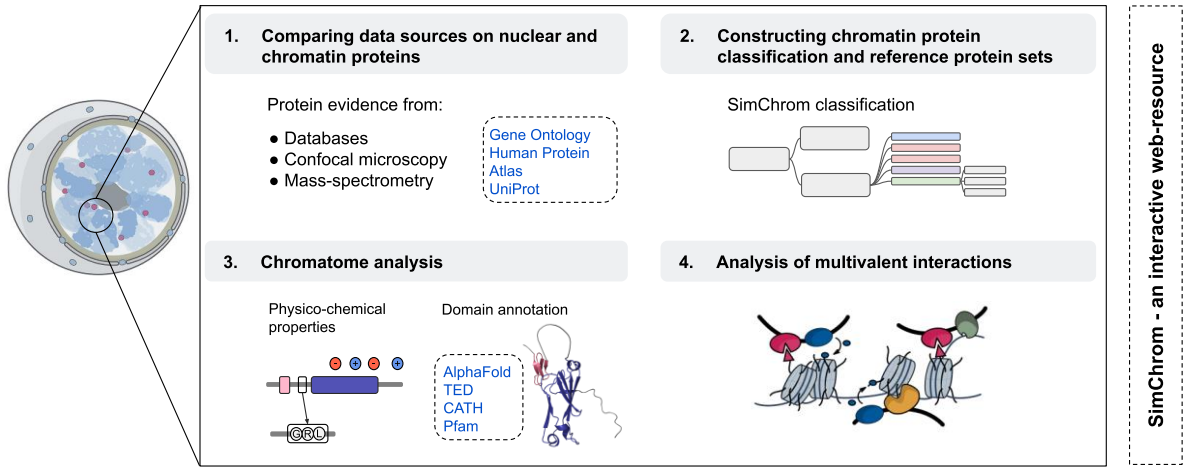
(Re)defining the human chromatome: an integrated meta-analysis of localization, function, abundance, physical properties and domain composition of chromatin proteins

Anna K. Gribkova[1,2], Grigoriy A. Armeev[1,2], Mikhail P. Kirpichnikov[1,3], Alexey K. Shaytan[1,2,4*]

[1] Department of Biology, Lomonosov Moscow State University, Moscow, Russia
[2] Vavilov Institute of General Genetics, Moscow, Russia
[3] Shemyakin–Ovchinnikov Institute of Bioorganic Chemistry, Russian Academy of Sciences, Moscow, Russia
[4] International Laboratory of Bioinformatics, AI and Digital Sciences Institute, Faculty of Computer Science, HSE University, Moscow, Russia
* To whom correspondence should be addressed. E-mail: shaytan_ak@mail.bio.msu.ru

GRAPHICAL ABSTRACT



ABSTRACT

The full complement of chromatin-associated proteins—collectively referred to as the *chromatome*—enables genome functioning in eukaryotes by participating in a wide range of physico-chemical processes. These include mediating diverse specific and non-specific intermolecular interactions, catalyzing *in situ* synthesis and modification of macromolecules, facilitating ATP-dependent chromatin remodeling, *etc.* Despite considerable progress in epigenomics and the structural characterization of many nuclear proteins and their complexes, our understanding of chromatin organization at the proteome scale remains incomplete. This gap hinders the development of a holistic view of genome regulation. In this study, we present a state-of-the-art characterization of the human

chromatome based on an integrative meta-analysis of diverse data sources describing the composition, abundance, and sub-nuclear localization of chromatin proteins. This effort is complemented by original analyses of their physico-chemical properties, domain architectures, and interaction patterns. To support and streamline these analyses, we developed a reference dataset of chromatin proteins, integrated with an empirical, function-based classification ontology and an associated interactive web resource — **SimChrom** — accessible at https://simchrom.intbio.org/. The reference dataset was carefully curated by reconciling data among protein databases, localization, and mass spectrometry-based experimental studies. Sequence-based and AI-assisted structural analyses revealed previously unannotated domains within chromatin proteins that warrant experimental validation, as well as the widespread use of multivalent interaction strategies that underpin chromatin organization. Together, our findings establish a robust framework for future studies aimed at elucidating genome function through detailed analysis of protein–protein and protein–nucleic acid interactions within chromatin.

KEY POINTS

- The first comprehensive meta-analysis of human chromatin proteins that bridges diverse data types
- Established an interactive SimChrom framework for chromatome research available for the community
- Identified functionally relevant hallmarks of chromatin protein organization

Keywords: chromatin, chromatome, epigenomics, proteomics, meta-analysis, genome functioning, protein domains, AI-based protein structure prediction, multivalent protein-protein interactions, intrinsically disordered proteins

1. INTRODUCTION

Chromatin, according to the generally accepted definition, is the complex of DNA, proteins, and associated RNA molecules found in the nuclei of eukaryotic cells [1,2] (**Fig. 1A**, **Interactive Fig. 1** at https://simchrom.intbio.org/#nucleus). However, in the crowded nuclear environment, it is challenging to establish stringent criteria that clearly distinguish between macromolecules that form a complex and those that do not, leaving room for interpretation of this definition. Chromatin proteins, collectively called the chromatome [3,4], enable genome functioning in space and time through active ATP-dependent processes and passive protein-DNA/RNA interactions. This functioning employs non-trivial physical phenomena such as liquid-liquid phase separation [5,6], topological constraints on the DNA, DNA looping and loop extrusion [7,8], diffusion in the crowded macromolecular environment [9], multivalent cooperative interactions [10,11], *etc.*, all of which are regulated by the chromatome composition at specific locations and the properties of individual proteins including their post-translational modifications (PTM), domain architecture and intrinsically disordered regions (IDR).

After the discovery of nucleosomes in 1970-ies chromatin research focused on elucidating molecular underpinnings of the genome organization and function at the nucleosome and supranucleosome levels [1]. During recent decades through the contributions of cryo-EM, epigenomics and 3D genomics much more details on the organization of large macromolecular assemblies [12], protein-DNA interactions and DNA topology [13,14] within chromatin have become available. We are at a point when holistic quantitative or at least qualitative models of the genome functioning based on integrating our knowledge about numerous molecular interactions and processes may seem to be within reach [15,16]. The scope of the data required for such models would lie beyond the one provided within the typical frameworks of genomics and epigenomics, and should also rely on what is sometimes referred to as "chromatomics" [3] - the systematic study of the entire content of the eukaryotic nucleus, including chromatin proteins, their spatio-temporal distribution and interactions. However, our understanding of the protein content of chromatin and its functioning at the "omics"-level is currently lagging behind our ability to probe DNA sequence, its epigenetic markup and 3D contacts. It faces certain challenges, which we detail below, with the human chromatome in mind.

The first set of challenges lies in precisely defining the set of proteins that make up the chromatome. Historically, the definition was operational in nature relying on experimental chromatin extraction followed by the analysis of the protein content via physico-chemical methods (see a historical account by K.E. van Holde [1]) and later by various flavours of mass spectrometry analysis combined with different chromatin extraction and treatment techniques (review by van Mierlo and Vermeulen 2021 [17]). Unsurprisingly, the results of such studies depend on several factors – the details of the chromatin extraction techniques (*e.g.*, non-strongly associating proteins may be not extracted), alternatively cytoplasmic proteins may contaminate the sample [18,19], the sensitivity and the resolution of the analysis method (*e.g.*, low abundant proteins may be not detected, variations in post-translational modifications, alternative splice isoforms) [20–22], and the transient nature of the expression of some nuclear proteins. An additional complication is the heterogeneous and dynamic composition of chromatin (sometimes called a fuzzy organel) – it depends on the cell type, cell cycle phase, as well as on the conditions experienced by the cell [23]. One has to keep in mind also that many proteins shuttle between nucleus and cytoplasm. Recent proteomics studies have estimated the number of chromatin proteins to be around 200 - 3800 [4,23–33]. Despite the above mentioned challenges, for many chromatin analysis tasks having a list of chromatin-associated proteins in the starting point.

Since the human chromatome contains at least several thousand entries, any attempt at the rational understanding and description of its functioning requires some dimensionality reduction approaches. Hence, certain grouping or classification of chromatin proteins that considers their functional properties is desirable. Yet, obtaining such a classification is currently challenging. There are certain historically established classes of chromatin proteins that can be clearly defined (*e.g.*,

histones, high mobility group proteins, *etc*.) [1], however, others have become obsolete (*e.g.*, nuclear matrix proteins [34] or cannot be easily defined (*e.g.*, nucleosol proteins). GeneOntology (GO) currently provides the most comprehensive set of annotations related to different aspects of gene products and is routinely used to interpret large-scale biological data, such as transcriptomics and proteomics results [35,36]. However, it cannot *per se* provide a straightforward and easy to comprehend classification of chromatin proteins due to the presence of a large number of chromatin related GO terms connected into a complex cumbersome hierarchy, which may be incomplete in some cases (*e.g.*, lacks a term for histones) or include obsolete terms in certain cases (*e.g.*, nuclear matrix). While ultimate functional classification of chromatin proteins may likely not be possible (due to the complexity of the genome functioning, different proteins contributing to many different functional processes, *etc.*), some approximation is at least needed to establish a framework for a rational reductionist-wise understanding of chromatin by us humans.

The third set of challenges, in our mind, relates to the need for developing systems biology approaches to describe and study chromatin in a holistic way as a complex functioning system [37,38]. Considerable advances in methodology are currently needed to move from studying the structure of individual macromolecular complexes and analyzing sequence-level (albeit genome-wide) epigenomic data towards the quantitative models of chromatin operation that can grasp the emergence of complex organismal functions. Chromatin functioning relies on complex dynamics networks of multivalent interactions between macromolecules. These interactions depend on the abundance of chromatin proteins in a given compartment, their physico-chemical properties, and domain architectures that mediate specific or non-specific interactions. Understanding these issues at the chromatome-wide scale is a prerequisite for building holistic functional chromatin models.

Motivated by the above mentioned challenges and the overall need to build complex models of chromatin functioning, in this work we attempted to provide the state-of-the-art meta-analysis of what is known about chromatin proteins, their localization, abundance, and properties. The uniqueness of this study is in cross-comparison of different data sources including database information, mass spectrometry data, and protein localization data. Our analysis was challenged by a common problem – the limited congruence between different data sources. To address it we developed several reference datasets for chromatin and nuclear proteins based on cross-comparison of different datasets and manual curation. Next we developed a relatively simple empirical function-based hierarchical classification of chromatin proteins (SimChrom classification) which was instrumental for all downstream analyses by allowing to compare properties between different groups of chromatin proteins. Using this framework we: 1) systematically analyzed the abundance of chromatin proteins, identified potential pitfalls in MS-based datasets, and using whole cell proteomics data quantified the presence of different chromatin proteins and chromatin protein groups in the cell; 2) characterized the interplay between amino acid composition of chromatin proteins, the prevalence of intrinsically disordered regions and specific

distribution of charged amino acids in their sequences; 3) analyzed the current state of structural characterization and domain annotation of chromatin proteins and based on novel AI-enabled protein structure prediction tools identified more than 200 domains in chromatin proteins that belong to currently unknown structural superfamilies and await experimental characterization, 4) characterized typical patterns of multivalent interactions employed by chromatin regulator proteins mainly engaging combinations of histone methylation, acetylation and DNA binding modes.

Finally, we supplement our analyses with an interactive web resource — **SimChrom** — accessible at https://simchrom.intbio.org/. SimChrom harmonizes data from different sources and may be used for exploration of different chromatin protein groups and properties of individual proteins.

The overall scheme of our work is outlined in **Fig. 1B**. Although our analysis in fact required many iterations to achieve self-consistency (*e.g.*, development of SimChrom classification was performed concomitantly with the development of reference chromatin protein sets, and classification was employed in analysis of the quality and content of different data sources) below we present the logic of our analysis as a series of consecutive steps to the extent possible, offloading detailed consideration of certain aspects that may require the familiarization with the whole manuscript to the **Supplementary Results and Discussion** (Suppl. R&D) sections.

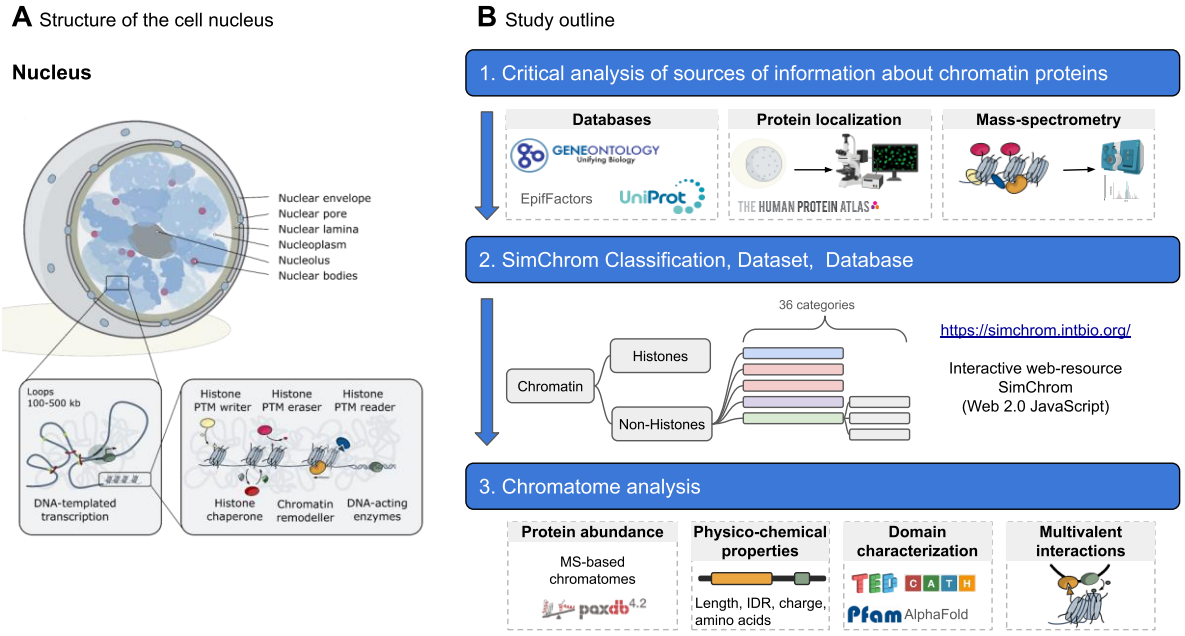


Figure 1. (A) The structure of the nucleus with details of chromatin organisation at the levels of chromatin domains and chromatin regulatory proteins. **(B)** The overview of this study shows sources of information about chromatin proteins (*e.g.*, their functions, subcellular localization, identification by MS-based methods or specific protein functional domains) and their use in the current study.

2. MATERIAL AND METHODS

For the purpose of all analyses in this study we used the set of human protein identifiers and corresponding amino acid sequences representing the canonical protein isoforms (usually corresponding to the major splice isoforms) as provided by the UniProtKB/Swiss-Prot database (also known as the reviewed section of the UniProt Knowledgebase) (UniProt proteome ID UP000005640, release 2022_2) [39]. The set contained 20,272 gene entries corresponding to 20,225 unique protein IDs (some genes code for identical protein sequences). Wherever needed the original datasets were mapped to the above described set of protein UniProt IDs.

2.1. Collection and processing of data about chromatin and nuclear protein repertoires, as well as other protein groups, from databases and MS-based studies

2.1.1. Protein localization data sources (UniProt, HPA, OpenCell)

Protein subcellular localization annotations were obtained from UniProtKB [39] (release 2022_2, subcellular location section), Human Protein Atlas (HPA) [40] (version 22, https://www.proteinatlas.org/download/subcellular_location.tsv.zip, accessed on 09.06.2022), the OpenCell dataset [41] was retrieved from the website (https://opencell.czbiohub.org/, accessed on 10.06.2022). To ensure high-confidence subcellular localization annotations, localization terms were filtered according to database-specific reliability criteria. In UniProt, only annotations supported by at least one evidence tag were retained. In HPA, only annotations with a reliability score exceeding the 'uncertain' threshold were included. In OpenCell, only annotations scoring above the lowest quality grade were retained.

For the analysis of protein multiple localization, localization annotations from HPA and UniProt were grouped into the following generalized categories: Nucleus, Cytoplasm, Endomembrane system, Other (including chromosome, secretory, and extracellular proteins in UniProt). Detailed grouping information is provided in **Suppl. Table ST3**.

To estimate overrepresentation or underrepresentation of a particular localization for a list of proteins enrichment analysis based on hypergeometric test (also known as Fisher's exact test) was used with significance threshold p-value < 0.05. To estimate the congruence of subnuclear localization annotations between UniProt (protein set denoted by A) and HPA (protein set denoted by B) following metrics were used: TP = $|A \cap B|$; FP = $|B \cap \neg A|$; FN = $|A \cap \neg B|$; Union = $|A \cup B|$; Jaccard similarity coefficient = TP / Union. Performance measures: Precision = TP / (TP + FP); Recall = TP / (TP + FN); and F1-score = 2 × Precision × Recall / (Precision + Recall).

2.1.2. GO and function-oriented databases

In **Section 3.1**, for a selected set of Gene Ontology (GO) terms, lists of associated human proteins were retrieved from the QuickGO database [42] (GO annotation set created on 2025-03-06) using its REST API. To capture all relevant proteins, annotations were obtained also for all their descendant terms with the relationships "is_a", "part_of", and "occurs_in". Information on human histone and non-histone epigenetic regulators was obtained from the EpiFactors database [43] (version 2.1, released September 10, 2024). DNA-binding transcription factors (dbTFs) were obtained from https://www.ebi.ac.uk/QuickGO/targetset/dbTF, accessed on 01.09.2024) [44].

2.1.3. MS-based studies

We conducted a comprehensive search on PubMed using keywords "chromatome", "chromatin proteins", "experimentally obtained chromatin proteins", "nuclear proteins", and "nuclear proteome" to gather relevant informational resources containing data on nuclear and chromatin proteins.

All protein entries identified in MS-based studies were mapped to the UniProt human reference proteome (release 2022_02), protein isoforms were collapsed to canonical entries. Gene names were used as secondary identifiers to facilitate mapping for records without a direct match. Outdated UniProt entry identifiers were updated to their current counterparts in the used proteome. Below the specific details of obtaining the protein lists from respective MS-studies are provided (see **Table 1**). From Kustatscher et al. (2014) study [23], which provides interphase chromatin probability scores (ICP) for 7,635 human proteins, proteins with ICP > 0.5 were selected. From Alabert et al. (2014) study [26] entries with missing UniProt ID, gene name, or intensity were excluded, unmapped entries were matched by gene name. No filtering by nascent enrichment or chromatin probability was applied. From Ginno et al. (2018) study [25] proteins quantified with at least two unique peptides and consistent signal across all three replicates of at least one cell cycle stage (G1, S, or M) were selected. Although mitotic chromatin was included, the protein composition across stages was nearly identical, and excluding M-phase would not significantly alter the dataset. From Shi et al. (2021) study [27], protein entries were taken from the experimental run that used conditions involving HaeIII digestion and no 1,6-hexanediol treatment ("condition 2" in the study), representing chromatin-associated proteins extracted under native conditions using the Hi-MS protocol. Only proteins with non-zero iBAQ values in all three replicates were retained. From Torrente et al. (2011) study [4] chromatin-associated protein lists obtained by three three extraction methods were combined (total chromatin extraction, high-salt extraction, and micrococcal nuclease digestion). Protein GI accession numbers were converted to UniProt reviewed entries. From Itzhak et al. (2016) study [45] all entries annotated as "mostly nuclear" by the "Global classifier 2" were taken. From Alvarez et al. (2023) study [28] a combined list of proteins obtained by NCC (Nascent Chromatin Capture) in HeLa S3 cells and iPOND (isolation of Proteins On

Nascent DNA) in TIG-3 fibroblasts was taken. From the Ugur et al. (2023) study [24] proteins with non-missing raw Log2 intensity values in all three chromatin replicates for human embryonic stem cells were taken.

2.1.4. Other sources

Housekeeping proteins were defined as proteins detected in all analyzed tissues by RNA-seq and downloaded from HPA (version 23) [46] (in total 8899 proteins). Manually curated protein complexes were downloaded from Complex Portal (accessed on 7 January 2025) [47]. Only complexes exclusively containing chromatin proteins were selected for domain co-occurrence analysis.

2.2 Construction of reference chromatin and nuclear protein datasets

2.2.1. The SimChrom ontology, SimChrom protein dataset, and SimChrom/SimChrom-SL classification

The SimChrom chromatin proteins classification ontology was developed simultaneously with the corresponding set of human chromatin proteins that were collected according to the developed classification. To this end, for almost every SimChrom classification term we attributed specific terms from GO classification that were manually selected to represent molecular functions and biological processes that happened exclusively inside the cell nucleus and were related to the respective SimChrom term. In certain cases cellular component GO terms were also used when they were directly related to the respective SimChrom term (*e.g.*, complexes of chromatin remodelers). The list of the GO terms attributed to every SimChrom term is given in **Suppl. Table ST4**. The dataset was then supplemented by proteins defined as chromatin proteins by several databases, review papers, and original studies (*e.g.*, HistoneDB 2.0 - histone proteins, Histome2 - for PTM writers, see details and corresponding data sources in **Suppl. Table ST4**). At the last step for certain SimChrom categories the contents of the dataset were additionally filtered to either remove the potentially non-nuclear proteins (in the case of RNA-binding proteins) or histones from the categories belonging to the “non-histone proteins” group (see details in **Suppl. Table ST4**).

To ensure unambiguous categorization, we constructed a SimChrom-SL (single labeled) classification, where each protein was assigned to exactly one category of the same SimChrom ontology. Assignment followed a predefined priority order: "Molecular function" and "Physico-chemical properties" categories were prioritized, followed by others ("Biological processes" and "Genomic location"). Within these groups, categories containing fewer proteins were ordered first (see priority hierarchy in **Suppl. Fig. SF3_2**). Each protein was labeled with its first eligible category in this sequence.

To assess the quality of the SimChrom dataset, Gene Ontology enrichment analysis was performed. Two groups of proteins were analyzed separately: (1) proteins present in SimChrom but absent from NULOC_CS (see below), and (2) proteins present in NULOC_CS but absent from SimChrom. GO enrichment was assessed using g:Profiler [48], with multiple testing corrections applied (Bonferroni correction, significance threshold 0.05). Only driver GO terms (defined by g:Profiler) were selected for further interpretation.

2.2.2. Construction of reference localization datasets

Reference datasets of nuclear (abbreviated as NULOC), non-nuclear (NON_NULOC) and cytoplasmic (CYTLOC) protein entries at different levels of confidence and uniqueness of localization were constructed (see set of localization terms in **Suppl. Table ST3**). The datasets were created using the combination of localization information from UniProt and HPA (full details are provided in **Suppl. Table ST6**). Datasets of proteins having only one specific localization (*i.e.* no other localization reported in the source databases) are denoted by the UL suffix. Datasets, where localization is simultaneously supported both by UniProt and HPA are denoted by CS (consensus) suffix, alternatively, if localization is supported by at least one database the JT (joint) suffix is given. The NECF ("no evidence code filtration") suffix denotes datasets where no preliminary filtration of localization information provided by the databases based on evidence codes or confidence levels were applied. We noticed that many histone protein entries in UniProt lack evidence codes for their localization (in release 2022_2). This is apparently due to the fact that these entries appeared in UniProt before the manually curation of evidence attribution was introduced, and they are still awaiting manual review and retrofitting. We manually added histone proteins in all constructed nuclear reference datasets.

2.3. Protein abundance data processing

2.3.1. PaxDB data

Two dataset from PaxDB [49] version 4.2 with protein abundance data were used: the dataset with the highest proteome coverage ("*H. sapiens* - Whole organism (Integrated)" - covers 99% of human proteome according to PaxDb, referred to as "PaxDb_INT" in this paper) and the dataset with the highest interaction consistency score ("Whole organism, SC (Peptideatlas,aug,2014)" - covers 84% of human proteome according to PaxDb, referred to as "PaxDb_PA" in this paper), see also **Suppl. R&D Sec. 3.1**. An abundance unit is protein per million, ppm, which describes protein abundance relative to all expressed molecules in the proteome. Protein abundance was obtained by aggregating abundance of individual genes with similar protein sequences (*e.g.,* in the case of canonical histones). Cumulative abundance was defined as the sum of protein abundances for a group of protein entries; cumulative weight was calculated as abundance multiplied by the protein molecular weight.

2.3.2. Abundance from MS-based studies

Protein abundance values were derived from mass spectrometry (MS) intensities or iBAQ values reported in individual studies. When a single intensity value was associated with multiple protein identifiers, it was divided equally among them. For proteins represented by multiple entries within a study, the median intensity (or abundance) value was used to obtain a single estimate per protein. For Kustatscher et al. (2014) study protein abundance was defined as the summed MS intensity. For Alabert et al. (2014) study protein abundance was based on MS intensity values. For Ginno et al. (2018) study abundance was defined as the median MS intensity across six replicates (G1 and S phases, three replicates each), based on the "Chromatome Reporters" dataset. For Shi et al. (2021) study abundance was calculated as the median iBAQ value across three replicates. For Itzhak et al. (2016) study for proteins with a single isoform, the "Estimated copy number per cell" was used as the abundance value. For proteins with multiple isoforms, the median copy number across isoforms annotated as "mostly nuclear" was used. For Ugur et al. (2023) study abundance was calculated as the median of the three replicate raw log2 intensity values.

2.4. Protein physico-chemical properties

Intrinsically disordered regions (IDRs) were identified based on solvent-accessible surface area (SASA) profiles, smoothed using a 20-residue sliding window, calculated from AlphaFold2-predicted protein structures [50]. Regions were classified as IDRs if they contained at least four consecutive residues with accessibility values greater than 0.55, and were separated from other IDRs only by non-IDR fragments longer than four residues. Protein tails were defined as N- or C-terminal IDRs. Protein charge was classified as positive (>=1), negative (<=1) and neutral (0) according to the number of positive charged amino acids (lysines and arginines) and negatives (aspartates and glutamates). Protein tail mean net charge was calculated as a sum of amino acid charges averaged per each amino acid position (first 80 amino acid position was analysed). Protein tail charge profile was constructed with averaged charges per amino acid residue with rolling window (size=10 aa).

To get the individual number of protein molecules of the respective charge we adjusted the mentioned number of charged protein entries by protein abundance (PaxDb_PA values were used). Additionally we took into account the overall net charge of every protein, resulting in the analysis of the cumulative charge conferred by positively or negatively charged proteins.

To characterize differences between chromatin and cytoplasmic proteins with respect to their amino acid composition we used the Uniform Manifold Approximation and Projection (UMAP) nonlinear dimensionality reduction technique. Protein amino acid fractions were standardized using StandardScaler and reduced with UMAP with default parameters. The median values of the amino acid fractions in the chromatin proteins and the proteins that were uniquely localized in the nucleus or

cytoplasm were compared using the Mann–Whitney test with Bonferroni correction. The statistical significance threshold was an adjusted p-value of less than 0.05. The fold enrichment of the amino acid fraction was calculated as the difference between the median value in chromatin or nuclear proteins and the median value in cytoplasmic proteins.

2.5. Protein structural characterization and protein domain analysis

Chromatin protein sequence coverage was assessed by identifying all experimentally resolved amino-acid residues for each target protein entry using structures deposited in the PDB database using PDB API (accessed on 15.01.2025). CATH structural domain assignments were used from [51] (CATH v4.3.0, AlphaFold v2). For Pfam domains we used Pfam-A regions from Pfam version 37.0, only the following type of regions were used: Domain, Family, Repeat, Coiled-coil. Data about protein domain annotation in other DBs (*e.g.*, InterPro, PANTHER, *etc*.) was obtained by InterPro API (InterPro version 103.0), only following types of domains were selected: domain, homologous_superfamily, family, repeat, coiled_coil. The percent of Pfam models presented in PDB was obtained by processing data from InterPro. TED domain annotation was downloaded using TED API v1 [52]. The domain annotations from different sources (TED, CATH and Pfam) were intersected if the length of intersection was more than half of the length of shortest annotation.

To look for and map TED domains and Pfam models to PDB or CATH database entries we used Foldseek Search (database versions with following labels were used “PDB100 20240101”, “CATH50 4.3.0”) [53] with easy-search module, --exhaustive-search option and default parameters with following filtration by probability (estimated probability for query and target to be homologous (*e.g.*, being within the same SCOPe superfamily)) higher than 0.9 and query coverage more than 0.5. To characterize representation of chromatin proteins’ TED domains in PDB, the best match PDB structure (with the FoldSeek highest probability score) for every TED domain was identified. TED domains were then classified by the level of sequence identity with their best PDB match. The taxa of the best matches were aggregated into several groups: 'Homo Sapiens', 'Mammalia', 'Other Vertebrata', 'Protostomia', 'Viridiplantae', 'Fungi', 'Archaea', 'Bacteria', 'Viruses', 'Other', 'N/A' (synthetic constructs or unclassified sequences). In the case of Pfam models, for each Pfam model sequence the best match by probability was selected, sequence identities for each Pfam model were averaged by median value. Additional searches of TED domains in CATH were also performed by sequence search using CATH/GENE3D annotations.

Novel structural domains (potentially representing new types of structural superfamilies or protein folds) were defined as TED domains that could not be matched to any known protein structure or protein structure superfamily (by FoldSeek search in PDB and CATH50). For these domains available functional or other annotations were downloaded from InterPro (only matches where more

than half of the TED domain was mapped to the annotated region were considered). DeepFri [54], a Graph Convolutional Network for predicting protein functions, was used to predict GO MF and GO BP terms for novel structural domains by both sequence (CNN model) and structure (GCN model).

Domain diversity of a SimChrom-SL category was calculated as the number of different Pfam domain models found in the proteins belonging to the respective SimChrom-SL category, relative domain diversity - domain diversity divided by the number of proteins in the corresponding SimChrom-SL category.

EMVI-domains were defined as those that were found in multiple copies or in combination with another Pfam domain in at least three chromatin regulator proteins (defined as the following SimChrom-SL categories: Histone chaperones, Histone PTM erasers, Histone PTM writers, Histone PTM readers, Histone modification, Chromatin remodelers, Methylated DNA binding, DNA (de)methylation, RNA modification). EMVI-domains were manually classified based on the information currently available in the literature into the following functional subgroups: Histone methylation, writer; Histone methylation, eraser; Histone methylation, reader; Histone acetylation, writer; Histone acetylation, eraser; Histone acetylation, reader; Histone phosphorylation, writer; Histone binding; DNA binding; DNA methylation; PPI; Dimerization/oligomerization; Chromatin remodeling; RNA binding; Other. The histone modification subgroups were additionally grouped together according to the type of the respective histone modification *(e.g.,* histone methylation). For domain composition scatter plots (*e.g.*,interactive graphs available at: https://simchrom.intbio.org/#domain_composition) Pfam domain models found in more than five chromatin proteins were selected. The conditional probability of finding a corresponding domain A in a chromatin protein given that another domain B is already present was estimated as $P(A|B) = P(A \cap B) / P(B)$.

3. RESULTS

3.1. Sources of information about chromatin and nuclear proteins and their critical evaluation

Given the ambiguity in exactly defining the criteria that delineate chromatin proteins from non-chromatin proteins, in analyzing and comparing the sources of information about chromatin proteins we chose to expand our analysis to include nuclear proteins. Nuclear localization data provides an independent way of validation since at least interphase chromatin proteins are also expected to be nuclear proteins. Hence, as the first step of our study we used database and literature mining to analyze the currently available sources of information about *human* chromatin/nuclear proteins and evaluated their congruency through cross-comparison.

The relevant sources of information were grouped into three classes: function-oriented protein/gene annotation databases, protein localization studies/databases, MS-based studies relying on chromatin extraction (**Fig. 2**). **Suppl. Table ST1** contains a representative list of 63 sources compiled for this study that include both historic and state-of-the-art datasets. It can be seen that although in the post-genomic era the size and completeness of the chromatin related proteomic datasets has grown considerably, there is still a lot of variation (*e.g.*, recent MS-based chromatin studies list from ~1000 to ~4000 chromatin proteins, the coverage of high-throughput localization studies ranges from ~1000 to ~12000 proteins), warranting datasets' validation, comparison and harmonization.

For our detailed analysis we chose (1) a number of general (GeneOntology) and specific (*e.g.*, EpiFactors [43], The Human Transcription Factors [55], Histone Database [56], *etc*.) protein/gene annotation databases or datasets that contain information about protein function (see **Fig. 2A**), (2) three major databases and/or high-throughput experimental studies that provide protein localization data (UniProt, HPA, OpenCell) (see **Fig. 2B**), and (3) eight state-of-the-art MS-based chromatin/nucleome studies (see **Fig. 2C**, **Table 1**). Below we proceed with presenting the results of our assessment of each class of information sources, followed by cross-comparing data between them.

A Function-oriented gene/protein annotation databases (DB)

The Gene Ontology resource

GENEONTOLOGY Unifying Biology

proteins: ~20 000

Classes of chromatin-related proteins (number of proteins):

Chromatin (CC) (2019) | Chromatin organization (825) | Chromatin remodeling (646) | Histone modifying activity (185) | Transcription by RNA polymerase II (445) | Nuclear DNA replication (27),
and others (more than 100)

Chromatin/epigenetic regulators

DBs: CR Cistrome, CR2Cancer, CRdb, FACER, EpiFactors

proteins: 400-850

Classes of chromatin-related proteins (number of proteins in EpiFactors):

Histones and protamines (65) | Histone PTM writers, readers, erasers (145,80,58) | DNA (and RNA) modification writers, readers, erasers | Chromatin remodeling (80) | Histone chaperones (23)

Protein class specific DBs

DBs: HistoneDB 2.0, SWI/SNF Infobase, The human transcription factors, TRRUST, HOCOMOCO, WERAM, HISTome2

proteins: 30-1500

Classes of chromatin-related proteins (number of proteins):

Histones (~60) | Chromatin remodelers (SWI/SNF family) (~30) | Histone PTM readers, writers, erasers (~100) | Transcription factors (~1500)

B Protein localization databases/studies

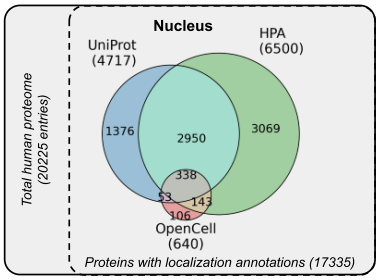
Main sources:

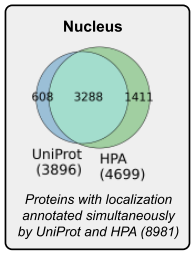
1. UniProt
2. THE HUMAN PROTEIN ATLAS
3. OpenCell

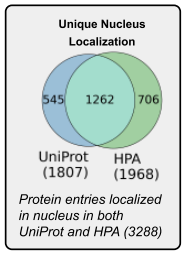
Common subnuclear localization terms:

Nucleoplasm; Nuclear bodies; Nucleolus; Nuclear envelope

Comparison of nucleus-localized protein sets according to UniProt, HPA, OpenCell

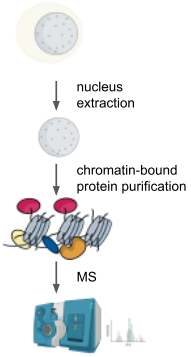




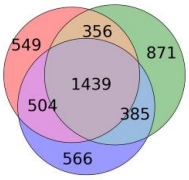


C Mass-spectrometry-based chromatomes

General scheme of experiments



Intersection of protein content from selected studies



Shi et al. (2021), 2848
Hi-MS: crosslinking + HaeIII digestion, biotin ligation)

Ginno et al. (2018), 3051
DEMAC: crosslinking + CsCl ultracentrifugation

Alvarez et al. (2023), 2894
NCC and **iPOND -** nascent chromatin

D Comparison of the list of chromatin/nuclear protein entries from different sources

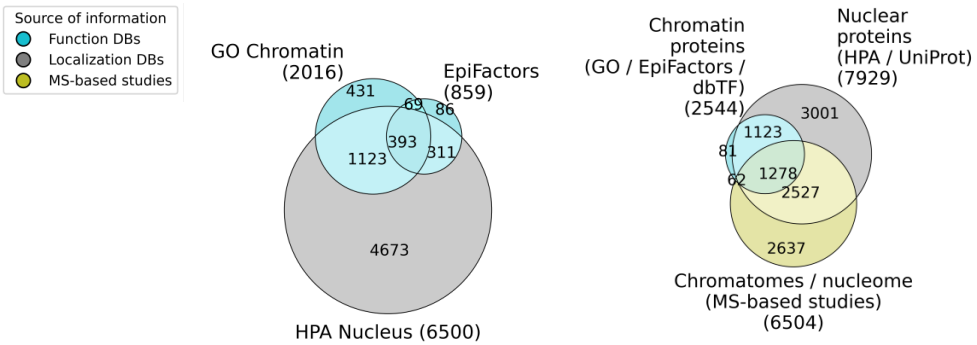


Figure 2. Source of information about chromatin proteins and their critical evaluation. (A) Function-oriented annotation databases include Gene Ontology, chromatin or epigenetic DBs (*e.g.*, EpiFactors), and protein class specific DBs (*e.g.*, HistoneDB, HISTome2). **(B)** Protein localization information was extracted from UniProt Knowledgebase (curated information about protein localization mainly from literature or computational analysis), the Human Protein Atlas (HPA, high-throughput immunohistochemical analysis of target proteins in fixed cells), and OpenCell (live-cell microscopy of proteins fused with fluorescent tags). The list of common terms describing subnuclear localization in the three databases is given. The Venn diagram **(top)** shows the intersection of nuclear proteins' lists from the three databases (with filtration by evidence codes/confidence levels, see Methods). The Venn diagrams at **the bottom** show the congruence among annotations in UniProt and HPA (see text). **(C)** The general scheme of mass-spectrometry-based studies for chromatin proteins' identification (left), the Venn diagram of chromatin proteins' lists from four different MS-based studies. **(D)** Comparison of chromatin/nuclear proteins' lists from different sources. **Left**: data from function-oriented DBs (proteins annotated by the term "Chromatin" in GO database and from EpiFactors database) versus localization databases (protein from the HPA database annotated by the term "Nucleus"), **right**: chromatin MS-based studies (proteins identified by three or more studies from the six analyzed chromatin MS-based studies, see Methods) versus data from function-oriented (proteins annotated by GO term "Chromatin") and localization databases (HPA term "Nucleoplasm").

Table 1. A representative list of chromatome and nucleome MS-based experimental studies (full description available in **Suppl. Table ST1**, datasets can be downloaded in **Interactive Table 1** at https://simchrom.intbio.org/#download).

Reference	Studied protein fractions	Chromatin purification methods and additional computation filtration	Type of cells / tissue	Proteins identified (by authors)	Number of proteins in our analysis
Torrente et al., 2011	Total chromatin, euchromatin, heterochromatin	(1) Total chromatin extraction with hypotonic lysis, Triton X-100 permeabilization, low-speed centrifugation, and EDTA-mediated nuclear lysis. (2) Salt extraction using high salt buffer (420 mM KCl), sonication, centrifugation, followed by dialysis. (3) Micrococcal nuclease (MNase) digestion, including total digestion and partial digestion to separate euchromatin and heterochromatin fractions.	HeLa S3	1038 (total chromatin extraction), 1388 (salt), 949 (MNase); 751 (partial MNase); 1912 (all identified chromatin proteins)	1501
Kustatscher et al., 2014	Total interphase chromatin	Chromatin Enrichment for Proteomics (**ChEP**): in vivo crosslinking with 1% formaldehyde, followed by differential extraction under denaturing conditions (SDS, urea), RNase A treatment, centrifugation-based chromatin pelleting, and sonication. To assess chromatin association, the study applied Multiclassifier Combinatorial Proteomics (MCCP), which integrates SILAC-based quantitative proteomics from 35 biochemical and biological perturbation experiments. A random forest machine learning algorithm was trained on curated chromatin and non-chromatin reference proteins to assign each detected protein an interphase chromatin probability score (ICP).	HeLa, MCF-7, HepG2, HEK293, U2OS, DT40	1980 (chromatin proteins with ICP>0.5); 7635 (total chromatin proteins with ICP values)	1956
Alabert et al., 2014	Nascent vs. mature chromatin	Nascent Chromatin Capture (**NCC**): biochemical isolation of newly replicated chromatin using biotin–dUTP incorporation. Cells were pulse-labeled with biotin–dUTP during DNA replication and fixed after either 20 min (nascent chromatin) or 2 h (mature chromatin). Chromatin was crosslinked with 2% formaldehyde, nuclei were isolated, and chromatin was sheared to 2–3 kb by sonication. Biotin-labeled DNA–protein complexes were isolated using streptavidin beads. For proteomic analysis, nascent and mature chromatin were metabolically labeled by SILAC and processed together.	HeLa S3	426 (nascent-enriched); 3995 (all identified chromatin proteins)	3861
Itzhak et al., 2016	Total nuclear proteome	Cells were metabolically labeled with SILAC and gently lysed under hypo-osmotic conditions to preserve organelle integrity. Post-nuclear supernatants were fractionated by differential centrifugation into five sub-organellar fractions plus cytosolic and nuclear pellets. Protein abundance profiles across SILAC fractions were processed by PCA and classified using a supervised SVM algorithm trained on curated organelle markers. In parallel, total intensities in the nuclear, cytosolic, and organellar fractions (from label-free MS) were used to assign each protein to global classes such as "mostly nuclear", based on relative signal distribution.	HeLa	1133 (nuclear); 672 (nucleo-cytosolic); 8710 (total)	1092

Ginno et al., 2018	Total chromatin: time-resolved (G1, S, M)	Density-based enrichment for mass spectrometry analysis of chromatin (**DEMAC**): formaldehyde-fixed cells were sonicated, subjected to cesium chloride (CsCl) gradient ultracentrifugation to isolate DNA–protein complexes by buoyant density (1.39 g/cm³). Chromatin fractions were collected, dialyzed, decrosslinked, digested with DNase I.	Human T98G (glioblastoma)	3065 (chromatome); 6242 (total proteome)	3051
Shi et al., 2021	Promoter-proximal chromatin	**Hi-MS** (Hi-C-based proteomics, adapted from BL-Hi-C): cells crosslinked with 1% formaldehyde; genomic DNA digested with HaeIII (GGCC sites); ligated with biotinylated bridge linkers; nuclei lysed in 0.2% SDS; chromatin sonicated; chromatin-DNA complexes captured on streptavidin beads. Quantified sensitivity to 1,6-hexanediol evaluated via AICAP index (Anti-1,6-Hexanediol Index of Chromatin-Associated Proteins).	K562	3228	2848
Ugur et al., 2023	Total chromatin	Chromatin Aggregation Capture (**ChAC**): nuclei fixed with 1% formaldehyde, lysed with SDS and urea, sonicated, and purified by protein aggregation capture (PAC) on magnetic beads. DIA-MS with DIA-NN used for quantification.	Human ESCs (H9)	2487	1730
Alvarez et al., 2023	Time-resolved (nascent, G2/M, early and late G1) chromatin	Nascent Chromatin Capture (**NCC**) method, which relies on pulse-labeling newly replicated DNA with biotin-dUTP, followed by formaldehyde crosslinking and sonication-based chromatin fragmentation. Biotinylated DNA-protein complexes were affinity-purified using streptavidin magnetic beads. HeLa S3 cells were synchronized and harvested at five post-replication time points (Nasc, Late S, G2/M, early G1, late G1) across six biological replicates.	HeLa S3	1454 (present at all time points in all 6 replicates; from total of 5770)	1478 (2894 total)
		Isolation of Proteins On Nascent DNA (iPOND): formaldehyde crosslinking (1%), EdU labeling for 15 minutes, click chemistry with biotin-azide, chromatin fragmentation, streptavidin bead enrichment.	TIG-3 fibroblasts	2351 (detected in ≥4 of 5 replicates)	2397 (2894 total)

Among the protein/gene annotation databases the GeneOntology (GO) database stands out as a comprehensive attempt in describing the functions of gene products in an ever growing number of organisms [35,36]. Within the GO framework genes are annotated according to their involvement in certain molecular functions, biological processes, and cellular components. The ontology itself forms a complex interlinked hierarchy with more than 40,000 GO terms and offers annotations to nearly the entire human reference proteome. However, despite its apparent comprehensiveness the GO database could not *per se* provide answers to the questions that were instrumental to this study, namely, to provide a set of chromatin genes/proteins and a relatively simple functional classification of these proteins that could be used for further analysis.

The GO cellular component term "chromatin" is defined broadly as "the ordered and organized complex of DNA, protein, and sometimes RNA, that forms the chromosome" and encompasses around *2000* proteins. Comparisons with other databases suggest that this number is a rather conservative estimate. For instance, up to 528 proteins listed in specialized databases of epigenetic factors (EpiFactors) and transcription factors (GO catalogue of TFs [44]) are missing from this set (see **Suppl. Fig. SF2_1A**); concomitantly the HPA protein localization database suggests that there are around *6000* proteins located in nucleoplasm (see **Suppl. Fig. SF2_2A**). Furthermore many proteins annotated by GO terms that are *bona fide* related to chromatin (*e.g.*, chromatin binding) are missing from those annotated by the GO term “chromatin” (see **Suppl. Fig. SF2_1C**). This stems in part from the complexity of the relations between the GO terms belonging to different annotation aspects, in this example the terms “chromatin organization” and "chromatin remodeling" are not connected to the term “chromatin” within the ontology tree. The manual search and identification of the chromatin related

GO terms and relevant proteins is challenging because (1) of the shear number of GO terms (*e.g.*, the word "chromatin" is found in the names of more than 60 terms, see **Suppl. Fig. SF2_1C**), (2) the fact that apparently relevant terms may include also non-chromatin associated entries (*e.g.*, transcription may also include mitochondrial transcription), (3) the fact that terms describing certain historically established chromatin protein groups may be missing (*e.g.*, histones, HMG proteins), (4) the fact that GO database is not chromatin-specific and may not be up-to-date in certain aspects (*e.g.*, contain obsolete terms such as "nuclear matrix" or lack annotations for proteins that are available in recent literature reviews). **Suppl. R&D Sec. 1.1** provides further details and examples from our analysis.

A number of epigenetic/chromatin regulators/factors databases (*e.g.,* EpiFactors [43], CRdb [57]) provide carefully curated information about chromatin proteins that are involved in what is historically assumed to be molecular mechanisms of epigenetic regulation **(Fig. 2A, Suppl. Table ST1)**. However, they annotate only around 400-800 chromatin proteins, which is much less than is expected to be in chromatin (see **Fig. 2D**). Unlike GO, these databases introduce a rather simple classification of proteins (the respective categories are highlighted in **Fig. 3**), but lack many essential chromatin categories (*e.g.*, histones, histone chaperones, *etc.*). Protein class specific databases and datasets available in published papers provided an even more trustworthy but narrow sets of information about certain classes of chromatin proteins. Of particular impact here by the number of provided entries are the databases of transcription factors. Recent databases (*e.g.*, The Human Transcription Factors [55], GO catalogue of TFs [44]) comprise around 1500 transcription factors. Additionally, several chromatin-related protein classes have been reviewed in the literature but lack dedicated database resources [58–64].

The other alternative and powerful source of information about nuclear/chromatin proteins is localization databases and proteome-wide studies – UniPot, HPA and OpenCell projects are currently regarded as the most comprehensive and trusted resources on protein intracellular localization (see **Fig. 2B**). From each resource we extracted the sets of proteins whose localization was annotated (only annotation with sufficient confidence levels was considered for our analysis - see **Methods Section 2.1.1**) as belonging to the nucleus or sub-nuclear compartments according to the localization ontologies specific to each resource. The detailed cross-comparison of the datasets is presented in **Suppl. Fig. SF2_2**, its interactive version (**Interactive Fig. 2** available at https://simchrom.intbio.org/#localization), and at length discussed in **Suppl. R&D Sec. 1.2** using **Suppl. Fig. SF2_2, SF2_3, SF2_4.** In summary there is a considerable degree of variation both between the sets of nuclear proteins, their sub-nuclear localization annotation and the annotation ontologies themselves between the resources.

It has to be kept in mind in the first place that localization annotation coverage is not complete – collectively the three resources cover 86% of the human reference proteome (*with sufficient*

confidence - see above), while only 44% of proteins are simultaneously annotated by UniProt and HPA. The resources also differ by the number of localization annotations they provide on average for each protein (median number is two and one, for HPA and UniProt, respectively), suggesting HPA is more complete with respect to annotating multilocalization of proteins. Hence, although the protein space coverage by UniProt is larger compared to HPA (70% vs 60%), the nucleome provided by the former is considerably smaller (~4700 vs 6500 proteins). Together the three resources annotate ~8000 proteins as having nuclear localization (see Venn diagrams in **Fig. 2B**), which *amounts to 47% of proteins* that have localization information according to at least one resource. *Hence, as a rough estimate it is tempting to conclude that current protein localization databases suggest that around half of human proteins have some evidence of nuclear localization.* However, it has to be kept in mind that the congruency between the resources remains mediocre. Among ~9000 proteins whose localization is simultaneously available in UniProt and HPA, ~3300 are annotated as nuclear by both HPA and UniProt, while another ~2000 are annotated as such only by one of the two resources (~600 by UniProt and ~1400 by HPA) (see **Fig. 2B**, lower left Venn diagram). The discrepancies are in part due to (1) incomplete annotation of multiple localization possibilities by the databases (among the ~2000 proteins, ~60% have a matching localization annotation between the databases other than nuclear), (2) potential biases in localization annotations (HPA tends to label nuclear proteins as vesicular proteins and UniProt tends to label nuclear proteins as secreted proteins and proteins of the extracellular matrix) (see **Suppl. R&D Sec. 1.2**, **Suppl. Fig. SF2_3**). Extrapolating the above estimates to the whole proteome (with all caveats in mind about the non-uniform annotation coverage of different protein groups) one can suggest that *between 35 to 60% gene protein products may be ascribed nuclear localization depending on the chosen degree of certainty*. Hence, a combination of the datasets provided by the resources may be used to construct reference nucleome datasets of varying confidence (see **Results Sec. 3.2**).

Many nuclear proteins were found to have multiple localization annotations belonging to different cellular compartments (see **Suppl. Fig. SF2_2B**, **Methods Section 2.1.1**, and **Suppl. R&D Sec. 1.2**). On one hand this reflects the functionally important property of nuclear proteins to shuttle between compartments. For example, many transcription factors and coactivators (*e.g.*, NF-κB, STAT, p53, TAF7, YAP/TAZ) regulate their action through cytoplasm/nucleus shuttling [65–67], even some histones, such as H2B, may relocalize to cytoplasm under stress and perform unconventional functions [68,69]. On the other hand, functionally irrelevant multiple localization annotations may arise due to experimental artefacts or suboptimal signal-to-noise thresholds, keeping in mind that all nuclear proteins are in fact synthesized in cytoplasm and imported to the nucleus. According to our analysis UniProt and HPA estimate separately that ~50% of proteins with nuclear localization may be also localized in other compartments (~40% in cytoplasm, 12%–22% in the endomembrane system), annotating around 48%-50% to be localized solely in the nucleus. However, once the annotations of UniProt and HPA are compared within the shared common set of proteins (nuclear localization

annotation is available in both databases) it turns out that only for ~40% of proteins the two databases reach consensus for their unique nuclear localization (see **Fig. 2B**, lower right Venn diagram). In other words (see **Suppl. R&D Sec. 1.2**), *approximately for every five proteins identified as uniquely localized in the nucleus by one database, it is likely that two of them will have non-nuclear localization annotation in the other database (per se or in addition to the nuclear localization)*. The same tendency was observed for the annotations of uniquely localized cytoplasmic proteins. The obtained estimates likely reflect the suboptimal specificity of the localization information provided by the databases (additional localizations of nuclear proteins may not be always captured) and potential presence of spurious localization annotations (artifacts or incorrect localization assignments). It is, however, non-trivial to deconvolute between these two types of errors.

Our analysis of sub-nuclear localization ontologies showed that the one of Uniprot is more diverse comprising 20 terms (this number includes chromosome localization which is distinct from the nuclear localization according to UniProt, although the majority – 84% – of chromosome proteins are also annotated as nuclear), while HPA and OpenCell comprize, 9 and 6 terms, respectively. However, in terms of annotation specificity only 19% of nuclear proteins in UniProt are annotated with sub-nuclear localization terms, while for HPA all of the nuclear proteins have some sub-nuclear localization (although 92% are considered a part of nucleoplasm, 33% bear localization annotations other than nucleoplasm). There are certain parts of the ontologies that do not match between the resources or to the state-of-the-art knowledge. For instance, HPA considers mitotic chromosomes as a part of nucleoplasm, while UniProt uses outdated "nucleus matrix" term. OpenCell is the only resource that explicitly considers "chromatin" as the possible localization for nuclear proteins, while UniProt explicitly lists "Chromosome" as the possible localization, which was in turn inherited from GO cellular compartment ontology where "chromatin" has a child-parent relation with the term "chromosome". The above-mentioned discrepancies reflect the dynamic complexity of cellular organization, our constantly evolving understanding of nuclear organization, and the resulting difficulty in describing subcellular localization in a form of a simple hierarchical tree-like ontology. While the exact names may differ, all resources converge on the presence of the following localization terms: Nucleoplasm; Nuclear bodies; Nucleolus; Nuclear envelope. Among these terms the nucleoplasm localization is the one most related to chromatin proteins (according to HPA nucleoplasm is what is found within the nuclear membrane, but excludes nucleoli according to the respective localization ontology). If one interprets the definition of chromatin broadly (treats proteins that localize with the interphase chromosomes to be part of the chromatin "complex") the set of proteins with nucleoplasm localization is a direct source of information about chromatin proteins. HPA lists around six thousand nucleoplasm proteins (**Fig. 2B**). The analysis of subnuclear multilocalization is available in **Suppl. R&D Sec. 1.2** and **Suppl. Fig. SF2_4**.

MS-based studies of chromatin extracts are another key source of information about the protein content of chromatin. Despite being the ultimate direct source of data about the composition of

chromatin it unfortunately has certain limitations (see **Introduction**, and relevant reviews [17,70]). To gain quantitative understanding into the utility of MS-based studies for our goals we have selected data from several studies in human cell lines (see **Table 1** and **Methods Section 2.1.3**) for analysis. The selected datasets included *five* studies that aimed at total interphase chromatin characterization using different methods of chromatin purification and post-MS data analysis, *two* studies characterizing nascent chromatin, and *one* study characterizing total nuclear proteome. The more than twofold variation (from 1.5 to 3.5 thousand entries) in the number of detected chromatin proteins in various MS-based studies highlights the varying sensitivities of different chromatin purification/MS-detection setups (**Table 1**). The pairwise comparison of different chromatin datasets of comparable size (having around 3000 proteins) suggests that for any given set its fraction overlapping with any other set does not exceed 68% (see **Fig. 2C**). The number of chromatin proteins present simultaneously in all total chromatin datasets is 179 (**Suppl. Fig. SF2_5B**). These facts highlight considerable variation of MS-based data due to different sample sources and chromatin extraction techniques.

We next thoroughly analyzed these protein datasets through cross-comparison between themselves, comparison with protein localization data, and tested enrichment of different chromatin protein categories (according to SimChrom classification described in **Results Section 3.2**). The detailed analysis is provided as **Suppl. R&D Sec. 1.3,** and we only succinctly summarize our conclusions below. From 10% to 38% of proteins identified in MS-based chromatin datasets currently do not have any support from localization databases through their annotated nuclear localization (see **Suppl. Fig. SF2_5A,C**), suggesting that even for chromatin purification protocols based on protein-DNA cross-linking there still might be a certain degree of contamination with non-nuclear proteins, mainly cytoplasmic ones (see **Suppl. Fig. SF2_6**). Yet, MS-based techniques may have predictive power to identify new chromatin proteins that are not annotated in the localization databases. For example, among 195 proteins reported simultaneously by at least five out of seven chromatin MS-based studies we estimated that around ~30% of proteins may have indications in the literature supporting their nuclear localization. MS-based studies are biased towards identifying the housekeeping proteins - more than 80% of nuclear/chromatin proteins reported by the MS-based studies were from the housekeeping pool, while the average expected fraction of nuclear housekeeping proteins is around 62% (see **Suppl. Fig. SF2_7A,B**). This is expected since many non-housekeeping proteins are conditionally expressed. However, MS-based studies tend to miss the housekeeping transcription factors too (and even to a greater extent non-housekeeping TF) apparently due to their low abundance and dynamic nature of interactions (**Suppl. Fig. SF2_7C**, **SF2_8A,** see also **Results Section 3.3.1** for discussion of chromatin protein abundance). MS-based studies also struggle to recover as separate gene products proteins with very similar sequences, *e.g.* canonical histone isoforms (see **Suppl. Fig. SF2_7C**, **SF2_8B**).

To finalize our analysis we compared the datasets from three types of data sources about chromatin proteins examined above (**Fig. 2D**, **Suppl. Fig. SF2_9**). One can see that localization databases are currently leading by the number of proteins that may be considered as chromatin proteins in the broad sense (*e.g.,* the proteins of the nucleoplasm). However, there is still limited congruence with the other data sources. For instance, 25% of GO "Chromatin" proteins are not localized in the nucleus according to HPA, moreover, of these 500 proteins, only 115 have any localization information in HPA. Notably, 42% (2699) of the proteins identified in MS-based chromatome and nucleome studies lack nuclear localization annotations in both UniProt and the HPA, whereas only 254 proteins remain entirely unannotated for subcellular localization in these databases.

Taken together our analysis of different chromatome data sources revealed considerable heterogeneity of information and limited congruence between the available datasets. The available functional databases while providing functionally supported data are either limited in scope or suffer from historically-contingent complexity and sometimes discrepancies in their classification ontologies that are not tailored to provide comprehensive straightforward information about interphase chromatin proteins. The localization databases is a powerful alternative source of information that can give an upper bound for the set of chromatin proteins (since they should have nuclear/nucleoplasm localization), provide a relatively reliable estimate of the lower bound for the number of nuclear proteins, however, they suffer from an incomplete coverage of the proteome-localization space and hence difficulties in estimating false-positive and false-negative annotation rates (keeping in mind the multi-localization of proteins) and limited congruence of subnuclear localization ontologies. The MS-based studies of chromatin extracts are the most direct source of information about chromatome, they may identify new chromatin proteins not annotated currently into the databases, however, they are limited in scope (many proteins are conditionally expressed or have low expression levels) and suffer from contamination with non-nuclear proteins.

3.2. The SimChrom chromatin protein classification, the SimChrom dataset and other reference datasets

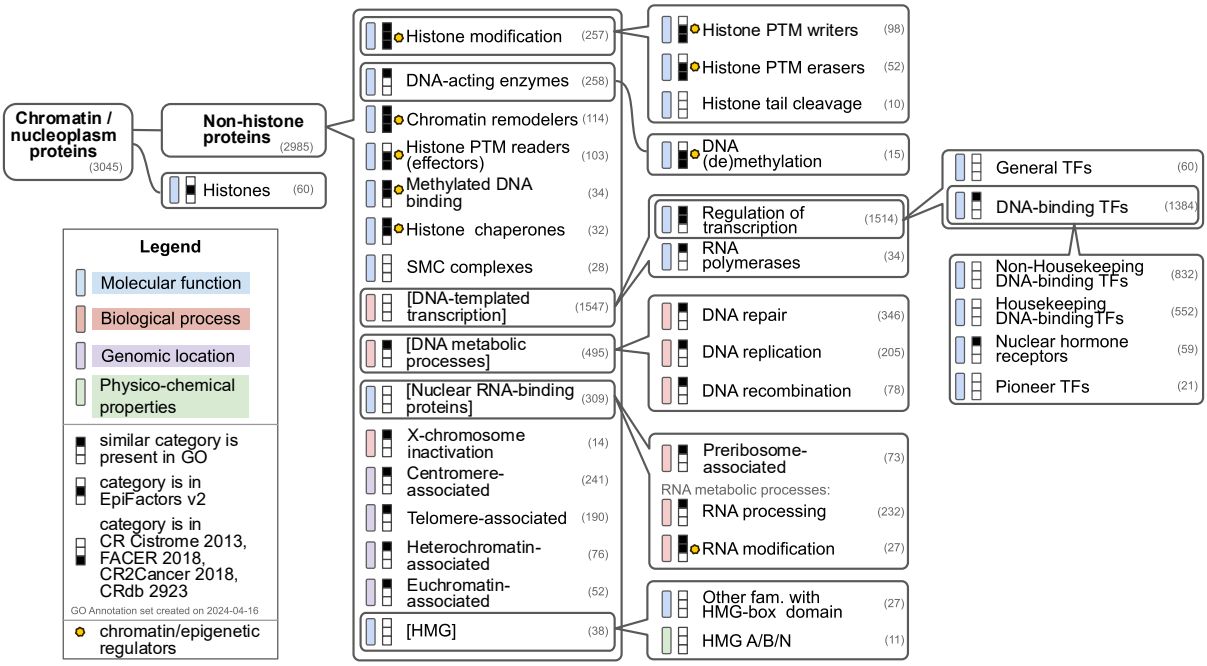


Figure 3. The SimChrom empirical chromatin classification ontology and the SimChrom chromatin proteins dataset. The hierarchical tree-like classification organizes 39 SimChrom categories. For each category the respective number of proteins from the SimChrom dataset is given in parentheses (proteins can simultaneously belong to more than one category with the exception of histones). The pictograms on the left of each category name provide the information about the presence of similar categories in the ontologies of other databases (see legend). The colored bars show whether the specific category was derived from grouping the proteins according to certain aspects: the similarity of their molecular functions, physico-chemical properties, involvement in similar biological processes or localization in similar genomic locations (see legend). Note: the latter annotations may deviate from GO annotation aspects.

Taking into account the advantages and disadvantage of different sources of information about chromatin proteins presented above, we aimed at constructing a reference set of chromatin proteins together with a classification ontology and several supplementary nuclear localization protein datasets that can be later used in analyzing the repertoire, abundance, functional, structural, and physico-chemical properties of chromatin proteins. Our aim was to create a relatively simple classification ontology that while potentially sacrificing the details will enable a holistic human-understandable overview of the chromatome (see **Suppl. R&D Section 1.1** for the discussion of GO complexity and ensuing challenges). The current version of SimChrom classification focuses on classification of chromatin/nucleoplasm proteins leaving aside the classification of the nuclear envelope proteins, which are historically not considered to be a part of chromatin. The SimChrom classification ontology was created by manually analyzing, critically evaluating, selecting and combining into a tree-like classification scheme information from (1) the historically established consensus on chromatin proteins classification (*e.g.*, histone, non-histone proteins, HMG-proteins [1]), (2) classification used in major

databases of chromatin and epigenetic regulators (*e.g.*, EpiFactors, FACER), (3) classification used in the three aspects of Gene Ontology, (**Suppl. Fig. SF3_1**). Our hierarchical SimChrom classification is presented in **Fig. 3**. The majority of classification terms used in SimChrom was inspired by GO-based classification, yet only a small subset of terms was used. The main focus of the classification was to classify chromatin proteins according to their functions and biological processes that they are involved in, but genomic-location (which is also indirectly related to function) and physical properties (*e.g.*, high-mobility group proteins of A, B and N families) were also considered (**Fig. 3** highlights what classificational aspects are most relevant for each term using a color bar). The SimChrom ontology was developed simultaneously with the SimChrom dataset in an iterative manner by obtaining sets of proteins annotated by various GO terms, extracting them from literature and domain specific databases, manually curating, validating and filtering (see **Methods Section 2.2**). Only major splice isoforms of genes are included in SimChrom. The resulting SimChrom dataset contains 3045 proteins, is available as a **Suppl. Table ST5** and viewable in the **Interactive Fig. 3** at the SimChrom web-site (https://simchrom.intbio.org#classification). The descriptive details about the SimChrom dataset are available in **Suppl. R&D Sec. 2**.

In the default SimChrom classification (depicted in **Fig. 3**) every protein from the SimChrom dataset may belong to more than one SimChrom ontology category. This provides the needed degree of flexibility since many proteins indeed may *bona fide* belong to several categories due to their complex functional, physico-chemical or structural properties. However, in certain cases of holistic analysis an even simpler classification may be useful, which ascribes every protein to only one category. Such single label classification (**SimChrom-SL**) based on the same SimChrom ontology was also developed (see **Methods Section 2.2**, **Suppl. Fig. SF3_2**). Briefly, if the protein belonged to several categories by default it was ascribed to the category with the least number of other proteins (*i.e.* the most specific category for this protein) with functional categories taking priority (see **Suppl. Fig. SF3_2** for category priority order).

As auxiliary datasets based on the results of **Section 3.1** we have compiled several reference datasets of nuclear and non-nuclear proteins at different levels of support (depending on whether nuclear localization is supported by one or several localization databases), confidence (depending on the evidence codes and reliability scores provided by the databases), and also whether proteins are uniquely localized in the nucleus or have multiple localization in the nucleus and other cellular compartments (see **Methods Section 2.2.2**). The list of the datasets and their definition is presented in **Suppl. Table ST6**, the datasets are available for download in the **Interactive Table 2** at https://simchrom.intbio.org/#download. Instrumental to our further analysis will be the "nuclear localization consensus" (NULOC_CS) dataset – the set of nuclear proteins, whose nuclear localization is supported (with sufficiently good confidence levels) both by UniProt and HPA and does not contradict the data from OpenCell, and the "nuclear localization joint dataset with no evidence code

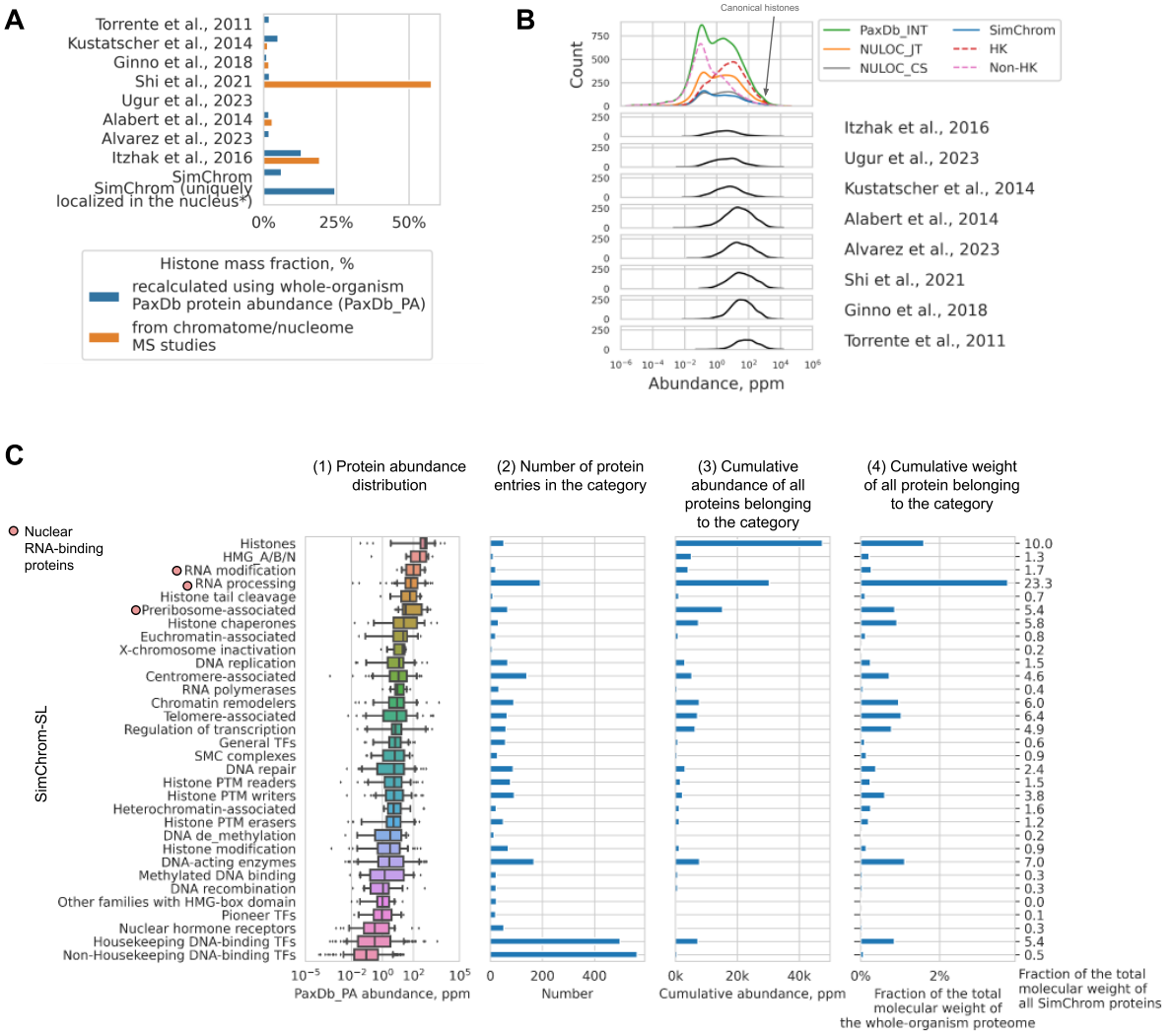
filtering" (NULOC_JT_NECF) dataset - the maximally broad set of nuclear proteins, which includes proteins whose nuclear localization is supported by any of the localization databases at any levels of confidence. The NULOC_CS dataset contains 3296 entries, while NULOC_JT_NECF contains 8912 entries.

To evaluate the contents of our SimChrom dataset we performed its cross-comparison to the localization based datasets described above (NULOC_CS and NULOC_JT_NECF) (see **Suppl. Fig. SF3_3**). Detailed discussion of the results is provided in **Suppl. R&D Sec. 2**. Briefly, almost all SimChrom proteins had some evidence of nuclear localization (95% were present in NULOC_JT_NECF dataset, 60% in NULOC_CS dataset, see **Suppl. Fig. SF3_3**). For the SimChrom proteins that did not have high confidence support of nuclear localization (non present in NULOC_CS) GO enrichment analysis of SimChrom-exclusive proteins revealed minimal association with non-nuclear functions, with only a minor subset (~10 centromere-associated proteins) linked to such categories (**Suppl. Fig. SF3_4**, **Suppl. Table ST8**). Moreover, no additional chromatin-related GO categories were found to be underrepresented in SimChrom, indicating its broad coverage of chromatin-associated functions (**Suppl. Table ST9**). 60% of SimChrom proteins were successfully identified in MS-based chromatomes and nucleomes (see **Suppl. R&D Sec. 1.3**, **Suppl. Fig. SF2_7**). The remaining 40%, predominantly low-abundant transcription factors, were likely undetected due to their transient nature and dynamic interaction properties, which pose challenges for MS-based detection. Furthermore GO enrichment analysis of MS-derived proteins absent in SimChrom or nuclear reference sets did not reveal a *bona fide* chromatin-associated category (see **Suppl. Table ST7**). Together, these results support the quality of the SimChrom dataset, suggesting that SimChrom is sufficiently comprehensive in its coverage of chromatin-related proteins and its categories.

3.3. Analysis of the human chromatome

Equipped with the datasets described above we aimed at a comprehensive characterization of the chromatome, including characterization of it composition (numbers of proteins belonging to different chromatin categories), abundance (the number of individual protein molecules present in the cells), physico-chemical properties of the amino acid sequences of the proteins, their domain architectures and interaction patterns (including engagement in multivalent interactions). The full discussion of the results is presented in **Suppl. R&D Sec. 3** and **Suppl. Fig. SF4_1 - SF 4_4, SF5_1 - SF5_5, SF6_1 - SF6_3, SF8_1, SF8_2.** The sections below summarize our analysis.

3.3.1. The chromatome composition and abundance of chromatin proteins



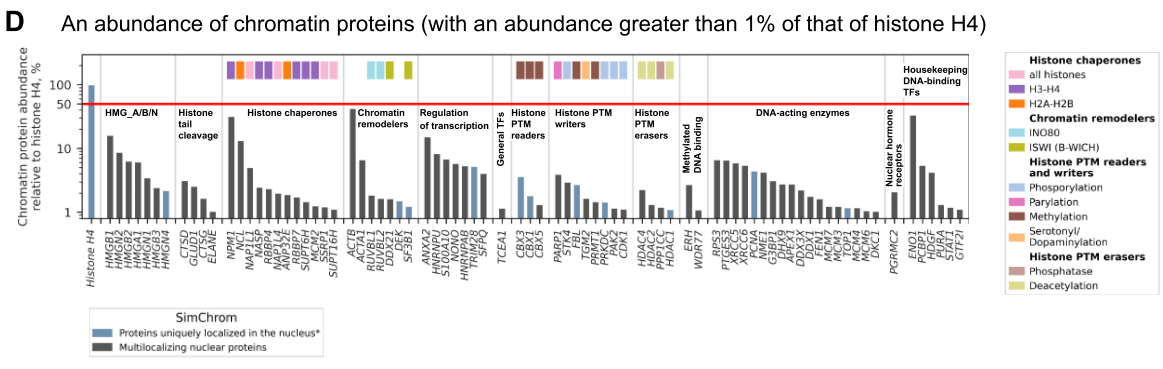


Figure 4. Analysis of the chromatome composition and abundance of chromatin proteins. (A) Fraction of histone proteins by mass in different datasets (according to experimental MS-based intensities and recalculated using abundance values from PaxDb_PA). **(B)** Distribution of proteins from different datasets according to their relative abundance values from PaxDb_INT. The distribution was constructed by taking the logarithm of the abundance values, making a histogram (bin size of 0.15) and smoothing it with a gaussian kernel for visual clarity. **(C)** Characterization of protein abundance (according to PaxDb_PA) for each SimChrom category and its contribution to the total abundance and mass of proteins. Each protein appears only in one category - SimChrom-SL classification is used. Subpanels show: 1 - distribution of chromatin proteins according to their relative abundance, box plots represent 25th and 75th percentiles, whiskers represent 5-95 percentiles, 2 - number of protein entries in each SimChrom-SL category, 3 - cumulative relative abundance of all proteins in the category, 4 - cumulative weight of all proteins belonging to the category relative to the whole human proteome and to the weight of all SimChrom proteins (numbers on the right side of the plot). The SimChrom categories are sorted according to the decrease of the median value of protein abundance. **(D)** The most abundant non-histone chromatin proteins. The barplot represents abundance (according to PaxDb_PA) for the most abundant non-histone chromatin proteins (with relative abundance of more than 1% of the histone H4 abundance). Proteins are referred to by the names of their genes grouped by SimChrom categories. The upper panel of colored rectangles provides additional functional classification (see legend). The red line at 50% abundance indicates the presumed abundance of nucleosomes (each containing two copies of the H4 histone).

To understand chromatin functioning it is important to know the chromatome content not only in terms of the set of proteins associated with chromatin, but also in terms of their abundance (*i.e.*, the (relative) number of proteins per cell or organelle). The analysis of MS protein intensities from the experimental chromatome/nucleome studies discussed above, revealed a high degree of variability (see **Fig. 4A**, **Supp. Fig. SF4_1**). For instance, the estimated relative mass of histone proteins varied from 0.1% to 58% depending on the study, suggesting a high degree of bias due to different experimental techniques and analysis pipelines used to process raw mass spectrometry data (see **Fig. 4A**). Hence, for further analysis we relied on the "whole-organism" protein abundance information available in PaxDb for *H. sapiens*, which for every protein reports its relative abundance in ppm (parts-per-million) [71]. See **Methods Section 2.1.3** and **Suppl. R&D Sec. 3.1** for the discussion of the quality and applicability of the data.

The overall distribution of the human proteome with respect to protein abundance is bimodal (**Fig. 4B**), with roughly half of the human proteins having low abundance (LA) values below 1 ppm. The corresponding distributions of chromatin and nuclear proteins follow the same distribution (the fraction of LA proteins is 48% in SimChrom and 40-45% for NULOC_CS/JT). On the contrary, the sets of chromatin/nuclear proteins extracted from the MS-based studies demonstrate abundance distributions that are mainly unimodal and centered around high abundant proteins (**Fig. 4B**, **Suppl. Fig. SF4_1C**). The fraction of LA proteins identified in MS-based studies ranges from 1 to 27%, suggesting that while there is an overall bias towards highly abundant proteins, there is certain variation due to the methods of protein extraction and analysis (see **Suppl. R&D Sec. 3.1**, **Suppl. Table ST10**).

Our analysis of the number of house-keeping (HK) and non-house-keeping (non-HK) genes (see **Methods Section 2.1.4**) showed that the set of chromatin proteins is enriched in HK-gene products

(the proteome wide fraction of HK genes is 40%, while in chromatin the fraction is 60%). It has to be kept in mind that around half (45%) of proteins in SimChrom are transcription factors, among which the non-HK TF constitute 60%. Hence, the non-TF fraction of SimChrom is even more enriched in HK-genes (74%). With respect to all-proteome distribution the increase in HK-proteins in chromatin is both due to the increase in LA and HA proteins (see **Suppl. Table ST10**). The bias towards HA proteins in MS-based studies of nuclear/chromatin proteins may be traced both due to the increased proportion of HK-proteins (which in turn have more HA representatives) and the difficulties in detecting the LA proteins among both HK and non-HK-proteins. The studies whose datasets are highly biased towards HA proteins mainly suffer from the latter problem (see **Suppl. R&D Sec. 3.1**).

We next aimed at understanding the abundance of different chromatin protein groups and individual chromatin proteins in the cell relying on our SimChrom-SL classification using PaxDb abundance data. The resulting diagrams depicting abundance variations of chromatin proteins, belonging to different SimChrom-SL categories, the number of proteins belonging to the respective categories, and the cumulative abundances (calculated both as the total number of protein molecules and the total molecular weight of protein molecules belonging to each SimChrom-SL category) are presented in **Fig. 4C**. To gain additional insights into the functioning of chromatin in **Fig. 4D** we plotted the abundance values of highly expressed chromatin proteins (abundance of more than 1% of the H4 histone abundance) belonging to SimChrom-SL categories of the "Molecular function" or "Physico-chemical properties" type. It is important to note that many chromatin proteins have additional localization in other cellular compartments, hence the presented data reflects the overall abundance of the chromatin proteins in the cell rather than their abundance in the nucleus (see **Suppl. Fig. SF4_3A** for an analogous analysis of chromatin proteins with unique nuclear localization).

As seen in subpanel 1 of **Fig. 4C** chromatin categories vary substantially by their median abundance from 0.09 ppm to 570 ppm and there is still considerable variation in the abundance values within the categories. The most abundant chromatin protein is histone H4 (~11000 ppm), and it is convenient to measure the abundance of all other proteins in fractions of its abundance. For detailed analysis of histone protein abundance see **Suppl. Fig. SF4_2A,C** and **Suppl. R&D Sec. 3.1.** Briefly, the most abundant histone variants are H3.3, H2A.X, H2A.Z, the least abundant are H2A.B and H1.7 (see **Suppl. Fig. SF4_2C**). Despite the relatively small number of protein coding human histone genes (108), many of which code for identical sequences, the cumulative abundance of histone proteins exceeds that of all other chromatin protein categories even if proteins with multiple localization are taken into account (see panel 2,3 in **Fig. 4C**). However, when the total molecular weight of proteins belonging to different categories is compared, the relatively small size of histone proteins (median ~15 kDa) results in them yielding the first place to RNA modification proteins (see panel 4, **Fig. 4C**). Collectively the cumulative weight of proteins belonging to "Nuclear RNA binding proteins" category

(that combines Preribosome-associated, RNA modification, and RNA processing categories) amounts to 30.4% of all SimChrom proteins weight (4.8% of whole-organism proteome weight). However, many proteins from these categories are also localized in cytoplasm, and the major contribution to their cumulative molecular weight likely comes from the cytoplasmic fraction. Among the proteins uniquely localized in the nucleus (**Suppl. Fig. SF4_3**), the mass fraction of histones is in the first place (38%), however, the mass fraction of RNA processing proteins (around 70 uniquely localized proteins, including splicing factors, pre-mRNA binding proteins, etc.) remains high (27%) it part due to higher average molecular weight of proteins in this group (see **Suppl. Table ST11** for a list of these proteins).

Other functional chromatin protein groups (or groups with specific properties) with high values of median abundance include HMG A/B/N, histone tail cleavage, histone chaperones, RNA polymerases, Chromatin remodelers and other categories (see **Fig. 4C**). The high mobility group proteins (HMG A/B/N) are the second group after histones ranked by their median abundance, the estimated ratio of HMG proteins to nucleosomes is 1:8, 1:2, 1:3 for proteins of main HMG superfamilies, HMGA, HMGB, or HMGN proteins, respectively. Among histone chaperones the H3-H4 histone chaperone NPM1 and H2A-H2B histone chaperone NCL have the highest abundance, 32% and 13% of H4 abundance, respectively.

The groups with least median abundance are those related to "DNA-binding transcription factors" (pioneer TFs, nuclear hormone receptors, housekeeping and non-housekeeping TFs). Non-housekeeping DNA binding transcription factors have an abundance between 0.00008 and 23.9 ppm, suggesting their expression at minimal levels when averaged across all body tissues. The majority of housekeeping transcription factors is also expressed only marginally (median abundance 0.314 ppm). However, the abundance of certain proteins classified as housekeeping TF may reach 91.8 ppm (RURB1 gene, 0.84% of H4) for TF uniquely localized in the nucleus or 3667 ppm for proteins that have multiple localization (ENO1 gene, 33.6% of H4). *The DNA-binding transcription factor groups have the largest number of genes in SimChrom-SL (1385 genes in total), however, their contribution to the cumulative weight of chromatin proteins is rather small (6.3%)*. See **Suppl. R&D Sec. 3.1** for a full discussion of the abundance of other protein groups in SimChrom.

3.3.2. Physico-chemical properties and amino acid composition

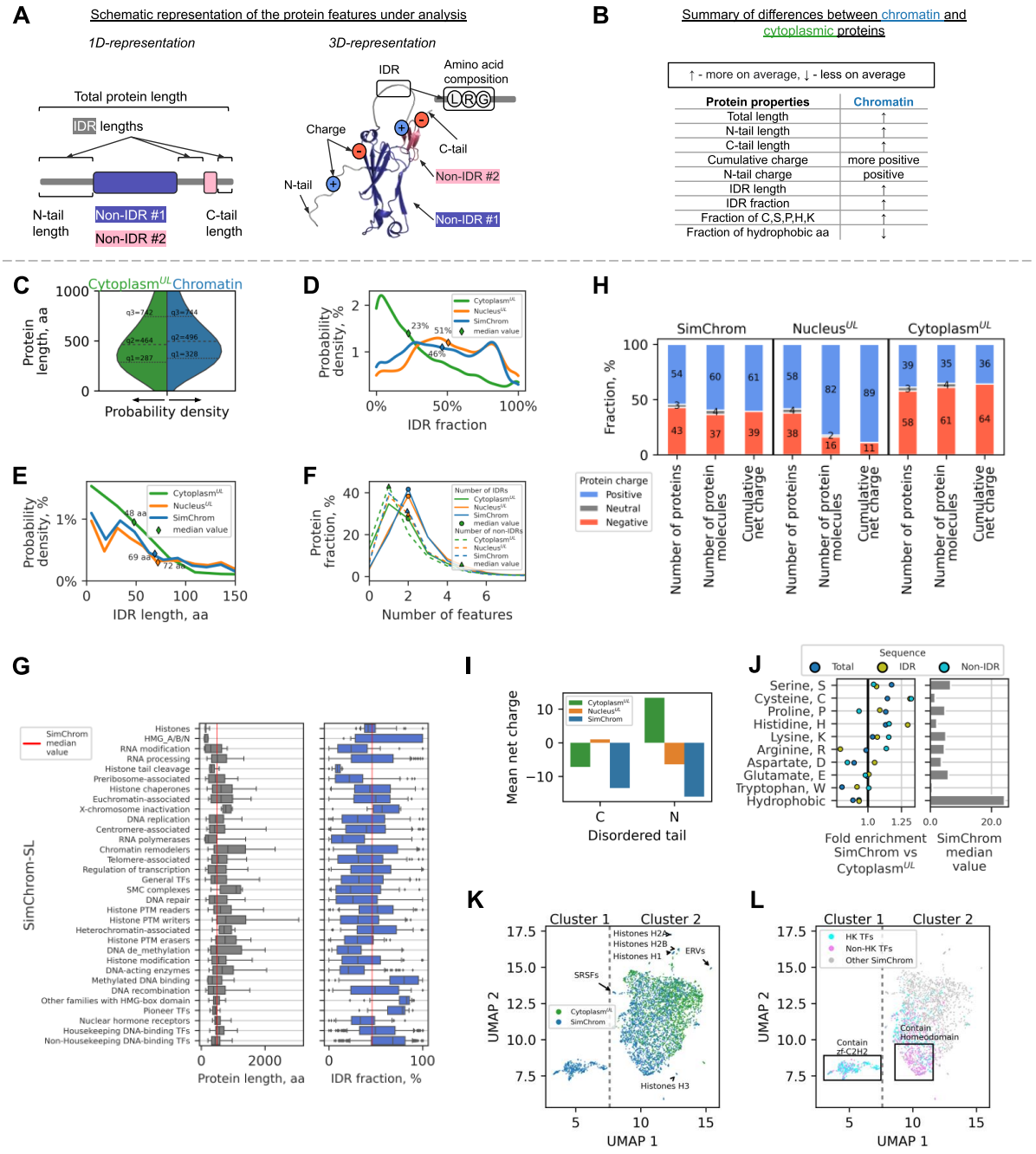


Figure 5. Physico-chemical properties and amino acid composition of chromatin proteins. *(A)* Schematic representation of the analyzed protein features/properties and (**B**) summary of the differences in properties of chromatin versus cytoplasmic proteins. Following datasets and designations are used in this figure: SimChrom - chromatin proteins, CYLOC_CS_UL (Cytoplasm^{UL} - uniquely localized cytoplasmic proteins), NULOC_CS_UL (Nucleus^{UL} - uniquely localized nuclear proteins). (**C-E**) Comparative distributions of chromatin, cytoplasmic and nuclear proteins with respect to their (**C**) protein length, **(D)** IDR fraction, **(E)** IDR length, (**F**) number of IDRs/non-IDRs. (**G**) The distribution of proteins by length and IDR fraction for individual chromatin protein categories (based on SimChrom-SL classification). Box plots represent 25th and 75th percentiles, whiskers represent 5-95 percentiles. (**H**) Ratio of the number of: (1) negatively and positively charged protein entries in the datasets (SimChrom, uniquely localized nuclear and cytoplasmic proteins), (2) individual protein molecules of the respective charge (estimated using PaxDb abundance data), and (3) ratio of the total net charges contributed by these molecules (calculated separately for all positively and all negatively charged protein molecules). (**I**) The average net charge of C- and N- disordered protein tails for different protein groups (see **Methods Section 2.4**). (**J**) Difference in amino acid composition of proteins (including subdivision into IDR and non-IDR regions) for chromatin proteins in comparison with cytoplasmic ones. The left panel represents statistically significant fold enrichment of median values (see Methods and **Suppl. Fig. SF5_4**). The right panel represents the median values of amino acid or group of amino acids fraction in chromatin proteins. (**K, L)** Grouping of chromatin and cytoplasmic proteins by their amino acids composition using UMAP dimensionality reduction technique. Different groups and individual proteins that stand out in the UMAP space are highlighted in panels **K** and **L** (mainly transcription factors). The dashed line denotes a visible border between clusters 1 and 2: cluster 1 mostly contains TFs with zf-C2H2 domains, whereas cluster 2 contains other chromatin proteins, including TFs with Homeodomain (see text for details).

The physico-chemical properties of chromatin proteins encoded in their primary sequence were analyzed in comparison with those of the uniquely localized nuclear and cytoplasmic ones (SimChrom, NULOC_CS_UL, and CYTLOC_CS_UL datasets were used), see **Fig. 5C-I** and **Suppl. Fig. SF5_1 - SF5_5**. **Fig. 5A,B** summarizes the properties that were under study and the main differences of chromatin proteins from cytoplasmic ones.

First we analyzed the total protein length, and presence of intrinsically disordered (IDR) and non-IDR regions (**Fig. 5C-G**, **Suppl. Fig. S5_1**). The average chromatin protein is somewhat longer than a cytoplasmic one (median length value increases from 460 to 490 amino acids (aa)), however, there is considerable variation among the SimChrom categories (**Fig. 5G**). The smallest proteins are histones and HMG group proteins (median values of 130 and 109 aa), while the SMC complexes and chromatin remodelers are on average much longer (median values of 1096 and 835). The longest chromatin proteins reach the length of five thousand aa (*e.g.*, preribosome associated protein MDN1 - 5596 aa, histone methyltransferase KMT2D - 5537 aa). It has to be kept in mind that the largest contributors to all average characteristics of chromatin proteins are transcription factors (TF) because of the number of proteins in these categories. Interestingly, the non-HK TF are on average shorter (median 432 aa) than cytoplasmic proteins, while HK TF are longer (median 563 aa). Chromatin and nuclear proteins differ drastically from cytoplasmic proteins in the fraction of their length occupied by IDR regions (~50% vs ~25%). This is both due to the increase in the number of IDRs (median is two vs one) and the length of individual IDRs (~70 vs ~50 aa). Interestingly, non-HK TFs have an especially

high fraction of IDRs (median ~68%, and contribute a pronounced peak in the distribution at around 80%, **Fig. 5D**, **Suppl. Fig. S5_1B**), some other groups are also especially rich in IDRs (HMG, pioneer TFs, Methylated DNA binding proteins). More details can be found in **Suppl. R&D Sec. 3.2**.

It is generally assumed that chromatin proteins are on average positively charged to compensate for the negative charge of the genomic DNA. In our datasets the entries related to positively charged proteins dominate among nuclear and chromatin proteins, while the ones related to negatively charged proteins dominate among the cytoplasmic ones (see **Fig. 5H**). The dominant charge group in each case consists of around more than half of protein entries (54-58%), while the opposite one consists of 38-43% of protein entries. This metric, however, does not account for protein abundance or total positive and negative charge conferred by the proteins. The data adjusted for protein abundance (see **Methods Section 2.4** and **Fig. 5H**) suggests that among the proteins that are uniquely localized in the nucleus, positively charged protein molecules dominate (82% vs 16%). The total amount of net positive charge contributed by such proteins exceeds the net negative charge contributed by the negative ones (the ratio of the values is 89:11). Among different chromatin protein groups there are those that are significantly enriched in negatively charged proteins (see **Suppl. Fig. S5_1E**). Among the categories having the most number of negatively charged proteins are Histone chaperones (84%), RNA polymerases (76%), Histone PTM erasers (71%). These categories, however, have many proteins that are localized both in the nucleus and the cytoplasm. This surprising presence of many negatively charged proteins in these categories is likely explained by their preferential association not with DNA, but rather with positively charged histones.

To further elucidate the peculiarities of charge structure in chromatin proteins we analyzed the length and average charge profiles of protein N- and C-terminal tails (**Suppl. Fig. SF5_2**). For chromatin proteins both N- and C-terminal tails are negatively charged, while for the proteins uniquely localized in the nucleus N-tail is negatively charged (see **Fig. 5I**). Interestingly there is a clear difference with the cytoplasmic proteins, whose N-tails are positively charged. The average charge of the N-terminal tails was -16 for SimChrom proteins, -6 for nuclear proteins, and +13 for cytoplasmic ones. The analysis of the N- and C-terminal tail charge for different SimChrom categories revealed that they varied between the different categories (**Suppl. Fig. S5_2E,F**). Histones had the most positively charged tails, while histone chaperones and HMGs had the most negatively charged ones. Transcription factors on average also had negatively charged protein tails. This is an interesting fact because most transcription factors are positively charged (**Suppl. Fig. S5_1E**).

We next set to analyze in detail the amino acid composition of chromatin/nuclear proteins with respect to cytoplasmic ones and the variability of amino acids composition between different groups of chromatin proteins (see **Fig. 5J,K,L**, **Suppl. Fig. SF5_4**, **SF5_5**, **Suppl. Table ST13**). To this end we first used the UMAP nonlinear dimensionality reduction technique to see if significant variations between chromatin proteins can be identified in the space of their amino acids composition. The

resulting 2D projections onto the main UMAP components revealed that (1) chromatin and cytoplasmic proteins occupied overlapping domains on the 2D map, but with a visible shift between their centers, suggesting there is an overall difference in the average amino acid composition, (2) certain chromatin protein groups formed dedicated clusters on the map, suggesting significant distinctness in their composition (see cluster 1, 2 and outliers shown by the arrows in **Fig. 5K,L**). Further analysis revealed that in the 2D UMAP map transcription factors, containing zinc finger domains and homedomains formed distinct clusters (see **Suppl. Fig. SF5_3A,B**). The most distinct group (cluster 1) was **almost exclusively** (415 out of 422) composed of zinc-finger containing DNA-binding transcription factors (240 housekeeping and 175 non-housekeeping) with the median number of zinc-finger domains (ZFD) of around 10 (**Suppl. Fig. SF5_3C**). Zinc-finger containing DNA-binding transcription factors were also present in cluster 2, but the median number of zinc-finger domains (ZFD) in that cluster was only three, hence containing a lower proportion of amino acids specific to ZFD (**Suppl. Fig. SF5_3D**). ZFD are enriched in histidine and cytosine (see **Fig. 5J** and discussion below). Other protein groups that occupied distinct positions on the UMAP map, included (1) histones, (2) serine/arginine-rich splicing factors (enriched in serine and arginine), and (3) reverse transcriptases of endogenous retroviruses (enriched in isoleucine and threonine) (see **Fig. 5K**, **Suppl. Fig. SF5_3D**).

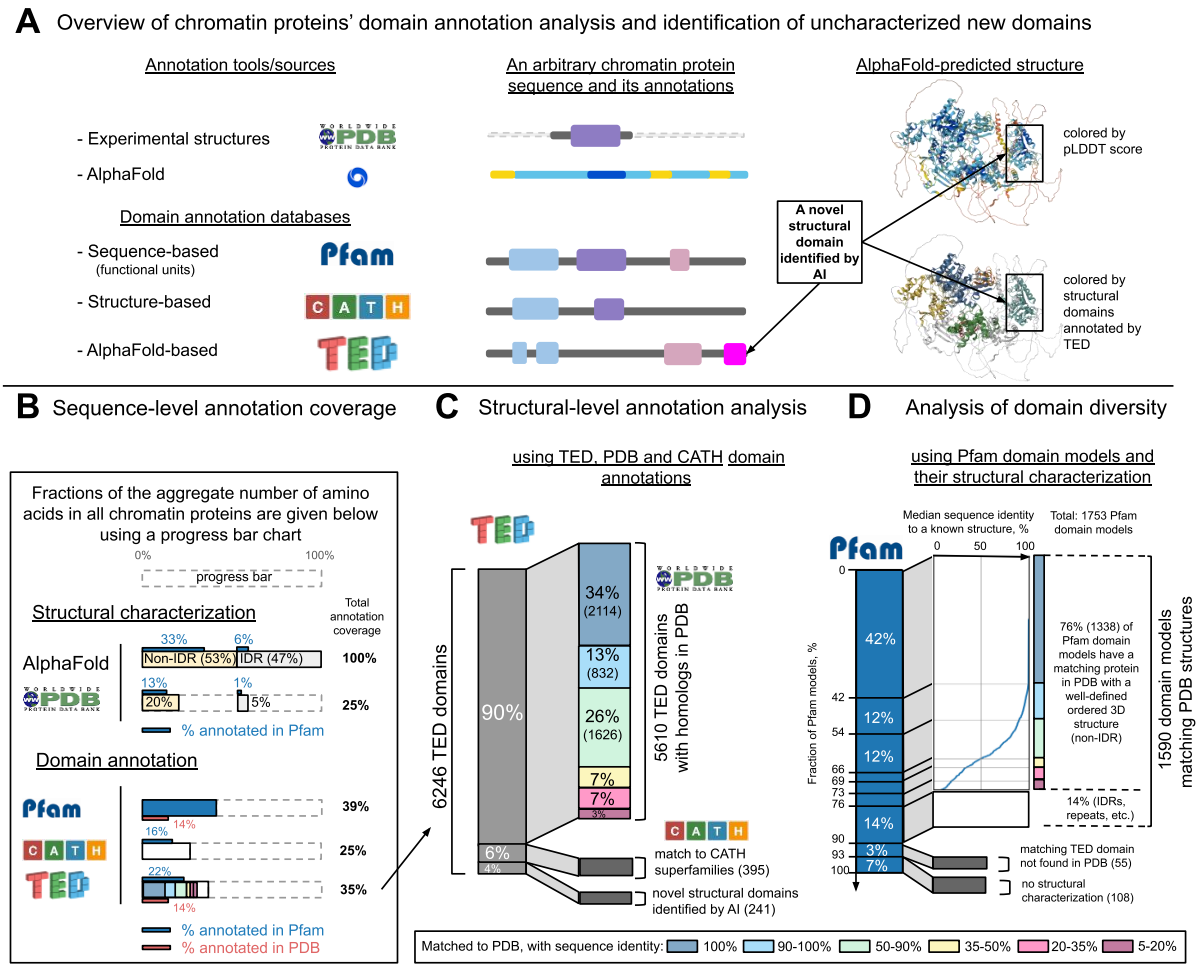
The detailed analysis of amino acids composition of different chromatin protein groups is presented in **Suppl. R&D Sec. 3.2**. Briefly, among the top four enriched amino acids in chromatin proteins are serine, cysteine, proline, and histidine (**Fig. 5J**). The enrichment of cysteine and histidine is solely contributed by the ZFD of transcription factors (**Suppl. Table ST13**, **Suppl. Fig. SF5_4B**, **Suppl. Fig. SF5_5A**). The total enrichment of serine and proline in chromatin proteins is attributed due to their enrichment in the non-IDR regions (relative to IDR and non-IDR regions of cytoplasmic proteins), and more importantly due to the higher proportion of IDR regions in chromatin proteins (46% vs 23%) that in turn have a considerably higher proportion of these amino acids than non-IDRs (**Suppl. Table ST13**). Serine was also enriched in non-IDR regions globally, while the enrichment of proline in non-IDRs was observed only in a few categories (*e.g.*, HMG-proteins) (**Suppl. Fig. SF5_4H**).

The enrichment of positively charged amino acids is only statistically significant for lysine, but not for arginine, and the enrichment is relatively moderate (1.03 in chromatin) (**Suppl. Fig. SF5_4A**). Arginine is highly enriched in non-IDRs, but it is the most depleted amino acid in IDRs of chromatin proteins *versus* the respective regions of the cytoplasmic ones. The depletion of negatively charged amino acids in chromatin/nuclear proteins is statistically significant for aspartate (fold enrichment is around 0.9), while the depletion of glutamate is statistically non-significant. Interestingly, aspartate is enriched in IDRs and significantly depleted in non-IDRs. This suggests that the increased positive charge of chromatin/nuclear proteins has its main contributions in the depletion of aspartate and enrichment of arginine in non-IDRs, and moderate global enrichment of lysine.

Among the most relatively depleted amino acids in chromatin/nucleus are hydrophobic aliphatic amino acids, they are relatively rare in IDRs and hence the large proportion of IDRs in

chromatin proteins accounts for their lower total fraction (**Suppl. Fig. SF5_4F,I**). Tryptophan, which is the rarest amino acid (~1% in proteins), is the most depleted amino acid on average in chromatin/nuclear proteins and almost in all chromatin categories, except for a few.

3.3.3. Domain composition of chromatin proteins and identification of new structural domains



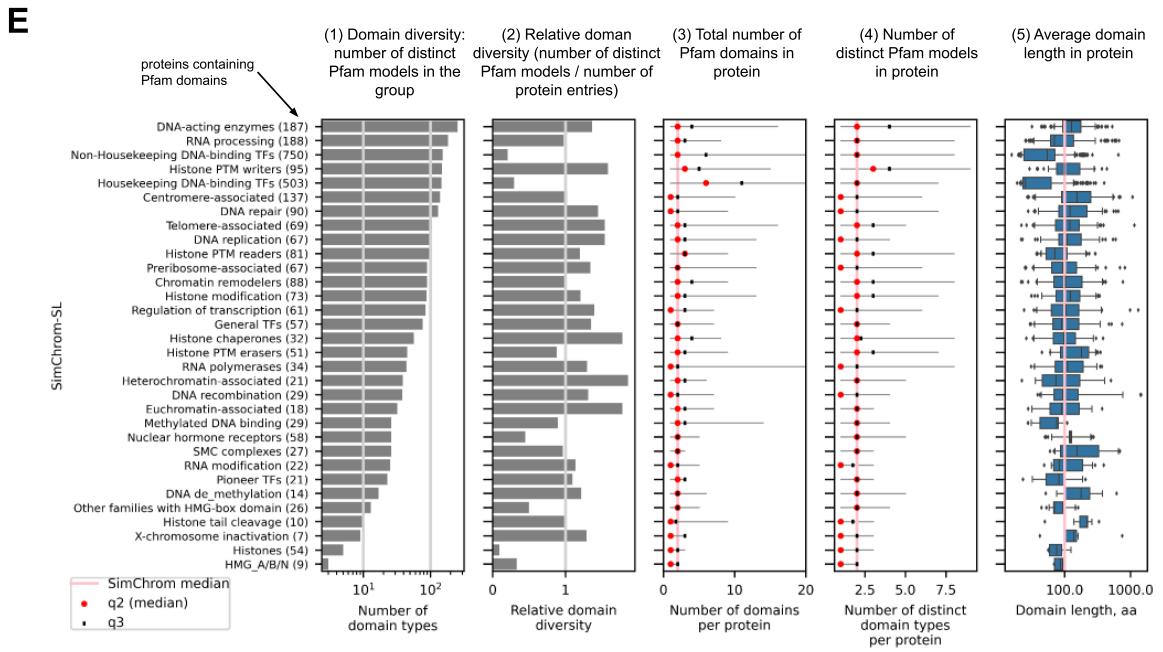


Figure 6. Domain composition of chromatin proteins and identification of new structural domains. (A) A schematic overview of chromatin proteins' domain annotation analysis and identification of uncharacterized new domains. Sources of annotation and typical annotation patterns for an abstract protein are schematically outlined. A structure with a novel domain identified using AI-based annotation pipeline implemented in TED resource is shown on the right. **(B)** Cumulative annotation coverage of all chromatin protein sequences combined at the amino acid level via different resources. Annotation coverage with experimental structures in the PDB database, the AlphaFold database, and three domain annotation databases (Pfam, CATH, TED) is presented. For AlphaFold and PDB additional information about the fraction of annotated amino acids belonging to IDRs and non-IDRs is depicted (see **Methods Section 2.4**). For all annotations additionally the fraction of amino acids belonging to annotated Pfam domain models in the Pfam database is also depicted, for Pfam and TED additionally the fraction of amino acids resolved in PDB is also depicted. **(C)** Analysis of the structural domains in chromatin proteins identified by the TED resource via AlphaFold-based algorithm. The number and fractions of structural domains that have matching structures in the PDB database at various levels of sequence identity are depicted. The structural matches were identified via FoldSeek (see **Methods Section 2.5**). For those domains that were not matched to PDB structures directly a few were annotated by CATH (depicted in orange), the remaining fraction (depicted in magenta) represent novel structural domains present only in TED. **(D)** Analysis of functional domain diversity in chromatin proteins as identified by the Pfam database. 11147 domains belonging to 1753 Pfam domain models were identified. The plot characterizes domain models with respect to the availability of a matching structure in PDB (the median sequence identities of the matches between the chromatin proteins' domains belonging to the respective Pfam model and their best structural match in the PDB database as identified by FoldSeek are shown), an annotated TED domain, or otherwise the absence of structural characterization (see **Methods Section 2.5**). **(E)** Analysis of functional domain diversity in chromatin proteins as identified by the Pfam database for proteins belonging to different chromatin categories according to SimChrom-SL classification. Subpanels 1-5 represent various characteristics.

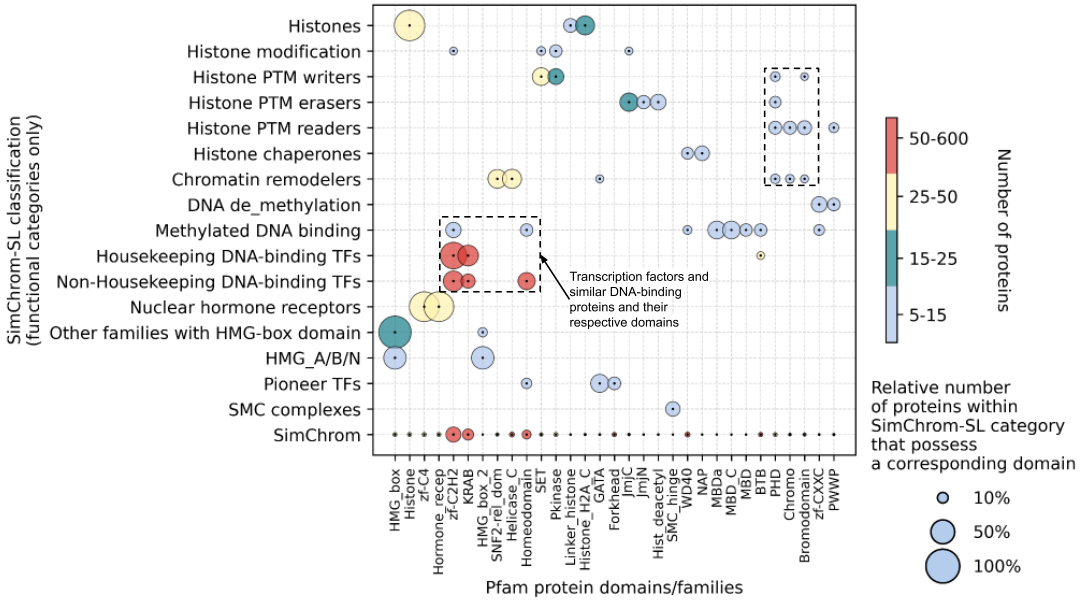


Figure 7. The most representative protein domains/families (according to Pfam) in proteins belonging to functional SimChrom categories. The dashed rectangles highlight the presence of particular groups of domains in certain categories of chromatin proteins (left - transcription factors and similar DNA-binding proteins, right - various histone interacting and modifying proteins). The plot is based on SimChrom-SL chromatin protein classification; only Pfam domain models present in more than five proteins were considered. Only datapoints with the size of more than 5% are displayed. The full size plots based on both SimChrom and SimChrom-SL classifications are available as **Interactive Fig. 4** (https://simchrom.intbio.org/#domain_composition).

Next we set out to systematically analyze the available data on structural characterization, domain annotation and domain composition of chromatin proteins. We specifically explored the structurally uncharacterized portion of the chromatome (the "dark" proteome) and identified potential new structural domains that are predicted by AI-based protein structure prediction tools (see **Fig. 6A**).

Historically, protein domains are loosely defined as evolutionary conserved units with similarities at functional, structural and/or sequence levels [72]. Related individual protein domains may be grouped and aligned to produce domain models, catalogued and annotated by a number of resources/databases such as PFAM [73], CDD [74], CATH [75], InterPro [76, p20], *etc* (see **Suppl. R&D Sec. 3.3** for a thorough discussion). The ultimate experimental structural characterization of chromatin proteins is available in the PDB database, however, recent progress in protein structure prediction spurred by AlphaFold resulted in new approaches to the structural characterization and discovery of new structural domains (*e.g.*, as implemented in the TED database used below [52]) (**Fig. 6A**).

Fig. 6B shows the fractions of the aggregate number of amino acids in all human chromatin proteins (referred below to as "aggregate chromatome sequence", or ACS) which are structurally characterized or have domain annotations in different databases. Detailed discussion is available in

Suppl. R&D Sec. 3.3. Briefly, despite recent tremendous progress in structural biology many human chromatin proteins still lack direct structural characterization. On one hand only 25% of ACS can be mapped directly to PDB structures and 25% can be mapped to known structural protein superfamilies (through CATH). On the other hand, AlphaFold 2 identifies 53% of ACS as belonging to non-IDRs, and TED predicts that 35% of ACS belong to domains having well-defined 3D structures. The latter is a conservative estimate of structurally characterizable ACS, since both partially ordered and disordered regions can become ordered in protein-protein complexes. For example, 9% of ACS is available in PDB (where protein complexes are present) while not being annotated with TED domains (which rely on single protein chain structure predictions). This is in line with the fact that among 6246 TED domains found in chromatin proteins almost half of them (42%) are directly covered by the PDB database. However, the majority of other domains (56%) can be matched to a PDB structure of a homologous protein at various levels of sequence identity (from 99% to 5%, see **Fig. 6C**). The majority of these homologous domains are in fact different paralogous sequences found within human genes (even for domains with sequence identity of 35-50% the fraction of human sequences among the matches was 51%), for matches with sequence identity above 35% the second largest contribution came from structures of mammalian homologues, for matches with sequence identity below 35% significant contributions were from structures derived from proteins of fungi, protostomia and bacteria (see **Suppl. Fig. SF6_1A** for details). Additionally, 6% of TED domains that lacked direct hits among the PDB structures were mapped to protein structural superfamilies in the CATH database. The remaining 4% (241) represented domains could not be matched to any known protein structure or protein structure superfamilies and potentially represent new types of structural superfamilies/folds. These domains are presented in **Suppl. Table ST14** (see also **Interactive Table 3** at https://simchrom.intbio.org/#novel_structural_domains), ranked via their structural complexity by the number of their secondary structure elements. Among these domains, 123 domains have annotations in Pfam or other domain annotation databases present in InterPro, leaving 118 domains that are completely without annotations. The latter domains belong to 106 chromatin proteins, which may be considered as prospective new targets for experimental studies of their function and structure. Among such proteins are, for example, (1) a protein encoded by the GTF3C1 gene (it has a previously unannotated and uncharacterized structural domain with a length of 233 amino acids, see detailed characterization in **Suppl. Fig. SF6_2A**), (2) the globular domain of the testis specific linker histone H1.7, which has a quite different sequence from other H1 proteins resulting in a predicted structure that has a different topology (the "wing" of the globular domain consists of three beta-sheets rather than two [77], see **Suppl. Fig. SF6_2B** and **Suppl. R&D Sec. 3.3**).

We used the sequence-based Pfam domain annotation to characterize the diversity of different types of evolutionary related protein domains (hereafter referred to as Pfam domain models or Pfam domain types) found in chromatin proteins and typical domain composition thereof. In total 1753

different Pfam domain models matched various parts of chromatin proteins (**Fig. 6D**). 42% of these were considered fully structurally characterized, *i.e.*, every individual domain in chromatin proteins belonging to these models can be found in PDB. 34% of domain models are partially characterized – their domains could be matched to a PDB structure of a homolog (using FoldSeek, see **Methods Section 2.5**). 14% of these Pfam domain models were not matched by FoldSeek to PDB structures with our strict criteria (see **Methods Section 2.5**), but could be still identified in PDB via sequence search methods – these represented more flexible domains with IDR regions, repeats and coiled-coils (34 Pfam models), DNA-binding motifs, *etc*. 3% (55 domain models) could be matched to structural domains predicted by AlphaFold and found in the TED database. These represent prospective targets for validation with structural biology methods and further investigation of their interactions. For instance, among these domain models are domains, potentially associated with chromatin remodeling (SANTA, zf-C3Hc3H), histone PTM writing (DUF7030, COMPASS-Shg1), zinc fingers (zf_CCCH_4, zf-LITAF-like, zf-WIZ, SWIM), *etc*. 7% of Pfam domain models currently have no structural information that can be assigned either through the PDB or TED databases.

We next analyzed the diversity of Pfam domain models in various SimChrom-SL protein categories (**Fig. 6E**, subpanels 1,2) and the domain content of individual proteins belonging to these categories (**Fig. 6E**, subpanels 3-5). Detailed discussion is available in **Suppl. R&D Sec. 3.3**. Briefly, the number of distinct Pfam domain models found in chromatin proteins (~1700) is comparable to the number of chromatin proteins (~3000), at the same time an average chromatin protein usually contains two Pfam domains representing two different domain models. The majority of Pfam domain models are present only in a single chromatin protein, but there are also those that are present in dozens or even hundreds of proteins (**Suppl. Fig. SF6_1B**). Certain chromatin groups stand out in terms of their domain composition in some aspects: the number of individual domains is high in housekeeping TF (due to ZFDs); transcription factors, histones and HMG proteins are relatively poor in their domain diversity (*i.e.*, the proteins in these categories harbour a limited number of distinct Pfam domain models); histone PTM writers on average have domains belonging to three different domain models (while this number is one or two for all others). Still a considerable number of chromatin proteins may harbor domains belonging to several domain models. DNA-acting enzymes, histone PTM writers, chaperones, remodelers, transcription factors may have as much as 8-9 Pfam domain models present in their sequence (see **Suppl. Table ST15**). There are 118 chromatin proteins harboring at least five different domain types (see **Suppl. Fig. SF6_1C**, leftpanel). This highlights the multivalency of protein interactions in chromatin, keeping in mind that many proteins further form protein-protein complexes increasing their interaction potential (see next section). The average individual domain length in chromatin proteins is around 65 amino acids (the median is 28 aa), however, this number is biased by the presence of many zinc-finger domains (around 22 aa in length). Subpanel 5 in **Fig. 6E** gives a more balanced view for each SimChrom category. For the majority of protein groups the median domain

length in protein is around 100 amino acids (mean is 137, median is 134). Only 70 chromatin proteins had no domain annotation at all.

The birds-eye view of the most frequently matched Pfam domain models in proteins of various functional SimChrom-SL categories is presented in **Fig. 7**. The data is presented for domain models that occur in at least five chromatin proteins and in at least 10% of proteins in a category (the threshold for data point depiction is 5%). The comprehensive interactive analysis figure with the ability to alter these thresholds and switch between SimChrom and SimChrom-SL classifications systems is available at **Interactive Fig. 4** (https://simchrom.intbio.org/#domain_composition). In **Fig. 7** the following categories and their respective domains can be grouped revealing their partially shared domain composition: 1) the categories containing transcription factors and their zinc finger, homeodomains and KRAB domains form the most frequently occurring entities, 2) some chromatin regulators, such as PTM writers, readers, erasers and chromatin remodelers together with their Chromo-, Bromo-, and PHD domains.

3.3.4. Multivalent interactions in chromatin protein

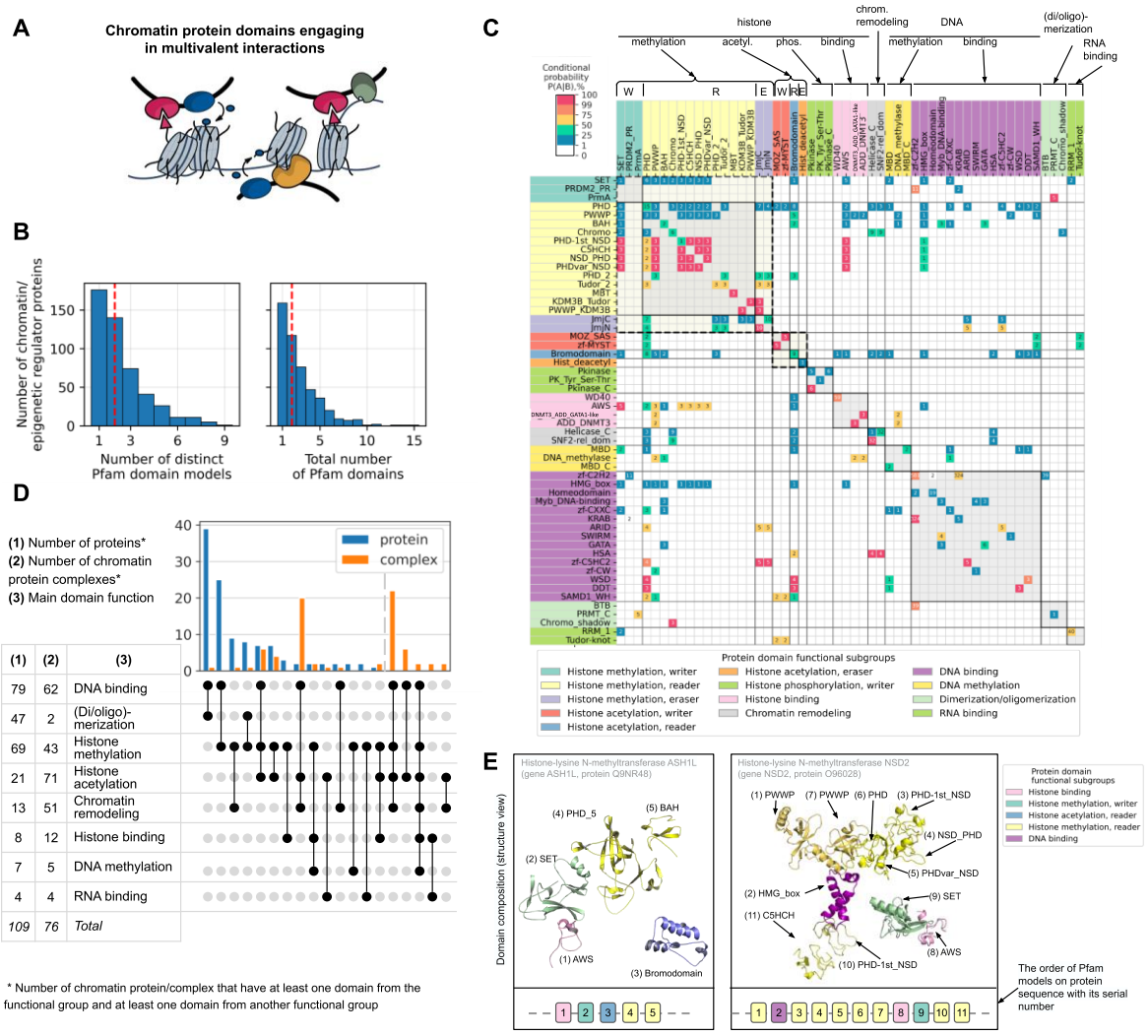


Figure 8. Analysis of multivalent interactions in chromatin proteins. (**A**) Schematic illustration of multivalent interactions. (**B**) Distribution of chromatin proteins from the chromatin/epigenetic regulators group with respect to the total number of Pfam domains (right) and distinct Pfam domain models (left), red lines indicate median values. The distributions of all chromatin proteins are shown in **Suppl. Fig. SF6_2C**. (**C**) Co-occurrence of domains corresponding to different Pfam domain models in chromatin proteins. Only domains found in proteins belonging to chromatin/epigenetic regulator groups are depicted (see **Methods Section 2.5** and **Fig. 3**). Domains are grouped into several functional classes (see description at the top of the plot). The values indicate the conditional probabilities of a domain in column (*A*) occurring alongside a domain in row (*B*) in chromatin proteins. Along the diagonal, data belonging to individual domain groups are highlighted with shading, dashed lines highlight groups associated with histone methylation or acetylation. The following abbreviations are used in the domain subgroup names of the latter: W - writers, R - readers, and E - erasers. (**D**) Co-occurrence of domains from different functional classes in chromatin proteins and protein complexes. Only combinations that are present in more than one protein or complex are shown, see full version in **Suppl. Fig. SF8_2**. (**E**) Examples of domain architectures in chromatin proteins containing the largest number of chromatin/epigenetic regulator domains. The top shows domains from 3D structures colored by their main function; the links between domains are not shown. The bottom shows the order of domains at the sequence level.

The presence of multiple domains (belonging to the same or different domain models) in chromatin proteins is a known feature contributing to their ability to engage in multivalent interactions (**Fig. 8A**) [10]. Below we present the analysis of such domains engaged in multivalent interactions (referred to as EMVI-domains hereafter) that are found in chromatin/epigenetics regulator proteins (see **Fig. 3** for definition of this group). These proteins often contain many domains (**Fig. 8B**). The median total number of domains found in chromatin proteins and in chromatin/epigenetic regulators is two. Nevertheless, many chromatin proteins contain more (16% - have three domains, 10% - four domains, 7% - five, six to fourteen - 10%). There are 409 Pfam domain models that are found in combination with other models or in multiple copies in at least one chromatin regulator protein. To limit our analysis to a manageable set of EMVI-domains, we selected those that were found in multiple copies or in combination with another Pfam domain in at least three chromatin regulator proteins (94 Pfam domain models in total), and from those we selected 59 domain models that we were able to manually classify based on the information currently available in the literature according to their functional binding modes. The following *functional groups* of domains were used: histone methylation/acetylation/phosphorylation, chromatin remodeling, histone binding, DNA binding, DNA methylation, protein dimerization/oligomerization, PPI, RNA binding. Histone post-translational modifications were further subdivided into readers, writers and erasers *functional subgroups* (see **Fig. 8C** and **Suppl. Table ST16** for the list of domains and their detailed classification). Domains involved in histone methylation are most present in chromatin regulators, followed by DNA binding, Histone acetylation, Histone phosphorylation and Chromatin remodeling associated domains (**Suppl. Fig. SF8_1A**).

We analyzed the co-occurrence of selected EMVI-domains in all chromatin proteins. There were in total 851 chromatin proteins (589 of these are transcription factors) that had more than one EMVI-domain. The conditional probability of finding a corresponding domain A in a chromatin protein given that another domain B is already present was estimated and is presented in **Fig. 8C** (columns and rows correspond to domains A and B, respectively). The **Interactive Fig. 5** is available at https://simchrom.intbio.org/#domain_co-occurrence (also extends the analysis to unclassified potential EMVI-domains found in at least two chromatin regulator proteins). The matrix in **Fig. 8C** allows to trace the interplay between different domains employed in architectures of chromatin proteins. The largest groups of domains in **Fig. 8C** are those involved in histone methylation and DNA binding, suggesting that these mechanisms are the most represented and employed in chromatin functioning regulation. See **Suppl. R&D Sec. 3.4** for detailed discussion of the results. Briefly, in certain cases one can see 100% association between the presence of various domains in chromatin proteins. This may be due to direct structural interactions between the domains or likely due to functional reasons. Among the Pfam domains that co-occur with the most number of other different Pfam domains is the PHD domain

(45 other domains), Bromodomain (38), SET (40) and PWWP (28) and chromatin remodeling Helicase_C (33) and SNF2-rel_dom (31), **Suppl. Table ST16**.

A more general view of multivalent interactions may be obtained if we trace the relationships within or between different functional groups of domains. One can see that domains from the same functional group (especially histone methylation) tend to co-occur (**Fig. 8C**, **Suppl. Fig. SF8_1B,C**) and may also be present in multiple instances in proteins (*e.g.*, in NSD2 there are nine histone methylation associated domains, see below). If a chromatin protein has a domain involved in histone methylation (either writing, reading or erasing) there is an estimated 38% chance that there will be another different functional domain from this group of domains (**Suppl. Table ST17**, **Suppl. Fig. SF8_1C**). For acetylation this estimated probability is 31%, for phosphorylation 18%. The associations between the occurrence of domains from different functional groups can also be observed. Domains involved in histone methylation (one of the most abundant groups by the number of Pfam models and the number of chromatin proteins) may be in a considerable number of chromatin proteins combined with other EMVI-domains (particularly DNA binding domains and histone acetylation), association with histone binding domains, chromatin remodeling, DNA methylation, (di/oligo)-merization and RNA binding domains was also observed (see **Suppl. Fig. SF8_2C**). The same can be said about domains involved in histone acetylation, although in a somewhat smaller number of cases, and with exclusion of their combination with dimerization domains. Notably, domains involved in histone phosphorylation were not found in combination with domains from other functional groups in our analysis. This may reflect an evolutionary strategy whereby combinations of histone methylation and acetylation evolved to delicately regulate gene expression at the epigenetic level, while phosphorylation remained as a more general mechanism affecting a broad number of proteins and pathways in the cell.

For a more comprehensive view of multivalent interactions it is reasonable to (1) analyze not only pairwise co-occurrence of different functional domains, but simultaneous co-occurrence of domains from several functional groups in proteins, (2) extend the analysis to complexes of chromatin proteins. The results of such analyses are presented in **Fig. 8D** (see **Methods Section 2.1.4** for our selection of 513 protein complexes (out of 2266) where all proteins are chromatin proteins from Complex Portal [47]). In our analysis at the level of individual proteins, proteins harbored domains only from up to four functional groups. Particularly, DNA binding domains may be combined with (di/oligo)-merization, chromatin remodeling, histone methylation, methylation and acetylation or histone acetylation and chromatin remodeling domains. The formation of chromatin protein complexes considerably enlarges the available combinations of functional domains. Among the 513 analyzed protein complexes, 181 complexes contained EMVI-domains from the analyzed functional domain groups, 101 complexes harboured more than one domain, 80 complexes harbored domains from different functional groups. From these 80 complexes the majority were various chromatin remodeling complexes (53 complexes), others representatives included histone acetyltransferase (13) and

deacetylase (4) complexes, and DNA-methyltransferase complexes (2). One can see that the largest number of analyzed complexes (24) simultaneously contained domains from four functional groups (DNA binding, histone methylation, histone acetylation, and chromatin remodeling), in a select number of complexes domains belonging to up to six functional groups were observed (all the above mentioned together with domains involved in DNA methylation and histone binding). These were all complexes involved in chromatin remodeling. For example, 'MBD2' or 'MBD3/NuRD nucleosome remodeling and deacetylase complex'. Notably, in chromatin complexes histone acetylation domains are found more often than histone methylation domains (unlike in the case when individual proteins are analyzed), this might, however, be biased by the current list of known chromatin complexes and their variability. Taken together chromatin protein complexes expand the multivalency of chromatin protein interactions and expand the functionality of the complexes.

As a final step we looked for chromatin proteins that matched to the most number of Pfam domain models that belonged to different functional groups/sub-groups and thus might engage in interactions of high multivalency (see **Suppl. Fig. SF8_2A**). According to this analysis, chromatin proteins could harbor domains from up to four functional groups/sub-groups. For example, histone-lysine N-methyltransferase ASH1L is involved in reading and writing of histone methylation, reading of histone acetylation and histone binding (**Fig. 8E**). Proteins could harbor up to nine Pfam domain models (some from identical functional categories). One of such examples is Histone-lysine N-methyltransferase NSD2 (**Fig. 8E**). This protein combines domains that likely engage in methylated histone binding (PWWP, PHD-1st_NSD, NDS_PHD, PHDvar_NSD, PHD, C5HCH), histone methylation (SET), histone binding (AWS) and may be DNA binding (HMG_box).

4. DISCUSSION

The presented study is a comprehensive effort to summarize, critically evaluate, systematize various types of information sources about chromatin proteins and use this meta-analysis to perform the state-of-the-art characterization of chromatin proteins' composition and properties in order to describe the functioning of the chromatome as a holistic entity.

To our knowledge this is the first meta-analysis study aimed at identifying which human proteins are present in chromatin (the chromatome) where we cross-compared conceptually different sources of experimental (mass-spectrometry-based and microscopy based) and database derived information (including databases of protein localization and database of known classes of chromatin proteins). The results of our meta-analysis highlight challenges both in establishing the *per se* definition of the chromatome and using various approaches to collect the list of proteins according to the said definition. In fact, historically, chromatome was operationally defined either through the sets of proteins that were known to contribute to the structure and functioning of chromatin (which itself is an evolving

term) [1] or through proteomics analysis of chromatin extracts (which depend on the experimental techniques) [17]. In our view, the chromatome—under a broad definition—may include all proteins with nuclear localization that can interact with the genomic DNA either directly or indirectly, with the possible exception of proteins tightly bound to the nuclear envelope. As we showed, within this definition, there is currently limited congruence between the sources of information and further efforts are needed to improve the quality and comprehensiveness of the datasets. Particularly, the MS-based studies suffer from an inherent limitation due to the fact that chromatin is a "fuzzy organelle" [23] (whose composition depends on the type of the cell line used and other conditions, given that many genes are conditionally expressed), supplemented by the technical problems in identifying the low abundant proteins and contamination during chromatin extraction. Localization studies have limited coverage of the proteome space, and may report only certain localizations for proteins, while according to our estimates ~50% of nuclear proteins may be also localized in other cellular compartments.

The processes happening within the nuclei of eukaryotic cells are perhaps ones of the most complex that Nature has created. While a lot is already known about the molecular mechanisms of processes happening in chromatin, the holistic understanding of its functioning that would be able to explain how complex and tightly regulated organismal traits and behavior originates from these molecular processes remains a big challenge. While a lot of expectations were based on application of cybernetics [78], and now are based on big data and AI technologies, one of the main approaches to rational understanding is the reductionist analysis of the system. Hence, keeping in mind all the known limitations (such as the multiple functions performed by many proteins) in order to understand and to analyze the properties of the human chromatome we proposed the hierarchical SimChrom classification ontology of chromatin proteins based on a manual synthesis and analysis of the established body of knowledge. The main advantage of SimChrom is its relative simplicity which enabled us to analyze and compare the properties of different groups of chromatin proteins without being overwhelmed by the amount of details. Other classifications, such as GeneOntology [35,36] do not offer such an advantage to this end, other databases do not offer the comprehensiveness of SimChrom and focus on only specific protein classes [55,79,43]. The SimChrom classification ontology would not be useful without the actual classification of ~3000 of chromatin proteins according to this ontology which form the SimChrom dataset. Another approximation/tradeoff useful for the rational analysis of the chromatome introduced in this study is the use of the SimChrom-SL classification, where each protein is ascribed only to one SimChrom category. Such an approach allowed us to highlight the differences in protein properties that may be essential for the particular protein class.

In our analysis of chromatin protein abundance we observed that MS-based studies of chromatin protein extracts yield very fluctuating abundance estimates (due to their analysis of specific cell lines and apparent bias to detect highly abundant proteins) and hence establishing the exact

abundance profiles of chromatin proteins inside the cell nucleus is in our opinion an unsolved challenge. Various chromatin MS-based studies provided highly varying intensities for histones. Hence techniques that attempt to convert MS-intensities to protein copy numbers based on known histone mass in the cell (such as "proteomic ruler") [80] are likely not applicable in the case of chromatin extracts studies. To analyze chromatin proteins' abundance, we used the normalized data from whole organism proteomes collected by PaxDb, which allowed us to gain certain insights but could provide only aggregate abundance for proteins with multiple localizations. According to our analysis, the distribution of chromatin proteins with respect to their abundance is bimodal with approximately one half representing the highly abundant ones (> 1 ppm) and another half the low abundant ones (< 1 ppm). This is quite similar to that observed for the whole proteome. Chromatin proteins are moderately enriched by those performing housekeeping functions in comparison with the whole proteome (60% of chromatin proteins are housekeeping ones). This suggests that the presence of many low-abundant chromatin proteins in the whole organism proteome may be attributed not only to their mere conditional expression in certain cells, but at least in certain cases to their low levels of expression in the cells *per se* (assuming that housekeeping genes are expressed in all cells*)*. A significant fraction of the chromatome (45 %) is represented by transcription factors. Proteomics and systems-biology studies of TF regulatory networks have shown that TFs can be expressed at very low levels and still play essential roles in cellular processes such as environmental responses, development, and differentiation [81,82]. Using our SimChrom classification and combining data on whole organism protein abundance and protein localization we were able to characterize the main chromatin proteins and chromatin protein groups that dominate by their mass fraction in the cell and the cell nucleus. Early studies established that histones are present in approximately the same mass as DNA, whereas non-histone chromosomal proteins contribute about 0.3-0.8 g per gram of DNA [1]. However, results can vary substantially depending on the experimental approach, methodology, and resolution [83]. To what extent these estimates are accurate within the definition of the chromatome used in this study still remains to be quantified. Our analysis is consistent with the fact that histones (with H4 histone being the sole leader due to the identity of its protein sequence among the family of encoding genes) dominate the protein contents of the cell nucleus (by the number of molecules and likely by the total mass), however, the proteins involved in RNA processing (including splicing factors, pre-mRNA binding proteins) rivals them once the mass fractions are compared (these proteins are on average considerably longer than histones). The latter fact highlights the functional importance of the nucleus as not only the DNA storing and processing entity, but also as a factory to produce and mature RNA molecules. This idea is also supported by the fact that among the most abundant individual proteins in the nucleus are histone chaperones nucleophosmin (NPM1) and nucleolin (NCL), which are also involved in nucleolar organization and ribosome biogenesis. Our analysis is consistent with earlier established knowledge that among the most abundant *bona fide* DNA/nucleosome interacting proteins are those from the HMG protein group. This is a

diverse group of small, basic and abundant proteins that modulate chromatin architecture and dynamics through non-sequence-specific and dynamic interactions with DNA and nucleosomes [84].

Using our carefully prepared datasets and chromatin proteins classification we were able to systematically and quantitatively address the questions of chromatin proteins composition both at the amino acid and domain level. The importance of intrinsically disordered regions in chromatin proteins have recently been the focus of many studies, suggesting their properties are essential to the dynamic nature of chromatin functioning. IDRs enable proteins to interact transiently and multivalently with multiple partners [85,86], recruit them through short linear motifs (SLiMs) [87,88] and "fuzzy" interactions where proteins can adapt to multiple partners [89,90], facilitate DNA binding and DNA motif search and recognition [91,92], promote liquid-liquid phase separation and formation on non-membrane organelles enriched with certain proteins [6,93]. We showed that both the fraction and the number of IDRs in chromatin proteins is higher than in cytoplasmic ones, which is reflected in particular changes in their amino acids composition. Specifically, they are enriched in serine and proline and are depleted in hydrophobic aliphatic amino acids. It is generally assumed that chromatin proteins are positively charged to compensate for the negative charge of DNA and RNA molecules. We show that this is true in general, but certain important protein groups are negatively charged (histone chaperones, RNA polymerases, Histone PTM erasers), suggesting charge-charge interactions are an important factor in organizing the structure and functioning of chromatin. Moreover, we find that there are certain specific charge patterns along the protein sequence – tails of chromatin proteins (especially the N-tails) are negatively charged on average. This property may influence translation efficiency and protein expression, consistent with the general relationship between peptide charge and translation [94]. We showed that the increased positive charge of chromatin proteins is conferred through alterations in the occurrence of positive and negative amino acids which is different for IDRs and non-IDRs. Counterintuitively, certain positive amino acids may be depleted in certain regions (*e.g.*, arginine is depleted in IDRs of chromatin proteins), suggesting there may be other factors and physico-chemical properties affecting the presence of charged amino acids along the protein sequence. Arginine residues exhibit a strong affinity for DNA due to their positively charged guanidinium groups, facilitating electrostatic interactions with the negatively charged DNA backbone. This property is advantageous in structured DNA-binding domains but may be detrimental in intrinsically disordered regions. Arginine residues strongly interact with DNA and RNA via their positively charged guanidinium groups, stabilizing structured nucleic acid–binding domains. However, in IDRs, excessive arginine can cause nonspecific nucleic acid binding, restricting flexibility and disrupting LLPS [95–97]. This might be an explanation for the relative depletion of arginine in IDRs of chromatin-associated proteins, balancing nucleic acid interactions with the need for dynamic, phase-separating behavior. Our analysis of domain annotation of chromatin proteins suggests that while non-IDRs are relatively well annotated (~60% are annotated by Pfam), the IDRs are poorly annotated (~87% are unannotated) - highlighting the

challenges in understanding their functions and interactions. While IDRs resist direct structural characterization with conventional structural biology methods, we envision that the application of molecular modeling in combination with the state-of-the-art machine learning models that can link protein sequence, dynamics, and function is a perspective way to move forward in deciphering the role of IDRs in chromatin functioning [98–101].

The structural characterization of chromatin protein domains is also an important task that is not completely solved. According to AlphaFold ~50% of the aggregate chromatin protein sequence are non-IDRs, while 35% form *bona fide* structural domains in monomeric proteins (as predicted by AphaFold and TED). Among these domains around 70% have relatively close homologs (with sequence identity of more than 50%) available in PDB. The advances in AI-based protein structure prediction currently open exciting opportunities to bridge this gap in the structural characterization of chromatin proteins. Particularly, using such tools we identified 241 structural domains in chromatin proteins whose 3D structures (or structure of their detectable homologs) were not previously experimentally solved, 106 proteins have domains that are also not functionally annotated by Pfam. These domains await experimental structural and functional characterization. We envision that the development and application of AI-based structure prediction tools will facilitate chromatin research, particular in analysis of protein-protein interactions, which is already happening [102,103].

The importance of multivalent protein interactions in chromatin was highlighted by D. Allis and colleagues around 20 years ago [104]. Here we performed systematic analysis of domains engaged in multivalent interactions present in chromatin regulator proteins (histone PTM readers/writers/erasers, chromatin remodelers, histone chaperones, proteins involved in DNA or nuclear RNA modifications). We showed that chromatin regulators indeed actively use such domains and may contain domains corresponding to up to nine distinct Pfam domain models. In chromatin protein complexes the number of such domains is significantly increased. Chromatin regulators contain an especially large number of domains involved in histone methylation that may be combined with histone acetylation, histone binding, DNA binding and DNA methylation domains. The data collected in this study provide further framework for understanding the engagement of multivalent interaction in chromatin functioning.

The hallmark feature of this work is the SimChrom web-resource which allows the user to explore the results of our analysis interactively, query and compare the information about the proteins the user is interested in from the curated datasets. This is a unique tool that uses interactive web-based data representation to address the multidimensionality and heterogeneity of the available data about chromatin proteins and allows to extract knowledge from the data. Inspired by the best practices of collaborative research in data science, SimChrom is implemented as a GitHub repository that is rendered as a web-site directly from GitHub, making it a secure, reliable, open source tool that may be easily copied and modified by the community.

Taken together we hope that our work establishes a holistic framework for further advances in the field of chromatin research which will help to understand genome functioning though deeper appreciation of the complex role played by the chromatome.

ACKNOWLEDGEMENTS

We thank A.L. Sivkina, D.K. Malinina, N.S. Gerasimova, A.V. Lyubitelev, S.V. Ulianov, and A.A. Gavrilov for valuable discussions that helped to improve this work.

AUTHOR CONTRIBUTIONS

AKG: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Visualization. GAA: Resources, Software. MPK: Conceptualization. Writing – review & editing. AKS: Conceptualization, Formal analysis, Funding acquisition, Methodology, Supervision, Validation, Writing – original draft, Writing – review & editing.

SUPPLEMENTARY DATA

Supplementary material is available online, including supplementary figures, tables, supplementary results and discussion.

CONFLICT OF INTEREST

Non declared.

FUNDING

This work was funded by the Russian Science Foundation grant #25-14-00046 (https://rscf.ru/en/project/25-14-00046/) (construction of chromatin protein classification, analysis of chromatin proteins domain composition, AI-based prediction of new structural domains), the Russian Science Foundation grant #23-74-10012 (https://rscf.ru/en/project/23-74-10012/) (analysis of physicochemical properties of chromatin proteins), and within the framework of the Ministry of Science and Higher Education of the Russian Federation project “Whole-Genome Epigenetic Analysis as the Basis for the Development of Genetic Technologies for the Prevention and Treatment of COVID” (FFRW-2023-0007), no. 123120500032-9 (analysis of multivalent interactions of chromatin protein domains). A.K.S. was supported by the HSE University Basic Research Program (structural

characterization of chromatin proteins) and A.K.G. was supported by the Gennady Komissarov Foundation (construction of reference datasets about protein localization).

DATA AVAILABILITY

The SimChrom database including interactive supplementary materials about chromatin proteins' classification, localization, functions, domain composition are freely available at a GitHub hosted web site https://simchrom.intbio.org/. The SimChrom source code is available at GitHub https://github.com/intbio/SimChrom and archived via Zenodo.

REFERENCES

1. Van Holde KE. *Chromatin*. Springer; 1989. doi:10.1007/978-1-4612-3490-6

2. Bernstein E, Allis CD. RNA meets chromatin. *Genes Dev*. 2005;19(14):1635-1655. doi:10.1101/gad.1324305

3. Imhof A, Bonaldi T. "Chromatomics" the analysis of the chromatome. *Mol BioSyst*. 2005;1(2):112-116. doi:10.1039/B502845K

4. Torrente MP, Zee BM, Young NL, et al. Proteomic Interrogation of Human Chromatin. *PLOS ONE*. 2011;6(9):e24747. doi:10.1371/journal.pone.0024747

5. Ulianov SV, Velichko AK, Magnitov MD, et al. Suppression of liquid-liquid phase separation by 1,6-hexanediol partially compromises the 3D genome organization in living cells. *Nucleic Acids Res*. 2021;49(18):10524-10541. doi:10.1093/nar/gkab249

6. Rippe K. Liquid–Liquid Phase Separation in Chromatin. *Cold Spring Harb Perspect Biol*. 2022;14(2):a040683. doi:10.1101/cshperspect.a040683

7. Ulianov SV, Khrameeva EE, Gavrilov AA, et al. Active chromatin and transcription play a key role in chromosome partitioning into topologically associating domains. *Genome Res*. 2016;26(1):70-84. doi:10.1101/gr.196006.115

8. Davidson IF, Peters JM. Genome folding through loop extrusion by SMC complexes. *Nat Rev Mol Cell Biol*. 2021;22(7):445-464. doi:10.1038/s41580-021-00349-7

9. Kanada R, Terakawa T, Kenzaki H, Takada S. Nucleosome Crowding in Chromatin Slows the Diffusion but Can Promote Target Search of Proteins. *Biophys J*. 2019;116(12):2285-2295. doi:10.1016/j.bpj.2019.05.007

10. Ruthenburg AJ, Li H, Patel DJ, Allis CD. Multivalent engagement of chromatin modifications by linked binding modules. *Nat Rev Mol Cell Biol*. 2007;8(12):983-994. doi:10.1038/nrm2298

11. Armeev GA, Gribkova AK, Shaytan AK. NucleosomeDB - a database of 3D nucleosome structures and their complexes with comparative analysis toolkit. *bioRxiv*. Preprint posted online April 18, 2023:2023.04.17.537230. doi:10.1101/2023.04.17.537230

12. Armeev GA, Gribkova AK, Shaytan AK. Nucleosomes and their complexes in the cryoEM era: Trends and limitations. *Front Mol Biosci*. 2022;9:1070489. doi:10.3389/fmolb.2022.1070489

13. Doronin SA, Ilyin AA, Kononkova AD, et al. Nucleoporin Elys attaches peripheral chromatin to the nuclear pores in interphase nuclei. *Commun Biol*. 2024;7(1):1-18. doi:10.1038/s42003-024-06495-w

14. Pletenev IA, Bazarevich M, Zagirova DR, et al. Extensive long-range polycomb interactions and weak compartmentalization are hallmarks of human neuronal 3D genome. *Nucleic Acids Research*. 2024;52(11):6234-6252. doi:10.1093/nar/gkae271

15. Consens ME, Dufault C, Wainberg M, et al. Transformers and genome language models. *Nat Mach Intell*. 2025;7(3):346-362. doi:10.1038/s42256-025-01007-9

16. Hwang Y, Cornman AL, Kellogg EH, Ovchinnikov S, Girguis PR. Genomic language model predicts protein co-regulation and function. *Nat Commun*. 2024;15(1):2880. doi:10.1038/s41467-024-46947-9

17. van Mierlo G, Vermeulen M. Chromatin Proteomics to Study Epigenetics - Challenges and Opportunities. *Mol Cell Proteomics*. 2021;20:100056. doi:10.1074/mcp.R120.002208

18. Kustatscher G, Grabowski P, Rappsilber J. Multiclassifier combinatorial proteomics of organelle shadows at the example of mitochondria in chromatin data. *Proteomics*. 2016;16(3):393-401. doi:10.1002/pmic.201500267

19. Ohta S, Bukowski-Wills JC, Sanchez-Pulido L, et al. The Protein Composition of Mitotic Chromosomes Determined Using Multiclassifier Combinatorial Proteomics. *Cell*. 2010;142(5):810-821. doi:10.1016/j.cell.2010.07.047

20. Sinitcyn P, Richards AL, Weatheritt RJ, et al. Global detection of human variants and isoforms by deep proteome sequencing. *Nat Biotechnol*. 2023;41(12):1776-1786. doi:10.1038/s41587-023-01714-x

21. Guo T, Steen JA, Mann M. Mass-spectrometry-based proteomics: from single cells to clinical applications. *Nature*. 2025;638(8052):901-911. doi:10.1038/s41586-025-08584-0

22. Wierer M, Mann M. Proteomics to study DNA-bound and chromatin-associated gene regulatory complexes. *Hum Mol Genet*. 2016;25(R2):R106-R114. doi:10.1093/hmg/ddw208

23. Kustatscher G, Hégarat N, Wills KLH, et al. Proteomics of a fuzzy organelle: interphase chromatin. *EMBO J*. 2014;33(6):648-664. doi:10.1002/embj.201387614

24. Ugur E, de la Porte A, Qin W, et al. Comprehensive chromatin proteomics resolves functional phases of pluripotency and identifies changes in regulatory components. *Nucleic Acids Research*. 2023;51(6):2671-2690. doi:10.1093/nar/gkad058

25. Ginno PA, Burger L, Seebacher J, Iesmantavicius V, Schübeler D. Cell cycle-resolved chromatin proteomics reveals the extent of mitotic preservation of the genomic regulatory landscape. *Nat Commun*. 2018;9(1):4048. doi:10.1038/s41467-018-06007-5

26. Alabert C, Bukowski-Wills JC, Lee SB, et al. Nascent chromatin capture proteomics determines chromatin dynamics during DNA replication and identifies unknown fork components. *Nat Cell Biol*. 2014;16(3):281-291. doi:10.1038/ncb2918

27. Shi M, You K, Chen T, et al. Quantifying the phase separation property of chromatin-associated proteins under physiological conditions using an anti-1,6-hexanediol index. *Genome Biology*. 2021;22(1):229. doi:10.1186/s13059-021-02456-2

28. Alvarez V, Bandau S, Jiang H, et al. Proteomic profiling reveals distinct phases to the restoration of chromatin following DNA replication. *Cell Reports*. 2023;42(1). doi:10.1016/j.celrep.2023.111996

29. Chou DM, Adamson B, Dephoure NE, et al. A chromatin localization screen reveals poly (ADP ribose)-regulated recruitment of the repressive polycomb and NuRD complexes to sites of DNA damage. *Proceedings of the National Academy of Sciences*. 2010;107(43):18475-18480. doi:10.1073/pnas.1012946107

30. Federation AJ, Nandakumar V, Searle BC, et al. Highly Parallel Quantification and Compartment Localization of Transcription Factors and Nuclear Proteins. *Cell Reports*. 2020;30(8):2463-2471.e5. doi:10.1016/j.celrep.2020.01.096

31. Dutta B, Ren Y, Hao P, et al. Profiling of the Chromatin-associated Proteome Identifies HP1BP3 as a Novel Regulator of Cell Cycle Progression. *Mol Cell Proteomics*. 2014;13(9):2183-2197. doi:10.1074/mcp.M113.034975

32. Geladaki A, Kočevar Britovšek N, Breckels LM, et al. Combining LOPIT with differential ultracentrifugation for high-resolution spatial proteomics. *Nat Commun*. 2019;10(1):331. doi:10.1038/s41467-018-08191-w

33. Wang H, Syed AA, Krijgsveld J, Sigismondo G. Isolation of Proteins on Chromatin Reveals Signaling Pathway–Dependent Alterations in the DNA-Bound Proteome. *Molecular & Cellular Proteomics*. 2025;24(3). doi:10.1016/j.mcpro.2025.100908

34. Razin SV, Iarovaia OV, Vassetzky YS. A requiem to the nuclear matrix: from a controversial concept to 3D organization of the nucleus. *Chromosoma*. 2014;123(3):217-224. doi:10.1007/s00412-014-0459-8

35. Ashburner M, Ball CA, Blake JA, et al. Gene Ontology: tool for the unification of biology. *Nat Genet*. 2000;25(1):25-29. doi:10.1038/75556

36. The Gene Ontology Consortium, Aleksander SA, Balhoff J, et al. The Gene Ontology knowledgebase in 2023. *Genetics*. 2023;224(1):iyad031. doi:10.1093/genetics/iyad031

37. Gorski S, Misteli T. Systems biology in the cell nucleus. *Journal of Cell Science*. 2005;118(18):4083-4092. doi:10.1242/jcs.02596

38. Johnstone CP, Wang NB, Sevier SA, Galloway KE. Understanding and Engineering Chromatin as a Dynamical System across Length and Timescales. *Cell Systems*. 2020;11(5):424-448. doi:10.1016/j.cels.2020.09.011

39. The UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2025. *Nucleic Acids Research*. 2025;53(D1):D609-D617. doi:10.1093/nar/gkae1010

40. Thul PJ, Åkesson L, Wiking M, et al. A subcellular map of the human proteome. *Science*. 2017;356(6340):eaal3321. doi:10.1126/science.aal3321

41. Cho NH, Cheveralls KC, Brunner AD, et al. OpenCell: Endogenous tagging for the cartography of human cellular organization. *Science*. 2022;375(6585):eabi6983. doi:10.1126/science.abi6983

42. Binns D, Dimmer E, Huntley R, Barrell D, O'Donovan C, Apweiler R. QuickGO: a web-based tool for Gene Ontology searching. *Bioinformatics*. 2009;25(22):3045-3046. doi:10.1093/bioinformatics/btp536

43. Marakulina D, Vorontsov IE, Kulakovskiy IV, Lennartsson A, Drabløs F, Medvedeva YA. EpiFactors 2022: expansion and enhancement of a curated database of human epigenetic factors and complexes. *Nucleic Acids Research*. 2023;51(D1):D564-D570. doi:10.1093/nar/gkac989

44. Lovering RC, Gaudet P, Acencio ML, et al. A GO catalogue of human DNA-binding transcription factors. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*. 2021;1864(11):194765. doi:10.1016/j.bbagrm.2021.194765

45. Itzhak DN, Tyanova S, Cox J, Borner GH. Global, quantitative and dynamic mapping of protein subcellular localization. Hegde RS, ed. *eLife*. 2016;5:e16950. doi:10.7554/eLife.16950

46. Uhlén M, Fagerberg L, Hallström BM, et al. Tissue-based map of the human proteome. *Science*. 2015;347(6220):1260419. doi:10.1126/science.1260419

47. Balu S, Huget S, Medina Reyes JJ, et al. Complex portal 2025: predicted human complexes and enhanced visualisation tools for the comparison of orthologous and paralogous complexes. *Nucleic Acids Res*. 2025;53(D1):D644-D650. doi:10.1093/nar/gkae1085

48. Raudvere U, Kolberg L, Kuzmin I, et al. g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Res*. 2019;47(W1):W191-W198. doi:10.1093/nar/gkz369

49. Wang M, Herrmann CJ, Simonovic M, Szklarczyk D, von Mering C. Version 4.0 of PaxDb: Protein abundance data, integrated across model organisms, tissues, and cell-lines. *Proteomics*. 2015;15(18):3163-3168. doi:10.1002/pmic.201400441

50. Akdel M, Pires DEV, Pardo EP, et al. A structural biology community assessment of AlphaFold2 applications. *Nat Struct Mol Biol*. 2022;29(11):1056-1067. doi:10.1038/s41594-022-00849-w

51. Bordin N, Sillitoe I, Nallapareddy V, et al. AlphaFold2 reveals commonalities and novelties in protein structure space for 21 model organisms. *Commun Biol*. 2023;6(1):160. doi:10.1038/s42003-023-04488-9

52. Lau AM, Bordin N, Kandathil SM, et al. Exploring structural diversity across the protein universe with The Encyclopedia of Domains. *Science*. Published online November 1, 2024. doi:10.1126/science.adq4946

53. van Kempen M, Kim SS, Tumescheit C, et al. Fast and accurate protein structure search with Foldseek. *Nat Biotechnol*. Published online May 8, 2023:1-4. doi:10.1038/s41587-023-01773-0

54. Gligorijević V, Renfrew PD, Kosciolek T, et al. Structure-based protein function prediction using graph convolutional networks. *Nat Commun*. 2021;12(1):3168. doi:10.1038/s41467-021-23303-9

55. Lambert SA, Jolma A, Campitelli LF, et al. The Human Transcription Factors. *Cell*. 2018;172(4):650-665. doi:10.1016/j.cell.2018.01.029

56. Draizen EJ, Shaytan AK, Mariño-Ramírez L, Talbert PB, Landsman D, Panchenko AR. HistoneDB 2.0: a histone database with variants—an integrated resource to explore histones and their variants. *Database*. 2016;2016:baw014. doi:10.1093/database/baw014

57. Zhang Y, Zhang Y, Song C, et al. CRdb: a comprehensive resource for deciphering chromatin regulators in human. *Nucleic Acids Research*. 2023;51(D1):D88-D100. doi:10.1093/nar/gkac960

58. Hammond CM, Strømme CB, Huang H, Patel DJ, Groth A. Histone chaperone networks shaping chromatin function. *Nat Rev Mol Cell Biol*. 2017;18(3):141-158. doi:10.1038/nrm.2016.159

59. Reeves R. High mobility group (HMG) proteins: Modulators of chromatin structure and DNA repair in mammalian cells. *DNA Repair*. 2015;36:122-136. doi:10.1016/j.dnarep.2015.09.015

60. Mayran A, Drouin J. Pioneer transcription factors shape the epigenetic landscape. *J Biol Chem*. 2018;293(36):13795-13804. doi:10.1074/jbc.R117.001232

61. Sun H, Fu B, Qian X, Xu P, Qin W. Nuclear and cytoplasmic specific RNA binding proteome enrichment and its changes upon ferroptosis induction. *Nat Commun*. 2024;15(1):852. doi:10.1038/s41467-024-44987-9

62. Van Nostrand EL, Freese P, Pratt GA, et al. A large-scale binding and functional map of human RNA-binding proteins. *Nature*. 2020;583(7818):711-719. doi:10.1038/s41586-020-2077-3

63. Azad GK, Swagatika S, Kumawat M, Kumawat R, Tomar RS. Modifying Chromatin by Histone Tail Clipping. *Journal of Molecular Biology*. 2018;430(18, Part B):3051-3067. doi:10.1016/j.jmb.2018.07.013

64. Lee H, Noh H, Ryu JK. Structure-function relationships of SMC protein complexes for DNA loop extrusion. *BioDesign*. 2021;9(1):1-13. doi:10.34184/kssb.2021.9.1.1

65. Cartwright P, Helin K. Nucleocytoplasmic shuttling of transcription factors. *Cell Mol Life Sci*. 2000;57(8-9):1193-1206. doi:10.1007/pl00000759

66. Cheng D, Semmens K, McManus E, et al. The nuclear transcription factor, TAF7, is a cytoplasmic regulator of protein synthesis. *Science Advances*. Published online December 2021. doi:10.1126/sciadv.abi5751

67. Shreberk-Shaked M, Oren M. New insights into YAP/TAZ nucleo-cytoplasmic shuttling: new cancer therapeutic opportunities? *Mol Oncol*. 2019;13(6):1335-1341. doi:10.1002/1878-0261.12498

68. Kobiyama K, Kawashima A, Jounai N, et al. Role of Extrachromosomal Histone H2B on Recognition of DNA Viruses and Cell Damage. *Front Genet*. 2013;4:91. doi:10.3389/fgene.2013.00091

69. Zeng Z, Chen L, Luo H, Xiao H, Gao S, Zeng Y. Progress on H2B as a multifunctional protein related to pathogens. *Life Sciences*. 2024;347:122654. doi:10.1016/j.lfs.2024.122654

70. Sigismondo G, Papageorgiou DN, Krijgsveld J. Cracking chromatin with proteomics: From chromatome to histone modifications. *PROTEOMICS*. 2022;22(15-16):2100206. doi:10.1002/pmic.202100206

71. Huang Q, Szklarczyk D, Wang M, Simonovic M, von Mering C. PaxDb 5.0: curated protein quantification data suggests adaptive proteome changes in yeasts. *Molecular & Cellular Proteomics*. Published online August 31, 2023:100640. doi:10.1016/j.mcpro.2023.100640

72. Vogel C, Bashton M, Kerrison ND, Chothia C, Teichmann SA. Structure, function and evolution of multidomain proteins. *Current Opinion in Structural Biology*. 2004;14(2):208-216. doi:10.1016/j.sbi.2004.03.011

73. Paysan-Lafosse T, Andreeva A, Blum M, et al. The Pfam protein families database: embracing AI/ML. *Nucleic Acids Res*. 2025;53(D1):D523-D534. doi:10.1093/nar/gkae997

74. Wang J, Chitsaz F, Derbyshire MK, et al. The conserved domain database in 2023. *Nucleic Acids Res*. 2023;51(D1):D384-D388. doi:10.1093/nar/gkac1096

75. Waman VP, Bordin N, Alcraft R, et al. CATH 2024: CATH-AlphaFlow Doubles the Number of Structures in CATH and Reveals Nearly 200 New Folds. *Journal of Molecular Biology*. 2024;436(17):168551. doi:10.1016/j.jmb.2024.168551

76. Blum M, Andreeva A, Florentino LC, et al. InterPro: the protein sequence classification resource in 2025. *Nucleic Acids Res*. 2025;53(D1):D444-D456. doi:10.1093/nar/gkae1082

77. Lyubitelev AV, Nikitin DV, Shaytan AK, Studitsky VM, Kirpichnikov MP. Structure and functions of linker histones. *Biochemistry (Moscow)*. 2016;81(3):213-223. doi:10.1134/S0006297916030032

78. Wiener N. Cybernetics. *Scientific American*. 1948;179(5):14-19.

79. Shah SG, Mandloi T, Kunte P, et al. HISTome2: a database of histone proteins, modifiers for multiple organisms and epidrugs. *Epigenetics & Chromatin*. 2020;13(1):31. doi:10.1186/s13072-020-00354-8

80. Wiśniewski JR, Hein MY, Cox J, Mann M. A "Proteomic Ruler" for Protein Copy Number and Concentration Estimation without Spike-in Standards*. *Molecular & Cellular Proteomics*. 2014;13(12):3497-3506. doi:10.1074/mcp.M113.037309

81. Palii CG, Cheng Q, Gillespie MA, et al. Single-Cell Proteomics Reveal that Quantitative Changes in Co-expressed Lineage-Specific Transcription Factors Determine Cell Fate. *Cell Stem Cell*. 2019;24(5):812-820.e5. doi:10.1016/j.stem.2019.02.006

82. Baskar R, Chen AF, Favaro P, et al. Integrating transcription-factor abundance with chromatin accessibility in human erythroid lineage commitment. *Cell Rep Methods*. 2022;2(3):100188. doi:10.1016/j.crmeth.2022.100188

83. Shinohara K, Toné S, Ejima T, Ohigashi T, Ito A. Quantitative Distribution of DNA, RNA, Histone and Proteins Other than Histone in Mammalian Cells, Nuclei and a Chromosome at High Resolution Observed by Scanning Transmission Soft X-Ray Microscopy (STXM). *Cells*. 2019;8(2):164. doi:10.3390/cells8020164

84. Hock R, Furusawa T, Ueda T, Bustin M. HMG chromosomal proteins in development and disease. *Trends in Cell Biology*. 2007;17(2):72-79. doi:10.1016/j.tcb.2006.12.001

85. Holehouse AS, Alberti S. Molecular determinants of condensate composition. *Molecular Cell*. 2025;85(2):290-308. doi:10.1016/j.molcel.2024.12.021

86. Miao J, Chong S. Roles of intrinsically disordered protein regions in transcriptional regulation and genome organization. *Current Opinion in Genetics & Development*. 2025;90:102285. doi:10.1016/j.gde.2024.102285

87. Zanzoni A, Ribeiro DM, Brun C. Understanding protein multifunctionality: from short linear motifs to cellular functions. *Cell Mol Life Sci*. 2019;76(22):4407-4412. doi:10.1007/s00018-019-03273-4

88. Kumar M, Michael S, Alvarado-Valverde J, et al. ELM—the Eukaryotic Linear Motif resource—2024 update. *Nucleic Acids Res*. 2024;52(D1):D442-D455. doi:10.1093/nar/gkad1058

89. Ghitti M, Colley LS, Mantonico MV, Musco G, Bianchi ME. Intrinsic disorder and fuzzy interactions drive multiple functions of HMGB1. *Trends in Biochemical Sciences*. Published online September 1, 2025. doi:10.1016/j.tibs.2025.08.001

90. Hatos A, Monzon AM, Tosatto SCE, Piovesan D, Fuxreiter M. FuzDB: a new phase in understanding fuzzy interactions. *Nucleic Acids Res*. 2022;50(D1):D509-D517. doi:10.1093/nar/gkab1060

91. Jonas F, Navon Y, Barkai N. Intrinsically disordered regions as facilitators of the transcription factor target search. *Nat Rev Genet*. 2025;26(6):424-435. doi:10.1038/s41576-025-00816-3

92. Már M, Nitsenko K, Heidarsson PO. Multifunctional Intrinsically Disordered Regions in Transcription Factors. *Chemistry*. 2023;29(21):e202203369. doi:10.1002/chem.202203369

93. Sabari BR, Dall'Agnese A, Young RA. Biomolecular Condensates in the Nucleus. *Trends in Biochemical Sciences*. 2020;45(11):961-977. doi:10.1016/j.tibs.2020.06.007

94. Requião RD, Fernandes L, Souza HJA de, Rossetto S, Domitrovic T, Palhano FL. Protein charge distribution in proteomes and its impact on translation. *PLOS Computational Biology*. 2017;13(5):e1005549. doi:10.1371/journal.pcbi.1005549

95. Fisher RS, Elbaum-Garfinkle S. Tunable multiphase dynamics of arginine and lysine liquid condensates. *Nat Commun*. 2020;11(1):4628. doi:10.1038/s41467-020-18224-y

96. Hong Y, Najafi S, Casey T, Shea JE, Han SI, Hwang DS. Hydrophobicity of arginine leads to reentrant liquid-liquid phase separation behaviors of arginine-rich proteins. *Nat Commun*. 2022;13(1):7326. doi:10.1038/s41467-022-35001-1

97. Dang M, Li T, Zhou S, Song J. Arg/Lys-containing IDRs are cryptic binding domains for ATP and nucleic acids that interplay to modulate LLPS. *Commun Biol*. 2022;5(1):1315. doi:10.1038/s42003-022-04293-w

98. Amaro RE, Åqvist J, Bahar I, et al. The need to implement FAIR principles in biomolecular simulations. *Nat Methods*. 2025;22(4):641-645. doi:10.1038/s41592-025-02635-0

99. Armeev GA, Kniazeva AS, Komarova GA, Kirpichnikov MP, Shaytan AK. Histone dynamics mediate DNA unwrapping and sliding in nucleosomes. *Nat Commun*. 2021;12. doi:10.1038/s41467-021-22636-9

100. Fedulova AS, Armeev GA, Romanova TA, et al. Molecular dynamics simulations of nucleosomes are coming of age. *WIREs Computational Molecular Science*. 2024;14(4):e1728. doi:10.1002/wcms.1728

101. Kilgore HR, Chinn I, Mikhael PG, et al. Protein codes promote selective subcellular compartmentalization. *Science*. Published online March 7, 2025. doi:10.1126/science.adq2634

102. Yang X, Zhu H, Shi L, et al. AlphaFold-guided structural analyses of nucleosome binding proteins. *Nucleic Acids Res*. 2025;53(14):gkaf735. doi:10.1093/nar/gkaf735

103. Lim Y, Tamayo-Orrego L, Schmid E, et al. In silico protein interaction screening uncovers DONSON's role in replication initiation. *Science*. 2023;381(6664):eadi3448. doi:10.1126/science.adi3448

104. Ruthenburg AJ, Li H, Patel DJ, David Allis C. Multivalent engagement of chromatin modifications by linked binding modules. *Nat Rev Mol Cell Biol*. 2007;8(12):983-994. doi:10.1038/nrm2298

Supplementary materials

(Re)defining the human chromatome: an integrated meta-analysis of localization, function, abundance, physical properties and domain composition of chromatin proteins

Anna K. Gribkova[1,2], Grigoriy A. Armeev[1,2], Mikhail P. Kirpichnikov[1,3], Alexey K. Shaytan[1,2,4*]

[1] Department of Biology, Lomonosov Moscow State University, Moscow, Russia

[2] Vavilov Institute of General Genetics, Moscow, Russia

[3] Shemyakin–Ovchinnikov Institute of Bioorganic Chemistry,

Russian Academy of Sciences, Moscow, Russia

[4] International Laboratory of Bioinformatics, AI and Digital Sciences Institute,

Faculty of Computer Science, HSE University, Moscow, Russia

* To whom correspondence should be addressed. Email: shaytan_ak@mail.bio.msu.ru

Table of contents

Supplementary Tables

Supplementary Table ST1. Representative list of chromatome datasets from whole proteome projects, databases, literature sources and computational predictions about chromatin or nuclear proteins. A short description and the number of proteins is provided for every source. Columns: Source of information; DB, link, data availability; Description of source of information, details of experiment; Number of proteins (machine readable), Year; Reference; Doi; Notes about number of proteins (human readable); Number of processed proteins.

Supplementary Table ST2. The evaluation metrics for all localized, nuclear and subnuclear (nucleolus, nuclear envelope, nuclear bodies, nucleoplasm) proteins from UniProt (A) and HPA (B). Metrics include the number of proteins with overlapping annotation, union of annotations, Jaccard similarity coefficient (measuring overlap), False Positives, False Negatives, and performance measures (precision, recall, and F1-score).

Supplementary Table ST3. Localization terms from UniProt and HPA which were grouped into generalized localization categories: nuclear categories (nucleus, nucleolus, nuclear envelope, nuclear bodies, nucleoplasm), cytoplasm, endomembrane system and others (in UniProt).

Supplementary Table ST4. SimChrom classification terms and details about their protein contents: corresponding GO terms, databases and literature sources, additional filters. Columns: SimChrom term, GO terms used to build the SimChrom DB, Explanation of the selected GO terms to build SimChrom DB, Number of proteins initially extracted from GO without proteins in nested classes (before additional filtration), Databases and literature sources used to add proteins and explanation, Additional filtering applied, Total number of proteins (with proteins in nested classes), Additional notes.

Supplementary Table ST5. The SimChrom ontology and SimChrom/SimChrom-SL classification of chromatin proteins. The list of SimChrom protein categories is given together with the aspects that they represent, and the lists of proteins belonging to each category given via UniProt IDs. Total number of proteins per category is also given.

Supplementary Table ST6. The constructed protein localization reference datasets, their definition and the corresponding number of protein entries. Abbreviations: NULOC - nuclear localization, NON_NULOC - non-nuclear localization, CYTLOC - cytoplasmic localization (see definition in **Supplementary Table ST3**), UL - uniquely localized, CS - consensus (supported by several sources), JT - joint (supported by at least one source), NECF - no evidence code filtering (localization data with any evidence level from the databases was used).

Supplementary Table ST7. GO enrichment analysis for the set of protein entries from MS-based chromatomes that are absent in NULOC_JT_NECF and SimChrom (n=2232). Only the driver GO terms that were highlighted by g:Profiler were included. Columns: source, term_name, term_id, highlighted, adjusted_p_value, negative_log10_of_adjusted_p_value, term_size, query_size, intersection_size, effective_domain_size, intersections.

Supplementary Table ST8. GO enrichment analysis of SimChrom-exclusive proteins (absent in NULOC_CS), n=1208. Only the driver GO terms that were highlighted by g:Profiler were included. Columns: source, term_name, term_id, highlighted, adjusted_p_value, negative_log10_of_adjusted_p_value, term_size, query_size, intersection_size, effective_domain_size, intersections, unexpected_for_chromatin.

Supplementary Table ST9. GO enrichment analysis of NULOC_CS-exclusive proteins (absent in SimChrom), n=1459. Only the driver GO terms that were highlighted by g:Profiler were included. Columns: source, term_name, term_id, highlighted, adjusted_p_value, negative_log10_of_adjusted_p_value, term_size, query_size, intersection_size, effective_domain_size, intersections.

Supplementary Table ST10. The number and fraction of housekeeping (HK)/ non-housekeeping (non-HK), low-abundant (LA) and high-abundant (HA) proteins in the whole proteome (PaxDB_INT), NULOC_JT, NULOC_CS, SimChrom and MS-based chromatomes and a nucleome.

Supplementary Table ST11. A list of SimChrom-SL proteins annotated with "RNA processing" category that are uniquely localized in the nucleus (NULOC_CS_UL). Columns: Gene, Entry, Protein names, localization by UniProt, HPA, OpenCell with filtration by evidence codes (see Methods), protein identification in MS-based studies, abundance by PaxDB_PA (ppm).

Supplementary Table ST12. Abundance data for all histone proteins and non-histone chromatin proteins with abundance values of more than 0.01% of Histone H4. Columns: SimChrom-SL, Gene, Abundance relative to histone H4 (%), Number of nucleosomes per protein, Abundance PaxDb_PA (ppm), Entry, Uniquely localized in nucleus, Details (for **Figure 4D**).

Supplementary Table ST13. Comparison of amino acid composition between chromatin and Cytoplasm^{UL} proteins. Table includes columns: Feature (fraction of amino acid), Sequence (total, IDR, non-IDR), median fraction for chromatin and cytoplasmic proteins in percent, fold enrichment of median values for SimChrom vs Cytoplasm^{UL}, statistically significance according to Mann Uitney test with multiple test correction (TRUE, FALSE), fold enrichment in IDRs vs non-IDRs of SimChrom, Fold enrichment in IDRs vs non-IDRs of Cytoplasm.

Supplementary Table ST14. List of uncharacterized structural domains in chromatin proteins. Columns: Annotated in InterPro (YES/NO), Entry, TED sequence number, ted_range, ted_len, pLDDT, packing_density, norm_rg, consensus_level, num_segments, num_helix_strand_turn, num_helix_strand, num_helix, num_strand, num_turn, TED_link, InterPro_link, SimChrom-SL, ted_id. See also **Interactive Table 3** (https://simchrom.intbio.org/#novel_structural_domains).

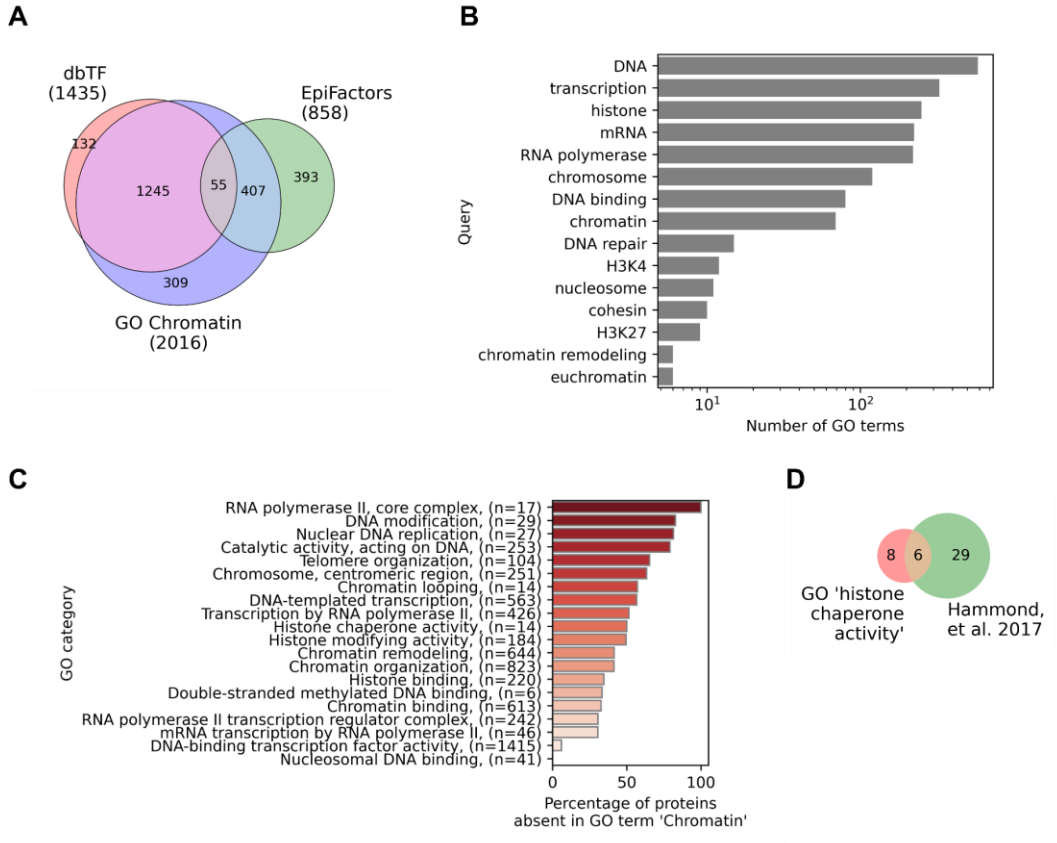
Supplementary Table ST15. List of Pfam domain models in chromatin proteins. Columns: Entry, SimChrom-SL, Total number of Pfam domains, Number of distinct Pfam models, pfamA_ids (only unique ids), Domain annotations (according to manual annotation of EMVI-domains, see **Methods**), Distinct domain annotations (only unique annotations).

Supplementary Table ST16. Pfam domain models that are present in three or more chromatin/epigenetic regulator proteins. Columns: pfamA_id, pfamA_acc, pfamA_name, domain model functional subgroup, domain model functional group, SimChrom-SL ChromReg (the category which contains a large fraction of proteins with the domain model, where available), number of SimChrom proteins with the domain model, number of ChromReg proteins with the domain model, number of co-occurring Pfam domain models in SimChrom proteins, number of co-occurring Pfam domain models in ChromReg proteins, number of SimChrom proteins (where the Pfam domain models co-occurs with other domain models), number of ChromReg proteins (where the Pfam domain model co-occurs with other domain models).

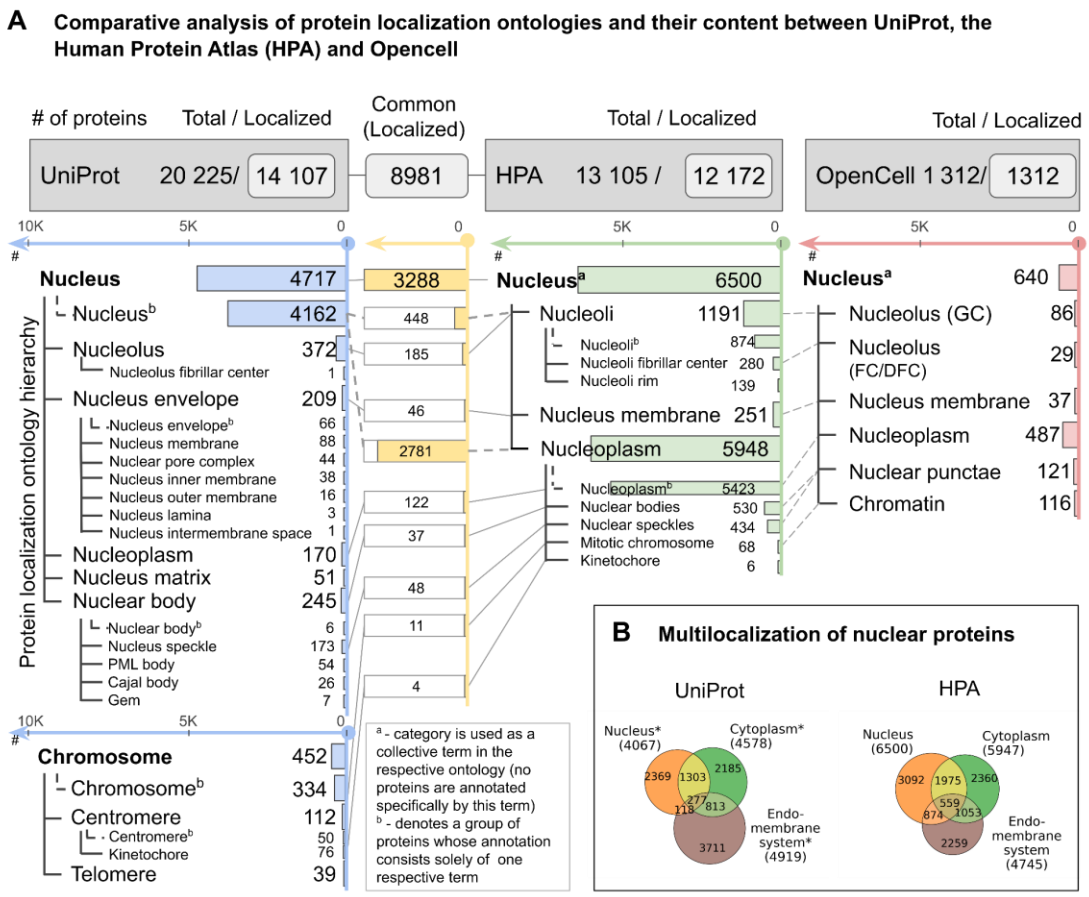
Supplementary Table ST17. The conditional probability of EMVI-domains to co-occur in chromatin proteins, grouped by their main functions. The conditional probability of finding a corresponding domain A in a chromatin protein given that another domain B is already present (columns and rows correspond to domains A and B, respectively).

Supplementary Figures

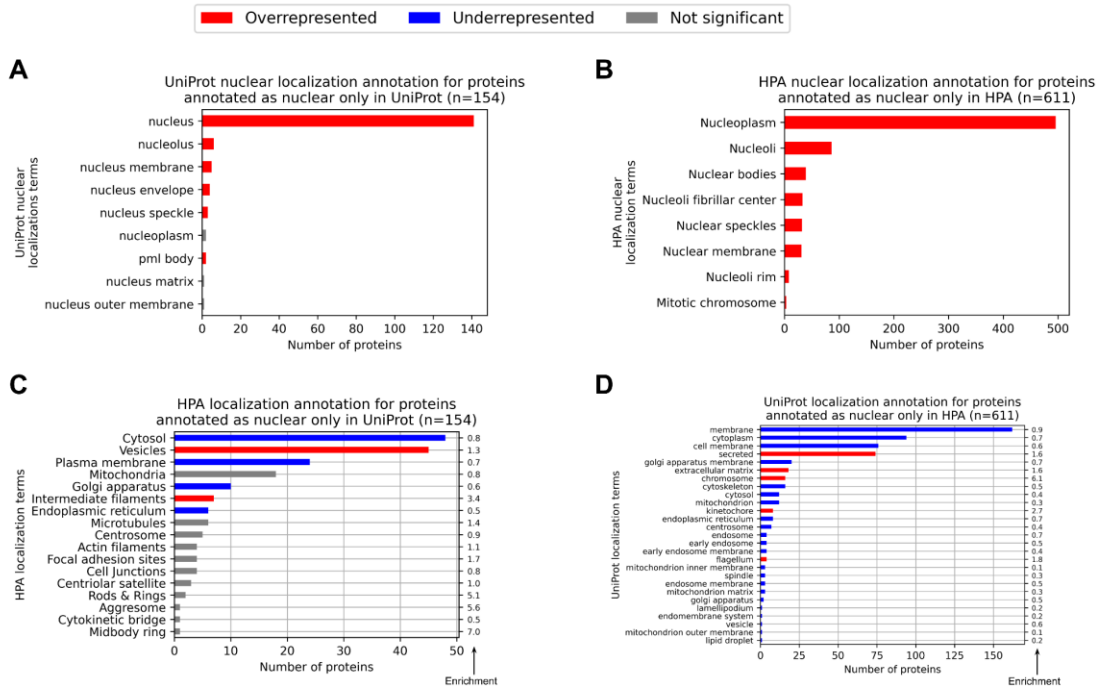
1. Sources of information about chromatin and nuclear proteins and their critical evaluation



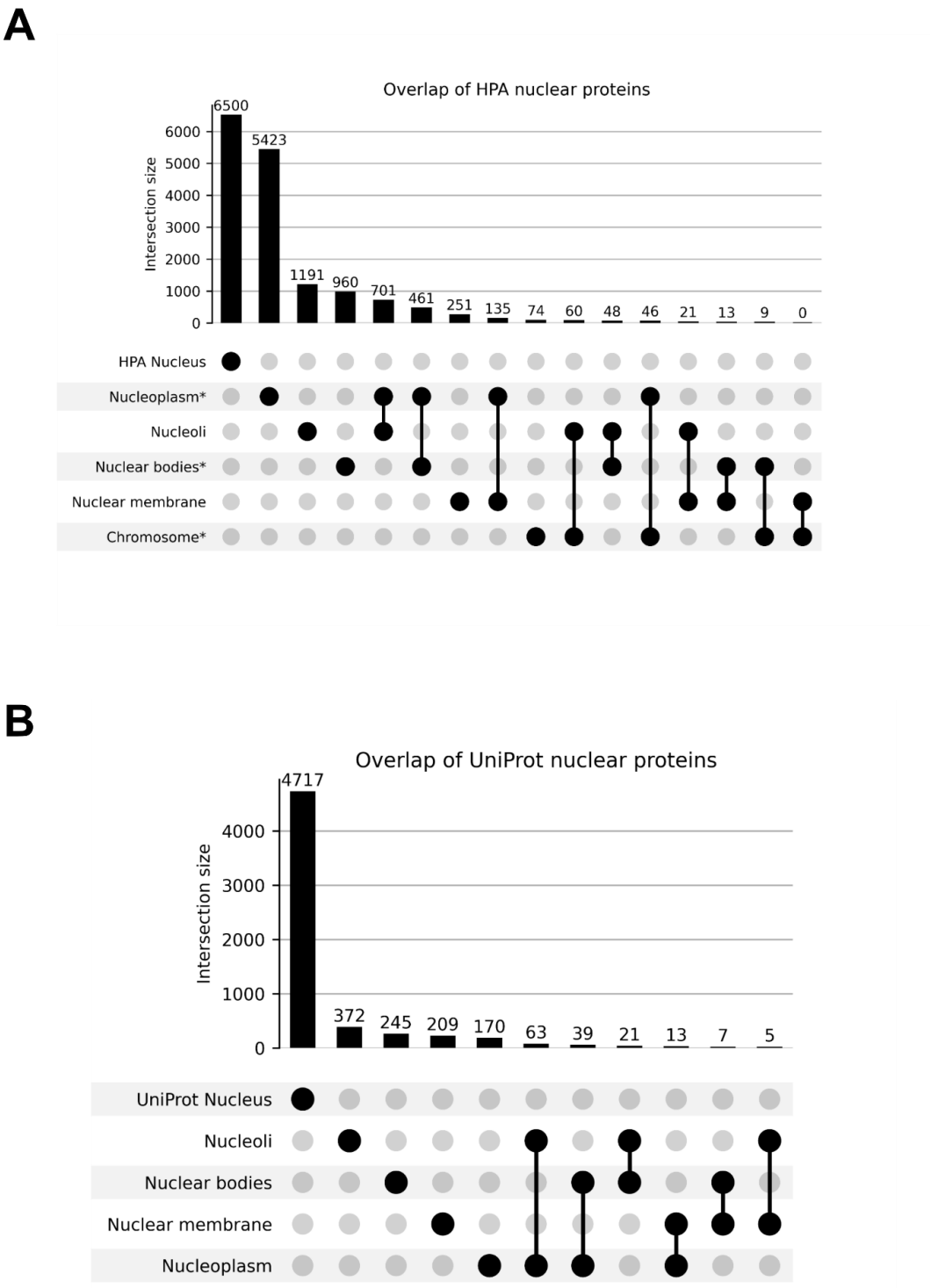
Supplementary Figure SF2_1. (**A**) The Venn diagram of protein repertoires from EpiFactors (epigenetic regulators), the GO catalogue of DNA-binding transcription factors, and GO term 'Chromatin'. (**B**) The number of GO terms which contain chromatin-associated keywords. (**C**) The fraction of proteins from suggested chromatin-associated categories and proteins annotated by the GO term 'Chromatin'. (**D**) The Venn diagram for histone chaperone proteins: proteins annotated by the GO term "histone chaperone activity" vs literature review by Hammond et al., 2017 [1].



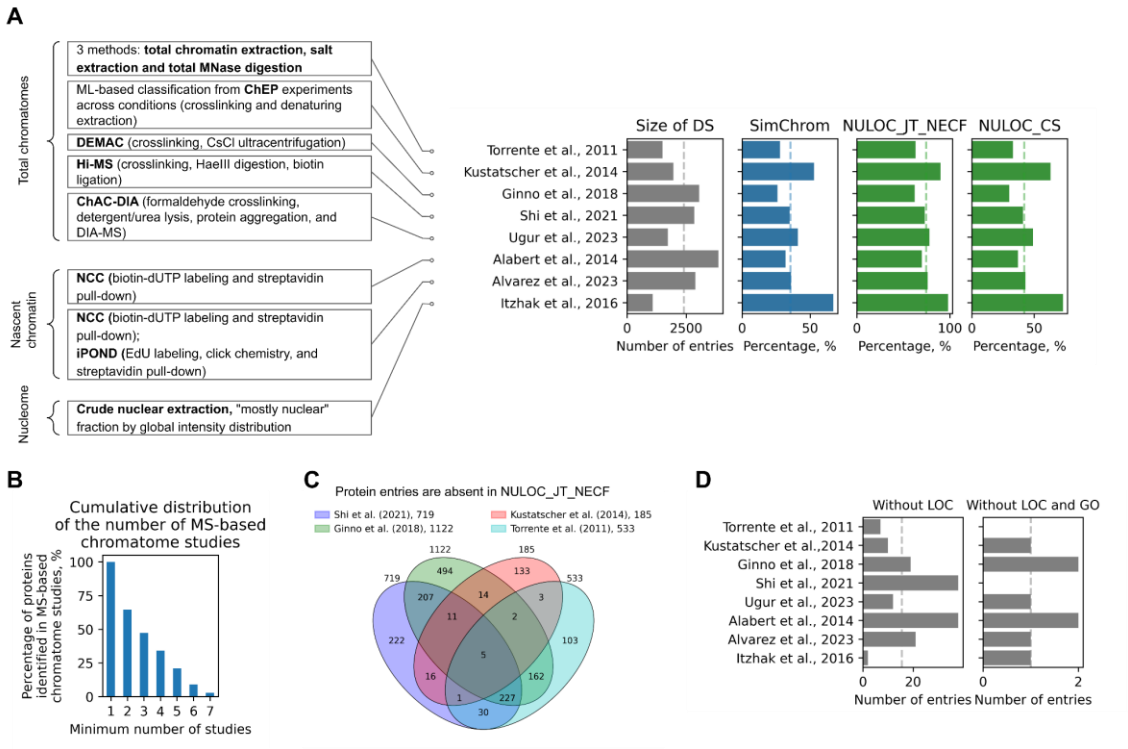
Supplementary Figure SF2_2. Nuclear proteome content and protein subnuclear localization according to three major databases/proteome-wide experimental initiatives: UniProt, HPA, OpenCell. See interactive version in **Interactive Figure 2** at https://simchrom.intbio.org/#localization. (**A**) A detailed comparative analysis of human nuclear proteome content and its subnuclear localization according to respective hierarchical ontologies between UniProt (blue), the Human Protein Atlas (HPA) (green) and Opencell (red). The scheme shows there ontologies for each localization annotation system, the number of proteins in each category (to the right of each category name), and the number of common proteins between different categories of UniProt and HPA (highlighted in yellow). Note that a protein may be present in several localization categories if it has multiple sublocalizations. On the top of the scheme the total number of human proteins in the database (total) and the number of proteins that have localization information supported by sufficient evidence (localized) is reported (see Methods) for the definition of the criteria). (**B**) A Venn diagram showing the overlap between the sets of nuclear, cytoplasmic and endomembrane proteins as defined by UniProt or HPA. UniProt proteins were considered with only three grouped localization tags; proteins with others localization were not considered.



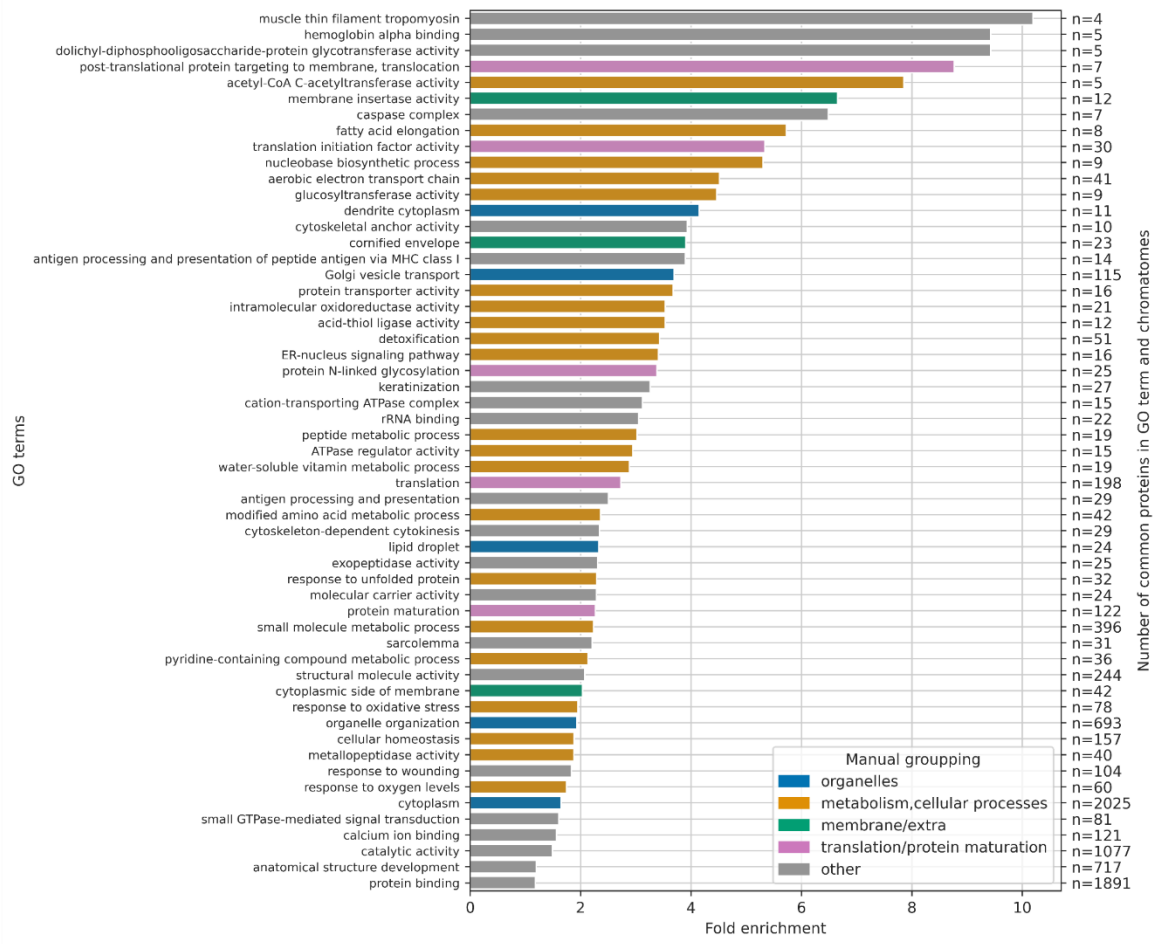
Supplementary Figure SF2_3. A cross-annotation analysis of localization terms for proteins annotated as nuclear by one resource - either UniProt or HPA, while having only non-nuclear localization annotations in the other resource. To identify significantly over- or under-represented localization terms, an enrichment analysis using the hypergeometric test (p-value < 0.05) was performed, see **Methods Section 2.1.1**. (**A**) The number of nuclear proteins in the UniProt with their subnuclear localization terms that lack nuclear localization annotation in HPA. (**B**) The number of nuclear proteins in the HPA with their subnuclear localization terms that lack nuclear localization annotation in UniProt. (**C**) The number of nuclear proteins in the UniProt that are non-nuclear in HPA, along with their non-nuclear localization annotation in HPA. (**D**) The number of nuclear proteins in the HPA that are non-nuclear in UniProt, along with their non-nuclear localization annotation in UniProt.



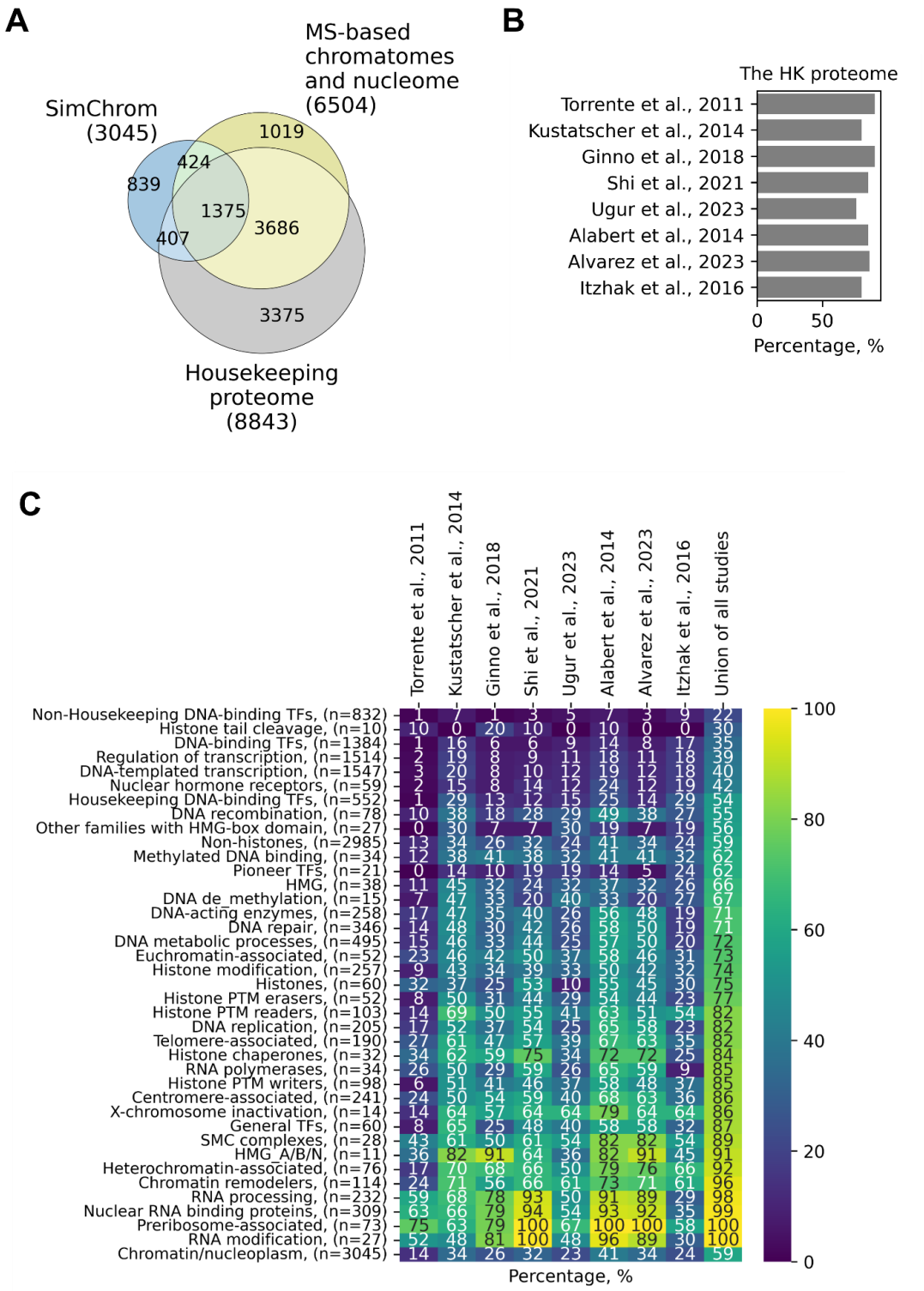
Supplementary Figure SF2_4. UpSet plots comparing sets of proteins according to their subnuclear localization as provided by HPA (**A**) and (**B**) UniProt. The generalized subnuclear localization terms are used as defined in **Supplementary Table ST3**.



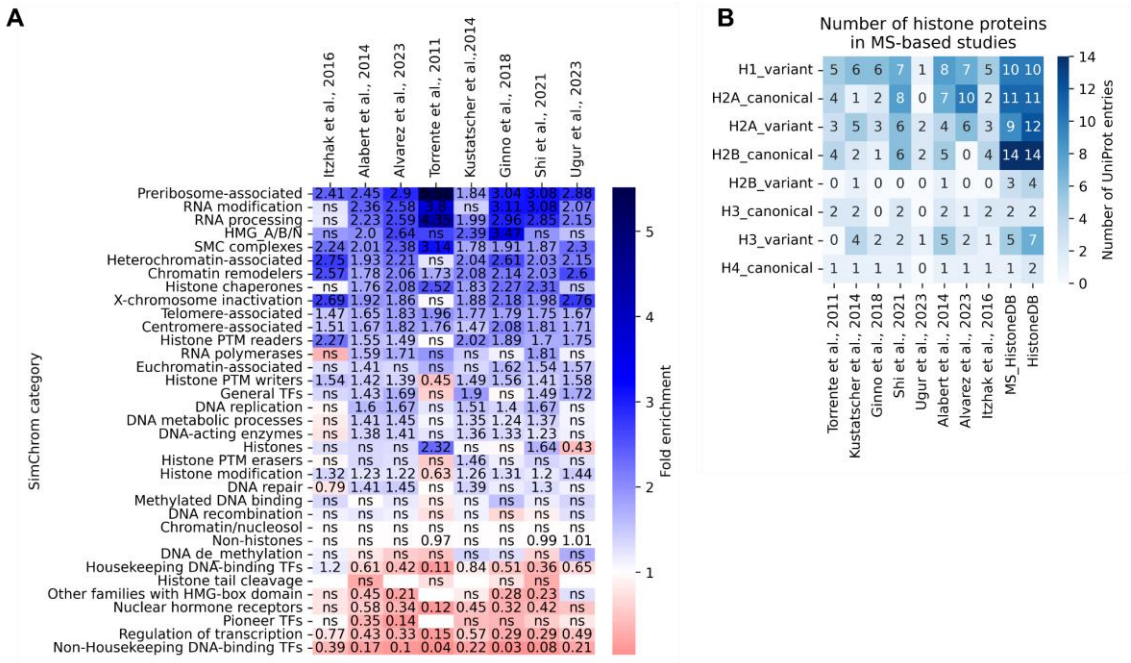
Supplementary Figure SF2_5. The MS-based human chromatome (and nucleome) datasets examined through the lens of annotations provided by the localization databases and the SimChrom chromatin protein classification. (A) The list of MS-based datasets of chromatin/nuclear proteins from the respective studies analyzed in this work together with a short description of the experimental/analysis workflow (left) and the plots (right) showing the size of the datasets and the fractions of the datasets that overlap with the SimChrom dataset or nuclear localization datasets (NULOC_CS and NULOC_JT_NECF). The median values are shown by dotted in the plots. (**B**) Cumulative fraction of proteins identified by at least N MS-based chromatin studies relative to the total number of chromatin proteins identified in at least one MS-based study. (**C**) Overlap of protein entries identified in MS-based studies that lack nuclear localization according to the database annotations (NULOC_JT_NECF dataset). (**D**) Number of protein entries identified in MS-based studies that have no annotations in the databases. Left: proteins lacking localization data in both UniProt and HPA. Right: proteins lacking both localization data and GO annotation.



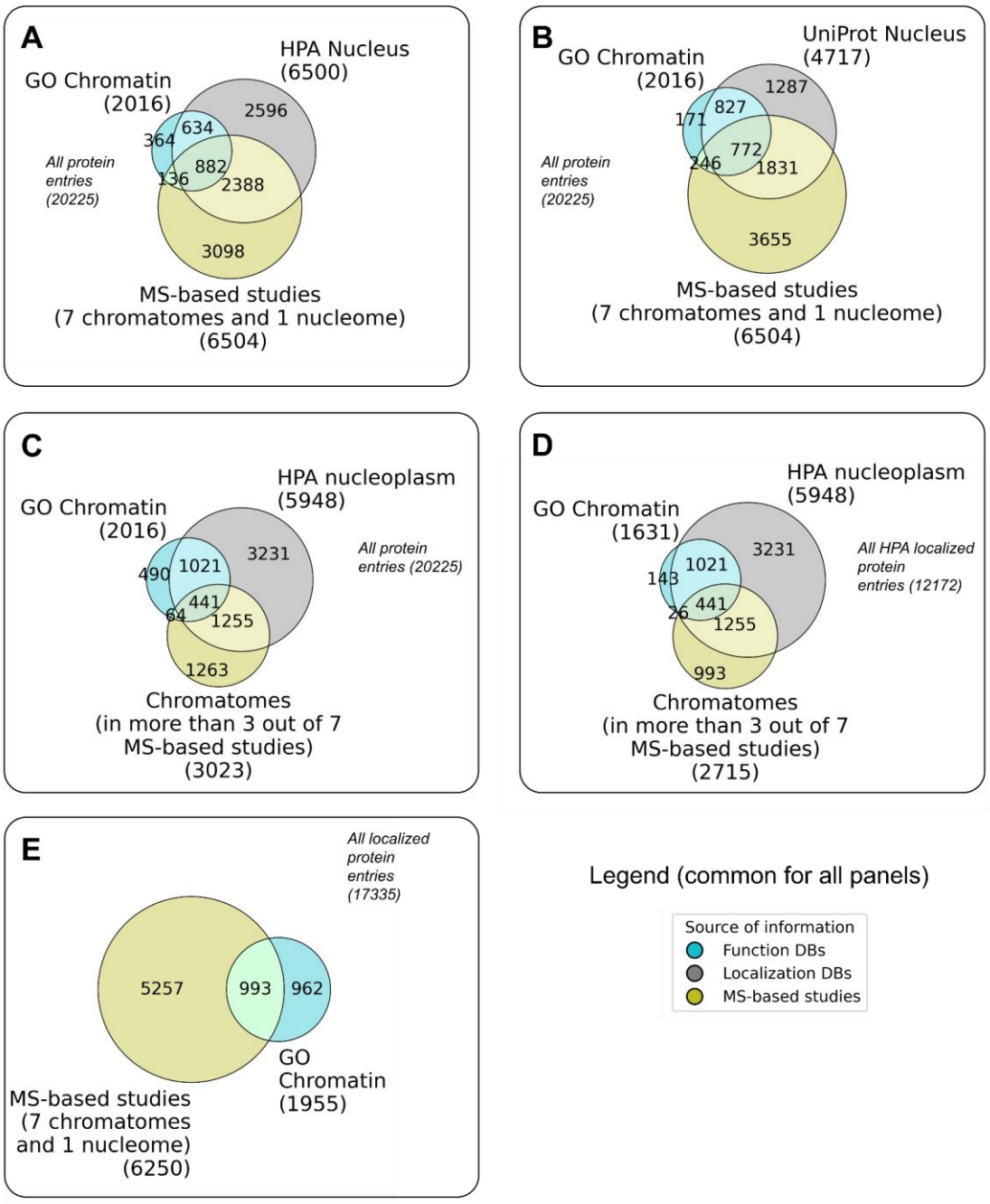
Supplementary Figure SF2_6. GO enrichment analysis for protein entries from MS-based chromatome datasets that are absent from both the NULOC_JT_NECF dataset and the SimChrom classification (n = 2232).



Supplementary Figure SF2_7. (**A**) The Venn diagram showing overlaps between the set of SimChrom proteins, the housekeeping proteome and the combined protein set of MS-based chromatome datasets (union of protein content from eight MS-based studies analyzed in this work). (**B**) The percentage of housekeeping (HK) proteins in MS-based chromatomes and nucleome (range: 76% - 90%). (**C**) The percentage of SimChrom proteins by SimChrom category in MS-based chromatomes and nucleome.

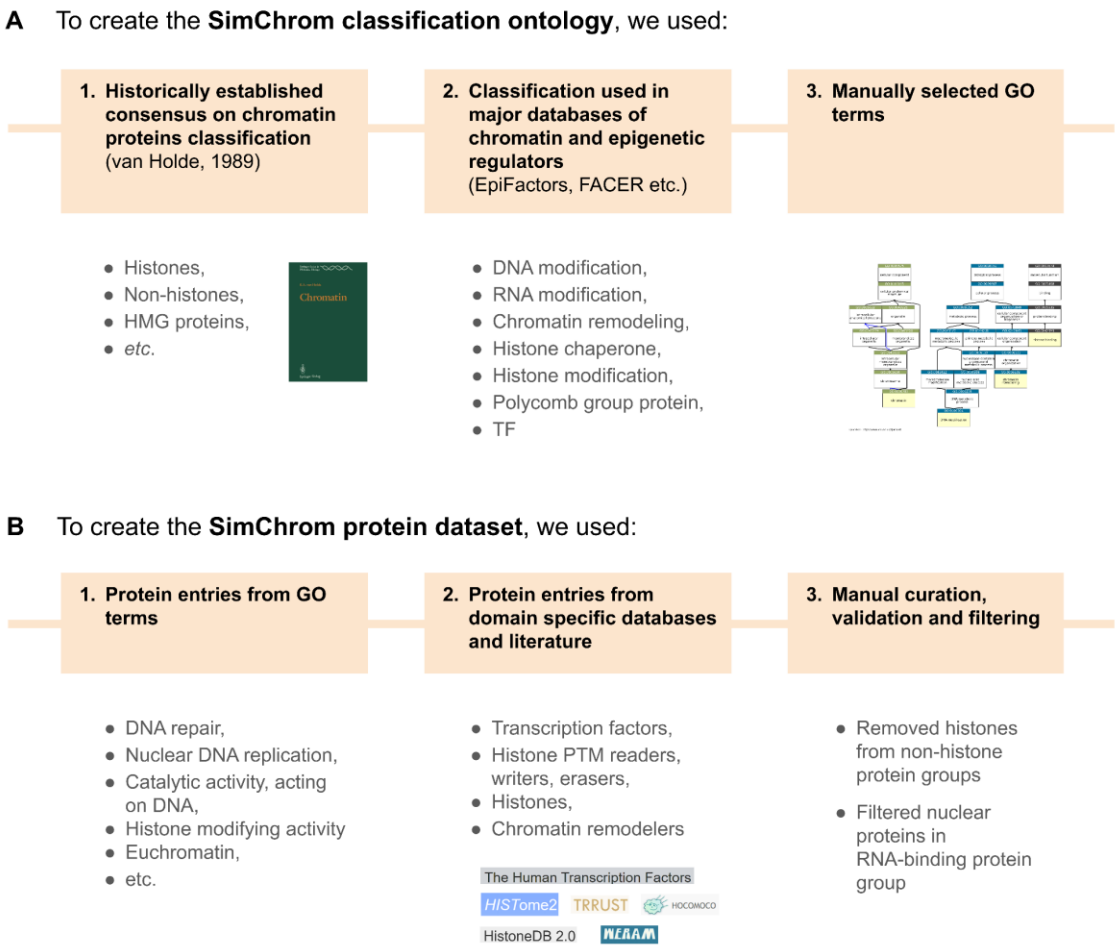


Supplementary Figure SF2_8. (**A**) Fold enrichment of chromatin-associated proteins identified in MS-based studies for every SimChrom category (fold enrichment is calculated with respect to the distribution of proteins among the categories in SimChrom). Only statistically significant values (p-value of Fisher exact test with Benjamini correction < 0.05) are shown. (**B**) Number of histone proteins detected in MS-based studies compared to the reference counts from MS_HistoneDB and HistoneDB 2.0.

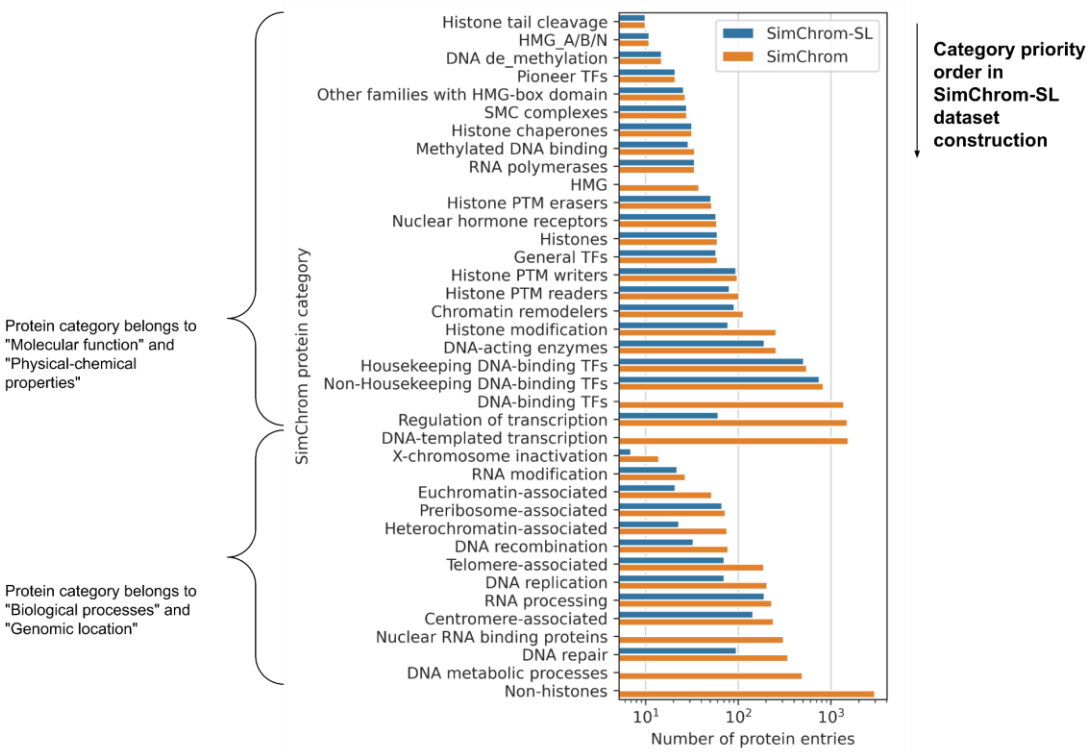


Supplementary Figure SF2_9. The overlap of chromatin/nuclear protein entries from different types of sources: protein function databases (GO "Chromatin"), protein localization DBs (Uniprot Nucleus, HPA Nucleus or Nucleoplasm), MS-based studies of chromatome and nucleome proteins (two protein sets are used - see legends: 1. union of protein entries from five total chromatin studies, two nascent chromatin and one nucleome; 2. protein entries that are present in three out of seven MS-based chromatin datasets). The “background” set of proteins for each panel is shown in italic.

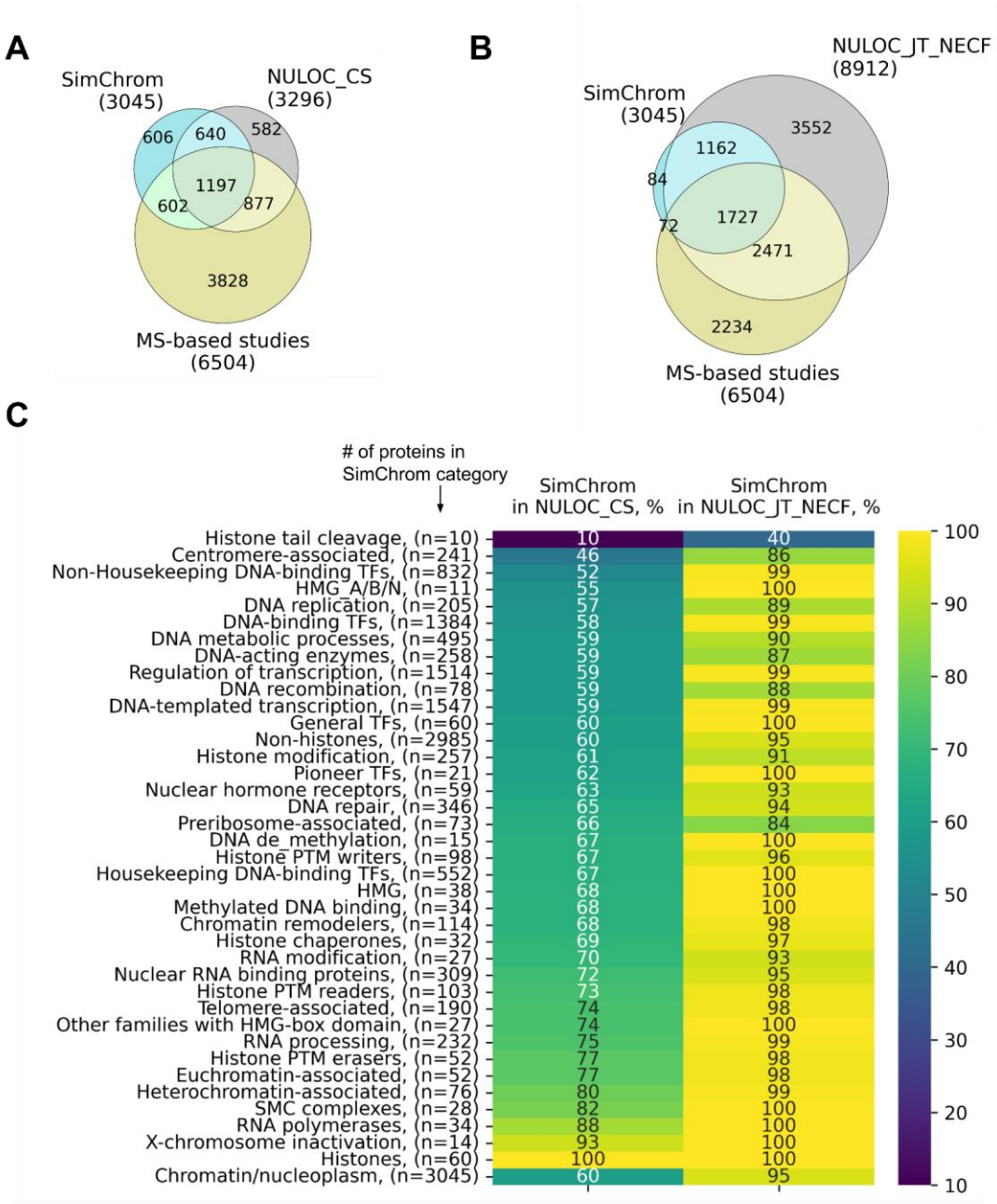
2. The SimChrom chromatin protein classification, the SimChrom dataset and other reference datasets



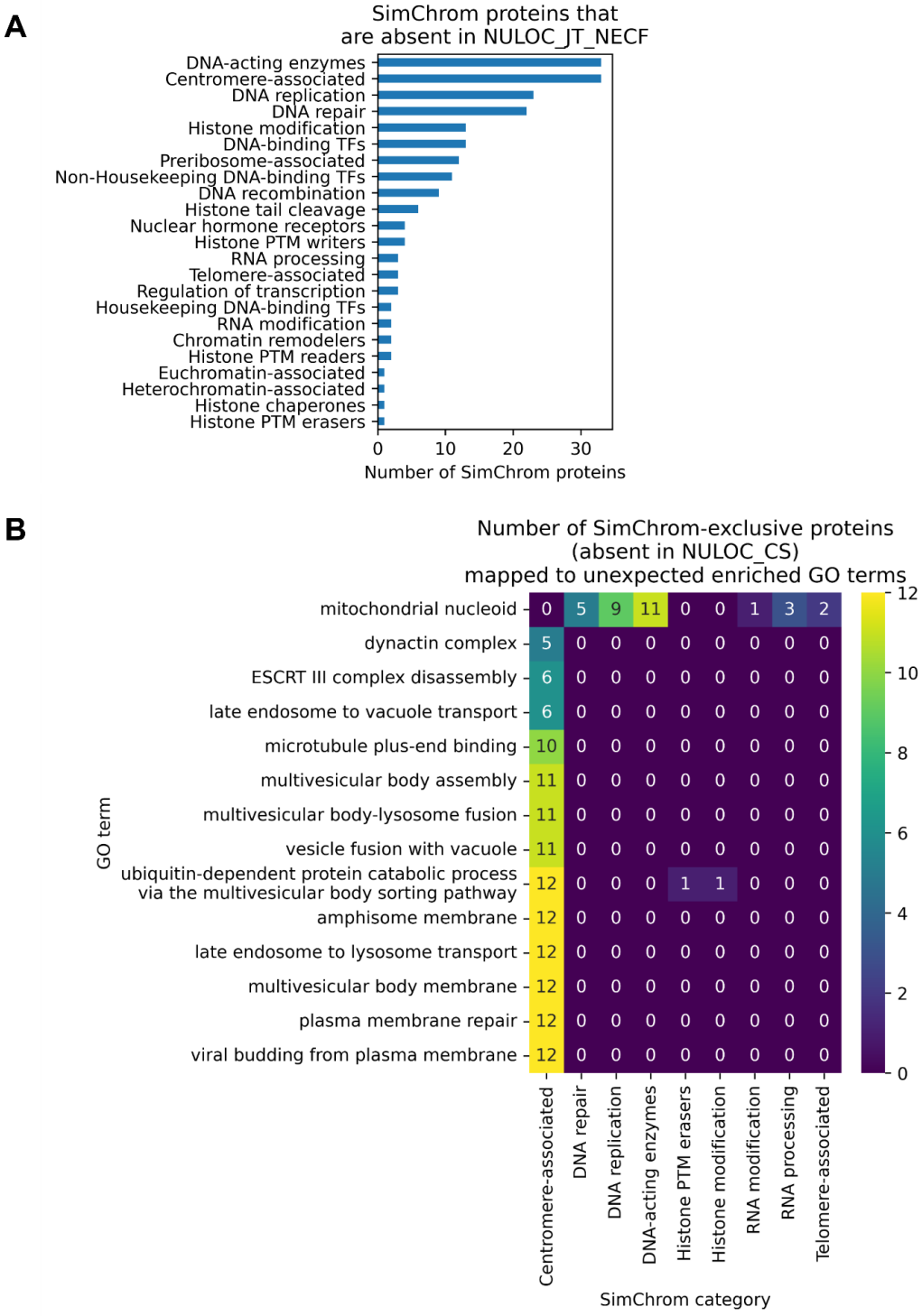
Supplementary Figure SF3_1. The scheme of creation of the SimChrom classification ontology and SimChrom protein dataset is shown in panels (**A**) and (**B**), respectively.



Supplementary Figure SF3_2. The order of SimChrom categories (from top to bottom) used to create the single-label SimChrom-SL classification of chromatin proteins. The categories were ordered as follows: molecular function and physicochemical properties were placed first, followed by the others. Among them, categories containing fewer proteins were ordered earlier. The number of proteins belonging to the respective SimChrom and SimChrom-SL categories is also shown.



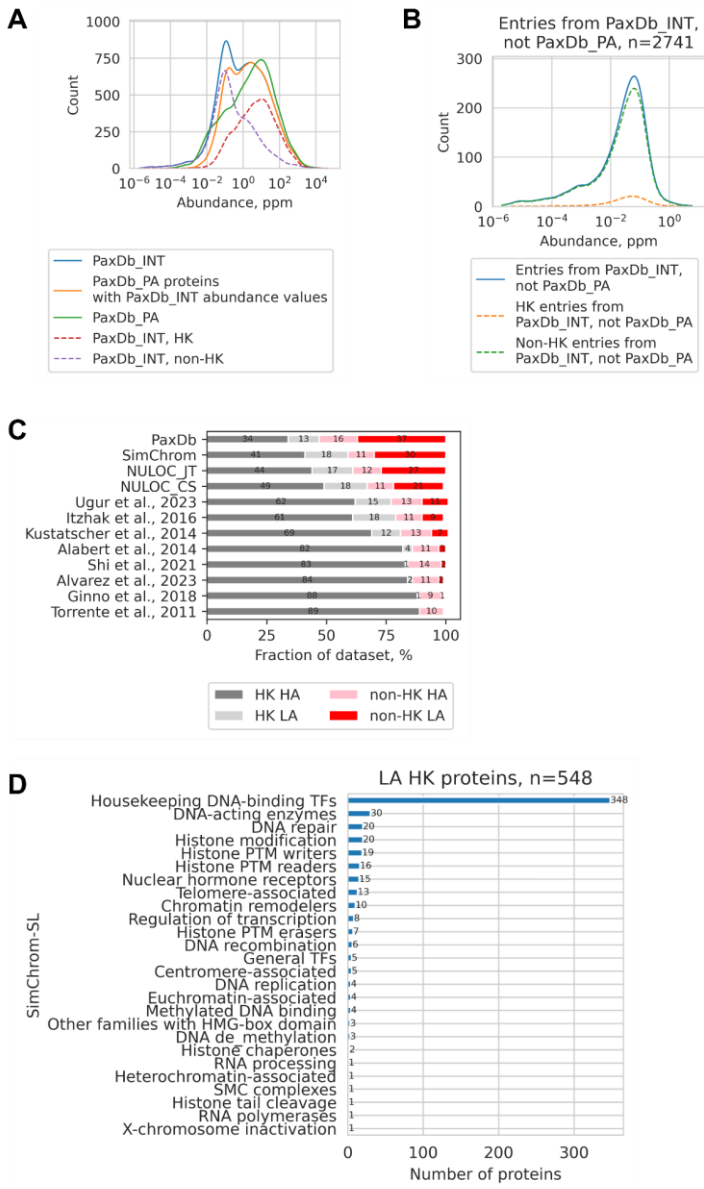
Supplementary Figure SF3_3. SimChrom protein analysis using protein localization information and MS-based chromatomes and nucleome. **(A-B)** A Venn diagram showing the overlap between the protein sets: SimChrom, proteins identified in MS-based studies and the reference nuclear protein sets NULOC_CS (**A**) or NULOC_JT_NECF (**B**). (**C**) The percentage of proteins from SimChrom categories that are found in NULOC_CS and NULOC_JT_NECF datasets.



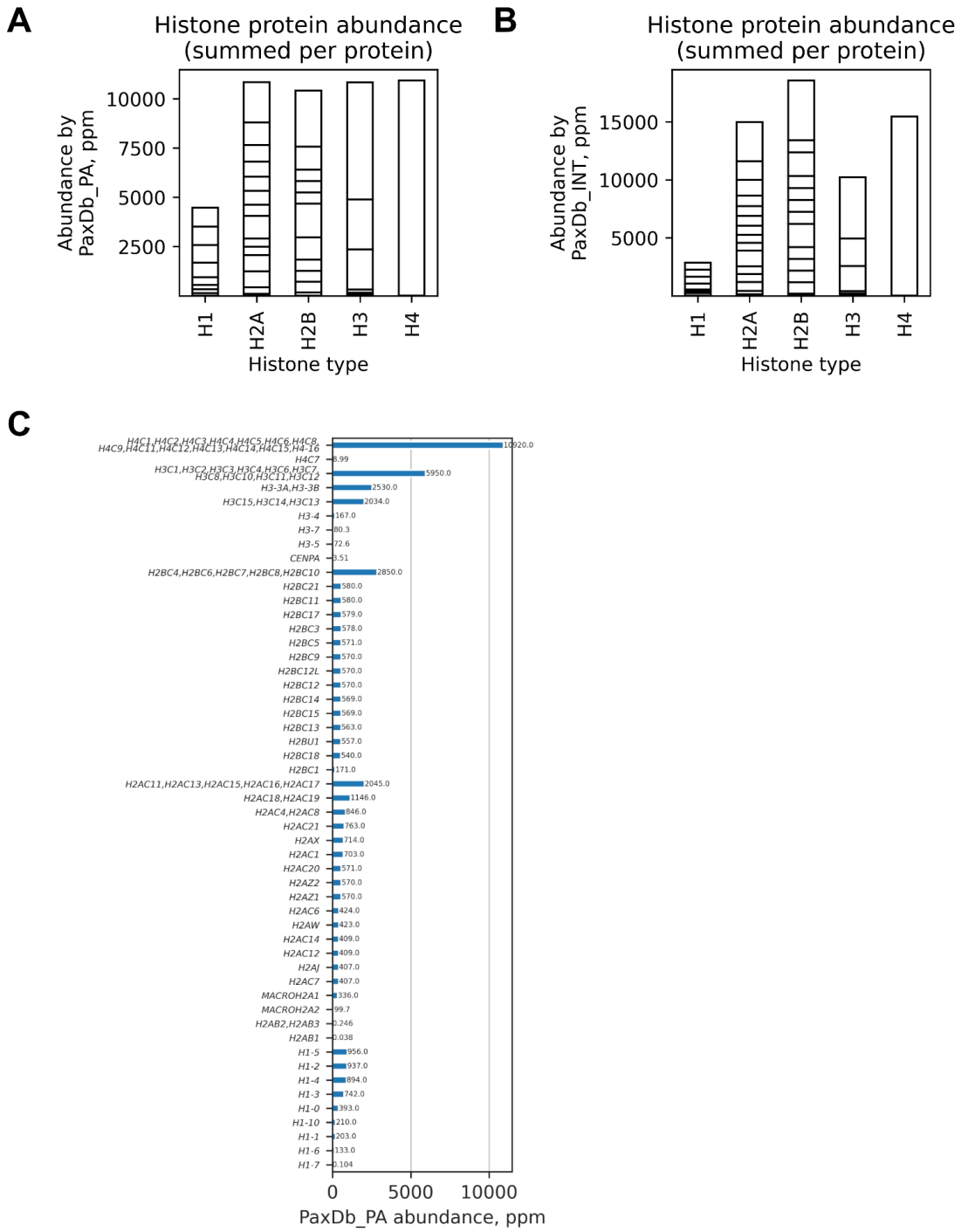
Supplementary Figure SF3_4. SimChrom protein dataset analysis using protein localization information. (**A**) The number of SimChrom-SL classified proteins without nuclear localization according to NULOC_JT_NECF (the broadest dataset that combined all nuclear protein entries from all protein localization databases at any level of confidence). (**B**) The unexpected enriched GO terms for proteins were identified for the SimChrom proteins that are absent in NULOC_CS.

3. Analysis of the human chromatome

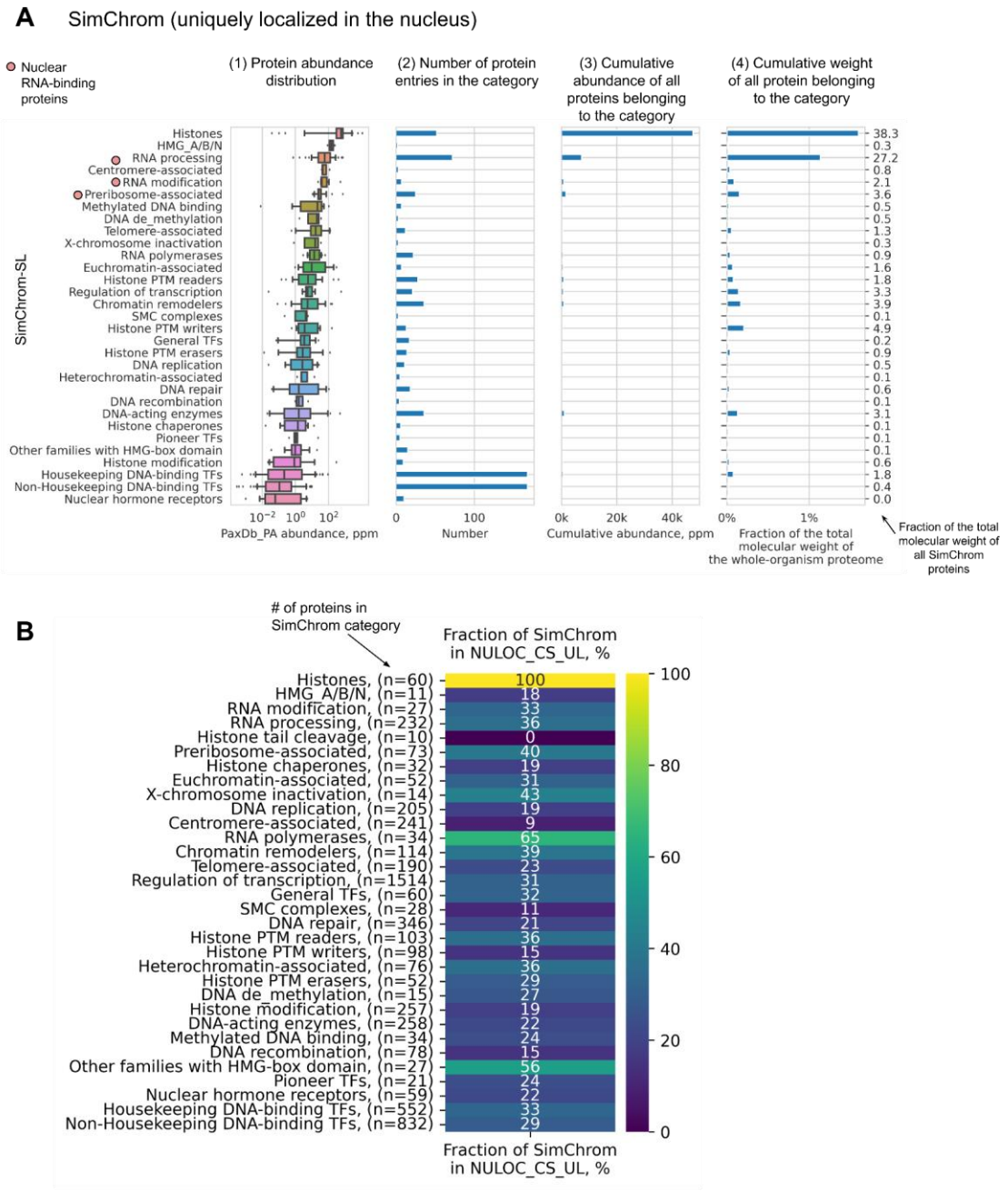
3.1. The chromatome composition and abundance of chromatin proteins



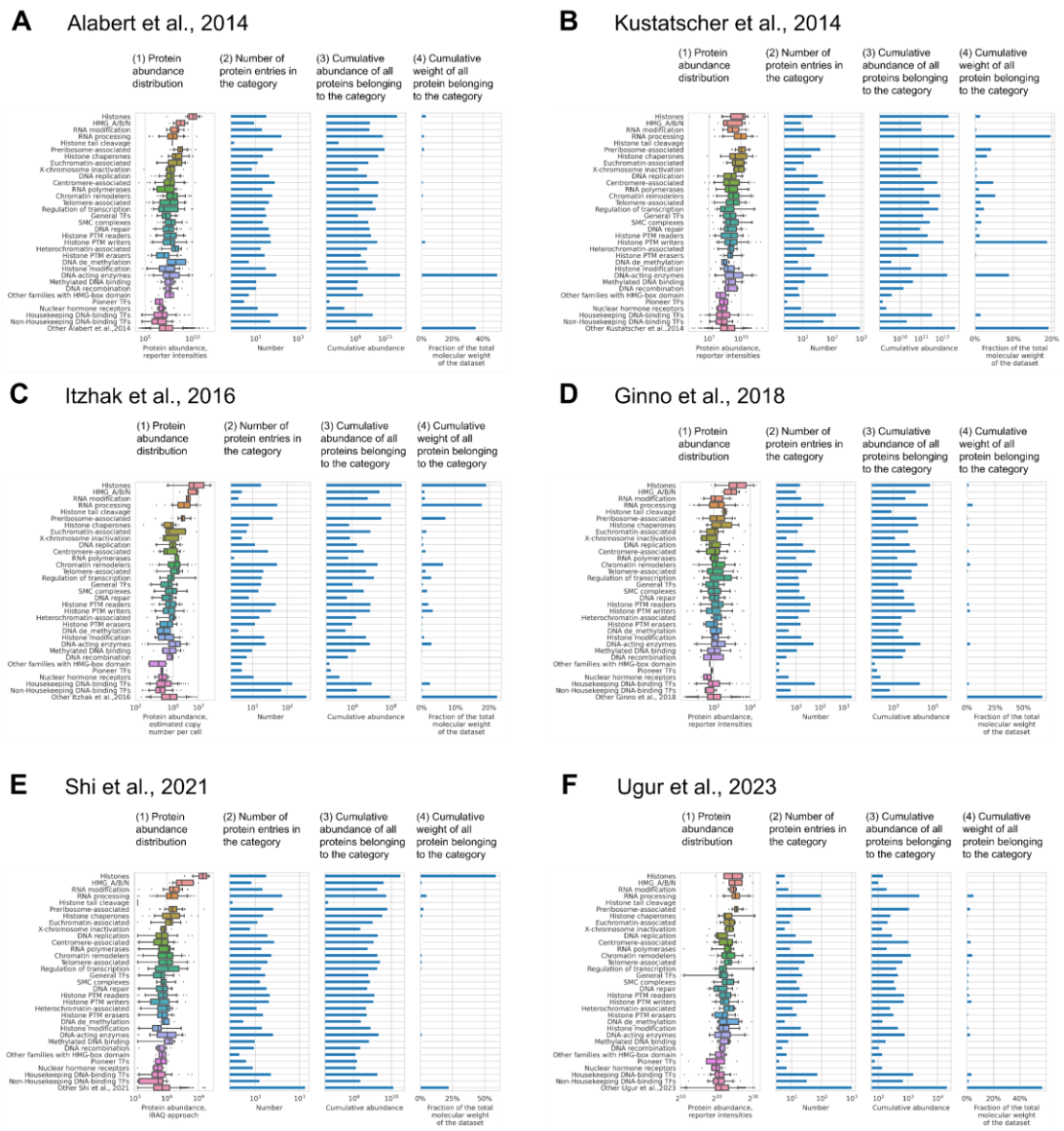
Supplementary Figure SF4_1. Chromatin proteins abundance analysis. (**A, B**) Distribution of proteins from PaxDb_INT and PaxDb_PA datasets according to their relative abundance values. The distribution for housekeeping (HK) and non-housekeeping (non-HK) are also shown (see legend). The distribution was constructed by taking the logarithm of the abundance values in ppm, making a histogram (bin size of 0.15) and smoothing it with a gaussian kernel for visual clarity. (**C**) Fraction distribution of low-abundant (LA) and high-abundant (HA) housekeeping (HK) and non-housekeeping (non-HK) proteins in the whole proteome (PaxDb_INT), protein localization datasets (NULOC_CS and NULOC_JT), and SimChrom and MS-based chromatomes. (**D**) The distribution of low-abundant (LA) housekeeping (HK) proteins among SimChrom-SL categories.



Supplementary Figure SF4_2. Abundance of histone proteins according to PAXDb datasets: PAXDb_PA "Whole organism, SC (Peptideatlas,aug,2014)" (**A**) and PAXDb_INT "H.sapiens - Whole organism (Integrated)" (**B**); ppm means part per million. Abundance of histone proteins grouped by histone types (each histone protein type is composed of one or several proteins that may in turn be encoded by several different genes) by PAXDb_INT. The horizontal lines on the bar plot delineate the contribution of different histone proteins belonging to the respective histone type. (**C**) Abundance of histone proteins according to PaxDB_PA.

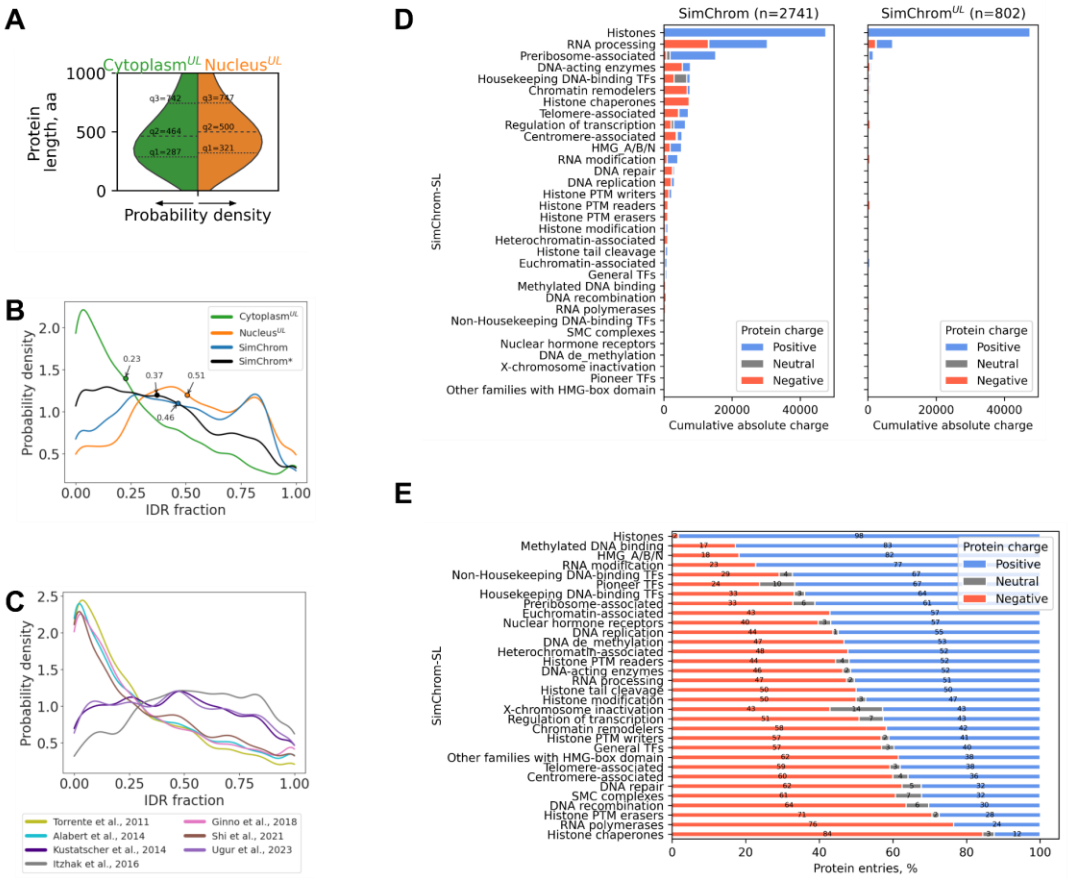


Supplementary Figure SF4_3. (**A**) This is a version of Figure 4C from the main text showing chromatin protein abundance for SimChrom categories, but created for the SimChrom proteins that are uniquely localized in the nucleus according to the NULOC_CS_UL dataset. (**B**) The fractions of proteins belonging to SimChrom categories (based on standard SimChrom classification) that are also present in NULOC_CS_UL dataset.

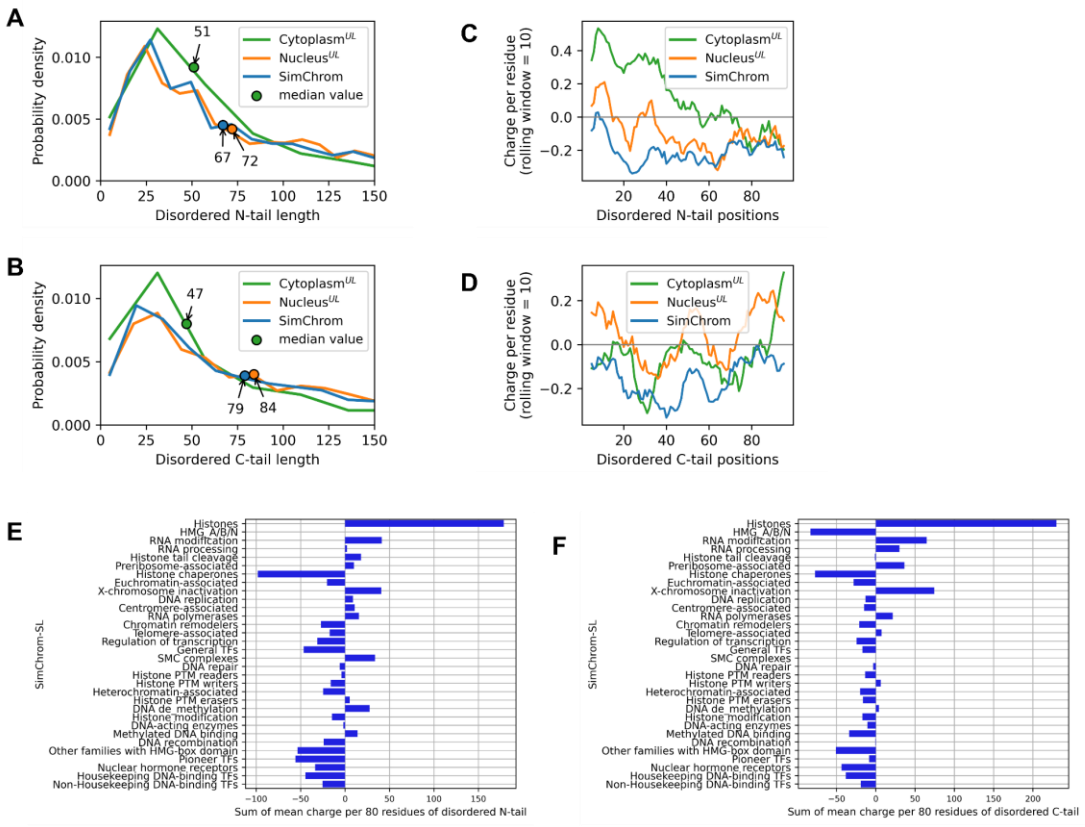


Supplementary Figure SF4_4. These are versions of Figure 4C from the main text showing chromatin protein abundance for SimChrom categories, but only for SimChrom proteins that were detected in MS-based chromatome/nucleome studies. The figure contains analysis for following chromatomes/nucleome (**A**) Alabert et al., 2014; (**B**) Kustatscher et al., 2014; (**C**) nucleome study from Itzhak et al., 2016; (**D**) Ginno et al., 2018; (**E**) Shi et al., 2021; (**F**) Ugur et al., 2023.

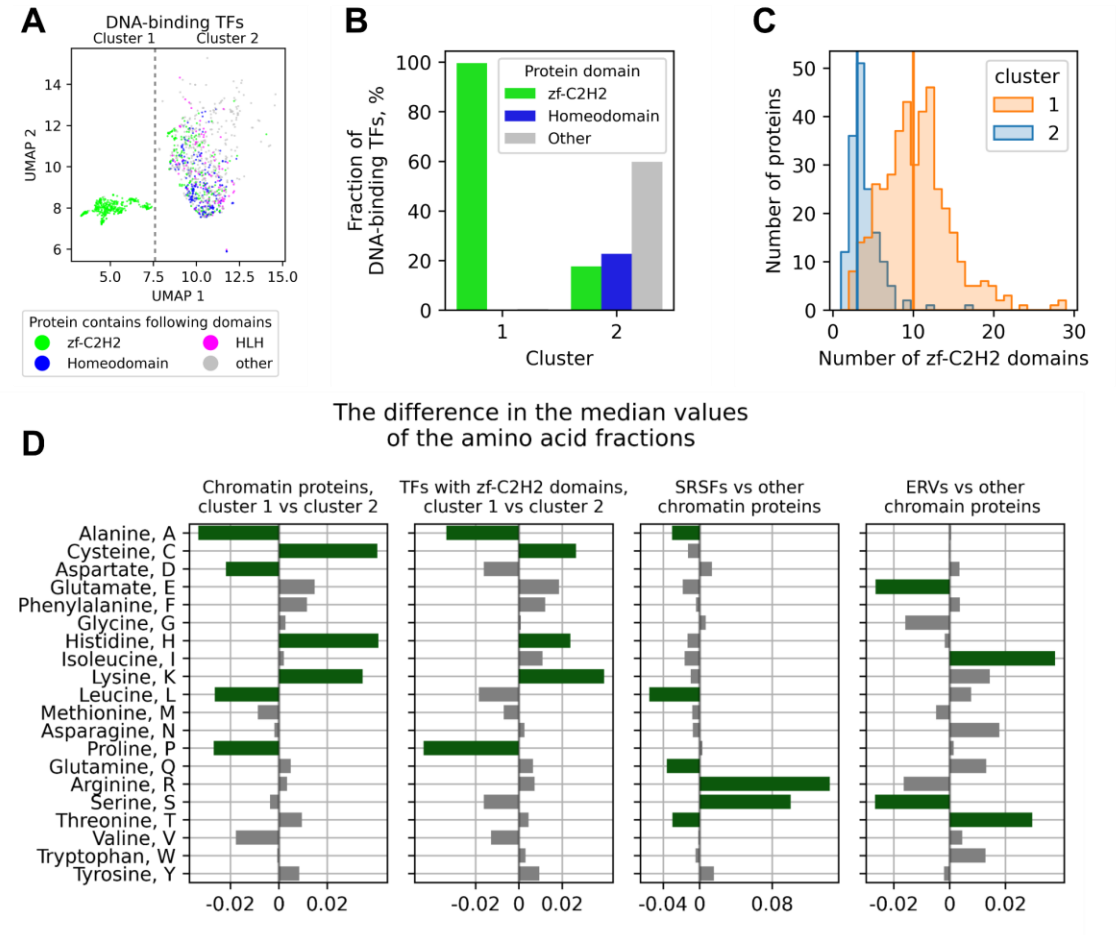
3.2. Physico-chemical properties and amino acid composition



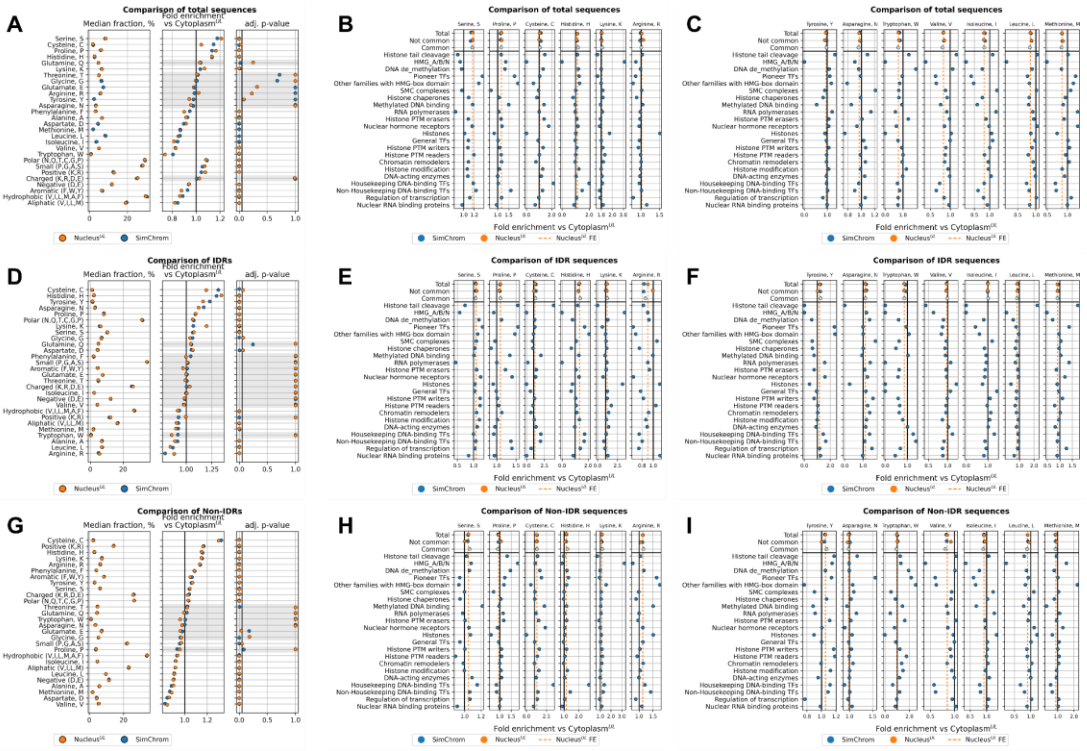
Supplementary Figures SF5_1. (A) The distribution of protein length for nuclear and cytoplasmic proteins. The q2, q3 (median) and q4 also shown. **(B)** IDR fraction distribution of chromatin, nuclear and cytoplasmic proteins (SimChrom, NULOC_CS_UL, CYLOC_CS_UL datasets). The median values are also shown. SimChrom* means SimChrom proteins without proteins from "Non-Housekeeping DNA-binding TFs" and "Housekeeping DNA-binding TFs" categories. **(C)** IDR fraction distribution for protein sets from MS-based chromatomes and nucleome. **(D)** Cumulative absolute charge of chromatin proteins from SimChrom (left) and chromatin proteins uniquely localized in the nucleus (right). **(E)** The fraction of protein entries with different types of charges (positive, negative, neutral) according to SimChrom categories.



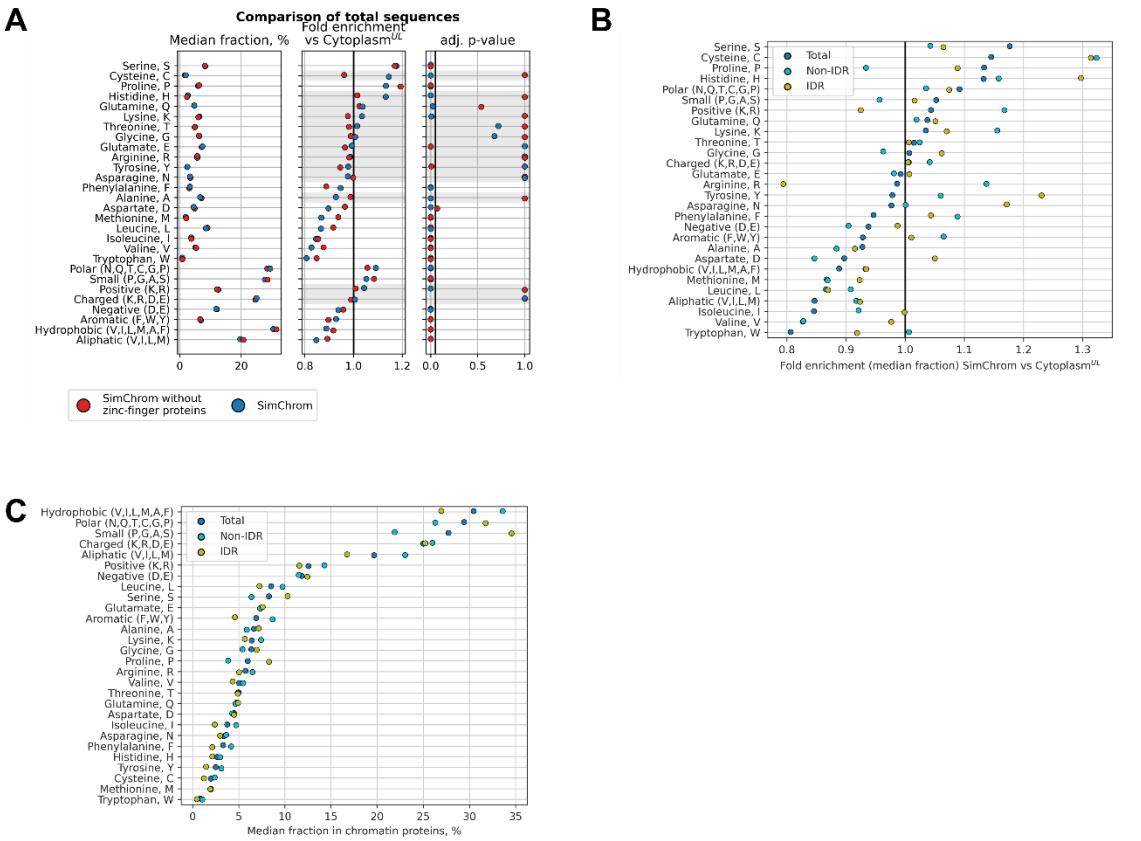
Supplementary Figures SF5_2. (A-B) The distribution of disordered tails' (N and C, respectively) length for chromatin, nuclear and cytoplasmic proteins. The medians are also shown. **(C-D)** Charge profile for N- and C-tails, averaged per amino acid residue with rolling window (10 residues) for each set of proteins (chromatin and uniquely localized in the nucleus and cytoplasm). **(E-F)** The charge of C- and N- disordered chromatin tails for each SimChrom category. For each set of proteins for each residue the mean charge was calculated, the bars represent the sum of these means. The analyzed length of disordered tails was less than 80 residues.



Supplementary Figures SF5_3. (A) This is a version of **Figure 5K** from the main text showing grouping of chromatin proteins by their amino acids composition using UMAP dimensionality reduction technique, but only transcription factors containing distinct DNA-binding domains are shown in the plot (the UMAP projections are the same as in **Figure 5K** calculated using the full set of chromatin proteins). **(B)** The fraction of DNA-binding transcription factors with zf-C2H2, homeodomain or others in the clusters from UMAP map in panel (A). **(C)** The number of DNA-binding TFs with different numbers of zf-C2H2 domains from clusters 1 and 2, median values are shown. **(D)** The difference in the median values of the amino acid fractions. The subpanels show the following comparison of: 1) chromatin proteins in cluster 1 with proteins in cluster 2; 2) TFs with the zf-C2H2 domain between clusters; 3) (Serine/Arginine-Rich Splicing Factor family (SRSFs) with other chromatin proteins; 4) proteins of the Endogenous retrovirus group (ERVs) with other chromatin proteins.

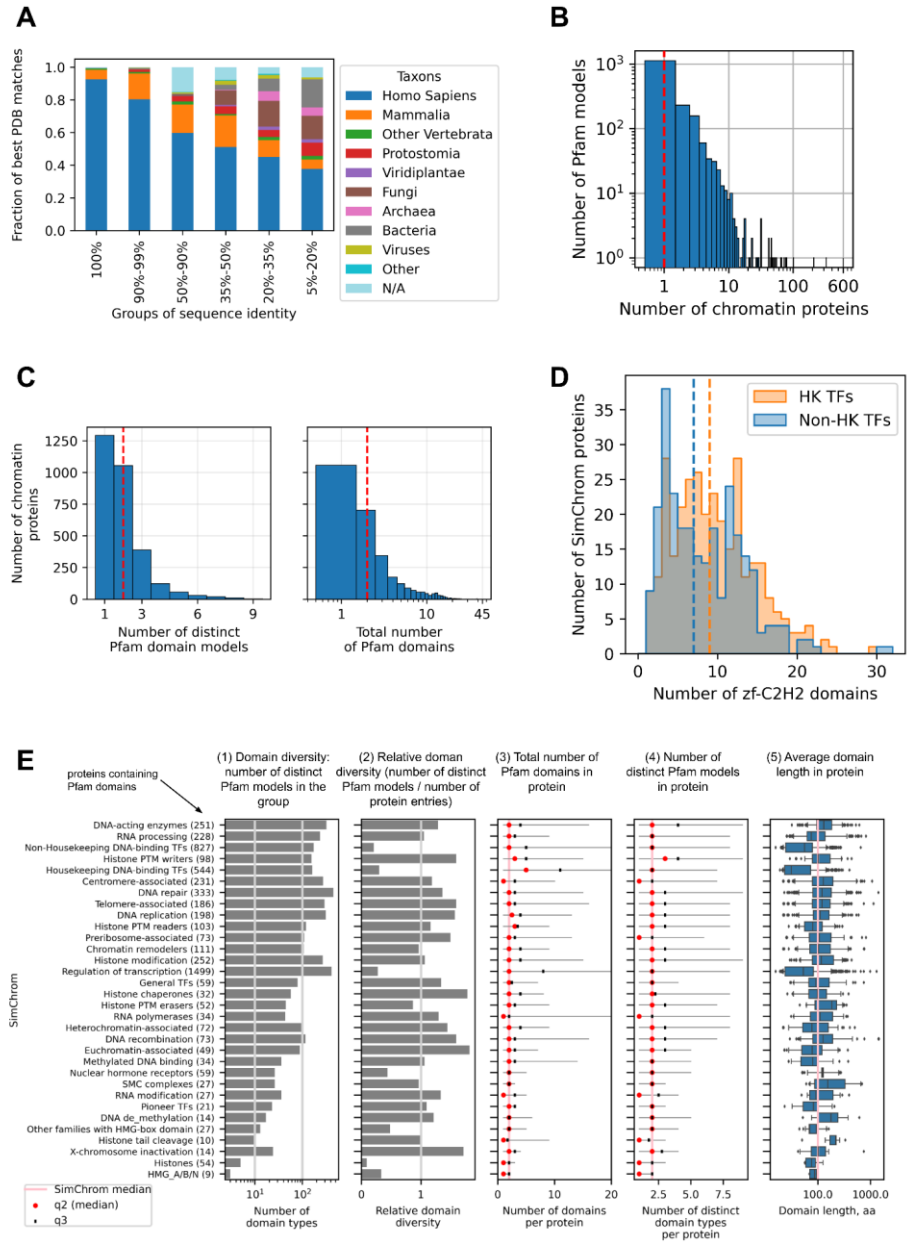


Supplementary Figures SF5_4. Comparison of amino acid composition between chromatin and uniquely localized nuclear proteins relative to cytoplasmic proteins (SimChrom, NULOC_CS_UL, CYLOC_CS_UL datasets). The comparison is done separately for the total protein sequence (**A,B,C**), IDRs (**D,E,F**), and non-IDRs (**G,H,I**). Subplots (**A, D, G**) presents the median fractions of amino acids for chromatin and nuclear proteins (subpanel 1 on each plot), the fold enrichment (FE) of these fractions relative to the cytoplasmic proteins (subpanel 2 on each plot), the black line indicates FE = 1. The adjusted p-value is shown for the statistical tests (Mann-Whitney test) comparing the median values of amino acid fractions for chromatin and nuclear proteins with the cytoplasmic ones (subpanel 3 on each plot). Gray highlights indicate a lack of statistical significance (adj. p-value > 0.05). Detailed analysis of the distribution of the selected amino acids in the total sequence of proteins belonging to respective SimChrom-SL protein categories is presented in panels **(B-I)**: enriched amino acids in chromatin proteins are shown in panels **(B,E,H),** depleted - in panels **(C,F,I)**. In the top of each plot (the first three rows) the following datapoints for the fold enrichment are given: “Total” - for all proteins from SimChrom or NULOC_CS_UL datasets (the latter also depicted by dashed line), “Common” – for common proteins among SimChrom and NULOC_CS_UL datasets, “Not common” – for proteins not present in the partner dataset (e.g., for SimChrom those present in SimChrom but absent in NULOC_CS_UL will be depicted, and vice versa for Nuclear_UL dataset).

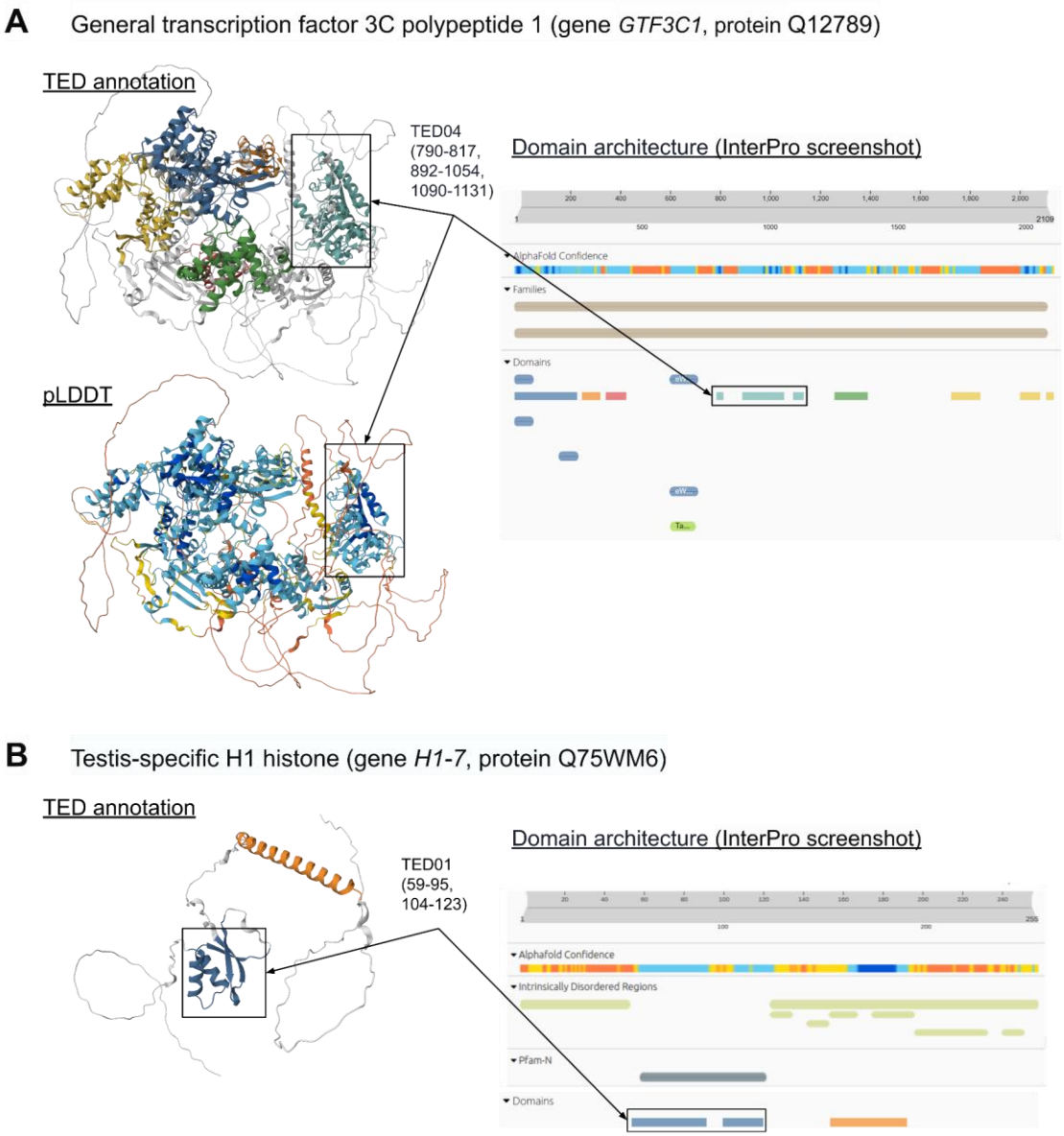


Supplementary Figures SF5_5. Additional comparisons of amino acids composition for different protein groups. (A) Comparison of amino acid composition in chromatin proteins and chromatin proteins without zf-C2H2 containing proteins relative to cytoplasmic proteins. The median fraction of amino acids for protein subsets (subpanel 1), the fold enrichment (FE) relative to the cytoplasmic proteins (subpanel 2), where the black line indicates FE = 1. The adjusted p-value is shown for the statistical tests comparing the median values of chromatin and chromatin proteins that lack zinc-finger domains with the cytoplasmic ones (subpanel 3). Gray shading indicates values that lack statistical significance (adj. p-value > 0.05). **(B)** Fold enrichment of amino acids' median fractions in chromatin proteins vs uniquely localized cytoplasmic ones, total sequences, IDRs and non-IDRs were analyzed separately. **(C)** Median value of amino acids' fractions in chromatin proteins for total protein sequences, IDRs and non-IDRs.

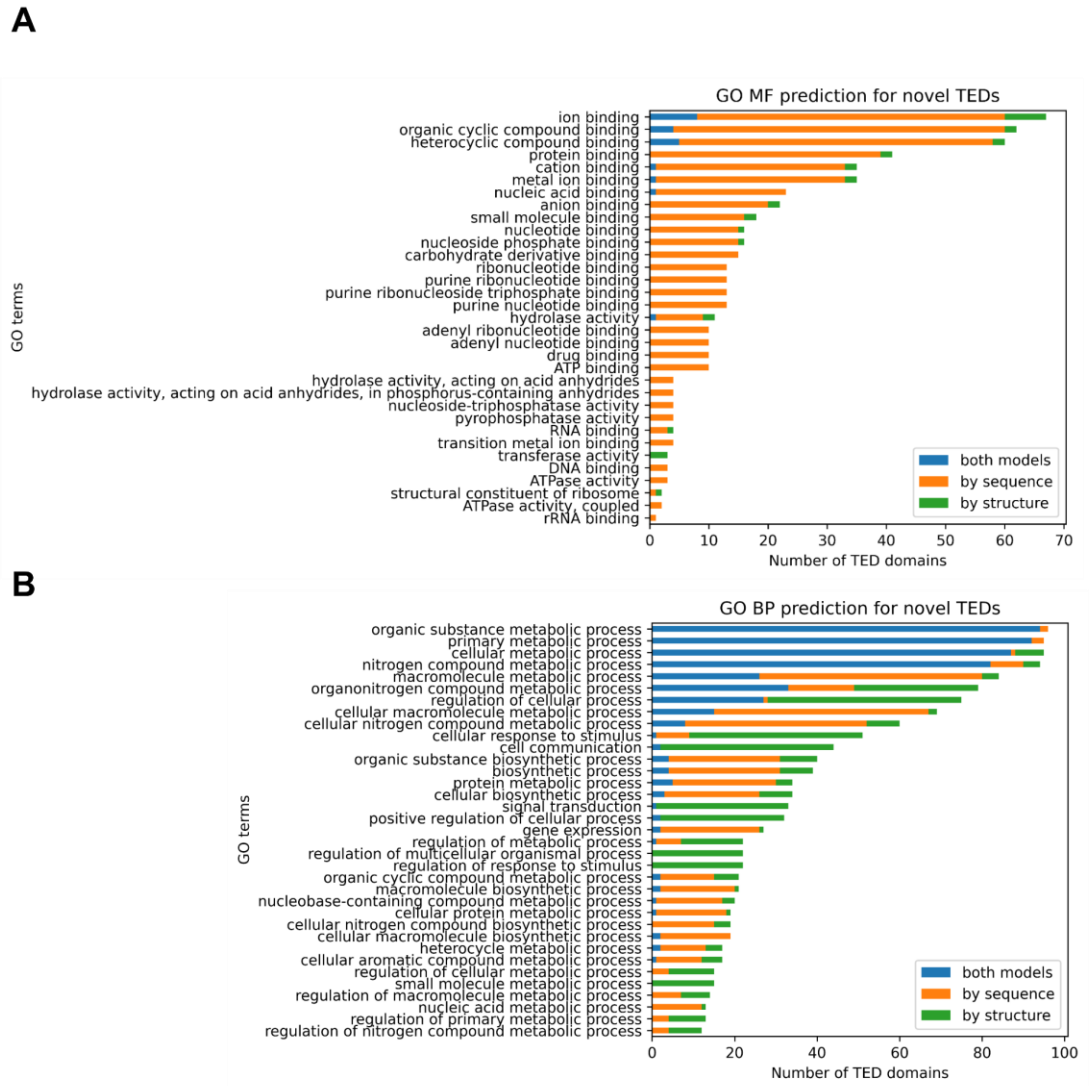
3.3. Domain composition of chromatin proteins and identification of novel structural domains



Supplementary Figures SF6_1. (**A**) Taxonomic distribution of source organisms for PDB structures with domains homologous to chromatin proteins (taxon of the best match to the structural domains identified by the TED resource, see **Figure 6**, **Methods**). (**B**) The histogram showing how many Pfam domain models (Y-axis) are found in exactly N (X-axis) chromatin proteins. One can see that the majority of Pfam domain models are represented only by domains found in one chromatin protein. The red line indicates median values. (**C**) The distribution of chromatin proteins according to the total number of Pfam domains identified in proteins (see also **Supplementary Table ST15**). (**D**) The distribution of proteins according to the number of zf-C2H2 domains in Housekeeping and Non-housekeeping DNA-binding transcription factors (HK TFs and Non-HK TFs, respectively). The lines indicate median values (7 and 9). (**E**) Analysis of functional domain diversity in chromatin proteins as identified by the Pfam database for proteins belonging to different chromatin categories according to SimChrom classification. Subpanel 1-5 represent various characteristics. This is the same as Figure 6E but SimChrom classification instead of SimChrom-SL classification is used.

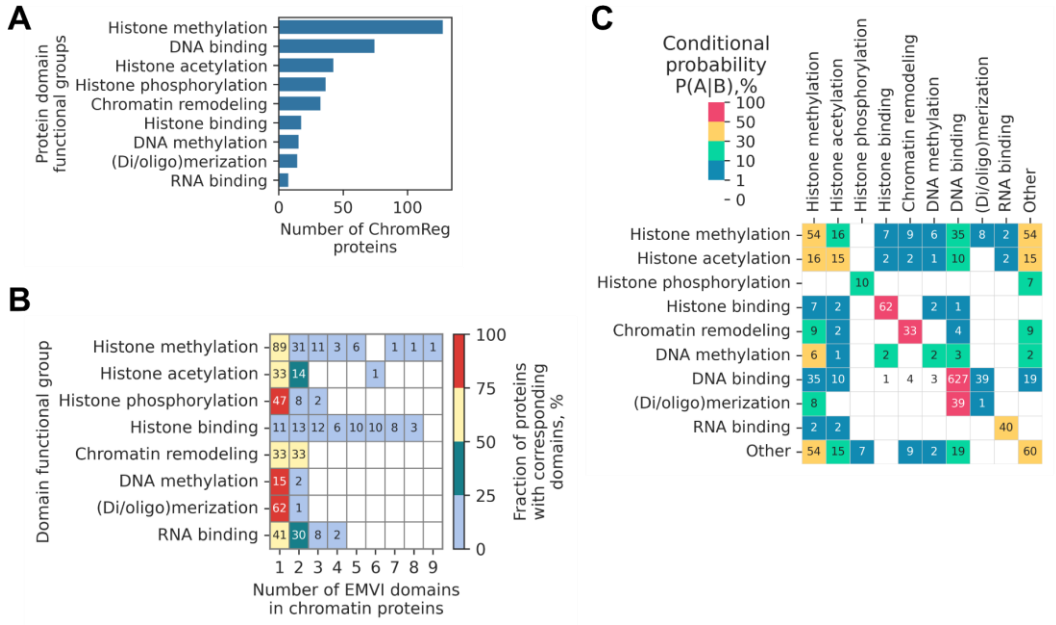


Supplementary Figures SF6_2. The examples of novel structural domains identified in chromatin proteins: structures, colored by TED domain annotation and AlphaFold2 pLDDT score, and its annotation in InterPro (screenshot). (**A**) General transcription factor 3C polypeptide 1 (gene GTF3C1, protein Q12789). (**B**) Testis-specific H1 histone (gene H1-7, protein Q75WM6).

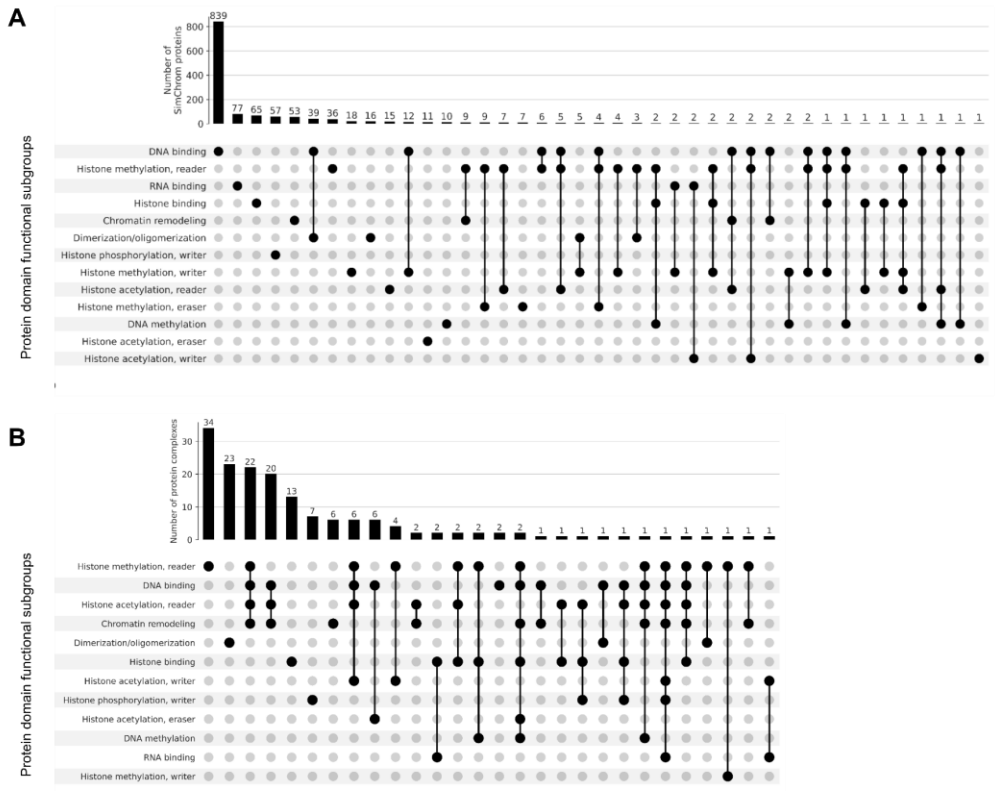


Supplementary Figures SF6_3. Predictions of GO molecular function (MF) (**panel A**) and biological processes (BP) terms (**panel B**) for novel structural domains without information in other DBs according to InterPro.

3.4. Multivalent interactions in chromatin proteins



Supplementary Figures SF8_1. (**A**) The number of chromatin regulator proteins that contain EMVI-domains of certain groups. (**B**) The number of chromatin proteins with different numbers of EMVI-domains belonging to different groups ('DNA binding' domain functional group is not shown). (**C**) Co-occurrence of EMVI-domains belonging to different functional groups in chromatin proteins. The values indicate the estimated conditional probabilities to find in a chromatin protein a domain specified in the column name given that a domain specified in the row name is already present.



Supplementary Figures SF8_2. The UpSet plot shows combinations of EMVI domains classified by their functional groups/subgroups in chromatin proteins (panel **A**) and protein complexes that exclusively contain chromatin proteins (panel **B**).

Supplementary Results and Discussion

1. Sources of information about chromatin and nuclear proteins and their critical evaluation

This section includes supplementary results and discussion to section **3.1. Sources of information about chromatin and nuclear proteins and their critical evaluation** in the main text.

A note on the distinction between and definition of nuclear proteome and chromatome

Nuclear proteome and chromatome are two terms that are historically used to describe the protein content of the nucleus and the proteins associated with genome packaging, maintenance and functioning (see **Figure 1A** for modern view of nucleus structure). The exact distinction between these two terms may be fuzzy and is often based on different consensus (protein localization or functional classification ontologies) or operational (experimental based extraction techniques) definitions. During interphase when the nucleus envelope is intact the chromatin proteins obviously reside inside the nucleus and are a part of the nuclear proteome. Hence, terminologically the nucleome seems to be more straightforwardly defined just by the protein contents of the nucleus. However, during mitosis and meiosis once the nucleus disintegrates as a distinct organelle, the situation becomes more complex. During these stages of the cell cycle there are no nuclear proteins *per se* while chromatin proteins can still be defined as those associated with the DNA in chromosomes.

Another debatable question is whether all of the proteins inside the nucleus can be considered chromatin proteins (even if nuclear envelope proteins are set aside). According to one modern view apart from the chromatin compartment the nucleus contains also interchromatin compartments [2] and nuclear bodies enriched in RNA and protein complexes (*e.g.*, nucleolus, nuclear speckles). Historically the soluble fraction of nuclear proteins was attributed to nucleosol or nuclear sap. However, to say that proteins localized in these compartments do not interact with genomic DNA at least transiently would be an oversimplification. Nucleoplasm proteins mostly also interact with genomes, some parts of the genome interact with the nucleolus (the so-called, nucleolar associated domains or NADs). Even proteins of the nuclear envelope – lamins do interact with the genomic DNA (*e.g.*, forming lamina-associated domains, LADs).

1.1. Analysis of chromatin proteins' representation in the GO database and other protein-function oriented databases

The most comprehensive gene and gene products classification resource to date is GeneOntology (GO), which classifies proteins according to the three interrelated ontologies describing molecular function, biological processes, and cellular components (called aspects). GO annotates 97% of all proteins of the human reference proteome (as provided by UniProt), but this classification also has several drawbacks. GO combines many categories (currently around 42 thousand terms) of various scope describing various aspects of gene functioning connected via different types of relationships (such as "A is B", "A is part of B", "A regulates B", "A occurs in B", *etc*.) in a non-treelike structure (directed acyclic graph). This complex intertwined hierarchy of GO makes it difficult to get a holistic picture of various chromatin protein groups, and apply a reductionist way of thinking while interpreting the results of bioinformatics analysis of protein sets made using GO classification. Another drawback is that GO omits categories that are historically well established in the community of chromatin researchers (*e.g.*, such categories as "histone proteins", "high-mobility group proteins", *etc.*), again hampering interpretation of GO-based data analysis using the established knowledge (see discussion below).

The GO cellular component term "chromatin" is defined broadly as "the ordered and organized complex of DNA, protein, and sometimes RNA, that forms the chromosome." Consequently, functionally relevant chromatin terms — such as "nucleosomal DNA binding", "DNA-binding transcription factor activity', 'Histone H3K27 DNA-binding transcription factor activity', 'Nucleus', 'Histone H3K27 monomethyltransferase activity" — may not be linked hierarchically to the chromatin GO node, resulting in incomplete overlaps between protein lists. Based on our analysis, over 500 functionally defined chromatin proteins (inferred from literature and other sources) are absent from GO annotations (see **Supplementary Figure SF2_1A**).

While the GO provides nearly comprehensive coverage of the proteome, its classification structure presents challenges for extracting or comparing specific protein sets. The GO hierarchy is complex, with heterogeneous relationships between nodes, overlapping protein annotations across terms, and varying levels of detail and completeness. For instance, manually inspecting GO terms associated with chromatin-related keywords (*e.g.*, "DNA", "transcription", "histone", "RNA polymerase") is impractical due to their sheer volume (>100 terms; see **Supplementary Figure SF2_1B**). Furthermore, GO terms inherently include proteins from all child terms, which can lead to unintended inclusions. For example: the term "DNA-templated" transcription incorporates "mitochondrial transcription"; "gene expression" encompasses functionally distinct processes like "protein maturation" and "translation".

When compared with the EpiFactors database [3] (containing epigenetic regulator protein entries obtained by text mining) 46% of entries in EpiFactors are missing from the list of GO 'chromatin' proteins, see **Figure 2D.** Moreover, while a specialized review by Hammond et al., 2017 [1] lists 35 proteins of histone chaperone category, the GO term "histone chaperone activity" includes only 14, with just 6 overlapping entries (see **Supplementary Figure SF2_1D**).

Several functionally important but small chromatin protein categories are entirely missing from GO, including: HMG proteins, Histone tail cleavage proteins, general TFs. Even for well-annotated classes like histones, inconsistencies persist. Most are classified under "structural constituent of chromatin", but this term also includes two non-histone proteins (*HMGA1* and *LMNTD2*).

The Gene Ontology offers the most comprehensive coverage, annotating nearly the entire human reference proteome. In contrast, chromatin/epigenetic regulator databases typically include 400–800 proteins, while protein class-specific databases vary significantly in scope, ranging from as few as 30 proteins (*e.g.*, chromatin remodelers of the SWI/SNF family) to over 1500 (*e.g.*, transcription factors). Despite their utility, none of these resources fully captures the complexity of chromatin-associated proteins, either in terms of protein coverage or functional classification. While chromatin/epigenetic regulator databases encompass key cofactors for certain protein complexes, they exclude critical categories such as transcription factors, RNA polymerase subunits, DNA-modifying enzymes, DNA repair machinery, and HMG proteins. Conversely, protein class-specific databases are limited to at most six functional categories in total, including histones, chromatin remodelers (from select families), histone post-translational modification (PTM) writers/readers/erasers, and transcription factors. Additionally, several chromatin-related protein classes have been reviewed as the gene group in HGNC (*e.g.*, 'High mobility group', 'DNA polymerases') or in the literature but lack dedicated database resources. Examples include histone chaperones [1], SMC complexes [4], HMG proteins [5], pioneer transcription factors [6], nuclear RNA-binding proteins [7,8], and histone tail cleavage enzymes [9]. The full list of chromatin-associated protein class specific sources are available in **Supplementary Table ST1**. The absence of centralized repositories for these protein classes highlights a critical gap in current bioinformatics resources.

The revealed problems for using GO directly for extracting a set of chromatin proteins arise from several factors (see also [10] for a broader discussion of GO applicability): 1) protein multifunctionality: many proteins participate in diverse processes or localize to multiple compartments, 2) ambiguous term definitions: brief GO descriptions may lead to inconsistent interpretations, 3) annotation delays and errors: lag times in updates and propagation of errors in curated datasets, 4) curation bias: well-studied proteins are annotated more thoroughly than niche categories.

1.2. Detailed comparative analysis of nuclear proteins subcellular localization between UniProt, HPA, and OpenCell

Note: **Interactive Figure 2** (https://simchrom.intbio.org/#localization) is the interactive version of **Supplementary Figure SF2_2,** which is the key source of information for the analysis presented below.

In this section we analyzed in detail the information on subcellular localization on nuclear proteins provided by three database/proteome wide studies: UniProt, HPA, OpenCell. Each resource uses its own hierarchical subcellular localization ontology and a system of evidence tags (UniProt), reliability scores (HPA) or annotation grades (OpenCell) to annotate proteins with localization terms with different levels of confidence or types of evidence. From the three databases we obtained the lists of proteins that had their localization annotation(s) supported by at least one evidence tag (in the case of UniProt), reliability score higher than "uncertain" (in the case of HPA) or annotation grades better than the worst one (in the case of OpenCell) (see **Methods 2.1.1** for details). UniProt is currently the main state-of-the-art reference data source about the human proteome, estimating it at 20225 proteins (release 2022_2). From these around 14 thousand are "localized" according to our criteria (see **Supplementary Figure SF2_2A**). HPA has so far experimentally analyzed around 13 thousand human proteins from which 12 thousand are localized. Interestingly, the localized protein sets of UniProt and HPA overlap only by 60-70% (8981 proteins, see **Supplementary Figure SF2_2A**), leaving around three thousand proteins, whose localization is characterized by HPA, but not available in UniProt. OpenCell has so far analyzed only a minor fraction of human proteome (all of them have been localized), although it provides localization information for another 37 proteins that are not "localized" by either UniProt or HPA (see **Supplementary Figure S2_2A**). In total the three databases collectively provide localization information for 17335 proteins amounting to 86% of the known human proteome. The fraction of "localized" proteins relative to each dataset constitutes 70%, 93%, 100% of the total number of proteins listed in UniProt, HPA, and OpenCell, respectively (see **Supplementary Figure S2_2A**).

We next set to analyze information provided by these resources about the nuclear/chromatin proteins including their sublocalization. The representation of different localization ontologies, the number of proteins annotated by the respective annotation terms (each individual term had to be supported by the criteria used to classify a protein as "localized" described above) and the overlap between the number of proteins annotated by different terms in UniProt and HPA is given in **Supplementary Figure SF2_2A**. We note that localization ontologies used by the three resources are somewhat different reflecting different approaches to protein classification, other ontologies used as basis (*e.g.*, GeneOntology (GO) in the case of UniProt), evolution of our understanding of chromatin organization (*e.g.*, some obsolete terms such as "Nucleus matrix" may be

present in UniProt [11]), resolution of experimental methods and data sources used to classify protein localization. All ontologies converge on the presence of the following four common localization entities within the nucleus "Nucleus envelope/Nucleus membrane", "Nucleolus", "Nucleoplasm", "Nuclear bodies/speckles/punctae". However, the hierarchical relations between these terms within each ontology and their further subdivision into sub-localization (subcompartment) entities differ. Moreover, the terms with the identical names may be used at different levels of the hierarchy as both the names of specific annotation terms and groups of annotation terms, these have to be treated as separate terms when using this ontology for the analysis of protein sets (*e.g.*, the terms "Nucleoplasm" in HPA, or "Nucleus envelope" in UniProt, see **Supplementary Figure SF2_2A**). The term "Chromatin" is only explicitly present in OpenCell, suggesting that chromatin proteins in UniProt and HPA should be found in localization subcategories (Nucleus, Nucleoplasm, etc.). However, UniProt has a separate localization term "Chromosome", which according to its classification is a separate entity not being a part of "Nucleus". This term was inherited by UniProt from GeneOntology (GO) cellular component classification. According to GO "chromatin" is a part of "chromosome", while the term "chromosome" may be used to describe both components located inside (e.g., "nuclear chromosomes") and outside (e.g., "mitochondrial chromosomes") the nucleus. To add to this complexity, HPA considers "mitotic chromosomes" a part of "Nucleoplasm", but does not explicitly include the term "Nucleus" in its cellular sublocalization ontology (the subcellular localization ontology at the most general level already is subdivided into "Nucleoli", "Nucleus membrane" and "Nucleoplasm", as well as, "Cytosol", "Mitochondria", etc). The above-mentioned discrepancies reflect the dynamic complexity of cellular organization, our constantly evolving understanding of nuclear organization, and the resulting difficulty in describing subcellular localization in a form of a simple hierarchical tree-like ontology.

We next performed comparative analysis of the quantitative and qualitative composition of nucleomes as provided by the three resources. We first addressed the question of the number of nuclear proteins known to date and the consistency in these estimates between the resources. The number of proteins having their localization assigned to the nucleus is estimated as 4717, 6500, and 640 by UniProt, HPA, and OpenCell respectively, which gives an estimate of 33%, 53%, and 49% of human proteome to be present in the nucleus according to these resources, respectively. The **liberal estimate** of the number of nuclear proteins (combined from three resources) is 8035, which amounts to 46% of the known localized proteome (17335 proteins in all resources combined).

Analysis of multi-localization of nuclear proteins in other cellular compartments

A significant proportion of the nuclear proteome is considered to be also localized in other cellular compartments. We have aggregated localization terms in HPA and UniProt ontologies into the following generalized categories Nucleus, Cytoplasm, Endomembrane system (see **Methods Section 2.1.1**). **Supplementary Figure SF2_2B** shows the multi-localization of proteins across these categories

in the form of Venn diagrams. According to UniProt and HPA 41% and 39% of nuclear proteins are localized in cytoplasm, respectively, 12% and 22% of nuclear proteins are localized in the endomembrane system, respectively. Among the nuclear proteins whole localization is available in both HPA and UniProt, only for ~40% of proteins the two databases reach consensus for their unique nuclear localization (1262/3288).

We next analyzed consistency between UniProt and HPA in annotating nuclear proteomes. The overlap between the nucleomes from UniProt and HPA amounts to only 3288 proteins, meaning that only 51% of nuclear proteins reported by HPA are supported by UniProt, and vice versa 70% of nuclear proteins reported by UniProt are supported by HPA. However, since both HPA and UniProt simultaneously provide localization information only for 8981 proteins, it is better to estimate consistency between the resources only within this set of proteins. Among these 8981 proteins, nuclear proteins constitute 3896, 4699, and 5307 according to UniProt, HPA, and HPA or UniProt, respectively. Hence, we estimate the consistency between UniProt and HPA in annotating the proteins as nuclear at 62% (Jaccard similarity measure) (3288/5307). Hence, 2019 proteins were annotated as nuclear by one resource but not the other, and at the same time had localization annotation in both resources. These discrepancies may in part come from the fact that proteins may have multiple localization, and one of the resources misses one of the localizations. We see that among these 2019 proteins, 1670 (83%) proteins were simultaneously localized in two or more major cellular/extracellular structures (Nucleus, Cytoplasm, Endomembrane system, Other (including Secretory/Extracellular) localization - see **Methods Section 2.1.1**) according to at least one of the resources. Among those for 1254 proteins HPA and UniProt were consistent in identifying at least one common localization, suggesting that the discrepancy in their nuclear localization between HPA and UniProt may be because one resource misses its additional localization in the nucleus, while they are consistent about their localization in other cellular compartments. For 765 proteins that were considered as nuclear only by one resource, no alternative common localizations were reported by both resources. To further understand potential sources of these discrepancies and annotation biases in HPA and UniProt we performed a cross-annotation analysis for this particular “discrepant” set of proteins: analyzed the annotations provided by one resource for the proteins that were considered nuclear by the other resource. The results of our analysis with respect to detailed annotation categories is presented in **Supplementary Figure SF2_3**. These “discrepant” proteins are mainly annotated as belonging to nucleus or nucleoplasm in one resource and fall into a wide range of categories in the other resource. Among this wide range of localization categories in both cases the cytoplasmic proteins form the largest category by size, but account only for 20%-30% of all “discrepant” proteins, and are statistically underrepresented (fold enrichment of 0.7-0.8, compared to the case if they were selected at random from the respective set of proteins - see **Methods Section 2.1.1**). Among nuclear proteins annotated by UniProt but whose annotation was missed by HPA (154 proteins), proteins with "vesicle" localization category according

to HPA were overrepresented (1.3 fold enrichment). Alternatively, among nuclear proteins annotated by HPA, whose annotation was missed by UniProt (611 proteins), "secreted" and "extracellular matrix" proteins (according to UniProt) were mainly overrepresented (1.6 and 1.6 fold enrichments, respectively) with largest number of proteins, while overrepresented "chromosome" and "kinetochore" (6.1 and 2.7 fold enrichments, respectively) contains less than 20 proteins. *Taken together this suggests that discrepancies in nuclear protein localization between UniProt and HPA have a complex origin, some effects may be in part due to 1) the tendency of HPA to mislabel nuclear proteins as vesicular proteins (may explain up to around 30% of proteins with mis-annotated localization), and 2) the tendency of UniProt to mislabel nuclear proteins as belonging to the secreted proteins and proteins of the extracellular matrix (may explain up to around 15% of proteins with mis-annotated localization). However, (see **main text section 3.1**) one of the main sources of these discrepancies is also the incompleteness of the databases (especially, UniProt) with respect to annotating multilocalization of proteins (e.g., the databases may agree on cytoplasmic localization of a protein, but one database may additionally annotate it as nuclear).*

We next performed comparative analysis for the number of nuclear proteins assigned to different subnuclear localization categories as provided by the three resources, and particularly analyzed the quantitative similarities and differences between the data provided by UniProt and HPA (see **Supplementary Figure SF2_2A**). A conceptual difference between UniProt and other resources, is that UniProt provide subnuclear localization data only for 19% (882 proteins) of its nuclear proteome, while HPA and OpenCell *per se* do not have a localization term “Nucleus” and provide protein localization by default at a finer level. Particularly, the largest subset of nuclear proteins in HPA and OpenCell (5948 and 487 proteins, respectively) belongs to nucleoplasm category, while only 170 nucleoplasmic proteins are found in UniProt (245 more proteins in UniProt are classified as belonging to nuclear bodies, which is not considered are part of nucleoplasm by UniProt, but is considered by HPA). Among the consensus set of nuclear proteins between UniProt and HPA (3288 proteins) 95% of proteins annotated as nucleoplasm by HPA are annotated solely by the term “Nucleus” in UniProt (2781 out of 3053). The second largest category of 1191 proteins according to HPA are proteins belonging to the Nucleolus (this is not the second largest category in OpenCell, but OpenCell is biased towards the most abundant proteins), the same is true for UniProt (372 proteins), however, the number of proteins is much less in accordance with the small proportion of proteins that UniProt provides subnuclear localization annotation for. Among these 372 proteins only around 50% (185 proteins) are also annotated as belonging to Nucleolus by HPA. The consistency between other subnuclear localization annotations is even less (see **Supplementary Figure SF2_2A** and **Supplementary Table ST2**). For example, in the case of nuclear envelope/nuclear membrane only 46 proteins have a consistent annotation between UniProt and HPA, while for 368 proteins the annotations do not match. Taken

together our current understanding of sub-nuclear localization annotations are not very much consistent between UniProt and HPA.

The subnuclear multi-localization of nuclear proteins was another aspect of our analysis. UpSet plots comparing the number of proteins that have multiple subnuclear localizations according to HPA and UniProt were used for this analysis (see **Supplementary Figure SF2_4**). To be able to compare the data between the two resources, given the difference in localization ontologies as described above, in this analysis we used a manually created common localization ontology consisting of four categories (nucleoplasm, nuclear membrane, nucleolus, nuclear bodies) separately mapped to HPA and UniProt as detailed in **Supplementary Table ST3**. According to HPA the majority of nucleoli proteins (59%) are also annotated as localized in nucleoplasm, which is not unexpected because nucleoli are suspended in nucleoplasm with proteins diffusing to and from the nucleoli. In the case of UniProt only 17% of nucleoli proteins are also localized in nucleoplasm, consistent with a small number of nuclear proteins annotated as nucleoplasmic proteins in UniProt. A similar situation is found when comparing multiple localization between nucleoplasm and nuclear bodies (48% and 16% overlap according to HPA and UniProt, respectively). Another interesting fact is that according to HPA the majority of nuclear membrane proteins (54%) have a second localization in the nucleoplasm (in the case of UniProt only 6% have the said location). This suggests that many of the inner nuclear membrane proteins may be tethered to the nuclear membrane transiently through interactions with the bona fide membrane proteins that have transmembrane domains.

Given the complexity of nuclear sublocalization ontologies, significant differences in certain aspects of data provided by UniProt, HPA, and OpenCell (as discussed above), and different comparative analyses between the datasets that may be done to answer various specific questions about nuclear localization/subnuclear localization/multi-localization of proteins, we have implemented an interactive data viewer based on the analysis present in this work. This viewer is available https://simchrom.intbio.org/#localization and allows to interactively select and analyze various subsets of nuclear proteome based on the data collectively provided by the three resources.

1.3. Detailed comparative analysis of sets of chromatin proteins identified in MS-based studies

An important source of information about chromatome content may be obtained from MS-based studies of chromatin or nuclear extracts (reviewed in [12,13]). We aimed at comparing these data to the chromatome and nucleome datasets developed above. We have chosen a number of high quality studies published since 2010 with publicly available data that analyzed human chromatin protein content (**Table 1**, **Supplementary Table ST1**, **Supplementary Figure SF2_5**, **Supplementary Figure SF2_6**, **Supplementary Figure SF2_7**). These studies used different human cell lines (Torrente et al.,

2011 - HeLa S3; Kustatscher et al., 2014 - HeLa, MCF-7, HepG2, HEK293, U2OS, DT40; Ginno et al., 2018 - T98G (derived from a human glioblastoma multiforma tumor); Shi et al., 2021 - K562, Ugur et al., 2023 - hESCs). For additional comparison we also included nascent chromatomes (Alabert et al., 2014 - HeLa S3; Alvarez et al., 2023 - HeLa S3 and TIG-3 fibroblasts) and MS-based study that reported a set of proteins predominantly localized in the nucleus (Itzhak et al., 2016 - HeLa cells).

While aiming at analyzing the chromatin content directly, MS-based studies are not without known limitations that should be kept in mind during the analysis. One previously reported limitation is the difficulty of extracting the chromatin associated proteins without the contamination from cytoplasmic and mitochondrial proteins, another one is the difficulty in retaining the proteins that associate with chromatin dynamically, transiently, or are present in small amounts [12,13]. Over the years different approaches to chromatin extraction have been suggested: (1) techniques based on crude differential detergent/salt extraction of chromatin proteins (which are known to result in a substantial contamination by cytoplasmic proteins, in part due to the sticky nature of chromatin which captures other proteins during extraction [14], (2) techniques that rely on formaldehyde cross-linking prior to the biochemical extraction of the chromatin in order to reduce contamination by cytoplasmic proteins (ChEP, DEMAC), (3) techniques that add additional steps of chromatin fractions' purification using ultracentrifugation (DEMAC), (4) techniques that are based on adaptation of *in situ* Hi-C methods that enrich DNA-protein complexes through extraction of biotinylated DNA fragments and are claimed to capture more transiently interacting chromatin proteins (Hi-MS), (5) techniques that use MNase digestion to aid in extracting chromatin associated proteins (total MNase digestion). Additionally, some studies compared protein MS-intensities between different cellular components to reliably establish the protein contents of a particular cellular compartment (e.g. the approach used by Itzhak et al., 2016) or supplemented MS-analysis with machine learning classification of proteins based on the changes in protein MS-intensities when cells under different biological and biochemical conditions were analyzed (e.g., the approach used by Kustatscher et al., 2014). The chosen datasets employed various of the above mentioned techniques (see **Supplementary Figure SF2_5A**, **Table 1**). Additional known limitations of MS-based studies are due to: (1) the variable sensitivity of MS-based analysis techniques combined with highly variable levels of expression of chromatin proteins, (2) the variation in chromatin protein expression between the cell lines used for analysis, and the fact that not all chromatin proteins are expressed in all cell types and during all stages of the cell cycle and development, (3) difficulties in interpretation of MS-data for proteins with high sequences similarity, such as histones [15].

First, we analyzed the absolute number of proteins in MS-based chromatin datasets and the fractions of those datasets that were present in the SimChrom dataset of chromatin proteins that was developed by us as a part of this study (see **Results Section 3.2**) and nuclear localization-based datasets (NULOC_JT_NECF, NULOC_CS see **Results Section 3.2**) (see **Supplementary Figure SF2_5A**).

The MS-based datasets showed high variability, only 179 proteins were in common between the chromatin datasets only (see **Supplementary Figure SF2_5B**). The largest MS-based datasets of total chromatin (Shi et al., 2021; Ginno et al., 2018) and nascent chromatin (Alabert et al, 2014; Alvarez et al., 2023) contained around three thousand proteins – approximately the same amount as the number of proteins in SimChrom. However, surprisingly the number of proteins for these datasets that were present in SimChrom was small (25-35%). This low consistency with SimChrom was also observed in the Torrenete et al., 2011 dataset. These three datasets (Shi et al., 2021; Ginno et al., 2018; Torrente et al., 2011) were based purely on experimental chromatin extraction techniques (**Supplementary Figure SF2_5A**). Alternatively, the consistency with SimChrom was twice as high (50-65%) for the Kustatscher et al., 2014 and Itzhak et al., 2016 datasets. These datasets stand out from the other datasets. Kustatscher et al., 2014 used a machine learning classification approach based on MS-data signals with a manually provided chromatin proteins training dataset, which was based on literature and database mining. In Itzhak et al., 2016 the proteins were considered nuclear if their MS-measured intensity in the crude nuclear extract exceeded 85% of the global intensity. We hypothesize that these two latter datasets have better consistency with SimChrom dataset in part because by their design they are initially biased either by the information already available in the literature and various databases (in the case of Kustatscher et al., 2014) or by the selection of proteins that are preferentially localized in the nucleus and hence having higher chances to be described in the literature and databases (in the case of Itzhack et al., 2016).

When MS-based datasets were compared to the nuclear localization datasets, we observed the same tendency. Around 95% of proteins in Kustatscher et al., 2014 and Itzhak et al., 2016 datasets may be found in our broad nuclear localization dataset (NULOC_JT_NECF), while for other MS-based datasets the proportion was around 60-70% (Torrente et al., 2011; Alabert et al., 2014; Ginno et al., 2018; Shi et al., 2021) and 75-80% (Ugur et al., 2023; Alvarez et al., 2023). The same tendency was observed for the consensus NULOC_CS dataset (the portion of the MS-based datasets present in NULOC_CS varied between 30% and 73%).

To further understand the origins of these discrepancies we analyzed proteins that were found in MS-based experimental chromatin datasets but not present in nuclear localization datasets or SimChrom (**Supplementary Figure SF2_5C**). The Itzhak et al., 2016 dataset had a 98% overlap with NULOC_JT_NECF and its inclusion would not affect the results presented below. In total there were 2232 of such proteins, and only a minor fraction of these (94 proteins) did not have localization annotation in the databases (see **Supplementary Figure SF2_5D**). Five proteins were present in four total chromatin MS-based datasets (see **Supplementary Figure SF2_5C**), they included proteins encoded by the FLNB (Filamin B, actin-binding protein), GMPS (Guanine Monophosphate Synthase), CHERP (Calcium Homeostasis Endoplasmic Reticulum Protein), ILKAP (ILK Associated

Serine/Threonine Phosphatase), PLEC (Plectin) genes. However, the Ugur et al., 2023 dataset includes only PLEC and CHERP. While according to UniProt and HPA all of these are non-nuclear proteins predominantly localized in cytoplasm (with additional localizations in cytoskeleton, intermediate filaments, endoplasmic reticulum, Golgi apparatus), manual literature mining confirmed experimental evidence supporting the presence of these proteins in the nucleus (e.g. [16]). We additionally randomly selected 20 proteins from a subset of proteins that were reported by at least five out of seven chromatin MS-based studies (there were 195 such protein coding genes) and manually performed literature searches. From those 15 for 5 genes literature evidence was found suggesting their nuclear localization (CALR [17], PDIA4 [18], ABCF2 [19], SEC23B [20], EIF3D [21]). Hence, it may be stated that MS-base studies currently have predictive power to identify new chromatin proteins that are not annotated as such by the localization and functional databases.

It is not straightforward to estimate the potential contamination of MS-based datasets with non-nuclear proteins since one cannot come up with an ultimate reference set of non-nuclear proteins. Even for well studied proteins there are still chances that they may have axillary functionality in the nucleus that have not yet been experimentally characterized. Still to address this problem we relied on analyzing the above mentioned set of 2232 proteins using information available in GO. GO enrichment analysis revealed that these proteins were mainly associated with a diverse set of GO-terms related to non-nuclear organelles/compartments, cellular metabolism, protein translation and maturation suggesting that no important chromatin associated categories were missed during the construction of SimChrom dataset that could account for this discrepancy (**Supplementary Figure SF2_6**). Previous studies suggest that the results of MS-based studies may be contaminated by cytoplasmic and mitochondrial proteins [12,22]. In our analysis 2025 proteins were related to cytoplasm, 115 to Golgi vesicle transport according to GO. No enriched terms include mitochondria proteins. It is still possible that among 2232 proteins there are still nuclear proteins, whose annotation by GO does not account for their additional, moonlighting functions in the nucleus. For instance, 30 proteins were associated with translation ("translation", “translation initiation factor activity”) according to GO, many proteins of the translational apparatus are known to be moonlighting proteins with functional roles in the nucleus [23]. Similarly, Golgi apparatus cooperating with nucleus and ER in vesicular transport.

In a different type of analysis we looked at chromatin proteins that were not identified by the MS-studies but were included in our SimChrom dataset. There were 1246 such proteins (or 41% of SimChrom). According to the HPA classification of housekeeping proteins, 67% (839 proteins) of these were not housekeeping, consistent with the idea that they were missed by MS-based studies, because they were not expressed in the cell lines. However, it was also found that MS-based studies are biased towards identifying the housekeeping proteins. More than 75% of nuclear/chromatin proteins reported by the MS-based studies were from the housekeeping pool, while the average expected fraction of

nuclear housekeeping proteins is around 62% (**Supplementary Figure SF2_7B**). Among the set of 1246 proteins not found in the MS-based studies the dominant SimChrom category was related to DNA-binding transcription factors (see **Supplementary Figure S2_7C**). 1148 DNA-binding TF were missed by MS-based studies, including 394 housekeeping TFs. This highlights another potential source of discrepancy - housekeeping TFs may be present in small amounts or be washed away during chromatin/nucleome extractions and thus be missed by MS analysis. To further understand the discrepancies between the MS-based datasets and SimChrom we performed enrichment analysis for the SimChrom categories in the experimental datasets (**Supplementary Figure SF2_8A**). It can be seen that categories related to DNA-binding transcription factors were mainly depleted in the MS-based datasets, consistent with the analysis described above. The separate analysis of housekeeping TF suggests that different experimental methods show high variability in their ability to recover TF in chromatin extracts. The Torrente et al., 2011 dataset includes only 8 housekeeping TF, while the Kustatscher et al., 2014 and Itzhak et al., 2016 datasets each contain 158. **The dynamic nature of transcription factors' interactions with chromatin likely explains these facts.** The most enriched categories were related to protein involved in RNA binding and metabolism, histone chaperone, remodelers and heterochromatin associated factors. This is likely related to the higher chances of these proteins to be detected due to their universal presence in the cells and high expression levels. Detailed analysis of the SimChrom categories representation in MS-based datasets further revealed some details of the differences between the datasets (**Supplementary Figure SF2_7C**). For instance, the ML-based Kustatscher et al., 2014 dataset was able to recover twice as many DNA transcription factors than Ginno et al., 2018 and Shi et al., 2021 datasets. The ratio of housekeeping and non-housekeeping TFs for Ginno et al., 2018 and Shi et al., 2021 datasets is the same, and is higher for the Kustatscher et al., 2014; Itzhak et al. 2016; Alabert et. al., 2014 datasets. Interestingly the nucleome Itzhak et al., 2016 dataset was also able to recover the same amount of TF as the Kustatscher et al., 2014 dataset, although the size of the dataset was three times smaller than Shi et al., 2021 and Ginno et al., 2018 datasets. This again points to the fact that TF may be lost during chromatin extraction. The representation of some other SimChrom categories had significant differences between the MS-based datasets likely attributed to the differential extraction probability. For instance, only 29% and 25% of RNA polymerases and histones, respectively, were present in the Ginno et al., 2018 dataset, while 50-60% were present in the Shi et al., 2021 dataset. A closer look at the histone proteins (for which currently information about all 62 expressed proteins is known [15,24,25] revealed that while certain tissue specific histone variants were missed as expected, all MS-based studies were not able to recover all canonical histones variants, especially for the H2B-histone (**Supplementary Figure SF2_8B**). For instance, from 14 canonical H2B proteins isoforms the Shi et al., 2021 dataset was able to recover six proteins (corresponding to the genes products of *H2BC1, H2BC3, H2BC4, H2BC13, H2BC18, H2BC26*), while the Itzhak et al., 2016 dataset four proteins (corresponding to genes products of *H2BC4, H2BC11, H2BC12, H2BC13*). While

the canonical histones are likely to be expressed simultaneously in the cell, these discrepancies may be due to the sensitivity of MS-based analysis or expression variation in cell lines.

Taken together, our analysis indicates that MS-based chromatin datasets exhibit inconsistencies with subcellular localization or functional annotations available in UniProt, HPA, or GO. Up to 35% of chromatin proteins identified by MS-based techniques lack known nuclear localization according to the databases. Among these, approximately 30% of proteins identified simultaneously by at least three experimental studies may possess an additional nuclear localization that is not currently captured by the databases, but may have been reported in research papers. For other proteins, the most parsimonious explanation of their presence is the considerable contamination of chromatin extracts by mainly cytoplasmic proteins. From another point of view, many chromatin proteins annotated in the databases are not present in the MS-based datasets. While this is partially due to the limited number of genes expressed in the analyzed cell lines, our analysis also suggests that many proteins (such as housekeeping transcription factors) are lost during chromatin extraction, likely due to the dynamic nature of their interactions and their low expression. Finally, we showed that MS-based studies that filter their results by selecting the proteins that are highly enriched in the nucleus with respect to other cellular compartments, or use ML-assisted classification based on database/literature data, show better consistency with the information in the databases. However, this comes at the expense of decreasing the size of their datasets and likely limiting their ability to identify new proteins associated with the nucleus/chromatin.

2. The SimChrom chromatin protein classification, the interactive SimChrom database and other reference datasets

This section supplements section **3.2. The SimChrom chromatin protein classification, the SimChrom dataset and other reference datasets** in the main text. Note: **Interactive Figure 3** (https://simchrom.intbio.org/#classification) is the interactive version of **Figure 3,** which is the key source of information for the analysis presented below.

The largest subgroup of the SimChrom “non-histone proteins" category is the “DNA-templated transcription” group (1547 proteins), which consists of the proteins belonging to the “Regulation of transcription” subgroup (1514 proteins) and "RNA polymerases" subgroup (34 proteins). The "DNA metabolic processes" form the second largest protein subgroup (495 proteins) and include DNA replication, repair, and recombination machinery. “Nuclear RNA binding proteins" is another major group of proteins (309 proteins) present in SimChrom. There are a lot of RNA binding proteins in the cell (more than 1500 [26]), in the “Nuclear RNA binding proteins" category we aimed at including only

those that are found inside the nucleus (see **Methods Section 2.2** and **Supplementary Table ST4**), including those involved in preribosome formation, RNA processing and modification inside the nucleus. Other major subgroups of our classification included “Histone modification” (257 proteins), "DNA-acting enzymes" (258 proteins including DNA methylation and demethylation enzymes), “Centromere-associated” (241 proteins) (see **Figure 3**). Several specific subgroups of various scope containing proteins important for chromatin functioning were also included in our classification, such as “Histone chaperones" (32 proteins), ATP-dependent "Chromatin remodelers" complexes (114 proteins), “SMC complexes” (28 proteins, including cohesins implicated in chromatin loop extrusion).

To evaluate the contents of our SimChrom dataset we performed its cross-comparison to the localization-based datasets described above (NULOC_JT_NECF and NULOC_CS) (see **Supplementary Figure SF3_3**). The vast majority of SimChrom entries was also present in the NULOC_JT_NECF dataset (the broadest dataset that combined all nuclear protein entries from all protein localization databases at any level of confidence), which is consistent with chromatin proteins being only a subset of nuclear proteins. However, a minor subset of SimChrom (156 proteins) was not classified as nuclear by the localization databases (**Supplementary Figure SF3_3A**). To further understand the nature of this minor discrepancy we analyzed the SimChrom categories which contributed the most entries to this subset (**Supplementary Figure SF3_3C**) or where a significant proportion of entries in the respective category was absent in the localization databases (**Supplementary Figure SF3_4A**). In this subset 32 entries out of 156 were not annotated by the localization databases at all, 27 were considered by UniProt as only chromosomal (this is consistent with “Centromere-associated” category of SimChrom having the most entries not present in NULOC_JT_NECF), and 124 had only non-nuclear localization according to the localization databases. A manual review of the latter entries suggested that they included both bona fide nuclear proteins (such as histone acetyl and methyltransferases), other proteins such as ribosomal proteins, various kinases and proteins involved in mitochondrial DNA processing. Further research and data would be needed to clarify the localization status of the latter entries. The SimChrom category having the largest proportion of proteins absent from the localization databases was “Histone tail cleavage” (**Supplementary Figure SF3_3C**). This is consistent with the fact that for many of the histone tail cleavage enzymes (*e.g.*, metalloproteinases, cathepsins, neutrophil elastase) the histone tail cleavage activity in the nucleus is not their primary function and manifests only in specific conditions and cell development stages [9]. The comparison of SimChrom with the NULOC_CS (our stringent high confidence consensus dataset of nuclear proteins) showed a sufficiently higher number of SimChrom proteins that were not included in the NULOC_CS dataset (1208) (**Supplementary Figure SF3_3A**). This is not unexpected since only 44% of the human proteome has simultaneous localization annotations at sufficient levels of confidence from UniProt and HPA (see **Results Section 3.1**). The intersection of SimChrom and NULOC_CS datasets included 1837 proteins (**Supplementary Figure SF3_3A**), and may be

considered as a set of chromatin proteins with a high level of confidence. To further validate that the subset of SimChrom that was not present in NULOC_CS (1208 proteins) represented nuclear proteins we performed GO enrichment analysis of this subset against a list of all GO-terms and then select non-nuclear associated terms for further analysis (see **Methods Section 2.2**, **Supplementary Figure SF3_4B**, **Supplementary Table ST8**). The analysis confirmed the low number of SimChrom proteins that were associated with *bona fide* non-nuclear GO categories (the "Centromere associated proteins" having the highest number – around a dozen out of 156 – of proteins that had "non-nuclear" GO-annotation terms, that belongs to charged multivesicular body proteins and dynactin complex subunits). Finally, as a byproduct of NULOC_CS and SimChrom comparison we find that 1459 proteins were present in NULOC_CS but not in SimChrom - this set may be considered as both a high-confidence set of nuclear non-chromatin proteins and a set of proteins that should be added to SimChrom. The GO analysis of this dataset did not reveal any clear GO categories that should have been included in SimChrom as categories related to chromatin functioning (see **Supplementary Table ST9**).

Limitations of the proposed chromatin classification SimChrom and its contents include the following: the absence of cell cycle control proteins and checkpoint signaling proteins, the lack of detailed classification for proteins involved in reading, writing and erasing DNA and RNA modifications. The 'Genomic location' categories require additional curation supported by experimental evidence to enhance the accuracy and reliability of their protein content. In addition, we did not consider the protein components of nonmembrane nuclear organelles whose proteins may also functionally interact with nucleic acids (directly or through phase separation). The classification does not include protein isoforms. Also, SimChrom is limited currently to human proteins only. These limitations will be addressed in the future versions of SimChrom.

3. Analysis of the human chromatome

3.1. The chromatome composition and abundance of chromatin proteins

This section supplements and expands section **3.3.1. The chromatome composition and abundance of chromatin proteins** in the main text.

To understand chromatin functioning it is important to know the chromatome content not only in terms of the set of proteins associated with chromatin, but also in terms of their abundance (*i.e.*, the (relative) number of proteins per cell or organelle). Hence, we aimed at analyzing the available mass spectrometry data to address this question. The analysis of MS protein intensities from the experimental chromatome/nucleome studies discussed above, revealed a high degree of variability (see **Figure 4A**, **Supplementary Figure SF4_1**). For instance, the estimated relative mass of histone proteins varied

from 0.1 to 58 % depending on the study, suggesting a high degree of bias due to different experimental techniques and analysis pipelines used to process raw mass spectrometry data (see **Figure 4A**). Hence, for further analysis we relied on the "whole-organism" protein abundance information available in PaxDB for *H. sapiens* [27]. PaxDb provides high quality information on protein abundance combined from many experiments with high coverage, dynamic range and interaction consistency (estimated consistency of abundance data with data on protein functional interactions) integrated over many cell types and conditions. Among the datasets available in PaxDb we have chosen two whole-organism datasets: the dataset with the highest proteome coverage ("H.sapiens - Whole organism (Integrated)" - covers 99% of human proteome according to PaxDb, referred to as "PaxDb_INT" in this paper) and the dataset with the highest interaction consistency score ("Whole organism, SC (Peptideatlas,aug,2014)" - covers 84% of human proteome according to PaxDb, referred to as "PaxDb_PA" in this paper), see **Supplementary Figure S4_1A**. Our analysis showed that with respect to PaxDb_PA, PaxDb_INT dataset has additional abundance information for around 2700 human proteins that almost exclusively have low levels of expression (less than 1 ppm, see **Supplementary Figure SF4_1B**). Among these proteins there are up to around 700 nuclear/chromatin proteins, hence we opted to use PaxDb_INT for general characterization of the abundance distribution of chromatin/nuclear proteins (presented in **Figure 4B**). PaxDb_PA dataset showed a higher consistency with respect to the relative abundance of functionally interacting chromatin proteins. The total abundance of different types of histone proteins (H3, H4, H2A, H2B) matched their expected equimolar ratio (see **Supplementary Figure SF4_2A,B**). Hence, PaxDb_PA was used for a detailed analysis of chromatin protein abundance distribution between chromatin protein groups and individual proteins (**Figure 4C,D**).

Around half of the whole-organism human proteome consists of low abundance proteins with expression levels of less than 1 ppm (~50%, see **Figure 4B**, **Supplementary Figure SF4_1C**, **Supplementary Table ST10**). The whole-proteome abundance distributions are positively skewed towards the low abundant proteins. In PaxDb_INT dataset this skewness is additionally supplemented by a second peak at the low abundance values (**Figure 4B**, **Supplementary Figure SF4_1A**). Among the low-abundance proteins only 25% of them correspond to the housekeeping proteins, while the proportion of housekeeping proteins among high-abundance proteins (abundance of more than 1 ppm) is 68% (see distribution in **Figure 4B**, **Supplementary Figure SF4_1C**, and **Supplementary Table ST10**). We used PaxDb_INT data to analyze abundance distributions of database-derived chromatome/nucleome protein sets discussed in the previous sections of the paper.

The NULOC_CS, NULOC_JT, and SimChrom datasets reported above all manifested distributions mirroring that of the whole proteome for PaxDb_INT data (see **Figure 4B**) with a significant proportion of proteins in these datasets still represented by the low-abundance proteins (40%, 44%, and 48%, for NULOC_CS, NULOC_JT, and SimChrom, respectively, see **Supplementary**

Figure SF4_1C, Supplementary Table ST10). We next aimed at understanding the types of proteins contributing to the low and high-abundance portion of the nucleome/chromatome. The proportions of house-keeping/non-housekeeping proteins in low- and high-abundance fractions for different database-derived datasets are given in Supplementary Figure SF4_1C and show that around 60% and 40% of chromatin/nuclear proteins are housekeeping ones for the high- and low-abundance fraction, respectively. As discussed above nuclear and chromatin databased-derived protein sets are on average enriched in housekeeping proteins with respect to the whole proteome (~58.5 % vs ~47%, see Supplementary Results and Discussion Section 1.3, Supplementary Figure SF4_1C, Supplementary Figure SF2_7A,B). This increase in the proportion of housekeeping proteins stems from both the increase of the number of low-abundance and high-abundance housekeeping proteins relative to the respective total numbers of low-abundance and high-abundance proteins in nucleome/chromatome datasets. A detailed analysis showed that the increase of the fraction of housekeeping proteins among the low-abundance ones was more than expected, while for the high-abundance ones was less than expected (under the assumption that high- and low-abundance fractions should contribute to the increase proportionally to the number of housekeeping proteins belonging to these fractions, see Supplementary Figure SF4_1C). For instance, under parsimonious considerations the overall increase in the fraction of housekeeping proteins for SimChrom with respect to the whole proteome (60% vs 47%) should imply the increase of house-keeping proteins' fraction among the low-abundance proteins from 25% to 32% (25*60/47=32), yet an increase to 38% was observed. **This highlights the important role that low-abundance housekeeping proteins play in chromatin functioning.** A more detailed analysis revealed that 64% of these housekeeping low-abundance chromatin proteins belong to the housekeeping DNA-binding transcription factors group (Supplementary Figure SF4_1D).

We next applied similar analysis to the sets of chromosome/nucleosome proteins identified in MS-based studies. The resulting distributions differed considerably from the distributions of database-derived protein sets discussed above, having a single maxima centered at higher values of abundance (see Figure 4B). **This fact again suggests that MS-based studies of chromatin are able mainly to recover highly expressed proteins and miss low expressed proteins** (low-abundance proteins are in the range of 1%-27% of the identified protein sets, Supplementary Table ST10, Supplementary Figure SF4_1C). The protein sets recovered by MS-based studies were significantly enriched in housekeeping proteins when compared to database-derived datasets (Supplementary Table ST10, Supplementary Figure SF4_1C). The abundance distributions varied between different MS-derived chromatin/nucleome protein sets. The chromatome studies based on extraction and/or cross-linking techniques (Alabert et al., 2023; Shi et al., 2021; Ginno et al., 2018; Torrento et al., 2011) had distributions shifted towards higher values of abundance, than the studies by Kustatscher et al., 2014; and Ugur et al., 2023 suggesting that the latter studies were also able to capture more chromatin proteins

with low abundance. The nucleome study by Itzhak et al., 2016 was also able to capture more lower-abundant proteins.

We next aimed at understanding the abundance of different chromatin protein groups and individual chromatin proteins in the cell relying on our SimChrom-SL classification using PaxDb_PA abundance dataset. The resulting diagrams depicting abundance variations of chromatin proteins, belonging to different SimChrom-SL categories, the number of proteins belonging to the respective categories, and the cumulative abundances (calculated both as the total number of protein molecules and the total molecular weight of protein molecules belonging to each SimChrom-SL category) are presented in **Figure 4C**. To gain additional insights into the functioning of chromatin in **Figure 4D** we plotted the abundance values of highly expressed chromatin proteins (abundance of more than 1% of the H4 histone abundance) belonging to SimChrom-SL categories of the "Molecular function" or "Physico-chemical properties" type. Abundance data for all histone proteins and non-histone chromatin proteins with abundance more than 0.01% of Histone H4 is presented in the **Supplementary Table ST12**. It is important to note that many chromatin proteins have additional localization in other cellular compartments, hence the presented data reflects the overall abundance of the chromatin proteins in the cell rather than their abundance in the nucleus. To shed more light on the protein abundance in the nucleus we have also built diagrams analogous to **Figure 4C** only for 802 SimChrom proteins that are uniquely localized in the nucleus (according to our NULOC_CS_UL dataset) (see **Supplementary Figure SF4_3A**). These proteins are also highlighted in **Figure 4D**. As seen in panel 1 of **Figure 4C** chromatin categories vary substantially by their median abundance from 0.09 ppm to 570 ppm and there is still considerable variation in the abundance values within the categories. The most abundant chromatin protein is histone H4 (~11000 ppm), which is expressed by a family of genes almost exclusively coding the same protein sequence (except for H4C7, which has a negligible abundance). It is convenient to measure the abundance of all other proteins in fractions of H4 abundance (see **Figure 4D**). Each nucleosome contains two copies of H4 histones, therefore the numbers are also easily converted to relative abundance of chromatin proteins per nucleosome. Detailed analysis of histone protein abundance is in **Supplementary Table ST12** and shown in **Supplementary Figure SF4_2A,C**. The total number of core nucleosomal histone types H3, H4, H2A, H2B expressed by various genes sums up to similar numbers (~10400-10900 ppm) consistent with their equimolar association within nucleosome core particles. The cumulative abundance of H1 histones (~4500 ppm) suggests that slightly less than one H1 histone is associated with each nucleosome. The most abundant core histone variants are H3.3 (23% of H4, 2530 ppm), H2A.X (6.5%, 714 ppm), H2A.Z (10%, 1140 ppm), H2A.W (3.8%, 423 ppm). The least abundant histone variants are H2A.B and H1.7 (less than 1 ppm). Despite the relatively small number of protein coding human histone genes (108), many of which code for identical sequences, the cumulative abundance of histone proteins exceeds that of all other chromatin protein categories even if proteins with multiple localization are taken into account (see panel 2,3 in

Figure 4C). However, when the total molecular weight of proteins belonging to different categories is compared, the relatively small size of histone proteins (median ~15 kDa) results in them yielding the first place to RNA processing proteins (see panel 4, Figure 4C). Collectively the cumulative weight of proteins belonging to "Nuclear RNA binding proteins" category (that combines Preribosome-associated, RNA modification, and RNA processing categories) amounts to 30.4% of all SimChrom proteins weight (4.8% of whole-organism proteome weight). However, many proteins from these categories are also localized in cytoplasm, and the major contribution to their cumulative molecular weight likely comes from the cytoplasmic fraction. If the same analysis is performed only for the SimChrom proteins that are uniquely localized in the nucleus (Supplementary Figure SF4_3A), the mass fraction of histone goes up to 38% of all the chromatin proteins that are uniquely localized in the nucleus.

Other functional chromatin protein groups (or groups with specific properties) with high values of median abundance and high level of individual protein abundance include HMG A/B/N, histone tail cleavage, histone chaperones, chromatin remodelers and other categories (see Figure 4D). The high mobility group proteins (HMG A/B/N) are the second group after histones ranked by their median abundance. Although grouped together due to historical reasons, they include three separate superfamilies: HMGA (contains AT-hook domains), HMGB (contains DNA binding HMG-box domain), and HMGN (contains nucleosome binding domain). Our analysis suggests that the ratio of HMG proteins to nucleosomes is 1:8, 1:2, 1:3 for HMGA, HMGB, or HMGN proteins, respectively. However, the majority of HMG proteins are not exclusively localized in the nucleus (Figure 4D). The histone tail cleavage proteins are another small group of proteins in our classification with high median abundance in the whole-organism proteome. These enzymes, however, are not exclusively specific for histone cleavage, and likely perform their main functions outside the nucleus by cleaving other proteins. Among histone chaperones the H3-H4 histone chaperone NPM1 and H2A-H2B histone chaperone NCL have the highest abundance, 32% and 13% of H4 abundance, respectively. The most abundant histone variant specific chaperone is ANP32E (specific for H2A.Z-H2B with abundance of 2%). The whole nucleosome chaperone FACT complex consisting of SSRP1 and SUPT16H gene products, has an abundance of around 1%, amounting to one FACT complex per around 50 nucleosomes. Among RNA polymerase subunits POLR2E the common subunit E of RNA polymerases I, II, and III is the most abundant protein (0.58% of histone H4 abundance or around 1 per 90 nucleosomes). The exclusive components of polymerase II (POLR2B, POLR2C, POLR2D, and others) have their abundances in the range of 0.05-0.3%. With a median human gene length of 24kb and nucleosomal repeat length of around 200 bp this gives a lower estimate of one polymerase II per approximately 10 genes. Among genes involved in chromatin remodeling actin encoding genes (*ACTB*, *ACTA1* and actin-like *ACTL6A*) are leading by the abundance of their protein products. While actin is a component of some chromatin remodeling complexes (e.g., SWI/SNF) the major contribution to its abundance clearly comes from its

cytoplasmic fraction involved in cytoskeleton formation [28]. The next by abundance are RUVBL1/RUVBL2, members of the family of ATPases associated with diverse cellular activities (the so-called AAA+ proteins), their abundance is around 1.6-1.9%. They form heterohexameric subunits in INO80 remodeller complexes [29], but this is not their exclusive function [30]. The same is likely the case for UCHL5, a putative and optional component of INO80 having relative abundance of 0.6%. The main ATPase of the INO80 family (*INO80* gene) has a dramatically lower abundance (0.006%, suggesting the presence of one INO80 complex per 8680 nucleosomes). The other more abundant chromatin remodeler families include the ISWI family (ATPase subunit SMARCA5 and auxiliary subunits DDX21, SF3B1, MYBBP1A having abundance of 0.7%, 1.6%, 1.2%, 0.6%, respectively), the CHD family (ATPase subunit CHD4 – abundance of 0.6%), the SWI/SNF family (ATPase subunits SMARCA4 or SMARCA2, and axillary subunits SMARCC2, SMARCC1, SMARCE, having abundance of around 0.2 and 0.6, respectively). By the abundance of their catalytic subunits we estimate the presence of one ISWI, CHD, or SWI/SNF remodeler complex per around 70, 80, or 200-300 nucleosomes, respectively.

Among genes involved in formation of SMC complexes the leading are SMC3 and SMC1A with abundance of around 0.8%. These genes are known to form mitotic cohesin complex together with RAD21 (abundance of 0.46%), which has been recently found to perform DNA loop extrusion process in the human genome needed to maintain 3D genome organization and topologically associating domains [31]. Based on SMC3 abundance we estimate the presence of one SMC1-SMC3 complex per around 60 nucleosomes. Among PTM readers the most abundant are the HP1α, HP1β, and HP1γ proteins involved in recognition of methylated H3K9 and formation of heterochromatin (respective abundances are: HP1γ (*CBX3* gene) 3.65%, HP1β (*CBX1* gene) - 1.82% , HP1α (*CBX5* gene) - 1.33%). Among Histone PTM writers the most abundant is PARP1 which PARylates many targets including histones (3.97%). Other nonspecific PTM writers with high abundance are a group of kinases (STK4, PRKDC, PAK2,CDK1) and FBL methyltransferase, which methylates both RNA and histones (particularly at H2AQ104me [32]). Other known histone PTM writers with relatively high abundance include PRMT1 (involved in H4R3me1, H4R3me2 deposition), PRMT5 (involved in H2A, H4R3, H3R8 methylation), NAA50 (involved in H4 acetylation), TGM2 (catalyzing serotonylation and dopaminylation of histone H3) having abundance of 1.5%, 0.7%, 0.87%, 1.67% respectively. The abundance of Polycomb group (PcG) subunits, in particular, PRC1 complex, which monoubiquitinates H2AK119, is around 0.02−0.18%, suggesting one complex per 275 nucleosomes (based on RNF2 and CBX2/4/6/7/8 subunits). Subunits of PRC2 complex which trimethylates H3K27 approximately twice lower: abundance of 0.08%−0.09% of histone H4 for core subunit EZH2 and EED, SUZ12 (suggesting one complex per 550 nucleosomes). The most abundant Histone PTM erasers contain deacetylases: nonspecific to histones HDAC4 (abundance of 2.28%), HDAC2 (abundance of 1.337%), HDAC1 (abundance of 1.108%), and component of remodeling complex NuRD: MTA2 (abundance of

0.597%)), phosphatases (PPP1CC, PPPCA, PPP2CB, PPP5C, abundance of 0.6%−1.2%). The most abundant demethylase is JMJD6 (acts on H3R2me, H4R3me) - abundance of 0.18%.

The groups with least median abundance are those related to "DNA-binding transcription factors" (pioneer TFs, nuclear hormone receptors, housekeeping and non-housekeeping TFs) (note that this SimChrom-SL category excludes this group). Non-housekeeping DNA binding transcription factors have an abundance between 0.00008 and 23.9 ppm, suggesting their expression at minimal levels when averages across all body tissues. The majority of housekeeping transcription factors is also expressed only marginally (median abundance 0.314 ppm). However, the abundance of certain proteins classified as housekeeping TF may reach 91.8 ppm (RURB1 gene, 0.84% of H4) for TF uniquely localized in the nucleus or 3667 ppm for proteins that have multiple localization (ENO1 gene, 33.6% of H4). The DNA-binding transcription factor groups have the largest number of genes in SimChrom-SL (1345 genes in total), however, their contribution to the cumulative weight of chromatin proteins is a rather small 6.3%.

We also performed a similar abundance analysis of chromatin proteins grouped by SimChrom-SL categories for the MS-based experimental datasets, where the abundance values were taken from the respective studies (see **Supplementary Figure SF4_4**). The mass fraction of SimChrom proteins in experimental chromatomes/nucleosomes was between 33% and 81% (Alabert et al., 2014 - 65%; Kustatscher et al., 2014 - 81%; Itzhak et al., 2016 - 77%; Ginno et al., 2018 - 33%; Shi et al., 2021 - 78%; Ugur et al., 2023 - 43%). This reflects the fact that among the proteins detected in MS-based studies, SimChrom proteins constitute only a fraction (see **Supplementary Figure SF2_5A**). In Ginno et al., 2018 and Ugur et al., 2023 the total mass fraction was dominated by these non-SimChrom proteins. In Shi et al., 2021 and Alabert et al., 2014 the significant mass fraction was provided by one group of proteins (e.g., by "Histones" in Shi et al., 2021 and "DNA-acting enzymes" in Alabert et al., 2014). Itzhak et al., 2016 dataset (in which proteins from predominantly nuclear fraction were estimated in copy number per cell) showed a more balanced distribution of mass fraction between different SimChrom-SL categories: Nuclear RNA-binding proteins (26%), Histones (19%), Chromatin remodelers (6.4%), and Histone PTM writers (3.4%). Taken together our analysis revealed a high level of variation and potential experimental biases in quantitative proteomics data provided by MS-based studies.

3.2. Detailed analysis of the physico-chemical properties and amino acid composition

This section supplements and expands section **3.3.2. Physico-chemical properties and amino acid composition** in the main text.

Using the constructed datasets of chromatin, and uniquely localized nuclear and cytoplasmic proteins (SimChrom, NULOC_CS_UL, CYTLOC_CS_UL) we analyzed the physico-chemical properties of chromatin/nuclear proteins, their distinction from cytoplasmic proteins and peculiarities of chromatin proteins belonging to different classes of the SimChrom classification. **Figure 5A,B** summarizes the results, details are provided in **Figure 5C-L** and **Supplementary Figures SF5_1 - SF5_5**.

The length distribution of chromatin proteins is positively skewed with respect to that of cytoplasmic proteins but only in the range of protein length less than around 800 amino acids (**Figure 5C**). The same picture is seen for nuclear proteins (**Supplementary Figures SF5_1A**). The median length of cytoplasmic proteins is 464, while 496 for SimChrom and 500 for nuclear proteins (**Supplementary Figure SF5_1A** for nuclear vs cytoplasm comparison). Hence the median length of chromatin proteins is 7% longer than that of cytoplasmic proteins. The fraction of relatively short proteins (less than 200 aa) is 1.6 times smaller for chromatin proteins than for cytoplasmic ones (13% vs 8%), suggesting that smaller proteins are on average underrepresented in chromatin with respect to cytoplasm. The longest chromatin proteins in our dataset were those from following SimChrom categories: 'Preribosome associated': (MDN1 - 5596 aa); 'Histone PTM writers' (Histone methyltransferases: KMT2D (5537 aa), KMT2C (4911 aa), KMT2A (3969 aa), E3 ubiquitin-protein ligase HUWE1 (4374 aa), kinase PRKDC (4128 aa)); 'TFs': TRRAP (3859 aa), ZFHX3 (3703 aa), ZFHX4 (3567 aa), ZNF292 (2723 aa), HIVEP1 (2718 aa); 'Chromatin remodeler' subunits: SRCAP (3230 aa), EP400 (3159 aa), CHD7 (2997 aa), CHD9 (2897 aa), CHD6 (2715 aa), while the shortest from 'HMG', 'Histones' (90-130 aa) and 'RNA polymerase' subunits: POLR2K (58 aa), POLR2L (67 aa).

The fraction of intrinsically disordered regions in chromatin and nuclear proteins (as calculated by AlphaFold2 SASA with 20-residue smoothing, see **Methods Section 2.4**) and its distribution was significantly different from the one for the cytoplasmic ones (see **Figure 5D**). 46% of chromatin and 51% of nuclear proteins had an IDR fraction of more than one half, while for cytoplasmic proteins the value was only 23%. The median IDR fractions also differed considerably (46%, 51%, 23%, correspondingly). Further analysis showed that the increase of IDR fraction was both due to the increase in average IDR length and the number of IDRs (see **Figure 5E,F**). The fraction of proteins with two, three or four IDRs was significantly higher for chromatin/nuclear proteins than for cytoplasmic ones (e.g., fraction of proteins with two IDRs was 38%, 41%, 27% for nuclear, chromatin and cytoplasmic proteins, respectively) mostly at the expense of a lower fraction of proteins without or with one IDR (e.g., fraction of proteins with IDR was 20%, 18%, 33%, respectively). A similar tendency was observed for the increase in the number of non-IDRs in chromatin/nuclear proteins with respect to cytoplasmic ones, albeit to a lesser extent (see **Figure 5F**).

We performed a detailed analysis of protein length and IDR fraction for different categories of chromatin proteins (see **Figure 5G**). The protein categories with the shortest median length are histones (130 aa), HMG proteins (109 aa), and RNA polymerase subunits (220 aa). The longest ones are SMC complex subunits (1096 aa), Chromatin remodelers (835 aa), Histone PTM writers (766 aa), Histone PTM erasers (747 aa), DNA-acting enzymes (655 aa). The categories with the largest variation in size (ratio of 5% and 95%-percentiles) include Histone PTM readers (1741), General TFs (1746), Histone chaperones (1768), DNA-acting enzymes (1810), DNA (de)methylation (1830), Centromere-associated (2013), Chromatin remodelers (2314), Histone PTM writers (2988). The categories with the smallest variation in protein length (calculated for protein categories with more than 15 entries) include Histones (144), Pioneer TFs (321), Nuclear hormone receptors (598), Non-Housekeeping DNA-binding TFs (766), RNA modification (768), Methylated DNA binding (856), Housekeeping DNA-binding TFs (874), Other families with HMG-box domain (943).

The median IDR fraction was smaller than that of cytoplasmic proteins (13%) for proteins in the following categories: RNA polymerase subunits (0%), Histone tail cleavage (3%), and Preribosome-associated (10%). Other groups with relatively small fraction of IDR included: DNA-acting enzymes (16%), RNA modification (16%), DNA de_methylation (17%), SMC complex subunits (18%), DNA repair (19%). The categories with the highest fraction of IDRs are HMG (99%, 7 out of 11 proteins are completely disordered), Other families with HMG-box domain (74%), Pioneer TFs (73%), Methylated DNA binding (62%). The analysis allowed us to identify the chromatin categories contributing most to the high IDR fraction proteins observed in the distribution in **Figure 5G**. The large categories that contain many proteins with high IDR fraction are those belonging to transcription factors, especially the “Non-Housekeeping DNA-binding TFs” (contributes 453 proteins with IDR fraction of more than 50%). The distribution of IDR fraction for chromatin proteins without the transcription factor groups (“Non-Housekeeping DNA-binding TFs” and "Housekeeping DNA-binding TFs") resembles that of the cytoplasmic ones much better (see **Supplementary Figure SF5_1B**). Overall our analysis suggests that chromatin and nuclear proteins are significantly enriched in the number and length of IDRs, which likely are important for fuzzy interactions happening in chromatin including liquid-liquid phase separation behavior. This enrichment is particularly evident for non-housekeeping transcription factors.

We also performed the analysis of IDR fraction distribution among the chromatin/nuclear protein sets identified in experimental MS-based studies discussed in **Results Section 3.1**. The Kustatscher et al., 2014; Itzhak et al., 2016; and Ugur et al., 2023 datasets (that were previously shown to include a substantial fraction of low abundant proteins) had a considerably higher proportion of proteins with high IDR fraction. This can be explained by the presence of a larger number of non-housekeeping TF in these datasets (see **Supplementary Figure SF5_1C**, **Supplementary Figure SF2_7C**).

It is generally assumed that chromatin proteins are on average positively charged to compensate for the negative charge of the genomic DNA. Using the protein datasets compiled above and the protein abundance data from PaxDB (see **Results Section 3.3.1**) we set out to analyze the distribution of chromatin/nuclear proteins with respect to their overall electrostatic charge. To this end we used three metrics (see **Figure 5H**). In the first approach we classified the proteins as negatively/positively charged or nearly neutral (see **Methods Section 2.4**) and calculated the number of protein entries in our datasets belonging to these classes. According to this approach positively charged proteins dominate among nuclear and chromatin proteins, while the negatively charged proteins dominate among the cytoplasmic ones. The dominant charge group in each case consists of around more than half of protein entries (54-58%), while the opposite one consists of 38-43% of protein entries. This metric, however, does not account for protein abundance or total protein charge. In the second approach we adjusted the first metric by protein abundance, while in the third approach we additionally took into account the overall net charge of every protein, resulting in the analysis of the cumulative charge conferred by positively or negatively charged proteins (see **Figure 5H**). Interestingly, only among the uniquely localized nuclear proteins the results were significantly affected by these two adjustments. The amount of positively charged proteins increased to 82% once the protein abundance was taken into account, and the overall net positive charge coffered by these proteins was 89%. This is mainly due to the high abundance and high net positive charge of histone proteins present in NULOC_CS_UL dataset. Surprisingly, for the SimChrom dataset the distribution of positively/negatively charged fractions did not change much as the result of these adjustments (43% vs 37% vs 39%). A detailed analysis of cumulative charges conferred by proteins from different SimChrom categories showed that significant cumulative negative net charge is contributed by proteins involved in RNA processing, Chromatin remodeling, Histone chaperones, and DNA-acting enzymes (see **Supplementary Figure SF5_1D**, left). These categories, however, have many proteins that are localized both in the nucleus and the cytoplasm, which reconciles these results with the charge distribution analysis of NULOC_CS_UL dataset discussed above. The same analysis of SimChrom proteins that are uniquely localized in the nucleus revealed a prevalence of positive charge by histone proteins (see **Supplementary Figure SF5_1D**, right). To further elucidate the nature of negatively charged proteins in chromatin we analyzed the fraction of entries belonging to negatively/positively charged proteins in each SimChrom category (**Supplementary Figure SF5_1E**). Among the categories having the most number of negatively charged proteins are Histone chaperones (84%), RNA polymerases (76%), Histone PTM erasers (71%). This surprising presence of many negatively charged proteins in these categories is likely explained by their preferential association not with DNA, but rather with positively charged histones.

To further elucidate the peculiarities of charge structure in chromatin proteins we analyzed the average charge profiles of protein N- and C-terminal tails. The median length of N-/C-terminal tails was 51/47, 72/84, and 67/79 for CYTLOC_CS_UL, NULOC_CS_UL, and SimChrom (see

Supplementary Figure SF5_2A,B). Only for the N-terminal tail there was a clear difference in the average charge profiles between cytoplasmic proteins and nuclear/chromatin ones (see Supplementary Figure SF5_2C,D, Figure 5I). The average charge of the N-terminal tails (less than 80 amino acids in length) was -16 for SimChrom proteins, -6 for nuclear proteins, and +13 for cytoplasmic ones (see Figure 5I). The analysis of the N- and C-terminal tail charge for different SimChrom categories revealed that they varied between the different categories (see Supplementary Figure SF5_2E,F). Histones had the most positively charged tails, while histone chaperones and HMGs had the most negatively charged ones. Transcription factors on average also had negatively charged protein tails. This is an interesting fact because most transcription factors are positively charged (Supplementary Figure SF5_1E).

We next set to analyze in detail the amino acid composition of chromatin/nuclear proteins with respect to cytoplasmic ones and the variability of amino acids composition between different groups of chromatin proteins (see Figure 5J,K,L, Supplementary Figure SF5_3, SF5_4, SF5_5, Supplementary Table ST13). To this end we first used the UMAP nonlinear dimensionality reduction technique to see if significant variations between chromatin proteins can be identified in the space of their amino acid composition. The resulting 2D projections onto the main UMAP components revealed that (1) chromatin and cytoplasmic proteins occupied overlapping domains on the 2D map, but with a visible shift between their centers, suggesting there is an overall difference in the average physico-chemical properties, (2) some chromatin proteins manifested a significant difference from others (see cluster 1 and cluster 2 in Figure 5K,L). Further analysis revealed that in the 2D UMAP map transcription factors, containing zinc finger domains and homedomains formed distinct clusters (see Supplementary Figure SF5_3A,B). The most distinct group (cluster 1) was **almost exclusively** (415 out of 422) composed of zinc-finger containing DNA-binding transcription factors (240 housekeeping and 175 non-housekeeping) with the median number of zinc-finger domains (ZFD) of around 10 (Supplementary Figure SF5_3C). Zinc-finger containing DNA-binding transcription factors were also present in cluster 2, but the median number of zinc-finger domains (ZFD) in that cluster was only three, hence containing a lower proportion of amino acids specific to ZFD (Supplementary Figure SF5_3C,D). ZFD are enriched in histidine and cysteine, which were among the top four mostly enriched amino acids in chromatin proteins (see Figure 5J and discussion below). Other protein groups that occupied distinct positions on the UMAP map, included (1) histones, (2) serine/arginine-rich splicing factors (SRSFs) (enriched in serine and arginine), and (3) reverse transcriptases of endogenous retroviruses (ERVs) (enriched in isoleucine and threonine) (see Figure 5K, Supplementary Figure SF5_3D).

The analysis of amino acids composition revealed that among the top four enriched amino acids in chromatin proteins were serine, cysteine, proline, histidine (Figure 5J, Supplementary Figure

SF5_5B). Their enrichment values are in the range 1.13-1.18. By classes of amino acids chromatin/nuclear proteins are mostly enriched in polar (N, Q, T, C, G, P), small (P, G, A, S), and positive (K, R) amino acids (**Supplementary Figure SF5_4A**, **Supplementary Figure SF5_5B**). It is important to note that such an analysis should be taken with a grain of salt, because the enrichment of certain amino acids may vary across different categories of chromatin proteins, and the categories with high number of protein entries have higher contribution to the overall average. To elucidate this variability we have also performed enrichment analysis for proteins in major functional SimChrom categories (see **Supplementary Figure SF5_4**).

Serine and proline are among amino acids that are relatively abundant in proteins (abundance of around 7-8% and 5-6% in chromatin and cytoplasmic proteins, respectively, **Supplementary Table ST13**). The total enrichment of serine in chromatin proteins is attributed simultaneously due to its enrichment in the IDR, non-IDR regions (relative to IDR and non-IDR regions of cytoplasmic proteins), and more importantly due to higher proportion of IDR regions in chromatin proteins (46% vs 23%) that in turn have a considerably higher proportion of serine than non-IDRs (**Supplementary Table ST13**). The slight enrichment of serine in IDRs was observed almost across all SimChrom categories (**Supplementary Figure SF5_4E**). In non-IDR regions serine showed both enrichment and depletion in certain categories, and the overall enrichment was driven mainly by transcription factors due to the large number of proteins in these categories. The total enrichment of proline in chromatin proteins is attributed due to its enrichment in IDRs (relative to IDRs of cytoplasmic proteins and more importantly due to higher proportion of IDR regions in chromatin proteins (proline is the most enriched amino acid in IDRs of both chromatin and cytoplasmic proteins versus the non-IDRs, fold enrichment 1.8-2.1, **Supplementary Table ST13**). The overall enrichment of proline in non-IDRs was close to one and statistically not significant. Certain small groups, such as histones and HMG proteins showed considerable deviations in proline content in their non-IDRs (**Supplementary Figure SF5_4H**). The enrichment of proline in IDRs was observed in many SimChrom categories (**Supplementary Figure SF5_4E**). In certain categories, such as Non-Housekeeping transcription factors and pioneer TFs enrichment was high (1.37 and 1.55 fold, respectively). Surprisingly, the enrichment of proline in IDRs of housekeeping TF was depleted (FE of 0.92). Suggesting that while there is still a considerable fraction of prolines in IDRs of housekeeping TF, this fraction is significantly lower than in IDRs of non-house keeping TF (7 % and 10.4 % median fractions, respectively).

Cysteine and histidine are among amino acids that have a relatively low abundance in proteins (abundance of around 1-2.5%, **Supplementary Figure SF5_5B**). The total enrichment of cysteine and histidine is mainly driven by the prevalence of zinc fingers containing transcription factors (**Supplementary Figure SF5_4B**). The exclusion of these proteins from analysis resulted in the disappearance of any statistically significant enrichment **Supplementary Figure SF5_5A**).

Interestingly, the enrichment of positive amino acids is only statistically significant for lysine, but not for arginine, and the enrichment is relatively moderate (1.03 in chromatin) (**Supplementary Figure SF5_4A**). Lysines are enriched in IDRs and non-IDRs of chromatin proteins, while arginines are depleted in IDRs and enriched in non-IDRs (when compared with IDRs and non-IDRs of cytoplasmic proteins) (**Supplementary Figure SF5_5B**). The overall higher positive charge of chromatin proteins stems also from the depletion of negatively charged amino acids in their sequence. The depletion of aspartate in chromatin/nuclear proteins is statistically significant (fold enrichment is around 0.9), while the depletion of glutamate is statistically non-significant (**Supplementary Figure SF5_4A**). This suggests that the increased positive charge of chromatin nuclear proteins has its main contributions in the depletion of aspartate, and moderate enrichment of lysine.

Within the IDR regions of chromatin proteins tyrosine and asparagine were also significantly enriched (FE of 1.23 and 1.17 versus the IDRs of cytoplasmic proteins, respectively), mainly due to the contribution of transcription factors (**Supplementary Figure SF5_4F**).

Among the most relatively depleted amino acids in chromatin/nucleus are tryptophan and hydrophobic/aliphatic amino acids like valine, isoleucine, leucine, and methionine (**Supplementary Figure SF5_4A**). Tryptophan is the rarest amino acid in proteins (around 1%). Certain categories like histone and HMG proteins lack it completely (**Supplementary Figure SF5_4C**). It is depleted in almost all chromatin categories, except for a few small categories such as DNA (de)methylation, histone tail cleavage, and histone modification, where the enrichment comes from non-IDRs (**Supplementary Figure SF5_4I**). Hydrophobic/aliphatic amino acids are depleted in IDRs vs non-IDRs of proteins and hence the large proportion of IDRs in chromatin proteins accounts for a lower fraction of these amino acids in chromatin proteins (**Supplementary Figure SF5_4F,I**).

The most enriched amino acids in the uniquely localized nuclear proteins were the same except for cysteine (this difference may be traced to the diminished number of transcription factors enriched in cysteines in the NULOC_CS_UL dataset, apparently because of their multiple localization, see **Supplementary Figure SF5_4A**, **Supplementary Figure SF4_3B**,).

3.3. Detailed analysis of the domain composition of chromatin proteins and identification of new structural domains

This section supplements and expands section **3.3.3. Domain composition of chromatin proteins and identification of new structural domains** in the main text.

Next we set out to systematically analyze the available data on structural characterization, domain annotation and domain composition of chromatin proteins. We specifically explored the structurally uncharacterized portion of the chromatome ("dark" proteome) and identified potential new structural domains that are predicted by AI-based protein structure prediction tools (see **Figure 6A**).

Historically, protein domains are loosely defined as evolutionary conserved units with similarities at functional, structural and/or sequence levels [33]. Domains may represent single proteins or exist in a variety of various sequence contexts. Sequences of related individual protein domains may be grouped and aligned to produce domain models. Domain models are catalogued and annotated by a number of resources/databases such as PFAM [34], CDD [35], CATH [36], InterPro [37], and may be further grouped into superfamilies, clans, folds, etc [38,39]. Domain models are usually defined through multiple sequence alignments (MSA) and corresponding hidden Markov models (HMM). In structure-based approaches (e.g. CATH/Gene3D database) domain superfamilies are assigned through grouping and alignment of available experimental 3D structures. The ultimate experimental structural characterization of chromatin proteins is available in the PDB database, however, recent progress in protein structure prediction spurred by AlphaFold resulted in new approaches to the structural characterization and discovery of new structural domains (*e.g.*, as implemented in the TED database) [40] (**Figure 6A**). Structure prediction algorithms combined with structure similarity search algorithms, such as FoldSeek [41], now allow to find remote homologs and assign individual domains to their respective superfamilies.

Figure 6B shows the fraction of the aggregate number of amino acids in all human chromatin proteins (referred below to as "aggregate chromatome sequence", or ACS) which are structurally characterized or have domain annotations in different databases. According to AlphaFold approximately one half of the ACS (47%) is predicted to be intrinsically disordered, or to become ordered within protein-protein complexes (the structurally uncharacterizable "dark" chromatome) (see **Methods**), and the rest as having distinct 3D structure. Direct experimental structural data in PDB is available for only one fourth of the ACS (20% of ACS are simultaneously considered ordered by AFDB and available in PDB). Hence, we envision that at least one third (34%) of the aggregate human chromatome sequence is amenable to characterization with structural biology methods but has not yet been characterized (constitutes the potentially structurally characterizable "dark" chromatome). The Pfam database (the largest sequence-based database of protein domains and protein families to date) has annotations for around 39% of the aggregate human chromatome sequence. The CATH database, which focuses on identifying and annotating structural domains, annotates 25% of ACS, while the automated AlphaFold-driven TED resource finds structural domains in 35% of ACS. The difference between the fraction of ACS annotated by TED and that considered ordered by AFDB was traced to at least several facts: 1) AlphaFold is known to be biased to predict long solitary alpha-helices which are

not considered domains by algorithms that identify structural domains, 2) the TED algorithm frequently fails to annotate repetitive regions that contain visually identifiable secondary structure elements within large multidomain proteins, 3) we identified non-IDRs as regions no less than 4 amino acids whereas median length of TED domain in human proteins were 108 aa. A caveat that has to be kept in mind, is that current automated analysis using AlphaFold is based only on predictions for single chain proteins, while in reality chromatin proteins engage in many intermolecular interactions. To some extent Pfam and CATH/TED are complimentary (see **Figure 6B**). In addition to 39% of ACS annotated by Pfam, TED annotates additionally 13% of ACS, and CATH adds annotations to 3% of ACS on top of it (yielding a combined annotation coverage of 55% by these three resources).

Next we analyzed the structural characterization of the aggregate human chromatome sequence from the point of view of structural domains present in chromatin proteins (as identified by the most comprehensive TED database, which automatically detects structural domains) (see **Figure 6C**). Chromatin proteins contain in total 6246 individual TED domains. Using FoldSeek and combination of FoldSeek and CATH resources (see **Methods**) we matched these domains to the structurally related domains in PDB or CATH superfamilies. The remaining domains were analyzed for the presence of previously uncharacterized structural folds/superfamilies and potential functional roles of these domains. Among the 6246 predicted structural domains constituting human chromatin proteins, 34% had exact matches in PDB structures (100% sequence identity, see **Methods**), 56% matched PDB structures of homologues with different levels of sequence identity (from 99% to 5%, for details see **Figure 6C**). The majority of these homologous domains were in fact different paralogous sequences found within human genes (even for domains with sequence identity of 35-50% the fraction of human sequences among the matches was 51%), for matches with sequence identity above 35% the second largest contribution came from structures of mammalian homologues, for matches with sequence identity below 35% significant contributions were from structures derived from proteins of fungi, protostomia and bacteria (see **Supplementary Figures SF6_1A** for details). Additionally, 6% of TED domains that lacked direct hits among the PDB structures were mapped to protein structural superfamilies in the CATH database (the information about potential sequence variation in each homologous superfamily collected in CATH database combined with AlphaFold structural predictions allowed to identify more distant structurally characterized homologues). The remaining 4% (241 TED domains) represented domains that could not be matched to any known protein structure or protein structure superfamily, potentially representing new types of structural superfamilies of even protein folds. These domains are presented in **Supplementary Table ST14** (see also **Interactive Table 3** at https://simchrom.intbio.org/#novel_structural_domains) and ranked via their structural complexity by the number of their secondary structure elements. Among these domains, 123 domains have annotations in Pfam or other domain annotation databases present in InterPro, leaving 118 domains that are completely without annotations. The latter domains belong to 106 chromatin proteins, which may be

considered as perspective new targets for experimental studies of their function and structure. Among such proteins are, for example, a protein encoded by the *GTF3C1* gene, a General transcription factor 3C polypeptide 1 (it has a previously unannotated and uncharacterized structural domain with length of 233 amino acids) (see detailed characterization in **Supplementary Figure SF6_2A**). Another instructive example is the globular domain of the testis specific linker histone H1.7 (product of *H1-7* gene, see **Supplementary Figure SF6_2B**). Despite the considerable amount of studies dedicated to the elucidation of the structure of H1-linker histones [42], the H1.7 histone variant (previously, named HANP1/H1T2) has a quite different sequence resulting in a predicted structure that has a different topology than other known H1 histones (the "wing" of the globular domain consists of three beta-sheets rather than two). The relation of this domain to the H1 histone family cannot be identified with conventional sequence analysis methods (such as those implemented in the Pfam database), however, it should be noted that new deep-learning-based annotation approaches (such as Pfam-N) are able to annotate it (see **Supplementary Figure SF6_2B**).

Next we predicted GO molecular functions and biological processes for mentioned above 118 not-annotated chromatin protein domains using DeepFRI [43], a Graph Convolutional Network for predicting protein functions by leveraging sequence features extracted from a protein language model and protein structures, see **Methods Section 2.5**. The top-7 common GO MF terms: ion binding, organic cyclic compound binding, heterocyclic compound binding, protein binding, cation binding, metal ion binding, nucleic acid binding. Top-10 GO BP terms: organic substance metabolic process, primary metabolic process, cellular metabolic process, nitrogen compound metabolic process, macromolecule metabolic process, organonitrogen compound metabolic process, regulation of cellular process, cellular macromolecule metabolic process, cellular nitrogen compound metabolic process, cellular response to stimulus. 9 out of 118 TED domains lacked the predicted GO molecular function by DeepFRI: two of them were in members of the regulatory factor X (RFX) family of transcription factors (encoded by genes *RFX1, RFX5*).

Many chromatin proteins contain similar, evolutionary related individual protein domains whose kinship may be identified by matching them to the same Pfam domain sequence models. Hence, we used the Pfam domain annotation to characterize the diversity of protein domains found in chromatin proteins and typical domain composition thereof. In total 1753 different domain types (sequence models) were identified in chromatin proteins (**Figure 6D**). Next we analyzed the structural information available for these models. 76% of domain models had at least one individual domain among chromatin proteins that could be matched to a PDB structure using FoldSeek (*bona fide* structural domain in **Figure 6D**). To characterize the comprehensiveness of the structural characterization of each domain model we estimated the median sequence identity between all individual domains in chromatin proteins belonging to the said domain model and their best matches in PDB found via FoldSeek (see **Methods**

Section 2.5, Figure 6D). 42% of domain models were considered fully characterized, *i.e.* every individual domain in chromatin proteins belonging to these models can be found in PDB, 34% of domain models are partially characterized. 14% of Pfam domains were not matched by FoldSeek to PDB structures, but could be still identified in PDB via sequence search methods – these represented IDR regions, repeats, etc. 3% (55 domain models) did not match any PDB structure but could be matched to structural domains predicted by AlphaFold and found in the TED database. These represent prospective targets for validation with structural biology methods and further investigation of their interactions. For instance, among these domains are domains, potentially associated with chromatin remodeling (SANTA, zf-C3Hc3H), histone PTM writing (DUF7030, COMPASS-Shg1), zinc fingers (zf_CCCH_4, zf-LITAF-like, zf-WIZ, SWIM) etc. 7% of Pfam domain models currently have no structural information that can be assigned either through the PDB or TED databases.

We next analyzed the diversity of Pfam domains in various SimChrom-SL protein categories (Figure 6E, subpanels 1,2) and the domain content of individual proteins belonging to these categories (Figure 6E, subpanels 3-5). Pfam identified 11147 individual domains in chromatin proteins belonging to 1753 domain types (Pfam models); only 70 chromatin proteins had no domain annotation at all. For the distribution of the total number of Pfam models and distinct Pfam models identified in chromatin proteins, see Supplementary Figure SF6_1B, Supplementary Figure SF6_1C. Expectedly, large SimChrom-SL categories consisting of more than one hundred proteins harbored the largest number of different domain types (*e.g.*, DNA-acting enzymes, histone PTM writers, transcription factors, etc.), while the smaller categories had less (see Figure 6E, subpanel 1). Although Pfam may not be comprehensive in its annotation, we estimated the number of distinct domain types per protein in each category (relative domain diversity, Figure 6E, subpanel 2). The average domain diversity was around one for all categories. The categories with the considerably lower domain diversity were transcription factors categories (their variability relies on different combinations of zinc-finger domains that are described through only a few Pfam domain models), histones (their functional variability is often conferred by only small changes in the sequence), and HMG-constainting proteins (this is a very small group of proteins with only nine proteins and three corresponding Pfam models). The median number of Pfam domains in human chromatin proteins was two (which corresponds to the structure based domain analysis presented above). Certain chromatin protein categories had a higher median number of domains, including Housekeeping TF, histone PTM writers and readers (Figure 6E, subpanel 3). Interestingly the median number of domains for Non-housekeeping TF was two (they more often rely on single homeodomains than on cassettes of zinc-finger domains), although the group is diverse and proteins with as many as 32 domains were present. This, however, is again explained by the large number of zinc-finger domains that may be present in such proteins (see Supplementary Figure SF6_1D). The presence of additional domains in PTM writers and readers may be hypothesized to have evolved due to the functional necessity for multivalent binding to different chromatin structures (see

below for a detailed analysis). Some categories mostly consist of single domain proteins, such as Centromere-associated, DNA repair, Regulation of transcription, RNA polymerases (but this group also includes proteins with a maximum of 42 domains), DNA recombination, RNA modification, Histones, HMG_A/B/N, etc. The analysis of the number of distinct different domain types present in chromatin protein categories corroborates the above mentioned analysis (**Figure 6E**, subpanel 4). Proteins from PTM writers group have the median number of three distinct domain types, while all other categories have less. Still many chromatin proteins harbor many distinct domain types, DNA-acting enzymes, histone PTM writers, chaperones, remodelers, transcription factors with as much as 8-9 distinct domains are present (see **Supplementary Table ST15**). There are 118 chromatin proteins harboring at least five different domain types (see **Supplementary Figure SF6_1C**). This highlights the multivalency of protein interactions in chromatin, keeping in mind that many proteins further form protein-protein complexes increasing their interaction potential. The average individual domain length in chromatin proteins is around 65 amino acids (the median is 28 aa), however, this number is biased by the presence of many zinc-finger domains (around 22 aa in length). Subpanel 5 in **Figure 6E** gives a more balanced view for each SimChrom category. For the majority of protein groups the median domain length is around 100 amino acids (mean is 137, median is 134).

The birds-eye view of the most frequently occurring Pfam domains' in various functional SimChrom-SL categories is presented in **Figure 7**. The data is presented for domains that occur in at least five chromatin proteins and in at least 10% of proteins in a category (the threshold for data point depiction is 5%). The comprehensive interactive analysis figure with the ability to alter these thresholds and switch between SimChrom and SimChrom-SL classifications systems is available in **Interactive Figure 4** (https://simchrom.intbio.org/#domain_composition). In **Figure 7** the following categories and their respective domains can be grouped revealing their partially shared domain composition: 1) the categories containing transcription factors and their zinc finger, homeodomains and KRAB domains form the most frequently occurring entities, 2) some chromatin regulators, such as PTM writers, readers, erasers and chromatin remodelers together with their Chromo, Bromodomain, PHD.

3.4. Detailed analysis of the multivalent interactions in chromatin proteins

This section supplements and expands section **3.3.4. Multivalent interactions in chromatin protein** in the main text.

The presence of multiple domains (belonging to the same or different domain models) in chromatin proteins is a known feature contributing to their ability to engage in multivalent interactions (**Figure 8A**) [44]. Below we present the analysis of such domains engaged in multi-valent interactions (referred to as EMVI-domains hereafter) that are found in chromatin/epigenetics regulator proteins (see

Figure 3 for definition of this group). To limit our analysis to a manageable set of EMVI-domains, we selected those that were found in multiple copies or in combination with another Pfam domain in at least three chromatin regulator proteins (94 Pfam domains in total), and from those we selected 59 domains that we were able to manually classify based on the information currently available in the literature according to their functional binding modes. The following *functional groups* of domains were used: histone methylation/acetylation/phosphorylation, chromatin remodeling, histone binding, DNA binding, DNA methylation, protein dimerization/oligomerization, PPI, RNA binding. Histone post-translational modifications were further subdivided into readers, writers and erasers *functional subgroups* (see **Figure 8C**, **Interactive Figure 5 (**https://simchrom.intbio.org/#domain_co-occurrence**),** and **Supplementary Table ST16** for the list of domains and their detailed classification). We consider this subset of chromatin proteins' domains as representative to illustrate the concept of multivalency in chromatin regulators interactions, since the selected domains are extensively characterized and their functions are known. A comprehensive analysis would require characterization of all 409 Pfam domain models that are found in combination with other models or in multiple copies in at least one chromatin regulator protein. As a compromise our online database includes the analysis of 163 Pfam domain models that are present in at least two chromatin regulator proteins (see https://simchrom.intbio.org/#domain_co-occurrence).

First, we analyzed the co-occurrence of selected EMVI-domains in all chromatin proteins. There were in total 922 chromatin proteins (306 with the exclusion of transcription factors) that had more than one selected EMVI-domain (including multicopies). The conditional probability of finding a corresponding domain A in a chromatin protein given that another domain B is already present was estimated and is presented in **Figure 8C** (columns and rows correspond to domains A and B, respectively). The **Interactive figure 5** is available at https://simchrom.intbio.org/#domain_co-occurrence (also includes unclassified potential EMVI-domains found in at least two chromatin regulator proteins). The matrix in **Figure 8C** allows to trace the interplay between different domains employed in architectures of chromatin proteins. The largest groups of domains in **Figure 8C** are those involved in histone methylation and DNA binding, suggesting that these mechanisms are the most represented and employed in chromatin functioning regulation.

There were 49 cases where association between the presence of various domains in chromatin proteins was 100% (red squares in **Figure 8C**), among them for 18 cases (9 domain pairs) the association was reciprocal (i.e. $P(A|B) = P(B|A)$). In certain cases this exclusive association between domains may be traced to the fact that they form a larger structural complex with direct structural interactions between the domains as judged by the visual inspection of AlphaFold based predictions (MOZ_SAS and zf-MYST, ADD_DNMT3 and DNMT3_ADD_GATA1-like). In other cases the association is likely due to functional reasons, in our analysis in the majority of cases such domains

were confined to the histone methylation readers subgroup (KDM3B_Tudor and PWWP_KDM3B; C5HCH, NSD_PHD, PHD-1st_NSD, and PHDvar_NSD found in Histone-lysine N-methyltransferases).

Among the EMVI Pfam domains that co-occur with the most number of other different Pfam domains in chromatin regulator proteins is the histone methylation/acetylation domains: PHD domain (45 other domains), Bromodomain (38), SET (40) and PWWP (28) and chromatin remodeling Helicase_C (33) and SNF2-rel_dom (31).

The diagonal elements in **Figure 8C** show that certain domains tend to be present in multiple copies, particularly often, MBT, WD40 and zinc finger (zf-C2H2) domains. However, all these are special cases of short repeat domains, where multiple copies are needed to form one functional unit. Not so often, but in a considerable number of proteins PHD and Bromodomain may be found in multiple copies.

Another more general view of multivalent interactions in chromatin proteins may be obtained if we trace the relationships within or between different functional groups of domains. One can see that domains from the same functional group (*e.g.*, histone methylation) tend to co-occur (**Figure 8C**) in chromatin proteins and may also be present in multiple instances in proteins (**Supplementary Figure SF8_1B**). For example, there may be up to nine histone methylation associated domains in chromatin proteins (**Supplementary Figure SF8_1B**). If a chromatin protein has a domain involved in histone methylation (either writing, reading or erasing) there is an estimated 38% chance that there will be another different functional domain from this group of domains (**Supplementary Table ST17**, **Supplementary Figure SF8_1C**). For acetylation this estimated probability is 31%, for phosphorylation 18%. Note, that domain categorization is not trivial. For example, WD40 is a repeat that folds into a higher order structure (median number in chromatin proteins are four). They can recognize both unmodified histone tails and methylated regions [45] and its classification may affect this part of study.

The associations between the occurrence of domains from different functional groups can also be observed. This can be seen in **Figure 8C**, and the upset plot in **Figure 8D** (see also **Supplementary Figure SF8_2A**). One can see that domains involved in histone methylation (one of the most abundant group by the number of Pfam models and the number of chromatin proteins) may be in a considerable number of chromatin proteins combined with other EMVI-domains (particularly DNA binding domains and histone acetylation), association association with with histone binding domains, chromatin remodeling, DNA methylation, (di/oligo)-merization and RNA binding domains was also observed, see **Supplementary Figure SF8_2C**. The same can be said about domains involved in histone acetylation, although in a somewhat smaller number of cases, and with exclusion of their combination with

dimerization domains. Notably, domains involved in histone phosphorylation were not found in combination with domains from other functional groups in our analysis. This may reflect an evolutionary strategy whereby combinations of histone methylation and acetylation evolved to delicately regulate gene expression at the epigenetic level, while phosphorylation remained as a more general mechanism affecting a broad number proteins and pathways in the cell. DNA binding domains are found to be associated with methylation, acetylation, dimerization, and chromatin remodeling domains although the relative number of proteins harboring combinations of such domains is small compared to the number of protein (mainly transcription factor) that have DNA binding domains in their sequence. Domains associated with catalytic subunits chromatin remodeling complexes in certain cases are found together with histone acetylation, methylation or DNA binding domains. A more detailed analysis reveals that one "reader" domains for acetylation and methylation are found in these proteins.

For a more comprehensive view of multivalent interactions it is reasonable to (1) analyze not only pairwise co-occurence of different functional domains, but simultaneous co-occurrence of domains from several functional groups in one proteins, (2) extend the analysis to complexes of chromatin proteins. The results of such analysis are presented in **Figure 8D** (see **Methods Section 2.1.4** for our selection of 513 protein complexes where all proteins are chromatin proteins from Complex Portal). In our analysis at the level of individual proteins, proteins harbored domains only from up to three functional groups. Particularly, DNA binding domains may be combined with (di/oligo)merization, chromatin remodeling, histone methylation, methylation and acetylation or histone acetylation and chromatin remodeling domains. The formation of chromatin protein complexes considerably affects the available combinations of functional domains. Among the 513 analyzed protein complexes, 181 complexes contained domains EMVI-domains from the analyzed functional domain groups, 101 complexes harboured more than one domain, 80 complexes harbored domains from different functional groups. From these 80 complexes the majority were various chromatin remodeling complexes (53 complexes), others representatives included acetyltransferase complexes (13), deacetylase (4), and DNA-methylation (2).

One can see that the largest number of analyzed complexes (22) simultaneously contained domains from four functional groups (DNA binding, histone methylation, histone acetylation, and chromatin remodeling), in a select number of complexes domains belonging to up to six functional groups were observed (all the above mentioned together with domains involved in DNA methylation and histone binding). These were all complexes involved in chromatin remodeling. For example, 'MBD2 or MBD3/NuRD nucleosome remodeling and deacetylase complex'. Notably, in chromatin complexes histone acetylation domains are found more often than histone methylation domains (unlike in the case when individual proteins are analyzed), this might, however, be biased by the current list of

known chromatin complexes and their variability. Taken together chromatin protein complexes expand the multivalency of chromatin protein interactions and expand the functionality of the complexes.

As a final step we looked for chromatin proteins that had the most number of domain types (Pfam models) that belong to different functional groups/sub-groups and thus may engage in interactions of high multivalency (see **Supplementary Figure SF8_2A**). According to this analysis chromatin proteins could harbor in their domain composition domains from up to four functional groups/sub-groups. For example, histone-lysine N-methyltransferase ASH1L is involved in reading and writing of histone methylation, reading of histone acetylation and histone binding (**Figure 8E**). Proteins could harbor up to nine individual domains (where several domains belong to identical functional categories). One of such examples is Histone-lysine N-methyltransferase NSD2 (**Figure 8E**). This protein combines domains that likely engage in methylated histone binding (PWWP, PHD-1st_NSD, NDS_PHD, PHDvar_NSD, PHD, C5HCH), histone methylation (SET), histone binding (AWS) and may be DNA binding (HMG_box).

References

1. Hammond CM, Strømme CB, Huang H, Patel DJ, Groth A. Histone chaperone networks shaping chromatin function. *Nat Rev Mol Cell Biol*. 2017;18(3):141-158. doi:10.1038/nrm.2016.159

2. Cremer T, Cremer M, Hübner B, et al. The Interchromatin Compartment Participates in the Structural and Functional Organization of the Cell Nucleus. *BioEssays*. 2020;42(2):1900132. doi:10.1002/bies.201900132

3. Marakulina D, Vorontsov IE, Kulakovskiy IV, Lennartsson A, Drabløs F, Medvedeva YA. EpiFactors 2022: expansion and enhancement of a curated database of human epigenetic factors and complexes. *Nucleic Acids Res*. 2023;51(D1):D564-D570. doi:10.1093/nar/gkac989

4. Davidson IF, Bauer B, Goetz D, Tang W, Wutz G, Peters JM. DNA loop extrusion by human cohesin. *Science*. 2019;366(6471):1338-1345. doi:10.1126/science.aaz3418

5. Reeves R. High mobility group (HMG) proteins: Modulators of chromatin structure and DNA repair in mammalian cells. *DNA Repair*. 2015;36:122-136. doi:10.1016/j.dnarep.2015.09.015

6. Mayran A, Drouin J. Pioneer transcription factors shape the epigenetic landscape. *J Biol Chem*. 2018;293(36):13795-13804. doi:10.1074/jbc.R117.001232

7. Sun H, Fu B, Qian X, Xu P, Qin W. Nuclear and cytoplasmic specific RNA binding proteome enrichment and its changes upon ferroptosis induction. *Nat Commun*. 2024;15(1):852. doi:10.1038/s41467-024-44987-9

8. Van Nostrand EL, Freese P, Pratt GA, et al. A large-scale binding and functional map of human RNA-binding proteins. *Nature*. 2020;583(7818):711-719. doi:10.1038/s41586-020-2077-3

9. Azad GK, Swagatika S, Kumawat M, Kumawat R, Tomar RS. Modifying Chromatin by Histone Tail Clipping. *J Mol Biol*. 2018;430(18, Part B):3051-3067. doi:10.1016/j.jmb.2018.07.013

10. Gaudet P, Dessimoz C. Gene Ontology: Pitfalls, Biases, and Remedies. In: Dessimoz C, Škunca N, eds. *The Gene Ontology Handbook*. Methods in Molecular Biology. Springer; 2017:189-205. doi:10.1007/978-1-4939-3743-1_14

11. Razin SV, Iarovaia OV, Vassetzky YS. A requiem to the nuclear matrix: from a controversial concept to 3D organization of the nucleus. *Chromosoma*. 2014;123(3):217-224. doi:10.1007/s00412-014-0459-8

12. van Mierlo G, Vermeulen M. Chromatin Proteomics to Study Epigenetics — Challenges and Opportunities. *Mol Cell Proteomics*. 2021;20:100056. doi:10.1074/mcp.R120.002208

13. Sigismondo G, Papageorgiou DN, Krijgsveld J. Cracking chromatin with proteomics: From chromatome to histone modifications. *PROTEOMICS*. 2022;22(15-16):2100206. doi:10.1002/pmic.202100206

14. Torrente MP, Zee BM, Young NL, et al. Proteomic Interrogation of Human Chromatin. *PLOS ONE*. 2011;6(9):e24747. doi:10.1371/journal.pone.0024747

15. El Kennani S, Adrait A, Shaytan AK, et al. MS_HistoneDB, a manually curated resource for proteomic analysis of human and mouse histones. *Epigenetics Chromatin*. 2017;10(1). doi:10.1186/s13072-016-0109-x

16. Zhou W, Cao H, Yang X, et al. Characterization of Nuclear Localization Signal in the N Terminus of Integrin-linked Kinase-associated Phosphatase (ILKAP) and Its Essential Role in the Down-regulation of RSK2 Protein Signaling *. *J Biol Chem*. 2013;288(9):6259-6271. doi:10.1074/jbc.M112.432195

17. Pronier E, Cifani P, Merlinsky TR, et al. Targeting the CALR interactome in myeloproliferative neoplasms. *JCI Insight*. 3(22):e122703. doi:10.1172/jci.insight.122703

18. Kuo T, Hsu S, Huang S, et al. Pdia4 regulates β-cell pathogenesis in diabetes: molecular mechanism and targeted therapy. *EMBO Mol Med*. 2021;13(10):e11668. doi:10.15252/emmm.201911668

19. Nishimura S, Tsuda H, Ito K, et al. Differential expression of ABCF2 protein among different histologic types of epithelial ovarian cancer and in clear cell adenocarcinomas of different organs. *Hum Pathol*. 2007;38(1):134-139. doi:10.1016/j.humpath.2006.06.026

20. Yehia L, Liu D, Fu S, Iyer P, Eng C. Non-canonical role of wild-type SEC23B in the cellular stress response pathway. *Cell Death Dis*. 2021;12(4):1-12. doi:10.1038/s41419-021-03589-9

21. Gao Y, Teng J, Hong Y, et al. The oncogenic role of EIF3D is associated with increased cell cycle progression and motility in prostate cancer. *Med Oncol*. 2015;32(7):196. doi:10.1007/s12032-015-0518-x

22. Kustatscher G, Grabowski P, Rappsilber J. Multiclassifier combinatorial proteomics of organelle shadows at the example of mitochondria in chromatin data. *Proteomics*. 2016;16(3):393-401. doi:10.1002/pmic.201500267

23. Kachaev ZM, Ivashchenko SD, Kozlov EN, Lebedeva LA, Shidlovskii YV. Localization and Functional Roles of Components of the Translation Apparatus in the Eukaryotic Cell Nucleus. *Cells*. 2021;10(11):3239. doi:10.3390/cells10113239

24. Seal RL, Denny P, Bruford EA, et al. A standardized nomenclature for mammalian histone genes. *Epigenetics Chromatin*. 2022;15(1):34. doi:10.1186/s13072-022-00467-2

25. Draizen EJ, Shaytan AK, Mariño-Ramírez L, Talbert PB, Landsman D, Panchenko AR. HistoneDB 2.0: a histone database with variants—an integrated resource to explore histones and their variants. *Database*. 2016;2016:baw014. doi:10.1093/database/baw014

26. Gerstberger S, Hafner M, Tuschl T. A census of human RNA-binding proteins. *Nat Rev Genet*. 2014;15(12):829-845. doi:10.1038/nrg3813

27. Wang M, Herrmann CJ, Simonovic M, Szklarczyk D, von Mering C. Version 4.0 of PaxDb: Protein abundance data, integrated across model organisms, tissues, and cell-lines. *Proteomics*. 2015;15(18):3163-3168. doi:10.1002/pmic.201400441

28. Clapier CR, Iwasa J, Cairns BR, Peterson CL. Mechanisms of action and regulation of ATP-dependent chromatin-remodelling complexes. *Nat Rev Mol Cell Biol*. 2017;18(7):407-422. doi:10.1038/nrm.2017.26

29. Ayala R, Willhoft O, Aramayo RJ, et al. Structure and regulation of the human INO80-nucleosome complex. *Nature*. 2018;556(7701):391-395. doi:10.1038/s41586-018-0021-6

30. Dauden MI, López-Perrote A, Llorca O. RUVBL1–RUVBL2 AAA-ATPase: a versatile scaffold for multiple complexes and functions. *Curr Opin Struct Biol*. 2021;67:78-85. doi:10.1016/j.sbi.2020.08.010

31. Ulianov SV, Khrameeva EE, Gavrilov AA, et al. Active chromatin and transcription play a key role in chromosome partitioning into topologically associating domains. *Genome Res*. 2016;26(1):70-84. doi:10.1101/gr.196006.115

32. Iyer-Bierhoff A, Krogh N, Tessarz P, Ruppert T, Nielsen H, Grummt I. SIRT7-Dependent Deacetylation of Fibrillarin Controls Histone H2A Methylation and rRNA Synthesis during the Cell Cycle. *Cell Rep*. 2018;25(11):2946-2954.e5. doi:10.1016/j.celrep.2018.11.051

33. Vogel C, Bashton M, Kerrison ND, Chothia C, Teichmann SA. Structure, function and evolution of multidomain proteins. *Curr Opin Struct Biol*. 2004;14(2):208-216. doi:10.1016/j.sbi.2004.03.011

34. Paysan-Lafosse T, Andreeva A, Blum M, et al. The Pfam protein families database: embracing AI/ML. *Nucleic Acids Res*. 2025;53(D1):D523-D534. doi:10.1093/nar/gkae997

35. Wang J, Chitsaz F, Derbyshire MK, et al. The conserved domain database in 2023. *Nucleic Acids Res*. 2023;51(D1):D384-D388. doi:10.1093/nar/gkac1096

36. Waman VP, Bordin N, Alcraft R, et al. CATH 2024: CATH-AlphaFlow Doubles the Number of Structures in CATH and Reveals Nearly 200 New Folds. *J Mol Biol*. 2024;436(17):168551. doi:10.1016/j.jmb.2024.168551

37. Blum M, Andreeva A, Florentino LC, et al. InterPro: the protein sequence classification resource in 2025. *Nucleic Acids Res*. 2025;53(D1):D444-D456. doi:10.1093/nar/gkae1082

38. Fox NK, Brenner SE, Chandonia JM. SCOPe: Structural Classification of Proteins--extended, integrating SCOP and ASTRAL data and classification of new structures. *Nucleic Acids Res*. 2014;42(Database issue):D304-309. doi:10.1093/nar/gkt1240

39. Chandonia JM, Guan L, Lin S, Yu C, Fox NK, Brenner SE. SCOPe: improvements to the structural classification of proteins - extended database to facilitate variant interpretation and machine learning. *Nucleic Acids Res*. 2022;50(D1):D553-D559. doi:10.1093/nar/gkab1054

40. Lau AM, Bordin N, Kandathil SM, et al. Exploring structural diversity across the protein universe with The Encyclopedia of Domains. *Science*. Published online November 1, 2024. doi:10.1126/science.adq4946

41. van Kempen M, Kim SS, Tumescheit C, et al. Fast and accurate protein structure search with Foldseek. *Nat Biotechnol*. 2024;42(2):243-246. doi:10.1038/s41587-023-01773-0

42. Lyubitelev AV, Nikitin DV, Shaytan AK, Studitsky VM, Kirpichnikov MP. Structure and functions of linker histones. *Biochem Mosc*. 2016;81(3):213-223. doi:10.1134/S0006297916030032

43. Gligorijević V, Renfrew PD, Kosciolek T, et al. Structure-based protein function prediction using graph convolutional networks. *Nat Commun*. 2021;12(1):3168. doi:10.1038/s41467-021-23303-9

44. Ruthenburg AJ, Li H, Patel DJ, Allis CD. Multivalent engagement of chromatin modifications by linked binding modules. *Nat Rev Mol Cell Biol*. 2007;8(12):983-994. doi:10.1038/nrm2298

45. Couture JF, Collazo E, Trievel RC. Molecular recognition of histone H3 by the WD40 protein WDR5. *Nat Struct Mol Biol*. 2006;13(8):698-703. doi:10.1038/nsmb1116