

МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
имени М.В. ЛОМОНОСОВА

БИОЛОГИЧЕСКИЙ ФАКУЛЬТЕТ
Кафедра биоинженерии

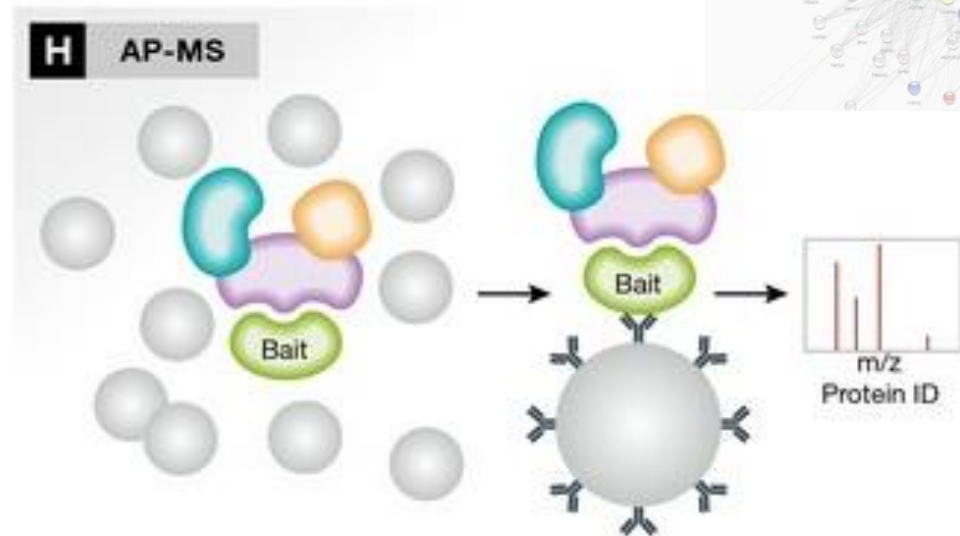
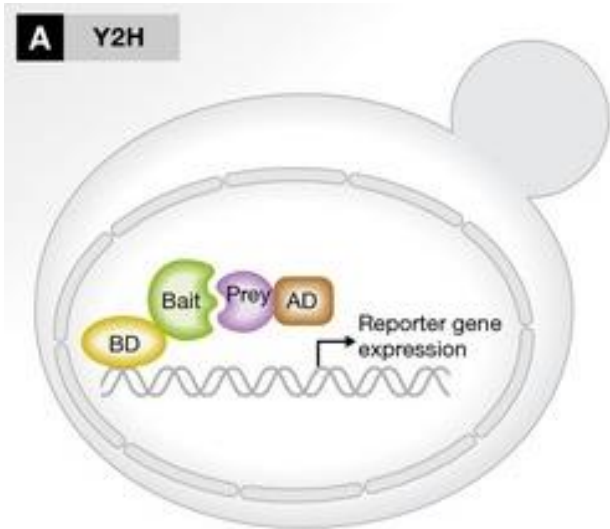
ПОСТРОЕНИЕ И АНАЛИЗ ИНТЕРАКТОМА НУКЛЕОСОМ И БЕЛКОВ ХРОМАТИНА

Грибкова Анна Кирилловна

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

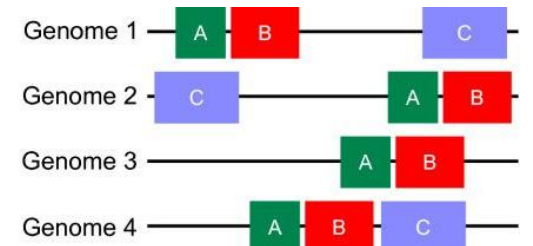
Москва - 2019

Экспериментальные методы

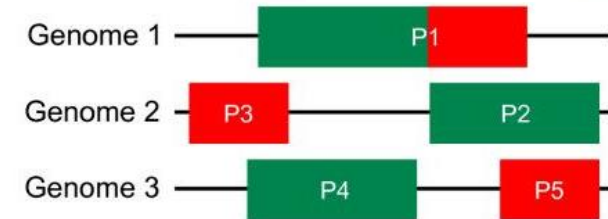


(Snider et al., 2015)

Предсказательные методы



(B) Gene Neighbourhood



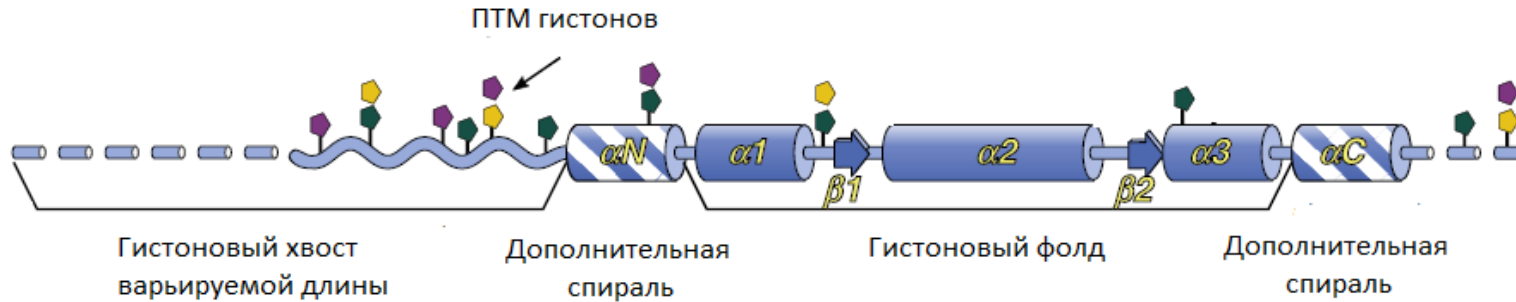
(A) Domain Fusion/Rosetta Stone

	A	B	C	D
Genome 1	1	1	1	0
Genome 2	0	0	1	1
Genome 3	1	1	1	0
Genome 4	1	1	0	1
Genome 5	0	0	1	0

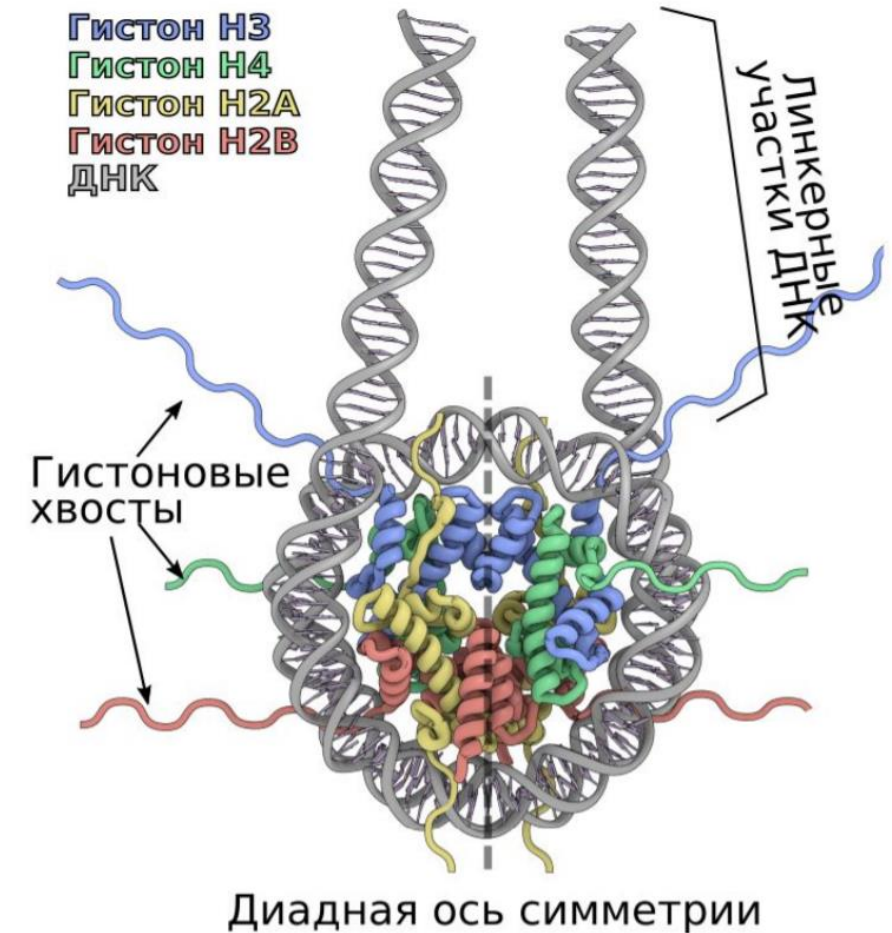
(C) Phylogenetic Profiles

(Raman, 2010)

Гистоны, нуклеосома



Тип и класс каноничности гистона	H2A кан.	H2B кан.	H3 кан.	H4 кан.	H2A вар.	H2B вар.	H3 вар.
Количество белок-кодирующих генов	14	17	13	15	13	4	5
Абсолютная консервативность последовательностей, %	91,5	91,3	99,3	100			



(Armeev et al., 2018)

Постановка задачи

Цель работы

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

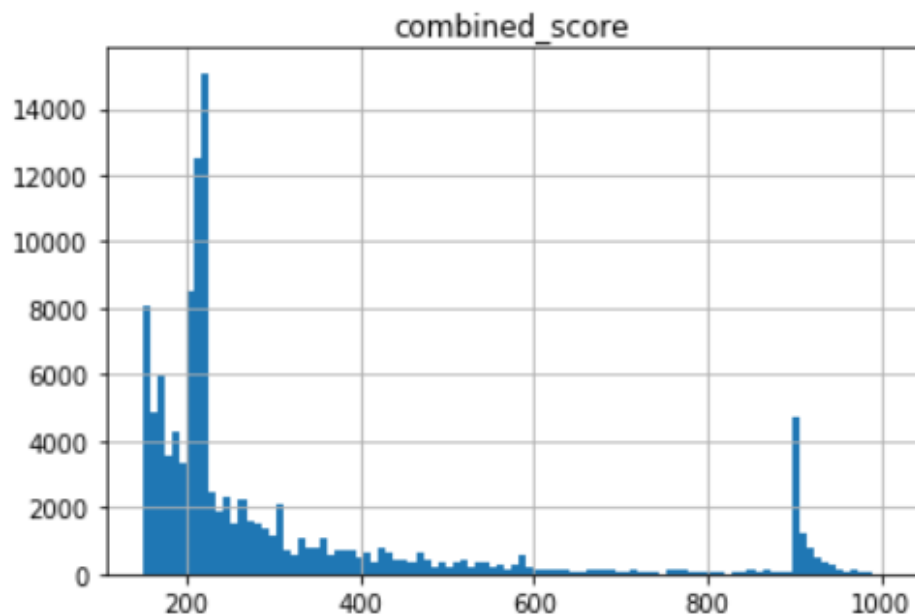
Задачи

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.

Методы. Базы данных ББВ.



- Экспериментальные данные
- БД сигнальных путей (KEGG, Reactome)
- Интеллектуальный анализ текстов
- Ко-экспрессия (ДНК- и РНК-микрочипы)



Анализ геномного контекста:

- Поиск совместно расположенных генов (neighbourhood)
- Поиск слияния генов (фьюжн-событий)
- Сравнение филогенетических профилей белков (cooccurence)

+ Перенесенные взаимодействия

Методы. Базы данных ББВ. BioGrid и IntAct

The screenshot shows the BioGRID 3.5 homepage. At the top, there's a navigation bar with links: home, help, wiki, tools, contribute, stats, downloads, partners, about us. The main heading is "Welcome to the Biological General Repository for Interaction Datasets". Below it, a paragraph describes BioGRID as an interaction repository with data compiled through comprehensive curation efforts. It mentions the current index version 3.5.171 and searches 69,031 publications for 1,676,780 protein and genetic interactions, 28,093 chemical associations, and 726,378 post-translational modifications. A search box is prominently displayed with the text "Q Search BioGRID:" and a dropdown menu set to "By Identifier". Below the search box is a text input field with "Homo sapiens" and a "Submit Identifier Search" button. To the right of the search box are links for "Advanced Search", "Helpful Search Tips", and "Featured Datasets". At the bottom left, there are buttons for "BioGRID Statistics" and "Latest Downloads". Below the main content, there's a "Related Projects" section with links to "BioGRID ORCS - An open repository of CRISPR screens", "Gene Info - Chrome Extension", and "The Kinase and Phosphatase Interactome in S. cerevisiae Curation Project". On the right side, there's a "Partners" section with logos for NIH, ORIP, CIHR IRSC, Mount Sinai Hospital, Princeton University, and Université de Montréal. Below that is a "Latest Updates" section with a BioGRID logo and text: "BioGRID Build 3.5.171 is released with 69,031".

The screenshot shows the IntAct Molecular Interaction Database homepage. At the top, there's a navigation bar with links: Home, Advanced Search, About, Resources, Download. The main heading is "IntAct Molecular Interaction Database". Below it, a paragraph describes IntAct as a freely available, open source database system and analysis tools for molecular interaction data. It mentions that all interactions are derived from literature curation or direct user submissions and are freely available. The IntAct Team also produces the Complex Portal. A search box is prominently displayed with the text "Search in IntAct" and a dropdown menu set to "Enter search term(s)...". Below the search box is a "Search" button and a link for "Search Tips". To the right of the search box, there's an "Examples" section with a list of search terms: "Gene, Protein, RNA or Chemical name: BRCA2, Staurosporine", "UniProtKB or ChEBI AC: Q06609, CHEBI:15996", "UniProtKB ID: LCK_HUMAN", "RNACentral ID: URS00004C95F4_559292", "PMID: 25416956", and "IMEx ID: IM-23318". Below the search box, there are three main sections: "Data Content", "Submission", and "Contributors". The "Data Content" section lists: Publications: 20550, Interactions: 882962, and Interactors: 109631. The "Submission" section has a link to "Submit your data to IntAct to increase its visibility and usability!". The "Contributors" section has a link to "Manually curated content is added to IntAct by curators at the EMBL-EBI and the following organisations:". Below these sections, there's a "Citing IntAct" section with a link to "The MINTAct project-- IntAct as a common curation platform for 11 molecular interaction databases." and a citation for Orchard S et al [PMID:24234451] with a link to the full text. At the bottom, there's a row of logos for various databases and organizations: MINT, UniProt, SIB, I2D, InnoDB, Molecular Connections, MatrixDB, MBInfo, AgBase, G, and IMEx. A note at the bottom right states: "IntAct is a member of the IMEx Consortium."

Результаты. Список гистоновых генов

- Количество белок-кодирующих генов - 93 (включая H1)
- Количество псевдогенов – 30
- Сплайс-изоформы для 23 белок-кодирующих гистоновых генов

Histone type	HGNC Symbol	NCBI gene ID	ENSG	ENST	ENSP	Uniprot	Biotype	Canonicity	Function
H1	HIST1H1A	3024	ENSG00000124610	ENST00000244573	ENSP00000244573	Q02539	Protein coding	canonical	COD
H1	HIST1H1C	3006	ENSG00000187837	ENST00000343677	ENSP00000339566	P16403	Protein coding	canonical	COD
H1	HIST1H1D	3007	ENSG00000124575	ENST00000244534	ENSP00000244534	P16402	Protein coding	canonical	COD
H1	HIST1H1E	3008	ENSG00000168298	ENST00000304218	ENSP00000307705	P10412	Protein coding	canonical	COD

Результаты. Построение интерактома

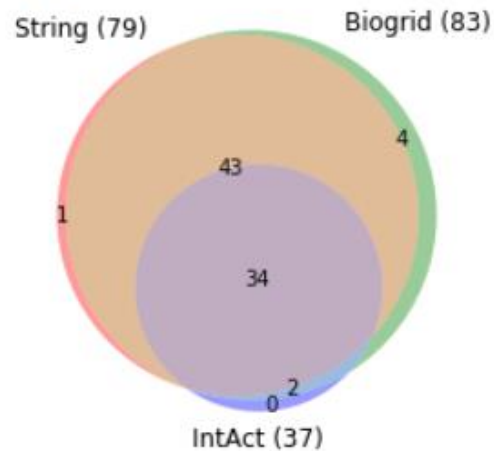
- 15402 экспериментальных/ предсказанных свидетельств
- 13887 взаимодействий (96 % из них уникальны для одной из БД)
- 85 коровых гистонов (3 псевдогена)
- 2333 взаимодействующих с гистонами партнера

DB	Method	Score	Source	Throughput	protein1	protein2	Histone type	Canonicity	class_new_new
BIOGRID	Affinity Capture-MS	0.953827695	BioGrid	High Throughput	HIST1H2BA	NAP1L1	H2B	variant	[Histone chaperones, Replication associated, G...
BIOGRID	Affinity Capture-MS	0.981664185	BioGrid	High Throughput	HIST1H2BA	RUVEL1	H2B	variant	[Hist mod: PTM writers, Chromatin remodellers,...
BIOGRID	Affinity Capture-MS	0.999999993	BioGrid	High Throughput	HIST1H2BA	MIER1	H2B	variant	[TF]
BIOGRID	Affinity Capture-MS	0.918775418	BioGrid	High Throughput	HIST1H2BA	HDAC1	H2B	variant	[Hist mod: PTM erasers, PTM readers, Chromatin...

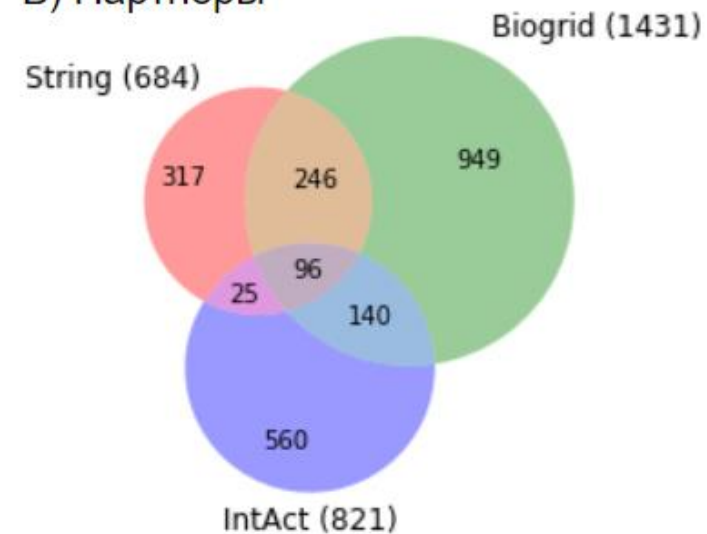
Результаты

Анализ содержания
баз данных (STRING,
BioGrid, IntAct)

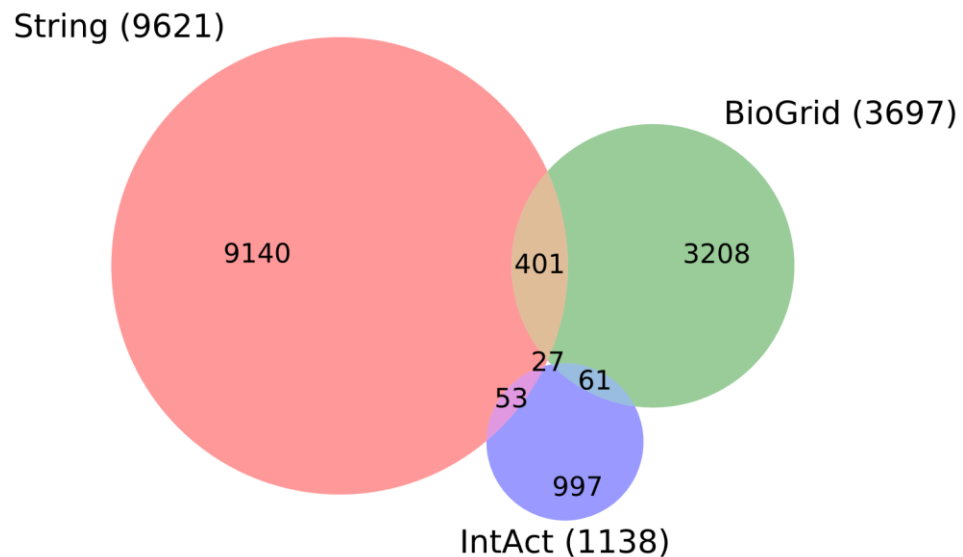
А) ГИСТОНЫ



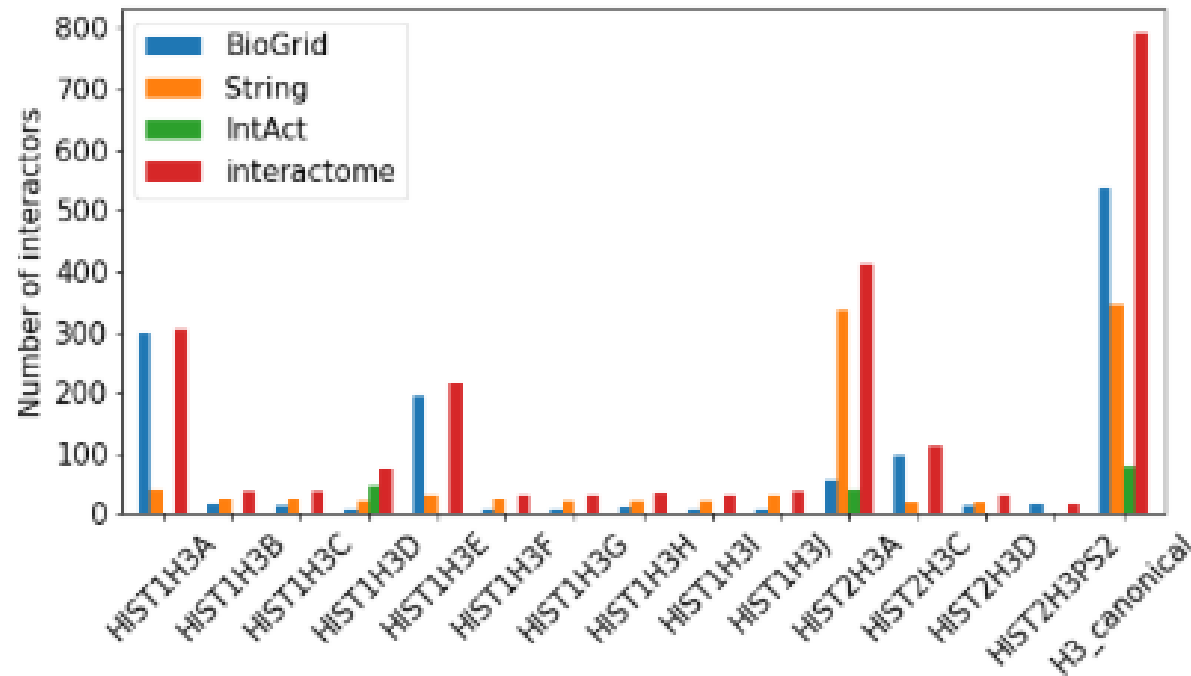
Б) Партнеры



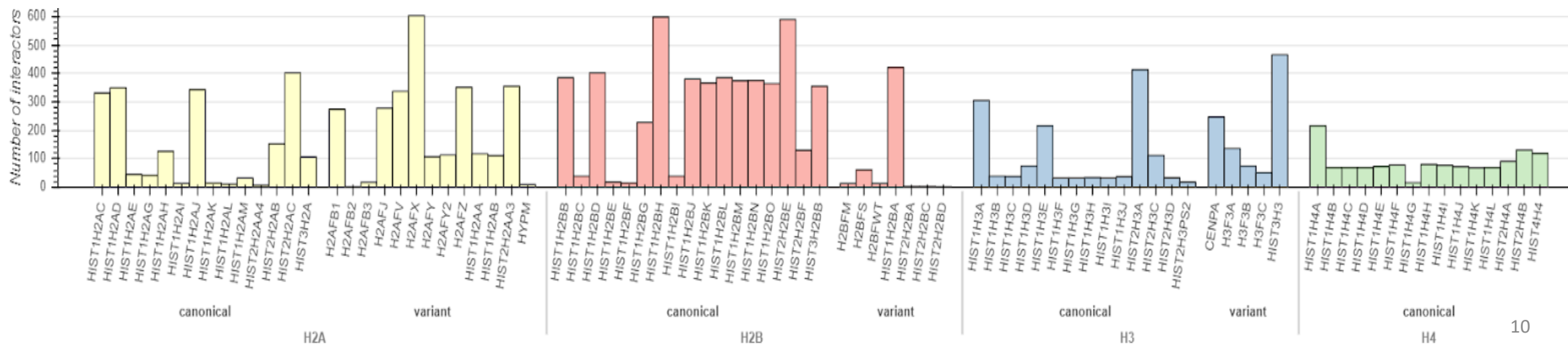
В) Парные взаимодействия



Количество белковых партнеров для канонических гистонов H3 из разных БД



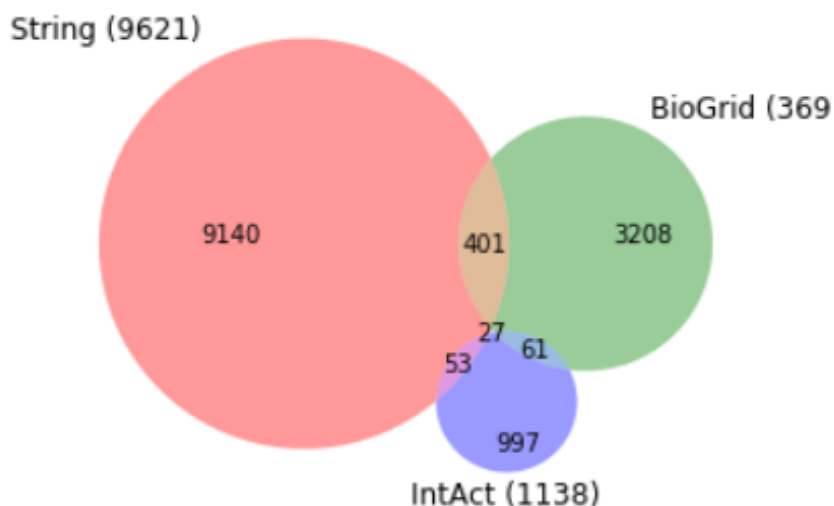
Количество белковых партнеров для всех гистонов в интерактоме



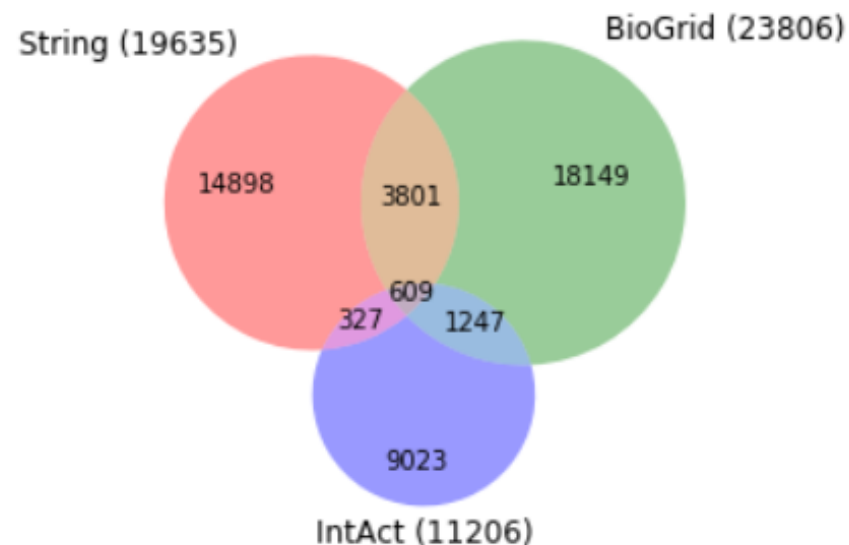
Учет мультигенности семейств канонических гистонов

После учета количество взаимодействующих партнеров для канонических H2A - 560, для H2B - 1154, для H3 - 793 и 350 для H4.

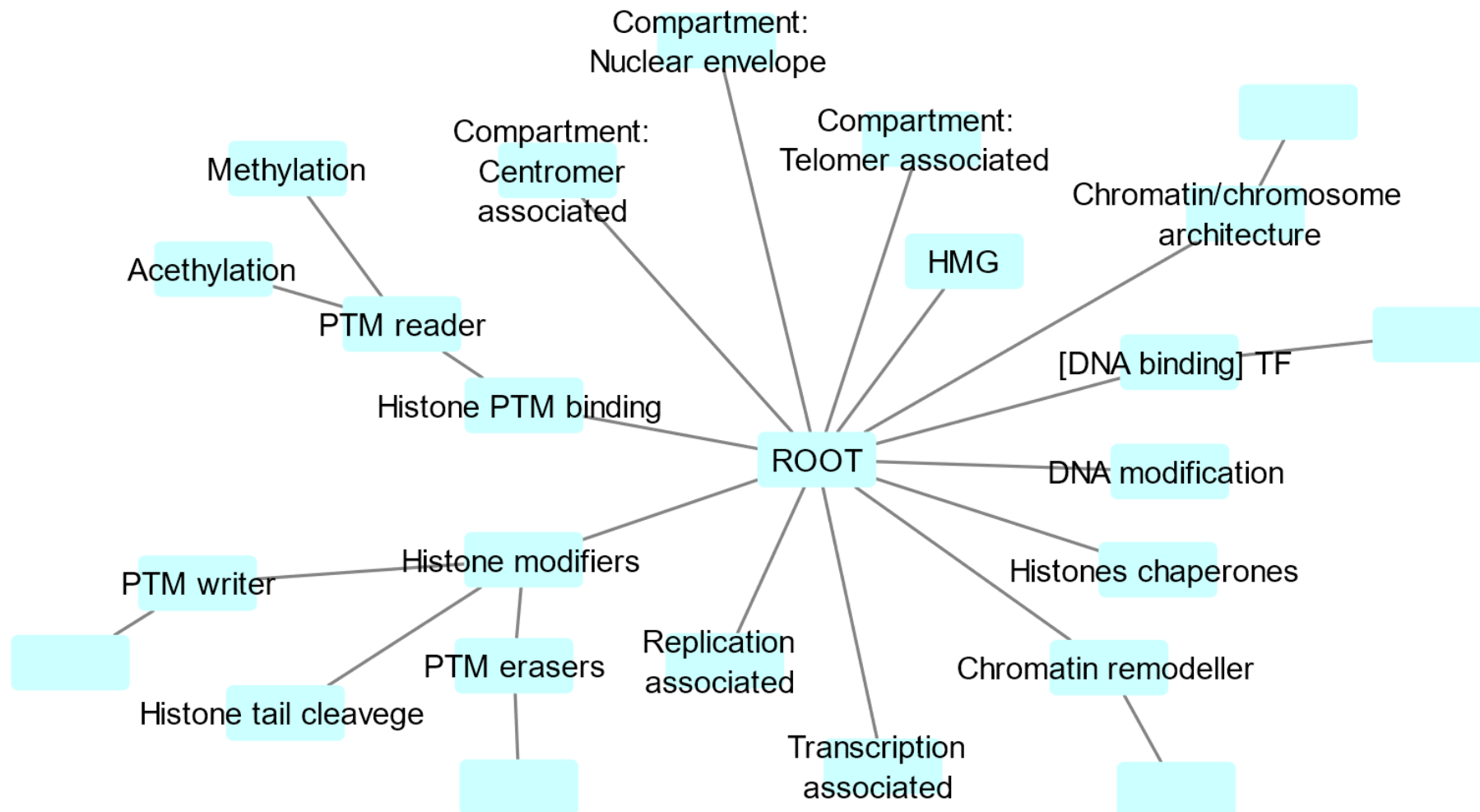
В) Парные взаимодействия
до коррекции



Г) Парные взаимодействия
после коррекции



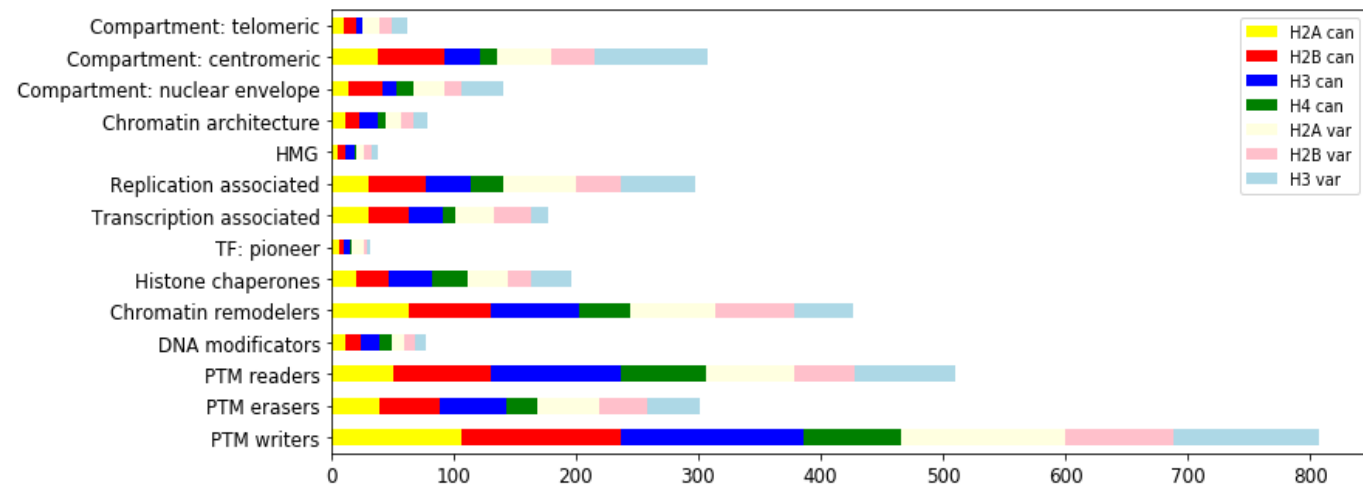
Иерархическая функциональная классификация белков-партнеров



Классификация гистоновых белков-партнеров

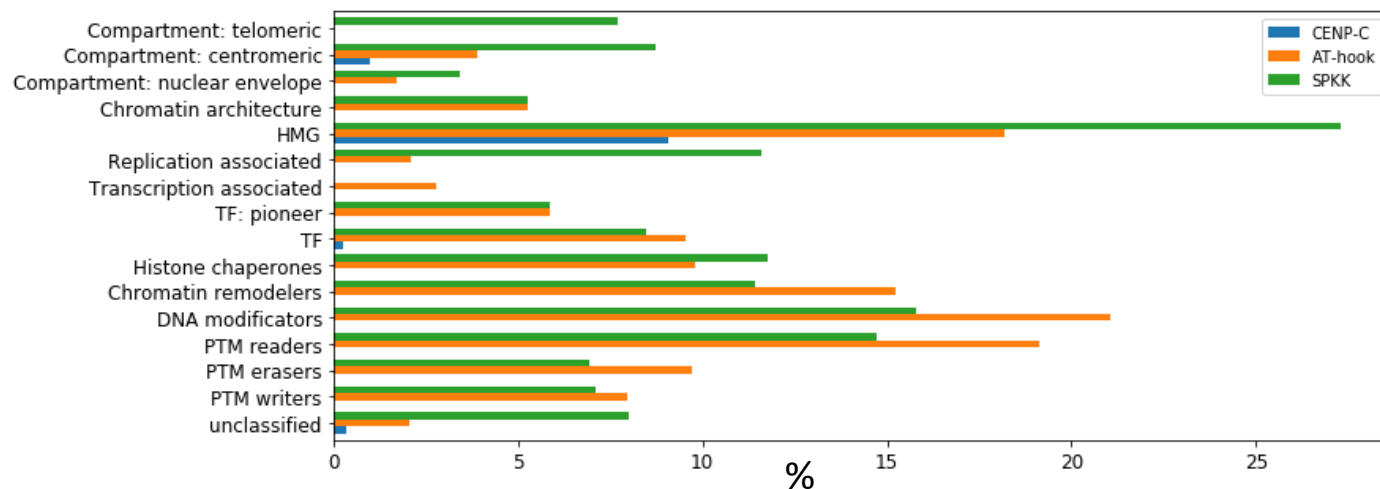
Функциональная категория	Количество генов по данным литературных источников	Количество пересечений с интерактомом	%
Гистоновые шапероны	36	34	94
Хроматиновые ремоделеры	82	69	84
Белки, удаляющие ПТМ	63	50	79
Белки, считывающие ПТМ	142	111	78
Белки, наносящие ПТМ	118	85	72
Транскрипционные факторы	795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

Количество белков определенных функциональных категорий, взаимодействующих с гистонами одного типа и класса каноничности



Результаты. Мотивы и домены

Процентное количество белков определенных функциональных категорий, содержащих рассматриваемые мотивы



CENP-C - 8

AT-hook - 108

SPKK - 176

Лидирующие доменные кланы в хроматиновых ремоделерах

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

Выводы

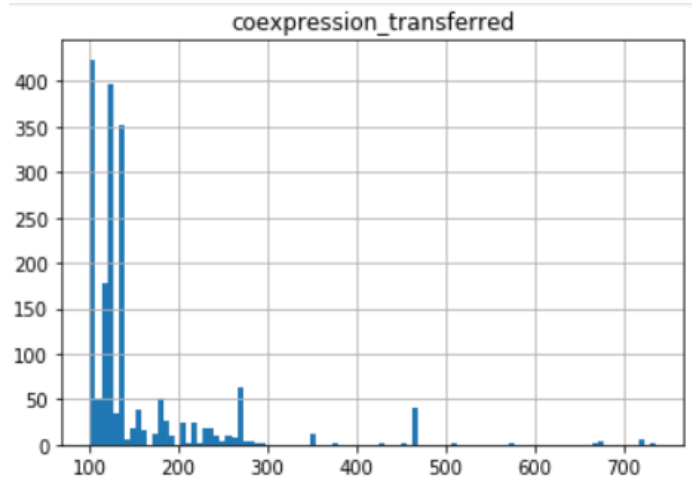
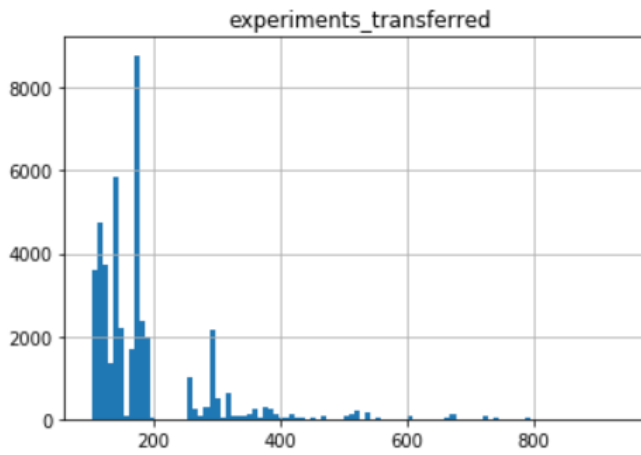
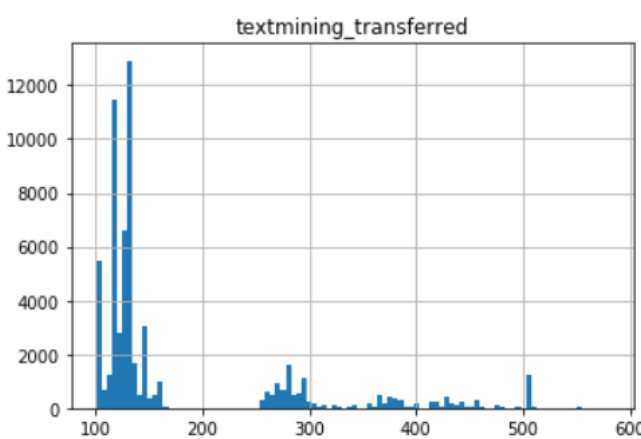
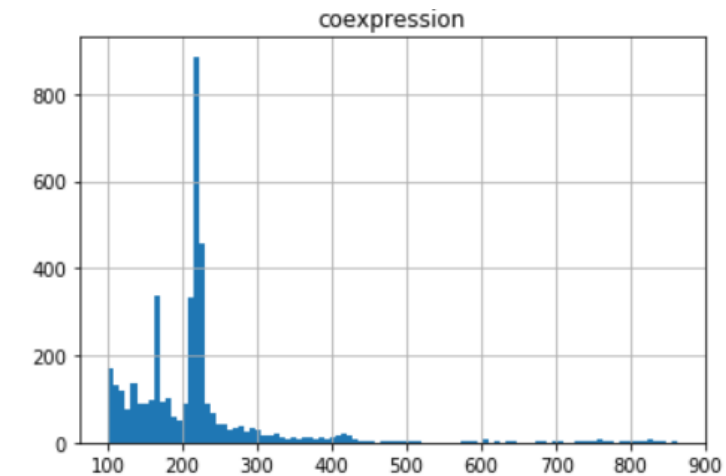
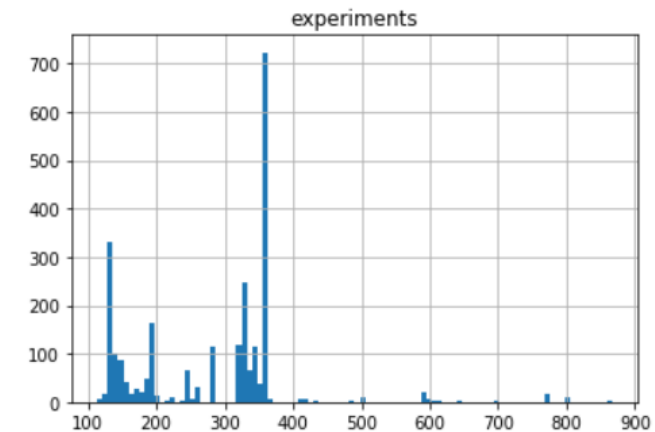
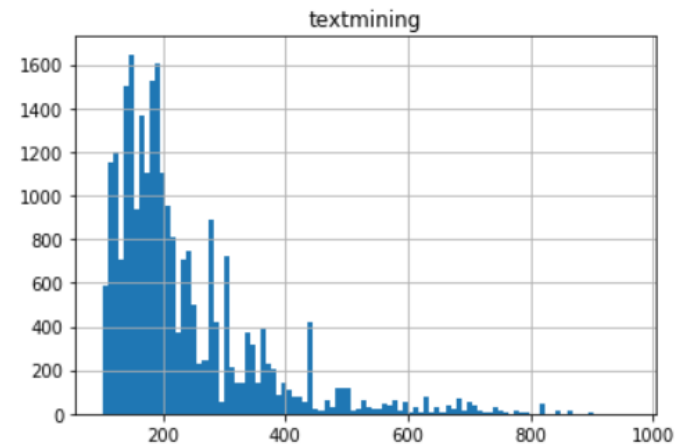
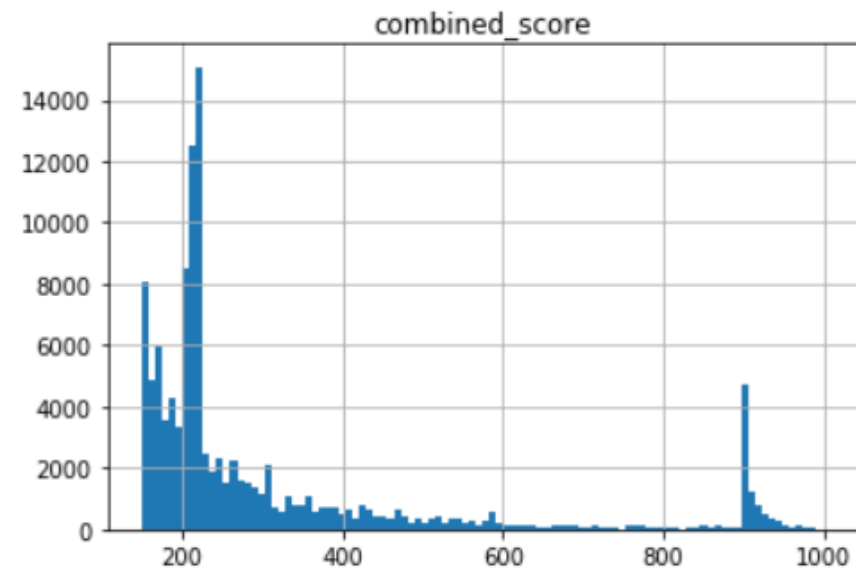
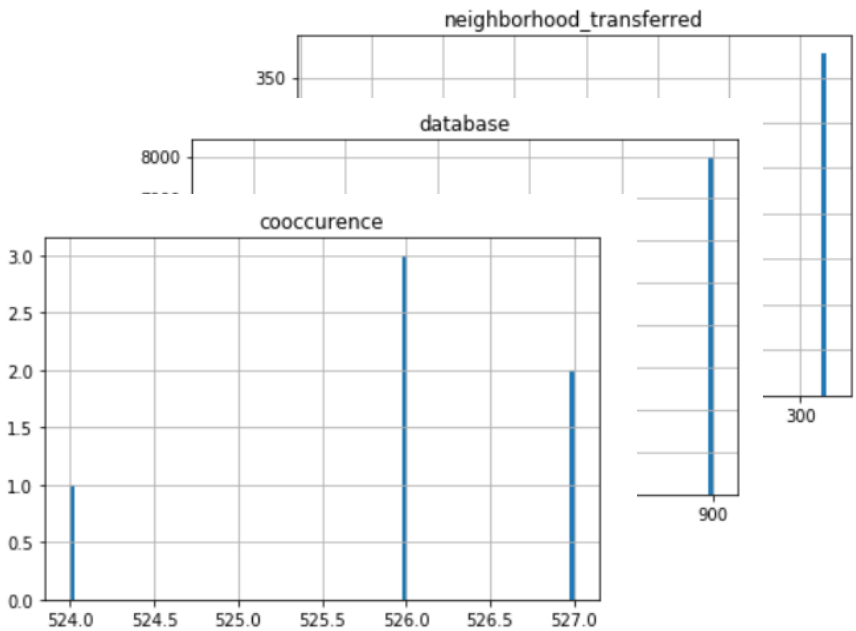
1. Разработан полуавтоматизированный программный код на языке Python, позволяющий загружать и анализировать данные ББВ из открытых БД.
2. Построен нуклесомный интерактом по материалам баз данных STRING, IntAct и BioGrid, включающий в себя 85 коровых гистона, 2333 белковых партнера, 15402 экспериментальных/предсказательных 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.

Благодарности

Выражаю благодарность научному руководителю в.н.с А.К. Шайтану и м.н.с. Г.А. Армееву за постановку задач, обсуждение результатов и помощь в выполнении работы

Дополнительные слайды

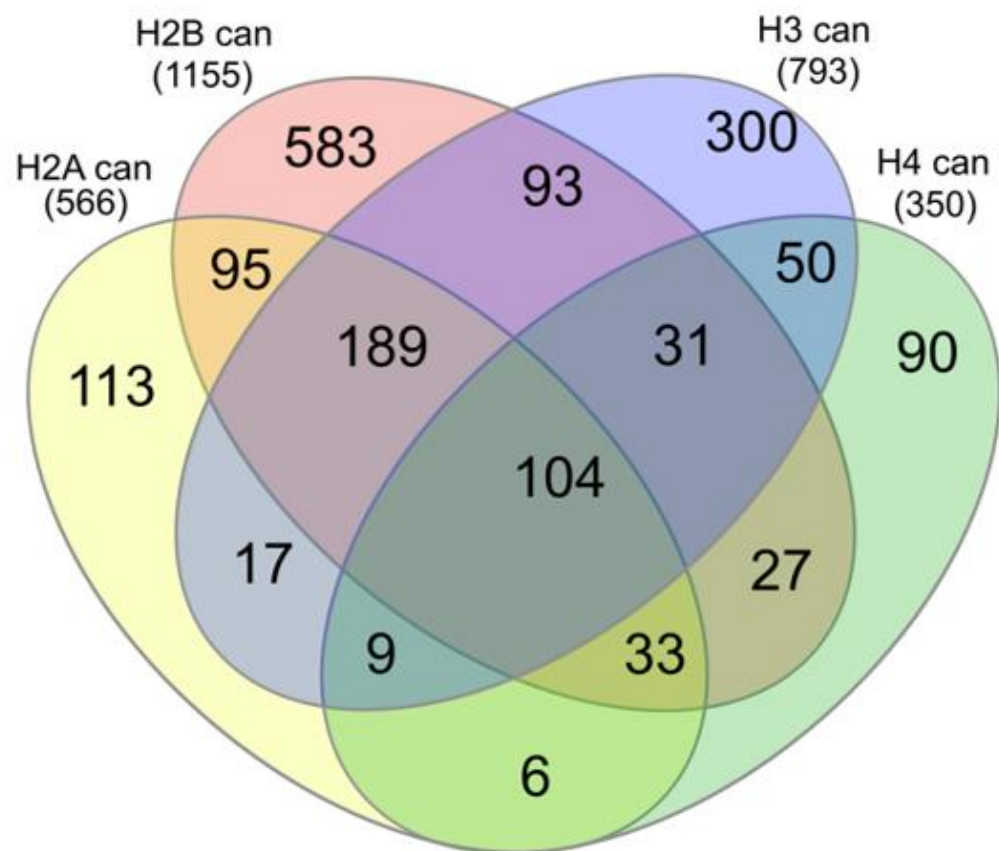
neighborhood, fusion, db_transf = None



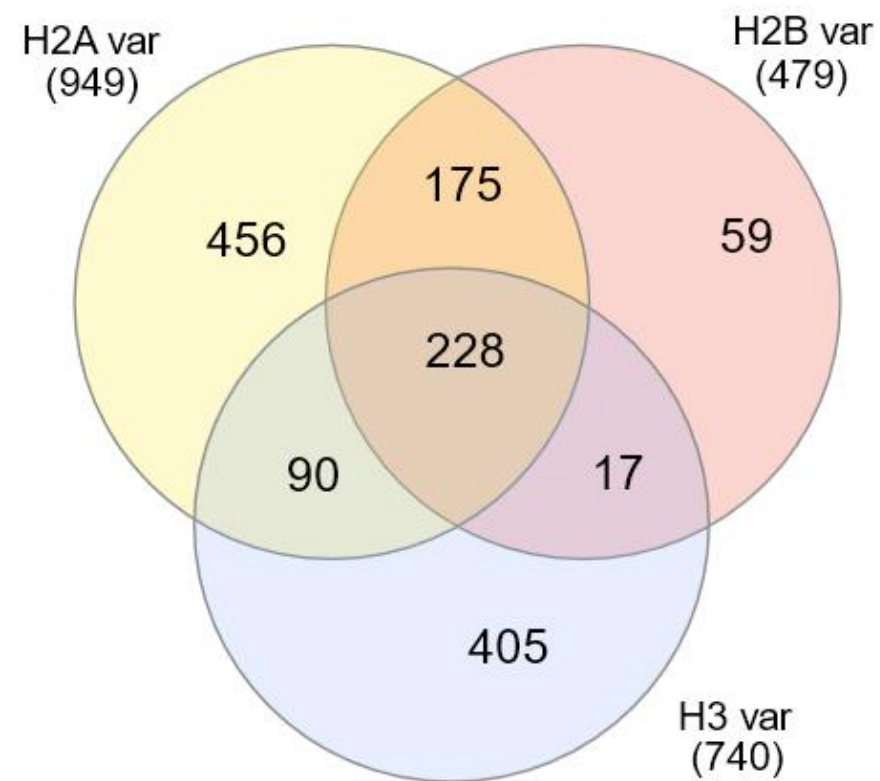
Фрагмент выравнивания (1-100 АК) для вариантных гистонов H2A, закрашенный фон – консервативность > 80 %

ENSG00000105968 ENST00000437072 H2AFV	-----MAGGKAGKDSGKAKAKAVSRSQ-----RAGLQV--LELAGN-----
ENSG00000246705 ENST00000501744 H2AFJ	-----GNYAERVVGAGPVYLAADVLE-----YLTAEI--LELAGN-----
ENSG00000105968 ENST00000350771 H2AFV	-----MFVPGRIHRHLKTRTTSHGRVGATAAVYSAAILE-----YLTAEV--LELAGN-----
ENSG00000113648 ENST00000423969 H2AFY	-----MSSRSGKKKSTKTSRSAKAGVIFPVGRMLRYIKKGHPKYRIGVGAPVYMAADVLE-----YLTGNT--LEKK-----
ENSG00000105968 ENST00000521529 H2AFV	-----MFVPGRIHRHLKTRTTSHGRVGATAAVYSAAILE-----YLTAEV--*
ENSG00000113648 ENST00000506671 H2AFY	-----MSSRSGKKKSTKTSRSAKAGVIFPVGRMLRYIKKGHPKYRIGVGAPVYMAADVLE-----YLTAEI--LELAGN-----
ENSG000000099284 ENST00000455786 H2AFY2	-----MSGRSGKKKMSKLSRSARAGVIFPVGRMLRYIKKGTFKYRISVGAPVYMAADVIE-----YLA AEI--LELAGNAARDNKKARIAPRHIL
ENSG00000113648 ENST00000510038 H2AFY	-----MSSRSGKKKSTKTSRSAKAGVIFPVGRMLRYIKKGHPKYRIGVGAPVYMAADVLE-----YLTAEI--LELAGNAARDNKKGRVTPRHIL
ENSG00000246705 ENST00000389078 H2AFJ	-----MSGRGKQGQKVRRAKAKSSRAGLQFPVGRVHRLLRKGNAYERVGAGPVYLAADVLE-----YLTAEI--LELAGN-----
1. ENSG00000105968 ENST00000446531 H2AFV	-MAGGKAGKDSGKAKAKAVSRSQRAGLQFPVGRIRHHLKTRTTSHGRVGATAAVYSAAILE-----YLTAEV--LELAGN-----
1. ENSG00000188486 ENST00000530167 H2AFX	-----MSGRGKTGGKARAKAKSSRAGLQFPVGRVHRLLRKGHYAEIRVGAGPVYLAADVLE-----YLTAEI--LELAGN-----
2. ENSG000000277858 ENST00000354514 H2AFB2	M P R R R R R R G S S G A G G R G R T C S R T V R A E L S F S V S Q V E R S L R E G H Y A Q R L S R T A P V Y L A A V I E Y L T A K V L E L A G N
1. ENSG000000099284 ENST00000373255 H2AFY2	-----MSSRSGKKKMSKLSRSARAGVIFPVGRMLRYIKKGTFKYRISVGAPVYMAADVIE-----YLA AEI--LELAGNAARDNKKARIAPRHIL
1. ENSG000000164508 ENST00000297012 HIST1H2AA	-----MSGRGKTGGKARAKAKSSRAGLQFPVGRIRHRLLRKGNAYEIRVGAGPVYLAADVLE-----YLTAEI--LELAGN-----
1. ENSG00000105968 ENST00000381124 H2AFV	-MAGGKAGKDSGKAKAKAVSRSQRAGLQFPVGRIRHHLKTRTTSHGRVGATAAVYSAAILE-----YLTAEV--*
1. ENSG00000113648 ENST00000511689 H2AFY	-----MSSRSGKKKSTKTSRSAKAGVIFPVGRMLRYIKKGHPKYRIGVGAPVYMAADVLE-----YLTAEI--LELAGNAARDNKKGRVTPRHIL
1. ENSG00000188486 ENST00000375167 H2AFX	-----MSGRGKTGGKARAKAKSSRAGLQFPVGRVHRLLRKGHYAEIRVGAGPVYLAADVLE-----YLTAEI--LELAGN-----
1. ENSG00000274183 ENST00000620016 H2AFB1	M P R R R R R R G S S G A G G R G R T C S R T V R A E L S F S V S Q V E R S L R E G H Y A Q R L S R T A P V Y L A A V I E Y L T A K V P E L A G N
1. ENSG00000105968 ENST00000308153 H2AFV	-MAGGKAGKDSGKAKAKAVSRSQRAGLQFPVGRIRHHLKTRTTSHGRVGATAAVYSAAILE-----YLTAEV--LELAGN-----
1. ENSG000000164032 ENST00000296417 H2AFZ	-MAGGKAGKDSGKAKTAKAVSRSQRAGLQFPVGRIRHHLKTRTTSHGRVGATAAVYSAAILE-----YLTAEV--LELAGN-----
1. ENSG00000277745 ENST00000615853 H2AFB3	M P R R R R R R G S S G A G G R G R T C S R T V R A E L S F S V S Q V E R S L R E G H Y A Q R L S R T A P V Y L A A V I E Y L T A K V L E L A G N
2. ENSG00000246705 ENST00000544848 H2AFJ	-----MSGRGKQGQKVRRAKAKSSRAGLQFPVGRVHRLLRKGNAYERVGAGPVYLAADVLE-----YLTAEI--LELAGN-----
1. ENSG00000105968 ENST00000222690 H2AFV	-MAGGKAGKDSGKAKAKAVSRSQRAGLQFPVGRIRHHLKTRTTSHGRVGATAAVYSAAILE-----YLTAEV--LELAGN-----
1. ENSG00000113648 ENST00000312469 H2AFY	-----MSSRSGKKKSTKTSRSAKAGVIFPVGRMLRYIKKGHPKYRIGVGAPVYMAADVLE-----YLTAEI--LELAGNAARDNKKGRVTPRHIL
1. ENSG00000113648 ENST00000304332 H2AFY	-----MSSRSGKKKSTKTSRSAKAGVIFPVGRMLRYIKKGHPKYRIGVGAPVYMAADVLE-----YLTAEI--LELAGNAARDNKKGRVTPRHIL
1. ENSG00000187516 ENST00000341016 HYPM	-----MSEKKKNKNSNTNNQTDPSRNEQLQVPRSFVDRVQDERDVQSQSSTINTLLTLLDCLADYIMERVG-----YLTAEV--LELAGN-----
1. ENSG000000164032 ENST00000651623 H2AFZ	-MAGGKAGKDSGKAKTAKAVSRSQRAGLQFPVGRIRHHLKTRTTSHGRVGATAAVYSAAILE-----YLTAEV--LELAGN-----

A)



Б)



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