 2013). Было показано, что использование доменов и их взаимодействий для прогнозирования белок-белковых взаимодействий и наоборот считается довольно надежным подходом (Wojcik and Schächter 2001).

Мотивы представляют собой короткие аминокислотные последовательности аминокислот, выполняющие определенные биологические функции, например, являются сайтами узнавания или сигнальными пептидами. В контексте нуклеосомного интерактома будут рассмотрены следующие мотивы: AT-hook (..[RPKST]GRP[RPKS]) (Aravind and Landsman 1998), CENP-C (R....P..[YFW]W) (Kato et al. 2013) и SPKK (SP[RK][RK]) (Churchill and Suzuki 1989), где точка - это любая аминокислота, а квадратные скобки обозначают любую из аминокислот, находящихся внутри.

Мотив AT-hook связывается с ДНК в малой бороздке, богатой аденином и тимином, и часто соседствует с функциональными доменами белков хроматина, например, с доменами гистоновой укладки, гомеодоменом и цинковыми пальцами. С АТ-богатыми регионами малой бороздки ДНК связывается и мотив SPKK, что впервые было продемонстрировано на гистонах H1 и H2B морского ежа. Мотив CENP-C используется для узнавания центромерными белками центромерной нуклеосомы, содержащей вариант гистона H3 CENP-A. Таким образом, поиск вышерассмотренных мотивов в последовательностях взаимодействующих с нуклеосомой белков является дополнительным мощным инструментом качественного анализа.

2.5. Белок-белковые взаимодействия

Взаимодействия белков в клетке обуславливают протекание различных биологических процессов. Изучение ББВ расширяет знания о функциональности белков, молекулярных основах различных заболеваний и способствует предсказанию новых мишеней для таргетных терапий. В настоящее время ББВ

является одной из ключевых тем системной биологии. К результатам ББВ можно отнести, например, изменение кинетических свойств ферментов, перенос субстратов между белками, образование новых сайтов связывания эффекторных молекул, инактивация белка, изменение специфичности белка в зависимости от его партнеров и регуляция выше- нижележащих уровней сигнальных путей.

С развитием высокопроизводительных экспериментальных методов выявления взаимодействующих белков, количество данных заметно увеличилось, впрочем, как и количество ложноотрицательных и ложноположительных результатов. Широкий набор экспериментальных методов позволяет выявлять физические взаимодействия, тогда как предсказательные методики способствуют по большей части накоплению знаний о функциональных взаимодействиях.

2.5.1 Экспериментальные методы

Экспериментальные методики определения ББВ можно разделить на генетические и биохимические. Используется следующая терминология: исследуемый белок называют «наживкой» (bait protein), а белки-партнеры, которые проверяются на взаимодействие с белком-«наживкой» - белки-«добыча» (prey protein). Если рассматривать экспериментальные методы, используемые для получения информации о ББВ нуклеосомного интерактома (материалы баз данных BioGrid, IntAct), то среди лидирующих техник будут аффинная хроматография белков со специфической меткой с последующей масс-спектрометрической идентификацией (АХ\МС) - около 40-ка % от всех экспериментальных взаимодействий и около 15 % от всех взаимодействий. К преимуществам АХ можно отнести высокую чувствительность, возможность проверить все пробы белков-«добычи» на взаимодействие со связанным исследуемым белком-«наживкой» в колонке и идентифицировать взаимодействующие мультисубъединичные белковые комплексы. К недостаткам относят возможное влияние метки на функцию белка и соблюдение строгих условий для предотвращения неспецифических взаимодействий.

Метод перекрестных сшивок (cross-linking study) предоставил 14 % экспериментальных взаимодействий, однако количество ложно-положительных взаимодействий увеличивается в связи с тем, что сшиться могут белки, находящиеся пространственно рядом, но не взаимодействующие. 14 % экспериментальных взаимодействий получено методами аффинного захвата с последующей детекцией вестерн-блоттингом (важно, что здесь используются только коэкспрессирующиеся белки) и реконструирования комплексов (подразумевается метод коиммунопреципитации и оптические биосенсоры на эффекте поверхностного плазмонного резонанса). В коиммунопреципитации используется экстракт цельных клеток, где белки присутствуют в их нативной форме в сложной смеси клеточных компонентов. Кроме того, использование эукариотических клеток способствует ПТМ, которые могут быть необходимыми для осуществления взаимодействия. Меньше 1-го % взаимодействий детектировано методом фагового дисплея.

Для нуклеосомного интерактома большое значение имеют ПТМ и белки, с ними взаимодействующие, а также сопутствующие этому молекулярные и биологические процессы. Методами выявления биохимической активности (термин из BioGrid) можно определить тип модификации, наносимой одним из белков. В нуклеосомном интерактоме лидерами являются следующие процессы - фосфорилирование, ацетилирование и метилирование (в диапазоне от 150-ти до 250-ти экспериментальных записей в интерактоме для каждой из модификации), для нескольких взаимодействий отмечено сумоилирование.

К наиболее используемым методикам выявления белков, считывающих гистоновые ПТМ относятся гистоновые микрочипы - гистоновые пептиды с различными ПТМ находятся на целлюлозной подложке или связаны с подложкой комплексом стрептавидин-биотин; белки, меченные глутатионтрансферазой (ГСТ) или His (гистидином) инкубируют и далее выявляют вестерн-блотом или флуоресценцией. Метод подходит для широкомасштабных скринингов, позволяет идентифицировать в том числе и связывание с комбинациями ПТМ, однако требует очистки исследуемых белков.

Подходы с магнитными частицами - создается библиотека случайных ПТМ-обогащенных гистоновых пептидов, далее исследуемые белки инкубируют с частицами, на которых закреплена библиотека, и детектируют МС. К преимуществам подхода можно отнести идентификацию синергических или ингибирующих комбинаций ПТМ, ограничения схожи с гистоновыми микрочипами. Изотопное мечение белков в культуре клеток (SILAC) - ядерный экстракт, выращенный на среде с меченными АК инкубируют с гистоновыми пептидами, содержащими ПТМ или рекомбинантными нуклеосомами, далее идентифицируют белки с помощью МС. Сложности метода заключаются в идентификации белков и доменов, ответственных за связывание.

Разумеется, одной из основных техник выявления ББВ до сих пор остается дрожжевой двугибридный анализ. Метод заключается в пришивание к исследуемому белку-«наживке» ДНК-связывающий домен транскрипционного фактора, а к белку-«добыче» - активационный домен того же самого транскрипционного фактора. При взаимодействии белка-«наживки» и белка-«жертвы» транскрипционный фактор будет активен, что приведет к транскрипции репортерного гена. Для последующей селекции тех дрожжевых колоний, где белки провзаимодействовали, обычно в качестве репортерного гена используют ген устойчивости к антибиотику (и рост колоний на среде с антибиотиком) или ген β-галактозидазы (для бело-голубой селекции). И хотя метод является относительно простым и позволяет проводить широкомасштабные скрининги, наличие большого количества ложноположительных и ложноотрицательных результатов, о которых речь пойдет дальше, а также низкий процент воспроизводимости результатов, дает веские основания для перепроверки взаимодействий более специфичными методами.

Ложноположительные взаимодействия обязаны своим появлением следующим экспериментальным ошибкам: неспецифические сшивки между белками, которые находятся рядом, но не взаимодействуют (метод перекрестных сшивок); высокая экспрессия белка, выполняющего роль «наживки»;

использование клеточного лизата, приводящее к детекции взаимодействий белков, находящихся в разных клеточных компартментах или экспрессирующихся в разные временные промежутки. Ложноотрицательные результаты могут быть следствием ошибок при конструировании белков слияния, особенно в широкомасштабных экспериментах, а также возможных изменений в экспрессии и фолдинге слитых белков. Под белком слияния подразумевается исследуемый белок или его взаимодействующий партнер из дрожжевого двугибридного анализа, слитый с одним из доменов фактора транскрипции.

Существует ряд методов проверки адекватности результатов высокопроизводительных экспериментов: надежность профиля экспрессии (EPR индекс - сравнивают профили экспрессии РНК взаимодействующих по данным эксперимента белков с профилями экспрессии для известных взаимодействующих и не взаимодействующих пар белков), метод верификации паралогов (PVM) - взаимодействие подтверждается, если белки имеют взаимодействующих паралогов. ERP оценивает наборы взаимодействий, а PVM отдельные взаимодействия (Deane et al. 2002). Также применяют методы локализации белка (PLM) (взаимодействующие белки должны принадлежать одному клеточному компартменту), меры общности взаимодействия IG1, IG2 (interaction generality). Оценка IG1 основывается на предположении, что если два белка взаимодействуют, а их другие партнеры при этом нет, то результат, скорее всего, является ложноположительным. Мера IG2 учитывает так же и топологию взаимодействий. Было показано, что методы, присваивающие оценки отдельным взаимодействиям, работают корректнее, чем методы, оценивающие наборы данных (Suthram et al. 2006).

Одна из проблем экспериментальных методов, подрывающая доверие к экспериментальным интерактомам - низкая воспроизводимость результатов, даже после исправления технических проблем и учета ложноположительных взаимодействий (von Mering et al. 2002). Например, использование методик AX и BioID (активированное мечение биотином белков-партнеров) для хроматин-

ассоциированных белков показало довольно малое пересечение белков-партнеров для канонических гистонов (Lambert et al. 2015). Использование разных методик помогает уловить взаимодействия, характерные для различных процессов и для разных стадий жизненного цикла клетки.

Таким образом вопрос, дополняют ли все экспериментальные данные друг друга или же считать действительным взаимодействием то, что показано в нескольких экспериментах, остается нерешенным.

2.5.2 Предсказательные методы

Идентифицировать ББВ можно не только экспериментальными методами, но и биоинформатическими. Здесь будет дан очень краткий обзор возможных направлений, список актуальных для нуклеосомного интерактома методов будет представлен в разделе описания базы данных STRING. Прежде всего, при поиске ББВ используют методы анализа геномного контекста: анализ консерватизма расположения генов в геномах; сравнительный анализ белковых пар и гомологичных белков слияния в геномах. Среди методов анализа эволюционных взаимоотношений белков выделяют: анализ корреляционной эволюции функционально связанных белков, оценку подобия филогенетических профилей.

Существует ряд сервисов для анализа возможных взаимодействий на основе сходства последовательностей, пространственной структуры белков и доменной организации: считается, что если два белка взаимодействуют, то гомологичные им, со структурами/последовательностями/доменами, похожими на первые два, тоже будут взаимодействовать.

2.6. Базы данных белковых взаимодействий

На данный момент существует огромное количество баз данных, содержащих информацию о ББВ. Есть как специализированные базы данных с информацией по определенному классу белков или заболеванию, так и общие. Базы данных можно разделить на следующие:

1. содержат экспериментально подтвержденные взаимодействия, данные из литературных источников или данные, добавленные авторами самостоятельно,
2. содержат предсказанные с помощью различных алгоритмов ББВ
3. объединяют в себе рассмотренные выше источники

Первичные базы данных независимо собирают информацию, мета-базы данных - агрегируют в себе несколько первичных.

Для выполнения данной работы были выбраны 3 БД (STRING, BioGrid, IntAct), в которые входят как первичные, так и мета-базы данных, как с предсказанными взаимодействиями, так и с экспериментальными.

Сравнивая информацию о ББВ из различных БД, нужно обращать внимание на частные особенности БД, методы и алгоритмы получения информации.

В последней сравнительной работе практически всех существующих БД с ББВ (Bajpai et al. 2019) для анализа использовался список генов, в котором были как ткане- и болезнеспецифичные гены, так и повсеместно встречающиеся. Именно такой разнородный специфичный набор может показать сильные и слабые стороны БД, а также может быть полезен для обобщения полученной информации на гистоны, среди которых встречаются тканеспецифичные гистоновые варианты. Так, в упомянутой выше работе из первично найденных 375 ресурсов для детального анализа были оставлены 16 БД. В зависимости от года исследования и тестового набора белков получаются абсолютно противоположные результаты. Например, при исследовании 6 БД (Lehne and Schlitt 2009) весомое преимущество по количеству и качеству белков-партнеров было у IntAct. Однако последний анализ (Bajpai et al. 2019) показал, что в IntAct меньше 10-ти процентов возможных взаимодействий и очень низкий процент (<1) уникальных взаимодействий.

В другой работе (Mathivanan et al. 2006) по количеству взаимодействий преобладала БД HPRD (Keshava Prasad et al. 2009), в дальнейшем же эта БД не показывала подобных успехов в исследованиях наполнения, что можно связать

с отсутствием обновлений после 2010-го года. Если сравнивать количество упоминаний в статьях используемых для различных анализов БД, то среди лидеров окажутся STRING, IntAct, HPRD, хотя последнее исследование (Bajpai et al. 2019) показало, что наибольший охват ББВ у BioGrid (69,5 %). Следовательно существует необходимость проведения актуальных сравнительных анализов баз данных белок-белковых взаимодействий.

Если говорить о выборе БД как источника информации для дальнейших исследований, то авторы последнего обзора рекомендуют использовать hPRINT (Elefsinioti et al. 2011) и STRING, показавшие не только большое количество уникальных взаимодействий по сравнению с другими БД, но и наибольшее количество предсказанных и экспериментально-подтвержденных взаимодействий, находящихся и в других БД. Более того, hPRINT и STRING являются мета-базами данных, собирая информацию из более специализированных источников, а также содержат свои предсказательные алгоритмы. Было отмечено, что hPRINT несмотря на большое количество пересекающихся с другими БД взаимодействиями, не так часто используется учеными для исследований, что может быть объяснено сравнительно недавним появлением этой БД или же затрудненным доступом по web-ссылке. Среди первичных БД при исследовании взаимодействий тестового набора генов, наибольшее количество уникальных для базы данных взаимодействий было найдено в BioGrid и IntAct. Эти же БД показали наибольшее количество взаимодействий, найденных в других БД: BioGrid (49,5 %) и IntAct (17,2 %).

2.6.1. STRING

База данных STRING (Search Tool for the Retrieval of Interacting Genes/Proteins) (http://string-db.org/) (Szklarczyk et al. 2017) содержит известные и предсказанные ББВ. Основной единицей взаимодействия в STRING является функциональная ассоциация, то есть рассматриваются не только физические взаимодействия, но и участие белков в одних и тех же сигнальных и

метаболических путях, в транскрипционной регуляции или совместное нахождение в мультисубъединичных комплексах.

БД центрирована на белок-кодирующих генетических локусах, соответственно альтернативные сплайс-изоформы и белковые пост-трансляционные модификации не рассматриваются. Идентификатором БД является Ensembl protein ID.

Экспериментальные источники информации о ББВ в базе данных STRING:

1. Данные высокопроизводительных экспериментов. Подавляющее большинство данных приходит из баз IMEx consortium (активные члены консорциума: DIP, HPIDB, IntAct, MBInfo, MINT, MatrixDB, Molecular Connections, I2D, InnateDB,UCL-BHF group, UCL London, UniProt group,Swiss-Prot group, SIB, EMBL-EBI) и базы банных BioGRID. Для белков из наборов данных высокопроизводительных экспериментов вероятностные оценки вычисляются раздельно, так как учитывается частота обнаружения взаимодействующих белков вместе или по отдельности, в то время как взаимодействия, импортированные из баз данных BIND, KEGG, MIPS, получают одинаковые оценки.
2. Базы данных: нахождение пары белков в одном сигнальном пути (данные из Kegg (Kanehisa and Goto 2000) и Reactome (Fabregat et al. 2018)).
3. Интеллектуальный анализ текстов - поиск названий белков в абстрактах PubMed, и в коллекции из более чем трех миллионов полнотекстовых статей. Пары белков получают оценку при упоминании вместе в одной статье/абстракте/предложении (относительно того, как часто они упоминаются отдельно). Значение оценки увеличивается при анализе предложений с помощью обработки методами анализа естественного языка (Natural Language Processing, NLP) и выявлении связующего слова для описания взаимоотношения между белками (Franceschini et al. 2012).
4. Коэкспрессия: данные экспрессии генов, полученные с использованием микрочипов. Используются данные, депонированные в NCBI Gene Expression Omnibus (Barrett et al. 2013)). Значения экспрессий из различных

экспериментов нормализуются и далее рассчитываются корреляции. Чем чаще экспрессия пары белков коррелирует в различных экспериментах при различных условиях, тем выше будет оценка. Подробнее алгоритм анализа данных генной экспрессии описан в (Szklarczyk et al. 2015). В дополнение к данным ДНК-микрочипов, с версии 10.5 STRING также учитываются данные экспрессии RNA-seq (что привело к добавлению информации из этого источника 16-ти видам живых организмов).

Биоинформатические подходы как источники информации о ББВ (анализ геномного контекста), характерны по большей части для геномов прокариот и архей. В STRING используются следующие подходы:

1. Поиск совместно расположенных генов (соседства генов), например, в случае консервативных ко-транскрибируемых оперонов. То есть эти пары генов находятся под общим эволюционным давлением и, следовательно, могут считаться функционально ассоциированными.
2. Поиск слияния генов (фьюжн-событий) - если ортологи исследуемых белков в другом виде слились в один белок-кодирующий ген.
3. Сравнительный геномный филогенетический анализ - то есть белки со схожими филогенетическими профилями (наличию/отсутствию в различных геномах) могут быть функционально связаны (Franceschini et al. 2016).

Еще один немаловажный вклад в оценку взаимодействия заключается в переносе белкового взаимодействия на ортологичные белки других видов. Поиск таких «интерологических» (комбинация английских слов «взаимодействие»и «ортологи») взаимодействий на практике затрудняется высокой частотой генных дупликаций, потерей генов и наличием генных перестроек. Наилучшие кандидаты в двух организмах - ортологи «один к одному», то есть гены, прослеживающиеся до одного общего гена у последнего общего предкового вида и не претерпевшие генных изменений, описанных выше.

Для переноса предсказанных взаимодействий в STRING используется две стратегии. Первая - модель кластеров ортологичных групп (COG) (использовалась до версии 9.1) - перенос по принципу «всё-или-ничего» на

основе заранее определенных ортологических отношений, описанных в высокоуровневых таксономических группах из базы данных COG. Начиная с версии 9.1 используется перенос «интерологов» на основе рассчитанных ортологичных взаимоотношений из базы данных eggNOG (Franceschini et al. 2012). Таксономическое дерево STRING охватывает 1133 вида и состоит из 495 ветвящихся узлов. Для переноса взаимодействий разработан следующий подход: белковые ассоциации переносятся ортологичным группам, к которым эти белки относятся (последовательно от низкого к более высокому уровню иерархии), далее ассоциации передаются наоборот, от ортологичных групп к белкам и осуществляется перенос взаимодействий «все-против-всех», сравниваемых по отдельности для каждого источника информации. Перенос между близкородственными видами осуществляется с большей оценкой, в то время как наличие паралогов (подразумевающих дупликации генов) снижает оценку. Наибольший вклад перенесенные взаимодействий вносят в интерактом наименее изученных видов.

Как референсные взаимодействия в базе данных STRING были выбраны данные из базы данных KEGG: если два взаимодействующих белка находятся в одном пути, то взаимодействие считается истинно положительным. Таким образом, итоговая вероятностная оценка STRING свидетельствует о нахождение двух белков в одном пути KEGG (von Mering et al. 2003).

После анализа рассмотренных выше источников информации о взаимодействии, вычисляется итоговая вероятностная оценка, состоящая из объединенных вероятностей из различных источников и скорректированная с учетом вероятности случайного наблюдения взаимодействия. Итоговая оценка вычисляется с предположением о независимости источников информации, такое возможно, потому что наборы данных пришедшие в результате использования одинаковых технологий объединены ранее и далее рассматриваются как один источник информации. Вычисление итоговой оценки состоит из трех шагов:

1. Вычитание из оценки каждого из источников информации (S_i) априорной вероятности взаимодействия двух случайно выбранных белков, параметр постоянный (p=0,041).

$$S_{i-nop} = (S_i - p)/(1 - p) \quad (1)$$

2. Объединение оценок из всех каналов

$$S_{top-nop} = 1 - \prod_{i-nop}(1 - S_{i-nop}) \quad (2)$$

3. Добавление априорной вероятности

$$S_{top} = S_{tot-nop} + p(1 - S_{tot-nop}) \quad (3)$$

Итоговая вероятностная оценка ранжируется следующим образом: 0,9-1 очень высокая вероятность взаимодействия, 0,7-0,8 - высокая вероятность, средняя 0,6-04 вероятность, и низкая 0,1 вероятность.

2.6.2. IntAct

IntAct (http://www.ebi.ac.uk/intact/) (Orchard et al. 2014) - открытая база данных курируемых молекулярных взаимодействий из литературных источников и загруженных пользователями самостоятельно. Основной идентификатор - Uniprot ID, и в качестве альтернативных приводится ряд различных идентификаторов, среди которых и HGNC Symbol. Для каждого взаимодействия предоставляется следующая дополнительная информация: экспериментальный метод, тип взаимодействия, ссылка на статью, БД источник взаимодействия, доверительная оценка, биологическая роль взаимодействующих белков (для подавляющего большинства взаимодействий гистонов не определена), экспериментальная роль взаимодействующих белков (нейтральный компонент, ловушка, жертва, акцептор/донор флуоресценции), система экспрессии и метод идентификации взаимодействующих белков, вид живых организмов.

Доверительная оценка для двух взаимодействующих белков рассчитывается из следующих нормализованных компонент: экспериментальный метод, тип взаимодействия, количество публикаций

(максимум 8). Итоговая оценка находится в диапазоне 0 - 1, где 1 - наивысшая достоверность взаимодействия.

2.6.3. BioGrid

The Biological General Repository for Interaction Datasets (BioGRID) (http://thebiogrid.org/) (Stark et al. 2006). БД белковых, генных и химических взаимодействий, изначально создавалась для хранения и обработки результатов высокопроизводительных экспериментов и данных литературы. Идентификатором белков является в том числе и HGNC Symbol. Каждая запись сопровождается следующими данными: тип взаимодействия, экспериментальный метод, производительность экспериментального метода, ссылка на статью, вид живых организмов, оценка (используется, если указана, статистическая оценка из оригинальной публикации). К уникальной особенности BioGrid можно отнести сведения о пост-трансляционных модификациях.

2.7. Интерактом

Интерактом представляет собой совокупность взаимодействующих молекул, например, белок - белковых, ген - белковых или белков с малыми молекулами. Целью интерактомики является изучение не только взаимодействий, но и их влияния на судьбу клетки/ткани/и т.д. в норме и при патологии. С помощью интерактома можно получать знания о функционировании молекулярных комплексов, идентифицировать функции неизвестных белков, исследовать нарушения во взаимодействии белков и разрабатывать соответствующие таргетные препараты, анализировать доменную организацию взаимодействующих белков и исследовать взаимодействия «патоген-хозяин».

Отдельное направление изучения интерактомов - это эволюционные изменения. Было отмечено (Zitnik et al. 2019), что устойчивость интерактома зависит от положения вида на филогенетическом дереве. В бактериях

интерактом более «подвижный», что связано с выживанием в переменчивых условиях среды. Было показано, что белковые семейства демонстрируют скоординированную перестройку взаимодействий с течением времени. Постепенные изменения белковых сетей в результате приводят к наиболее устойчивым интерактомам, способным компенисировать отдельные белковые нарушения. Проведенное изучение интерактомов 1844 видов (Zitnik et al. 2019) подтверждает гипотезу о стремлении белковых сетей противодействовать событиям, выражающимся в потере белков.

Сеть белковых взаимодействия обычно представляют в виде графовой структуры, где узлы - это белки, а соединяющие их ребра - свидетельства о взаимодействии.

Распространенным методом анализа графовых структур является анализ сетевых метрик. Такой анализ актуальнее проводить для графов с высокой плотностью (то есть с бОльшим количеством возможных связей), например, если бы интерактом отражал не только взаимодействия гистон - негистон, но и все возможные взаимодействия между гистоновыми партнерами и дополнительно, всех взаимодействующих партнеров негистонов. Очевидно, что при таком дополнении граф очень сильно разрастется, будут ли гистоновые белки хабами (узлами с наибольшим количеством связей) - открытый вопрос.

К наиболее частым метрикам интерактомных анализов относят: степень узла - количество узлов, взаимодействующих с рассматриваемым; степень центральности - нормализованная оценка степени узла; центральность по посредничеству - количество кратчайших путей, между всеми парами узлов, проходящих через данный узел; центральность по близости - насколько близко узел расположен к остальным узлам сети; центральность по собственному вектору - зависимость между центральностью узла и центральностями соседних узлов;

Кластеризация также является распространенным подходом исследования графов высокой плотности, где объекты в одном кластере наиболее похожи друг на друга, чем на представителей других классов. Кластеризация может быть

иерархической и неиерархической и включает огромное множество подходов; кластеризующий коэффициент - степень взаимодействия между собой ближайших соседей данного узла.

Для визуализации нтерактомов, графового анализа, кластеризации и т.д. удобно использовать сервис Cytoscape (Shannon et al. 2003).

2.8. Взаимодействия патоген-хозяин

Одно из интересных направлений исследования интерактома - исследование взаимодействий патоген-хозяин, то есть исследование пересекающихся белковых сетей нескольких организмов. В контексте исследования нуклеосомного интерактома данное направление чрезвычайно актуально, потому что известен ряд человеческих патогенов, чьи белки, непосредственно или косвенно, взаимодействуют как с гистонами, так и с белками нуклеосомного интерактома. Ниже в качестве обзора будут приведены некоторые примеры. Так, *Mycobacterium tuberculosis* (вызывает туберкулез) секретирует липобелок LpqH, связывающийся с хроматиновыми ремоделерами класса SWI/SNF и препятствующий его функционированию. *Listeria monocytogenes* (инфицирует центральную нервную систему и может вызвать менингит и энцефалит) регулирует конденсацию хроматина через белок LntA, который привлекает регулятор гетерохроматина BAHD1, образуя в результате гетерохроматин. Вирус иммунодефицита человека способен через белок vpr способствовать диссоциации от хроматина гистон-ацетилтрансферазного комплекса p300/HAT.

С другой стороны, патогены могут влиять и на ПТМ гистонов, как через взаимодействия с белками хозяина, так и непосредственно изменяя модификации. *Shigella flexneri* (вызывает диарею), регулирует фосфорилирование гистона H3 в положении S10 путем секретирования фофсфотреонин лиазы OspF, которая удаляет фосфатные группы с членов сигнального пути MAPK, что и предотвращает MAPK-зависимое фосфорилирование H3S10. *Chlamydia trachomatis* (в зависимости от серотипа

может вызывать трахому, конъюнктивит, урогенитальные инфекции, венерическую лимфогранулему), *Legionella pneumophila* (легионеллез) за счет своих метилтрансфераз наносят соответствующие ПТМ, причем *L. pneumophila* в определенное место: H3 K10. Уже упомянутая M. tuberculosis и *Streptococcus pyogenes* (возбудитель скарлатины и других инфекций) фосфорилируют гистоны (Silmon de Monerri and Kim 2014).

Примером так называемой молекулярной мимикрии может служить белок вируса гриппа А подтипа H3N2 - NS1, содержащий аминокислотную последовательность ARTK, напоминающую хвост гистона хозяина. За счет этого вирусный белок захватывает фактор элонгации hPAF1 транскрипции хозяина и избирательно подавляет выработку антивирусных белков (Marazzi et al. 2012).

3. Материалы и методы исследования

3.1. Список гистонов

Для создания обновленного списка всех известных генов гистонов человека (с разбивкой на гены и псевдогены) и соответствующих им белков, включая сплайс изоформы, проводился анализ данных из баз данных MS_HistoneDB (El Kennani et al. 2017), HGNC (Gene Family: Histones) и консультирование с группой консорциума HGNC, ведущей в настоящее время пересмотр номенклатуры генов гистонов человека. Была создана таблица, где для каждого гистонового гена имеется следующая информация: название по HGNC, NCBI gene ID, ENSG идентификатор по системе Ensembl, ENST идентификатор транскрипта, ENSP идентификатор белка, Uniprot ID, тип гистона, принадлежность к классу канонических гистонов, функциональность (белок-кодирующий или псевдоген), биотип гена по Ensembl, название белка по (Talbert et al. 2012) и краткое название белка для использования его в скриптах. Полученная таблица генов гистонов человека доступна в Приложении А, веб-страницы генов и белков доступны по ссылкам: список генов, список белков). Количество кодирующих генов гистонов человека идентифицированных на данный момент - 92 (включая H1), количество псевдогенов - 30. Конвертация между различными идентификаторами проводилась с помощью bioDBnet: db2db (Mudunuri et al. 2009).

Множественное выравнивание белковых последовательностей канонических гистонов было проведено с в пакете программ MEGA7 по алгоритму ClustalW (Kumar, Stecher, and Tamura 2016).

3.2. Обработка информации БД ББВ

Для обработки информации о ББВ из баз данных IntAct, BioGrid, STRING был разработан полуавтоматизированный программный код на языке Python, подгружающий информацию о взаимодействиях гистонов и других белков из баз

данных IntAct, BioGRID (версия 3.5.170), STRING (версия 10.5). Обработка загруженных данных заключается в следующем: фильтрация взаимодействий (оставляем только те, в которых один из взаимодействующих белков - гистон), приведение идентификаторов к виду имени гена по HGNC, дополнение информации о типе и «каноничности» (относится ли он к классу канонических гистонов или вариантным гистонам).

Для представления данных в итоговом интерактоме был выбран идентификатор имени гена по HGNC. В IntAct для каждого белка было представлено несколько идентификаторов, нужный выделялся с помощью regexp по кодовому слову: gene name. В BioGrid нужный идентификатор находился в колонке "Official Symbol Interactor A/B". Информация в STRING представлена в виде идентификаторов вида ‘9606.ENSP0..’, для приведения соответствия с именем HGNC на основе загруженного с сайта STRING файла соответствий между различными форматами, был выбран формат BioMart_HUGO, как дающий наибольшее соответствие. Далее был создан словарь соответствий между исходным форматом данных STRING и BioMart_HUGO, который для некоторых гистонов и их партнеров был дополнен в ручном режиме.

В STRING для каждого взаимодействия представлена его вероятностная оценка. Для выработки критериев отбора информации было проведено исследование количества взаимодействий в зависимости от вероятностной оценки в каждом из источников информации и в результате был выбран порог итоговой оценки = 0,7.

Оценка взаимодействия в других БД подразумевает следующее: в IntAct, по сути, количество свидетельств о данном взаимодействии с учетом «веса» источника, а в BioGrid - статистическая оценка, если указана в публикации.

3.3. Построение иерархической функциональной классификации

Классификация белков интерактома с помощью системы Gene Ontology (GO) (Gene Ontology Consortium 2015) в терминах словарей «участие в

биологическом процессе», «молекулярная функция» и «внутриклеточная локализация» показала неудовлетворительную глубину аннотирования и большое количество общих терминов, таких как, например, молекулярная функция «связывание с белками хроматина» или клеточный компартмент «ядро». Это послужило причиной разработки функциональной иерархической классификации белков, взаимодействующих с гистонами.

При составлении классификации и «референсного набора» взаимодействующих с гистонами белков были проанализированы следующие литературные источники: (Musselman et al. 2012; Xu et al. 2017) содержащие информацию о белках, считывающих пост-трансляционные модификации (ПТМ) гистонов; (Burgess and Zhang 2013) - информация о гистоновых шаперонах; (Mani et al. 2017; Zhang et al. 2016) - ремоделеры хроматина, (Khare et al. 2012) - белки, наносящие и стирающие ПТМ, (Han et al. 2018) - транскрипционные факторы, (Mayran and Drouin 2018) - пионерные транскрипционные факторы, (Cubeñas-Potts and Corces 2015) - архитектурные белки хроматина. Также использовались результаты функционального обогащения взаимодействующих с гистонами белков по GO. Высшие уровни иерархии составленной классификации представлены на рисунке 2.

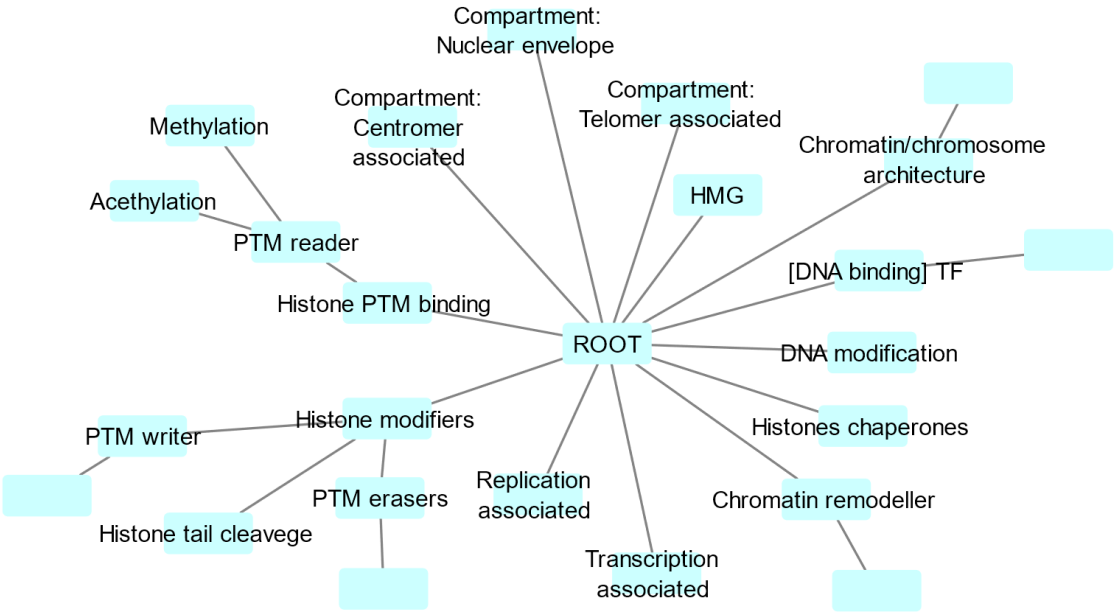

Рис. 2. Схема рациональной классификации для белков, с которыми взаимодействуют гистоны. Неподписанные прямоугольники означают следующие уровни классификации, не отображенные на рисунке.

Классификация белков проводилась следующим образом - пересечения генов с генами из литературных источников и специализированных баз данных, классификация с помощью системы Gene Ontology и выявление наиболее подходящих терминов для соотнесения с разработанной функциональной классификацией.

3.4. Построение интерактома

Для построения интерактома данные из БД STRING, BioGrid, IntAct были объединены. Для канонических гистонов одного типа была проведена процедура обобщения взаимодействующих партнеров с учетом мультигенности семейств канонических гистонов.

3.5. Анализ качества данных

Для оценки качества имеющейся информации было проведено выявление в итоговом интерактоме негистоновых белков хроматина, относящихся к определенным функциональным категориям, взятых из литературных источников, описанных выше. С другой стороны оценка качества имеющихся данных была проведена путем выявления белковых партнеров, точно не взаимодействующих с гистонами и не относящихся к ядерному компартменту клетки. Такие категории (например, миофибриллы, филоподии, реснички, микросомы, подосомы, фокальные контакты, белки клеточной адгезии и т.д.) были выявлены с помощью классификации Gene Ontology.

Для взаимодействующих с гистонами белков строились различные диаграммы, отражающие количество общих для гистонового типа партнеров, уникальных для гистона и взаимодействующих с несколькими гистонами типа. Также строились диаграммы пересечения гистонов, партнеров и парных

взаимодействий по трем БД. Исследовалось количественное распределение партнеров гистонов для каждого типа и класса в соответствии с функциональными категориями.

Поиск мотивов осуществлялся с помощью сервиса PROSITE (Sigrist et al. 2002), поиск доменов с помощью HMMER (Finn, Clements, and Eddy 2011).

4. Результаты и обсуждение

4.1 Анализ и классификация генов гистонов

Для выполнения работы требовалось составить список всех гистонов человека для дальнейшего поиска их в базах данных ББВ. Итоговый список включает 123 гена, состоящий из 93 белок-кодирующих генов и 30 псевдогенов. Так, 15 генов кодируют канонические гистоны H2A, 17 - канонические H2B, 14 - канонические H3, 15 генов - H4 и 5 генов H1. Количество белок-кодирующих гистонов вариантов H2A - 13, H2B - 4, H3 - 6, H1 - 6.

Множественное выравнивание последовательностей канонических белков человека каждого типа показало, что последовательности H2A абсолютно консервативны на 91.5%; H2B - на 91,3% (исключая короткий транскрипт HIST1H2BJ, и по одному транскрипту с дополнительными аминокислотами на C-конце HIST1H2BM и HIST2H2BF); H3 на 99,3%; и H4 без учета HIST1H4G на 100%, который отличается на 14,5% и имеет укороченный на 4 аминокислотных остатка C-конец. Канонические гистоны каждого типа объединяют в мультигенные семейства, то есть в группы структурно-родственных белков, возникших в эволюции в результате ряда последовательных дупликаций гена-предшественника.

Канонические линкерные гистоны H1 оказались более вариабельными, абсолютная консервативность белковых последовательность составила 40,4%. Коровые гистоны нуклеосомы и линкерный H1 различаются по расположению на нуклеосоме, по происхождению (коровые гистоны возникли у архейного предка, в то время как линкерный у эубактерий (Kasinsky et al. 2001), по степени консервативности канонических гистонов и доменной организацией. Далее под нуклеосомным интерактом будут пониматься именно взаимодействия между нуклеосомным кором, то есть гистонами H2A, H2B, H3, H4 и негистоновыми белками.

4.2 Анализ содержания баз данных (STRING, BioGrid, IntAct)

Для выполнения работы были выбраны три базы данных: BioGrid и IntAct, содержащие экспериментальные данные и STRING, использующая предсказательные алгоритмы. Последний сравнительный обзор наполнения БД ББВ (Bajpai et al., 2019) подтверждает, что на данный момент они лидируют и по покрытию всех взаимодействий, и по количеству уникальных.

Для STRING был проведен анализ количества взаимодействий в зависимости от вероятностной оценки для каждого из источников информации (канала). Оказалось, что для нуклеосомного интерактома каналы поиска совместно расположенных генов, фьюжн и базы данных для других организмов не дают никакой информации. Каналы базы данных, «перенесенные» совместно расположенные гены и данные сравнительного геномного филогенетического анализа содержат взаимодействия только при вероятностных оценках 0,3; 0,9 и 0,5, соответственно. Подавляющее количество экспериментальных данных оценивается STRING 0,1-0,4, то есть как низковероятные. Для итоговой оценки характерно два пика по количеству взаимодействий - в диапазоне вероятности 0,15-0,2, что связано с данными интеллектуального анализа текста, имеющих схожий пик, и при вероятности 0,9, что связано с данными канала базы данных. Для дальнейшего анализа был выбран порог итоговой вероятностной оценки 0,7.

Для всех белок-кодирующих гистонов человека из обновленного списка (см. Приложение А) были обнаружены взаимодействия (кроме вариантного гистона H1 HILS, для которого однозначно не установлена функциональность), в то время как в BioGrid были описаны взаимодействия и для 4 псевдогенов. В IntAct содержатся взаимодействия только для половины белок-кодирующих гистонов. Количество негистоновых партнеров в BioGrid в 2 раза больше, чем в STRING и в 1,7 раз больше, чем в IntAct. Однако несмотря на бОльшее число представленных гистонов и партнеров в базе данных BioGrid, общее количество пар взаимодействий гистон - негистон в три раза больше в STRING, по сравнению с BioGrid, и в 9 раз, по сравнению с IntAct.

Построенные диаграммы пересечения гистонов, уникальных взаимодействующих партнеров и парных взаимодействий по данным STRING, BioGrid, IntAct представлены на рис. 3.

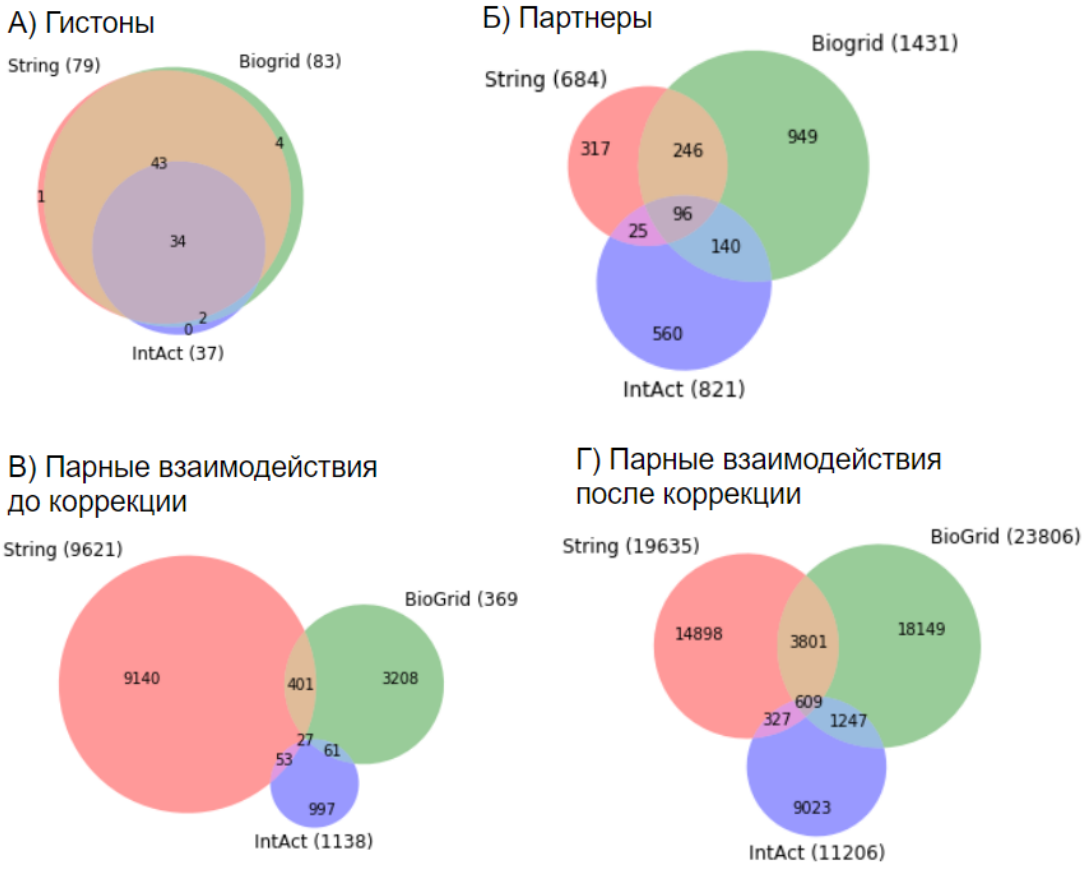

Рис. 3. Диаграммы Венна для данных STRING, BioGrid, IntAct. Показаны пересечения для А) коровых гистонов, Б) взаимодействующих с ними партнеров, В) парных взаимодействий гистон-негистон до учета мультигенности семейств канонических гистонов (подробнее в тексте), Г) парных взаимодействий гистон-негистон после учета.

4.3. Анализ интерактома

Всего в интерактоме содержится 15402 экспериментальных b предсказательных свидетельств о 13887 уникальных взаимодействиях, участие в которых принимают 85 коровых гистонов и 2333 негистоновых белка. Подробная количественная статистика представлена в Приложении Б.

Общее количество взаимодействий гистон-негистон в объединенном интерактоме составило 13887, причем 96 % из них уникальны для одной из БД, и только 27 взаимодействий найдены во всех трех БД, что составляет 0,19 % от всех взаимодействий.

Количество взаимодействий для каждого из гистонов, сгруппированных по типу и классу каноничности показано на рис. 4.

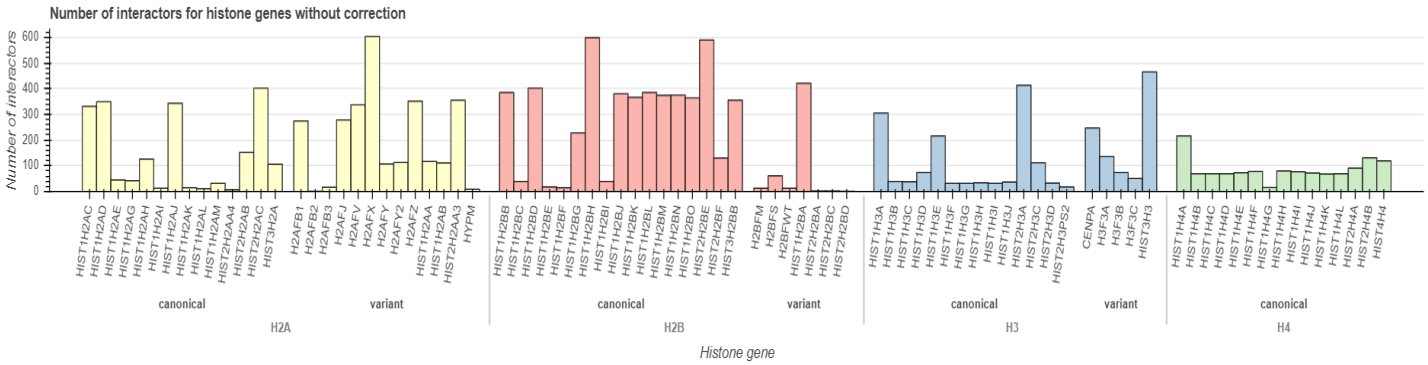

Рис. 4. Количество взаимодействий для каждого из гистонов, сгруппированных по типу и классу каноничности по материалам из STRING, IntAct, BioGrid.

При исследовании количества уникальных взаимодействий из разных БД для канонических гистонов одного типа (рис. 5), были отмечены резкие различия, подтверждающие предположение о затрудненной идентификации представителей мультигенных семейств канонических гистонов в высокопроизводительных экспериментах. Так, на рис. 5 В) можно заметить около 300 взаимодействий для HIST1H3A и около 200 для HIST1H3E из BioGrid, в то время как из STRING приходит сравнительно меньшее количество взаимодействий (меньше 50-ти). Однако для HIST2H3A преобладающее количество взаимодействий приходит, наоборот, из STRING (около 350).

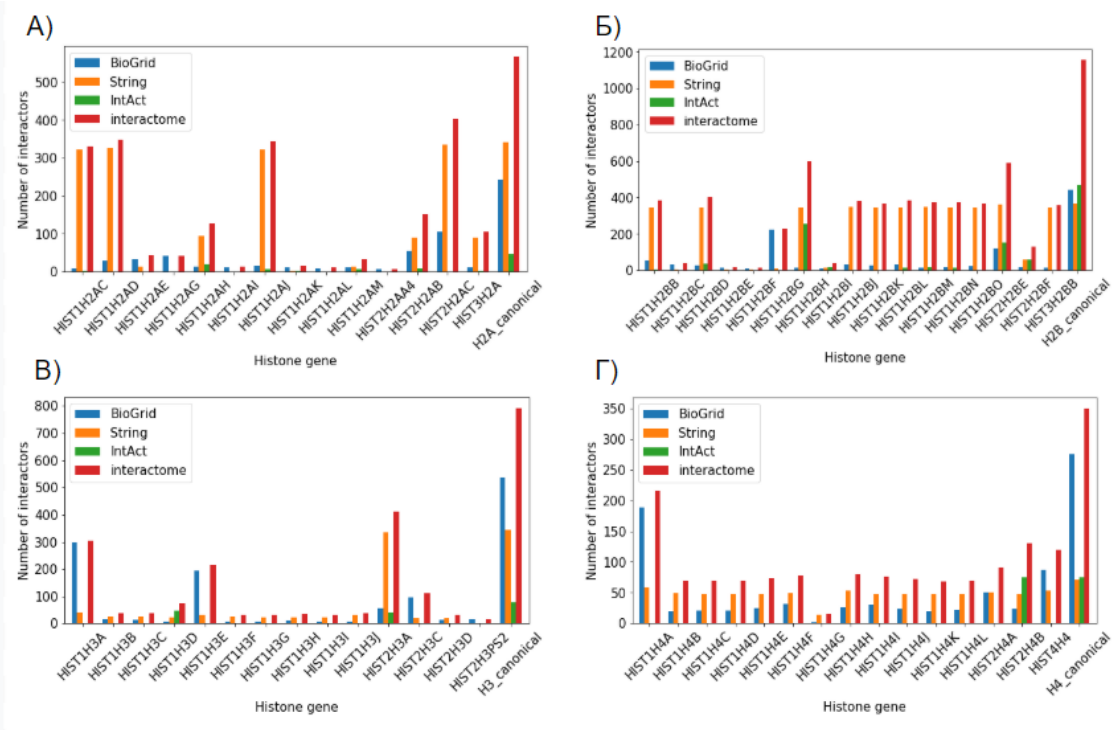

Рис. 5. Количество взаимодействий для канонических гистонов из трех БД и объединенного интерактома. HX_canonical - суммарное количество уникальных взаимодействий для канонических гистонов данного типа. А) канонические гены гистонов H2A, Б) - H2B, В) - H3, Г) - H4.

Для корректировки возможных неточностей при идентификации канонических гистонов одного типа, совершаемых в высокопроизводительных экспериментах, была разработана процедура учета мультигенности семейства канонических гистонов и последующее обобщение всех найденных взаимодействующих партнеров на все канонические гистоны данного типа. Таким образом, итоговое количество найденных взаимодействующих партнеров составило 566 для канонических H2A, 1154 для H2B, 793 для H3 и 350 для H4.

Интересно, что если сравнить парные взаимодействия после учета мультигенности канонических гистонов (рис. 3, Г), то в объединенном интерактоме их число составит 609 из 48054 (1,27 %). Получается, что все три базы данных действительно дополняют информацию о нуклеосомном интерактоме, а разные наборы взаимодействующих партнеров, как и парных взаимодействий, могут дополнительно поднять вопрос об отмеченных ранее

низкой воспроизводимости экспериментальных методов и ложных (как ложноположительных, так и ложноотрицательных) результатах.

4.3.1 Классификация гистоновых белков-партнеров

Взаимодействующие с нуклеосомой белки были рассмотрены с точки зрения взаимодействий с гистонами разных типов. Результаты показаны на рис. 6, так, 6 % белков взаимодействуют со всеми типами канонических гистонов и 16 % - с гистоновыми вариантами. Наибольшее количество уникальных партнеров среди канонических гистонов у H2A (50 %), среди вариантных - у H3 (54,7 %).

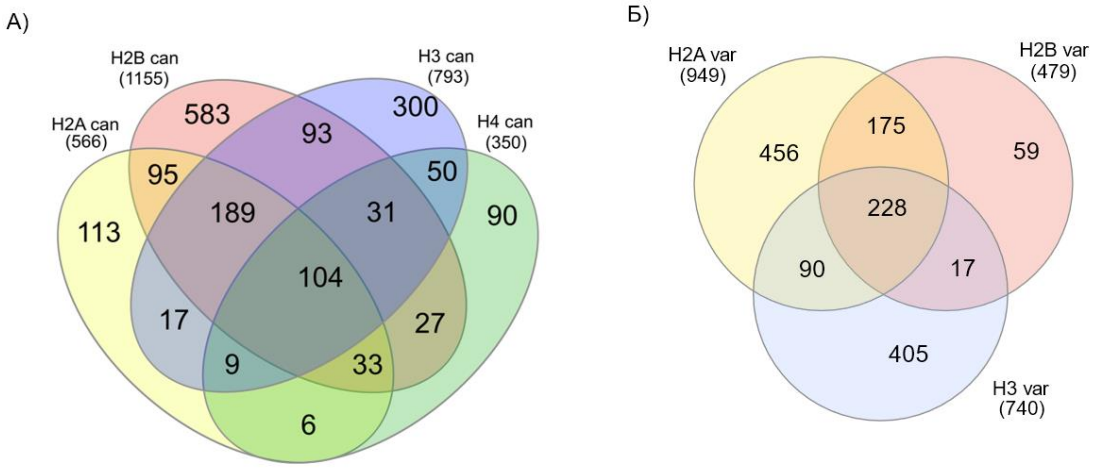

Рис. 6. Пересечения взаимодействующих с нуклеосомой белков, А) с каноническими гистонами, Б) с гистоновыми вариантами.

Взаимодействующие с нуклеосомой белки были классифицированы в соответствии с разработанной иерархической классификацией (см. рис.2), для 33% неклассифицированных генов требуется более тщательная ручная классификация или добавление новых категорий.

В таблице 1 представлено количество генов определенных функциональных категорий по литературным данным (подробнее описано в п. 2.4. Негистоновые белки интерактома) и количество этих же генов в интерактоме. Так, количество генов основных функциональных классов из литературных данных, представленных в интерактоме, находится в диапазоне

72-94 %. Низкую представленность транскрипционных факторов из базы данных TRUST (37%) и архитектурных белков (19%) в интерактоме можно объяснить пространственной и временной специфичностью данных взаимодействий.

Таблица 1. Количество пересечений генов интерактома с генами определенных функциональных категорий, взятых из литературных данных.

Функциональная категория	**Количество генов по данным литературных источников**	**Количество пересечений с интерактомом**	**%**
Гистоновые шапероны	36	34	94
Хроматиновые ремоделеры	82	69	84
Белки, удаляющие ПТМ	63	50	79
Белки, считывающие ПТМ	142	111	78
Белки, наносящие ПТМ	118	85	72
Транскрипционные факторы	795	292	37
Архитектурные белки хроматина	36	7	19

После функционального обогащения Gene Ontology были выявлены категории, не имеющие отношения к ядерному компартменту в целом и к взаимодействиям с гистонами в частности. К ним относится 324 белков (13,9 %), информация о большинстве из которых приходит из базы данных BioGrid.

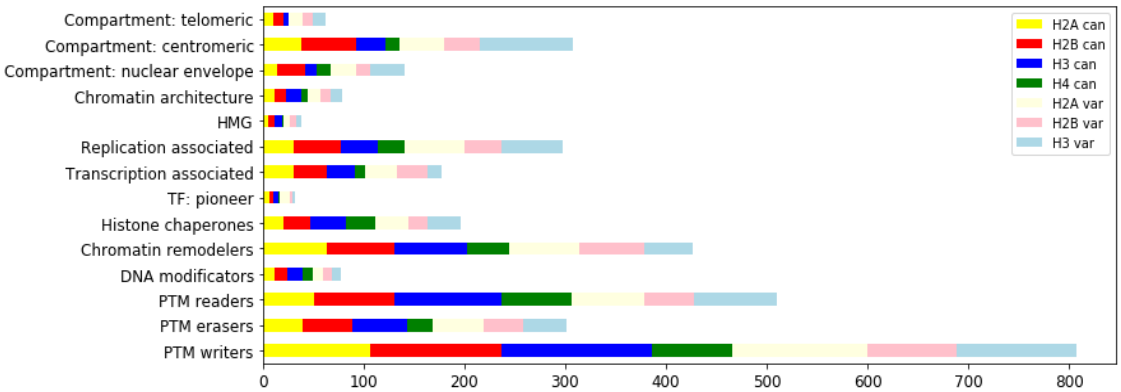

Рис. 7. Количество белков определенной функциональной категории, взаимодействующих с каждым из наборов гистонов (одного типа и класса каноничности: гистоновые варианты в светлых тонах). Категория

транскрипционных факторов не показана для увеличения масштаба остальных категорий.

Анализ интерактома с использованием разработанной классификации изображен на рис. 7, где показано количество партнеров определенных функциональных групп, взаимодействующих с каждым из типов гистонов одного класса. Наибольшее количество классифицированных белковых партнеров относятся к классам транскрипционных факторов, белков, взаимодействующих с пост-трансляционными модификациями гистонов, и ремоделерам хроматина.

4.4. Поиск мотивов и доменов

В последовательностях белков интерактома проводился поиск мотивов: CENP-C, AT-hook и SPKK. Общее количество белков с мотивом CENP-C - 8, AT-hook - 108, SPKK - 176. Интересно нахождение мотива CENP-C в FUT8 (фукозилтрансфераза). Процентное количество белков, содержащих рассматриваемые мотивы, для каждой из функциональных категорий изображено на рис. 8.

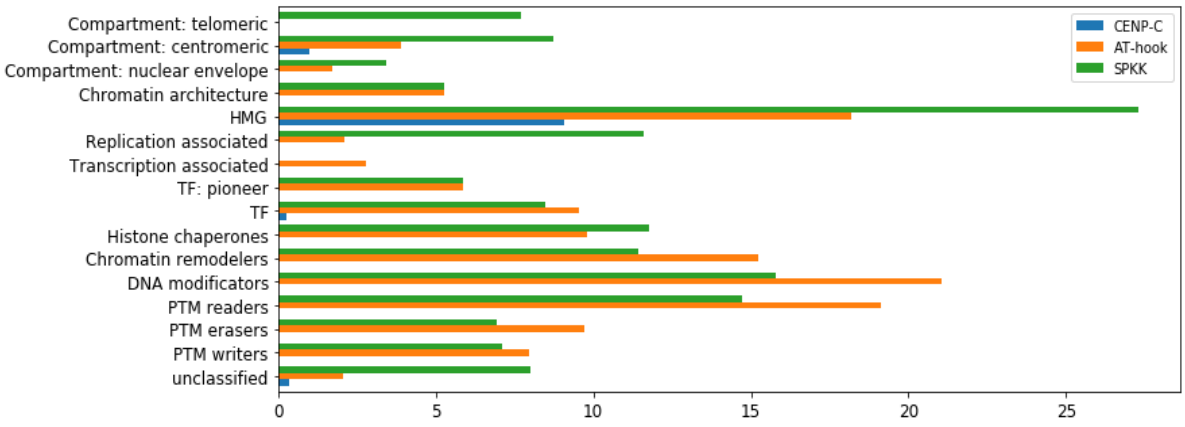

Рис. 8. Процентное содержание белков нуклеосомного интерактома, содержащих мотивы CENP-C, AT-hook и SPKK, к общему количеству белков, относящихся к данным функциональным категориям, unclassified - неклассифицированные белки интерактома.

Также последовательности белков интерактома, относящихся к категории хроматиновых ремоделеров, были проанализированы на предмет доменной организации. Было найдено 29 уникальных доменных клана по систематике Pfam и 51 доменное семейство, не объединенное в кланы. В таблице 2 показаны частоты встречаемости представителей доменных кланов среди ремоделеров хроматина, одинаковые домены в составе одного белка учитывались 1 раз.

Таблица 2. Кланы по Pfam для хроматиновых ремоделеров.

Клан по PFAM	**Расшифровка клана**	**частота встречаемости доменов клана*, %**
CL0023	P-loop containing nucleoside triphosphate hydrolase superfamily	16,7
CL0049	Tudor domain 'Royal family'	5
CL0123	Helix-turn-helix clan	5
CL0390	FYVE/PHD zinc finger superfamily	4
-	Bromodomain**	3,5
CL0167	Zinc beta-ribbon	3

Примечание * среди найденных кланов и доменных семейств, не объединенных в доменные кланы; ** доменное семейство

В клан CL0023 (Neuwald et al. 1999) входит 229 доменных семейств, белки данного класса - шапероноподобные АТФазы, участвующие в сборке/разборке белковых комплексов. Клан CL0049 (Maurer-Stroh et al. 2003) состоит из 33 семейств и включает в себя такие домены, как, например, chromo, MBT, PWWP и tudor. Белки данного клана узнают метилированные лизины на гистонах и других белках, а также метилированный аргинин. К клану CL0123 (Brennan and Matthews 1989) относится 340 семейств, содержащих ДНК-связывающий мотив спираль-поворот-спираль.

5. Заключение

В рамках данной работы был составлен обновленный список гистонов человека, включающий белок-кодирующие гены и псевдогены, а также белковые сплайс-изоформы. На основе этого списка были проанализированы взаимодействия гистонов из БД STRING, BioGrid, IntAct и построен нуклеосомный интерактом с использованием разработанного программного кода. Для преодоления экспериментальных трудностей идентификации консервативных канонических генов предложен метод учета мультигенности семейств канонических гистонов и обобщение взаимодействий внутри одного типа канонических гистонов. Для белков, взаимодействующих с нуклеосомой, составлена функциональная иерархическая классификация. Проведен качественный и количественный анализ белков-партнеров, а также выполнен поиск специфических мотивов в белках интерактома и исследована доменная организация хроматиновых ремоделеров.

6. Выводы

1. Разработан полуавтоматизированный программный код на языке Python, позволяющий загружать и анализировать данные ББВ из открытых БД.
2. Построен нуклесомный интерактом по материалам баз данных STRING, IntAct и BioGrid, включающий в себя 85 коровых гистона, 2333 белковых партнера, 15402 экспериментальных/предсказательных свидетельства о 13887 уникальных взаимодействия.
3. Показано, что 96 % взаимодействий уникально для одной из БД, и только 0,19 % встречаются в трех БД. После проведения разработанного учета мультигенности семейств канонических гистонов, количество общих для трех БД взаимодействий составило 1,27 %.
4. Разработана иерархическая функциональная классификация белков, взаимодействующих с нуклеосомой, для 67 % белков интерактома определены функциональные категории. В интерактоме содержится 72-94% белков основных функциональных категорий, описанных в проанализированных литературных источниках.
5. Проведен поиск мотивов CENP-C, AT-hook, SPKK в последовательностях белков интерактома. Количество белков с мотивом CENP-C - 8, AT-hook - 108, SPKK - 176.
6. Проведен анализ доменной организации белков, относящихся к хроматиновым ремоделерам. По частоте встречаемости лидируют доменные кланы шапероноподобных АТФаз, Тюдор домены, спираль-поворот-спираль.

Список публикаций по результатам работы:

1. Armeev, Grigoriy A, Anna K Gribkova, Iunona Pospelova, Galina A Komarova, and Alexey K Shaytan. 2019. ‘Linking Chromatin Composition and Structural Dynamics at the Nucleosome Level’. *Current Opinion in Structural Biology* 56 (June): 46–55. https://doi.org/10.1016/j.sbi.2018.11.006.
2. [accepted abstract] A.K. Gribkova, A.K. Shaytan. 2019. 'Construction and analysis of an interactome between nucleosomes and chromatin proteins'.

Supplement: 44rd FEBS Congress, Biochemistry Forever, Poland, Krakow, July 7-12, 2019.

7. Список литературы

1. Aravind, L., and D. Landsman. 1998. ‘AT-Hook Motifs Identified in a Wide Variety of DNA-Binding Proteins’. *Nucleic Acids Research* 26 (19): 4413–21. https://doi.org/10.1093/nar/26.19.4413.
2. Armeev, Grigoriy A, Anna K Gribkova, Iunona Pospelova, Galina A Komarova, and Alexey K Shaytan. 2019. ‘Linking Chromatin Composition and Structural Dynamics at the Nucleosome Level’. *Current Opinion in Structural Biology* 56 (June): 46–55. https://doi.org/10.1016/j.sbi.2018.11.006.
3. Bajpai, Akhilesh Kumar, Sravanthi Davuluri, Kriti Tiwary, Sithalechumi Narayanan, Sailaja Oguru, Kavyashree Basavaraju, Deena Dayalan, Kavitha Thirumurugan, and Kshitish K Acharya. 2019. ‘How Helpful Are the Protein-Protein Interaction Databases and Which Ones?’ *BioRxiv*, March. https://doi.org/10.1101/566372.
4. Barrett, Tanya, Stephen E. Wilhite, Pierre Ledoux, Carlos Evangelista, Irene F. Kim, Maxim Tomashevsky, Kimberly A. Marshall, et al. 2013. ‘NCBI GEO: Archive for Functional Genomics Data Sets--Update’. *Nucleic Acids Research* 41 (Database issue): D991-995. https://doi.org/10.1093/nar/gks1193.
5. Brennan, R. G., and B. W. Matthews. 1989. ‘The Helix-Turn-Helix DNA Binding Motif’. *The Journal of Biological Chemistry* 264 (4): 1903–6.
6. Burgess, Rebecca J., and Zhiguo Zhang. 2013. ‘Histone Chaperones in Nucleosome Assembly and Human Disease’. *Nature Structural & Molecular Biology* 20 (1): 14–22. https://doi.org/10.1038/nsmb.2461.
7. Churchill, M. E., and M. Suzuki. 1989. ‘“SPKK” Motifs Prefer to Bind to DNA at A/T-Rich Sites’. *The EMBO Journal* 8 (13): 4189–95.
8. Cubeñas-Potts, Caelin, and Victor G. Corces. 2015. ‘Architectural Proteins, Transcription, and the Three-Dimensional Organization of the Genome’. *FEBS Letters* 589 (20 Pt A): 2923–30. https://doi.org/10.1016/j.febslet.2015.05.025.
9. Davey, Curt A., David F. Sargent, Karolin Luger, Armin W. Maeder, and Timothy J. Richmond. 2002. ‘Solvent Mediated Interactions in the Structure of the Nucleosome Core Particle at 1.9 a Resolution’. *Journal of Molecular Biology* 319 (5): 1097–1113. https://doi.org/10.1016/S0022-2836(02)00386-8.
10. Deane, Charlotte M., Łukasz Salwiński, Ioannis Xenarios, and David Eisenberg. 2002. ‘Protein Interactions: Two Methods for Assessment of the Reliability of High Throughput Observations’. *Molecular & Cellular Proteomics: MCP* 1 (5): 349–56. https://doi.org/10.1074/mcp.m100037-mcp200.
11. Dekker, Job, Karsten Rippe, Martijn Dekker, and Nancy Kleckner. 2002. ‘Capturing Chromosome Conformation’. *Science (New York, N.Y.)* 295 (5558): 1306–11. https://doi.org/10.1126/science.1067799.
12. El Kennani, Sara, Annie Adrait, Alexey K. Shaytan, Saadi Khochbin, Christophe Bruley, Anna R. Panchenko, David Landsman, Delphine Pflieger, and Jérôme Govin. 2017. ‘MS_HistoneDB, a Manually Curated Resource for Proteomic Analysis of Human and Mouse Histones’. *Epigenetics & Chromatin* 10 (1). https://doi.org/10.1186/s13072-016-0109-x.
13. Elefsinioti, Antigoni, Ömer Sinan Saraç, Anna Hegele, Conrad Plake, Nina C. Hubner, Ina Poser, Mihail Sarov, et al. 2011. ‘Large-Scale de Novo Prediction of Physical Protein-Protein Association’. *Molecular & Cellular Proteomics: MCP* 10 (11): M111.010629. https://doi.org/10.1074/mcp.M111.010629.
14. El-Gebali, Sara, Jaina Mistry, Alex Bateman, Sean R. Eddy, Aurélien Luciani, Simon C. Potter, Matloob Qureshi, et al. 2019. ‘The Pfam Protein Families Database in 2019’. *Nucleic Acids Research* 47 (D1): D427–32. https://doi.org/10.1093/nar/gky995.
15. Fabregat, Antonio, Steven Jupe, Lisa Matthews, Konstantinos Sidiropoulos, Marc Gillespie, Phani Garapati, Robin Haw, et al. 2018. ‘The Reactome Pathway Knowledgebase’. *Nucleic*

Acids Research 46 (D1): D649–55. https://doi.org/10.1093/nar/gkx1132.
16. Finn, Robert D., Jody Clements, and Sean R. Eddy. 2011. 'HMMER Web Server: Interactive Sequence Similarity Searching'. *Nucleic Acids Research* 39 (suppl_2): W29–37. https://doi.org/10.1093/nar/gkr367.
17. Flanagan, John F., Li-Zhi Mi, Maksymilian Chruszcz, Marcin Cymborowski, Katrina L. Clines, Youngchang Kim, Wladek Minor, Fraydoon Rastinejad, and Sepideh Khorasanizadeh. 2005. 'Double Chromodomains Cooperate to Recognize the Methylated Histone H3 Tail'. *Nature* 438 (7071): 1181–85. https://doi.org/10.1038/nature04290.
18. Franceschini, Andrea, Jianyi Lin, Christian von Mering, and Lars Juhl Jensen. 2016. 'SVD-Phy: Improved Prediction of Protein Functional Associations through Singular Value Decomposition of Phylogenetic Profiles'. *Bioinformatics* 32 (7): 1085–87. https://doi.org/10.1093/bioinformatics/btv696.
19. Franceschini, Andrea, Damian Szklarczyk, Sune Frankild, Michael Kuhn, Milan Simonovic, Alexander Roth, Jianyi Lin, et al. 2012. 'STRING v9.1: Protein-Protein Interaction Networks, with Increased Coverage and Integration'. *Nucleic Acids Research* 41 (D1): D808–15. https://doi.org/10.1093/nar/gks1094.
20. Han, Heonjong, Jae-Won Cho, Sangyoung Lee, Ayoung Yun, Hyojin Kim, Dasom Bae, Sunmo Yang, et al. 2018. 'TRRUST v2: An Expanded Reference Database of Human and Mouse Transcriptional Regulatory Interactions'. *Nucleic Acids Research* 46 (D1): D380–86. https://doi.org/10.1093/nar/gkx1013.
21. Hetzer, Martin W. 2010. 'The Nuclear Envelope'. *Cold Spring Harbor Perspectives in Biology* 2 (3). https://doi.org/10.1101/cshperspect.a000539.
22. Horton, John R., Anup K. Upadhyay, Hank H. Qi, Xing Zhang, Yang Shi, and Xiaodong Cheng. 2010. 'Enzymatic and Structural Insights for Substrate Specificity of a Family of Jumonji Histone Lysine Demethylases'. *Nature Structural & Molecular Biology* 17 (1): 38–43. https://doi.org/10.1038/nsmb.1753.
23. Jacobson, R. H., A. G. Ladurner, D. S. King, and R. Tjian. 2000. 'Structure and Function of a Human TAFII250 Double Bromodomain Module'. *Science (New York, N.Y.)* 288 (5470): 1422–25.
24. Kanehisa, M., and S. Goto. 2000. 'KEGG: Kyoto Encyclopedia of Genes and Genomes'. *Nucleic Acids Research* 28 (1): 27–30. https://doi.org/10.1093/nar/28.1.27.
25. Kasinsky, H. E., J. D. Lewis, J. B. Dacks, and J. Ausió. 2001. 'Origin of H1 Linker Histones'. *FASEB Journal: Official Publication of the Federation of American Societies for Experimental Biology* 15 (1): 34–42. https://doi.org/10.1096/fj.00-0237rev.
26. Kato, Hidenori, Jiansheng Jiang, Bing-Rui Zhou, Marieke Rozendaal, Hanqiao Feng, Rodolfo Ghirlando, T. Sam Xiao, Aaron F. Straight, and Yawen Bai. 2013. 'A Conserved Mechanism for Centromeric Nucleosome Recognition by Centromere Protein CENP-C'. *Science (New York, N.Y.)* 340 (6136): 1110–13. https://doi.org/10.1126/science.1235532.
27. Keshava Prasad, T. S., Renu Goel, Kumaran Kandasamy, Shivakumar Keerthikumar, Sameer Kumar, Suresh Mathivanan, Deepthi Telikicherla, et al. 2009. 'Human Protein Reference Database—2009 Update'. *Nucleic Acids Research* 37 (Database issue): D767–72. https://doi.org/10.1093/nar/gkn892.
28. Khare, Satyajeet P., Farhat Habib, Rahul Sharma, Nikhil Gadewal, Sanjay Gupta, and Sanjeev Galande. 2012. 'HIstome—a Relational Knowledgebase of Human Histone Proteins and Histone Modifying Enzymes'. *Nucleic Acids Research* 40 (Database issue): D337–42. https://doi.org/10.1093/nar/gkr1125.
29. Kumar, Sudhir, Glen Stecher, and Koichiro Tamura. 2016. 'MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets'. *Molecular Biology and Evolution* 33 (7): 1870–74. https://doi.org/10.1093/molbev/msw054.
30. Lambert, Jean-Philippe, Monika Tucholska, Christopher Go, James D.R. Knight, and Anne-Claude Gingras. 2015. 'Proximity Biotinylation and Affinity Purification Are Complementary Approaches for the Interactome Mapping of Chromatin-Associated Protein Complexes'.

Journal of Proteomics 118 (April): 81–94. https://doi.org/10.1016/j.jprot.2014.09.011.
31. Lehne, Benjamin, and Thomas Schlitt. 2009. 'Protein-Protein Interaction Databases: Keeping up with Growing Interactomes'. *Human Genomics* 3 (3): 291–97. https://doi.org/10.1186/1479-7364-3-3-291.
32. Lieberman-Aiden, Erez, Nynke L. van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, et al. 2009. 'Comprehensive Mapping of Long Range Interactions Reveals Folding Principles of the Human Genome'. *Science (New York, N.Y.)* 326 (5950): 289–93. https://doi.org/10.1126/science.1181369.
33. Mani, Udayakumar, Alagu Sankareswaran S, Arun Goutham R N, and Suma Mohan S. 2017. 'SWI/SNF Infobase-An Exclusive Information Portal for SWI/SNF Remodeling Complex Subunits'. *PloS One* 12 (9): e0184445. https://doi.org/10.1371/journal.pone.0184445.
34. Marazzi, Ivan, Jessica S. Y. Ho, Jaehoon Kim, Balaji Manicassamy, Scott Dewell, Randy A. Albrecht, Chris W. Seibert, et al. 2012. 'Suppression of the Antiviral Response by an Influenza Histone Mimic'. *Nature* 483 (7390): 428–33. https://doi.org/10.1038/nature10892.
35. Mathivanan, Suresh, Balamurugan Periaswamy, TKB Gandhi, Kumaran Kandasamy, Shubha Suresh, Riaz Mohmood, YL Ramachandra, and Akhilesh Pandey. 2006. 'An Evaluation of Human Protein-Protein Interaction Data in the Public Domain'. *BMC Bioinformatics* 7 (Suppl 5): S19. https://doi.org/10.1186/1471-2105-7-S5-S19.
36. Maurer-Stroh, Sebastian, Nicholas J. Dickens, Luke Hughes-Davies, Tony Kouzarides, Frank Eisenhaber, and Chris P. Ponting. 2003. 'The Tudor Domain "Royal Family": Tudor, Plant Agenet, Chromo, PWWP and MBT Domains'. *Trends in Biochemical Sciences* 28 (2): 69–74. https://doi.org/10.1016/S0968-0004(03)00004-5.
37. Mayran, Alexandre, and Jacques Drouin. 2018. 'Pioneer Transcription Factors Shape the Epigenetic Landscape'. *The Journal of Biological Chemistry* 293 (36): 13795–804. https://doi.org/10.1074/jbc.R117.001232.
38. Medvedeva, Yulia A., Andreas Lennartsson, Rezvan Ehsani, Ivan V. Kulakovskiy, Ilya E. Vorontsov, Pouda Panahandeh, Grigory Khimulya, Takeya Kasukawa, The FANTOM Consortium, and Finn Drabløs. 2015. 'EpiFactors: A Comprehensive Database of Human Epigenetic Factors and Complexes'. *Database* 2015: bav067. https://doi.org/10.1093/database/bav067.
39. Memišević, Vesna, Anders Wallqvist, and Jaques Reifman. 2013. 'Reconstituting Protein Interaction Networks Using Parameter-Dependent Domain-Domain Interactions'. *BMC Bioinformatics* 14 (May): 154. https://doi.org/10.1186/1471-2105-14-154.
40. Mering, Christian von, Martijn Huynen, Daniel Jaeggi, Steffen Schmidt, Peer Bork, and Berend Snel. 2003. 'STRING: A Database of Predicted Functional Associations between Proteins'. *Nucleic Acids Research* 31 (1): 258–61.
41. Mering, Christian von, Roland Krause, Berend Snel, Michael Cornell, Stephen G. Oliver, Stanley Fields, and Peer Bork. 2002. 'Comparative Assessment of Large-Scale Data Sets of Protein-Protein Interactions'. *Nature* 417 (6887): 399–403. https://doi.org/10.1038/nature750.
42. Morinière, Jeanne, Sophie Rousseaux, Ulrich Steuerwald, Montserrat Soler-López, Sandrine Curtet, Anne-Laure Vitte, Jérôme Govin, et al. 2009. 'Cooperative Binding of Two Acetylation Marks on a Histone Tail by a Single Bromodomain'. *Nature* 461 (7264): 664–68. https://doi.org/10.1038/nature08397.
43. Mudunuri, Uma, Anney Che, Ming Yi, and Robert M. Stephens. 2009. 'BioDBnet: The Biological Database Network'. *Bioinformatics* 25 (4): 555–56. https://doi.org/10.1093/bioinformatics/btn654.
44. Musselman, Catherine A, Marie-Eve Lalonde, Jacques Côté, and Tatiana G Kutateladze. 2012. 'Perceiving the Epigenetic Landscape through Histone Readers'. *Nature Structural & Molecular Biology* 19 (12): 1218–27. https://doi.org/10.1038/nsmb.2436.
45. Neuwald, A. F., L. Aravind, J. L. Spouge, and E. V. Koonin. 1999. 'AAA+: A Class of Chaperone-like ATPases Associated with the Assembly, Operation, and Disassembly of Protein Complexes'. *Genome Research* 9 (1): 27–43.

46. Orchard, S., M. Ammari, B. Aranda, L. Breuza, L. Briganti, F. Broackes-Carter, N. H. Campbell, et al. 2014. 'The MIntAct Project--IntAct as a Common Curation Platform for 11 Molecular Interaction Databases.' *Nucleic Acids Research* 42 (Database issue): D358-63. https://doi.org/10.1093/nar/gkt1115.
47. Ruthenburg, Alexander J., Haitao Li, Thomas A. Milne, Scott Dewell, Robert K. McGinty, Melanie Yuen, Beatrix Ueberheide, et al. 2011. 'Recognition of a Mononucleosomal Histone Modification Pattern by BPTF via Multivalent Interactions'. *Cell* 145 (5): 692–706. https://doi.org/10.1016/j.cell.2011.03.053.
48. Ruthenburg, Alexander J., Haitao Li, Dinshaw J. Patel, and C. David Allis. 2007. 'Multivalent Engagement of Chromatin Modifications by Linked Binding Modules'. *Nature Reviews. Molecular Cell Biology* 8 (12): 983–94. https://doi.org/10.1038/nrm2298.
49. Shannon, Paul, Andrew Markiel, Owen Ozier, Nitin S. Baliga, Jonathan T. Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. 2003. 'Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks'. *Genome Research* 13 (11): 2498–2504. https://doi.org/10.1101/gr.1239303.
50. Sigrist, Christian J. A., Lorenzo Cerutti, Nicolas Hulo, Alexandre Gattiker, Laurent Falquet, Marco Pagni, Amos Bairoch, and Philipp Bucher. 2002. 'PROSITE: A Documented Database Using Patterns and Profiles as Motif Descriptors'. *Briefings in Bioinformatics* 3 (3): 265–74.
51. Silmon de Monerri, Natalie C., and Kami Kim. 2014. 'Pathogens Hijack the Epigenome'. *The American Journal of Pathology* 184 (4): 897–911. https://doi.org/10.1016/j.ajpath.2013.12.022.
52. Stark, Chris, Bobby-Joe Breitkreutz, Teresa Reguly, Lorrie Boucher, Ashton Breitkreutz, and Mike Tyers. 2006. 'BioGRID: A General Repository for Interaction Datasets'. *Nucleic Acids Research* 34 (Database issue): D535–39. https://doi.org/10.1093/nar/gkj109.
53. Suthram, Silpa, Tomer Shlomi, Eytan Ruppin, Roded Sharan, and Trey Ideker. 2006. 'A Direct Comparison of Protein Interaction Confidence Assignment Schemes'. *BMC Bioinformatics* 7 (July): 360. https://doi.org/10.1186/1471-2105-7-360.
54. Szklarczyk, Damian, Andrea Franceschini, Stefan Wyder, Kristoffer Forslund, Davide Heller, Jaime Huerta-Cepas, Milan Simonovic, et al. 2015. 'STRING V10: Protein-Protein Interaction Networks, Integrated over the Tree of Life'. *Nucleic Acids Research* 43 (Database issue): D447-452. https://doi.org/10.1093/nar/gku1003.
55. Szklarczyk, Damian, John H. Morris, Helen Cook, Michael Kuhn, Stefan Wyder, Milan Simonovic, Alberto Santos, et al. 2017. 'The STRING Database in 2017: Quality-Controlled Protein-Protein Association Networks, Made Broadly Accessible'. *Nucleic Acids Research* 45 (D1): D362–68. https://doi.org/10.1093/nar/gkw937.
56. Talbert, Paul B., Kami Ahmad, Geneviève Almouzni, Juan Ausió, Frederic Berger, Prem L. Bhalla, William M. Bonner, et al. 2012. 'A Unified Phylogeny-Based Nomenclature for Histone Variants'. *Epigenetics & Chromatin* 5 (1): 7. https://doi.org/10.1186/1756-8935-5-7.
57. Venkatesh, Swaminathan, and Jerry L. Workman. 2015. 'Histone Exchange, Chromatin Structure and the Regulation of Transcription'. *Nature Reviews. Molecular Cell Biology* 16 (3): 178–89. https://doi.org/10.1038/nrm3941.
58. Wojcik, J., and V. Schächter. 2001. 'Protein-Protein Interaction Map Inference Using Interacting Domain Profile Pairs'. *Bioinformatics (Oxford, England)* 17 Suppl 1: S296-305.
59. Xi, Qiaoran, Zhanxin Wang, Alexia-Ileana Zaromytidou, Xiang H.-F. Zhang, Lai-Fong Chow-Tsang, Jing X. Liu, Hyesoo Kim, et al. 2011. 'A Poised Chromatin Platform for TGF-β Access to Master Regulators'. *Cell* 147 (7): 1511–24. https://doi.org/10.1016/j.cell.2011.11.032.
60. Xu, Yang, Shuang Zhang, Shaofeng Lin, Yaping Guo, Wankun Deng, Ying Zhang, and Yu Xue. 2017. 'WERAM: A Database of Writers, Erasers and Readers of Histone Acetylation and Methylation in Eukaryotes'. *Nucleic Acids Research* 45 (Database issue): D264–70. https://doi.org/10.1093/nar/gkw1011.
61. Yun, Miyong, Jun Wu, Jerry L. Workman, and Bing Li. 2011. 'Readers of Histone Modifications'. *Cell Research* 21 (4): 564. https://doi.org/10.1038/cr.2011.42.

62. Zhang, Pingyu, Keila Torres, Xiuping Liu, Chang-gong Liu, and Raphael E. Pollock. 2016. ‘An Overview of Chromatin-Regulating Proteins in Cells’. *Current Protein & Peptide Science* 17 (5): 401–10.
63. Zitnik, Marinka, Rok Sosič, Marcus W. Feldman, and Jure Leskovec. 2019. ‘Evolution of Resilience in Protein Interactomes across the Tree of Life’. *Proceedings of the National Academy of Sciences* 116 (10): 4426–33. https://doi.org/10.1073/pnas.1818013116.

8. Приложения

Приложение А. Список генов гистонов человека.

Histone type	HGNC Symbol	NCBI gene ID	ENSG	Canonicity*	Function*
H1	HIST1H1A	3024	ENSG00000124610	canonical	COD
H1	HIST1H1C	3006	ENSG00000187837	canonical	COD
H1	HIST1H1D	3007	ENSG00000124575	canonical	COD
H1	HIST1H1E	3008	ENSG00000168298	canonical	COD
H1	HIST1H1B	3009	ENSG00000184357	canonical	COD
H1	H1F0	3005	ENSG00000189060	variant	COD
H1	HIST1H1T	3010	ENSG00000187475	variant	COD
H1	H1FNT	341567	ENSG00000187166	variant	COD
H1	H1FOO	132243	ENSG00000178804	variant	COD
H1	HILS1	373861	ENSG00000188662	variant	COD
H1	H1FX	8971	ENSG00000184897	variant	COD
H2A	HIST1H2AG	8969	ENSG00000196787	canonical	COD
H2A	HIST1H2AI	8329	ENSG00000196747	canonical	COD
H2A	HIST1H2AK	8330	ENSG00000275221	canonical	COD
H2A	HIST1H2AL	8332	ENSG00000276903	canonical	COD
H2A	HIST1H2AM	8336	ENSG00000278677	canonical	COD
H2A	HIST1H2AC	8334	ENSG00000180573	canonical	COD
H2A	HIST1H2AD	3013	ENSG00000196866	canonical	COD
H2A	HIST1H2AE	3012	ENSG00000277075	canonical	COD
H2A	HIST1H2AH	85235	ENSG00000274997	canonical	COD
H2A	HIST1H2AJ	8331	ENSG00000276368	canonical	COD
H2A	HIST2H2AB	317772	ENSG00000184270	canonical	COD
H2A	HIST2H2AC	8338	ENSG00000184260	canonical	COD
H2A	HIST3H2A	92815	ENSG00000181218	canonical	COD
H2A	HIST1H2APS4	8333	ENSG00000218690	canonical	PS
H2A	H2AFJ	55766	ENSG00000246705	variant	COD
H2A	H2AFX	3014	ENSG00000188486	variant	COD
H2A	H2AFZ	3015	ENSG00000164032	variant	COD
H2A	H2AFV	94239	ENSG00000105968	variant	COD
H2A	H2AFY	9555	ENSG00000113648	variant	COD
H2A	H2AFY2	55506	ENSG00000099284	variant	COD
H2A	HIST1H2AA	221613	ENSG00000164508	variant	COD

H2A	H2AFB1	474382	ENSG00000274183	variant	COD
H2A	H2AFB2	474381	ENSG00000277858	variant	COD
H2A	H2AFB3	83740	ENSG00000277745	variant	COD
H2A	HYPM	25763	ENSG00000187516	variant	COD
H2B	HIST1H2BB	3018	ENSG00000276410	canonical	COD
H2B	HIST1H2BC	8347	ENSG00000180596	canonical	COD
H2B	HIST1H2BE	8344	ENSG00000274290	canonical	COD
H2B	HIST1H2BF	8343	ENSG00000277224	canonical	COD
H2B	HIST1H2BG	8339	ENSG00000278588	canonical	COD
H2B	HIST1H2BI	8346	ENSG00000273802	canonical	COD
H2B	HIST1H2BD	3017	ENSG00000158373	canonical	COD
H2B	HIST1H2BH	8345	ENSG00000275713	canonical	COD
H2B	HIST1H2BJ	8970	ENSG00000124635	canonical	COD
H2B	HIST1H2BK	85236	ENSG00000197903	canonical	COD
H2B	HIST1H2BL	8340	ENSG00000185130	canonical	COD
H2B	HIST1H2BM	8342	ENSG00000273703	canonical	COD
H2B	HIST1H2BN	8341	ENSG00000233822	canonical	COD
H2B	HIST1H2BO	8348	ENSG00000274641	canonical	COD
H2B	HIST2H2BE	8349	ENSG00000184678	canonical	COD
H2B	HIST2H2BF	440689	ENSG00000203814	canonical	COD
H2B	HIST3H2BB	128312	ENSG00000196890	canonical	COD
H2B	H2BFS	54145	ENSG00000234289	variant	COD
H2B	H2BFM	286436	ENSG00000101812	variant	COD
H2B	H2BFWT	158983	ENSG00000123569	variant	COD
H2B	HIST1H2BA	255626	ENSG00000146047	variant	COD
H3	HIST1H3A	8350	ENSG00000275714	canonical	COD
H3	HIST1H3B	8358	ENSG00000274267	canonical	COD
H3	HIST1H3C	8352	ENSG00000278272	canonical	COD
H3	HIST1H3D	8351	ENSG00000197409	canonical	COD
H3	HIST1H3E	8353	ENSG00000274750	canonical	COD
H3	HIST1H3F	8968	ENSG00000277775	canonical	COD
H3	HIST1H3G	8355	ENSG00000273983	canonical	COD
H3	HIST1H3H	8357	ENSG00000278828	canonical	COD
H3	HIST1H3I	8354	ENSG00000275379	canonical	COD
H3	HIST1H3J	8356	ENSG00000197153	canonical	COD
H3	HIST2H3A	333932	ENSG00000203852	canonical	COD

H3	HIST2H3C	126961	ENSG00000203811	canonical	COD
H3	HIST2H3D	653604	ENSG00000183598	canonical	COD
H3	HIST2H3PS2	440686	ENSG00000203818	canonical	PS
H3	H3F3A	3020	ENSG00000163041	variant	COD
H3	H3F3B	3021	ENSG00000132475	variant	COD
H3		391769	ENSG00000269466	variant	COD
H3		340096		variant	COD
H3	H3F3C	440093	ENSG00000188375	variant	COD
H3	CENPA	1058	ENSG00000115163	variant	COD
H3	HIST3H3	8290	ENSG00000168148	variant	COD
H4	HIST1H4A	8359	ENSG00000278637	canonical	COD
H4	HIST1H4B	8366	ENSG00000278705	canonical	COD
H4	HIST1H4C	8364	ENSG00000197061	canonical	COD
H4	HIST1H4D	8360	ENSG00000277157	canonical	COD
H4	HIST1H4E	8367	ENSG00000276966	canonical	COD
H4	HIST1H4F	8361	ENSG00000274618	canonical	COD
H4	HIST1H4H	8365	ENSG00000158406	canonical	COD
H4	HIST1H4I	8294	ENSG00000276180	canonical	COD
H4	HIST1H4J	8363	ENSG00000197238	canonical	COD
H4	HIST1H4K	8362	ENSG00000273542	canonical	COD
H4	HIST1H4L	8368	ENSG00000275126	canonical	COD
H4	HIST2H4A	8370	ENSG00000270882	canonical	COD
H4	HIST2H4B	554313	ENSG00000270276	canonical	COD
H4	HIST4H4	121504	ENSG00000197837	canonical	COD
H2A	HIST1H2AB	8335	ENSG00000278463	variant	COD
H4	HIST1H4G	8369	ENSG00000275663	canonical	COD
H2A	HIST2H2AA3	8337	ENSG00000203812	variant	COD
H2A	HIST2H2AA4	723790	ENSG00000272196	canonical	COD
H2	H2AFVP1	654500	ENSG00000258741	variant	PS
H2	H2AFZP1	54049	ENSG00000213440	variant	PS
H2	H2AFZP2	346990	ENSG00000253173	variant	PS
H2	H2AFZP3	728023	ENSG00000218502	variant	PS
H2	H2AFZP4	100462795	ENSG00000255329	variant	PS
H2	H2AFZP5	100288330	ENSG00000234612	variant	PS
H2	H2AFZP6	100462800	ENSG00000233733	variant	PS
H3	H3F3AP1	654505	ENSG00000259389	variant	PS

H3	H3F3AP2	664611	ENSG00000270433	variant	PS
H1	HIST1H1PS1	387325	ENSG00000216331	variant	PS
H1	HIST1H1PS2	10338	ENSG00000224447	variant	PS
H2A	HIST1H2APS1	387319	ENSG00000216436	variant	PS
H2A	HIST1H2APS2	85303	ENSG00000242387	variant	PS
H2A	HIST1H2APS3	387323	ENSG00000218281	variant	PS
H2A	HIST1H2APS5	10341	ENSG00000234816	variant	PS
H2B	HIST1H2BPS1	100288742	ENSG00000178762	variant	PS
H2B	HIST1H2BPS2	10340	ENSG00000217646	variant	PS
H3	HIST1H3PS1	100289545	ENSG00000220875	variant	PS
H4	HIST1H4PS1	10337	ENSG00000217862	variant	PS
H2B	HIST2H2BA	337875	ENSG00000223345	variant	PS
H2B	HIST2H2BB	338391	ENSG00000240929	variant	PS
H2B	HIST2H2BC	337873	ENSG00000261716	variant	PS
H2B	HIST2H2BD	337874	ENSG00000220323	variant	PS
H2B	HIST3H2BA	337872	ENSG00000181201	variant	PS
H2A	HIST1H2APS6	42647	ENSG00000223383	variant	PS
H2B	HIST1H2BPS3	42633	ENSG00000226908	variant	PS
H2B	H2BFXP	25757		variant	PS
H3	HIST2H3DP1	43797	ENSG00000213244	variant	PS

Примечание * Canonicity - класс каноничности гистона (каноничный/гистоновый вариант), Function - функциональность (COD - белок-кодирующий ген, PS - псевдоген)

Приложение Б. Общая статистика по количеству гистонов, партнеров и парных взаимодействий для STRING, BioGrid, IntAct.

категория/БД	STRING	IntAct	BioGrid	пересечение трех БД	общее кол-во уник.
кол-во гистонов	79	40 (37*)	83	34	85
кол-во уник. гистонов	1	0	4		5
кол-во белков-партн.	684	821	1431	96	2333
кол-во уник. белков-партн.	317	60	949		1826
кол-во свидетельств о взаимод. до уч. идентичн.канонич.	9621	1138	3697	27	13887
кол-во свидетельств о взаимод. после уч. идентичн.канонич.	19635	11206	23806	609	55609
кол-во уник. взаиимод. до уч. идентичн. канонич.	9140	997	3208		13345
количество уникальных взаимодействий после уч. идентичн. канонич.	14898	9023	18149		48054

примечание * во взаимодействиях гистон-негистон