Imperial College London
Department of Chemical Engineering

Mining of Next Generation Sequencing (NGS) and Affinity-Purification Mass Spectrometry (AP-MS) data to identify chromatin targets of epigenetic regulators

Saulius Lukauskas

Submitted in part fulfilment of the requirements for the degree of
Doctor of Philosophy in the Imperial College London, October 2018

Declaration of originality

Even though pronoun *we* is used throughout this thesis, the work presented in this thesis is my own, unless explicitly stated otherwise.

Copyright declaration

The copyright of this thesis rests with the author. Unless otherwise indicated, its contents are licensed under a Creative Commons Attribution-Non Commercial 4.0 International Licence (CC BY-NC).

Under this licence, you may copy and redistribute the material in any medium or format. You may also create and distribute modified versions of the work. This is on the condition that: you credit the author and do not use it, or any derivative works, for a commercial purpose.

When reusing or sharing this work, ensure you make the licence terms clear to others by naming the licence and linking to the licence text. Where a work has been adapted, you should indicate that the work has been changed and describe those changes.

Please seek permission from the copyright holder for uses of this work that are not included in this licence or permitted under UK Copyright Law.

Abstract

In eukaryotes, modifications of DNA molecules and core histone proteins regulate core nuclear functions such as DNA transcription, DNA replication, and DNA repair. Individual modifications are associated with regions of active and repressed transcription, and their readers, writers, and erasers are known. In combination, these modifications are assumed to form a 'code' describing the chromatin regulation. The specifics of this combinatorial 'code', however, are still unknown. Recent advances in technology have allowed this regulation by nucleosomal modifications to be studied experimentally in high throughput. Such studies, however, face an information overload as more data is being produced than can be analysed. In this thesis, we attempt to provide methods to study this large-scale data globally. To do this, we describe a number of data mining approaches to extract, summarise and link the data in order to elucidate the language of nucleosomal modifications.

We base our analyses on a novel SILAC nucleosome affinity purification (SNAP) dataset produced in our laboratory, as well as data from Next-Generation Sequencing (NGS) studies. We apply a modified CLR network inference algorithm to predict novel interactions between epigenetic readers. We apply Factor Analysis (FA) to describe the key constituents of the code that drive the epigenetic reader binding. We further show that linear combination of individual responses to modifications can describe a significant proportion of the reader behaviour *in vitro*, and use NGS datasets to verify them *in vivo*. We find discrepancies between the *in vitro* and *in vivo* results and use Bernoulli Mixture Models (BMM) to analyse them. This provides us with a direction to improve the analysis and experimental methods.

Together, the results presented in this thesis provide a global overview of major aspects of the regulation by nucleosomal modifications, which brings us a step forward towards *cracking* the histone code.

Contents

Declaration of originality **3**

Copyright declaration **3**

Abstract **4**

List of Figures **9**

List of Tables **12**

Acknowledgements **14**

1 Background & introduction **16**

1.1 Epigenetics 16

1.1.1 Histones, nucleosomes and chromatin 17

1.1.2 Role of the chromatin in transcription 18

1.1.3 The language of covalent histone modifications 20

1.2 Studying the language of the nucleosome 31

1.2.1 Experimental approaches 31

1.2.2 List of notable studies 41

1.3 Aims of this thesis 43

2 Study of the SILAC nucleosome affinity purification dataset 47

2.1 Introduction 47

2.1.1 The dataset 48

2.2 Methods 51

2.2.1 Data visualisation 51

2.2.2 Data pre-processing 58

2.2.3 Network inference 69

2.2.4 Factor analysis 83

2.2.5 Identification of DNA-binding transcription factor binding sites 94

2.2.6 Linear modelling of epigenetic reader binding preferences 95

2.3 Results 98

2.3.1 Modified nucleosomes profile major epigenetic processes 98

2.3.2 Network interaction scores follow the expected distribution 100

2.3.3 Estimated networks recover major regulatory blocks of chromatin 102

2.3.4 High-confidence network predictions identify novel subunits of protein complexes 104

2.3.5 Eleven regulatory dimensions, two complementary viewpoints 114

2.3.6 Factors with complex loadings regulate DNA-binding transcription factor response 122

2.3.7 Linear models recover epigenetic reader binding profiles 128

2.4 Discussion 135

3 Next generation sequencing datasets **140**

3.1 Introduction . . . 140

3.2 Methods . . . 141

3.2.1 Pipeline for high-throughput ChIP-seq data processing . . . 141

3.2.2 Learning to rank the epigenetic reader binding preferences . . . 144

3.2.3 Clustering chromatin states using mixtures of independent Bernoulli variables . . . 148

3.3 Results . . . 161

3.3.1 ChIP-seq based models provide a complementary view of epigenetic reader binding profiles . . . 161

3.3.2 Bernoulli mixture models highlight the importance of chromatin-reader associated states . . . 169

3.4 Discussion . . . 176

4 Conclusion **185**

4.1 Summary of Thesis Achievements . . . 186

4.2 Applications . . . 189

4.3 Future work . . . 190

Bibliography **193**

Appendix A Reference **215**

List of acronyms . . . 215

List of protein complexes . . . 217

List of proteins . . . 221

List of software & databases 231

Appendix B Supplementary information **235**

B.1 Supplementary figures and tables 235

B.2 Large-scale mononucleosome purification from HeLa cells 250

B.2.1 Reagents & buffers 250

B.2.2 Protocol 251

Appendix C Accession lists **255**

Appendix D Permissions to reproduce figures **266**

List of Figures

1.1 Summary of chromatin packaging and regulation 17

1.2 Possible modes of combinatorial interaction between nucleosomal PTMs 24

1.3 Cross-talk and combinatorial read out of histone PTMs near the *FOSL1* enhancer 25

1.4 Summary of key lysine methylation sites on histones H3 and H4 27

1.5 The main mechanisms of ATP-dependent chromatin remodelling 31

1.6 Workflow of a ChIP-seq experiment 32

1.7 Visualisation of the effect of propionic anhydride treatment on histone tails . . . 35

1.8 Description of SNAP workflow and sample results 40

2.1 Design of the novel SNAP dataset analysed in this thesis 50

2.2 Scatterplot visualisation of a SNAP experiment 55

2.3 Comparative heatmap visualisation of SNAP experiments 56

2.4 Architecture of the interactive SNAP visualisation 57

2.5 A summary for the preprocessing workflow for the large scale SNAP dataset . . 59

2.6 Identifying the major axes of the data variance within a SNAP experiment . . . 61

2.7 Illustration of the residual correction procedure for SNAP data 65

2.8 Ratio projection imputation of missing data in SNAP datasets 68

2.9 Performance statistics of network inference algorithms 74

2.10 Derivation of the p-value statistic for the *pseudo z-score* statistic for CLR 76

2.11 Determination of *high-confidence* threshold for detection of novel protein interactions 82

2.12 Estimation of the number of factors for Factor Analysis 87

2.13 Choosing the best rotation for the hidden axes 91

2.14 Factor score estimation 92

2.15 Summary statistics for the training of linear model predictions 97

2.16 Histogram of the distribution of the scores from linear model analysis 98

2.17 GO enrichment analysis on the protein sets significantly regulated by the modified nucleosomes 99

2.18 Distribution of scores obtained from the CLR algorithm 101

2.19 Chromatin regulatory network identified from the network analysis 103

2.20 High-confidence predictions from the network analysis 105

2.21 Heatmap representation of model's predictions of substantial interest 107

2.22 Heatmaps for B1 and B2 high-confidence prediction clusters in Figure 2.20 . . . 113

2.23 Comparison between *bentlerT* and *bifactor* rotations 115

2.24 Protein domains, complexes or transcription factor families which have greater mean factor score than the rest 118

2.25 Summary figure for the results of Factor Analysis (FA) 119

2.26 TF binding clusters from SNAP data 124

2.27 TF binding clusters relationships with regulatory factors and CHD proteins . . . 126

2.28 Heatmap of members of the TF cluster 6 along of with their high confidence interactors 127

2.29 Visualisation of the TF hotspots on the nucleosomal core particle 128

2.30 Epigenetic reader motifs estimated from the SNAP data 130

2.31 Association between predicted epigenetic reader motifs and InterPro domains . . 134

3.1 Schematic for the ChIP-seq pipeline 143

3.2 Illustration of the LETOR model for chromatin reader binding preferences from ChIP-seq data 146

3.3 Summary of the dataset of ChIP-seq based predictions for protein interactions with histone PTMs 148

3.4 BMM model for establishing the interaction between chromatin states and epigenetic readers 151

3.5 Plot of the coefficients estimated from ChIP-seq and SNAP where the models agree 163

3.6 Relationship between performance statistics and agreement of the models 164

3.7 Profiles of chromatin interaction coefficients in a selection of cases where predictions from the ChIP-seq model do not agree with SNAP 168

3.8 Visualisation of SNAP data for the proteins shown in Figure 3.7 169

3.9 A summary of the inferred BMM model 171

3.10 A summary of the BMM model results for the BRD4 protein 175

3.11 Quantification of mononucleosome yield from incremental digestions using MARCC protocol 183

B.1 Complete set of complexes regulated by the hidden factors 236

B.2 Complete set of protein domains regulated by the hidden factors 237

B.3 Complete set of transcription factors regulated by the hidden factors 237

B.4 *bifactor* rotation equivalent to the factor maps of *bentlerT* rotation shown in Figure 2.25 238

B.5 Factor clusters for the top-left panels of Figure 2.25 and Figure B.4 239

B.6 DNA sequence used in the production of recombinant di-nucleosomes for SNAP pull-downs 240

B.7 Distribution of TF binding sites on the SNAP di-nucleosomes and TF clusters augmented with CHD proteins 241

B.8 Complete set of histone-association coefficients estimated from the ChIP-seq data 243

B.9 Heatmap visualisation of the strongest targets predicted for each of the proteins analysed using ChIP-seq prediction model 244

B.10 Main performance indicators for the BMM model training procedure 245

B.11 A summary of the inferred BMM model when trained using the data from H1-hESC cell-line only 246

B.12 A summary of the inferred BMM model when trained using the data from H9 cell-line only 247

B.13 A summary of the inferred BMM model when trained using the data from IMR-90 cell-line only 248

B.14 Profiles of the signal for a selection of ChIP-seq tracks around the top 10,000 highest BRD4 peaks in IMR-90 cell line 249

List of Tables

1.1 Summary of protein domains that associate with acetylated and methylated lysines 23

2.1 Conversion table between the estimated CLR scores and their adjusted p-values for our data 82

2.2 Readouts of loaded regulators for the 11 regulatory axes predicted by FA 116

B.1 Performance statistics for the network inference algorithms tested in section 2.2.3 235

B.2 Correlation between the ChIP-seq predictions of histone association coefficients across different cell-types and/or SNAP predictions 242

C.1 Accessions of datasets used in ChIP-seq experiments 255

C.2 List of protein complexes and their members, as well as the sources for annotation 258

Acknowledgements

First and foremost I would like to thank my supervisors Dr. Till Bartke and Dr. Peter DiMaggio for the opportunity to work in their labs. I am really grateful for all the scientific discussions that we have had, their passion for science that has radiated onto me, as well as measured criticisms all of which have helped to get the project to where it is now.

I would also like to thank my fellow PhD students (most of which are doctors by now) for providing both scientific and moral support throughout the PhD. Out of them I would like to explicitly thank Aaron, Anne, Ben, Diego, Kate and Thomas who have patiently answered every single question that I have had, no matter how stupid it was. Together you gave me both the confidence and the experience necessary to work in the "wet" lab, of which I had none to begin with, and therefore have opened me many more doors into biology than I can imagine.

I also want to thank Dr. Nhuong Nguyen for his work on developing and establishing the new SNAP pull-down dataset which now forms the basis of the chapter 2 of this thesis.

Finally, as much as it sounds like a cliché, I would like to thank all of the friends that I have made during this PhD, both in the academic and non-academic circles, as well as the members of my family. On the worst days you were the ones who helped me to get through the day, on the best days you were the first ones I'd share the joy with.

Dedication

To my family – for their overwhelming support throughout this journey.

Chapter 1

Background & introduction

1.1 Epigenetics

Traditionally, the term *epigenetics* has been associated with the heritable information that is not passed through the DNA sequence itself, drawing a parallel to the field of *genetics* which has flourished after the description of the structure of the DNA molecule by Watson & Crick.

Increased understanding of the molecular mechanisms associated with DNA-based gene regulation has stretched the definition of this term to its literal back-translation "above (*epi-*) genetics", which has encompassed the whole regulatory system beyond the genomic information. As the number of publications adopting this new meaning grew, the original definition has been somewhat forgotten, which has prompted some scholars to advocate for its return[1]. These scholars suggest that the term *epigenetics* which is used in the latter case should be replaced and the term *transcription regulation* should be used instead.

While in this thesis we do not focus on the heritability of epigenetic information, it seems inappropriate to call this chapter *transcription regulation*, as the epigenetic mechanisms described here regulate not only the DNA transcription, but also other processes such as DNA replication or DNA repair.

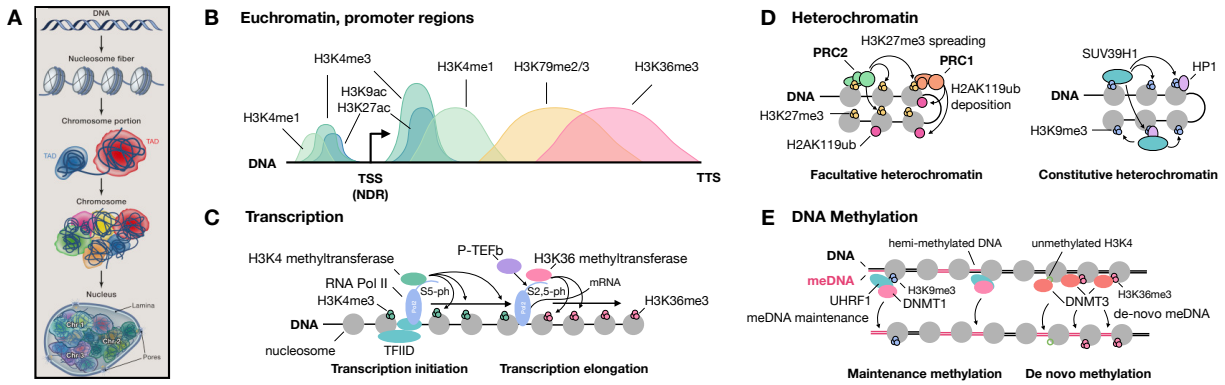

Figure 1.1: (A) DNA packaging inside eukaryotic nuclei. Reprinted from Cell, 160 (2015), Tom Sexton and Giacomo Cavalli, The Role of Chromosome Domains in Shaping the Functional Genome, Cell, p. 1050, Copyright 2015, with permission from Elsevier. [2] (B) Histone Post-Translational Modifications (PTMs) are landmarks of chromatin function. Based on a similar figure in [3]. (C) Model of DNA transcription and its relationship to activated histone marks. Based on a similar figure in [4]. (D) Canonical model of chromatin landmark establishment in Polycomb-dependent facultative heterochromatin[4, 5] as well as constitutive heterochromatin[4] (E) Canonical model of establishment of DNA methylation[6].

1.1.1 Histones, nucleosomes and chromatin

Eukaryotic DNA is tightly packed inside the nucleus via the help of histone proteins. 147 base pairs (bp) of DNA wrap around the octamer of four core histone proteins: Histones H2A, H2B, H3 and H4 into a structural unit called the nucleosome. These nucleosomes further arrange themselves into higher-order structures within the three-dimensional space of the nucleus. Spatially close regions of the DNA form Topologically Associated Domains (TADs) (see Figure 1.1 A) which represent distinct functional subunits of the chromatin [7].

The formation of TADs is mediated by the Cohesin complex which plays a part in forming and maintaining the DNA-DNA interactions of the TADs[7, 8]. The borders of TADs are enriched in CTCF binding sites. These borders act as a barrier between nearby (DNA sequence-wise) domains and prevent the spread of transcriptional activity in between the TADs[7]. At the higher level, the TADs are arranged into chromosomes, DNA-molecules of a size ranging from tens to hundreds of megabases, which occupy distinct chromosomal territories in the nucleus (Figure 1.1 A). This model has been first proposed a century ago, then abandoned, and then re-established again recently[7], resulting in an unexpected infancy of the studies of the dynamics of such chromosome domains.

Some recent evidence show that domains interacting with the nuclear lamina are heterochromatic (i.e. domains where transcription is inactive, repressive domains)[9], however it is not clear whether all heterochromatic TADs are associated with the nuclear lamina. Recently, it has been shown that heterochromatin domains are created via phase separation-like behaviour with the help of HP1a (CBX5) which interacts with H3K9me3 - a mark associated with constitutive heterochromatin[10]. Euchromatic domains that are associated with active transcription have also been shown to rely on phase separation[11, 12], entangling the study of biology with polymer physics.

At the nucleosomal level, histone proteins are more than just purely structural components. These proteins, while globular, have an unstructured N-terminal *tail* which is frequently modified post-translation, by covalent addition of methyl, acetyl, phospho or other chemical groups. The number of such known Post-Translational Modifications (PTMs) is increasing every year, quickly making the comprehensive review papers (e.g. [13]) obsolete. For instance, lysine benzoylation has been added to the list of known histone PTMs recently[14].

1.1.2 Role of the chromatin in transcription

While little is known about such novel histone PTMs, the role of other histone PTMs in various cellular processes has already been established and they are generally thought to act as the landmarks of transcriptional activity[3]. Transcription Start Sites (TSSs) are known to contain Nucleosome Depleted Region (NDR), which exposes the free DNA so that Transcription Factors (TFs) can associate. Around these sites devoid of nucleosomes, especially towards the direction of transcription, lysine four of histone H3 is trimethylated (H3K4me3). The bodies of active genes are then marked with H3K79me3 and H3K36me3 with the distribution similar to the one illustrated in Figure 1.1 B[3]. Enhancer regions (i.e. the regions that stimulate transcription, but do not produce coding transcripts themselves) are similarly marked with H3K4me1 and the acetylation of lysine 27 of histone H3 (H3K27ac)[3]. Additionally, depending on the context, the core histone proteins can be exchanged with their non-canonical variants. For instance, replacement of the canonical H2A histone with its variant H2A.Z at the nucleosome next to

the TSS eases the path of RNA Polymerase II[15, 16].

The current evidence suggests that at least some of these landmarks are established as a consequence of the transcription itself (Figure 1.1 C). In eukaryotes establishment of transcription is a multi-step process. First, the initiation factor TFIID assembles at the gene promoter with the help of TFs. It then phosphorylates the 5th serine in the unstructured repeats on the C-terminal domain of RNA Polymerase II initiating transcription. The phosphorylated polymerase travels a short distance along the DNA, which is tracked by methyltransferases. These methyltransferases are homologous to the COMPASS methyltransferase in yeast (e.g. COMPASS, COMPASS-like MLL1,2, COMPASS-like MLL3,4 in humans, see text below). This methyltransferase deposits the H3K4 methylation in the neighbouring nucleosomes in an incremental fashion that is indicative of the time spent at the locus[17] (Figure 1.1 C). The recruitment of the P-TEFb complex, often by the protein BRD4, leads to phosphorylation of the second serine on the C-terminal domain repeats of RNA Polymerase II, which allows the RNA polymerase to proceed into the elongation stage[18]. The progress of this doubly-phosphorylated polymerase is then tracked by another methyltransferase which methylates H3K36[4], resulting in a general modification pattern displayed in Figure 1.1 B.

Similarly, the transcriptionally repressive facultative heterochromatin is often marked with H3K27me3, deposition of which depends on the enzymes belonging to the Polycomb group of proteins (PRC1 and PRC2 complexes, see below), while constitutive heterochromatin is enriched in H3K9me3 (Figure 1.1 D). Interestingly, these are the only two histone PTMs whose mechanism of inheritance across cell division is known[4]. Both of these inheritance mechanisms rely on chromatin factors that can both recognise and write the chromatin modification. Facultative heterochromatin is believed to spread via PRC2's ability to recognise and deposit H3K27me3, which is also recognised by PRC1 complex, leading to ubiquitinylation of histone H2A (Figure 1.1 D, left)[5]. Similarly, constitutive heterochromatin spreads via SUV39H1 methyltransferase, which can both recognise and deposit H3K9me3 and the subsequent binding of HP1 proteins, which can then facilitate further recruitment of SUV39H1[4, 19, 20] (Figure 1.1 D, right).

Another heritable modification that is generally repressive (with some exceptions[21]) is the methylation of cytosines on the DNA. In eukaryotes, majority of the DNA methylation occurs at 5-methylcytosine (5-mC) sites at CpG dinucleotides[21]. The canonical model (which is a significant oversimplification[6, 21]) splits the human DNA methyltransferases into two groups: maintenance (DNMT1) and *de novo* (DNMT3A, DNMT3B) methyltransferases.

The maintenance methyltransferase DNMT1 preferentially methylates CpG sites that are already hemi-methylated, thus playing a role in re-establishment of methylated chromatin post DNA replication[21]. This interaction is dependent on UHRF1, which is recruited to such hemi-methylated sites as well as H3K9me3[21, 22] (Figure 1.1 E).

The *de novo* methyltransferases DNMT3A and DNMT3B instead deposit methylation at regions where it has not previously been established at. These methyltransferases are preferentially targeted to regions that are not marked by H3K4 methylation[21] (Figure 1.1 E). Evidence shows that DNMT3B is also targeted to bodies of active genes via PWWP-domain mediated interaction with H3K36me3, to prevent the transcription from cryptic promoters[23]. Interestingly, this DNA methyltransferase can therefore be linked to active transcription of genes while also having a generally repressive function.

1.1.3 The language of covalent histone modifications[1]

The observations of the apparent association between histone PTMs and DNA methylation with transcription have prompted the idea that these modifications exhibit a *code* that, once understood, could describe the chromatin behaviour completely. This idea has been proposed for the first time in the seminal paper by Strahl and Allis[24], where the authors suggested that "distinct histone modifications, on one or more tails, act sequentially or in combination to form a 'histone code' that is, read by other proteins to bring about distinct downstream events".

While certain modifications such as H4K16ac have been shown to directly impact the chromatin compaction[25], this hypothesis has directed the field to search for such *reader* proteins

[1] Intentional reference to Strahl and Allis: "The language of covalent histone modifications"[24]

that could recognise (i.e. *read*) post-translationally modified amino acids on the histones and modified nucleotides in the DNA sequence. Numerous proteins that catalyse the deposition (i.e. *writing*) or removal (i.e. *erasing*) of such modifications have also been identified and linked to biological functions, providing further basis to this theory.

At the same time, the field has taken the word *code* perhaps too literally which has lead to a debate of what constitutes the code and what does not. For instance, Oliver J. Rando, in his 2012 review[26], highlights that histone PTMs have a very low complexity *in vivo*, suggesting that the code is not utilised. While in theory there are more possible combinations of PTMs than nucleosomes in the nucleus, the majority of the variance can be explained by blocks of modifications that vary together[26, 27]. He raises a valid concern that *in vitro* studies examining possible combinations of modifications may be irrelevant, if these combinations do not exist *in vivo* and therefore should not be considered as evidence for the histone code.

The low complexity of histone modifications serves to show that the histone code is built upon a set of constituents, principal components, or factors, which is more indicative of how *languages* work. If the histone *code* was compared to a *language* instead, the chromatin modifications would serve as words, their combinations would form sentences. The chromatin context, such as gene coding regions or TADs therefore would reflect paragraphs and chapters. Now, the low complexity is fairly natural: certain key PTMs (keywords) are used more than the others. Different chromatin contexts may also have a different set of PTMs just how different texts use slightly different vocabulary. In English, it is also common for words to occur with others more frequently than not (such constructs are called collocations, e.g. "[to] shrug [one's] shoulders"), similarly one PTM may occur with another PTM more often than not. Breaking down a collocation will make it lose the meaning, but linguists are still interested in why this happens, so should we be.

Given the title of the original publication by Strahl and Allis, *The language of covalent histone modifications*[24], it is likely that this view is what the authors intended all along. In this framework the concern raised by Rando is still valid. It is just not a concern about what constitutes a histone code, and what does not. It is clear that both *in vitro* and *in vivo* results

highlight different aspects of the nucleosomal language, just like the study of grammar and the study of discourse help shape our understanding of how the English language works.

Reading histone modifications

The first reader domain for the Post-Translational Modification (PTM) on the histone proteins has been identified as early as in 1999[28]. This reader domain, a bromodomain, has been shown to bind acetylated lysine residues. Since then, the binding preferences of the human bromodomain family have been exhaustively characterised[29]. All bromodomains bind acetylated lysine residues with varying specificity towards the neighbouring residues. For instance, the transcription elongation factor BRD4 (which we have mentioned previously) contains two bromodomains. The first bromodomain specifically binds acetylated lysines on the histone H4, while the second bromodomain is more general[29]. It is common to observe the binding of bromodomain-containing proteins to be enhanced by presence of multiple acetylations on nearby lysines. This is likely due to the ability of a single bromodomain to fit two acetylated residues into the binding pocket[30]. Besides bromodomains, double PHD finger and double pleckstrin homology domains have also been shown to target acetylated lysines (reviewed in [28]).

Other nucleosome modifications have been shown to be targeted by other protein domains (see Table 1.1, reviewed in [28]): chromodomain-containing proteins have been shown to associate with H3K9me3 or H3K27me3. PWWP domains weakly associate with H3K36me3 or H4K20me1. Some PHD (plant homeodomain) finger containing proteins associate strongly with H3K4me2/3, though this domain is more versatile. Many proteins (and protein complexes) harbour multiple such chromatin interacting domains, which provides direct mechanisms of how the histone modification language can be read in combinatorial fashion.

Figure 1.2 illustrates four scenarios which describe how such combinatorial read-out could be achieved. Multiple chromatin binding domains within the same protein may each associate with different PTMs on the nucleosomes, and activation of each may be required for successful recruitment of the protein (panel A). Chromatin complexes, where binding domains are

	me1	me2	me3	ac
H3K4	Chromo-barrel, DCD, MBT	DCD, MBT, PHD	DCD, PHD, TTD, zf-CW	Bromodomain
H3K9	Ankyrin, MBT	Ankyrin, Chromodomain, MBT	ADD, Chromodomain, PHD, TTD, WD40	
H3K14				Bromodomain
H3K18				Bromodomain
H3K23				Bromodomain
H3K27		Chromodomain	Chromodomain, WD40	Bromodomain
H3K36		Chromo-barrel	Chromo-barrel, PWWP, Tudor	Bromodomain
H3K79			PWWP	
H4K5				Bromodomain
H4K8				Bromodomain
H4K12				Bromodomain
H4K16				Bromodomain
H4K20	Chromo-barrel, PWWP	BAH, TTD	PWWP	Bromodomain

Table 1.1: Protein domains that associate with methylated and acetylated lysines on histones H3 and H4. Abbreviations: ADD (ATRX-DNMT3-DNMT3L domain), BAH (Bromo-adjacent homology domain), DCD (double chromodomain), MBT (malignant brain tumour domain), PHD (plant homeodomain), TTD (tandem tudor domain), zf-CW (zinc finger, CW type). Data from [28]

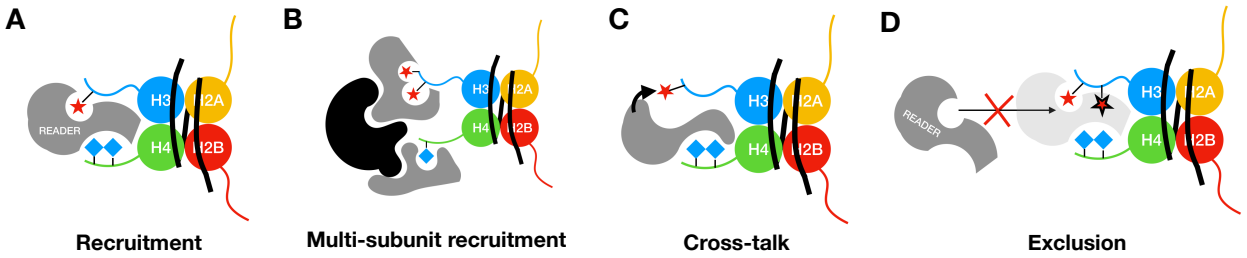

Figure 1.2: Possible modes of combinatorial interaction between nucleosomal PTMs: (A) Co-recruitment of epigenetic readers, either in *cis* or *trans*, where multiple modifications are necessary to ensure stable interaction; (B) Co-recruitment of a protein complex by activating individual reader domains in different subunits of the complex; (C) Cross-talk between two modifications, where successful recruitment of an epigenetic reader to one modification, either deposits or removes another one; (D) Exclusion of protein binding, where a presence of a modification may prevent a protein from binding to a set of PTMs it would normally interact with.

harboured in different subunits, can be recruited via a similar mechanism (B), however here the PTMs may direct the exact subunit composition of the complex. If the recruited protein contains enzymatic activity, cross-talk between histone PTMs may be established (C). Finally, PTMs can also prevent the binding of epigenetic readers (Figure 1.2 D).

The first three processes (Figure 1.2 A-C) are nicely illustrated by the mechanism of P-TEFb recruitment to the enhancer of *FOSL1* gene (Figure 1.3). Briefly, transcription at the *FOSL1* locus is primed by pausing RNA Polymerase II at an initiated, but not yet elongating state. At this state, the enhancer of this gene is already marked with H4K5ac, H4K8ac, H4K12ac and H3K9ac. Upon a stimulus, MYC recruits PIM1 kinase which deposits H3S10ph[31] (Figure 1.3 A). The deposition of H3S10ph leads to the recruitment of KAT8 (MOF) which in turn deposits H4K16ac creating a cross-talk between the two PTMs (Figure 1.3 B). Finally, the double-bromodomain protein BRD4 binds combinatorialy to the acetylated histone H4 and H3K9ac, which in turn recruits P-TEFb that phosphorylates serine 2 on RNA Polymerase II C-terminal repeats (Figure 1.3 C), initiating the transcriptional elongation.

The most frequently cited example for the exclusion by chromatin modifications (Figure 1.2 D) is is the apparent inability for the HP1 proteins (e.g. CBX5) to bind their preferred substrate H3K9me3 when the neighbouring serine 10 on histone H3 is phosphorylated (H3S10ph)[26, 28]. Similarly, ATRX binding to H3K9me3 is inhibited by H3K4me3[26, 33]. More recently, it has been shown that TONSL is excluded from nucleosomes containing H4K20me via its ankyrin-

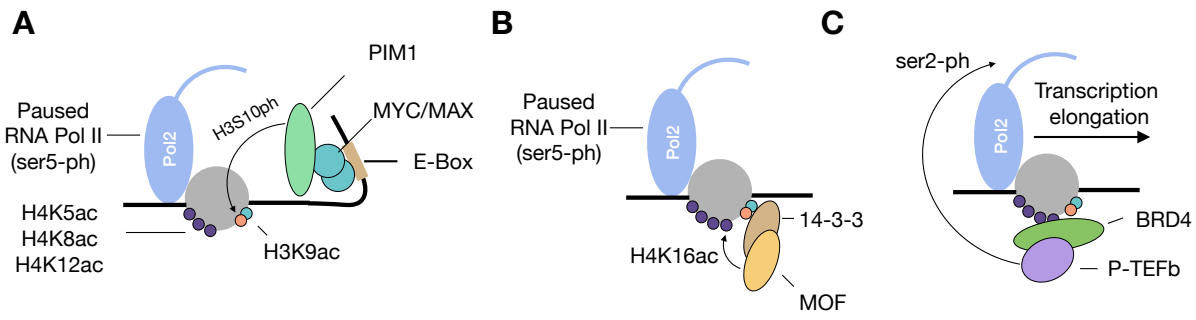

Figure 1.3: Cross-talk and combinatorial read out of histone PTMs near the *FOSL1* enhancer leading to transcription elongation, model based on [31, 32]. *FOSL1* gene transcription is mediated by the recruitment of the PIM1 kinase by the Transcription Factor MYC. Recruitment of PIM1 deposits H3S10ph on histone H3 already acetylated at lysine 9 (H3K9ac) (A). This in turn recruits KAT8 (MOF) acetyltransferase which deposits H4K16ac on the neighbouring histone H4 (B). The combinatorial binding to H4K16ac and H3K9ac recruits the double bromodomain-containing protein BRD4, which in turn recruits P-TEFb. Successful recruitment of P-TEFb phosphorylates serine 2 on RNA Polymerase II, which releases the polymerase from a paused state allowing for successful transcription. Note that the DNA transcription does not occur at this enhancer locus, but the proximally close promoter (due to chromatin looping) is transcribed instead, the promoter was omitted from this figure in order to simplify it.

repeat domain[34], though in such cases it may be more accurate to say that TONSL is a reader for unmodified lysine K20 instead.

Writing and erasing histone modifications

The histone modifications described previously are deposited by histone modifying enzymes ("epigenetic writers") and removed by de-modification enzymes ("epigenetic erasers"). We briefly review the key enzymes responsible for this regulation in this section.

Lysine acetylation Lysine acetyltransferases (i.e. the enzymes that catalyse the addition of acetyl residues on histone tails) can be divided into three families: GNAT, MYST and p300/CBP family[35]. These enzymes generally do not function as individual proteins, but instead act as catalytic subunits of protein complexes. The GNAT family of histone acetyltransferases reside in ATAC, NuA4, PCAF, SAGA and TFTC-HAT protein complexes[36]. MYST family histone acetyltransferases (HATs) locate in HBO1, MOZ/MORF, MSL, NuA4

and NSL complexes[36]. The members of third family of acetyltransferases – p300/CBP – consist of EP300, CREBBP which act together, and TAF1 which is part of the TFIID complex.

These large protein complexes both guide the histone acetyltransferases to the appropriate regions of the genome, and change their substrate specificity. For instance, the KAT8 (MOF) acetyltransferase acetylates H4K16ac as part of the MSL complex, but can also acetylate H4K5ac or H4K8ac when assembled into the NSL complex instead[37]. Other acetyltransferases are regulated in a similar way. This extra layer of regulation allows for extra redundancy among lysine acetyltransferases, making their substrate specificity harder to define[36]. Nevertheless, KAT2A (GCN5) acetyltransferase is commonly associated with H3K9, H3K14 and H3K36 acetylation[38]. KAT2B (PCAF) and KAT6A (MOZ) are typically associated with H3K14ac[38].

Acetylation can be removed via histone deacetylase (HDAC) enzymes. These enzymes are classified into four different classes: I, II, III and IV. Class III histone deacetylases (HDACs) (i.e. sirtuins) are different from the rest as members of this class catalyse the reaction using NAD+ dependent chemistry[35]. Class IV HDACs contain only one HDAC protein: HDAC11. Like histone acetyltransferases (HATs), HDACs are very versatile and typically associate with a large number of protein complexes, which shape their function. For instance, HDAC1 and HDAC2, occur in most of the nuclear repressive complexes such as the BHC, LSD-CoREST, MeCP1, N-CoR2, NuRD and mSin3A[39]. This versatility necessarily redefines their function based on the associated context. Class III deacetylases, on the other hand, are thought to be more specific with SIRT6 associated with deacetylation activity on H3K9 and H3K56, and SIRT7 catalysing removal of the acetyl group from H3K18[40].

Lysine methylation Unlike acetylation, lysine methylation is considered to be more residue-specific. H3K4, H3K36 and H3K79 methylation is generally associated with active transcription, H3K9, H3K27 and H4K20 methylation is considered repressive[42] (see also the beginning of this chapter, section 1.1.1).

Histone H3 lysine 4 methylation (H3K4me) is primarily deposited by *SET1* methyltransferase

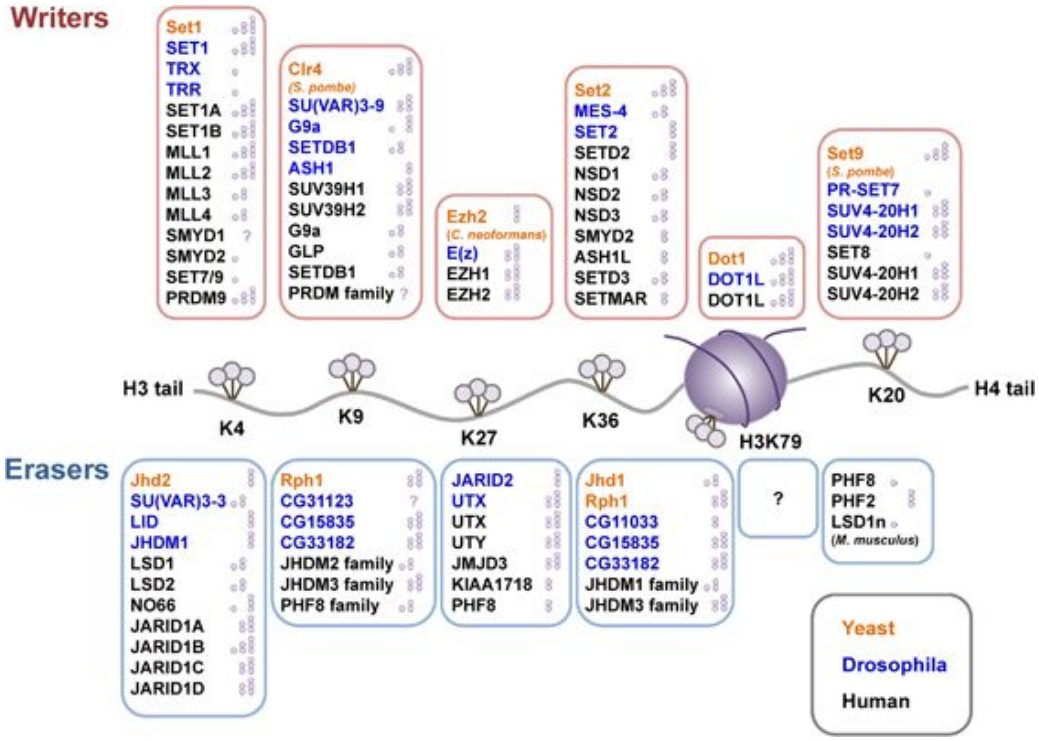

Figure 1.4: Summary of key lysine methylation sites on histones H3 and H4 along with the associated methyltransferases. Depicted next to the names are the mono, di- and tri- methylation states the methyltransferases can catalyse. H3K79 demethylase has recently been discovered to be KDM2B[41]. Reprinted by permission from Nature publishing Group: Nature Experimental & Molecular Medicine, Writing, erasing and reading histone lysine methylations, Kwangbeom Hyun, Jongcheol Jeon, Kihyun Park & Jaehoon Kim, 2017[42]

in yeast, but in human cells it is catalysed by 10 enzymes[42] (see Figure 1.4). The six human homologs of the yeast *SET1*: SETD1A, SETD1B, KMT2A (MLL1), KMT2B (MLL2), KMT2C (MLL3) and KMT2D (MLL4) are perhaps the most studied writers of this modification. These methyltransferases, including *SET1*, operate as part of multi subunit complexes, specifically COMPASS (both in humans and yeast, which includes SETD1A and SETD1B in humans), as well as COMPASS-like MLL1,2 and COMPASS-like MLL3,4 which are not present in yeast.

Research on the enzymatic mechanisms of these methyltransferases suggests that these distinct protein complexes catalyse the methylation to a different degree, i.e. restricting the action of the protein complex to deposit one or two methyl groups only. The *in vivo* evidence, on the other hand, often contradicts this: indicating a more complicated interdependency between these methylation states that is still not clear.

In a recent paper, Soares et al.[17] presented a way to combine these observations. The authors

shown that H3K4me is an incremental modification, deposited by COMPASS-like methyltransferases, which is indicative of the time RNA Polymerase II spends at a given locus in yeast, at the same time showing that multiple passes of the *Set1* methyltransferase are required to establish a trimethylated state. In mammals, who rely on multiple homologs of this methyltransferase, it is likely that the incremental deposition is established in a cascading fashion via a combination of enzymes. Given how important the methylation states of this lysine are at distinguishing chromatin regions (e.g. me1 marking enhancers and me3 marking promoters), it is therefore desired that the future research addresses this issue in more detail.

Other residues on histone H3 are methylated by different methyltransferases (Figure 1.4). SUV39H1 and SUV39H2 are responsible for spreading the H3K9me2/3 modification through heterochromatic regions, which is thought to depend on the deposition of monomethylation by other enzymes, such as SETDB1[42], while EHMT1 and EHMT2 (G9a and GLP) mono- or di-methylates H3K9 in euchromatin (Figure 1.4)[42]. H3K27me3 – a well established mark of facultative heterochromatin – is established by the PRC2 methyltransferases EZH1 and EZH2[42]. H3K27me2 is also associated with gene repression and mediated by PRC2[43]. While PRC2 reportedly also deposits H3K27me1[44], this mark is primarily associated with active transcription and can also be established by EHMT1[45]. In yeast, H3K36me is established through the Set2 enzyme, which follows the path of elongating RNA Polymerase II (see Figure 1.1 C)[42]. Mammals contain at least eight H3K36me methyltransferases, out of which NSD1, NSD2 and NSD3 are considered primary mono- and di- methyltransferases, whereas trimethylation can only be catalysed by SETD2[42]. Interestingly, H3K36me3 has been shown to inhibit PRC2 dependent dimethylation of H3K27 leading to H3K27me1 accumulation[44], while H3K36me2 negatively regulates the transition from H3K27me2 to H3K27me3, describing a cross-talk between these PTMs[46]. All three methylation states on H3K79 are mediated by DOT1L[42].

Previously established lysine methylations can be removed by demethylase enzymes. The first discovered lysine demethylase, KDM1A (LSD1) can demethylate mono- and di- methylated lysine 4 on histone H3 (H3K4me1 and H3K4me2)[42]. This demethylase uses FAD (flavin adenine dinucleotide) as a cofactor and interacts with LSD-CoREST complex which also con-

tains a histone deacetylase, linking the two PTMs together. Six other enzymes have also been implicated in the removal of Histone H3K4 as of today (Figure 1.4). Enzymes containing jumonji (Jmj) domains catalyse the demethylation reaction using α-ketoglutarate and Fe(II) as cofactors, which allow the all three states of methylation to be removed[42].

Removal of methylation on other histone H3 lysines is catalysed by lysine demethylases whose activity is often regulated by the presence or absence of other histone PTMs. For instance, PHF8 enzyme preferentially removes H3K9me1/2 from sites marked with H3K4me3, which is recognised by its PHD finger domain[42]. Similarly, KDM6A (UTX) is a histone H3 lysine 27 demethylase (me2/me3), which is part of COMPASS-like MLL3,4 H3K4 methyltransferase complex[42]. These examples serve to illustrate the direct cross-talk between methylated lysine residues that provides a causal mechanism to the establishment of distinct chromatin domains.

Other modifications Arginine methylation (Rme), serine/threonine phosphorylation (Sph/Tph), lysine ubiquitinylation and biotinylation (Kub and Kbio) and ADP-ribosylation are other commonly studied histone PTMs for which the writers, erasers and readers have already been established. We direct the reader to [13] for a review of such modifications.

There are many other novel modifications whose function or the precise regulatory mechanisms we do not yet know. Lysine acylations, such as propionylation (pr) or butyrylation (bu), are perhaps the most interesting novel histone PTMs as of today. These modifications are chemically very similar to lysine acetylation (propionyl has one extra methyl group as compared to acetyl, while butyryl has two) and therefore are able to be written and erased by HATs and HDACs[47, 48]. The study of these novel modifications is still in its infancy, but they serve as good candidates to provide the link between chromatin state and metabolism[47].

Modifying the structure of chromatin

Some histone PTMs are known to directly affect the structure of chromatin in the epigenetic reader independent fashion. The majority of such modifications occur within the core globular domain of the histones, as opposed to the nucleosome tails. For instance, H3K56ac, H4K77ac,

H4K79ac are proximally close to the DNA entry region to the nucleosome. Addition of an acetyl group removes a positive charge from these lysines, which reduces the interaction with negatively charged DNA causing it to unwrap[49]. Phosphorylation of the nearby serines or threonines can cause a similar effect by introducing a negative charge. The modifications that occur closer to the dyad axis of the nucleosome (e.g. H3K64ac, H3K115ac, H3K122ac) can destabilise nucleosomes allowing them to adopt alternative conformations[49]. H4K91ac can also destabilise nucleosomes, albeit by breaking down the contacts between the histones[49].

Interestingly, some histone tail PTMs also exhibit these biochemical properties. Unmodified histone H3 tails can interact with the linker DNA between the neighbouring nucleosomes, charge-altering modifications on such tails disrupt their interaction with the DNA[50]. Addition of the linker histone H1 shelters such interactions, preventing the writer enzymes from accessing these tails[50].

The histone H4 tail, on the other hand, interacts with the histone surface of nearby nucleosome. It has been shown that lysine 16 of histone H4 (H4K16) preferentially interacts with the acidic patch on histones H2A and H2B on the neighbouring nucleosome[25]. Acetylation of this residue breaks down this interaction causing chromatin to decompact[25].

Chromatin remodellers While biochemical properties of the nucleosome can modify the DNA accessibility by altering the structure of the chromatin, more drastic effects are achieved with the help of adenosine triposphate (ATP)-dependent chromatin remodellers. Such chromatin remodellers are split into four families: ISWI, CHD, SWI/SNF and INO80. Figure 1.5, which is based on the review in [51], presents a summary of such ATP-dependent remodeling activities. The ISWI and CHD family of remodellers function to establish regular spacing of nucleosomes upon the histone deposition after replication (panel A). SWI/SNF family of remodelers regulate the DNA accessibility required for transcription by sliding or removing nucleosomes (or parts thereof) (panel B). INO80 family of remodelers, on the other hand, catalyse ATP to exchange histone variants (panel C). Specifically, the SRCAP complex deposits histone variant H2A.Z, while INO80 removes it. It is important to note that this view is oversimplified

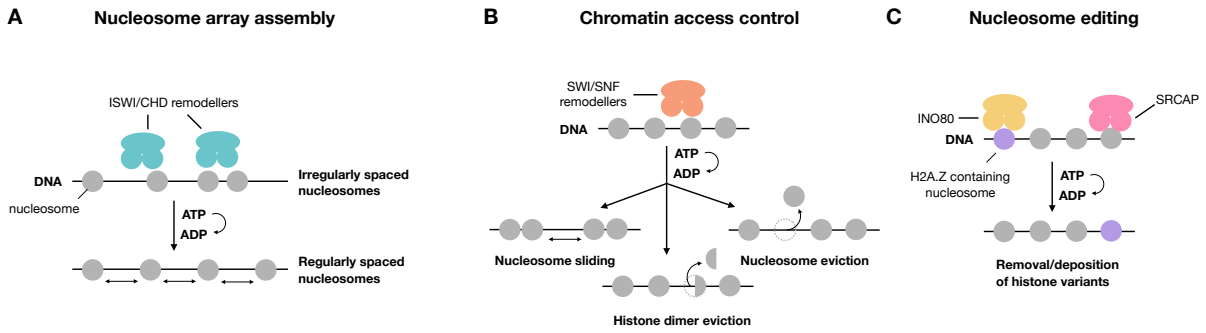

Figure 1.5: Simplified description of the main mechanisms of ATP-dependent chromatin remodelling. (A) ISWI/CHD remodellers play a part in histone deposition and spacing of the nucleosomes. (B) SWI/SNF chromatin remodellers control the access to chromatin by either sliding the nucleosomes to expose the DNA for TF binding or by evicting histone dimers, or whole nucleosomes. (C) INO80 and related SRCAP chromatin remodelers can modify the nucleosome structure by replacing Histone H2A with the histone variant H2A.Z (SRCAP) and back (INO80). Note that this is an overly simplified view, as INO80 can also promote chromatin access, so can NuRF and some CHD remodellers. Figure and caption based on a similar figure in [51, Figure 1]

as these chromatin remodelers are not limited to these functions, for instance INO80 is also known to facilitate access to chromatin[51].

1.2 Studying the language of the nucleosome

1.2.1 Experimental approaches

ChIP-seq

Most results presented in this introduction have been obtained from chromatin immunoprecipitation and next-generation sequencing (ChIP-seq) based experiments. The utility of the method and significant reduction of the costs of Next-Generation Sequencing (NGS) analyses over the last decade, have quickly made it one of the most popular tools to study the chromatin.

ChIP-seq is based on a relatively simple, yet versatile, idea which hijacks the immune system's ability to recognise foreign antigens. It relies on the availability of purified antibodies which recognise specific histone PTMs or chromatin-interacting proteins, which can be purchased

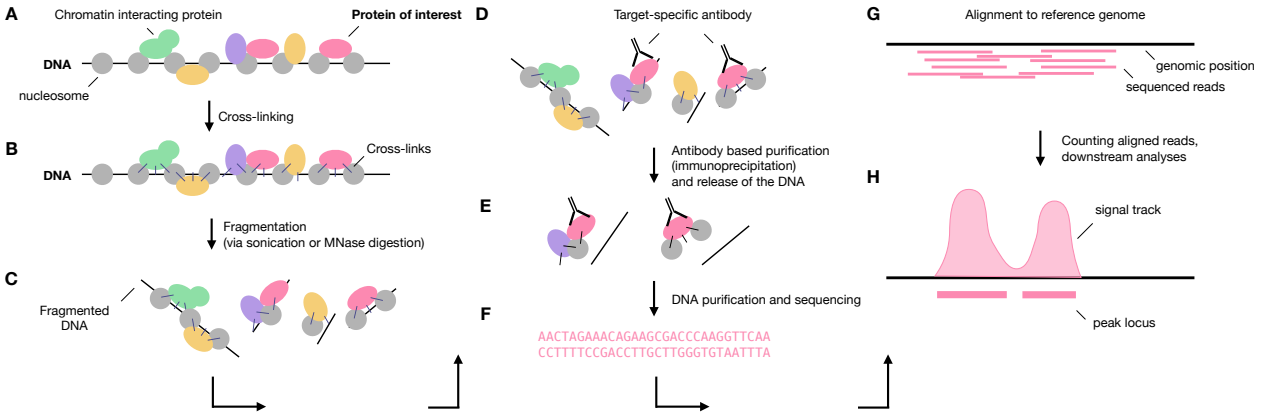

Figure 1.6: General workflow of Chromatin immunoprecipitation sequencing (ChIP-seq) experiment. Chromatin interacting proteins are cross-linked to the DNA and other proteins to preserve their interactions (A,B). Chromatin is then fragmented either via sonication or by treatment with Micrococcal Nuclease (MNase) enzyme which digests the internucleosomal DNA (C). An antibody, which is specific for the chromatin binding protein, or a histone PTM is then used to purify the DNA associated with the target (D,E). This DNA is then sequenced and aligned to the reference genome (F,G). The number of reads that associate with given chromatin locus is then counted and subjected for downstream analyses such as peak calling (H).

from multiple biotechnology companies.

These antibodies are used to recognise specific modifications or transcription factors bound to chromatin. Prior to the experiment, the chromatin is cross-linked to preserve the interactions and fragmented via sonication or MNase digestion to facilitate access (Figure 1.6 A-C). Antibody-based recognition allows the DNA associated with the protein or histone PTM to be purified (Figure 1.6 D-E). This purified DNA is then sequenced using short (30-60bp) read sequencing (Figure 1.6 F).

The sequenced reads are mapped to the reference genome (e.g. *hg38*, Figure 1.6 G). The number of reads that map to a particular locus is counted and compared to the number of reads that map to the same locus from a control (input) dataset (i.e. either the bulk fragmented DNA or the control produced using unspecific immunoglobulin). It is assumed that the number of reads associated with a chromatin locus is indicative of how often a TF or a PTM is present at this particular locus.

Downstream, the read counts are most frequently analysed for the enrichment over the control. Regions containing significantly more reads than the control are generally referred to as *peaks* (Figure 1.6 F). While the overall enrichment of reads over the control is referred to as the *signal*

track[52]. The MACS algorithm[53] can be used to identify both.

The major criticisms of this method revolve around the specificity of the antibodies, the biases associated with the fragmentation procedures, and the limit to a single PTM or chromatin-interacting protein per experiment.

Issues associated with antibodies are a concern that is shared among many biologists. Vivien Marx jokes that "[antibody] specificity makes antibodies versatile lab tools. Unless it's Tuesday. Or there's a full moon."[54]. While Marx here refers to the frustrating ability for an antibody-based experiment to fail for no apparent reason, even successful experiments may raise concerns. Multiple studies have assessed the specificity of commercially-available antibodies[55, 56]. The most recent of them[57] has re-examined results from previously published ChIP-seq studies and was able to provide an example where antibody cross-reactivity has caused incorrect interpretation of the results as the apparent H3K4me3 pattern could be explained by the antibody binding to H3K4me2 instead.

Another set of concerns regarding ChIP-seq revolve around the fragmentation efficiency. DNA in the heterochromatic regions are by definition less accessible, especially when cross-linked to proteins. Fragmentation methods, such as sonication or MNase digestion, struggle to fragment such DNA which results in less reads sequenced in the heterochromatic regions. Becker et al. have shown that this reduction occurs both in the input (control) and in the target experiments[58]. To study heterochromatin in more detail, the authors suggest a two-step purification procedure instead. In this procedure, the heterochromatin will first be separated from the easy-to-fragment euchromatin by sucrose gradient purification, and then re-fragmented to allow higher resolution in these regions[58].

Finally, classic ChIP-seq experiments cannot measure direct co-occupancy of two (or more) target histone PTMs at the same nucleosome, as each immunoprecipitation only samples one of them. We can do no better but say the modifications are likely to co-occur at a locus if their peaks overlap, but cannot claim they do co-occur on the same nucleosome. For instance, a scenario where two PTMs A and B were mutually exclusive, yet equally distributed (i.e. 50% of nucleosomes containing the modification A, and 50% containing the modification B), the

resulting ChIP-seq signal would be no different from that where the modifications were strictly collocating. Sequential ChIP technology, where the sample purified by the first antibody (for PTM A), would sequentially be subjected to the treatment with the second antibody (for PTM B) in theory could identify such interactions[59]. In practice, however, this method would require a significant amount of starting material as well as the antibody which is not feasible for anything but small scale studies (especially when proteomics can be used to answer a similar question instead).

Bottom-up proteomics on histones

Some of the questions that cannot be answered by ChIP-seq can be overcome by proteomic approaches. The direct combinatorial pattern of PTMs on the histones can be studied via Mass Spectrometry (MS) approaches. Mass spectrometers measure the mass to charge (m/z) ratio of ionised small molecules in the sample with a high accuracy. The intensity of this signal for a given m/z ratio is indicative of the abundance of the ion. The instrument can isolate particular ions of interest (based on a set of pre-defined rules) and subject them to Collision-Induced Dissociation (CID) (or one of the other forms of dissociation), which breaks the precursor ion down in a very predictable way. Protonated peptides tend to break at their peptide backbone, which allows them to be directly identified from such ions[60]. Coupling the mass spectrometer to a Liquid Chromatography (LC) system (usually reversed phase HPLC) allows complex mixtures to be analysed as peptides are separated based on their biochemical properties prior to the analysis.

In the so called *bottom-up* approach, proteins are digested into smaller peptides using proteolytic cleavage prior to the LC separation and Mass Spectrometry (MS) analysis. Smaller, digested peptides provide fewer technical challenges for identification and therefore this approach is more common than others. Typically, trypsin protease is used as it cuts strictly at the C-terminal side of lysine (K) and arginine (R) residues, which are common. Tryptic digestion further ensures that peptides contain at least one positively-charged amino acid (i.e. the lysine or the arginine) allowing for efficient ionisation of the peptides.

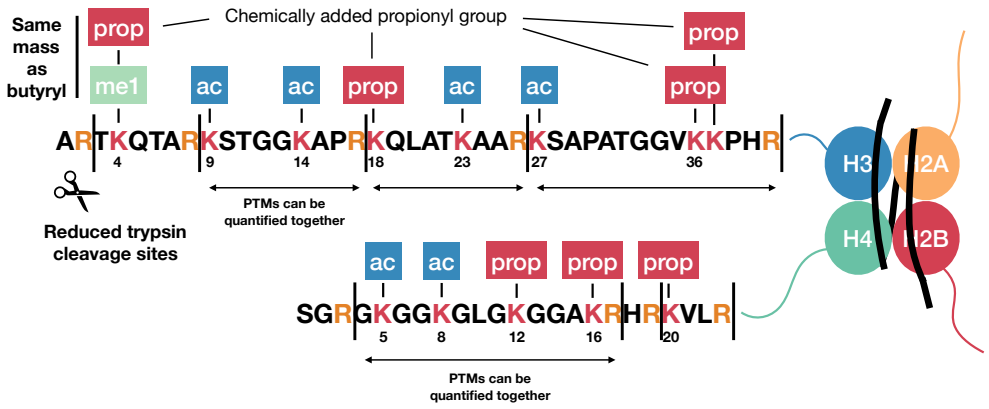

Figure 1.7: Visualisation of the effect of propionic anhydride treatment of a hypothetical H3K4me1K9acK14acK23acK27ac and H4K5acK8ac nucleosome. After this treatment, unmodified lysines, as well as H4K4me1, are labelled by additional propionyl group. This effectively blocks trypsin from cutting at the C-termini of lysine (K) residues, leaving only arginine (R) residues exposed to this digestion (i.e. turning it into an Arg-C enzyme). PTMs that occur between arginine residues can therefore be quantified together.

Histone peptides, which are rich in lysines, provide a challenge to the canonical bottom-up approach as peptides resulting from such digestions are often too small to retain on chromatography columns[61]. More importantly, the goal of proteomic analyses of histone proteins is not the identification and quantification of the histone proteins themselves, but rather the quantification of histone PTMs. Trypsin digestion efficiency is influenced by the presence of such modifications, which often occur on lysines, to the extent that addition of even the lightweight methyl groups compromises it. This results in a highly complex mixture of peptides cut at different residues which provides a challenge for analysis.

The ability to purify the histone proteins with high efficiency[62], however, has allowed the development of a specialised bottom-up approach for the analysis of such PTMs. In this approach the purified histone proteins are treated with propionic anhydride, which chemically labels the unmodified (and monomethylated) lysines with a propionyl group (Figure 1.7)[61, 63]. This prevents the trypsin digestion at the C terminus of all lysine residues, turning it into a protease which only cuts at arginine residues (i.e. similar to the Arg-C protease).

By limiting trypsin digestion to arginines only, the search space for modified peptides is reduced significantly, which allows for efficient identification and quantification of their modifications for which computational tools are available[64, 65]. This also allows accurate quantification of the

co-occurrence of nearby PTMs that occur between the arginine residues. For instance, the global levels of H3K9ac and H3K14ac co-occurrence (H3K9acK14ac) can be directly compared to the levels of the co-occurence of both of these lysines in their unmodified state (H3K9unK14un), as well as to the levels of H3K9acK14un and H3K9unK14ac. But the co-occurrence of, say, H3K27ac and H3K9ac, which occur on different tryptic peptides (see Figure 1.7), cannot be directly estimated via this method. Alternative proteomic strategies (termed *middle-down* and *top-down*) that use different proteases that cut less frequently have been developed to address these issues, but they are beyond the scope of this thesis (see review in [66] instead).

Importantly, the ability to directly quantify the relative levels of nearby PTMs provides an advantage of this method over ChIP-seq. Yet the *bottom-up* approach, as well as the *middle-down* and *top-down* approaches, do not provide any information about the co-occurrence of modifications across the nucleosomal histones, nor do they provide information about the genomic location of these modifications. ChIP-seq, on the other hand, even if it does not directly measure it, provides a correlative measurement of such co-occurrence, highlighting the complementarity of the two datasets and providing a link to combine them, which we briefly discuss in section 3.4.

It is further important to note that the described bottom-up approach cannot distinguish naturally occurring lysine propionylation and butyrylation[47, 48] from unmodified and monomethylated lysines respectively, as these modifications have the same mass as the equivalent chemical derivatives. If accurate quantification of such residues is desired, isotopically labelled propionic anhydride can be used instead as it will produce a mass shift that is distinguishable from the naturally occurring propionylation.

Nucleosome affinity pull-downs

The method that forms the basis of the chapter 2 of this thesis relies on such isotopic labelling of amino acids.

Mass Spectrometry (MS), while being a very precise technique, is sensitive to the differences

in sample preparation and is biased by other associated technical issues such as the protease digestion efficiency and the biochemical properties of peptides. These biases complicate the relative comparison of protein abundances across the experimental conditions as the effect from the biases could not be decoupled from the true measured effect.

The Stable Isotope Labelling with Amino acids in cell Culture (SILAC) method allows relative abundances between two samples to be compared directly, even in presence of these biases[67]. In this experiment, two populations of cells (usually treated and control) are grown. The media in one of the two batches is supplemented with isotopically labelled heavy amino acids (e.g. arginine C_6^{13}, N_4^{15} and lysine C_6^{13}, N_2^{15}) which get incorporated into the proteins. This unique labelling allows the experimental samples to be pooled in equal proportions before processing them for MS, and re-identified during the analysis. Early mixing of the samples ensures that the biases associated with the sample preparation or mass spectrometry based analysis are identical for both samples so they can be compared directly.

In SILAC nucleosome affinity purification (SNAP) experiments, isotopic labelling is used to produce two separate nuclear extracts, one labelled with *heavy* amino acids (Arg C_6^{13}, N_4^{15}, Lys C_6^{13}, N_2^{15}), the other with *light* (Arg C_6^{12}, N_4^{14}, Lys C_6^{12}, N_2^{14}). Unlike the classic SILAC approach, in this experiment the cells are not subjected to any treatment prior to harvesting, but are rather grown in as similar conditions as possible as it is necessary for accurate Affinity Purification (AP) experiments from the nuclear extracts.

SILAC nucleosome affinity purification (SNAP) experiments use a recombinant nucleosome as a substrate for such Affinity Purification (AP). The procedure of recombinant nucleosome assembly is described in [68]. Briefly, the recombinant human histone proteins were expressed and purified from *E. coli*. N-terminally modified histone H3 was produced by expressing a truncated H3.1Δ1-31T32C construct, which lacks the amino acids 1-31, and introduces a cysteine (instead of a threonine) at the 32nd position. Thioester peptides containing the remaining N-terminal H3 sequence (i.e. H3.1 1-31) together with the desired post-translational modifications were synthesised commercially and ligated to the purified truncated histone to generate a full-length protein. Modified histone H4 can be obtained by a similar chemical ligation of a

desired N-terminal peptide to H4Δ1–28I29C construct[69]. Histone proteins with the desired PTMs were then refolded into histone octamers, together with the remaining histones and then further assembled with DNA into nucleosomes. The Widom 601 DNA sequence was used throughout as it has been specifically designed to provide strong positioning of the octamer on the DNA, which is required for stable nucleosome reconstitution *in vitro*[70]. Bartke describes how this sequence could be produced in *E. coli* and purified[68]. To facilitate the pull-down experiments, the purified DNA sequence was modified by incorporation of dUTP-biotin at one of the sequence ends, which acts as a binding site for streptavidin. This approach also allows creation of nucleosomes that mimic eukaryotic DNA methylation, for which the purified 601 DNA sequence can be treated with M.SssI CpG methyltransferase which deposits a methyl group at the C5 site of cysteines in a CpG-context, similar to 5mC found in mammalian DNA.

During the experiment the heavy and light nuclear extracts are subjected to a different recombinant nucleosome each. In the forward experiment, light nuclear extract is applied to the unmodified recombinant nucleosome, and the heavy nuclear extract is subjected to the nucleosome modified post-translationally. The streptavidin-biotin interaction then allows the proteins binding to each of the two nucleosomes to be purified. The two protein samples are combined and analysed on the mass spectrometer after SDS-PAGE separation and tryptic digestion. The relative difference between the measured abundance of the heavy-labelled protein and the measured abundance of the light protein (i.e. ratio H/L) is therefore equivalent to the relative amounts of the protein bound to modified nucleosome (heavy) as compared to the unmodified one (light, see Figure 1.8 A), which provides a direct view of the epigenetic reader interactions at a nucleosomal scale.

For each modified nucleosome this experiment is run a second time, in a *reverse* setting, where the light nuclear extract is applied to the modified nucleosome, and heavy nuclear extract is applied to the unmodified one. This produces a reverse H/L ratio, which is, in this case, indicative of the ratio between the measured protein abundance bound to the unmodified nucleosome (heavy) compared to the abundance bound to the modified nucleosome (light), and is therefore an inverse of the forward H/L ratio described in the previous paragraph. It is desirable that the heavy and light nuclear extracts used in all of the experiments come from

the same batch, as it was described in [68], since this removes any confounding batch effects that would be created otherwise.

Figure 1.8 B and C provides an example of results from this AP procedure: results for H3K9me3 modified nucleosomes (B) and H3K9me3 + 5-mC modified nucleosomes (C) plotted on a log-log scale. This figure is reproduced from [68], which used HeLa nuclear extract. In these scatter plots, the proteins recruited by the modified nucleosome aggregate in the bottom right quadrant as this quadrant corresponds to non-negative $\log_2$ forward ratio, and negative $\log_2$ reverse ratio. For instance, CBX5 and UHRF1 are strongly recruited by both of the nucleosomes in Figure 1.8 B and C. The proteins excluded by the modified nucleosomes are visible in the top-left quadrant instead. It is clear that no proteins are significantly excluded by H3K9me3 (Figure 1.8 B), yet a subset of proteins is strongly excluded from binding the same nucleosome when the DNA is methylated (Figure 1.8 C).

As part of one of the main goals of this thesis, the decoding of the chromatin modifications, in chapter 2 we explore a novel, unpublished, dataset produced using this experimental technique. For this dataset, the experimental procedure for the original SNAP[68] has been modified slightly to increase the efficiency and data quality, which has allowed the scale of the original dataset to be increased five to ten fold, both in terms of sensitivity as well as throughput. The specific changes are described in the next chapter (section 2.1.1).

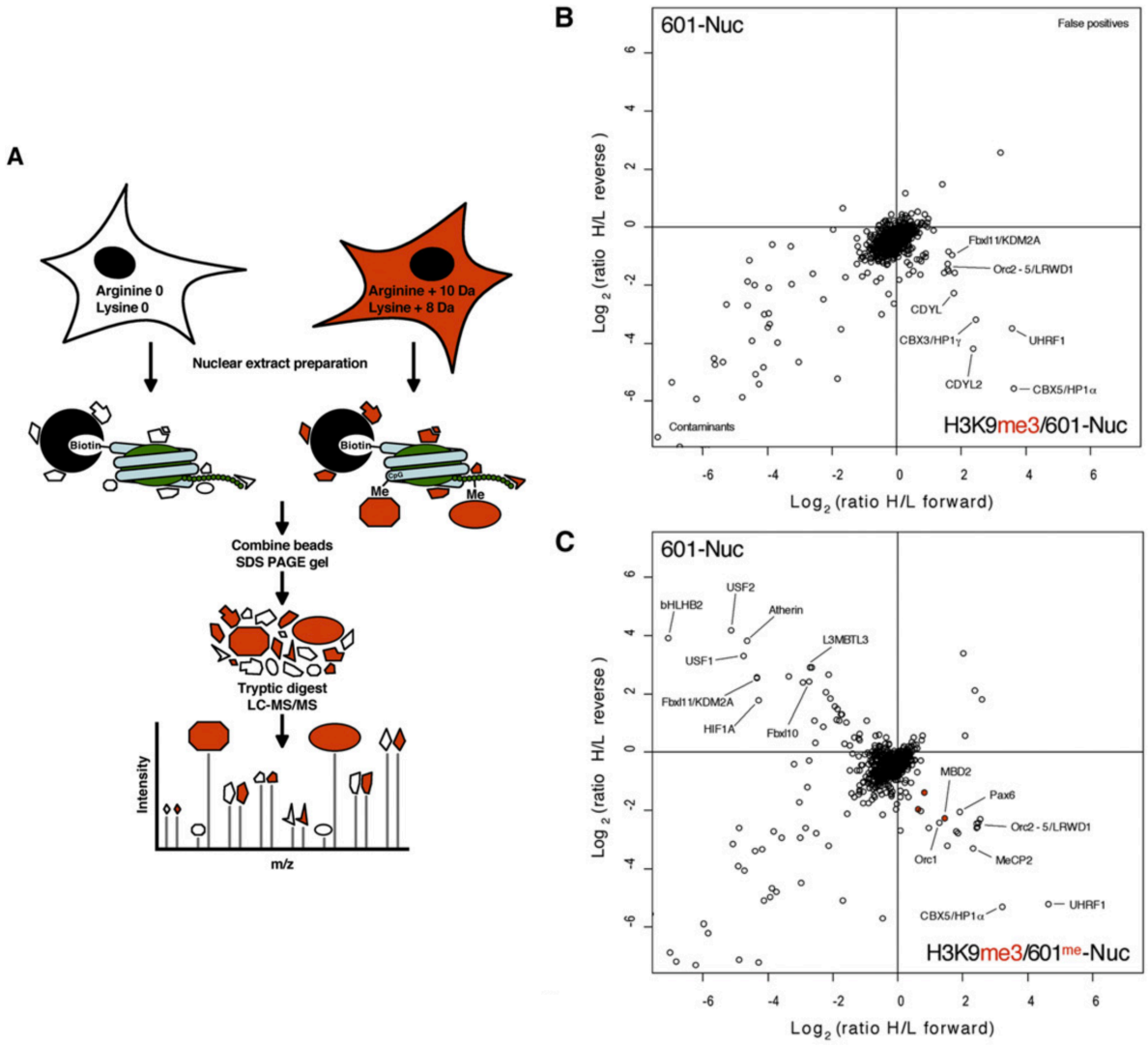

Figure 1.8: Description of SILAC nucleosome affinity purification (SNAP) workflow and sample results. (A) Description of the experimental procedure. Briefly, HeLa cells are grown in heavy and light media. The nuclear extracts from such cells are subjected to recombinant nucleosomes. One of the recombinant nucleosomes contains no post-translational modifications and/or DNA methylation, while the other nucleosome (red) is modified. In this figure, the heavy nuclear extract is subjected to the modified nucleosome. The proteins bound to the nucleosomes are purified via the biotin tag incorporated to the nucleosomal DNA. Purified proteins are pooled and analysed by mass spectrometry. The ratio between heavy and light protein intensities is indicative of the binding preferences to the modified nucleosome. The experiment is then performed in the reverse pattern, where modified nucleosome is subjected to the light nuclear extract, and the unmodified nucleosome subjected to the heavy extract (not pictured), producing a second, inverse, association ratio. (B), (C) Example results from such purification procedures for two modified nucleosomes. Nucleosome modified by H3K9me3 (B) and nucleosome modified by H3K9me3 + m5C DNA methylation (C). Proteins preferentially associating with modified nucleosome are in the lower-right quadrant of the plot, while proteins excluded by the modification are present in the upper-left quadrant. Figure reprinted from Cell, Volume 143, Issue 3, Bartke et al., Nucleosome-Interacting Proteins Regulated by DNA and Histone Methylation, 470-484, 2010[68, Figure 2], with permission from Elsevier.

1.2.2 List of notable studies

The previous section has described the key experimental techniques for studying the language of the nucleosomes. In this section we briefly review a set of publications that study chromatin regulation in high-throughput which, we believe, help put this thesis into the context.

Experimental studies

The first large scale profiling of the histone PTMs using ChIP-seq was published in 2007[71]. In this study, the authors have mapped the localisation of 20 lysine and arginine methylations on the genomic sequence in the CD4+ T-cells. This publication was quickly followed by another publication which increased the number of profiled chromatin marks to 39 by adding the data for 19 additional histone acetylations for the same cell line[72]. Together these studies established ChIP-seq as as a method for interrogating chromatin, and introduced the field to the now customary bimodal shape of H3K4me3 enrichment at the promoters which has established the histone PTMs as chromatin landmarks.

This dataset has kept its status of the biggest histone PTM ChIP-seq dataset until the data produced by the Roadmap consortium was released in 2015[52]. In this release, the Roadmap consortium profiled 111 human reference epigenomes, which not only provided functional annotations to the genetic regions in a wide range of cell types, but also provided unprecedented insights into the variation between them.

At approximately the same time as the publication of the first histone ChIP-seq paper, the bottom-up Mass Spectrometry (MS) method for characterisation of histone PTMs was published[61]. This publication eventually lead to the identification of many novel histone PTMs in high-throughput[73]. A similar MS based technique was used to establish the relative levels of PTM co-occurrence across the histone tails[74] allowing to decode modification patterns on individual histones.

MS has also been extensively used to establish human protein-protein interactions and define protein complexes. In 2015, Hein et al. have published the predicted human interactome dataset

consisting of 5,400 proteins and 28,500 interactions inferred from a library of 1,125 APs[75]. The release of this dataset has substantially increased the knowledge about the interactions between proteins and their complexes, including epigenetic ones.

Five years prior to this publication, Vermeulen et al. characterised protein binding to the peptides methylated at H3K4, H3K9, H3K27, H3K36 and H4K20 using SILAC AP MS[76]. At the same time, Bartke et al. characterised the epigenetic reader response to nucleosomes modified by H3K4, H3K9, H3K27 and DNA methylation using the SNAP approach[68]. Similarly, in 2012 Filippakopoulos et al. characterised the peptide substrate preference of the complete human bromodomain family[29].

Together these studies provided the scaffold of information about epigenetic processes within chromatin on which further studies can build upon.

Computational studies

The majority of the experimental studies mentioned earlier have also employed a significant computational aspect. For instance, the publication of the Roadmap epigenome dataset required a bespoke data processing pipeline to be established along with a significant number of downstream analyses that had to be performed. Similarly, the work of Hein et al.[75] has significantly relied on the data science based methods to establish the protein-protein interaction network as it is inevitable at this scale.

Completely data-driven studies, which rely on the data published by others, are rare. Such studies are almost exclusively based on the NGS datasets as they are more accessible.

The work that analysed the early ChIP-seq datasets, for instance, has shown that histone PTMs are predictive of gene expression[77, 78]. Similarly, Ernst & Kellis have pioneered Hidden Markov Models (HMMs) in the data-driven identification of chromatin states based on the co-occurrence of epigenetic landmarks which allowed the epigenome to be functionally annotated[79].

A handful of studies have attempted to model the interactions between chromatin readers and

histone PTMs computationally. A family-dependent relationship between DNA-binding TFs and PTMs has been shown recently[80]. Martin Vingron's group have used a combination of partial correlation and regression-based methods to infer the network of the chromatin reader interactions with PTMs, which has allowed the identification of the PRC2 relationship with H4K20me1[81, 82]. Two other groups have also described methods to infer similar co-occurrence networks[83, 84].

Literature-based studies

Besides experimental and computational approaches, multiple literature based studies have increased our understanding of epigenetic processes. Particularly, the EpiFactors database has provided a global manually curated list of epigenetic proteins and complexes along with their functional annotations[39]. It has not been maintained since the original publication and therefore is out of date, but it still serves as a useful resource.

More up to date descriptions of protein function can be found in UniProt[85]. Epigenetic complexes described in this database are slowly getting deposited into the Complex Portal[86] as well. The BioGRID database, on the other hand, collects the protein-protein interaction information from multiple studies and displays it in an unified format[87].

Major review papers such as [4, 5, 13, 21] attempt to consolidate this knowledge in a similar way, yet usually for a very narrow topic.

Together these studies and databases provide access to the data produced in the experiments described previously, and provide the context of the relevance of such results.

1.3 Aims of this thesis

The general rules for how epigenetic readers interact with nucleosomal modifications are well understood, but it is still far from clear how multiple modifications are interpreted. Nucleosomal modifications are considered to form a *code*, or, as we argue in this introduction, a *language*,

which is still largely not understood. In this thesis we aim to generate novel insights into how epigenetic proteins *read* this language, using a set of data-driven approaches.

Specifically, we are interested in the exact constituents of the nucleosome code, and whether they can be identified using the data available. We are interested in the proteins that read and respond to these constituents of the code and in what way, as well as interactions between such epigenetic proteins.

We base our observations on the data from a recent large-scale SILAC nucleosome affinity purification (SNAP) experiment carried out in our group, which provides a direct read-out of epigenetic reader interactions with modified nucleosomes, and complement it with the data from Next-Generation Sequencing (NGS) experiments.

In this thesis we present approaches to handle this data set and to analyse the results. In order to identify the interactions between the epigenetic proteins we profile a number of network-inference approaches on the SNAP dataset and explore the results inferred by the best performer to identify novel interactions.

In order to establish the main constituents of the nucleosome code we employ a Factor Analysis (FA) approach. FA allows us to estimate the main regulatory dimensions that govern this protein interaction network, like principal component analysis would, but at the same time allows us to trade some of the regulation explained for interpretability – something that is crucial in the context of the epigenetic code.

In order to describe how the proteins respond to the nucleosomal modifications, we establish a linear model based approach to identify the Post-Translational Modifications (PTMs) regulating the individual chromatin readers. This linear model is intended to provide a building block for higher order models for further analysis of the complex modification codes.

To verify the accuracy of this linear model we also describe an equivalent model for the Chromatin immunoprecipitation sequencing (ChIP-seq) dataset and compare the results between the models.

To investigate the cases where the predictions of the model differ we also describe a Bernoulli

Mixture Model (BMM)-based clustering approach for ChIP-seq datasets. This model allows us to explore the chromatin reader interactions with the genome and provides an insight into how the *in vivo* regulatory mechanisms impact the behaviour of chromatin readers.

We find that the data described by the SNAP dataset provides an immediately useful resource for the researcher and therefore describe an interactive interface for querying this dataset.

We further discuss that the novel insights that were generated form the network and factor analyses are of a substantial interest to the community, and use the published literature to describe their validity. Similarly, we suggest that the results from the linear model analyses could be used to classify the epigenetic readers based on their PTM preference. We also suggest that ChIP-seq data can be used for this purpose too. Finally, we hypothesise that the BMM clustering method could in theory be combined with data from Affinity Purification Mass Spectrometry (AP-MS) experiments as well in order to provide additional insights into the composition of the epigenetic code.

In the context of the published literature, this study, in its spirit, is most similar to the EpiFactors database. However, unlike the EpiFactors database, we employ a completely data and experiment-driven approach for recovering the regulatory landscape of chromatin, rather than solely relying on the published literature. This way we are not only describing what is already known, but also provide novel insights into chromatin regulation, simultaneously providing the data to back them up.

We also recover protein-protein interactions within the epigenetic readers, which draws parallels to the work presented in the BioGRID database including the work in [75]. Further, we describe a method to mine ChIP-seq datasets for insights into epigenetic reader interactions with nucleosomes, some of which, is conceptually similar to the analyses performed as part of the Roadmap project[52] or the predictive analyses presented in [77, 78, 80].

Taken together, in this thesis we aim to bring the study of the language of nucleosomal modifications to a new level. Instead of tackling a single aspect of the epigenetic code, we mine the datasets available to us, as well as integrate the results with other databases, to provide

a multi-dimensional view into the chromatin regulation. We illustrate the interconnectedness across these dimensions and highlight how they complement each other. We view the analyses described in this thesis as building blocks for novel hypotheses about epigenetic regulation that can only be made from an integrated analysis. We therefore explore the results of the analyses ourselves to illustrate the utility of this approach by example. We highlight the key findings from our analyses ourselves and thus contribute to the understanding of this language of nucleosomal modifications directly. At the same time we hope to make an indirect contribution as well by providing a resource to the scholars of epigenetics that enables them to produce further findings.

Chapter 2

Study of the SILAC nucleosome affinity purification dataset

2.1 Introduction

We begin the description of our attempts to understand the language of the nucleosomes with the SILAC nucleosome affinity purification (SNAP) dataset. The SNAP method, originally published in 2010[68], directly queries chromatin reader interactions with modified nucleosome substrates. This is achieved by comparing the abundance of proteins bound to a modified nucleosome to the abundance of proteins bound to a generic, unmodified one.

In this chapter we analyse the next iteration of this dataset which has recently been produced by Dr. Nhuong Nguyen (postdoc, Bartke lab) and is currently unpublished. This new iteration of the dataset queries 55 differentially modified di-nucleosomes – a so far unprecedented scale.

We use this dataset to study the epigenetic regulation by the nucleosomal modifications in a high throughput fashion. We aim to identify the principle nucleosomal modifications that define the majority of the regulation in the dataset, as well as to characterise the chromatin reader proteins by their preference to various modifications. We further aim to determine the interplay between epigenetic reader proteins, and to characterise how this interplay is regulated

by epigenetic modifications.

2.1.1 The dataset

The library of 55 recombinant di-nucleosomes profiled in this novel SNAP dataset has been designed to capture a variety of nucleosomal modification states present *in vivo*: promoter states (marked with H3K4me3), enhancer states (marked with H3K4me1), heterochromatic states (marked with DNA methylation and/or H3K9me3/H3K27me3) as well as combinations thereof. The distinct recombinant octamers have been numbered in an increasing fashion, prefixing their number with the letter "H" (for "**H**istone octamer"), e.g. H01. Nucleosomes reconstituted in the presence of 5-methylcytosine (5-mC) methylated DNA (at CpG sites) have been encoded with the suffix "M" (e.g. H01M). Figure 2.1 A displays the complete library of the recombinant di-nucleosomes.

This library has been produced using modified histones generated by native chemical ligation, essentially as in [68] (see also section 1.2.1), extending the method to allow for modifications on histone H4, and creation of di-nucleosomal arrays with or without H2A.Z. A detailed description of the methods is provided in [69]. The use of di-nucleosomal substrates adds a small linker sequence between the two Widom 601 sequences, which provides a potential docking site for DNA-binding proteins. Di-nucleosomes also provide a more biologically relevant substrate as it allows us to study the inter-nucleosomal interactions of histone Post-Translational Modifications (PTMs) such as the H4K16ac interaction with the acidic patch of the neighbouring nucleosome[25]. Finally, the additional histone octamer increases the signal-to-noise ratio by providing two substrates for every DNA segment immobilised to streptavidin and reduces the effect of DNA repair factors that bind to the blunt end of the DNA.

In total, the library of nucleosomes profiles the modifications on six lysine residues on the tails of histone H3 (K4, K9, K14, K18, K23 and K27) and five on histone H4 (K5, K8, K12 and K20). These lysines are modified with acetylation (ac) or mono-, di- or tri-methylation (me1, me2 or me3) or combinations thereof. Five nucleosomes also contain the variant histone H2A.Z, substituted in place of canonical H2A. Nine nucleosomes contain 5-mC DNA at CpG

sites, while the DNA in others is free from methylation. As seen in Figure 2.1, the nucleosome library can be divided into two groups: a group containing euchromatic modifications, such as H3K4me and acetylations, and a group containing repressive modifications, such as DNA methylation and H3K9me3.

The reader is encouraged to note the format of the Figure 2.1 A as it will be used consistently throughout the thesis. In this representation each row corresponds to one of the 55 di-nucleosomes used in this experiment. The columns of this figure list the potential modification sites on the di-nucleosomes. Filled squares indicate the presence of a modification on a given nucleosome, while empty squares highlight unmodified residues (see Figure 2.1 which provides a sketch for the nucleosome H45). Mono-, di-, and tri-methylations of the lysines are colour coded in green colours, acetylations are highlighted in blue. Histone variant H2A.Z and DNA methylation form the first and last columns respectively, and are coloured purple and cyan. We use this format as it makes easy to compare the nucleosomes between pull-downs visually.

The heavy and light nuclear extracts used in this SNAP experiment have been prepared from HeLa S3 cells as described before[68, 69]. The same batch of each of the extracts has been used in all of the 55 pull-downs as to eliminate the batch effect. In order to maximise the reproducibility, this batch has been prepared as a mixture of three independent nuclear protein extractions from three independent HeLa cell populations grown in the appropriate SILAC media. Nucleosomes corresponding to 12.5 μg of octamer were immobilised on streptavidin beads and incubated with 0.5mg of Stable Isotope Labelling with Amino acids in cell Culture (SILAC) labelled nuclear extract as described in [69]. Unlike in [68], SDS-PAGE purification step has been skipped, eluting the bound proteins directly via the digestion on beads instead[69].

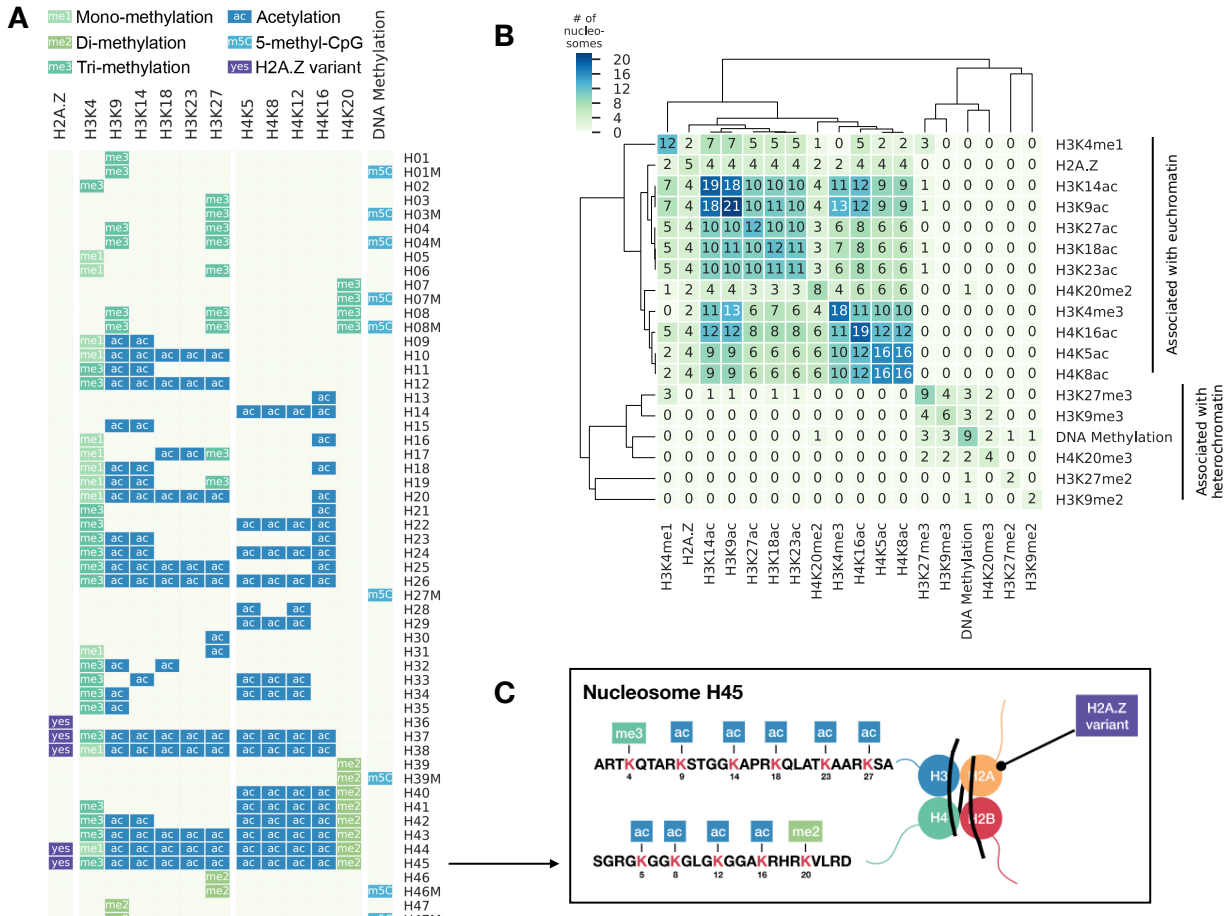

Figure 2.1: Design of the novel SNAP dataset analysed in this thesis. (A) The library of 55 recombinant di-nucleosomes that were used to generate the SNAP dataset. Nucleosomes are listed in rows, and encoded using the numbered notation prefixed with an H. Modified locations on the recombinant nucleosomes are listed in the columns. Filled squares in the table indicate the presence of modification on a particular nucleosome and should be combined row wise. An example how to read this table is shown in panel (C). (B) Co-occurence of the nucleosomal modifications in the dataset. Numbers in the heatmap indicate the number of nucleosomes two modifications co-occur in. The diagonal entries show the total number of nucleosomes containing the modification. (C) a sketch of the nucleosome H45.

2.2 Methods

2.2.1 Data visualisation

Scatter plots

Since each of the nucleosome pull-downs has been executed in two settings: forward and reverse, the experiments produce two values for each of the proteins. The first value, forward H/L ratio (where H/L stands for "Heavy/Light"), is the ratio between the measured protein abundance (estimated using the heavy nuclear extract) bound to the modified nucleosome and the measured protein abundance (from light nuclear extract) interacting with the unmodified nucleosome. The reverse ratio is the inverse of this quantity: the ratio between the measured protein abundance (from heavy nuclear extract) bound to the unmodified nucleosome and the measured protein abundance (light nuclear extract) bound to the modified nucleosome.

It is convenient to plot these two numbers in $\log_2$ scale on a two-dimensional scatter plot as the one shown in Figure 2.2 A. Here the forward ratio is shown on the horizontal (x) axis and the reverse – on the vertical one. In the idealised setting (which is not always the case, see section 2.2.2) the observed proteins fall on the main diagonal of this plot. Proteins enriched by the modified nucleosome are thus located in the lower-right quadrant of the plot, while proteins excluded are located in the top-left. The distance along the diagonal is proportional to the strength of recruitment/exclusion as the values are direct indicators of the enrichment ratios.

Comparing such scatter plots produced between two pull-downs can highlight the differences in binding preference to modified nucleosomes. In the Figure 2.2 B a scatter plot from H01M (H3K9me3 + 5-mC) experiment is shown. If we were to compare it to the scatter plot of the H01 (H3K9me3) pull-down (Figure 2.2 A), it is clear that UHRF1 and CBX5 levels bound to nucleosomes increase upon the addition of methylated DNA, while KDM2A levels are strongly reduced indicating the direct influence of these modifications on the nucleosomes to binding of these factors.

It is easy to imagine that such visualisation will break down quickly when exploring more than

a handful of proteins. Similarly, fitting all 55 pull-downs on a single page is only possible if the relevant details are omitted (Figure 2.2 C). In order to visualise the data better we employ a heatmap-based approach instead.

Interactive heatmap visualisation

In order to capture the global binding profiles for proteins we display them in a heatmap format. While visualisation of the complete data set is possible this way, to keep the visualisation compact we have developed an online visualisation tool which allows one to select a protein (or protein complex) of interest with the help of a browser as seen in the screenshot in Figure 2.3 A. The system then automatically selects a subset of the relevant proteins (i.e. proteins showing similar behaviour across the dataset) based on the results obtained form the network analysis (section 2.2.3) to ensure the necessary context is always provided.

The query protein and the relevant proteins are then presented in the format shown in Figure 2.3 B. At the top of this heatmap the experiment design matrix is shown, which is based on the one already seen in the Figure 2.1, but rotated sideways, and clustered based on the visualised chromatin profiles.

Initially we have clustered the columns of this heatmap locally, based on the binding patterns of the proteins visualised. We have, however, found that users preferred the columns to be clustered globally (i.e for the whole dataset) by default, as it does not change the order of the columns from one query to another. This global clustering is therefore used throughout this thesis, however, an option to re-cluster the heatmap locally is presented to the user of the online visualisation.

In the panel below the nucleosome design matrix, a small map of the functional annotations of the nucleosomes is present. Red colours encode promoter states, orange – enhancer states, green – euchromatin, blue – heterochromatin and brown encoding other states. We omit this matrix in the figures in this thesis and let the clustered design matrix speak for itself, but it is included in the interactive interface.

Finally, in the bottom part of the plot we represent the forward and reverse ratios of queried proteins in the experiments. Each cell in the plot corresponds to a protein and experiment pair, with the left half of the circle representing the reverse pull-down and the right-half of the circle representing the forward one (Figure 2.3 D). Red colours indicate recruitment to modified nucleosome, while blue colours indicate exclusion. The colours are directly related to the measured ratios of proteins. The scale is provided in the top-right corner of the heatmap, with the colour scale for reverse pull-downs inverted to match the scale for forward ones. The size of half-circle is indicative of the MS1 peak intensity of the protein i.e. the total intensity of the protein's peptides, which acts as an indication of the protein abundance, and therefore the confidence in the measured ratio for the protein (as less intense signals are more likely to produce spurious results). This intensity information is used only for the visualisation purposes in this thesis and not considered in any analyses. In some of the plots presented in the thesis, especially the plots that have been generated automatically, it is omitted.

The right-hand side of this plot is reserved for supporting information, such as the complex memberships of the proteins. The researcher can alternatively choose to display the protein domain information in a similar format instead, with the help of the control panel in the interactive interface.

Architecture of the web application

The interactive interface described previously is powered via a collection of web services, and is designed to be released together with the publication of this work. Currently the complete visualisation runs on a single virtual server with two CPUs and 1 GB of memory, but it has been designed to be easily scalable across multiple servers in order to cope with increased load upon publication.

The scalable architecture of this interface is summarised in Figure 2.4. Every component in this figure can be run on an individual server or, even multiple servers, as needed. With the help of containerisation software Docker these servers can be configured and started in minutes. The front-facing server runs an instance of NGINX which acts as a load balancer

(dividing the load across multiple downstream servers equally), as well as directly serves the static content, such as images, scripts or style information. Behind this server an instance of Flask application provides the standard navigation controls and generation of welcome pages. This is a low-resource application and is unlikely to cause any scaling concerns.

The interactive visualisation itself is powered via the Bokeh framework, which allows for easy communication between the visualisation in the browser and logical processing back-end (on server) in python. Due to this back-end interaction the visualisation itself is the most resource-demanding component. A lot of work has been put into optimisation of this back-end processing, offloading as much work as possible into an off-line pre-computation script. The modest computing server we use is sufficient for the visualisation use along our small group, however, upon its publication, it is likely that the number of requests will outgrow the computing resources available. With this in mind, the visualisation is designed to be completely self-contained. Depending on demand, a number of similar servers hosting the visualisation can be created, allowing the load-balancer node to split the traffic equally among them, thus achieving a linear scalability.

Similarly, this architecture allows to reduce the server usage once the initial peak of traffic declines, providing monetary savings.

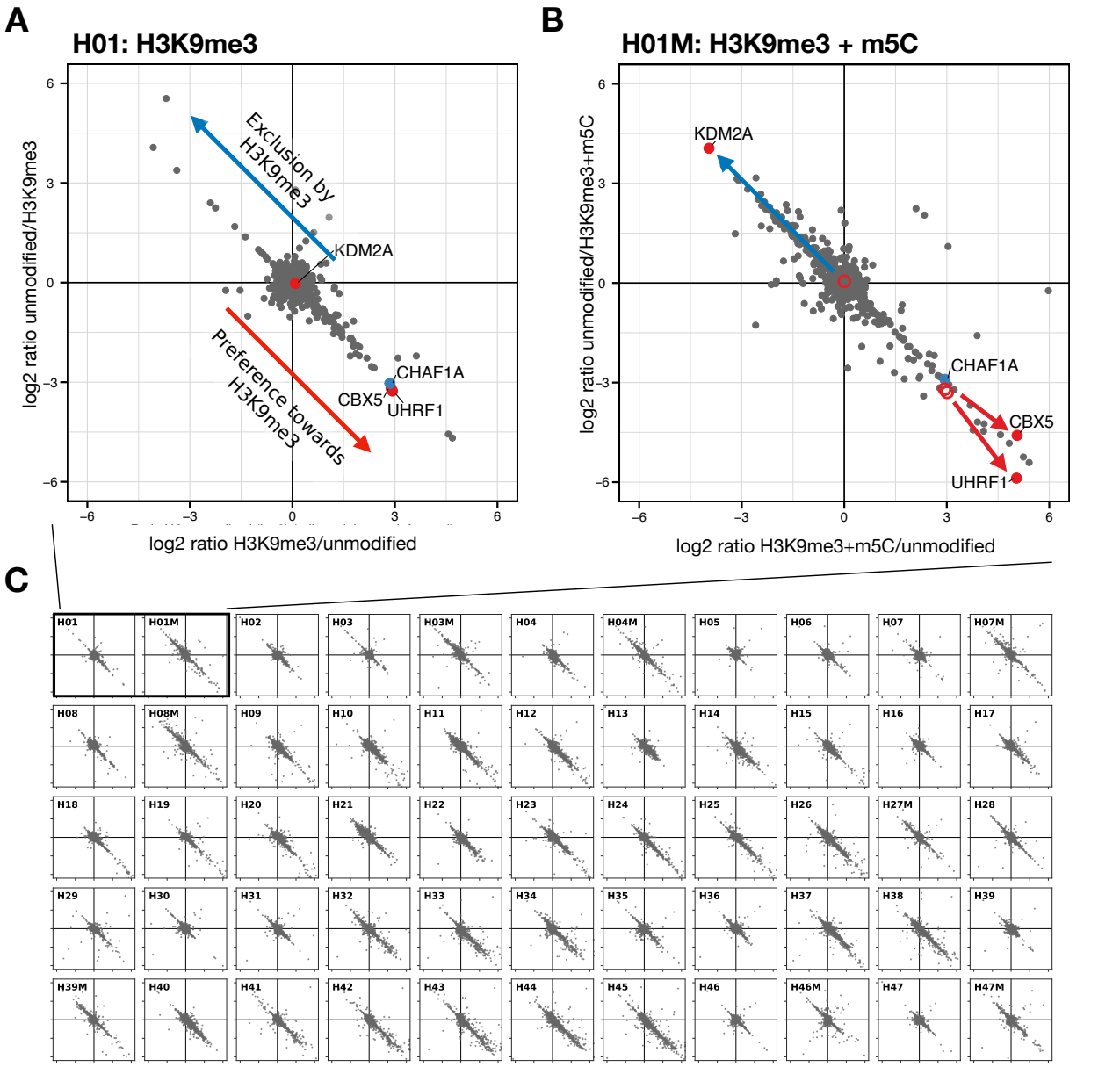

Figure 2.2: Scatterplot visualisation of a SNAP experiment. (A) Visualisation of the post-processed (as described in see section 2.2.2) result from H01 (H3K9me3) nucleosome pull-down. Proteins showing preference to the modified nucleosome appear in the lower right quadrant of the plot, while proteins excluded by this modification appear in the top-left quadrant. The other two quadrants typically contain outlier proteins (see section 2.2.2), whose influence is reduced by the preprocessing procedure. Preference/exclusion strength increases along the direction of the arrows. (B) Visualisation of the result from H01M (H3K9me3 + 5-mC) pull-down. Comparing this panel to panel A of this figure it is clear that more CBX5 and UHRF1 is associated with the nucleosome when the m5C is present. Similarly, binding of KDM2A is not influenced by H3K9me3 nucleosome at all, but is strongly excluded by the addition of DNA methylation. (C) Scatter plots for all of the pull-downs analysed, showing the scale of the experiment.

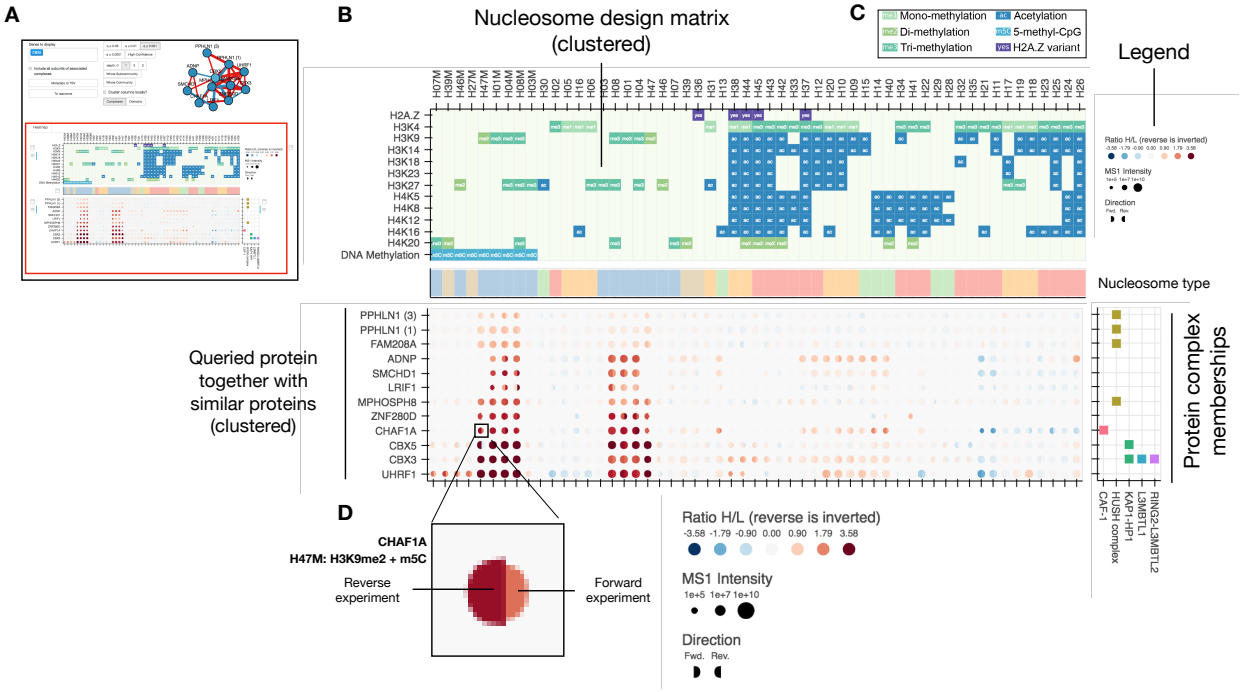

Figure 2.3: Comparative, interactive heatmap representation of protein binding profiles across multiple SNAP experiments. (A) Screenshot of the interactive interface, where the protein of interest can be queried, and the visualisation adjusted. Area highlighted in red enlarged in (B). (B) Breakdown of the heatmap visualisation used in the website and this thesis. The top part of the heatmap encodes the nucleosome design matrix, which is simply a rotated and clustered version of Figure 2.1. The legend for this figure is provided in the panel (C) and is used consistently across the thesis. Below the figure lies a nucleosome type indicator diagram where red encodes promoter states, orange – enhancer states, green – euchromatin, blue – heterochromatin and brown – other states. We frequently omit this plot in the figures in the thesis. The heatmap for each of the proteins is indicated directly below. Each row in this heatmap represents a protein, while the columns represent the nucleosome pull-down experiments. The circles inside the cells (D) encode the $\log_2$ forward/reverse ratios, with the left half of the circle representing the reverse pull-down, and right half of the circle representing the forward one. The colour scale for reverse pull-downs has been inverted to match the forward ones. Size of the half-circle corresponds to the MS1 intensity of the protein, which should be interpreted as a measure of protein's abundance, and the confidence in the H/L ratio estimate. Finally, the panel on the right hand side of the heatmap provides information about the protein complex memberships of the queried proteins. This information can be exchanged with protein domain information via the controls in the page.

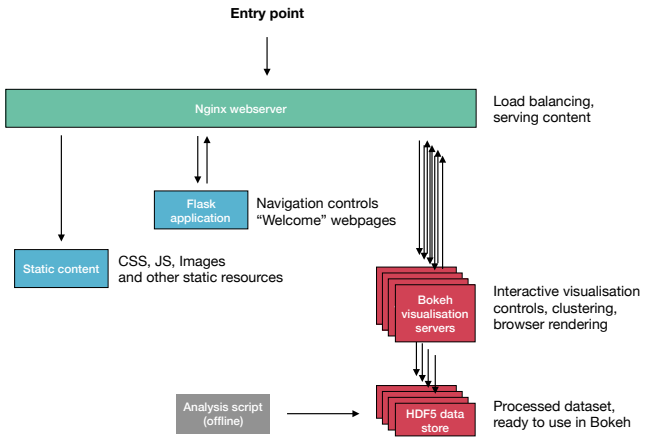

Figure 2.4: Architecture of the interactive SNAP visualisation. The NGINX layer deals with the initial request from the user and either serves the static content to the user directly, or redirects it to the Flask or Bokeh application, depending on the request. Each of these applications, including NGINX can be run in multiple instances on multiple servers, depending on the demand, as illustrated with multiple bokeh application nodes in the diagram. These nodes each store the web-ready results from data analysis in HDF5 format.

2.2.2 Data pre-processing

The scale of such nucleosomal SILAC affinity pull-down is unprecedented. For a proteomics experiment of such scale, the data preprocessing step is extremely crucial. As with any comprehensive analysis, it is important to ensure that the biases that are prominent across the different parts of the experiment are accounted and corrected for.

For instance, production of heavy and light SILAC labelled protein extracts has necessarily come from two populations of cells: one grown in light, the other in heavy medium. Even if we were to assume the isotopic labelling to be hundred percent efficient, we would still expect the intensities of proteins to differ substantially, which needs to be accounted for. Similarly, great care has been put into each and every pull-down experiment, yet the experiment-to-experiment variation still remains, which we also correct for.

The summary of the pre-processing workflow is provided in Figure 2.5 and described in detail below.

Processing metadata

This project has grown to span across multiple years, during which a great deal of knowledge has been generated. Genome assemblies have been released and updated, papers have been published, proteins have been renamed. For instance, protein *C17orf96* has since been renamed EPOP[88], *C11orf84* is now known as SPINDOC[89].

In order to maintain the appeal of this data and not let it get out of date we must periodically update this protein-associated metadata. Re-processing the raw output from the mass spectrometer with more recent proteome assemblies is infeasible with the resources available to us. Even if it was, reprocessing the dataset could result in a change of the outputs, which is not desirable. On the other-hand, updating the metadata only is a cheap, simple and efficient way to maintain the relevance of the resource.

Metadata updates have recently been made exceptionally easy via the introduction of the

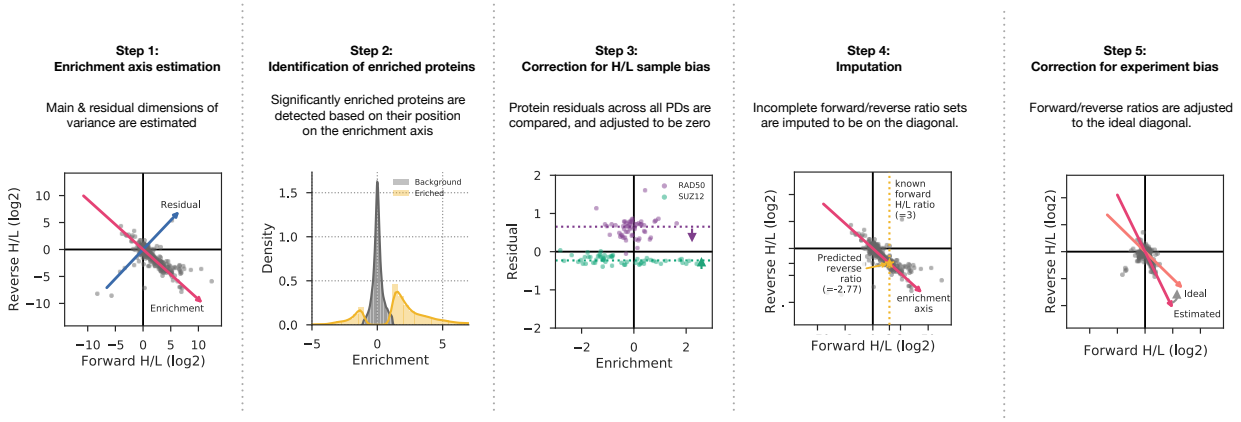

Figure 2.5: A summary for the preprocessing workflow for the large scale SNAP dataset. The main and residual variance axes are estimated for each of the individual pull-downs. The first variance axis corresponds to the degree of enrichment of the protein towards differentially modified nucleosome. Significantly enriched (both positively and negatively) proteins are then identified via their position on this axis. The residual variance direction corresponds to sample-induced bias from differentially labelled populations of cells. In the third step this bias is corrected by adjusting the proteins to have mean residual of zero. At this point the missing values, such as the cases where only one of the two ratios is known, are imputed. Finally, the datasets are corrected for mixing bias between the heavy and light populations by adjusting their main direction of variance to match the idealised case

Mygene.info API service[90]. From this service we are able to fetch the most recent protein names, gene names, as well as descriptions in a machine readable format that we can integrate with our data. Additional metadata such as Ensembl, Entrez gene IDs or Interpro annotations can also be obtained via this API.

Upon arrival, the updated information is processed through a consolidation algorithm whose primary goal is to assign an identifier for each of the rows in the dataset. This identifier is intended to serve as a natural primary key for all of our data tables, and therefore has to necessarily be unique and human readable. While UniProt identifiers, e.g. *O60885*, are unique by design they are often hard to read by humans. Gene names, for instance SMARCA4, are often easy to comprehend, but do not define protein products uniquely. In some cases, two or more proteins are indistinguishable by mass spectrometer and are therefore reported as one row in the dataset. In these scenarios we desire the key to accurately represent all of them, yet we would still like to have this key to be as short as possible so it can be used in visualisations directly.

In order to achieve all of these seemingly conflicting goals we apply the following algorithm to

each of the protein groups (i.e. rows) in the dataset:

1. Use the metadata from Mygene.info to identify all the recent gene names for the protein group.
2. Identify longest common prefixes across these gene names, for instance prefix *SMAD* that is common among proteins SMAD2, SMAD3 and SMAD9.
3. If three or more proteins within a group share a common prefix, merge them into a single annotation using square bracket notation. The aforementioned *SMAD* protein group would therefore result in the label SMAD[2,3,9].
4. Where it is possible, further shorten the annotations by combining consecutive number spans (e.g. CALM[1-3] vs CALM[1,2,3]).
5. In cases where no gene name annotations are associated with proteins, fallback to their UniProt identifiers.
6. If above process results in duplicate labels across proteins, append a parenthesised digit to the duplicated proteins, e.g. SMARCA (1) vs SMARCA (2). Assign lower numbers to proteins with higher confidence score as reported from MaxQuant.

Once the unique labels are established for the proteins, the indices of the interim data tables are replaced and the subsequent data pre-processing steps occur. For the reader's benefit we provide an appendix to this thesis listing all the proteins mentioned in this thesis, their short summaries and any additional information necessary.

Step 1: Data normalisation, inference of the main diagonal

As described before SILAC experiments produce two quantities for each of the proteins: the forward Heavy/Light (H/L) and the reverse H/L ratio. Forward H/L ratio is defined as the protein abundance (from heavy nuclear extract) bound to the modified nucleosome divided by the measured protein abundance (from light nuclear extract) interacting with the unmodified

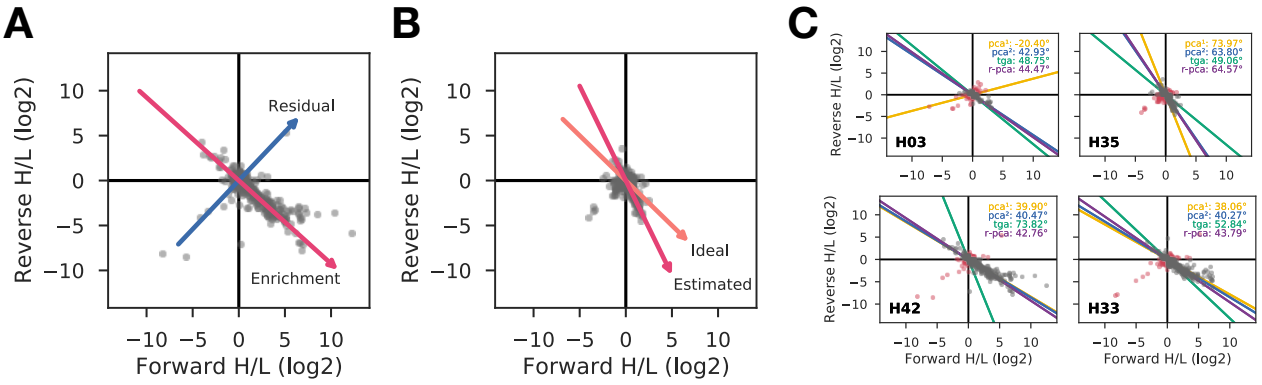

Figure 2.6: Identifying the major axes of the data variance within a SNAP experiment. (A) Visualisation of the two axes, enrichment axis follows the main scatter of the data, along the $y = -x$ diagonal and corresponds to the protein's preference to the nucleosomes. The residual axis is perpendicular to the aforementioned axis and describes the heavy/light sample bias for a protein. (B) H35 Pull-Down: an example of extreme deviation of the main enrichment axis from the idealised diagonal. (C) Comparison of the performance of multiple variance decomposition algorithms. Key: pca^1 - Principal Component Analysis (PCA) applied to the whole dataset, pca^2 PCA applied to only inlier points (black), tga - Trimmed Grassmann Averages (TGA) algorithm, applied to inlier points, r-pca repeated iterative PCA algorithm described in this thesis. Inferred angles of the lines are listed in the top-right corner, Pull-Down IDs are listed in the bottom left corner.

nucleosome. The reverse ratio is the inverse of this quantity, measuring the change in abundance between the unmodified and modified nucleosomes.

Since both of these ratios measure the same thing: the enrichment of the protein towards the modified nucleosome, the data is inherently one-dimensional. If we were to assume the existence of this dimension, which we will call *enrichment* dimension, in an ideal scenario, this single dimension would manifest as a diagonal $x = -y$ line in the SILAC plot. Due to unequal mixing of the heavy and light nuclear extracts between the experiments, this line shifts up or down. For instance, pull-down *H35* in Figure 2.6 B has an extreme skew due to an experimental error. Fortunately, as we describe later on in the text, this bias is both detectable and correctable.

We call the dimension perpendicular to the aforementioned *enrichment* dimension the *residual* dimension (Figure 2.6 A). This dimension captures both the random experimental noise as well as the differences between heavy and light samples. Positive value on this dimension indicates more of the protein's intensity coming from the heavy sample than expected, which could happen due to problems with protein labelling, differences in protein expressions across

the nuclear extracts, or bias in quantification by mass spectrometry. Core histone proteins are specific outliers in this dimension as the recombinant nucleosomes are explicitly light-labelled, and there are almost no heavy-labelled histones in the nuclear extract, causing their residuals to be extremely negative.

Standard variance decomposition approaches, for instance PCA, can be used to identify both the *enrichment* and *residual* dimensions. Naive application of PCA on our dataset puts too much weight on the outliers with extreme residuals, such as histones, sometimes even detecting the residual axis as the main (Figure 2.6 C, H03, yellow line). To counter this, we only perform PCA analysis on points falling within the top-left and bottom-right quadrants of the data (Figure 2.6 C, black points and blue lines) as these points contain the information we are interested in. This removes the extreme bias from the histone proteins, but does not discount the bias of other significant outliers. For instance in the pull-down *H42*, shown in the same figure, the PCA estimate is skewed towards the group of off-diagonal outliers in the bottom-right quadrant. Sensitivity to outliers is a fairly documented behaviour of PCA, inspiring an outlier detection algorithm based on this decomposition[91]. To counteract this we have profiled the Trimmed Grassmann Averages (TGA) approach which adds outlier-robustness to PCA by averaging the results from multiple subspaces spanned by the data[92]. In our tests, however, it has performed even worse than PCA, as shown by the green lines in Figure 2.6.

Realising that we need to solve a much simpler problem than robust variance decomposition in a general case, we have opted to use a heuristic algorithm instead. In this heuristic we keep repeating the PCA procedure multiple times, removing outlier proteins on the residual dimension between every iteration. Specifically, the algorithm is defined below:

1. Remove all data points not within the top-left and bottom-right quadrants;
2. Fit PCA to the remaining data points;
3. Find the outliers on the residual dimension (PC2) that are more than two standard deviations away from the mean, and remove them;
4. Keep repeating steps 2 and 3 until the angle of the enrichment dimension does not change.

As shown in purple lines in Figure 2.6 C, this algorithm appears to be less responsive to the group of outliers above the diagonal (lower panels) as compared to others. We have therefore heuristically selected it for the estimation of the enrichment and residual axes for the subsequent preprocessing steps.

Step 2: Detecting significant enrichments

Naively, in order to detect significantly enriched proteins in the pull-downs we would need to test both forward and reverse ratios for enrichment and then combine their results. Decoupling the enrichment dimension from the residual dimension, as described in the previous section, allows us to perform the test directly on this estimated enrichment value instead. The log2 protein ratios obtained from the SILAC experiments generally do not follow a standard normal distribution, but in the leading protein analysis software MaxQuant, the distribution tails are assumed to be approximately normal[93]. Specifically, in this seminal paper, Cox and Mann[93] compute *left* and *right* standard deviations (s_L and s_R) of this data by estimating the 15.87, 50 and 84.13 percentiles of the ratios (r_{-1}, r_0, r_1):

$$s_L = r_0 - r_{-1} \quad s_R = r_1 - r_0 \tag{2.1}$$

Reproducing the equations from [93] here, the p-value for a given ratio r can then be estimated from the following statistic as:

$$z = \begin{cases} \frac{r - r_0}{s_R}, & \text{if } r \geq r_0 \\ \frac{r - r_0}{s_L}, & \text{if } r < r_0 \end{cases} \tag{2.2}$$

$$\text{p-value} = \frac{1}{2} \times \text{erfc}\left(\frac{z}{\sqrt{2}}\right) = \frac{1}{\sqrt{2\pi}} \int_z^\infty e^{-t^2/2}\, dt \tag{2.3}$$

The reasoning for separate *left* and *right* standard deviations lies in the observation that the design of SILAC experiments forces the spread of protein intensities to favour one side of the

distribution more proteins are expected to bind the bait (in our case the nucleosome), than to be excluded by it. This is clearly the case with ratios obtained from our experiments too: the ratios for proteins recruited by the nucleosomes are more extreme than excluded ones. Pooling these variations together would not result in an accurate estimate.

On the other hand, after the enrichment decomposition, the variance due to true noise is likely to be similar in both directions, given our bidirectional (forward and reverse) experiment design. The variance *estimates* in both directions will likely be different, as the enriched direction would have a larger spread of proteins associated with it, but not the true variances. We would therefore expect $s_R \approx s_L$, but $\hat{s}_R > \hat{s}_L$ in the general case. This way, applying Equation 2.2 would result in a number of false-negatives associated with the enrichment direction as it's variance is estimated inaccurately. To counter this we change the test to consider the minimum of the two estimated variances, irrespective of the directionality, for our dataset:

$$z = \frac{r - r_0}{min(s_R, s_L)} \tag{2.4}$$

The obtained test statistics for any protein/pull-down pairs are then computed using a combination of Equation 2.4 and Equation 2.3, where S_L and S_R are estimated from a pooled set of experiments. The p-values are then corrected using Benjamini-Yekutieli method and the enrichment is called at 0.05 threshold.

Note that while the p-value estimation procedure described here provides a way of generating a list of proteins that are up- or down- regulated in a given dataset, it is rarely used in the subsequent analyses. Specifically, the protein lists are used only in section 2.3.1. In other analyses, we only use the number of experiments the protein is significantly enriched in as a summary statistic for the protein, which is used for data pre-filtering and weighing of the subsequent statistics. The values analysed in other analyses are always the corrected forward/reverse ratios as described further in this section.

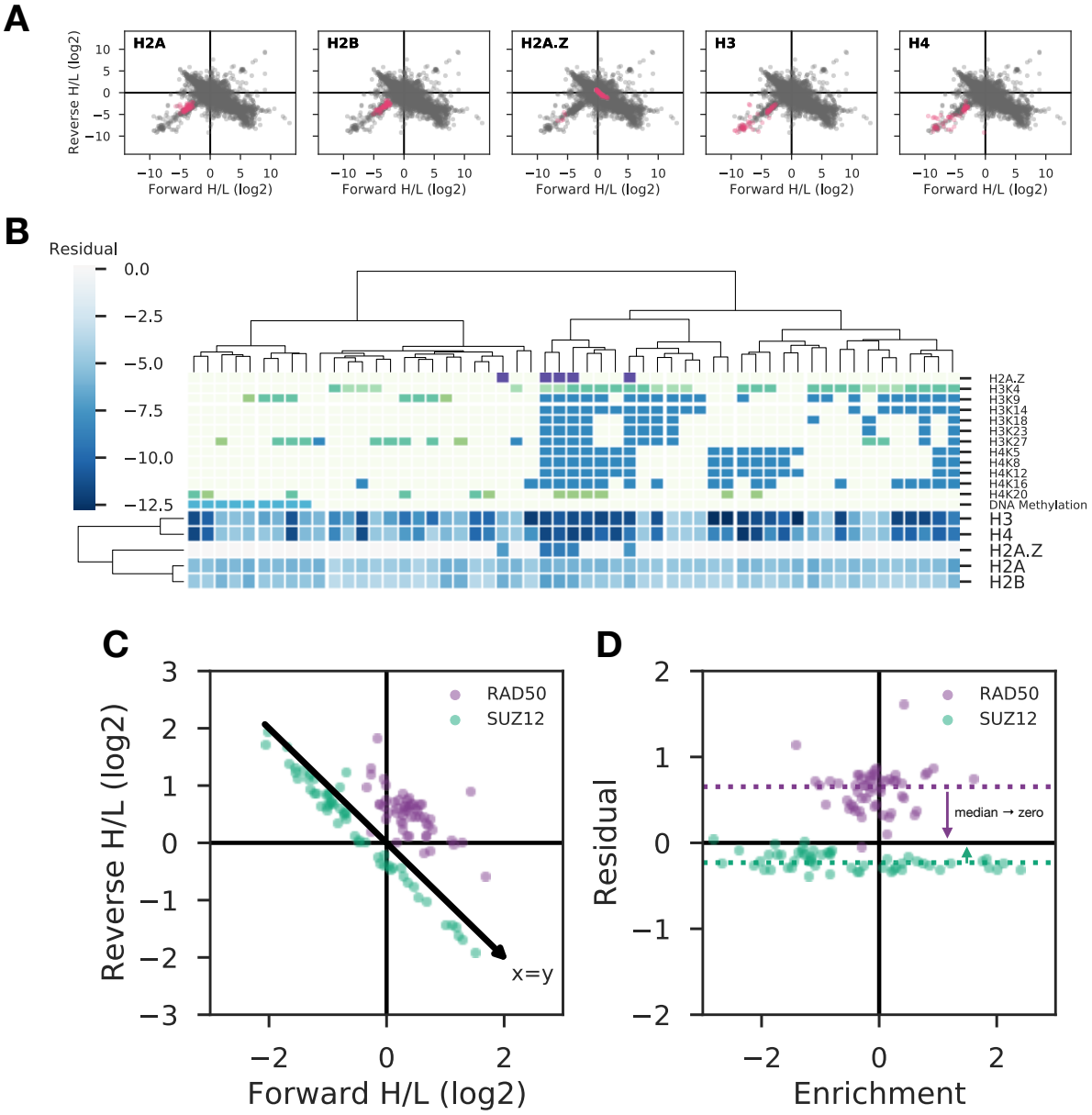

Figure 2.7: Illustration of the residual correction procedure for SNAP data. (A) Position of histone proteins on the two-dimensional ratio plots. Forward and reverse ratios for all proteins in all experiments are plotted on a single plot. Histones are marked in red and found exclusively in the bottom-left corner of the scatter plot, perpendicular to the main diagonal. Since histone H2A.Z is only incorporated to nucleosomes for a small number of pull-downs it is positioned off the main diagonal only for a subset of pull-downs. (B) The information in (A) presented in a heatmap format, where we plot the values of residual dimension. It is now clear that H2A.Z residuals are near-zero for all pull-downs except the H2A.Z containing ones. (C) Non-histone proteins tend to fall onto an offset diagonal across the experiments and therefore appear to have a residual value associated with the protein. For instance SUZ12 is consistently below the diagonal for all experiments, while RAD50 lies consistently above. To correct for this, the proteins are adjusted to have a median residual of zero (D).

Step 3: Correction for heavy/light sample bias

We have only focused on the enrichment axis of the data so far as it directly captures the desired output of the experiment. The other, residual, dimension is arguably important as well. Residual bias is perhaps best illustrated with the histone proteins, especially histone H2A.Z (H2AFV/H2AFZ). Inevitably a trace amount of this histone is present in the nuclear extract, which can be detected by the mass spectrometer. In five experiments, however, we have also added a recombinant Histone H2A.Z to the nucleosome, which is also detected. The added recombinant H2A.Z is necessarily light labelled which causes a significant surplus of the light labelled protein intensity, and therefore a strongly negative value on the residual dimension. This can be clearly seen in Figure 2.7 A and B where H2A.Z behaves like a regular protein for most pull-downs, with its values following the main diagonal and near-zero residuals, except for five pull-downs where H2A.Z has been added to the nucleosome.

Histones are easily noticeable, but not representative, examples of the protein relationships with the residual dimension. The residual skew in other proteins is introduced by the variation in the experiments which is caused by the differences in extraction efficiency from the two cell populations, differences in labelling efficiency, differences in protein identifiability upon labelling, inaccurate protein concentration measurement of the samples, etc. This observed bias is consistent across all of our experiments, and therefore the residual values for a protein tend to be the same across the experiments. This has the effect of proteins consistently appearing slightly above or slightly below the main diagonal (Figure 2.7 C). An enrichment/residual plot provides a rotation of the same data (Figure 2.7 D). In this format we only need to subtract the protein median to force the distribution back onto the main diagonal.

Step 4: Imputation of missing values

Missing values provide a problem common to all large-scale proteomics experiments. Multiple strategies of dealing with them have been developed ranging from ignoring proteins with missing values completely to machine learning based approaches [94]. In our case, there are four direct

measurements for each of the protein and experiment pairs: heavy and light intensities in the forward experiment and heavy and light intensities in the reverse experiment. Since the forward and reverse experiments are measured independently, their missing values are also independent. Combining the two experiments we obtain a set of 15 possibilities of missing intensity patterns in our data, these possibilities are listed below:

All four intensities missing – 25,836 (or 24.47%) of intensities measured. This occurs when a protein is not detected in particular pull-down experiment at all (but is detected in a pull-down using some other nucleosome).

Three out of four intensities missing – 68 (0.06%). This is a fairly rare occurrence similar to the one above, where a protein is only detected in all but one intensities for a given experiment.

Only one ratio present – 9,619 (9.11%). Protein detected only in forward or reverse experiment, which is sufficient to estimate one ratio, but not both.

Only forward or reverse intensities observed – 37 (0.04%). Same as above, intensities for only one of the experiments observed, however without sufficient confidence to estimate a ratio.

Forward light and reverse heavy intensities missing – 6 (0.01%). Protein detected only in the presence of a modified nucleosome, but not detectable in the presence of unmodified one.

Both heavy or both light intensities are missing – 3 (0.00%) Protein not detected in one of the labelled samples.

From the numbers above it is clear that a substantial portion of the data contains missing values. In fact 68.28% of detected proteins have at least one value missing. In the 9.11% of cases where only one ratio is present, we can infer the other one using the information obtained in the previous post-processing steps. Given what we know about the nature of our data, the most-likely position for this ratio will occur on the main diagonal itself. We can estimate this

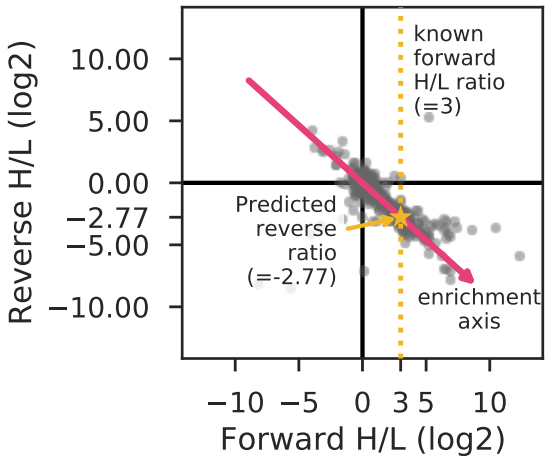

Figure 2.8: Ratio projection imputation of missing data in SNAP datasets. In cases where only one ratio could be estimated by the mass spectrometer, the missing ratio is imputed by projection onto the main diagonal estimated from other protein ratios in this pull-down.

by drawing a straight line at the known ratio and see where it intersects with the estimated diagonal, as illustrated in Figure 2.8, providing an estimate for our imputed ratio.

Similarly, in 6 cases where protein is detected with the modified nucleosome, but not the unmodified one, the ratios, and therefore enrichment values are infinite. In these six cases we impute the missing values to have the maximum enrichment observed for the associated protein across all pull-downs, and median residual for this protein (which is now necessarily zero, as described in step 3). Data imputation is not performed for any of the other missing data cases, and they are all assumed to be zero.

Step 5: Rotating the data scatter onto the ideal diagonal

While the enrichment decomposition described above provides an elegant way to disentangle the experimental effects and their biases, we found the ratio-based framework to be more intuitive to visualise and understand. Working with enrichment/residual dimensions directly also provides a statistical problem since the two axes are on very different scales (with residual dimension being much smaller). We therefore re-encode the decomposed dimensions as ratios for future analyses.

This can be done via a simple rotation operation. In order to make the forward/reverse ratios

comparable across the experiments we rotate the enrichment decomposition axes to be on the ideal, 45 degree diagonal. As described in step 1, original data points do not usually lie on this ideal diagonal, this step has the effect of correcting for this bias, a sketch of this can be seen in Figure 2.5. After this rotation the estimated ratios have intuitive meaning, become comparable across pull-downs and therefore become easy to visualise. On the other hand the values are still strongly related to the decomposed terms and therefore have their biases corrected.

2.2.3 Network inference

The scale of this SNAP dataset allows the recovery of the regulatory network of chromatin across the sampled nucleosomes. Which in turn allows us to not only study the relationship between the epigenetic readers and the chromatin, but also the relationships between the chromatin readers themselves.

In this section we describe the method for the recovery of this regulatory network, including the profiling and improvement of the network inference algorithms, as well as establishment of the high-confidence set of novel protein interactions from their results (section 2.2.3).

Selecting a network inference algorithm

The rise of high throughput technologies for RNA-seq has fuelled an increase of interest in the systems biology based approaches to study gene regulation. A number of gene network inference methods have been proposed since. These algorithms can be split into two categories: algorithms inferring networks containing directed edges that indicate causation and indirected ones, which only indicate relationship. Since our data contains no time-based link between different states of the variable, we cannot tell what protein interactions occur before the others, and thus cannot infer the causality of such relationships. Indirected network inference algorithms assume the network structure to be an undirected graph. Nodes on such graphs represent proteins in our case (and genes/transcripts in RNA-seq applications), while the presence of an edge between two nodes correspond to an interaction between them.

It is more convenient mathematically to work with the network in the adjacency matrix format. In this format the whole network is defined via one symmetric matrix $\mathbf{A}$ whose entry $\mathbf{A_{i,j}}$ describes the edge between i and j. In the simplest case $\mathbf{A_{i,j}} = \mathbf{A_{j,i}} = 1$ would represent the edge between i and j, while a value of zero would indicate that these nodes are not connected. Network inference algorithms adopt a slightly less black-and-white approach where the entries $\mathbf{A_{i,j}}$ are real valued and correspond to the *likelihood* that i and j are connected by an edge. The exact meaning of this likelihood depends on the statistic used in the network algorithm, but it usually is a ranking (i.e. not an absolute measurement) where node pairs with higher scores are more likely to interact than lower ones. How this likelihood, or score, is calculated defines the network inference algorithm.

Most inference algorithms work in a two-step procedure consisting of a local and global step. The local step estimates the initial score for each protein pair individually using one metric or another. The global step then re-evaluates the estimated scores in combination to reduce the bias in the metric.

The simplest method, for instance, computes the correlation between parameters in the local step, and skips the global step. This estimation can be performed either naively or using a shrinkage operator, such as the Ledoit-Wolf[95] method. The direct result of this is often a network where transient effects, such as correlation between the nodes A and C through mutual interaction with the node B, are as strong as the direct ones. This particular problem comes from the inability of the local step to distinguish any such effects.

The global step of inference therefore attempts to remove this bias. Conversion of correlations to partial correlations, that is – residual correlations left after explaining away the other variables, is a direct solution to the transient interaction problem illustrated by the example in the previous paragraph. While extremely powerful in $n > p$ cases[81], the models pose significant problems in scenarios with few data points but many variables (which most network inference problems are)[96, 97].

Other algorithms approach the problem from a different angle. For instance, the Weighted Correlation Network Analysis (WGCNA) method instead looks into the properties of the estimated

graph directly, to make its topological properties reflect those seen in biological networks[98]. Context-Likelihood of Relatedness (CLR) attempts to model the background distribution of interaction scores for each gene, and then re-weighs the scores based on the background model[99, 100]. The Algorithm for the Reconstruction of Gene Regulatory Networks (ARACNE) uses an information theoretic result to assign a most probable set of edges between a group of nodes[101].

Both of the latter two methods, CLR and ARACNE, employ a Mutual Information (MI) based metric instead of correlation. MI is a metric which computes the amount of information one random variable provides about the other[99]:

$$MI(X,Y) = \sum_{i,j} P(x_i, y_j) \log \left(\frac{P(x_i, y_j)}{P(x_i)P(y_j)} \right) \tag{2.5}$$

In this expression, X and Y represent two proteins whose observed forward H/L ratios across multiple experiments are treated as random values. Since these ratios are real-valued, the sum should be replaced with an integral. In practice, however, it is more convenient to define a set of intervals in which a given value may fall into and treat those bins as discrete variables. Robust estimators of mutual information in discrete case are reviewed in [102].

In our application this is equivalent to measuring how much we can tell about the behaviour partner of a protein, given that we know the exact behaviour of another. Mutual Information is a metric closely related to entropy (H) and conditional entropy ($H(\cdot|\cdot)$), in fact the following properties hold:

$$MI(X,Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) \tag{2.6}$$

Equation 2.6 expresses the explainability formulation of MI explicitly. If knowing the profile of X told us about the binding profile for Y, the conditional entropy term $H(Y|X)$ would be small and therefore the mutual information would be large. In either of these formulations, the MI criterion does not assume anything about the nature of the relationship between random variables as it only measures the change in entropy. This allows it to capture non-linear

relationships. Anti-correlated relationships can also be captured by the MI metric, but, in contrast to conventional correlation, it provides no method to distinguish them from the positively correlated ones.

It has recently been proposed that MI can be directly substituted with Distance Correlation (DC) in the local step of several MI-based algorithms[103]. This observation only aggravates an already existing problem of the sheer number network inference algorithms to choose from. DREAM2 challenge has previously attempted to provide an unbiased evaluation of multiple algorithms[104]. It is very likely, however, that the algorithms tested would perform very differently on our data from proteomics datasets than they perform in this benchmark or any other other microarray/RNA-seq derived dataset. We have therefore opted to benchmark a subset of available algorithms on our data ourselves.

To benchmark the algorithms we rephrase the network inference problem as an unsupervised classification problem. Given a matrix of protein profiles across the nucleosome pull-downs, for each pair of the proteins in the dataset we aim to assign a binary label corresponding to interaction or no interaction via the algorithm. We then attempt to test these predicted interactions across a validated reference dataset.

We take the set of human interactions in the BioGRID database[87] to be this reference. BioGRID was chosen over alternatives such as STRING[105] due to its frequent updates and policy to contain only experimentally verified interactions. To represent our dataset fairly, the BioGRID interactions database has been pre-filtered to not contain any predicted interactions with histone proteins or self-loops (i.e. protein interactions with themselves).

In a network inference setting the classes to be predicted are extremely uneven: there are substantially fewer true interactions than false ones. In scenarios like these Receiver Operating Characteristic (ROC) is not an appropriate statistic for model comparison, and Precision-Recall Curve (PRC) should be used instead[106]. Specifically, in such biased setting, the PRC curves are more interpretable as the classification problem is treated as a ranking one[106]. The question that is asked is not what is the performance of the algorithm across the whole dataset, but what is the performance of the algorithm within the first 10, 100, 1000, etc. best-

scoring predictions. PRC provides a direct answer to each of these questions by plotting the two quantities below against each other for all possible thresholds:

$$\text{recall} = \frac{TP}{TP + FN} \tag{2.7}$$

$$\text{precision} = \frac{TP}{TP + FP} \tag{2.8}$$

where TP is the number of true positives, FP is the number of false positives and FN is the number of false negatives.

Unlike ROC, PRC curves cannot be interpolated linearly[106]. This requires special care of the estimation of area under this curve as it is often implemented incorrectly. For instance, at the time of writing the implementation of area under Precision-Recall Curve (auPRC) metric in scikit-learn[107] – a popular machine learning library in Python – is implemented incorrectly[1]. Until this issue is resolved, we suggest the R package precrec[108] instead, which is what we also use in this project. In another paper[109], Flach and Kull list several other drawbacks of PRC curves, including the lack of defined baseline and the fact that area under is a heuristic that is not truly interpretable. The authors propose a Precision-Recall Gain metric as a solution, which they define as:

$$\text{precG} = \frac{\text{precision} - \pi}{(1 - \pi) \times \text{precision}} = 1 - \frac{\pi}{1 - \pi}\frac{FP}{TP} \tag{2.9}$$

$$\text{recG} = \frac{\text{recall} - \pi}{(1 - \pi) \times \text{recall}} = 1 - \frac{\pi}{1 - \pi}\frac{FN}{TN} \tag{2.10}$$

where π is the proportion of the entries marked as the positive class. We have initially considered this metric, but as $\pi \to 0$, the term $\frac{\pi}{1-\pi}$ also $\to 0$ meaning both precG $\to 1$ and recG $\to 1$ regardless of the false positive/false negative rates. Intuitively, this provides an implementation

[1]See discussion in `https://github.com/scikit-learn/scikit-learn/issues/4577`

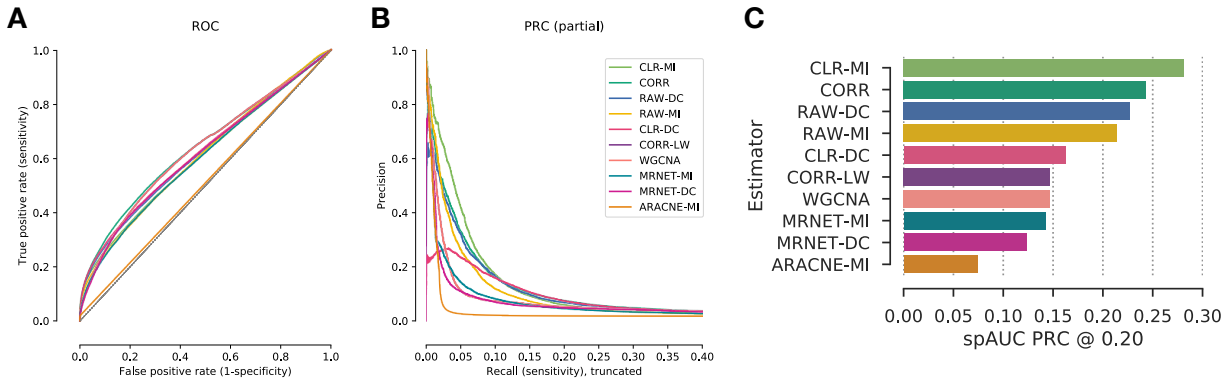

Figure 2.9: Performance statistics of network inference algorithms. (A) ROC, (B) PRC for the estimated networks. (C) Barplot of spAUC at 0.2 recall for the network algorithms, which has been used to select the best algorithm. Key: *CLR-MI*, *CLR-DC*: Context-Likelihood of Relatedness (CLR) algorithm using mutual information and distance correlation metrics respectively. *CORR*, *CORR-LW*: Network estimated using correlation as adjacency metric without a global step, as well as network estimated with correlation estimated using Ledoit-Wolf shrinkage criterion. *RAW-MI*, *RAW-DC*: networks estimated using mutual information and decision correlation metrics without a global step. *WGCNA*: network estimated using WGCNA algorithm. *MRNET-MI*, *MRNET-DC*: networks estimated using ARACNE algorithm with mutual information and distance correlation as inputs. *ARACNE-MI*: networks estimated using ARACNE algorithm with mutual information as input.

challenge due to the limited amount of precision of floating point numbers. At the same time, the more the classes are imbalanced (the smaller the π), the less weight do the performance metrics hold in this setting. Arguably, this is a property that is not desired.

In our experiments we therefore use PRC as the evaluation statistic. To be more precise, we compute the area under the partial segment of this curve for the precision range between 0 and 0.20, and rescale it to be between 0 and 1 (as if the data points outside of 0.20 recall threshold were never in the dataset to begin with), which is known as standardised partial Area Under Curve (spAUC) method[108]. We chose a partial segment of the curve as we do not expect to reconstruct the full BioGRID network from our limited dataset but we expect the model to reproduce at least a small subset of this network (i.e. up to 20% as per our criterion) faithfully.

The network has been trained on the binding profiles of all proteins that have at least one non-zero forward and reverse ratio. The results from the benchmarking procedure are shown in Figure 2.9 and listed in Table B.1, the legend is explained in the figure caption. From this data it is clear that the CLR algorithm, which uses Mutual Information (MI) in the local step performs best. We have used the following parameters for all of the mutual information based algorithms

in the MINET package, including CLR: `estimator='mi.shrink'`, `disc='equalwidth'` as well as the `nbins` parameter set to the square root of the number of columns in the data. Rather unexpectedly, Pearson correlation in its simplest form has been the second best performer on our data, beating every other inference algorithm we have tested, including CLR with DC metric substituted instead of MI.

Derivation of p-values for Context-Likelihood of Relatedness (CLR) algorithm

The best performer in our benchmark, the CLR algorithm has been first described in [99] and then again reimplemented as part of the MINET package[100]. Like all algorithms in the MINET package, CLR relies on MI scores in the local estimation step. In the global step the algorithm assumes that mutual information scores for false positive interactions of a gene come from a random normal distribution. Since a given protein is expected to interact with only a small subset of other proteins, the complete set of MI values associated with a given protein are expected to provide a good estimate of this underlying noise distribution. Once the noise distribution is estimated, a standard z statistic is computed to test whether the observed MI value is likely to come from this background noise distribution. Note that this explicitly assumes that the background distribution of the MI is normally distributed. The authors of the original manuscript acknowledge that this approximation is not entirely accurate ([99, Supplementary Figure S4]), but claim that attempts to model the distribution more accurately produced substantial computational overhead, yet produced results that were very similar. As we show in this section, the assumption of normality also allows for derivation of the p-values function for the algorithm scores.

Since each edge is associated with two proteins, the CLR algorithm performs the z-test twice, once for each of the proteins in the pair, each associated with their own background distribu-

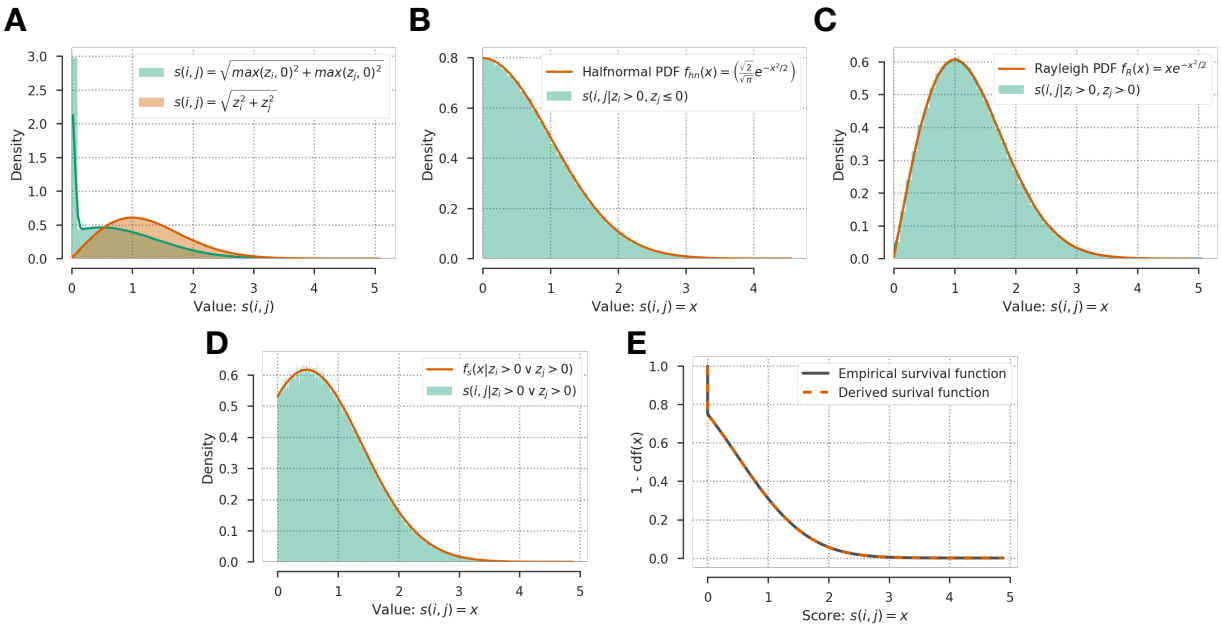

Figure 2.10: Derivation of the p-value statistic for the *pseudo z-score* statistic for CLR. (A) Comparison between the distribution of the score used in CLR and a simpler version of the same function. (B) Halfnormal distribution fit to a subset of randomly generated scores that come from cases where only one z_i and z_j is non-zero. (C) Rayleigh distribution fit to subset of scores where both z_i and z_j is non-zero. (D) The fit of the derived Probability Density Function (PDF) to the full set of randomly generated scores. (E) overlap between empirically estimated survival function and the derived survival function.

tions. The two test statistics, z_i and z_j are then combined using the following function[2]:

$$s(i,j) = \sqrt{max(0,z_i)^2 + max(0,z_j)^2} \tag{2.11}$$

The $s(i,j)$ statistic is sometimes referred to as *pseudo z-score*. Note that this is a fairly unusual way to combine the results of two z tests, unlike Fisher's omnibus test or Stouffer's z-test which have a solid founding in the literature[110]. As far as we are aware, the statistical properties of this score have not yet been explored up until now. Below we derive the p-value for this statistic under the null hypothesis that the two combined z scores, z_i and z_j follow the standard normal distribution $z_i, z_j \sim N(0,1)$.

If there was no *max* operator in Equation 2.11 the PDF of the pseudo-z score would follow the Rayleigh distribution by definition. It is not clear when exactly this operator was introduced,

[2]In [99] this is reported differently, as $s(i,j) = \sqrt{z_i^2 + z_j^2}$, but in the source code of MINET[100] it is implemented as we describe in Equation 2.11. It is not clear where the discrepancy comes from.

as in was not included in the original publication of the algorithm[99]. Regardless, the use of it is beneficial, as it forces the *pseudo z-score* statistic to act as a single-tailed statistic as the significant under-representations are ignored. To illustrate this, we have generated 1,000,000 random pairs of variables from the standard normal distribution (i.e. a million z_i and z_j pairs) and applied the statistical functions to each of them. As seen in Figure 2.10 A, the capping using the maximum operator skews the distribution of expected scores to the left significantly. The skew is due to the *pseudo z-score* coming from negative z_i or z_j distributions receiving smaller scores as their individual contributions are removed.

The *pseudo z-scores* are therefore truly coming from a mixture of three different distributions:

1. Point mass on zero when both $z_i \leq 0$ and $z_j \leq 0$;
2. Half-normal distribution when one of z_i and z_j are greater than zero;
3. Rayleigh distribution when both $z_i > 0$ and $z_j > 0$.

Which we explain below.

Note that due to the symmetry of the standard normal distribution $P(z_i \leq 0) = \frac{1}{2}$. Assuming that z_i and z_j were independent, $P(z_i \leq 0, z_j \leq 0) = \frac{1}{4}$ the three options above have the mixing proportions of $\frac{1}{4}$, $\frac{2}{4}$ and $\frac{1}{4}$ respectively.

Case $z_i \leq 0, z_j \leq 0$

It is clear that from Equation 2.11 that necessarily:

$$s(i, j|z_i \leq 0, z_j \leq 0) = 0 \tag{2.12}$$

In this case, we can express the probability density function f_s as:

$$f_s(x|z_i \leq 0, z_j \leq 0) = \delta(x = 0) \tag{2.13}$$

where δ is the Dirac delta function.

Case $z_i > 0$, $z_j \leq 0$ (and the equivalent $z_i \leq 0$, $z_j > 0$)

In the second case, where only one of the z scores is non-zero, the score function collapses to:

$$s(i, j|z_i > 0, z_j \leq 0) = \sqrt{z_i^2 + 0} = |z_i| \tag{2.14}$$

We used z_i here as an example, but the equation is analogous for z_j. Since z_i is a normally distributed variable, $|z|$ follows the half-normal distribution by definition. This distribution has the following PDF f_{hn}:

$$p(|z_i| = x) = f_{hn}(x) = \frac{\sqrt{2}}{\sigma\sqrt{\pi}} e^{-x^2/(2\sigma^2)} = \frac{\sqrt{2}}{\sqrt{\pi}} e^{-x^2/2} \quad \text{where } x > 0 \tag{2.15}$$

where the last simplification comes from the fact that $\sigma = 1$ for standard normal variables. Combining the independence assumption between z_i and z_j and Bayes' theorem we can show that conditioning on the sign of z_i has no effect on the distribution:

$$f_{hn}(x|z_i > 0, z_j \leq 0) = \frac{P(z_i > 0, z_j \leq 0 \mid |z_i| = x) f_{hn}(x)}{P(z_i > 0, z_j \leq 0)} \tag{2.16}$$

$$= \frac{P(z_j \leq 0 \mid |z_i| = x) P(z_i > 0 \mid |z_i| = x) f_{hn}(x)}{P(z_i > 0, z_j \leq 0)} \tag{2.17}$$

$$= \frac{\frac{1}{2} \times \frac{1}{2} \times f_{hn}(x)}{\frac{1}{4}} \tag{2.18}$$

$$= f_{hn}(x) = \frac{\sqrt{2}}{\sqrt{\pi}} e^{-x^2/2} \tag{2.19}$$

In Equation 2.18 we have used the symmetry of absolute value operator to infer that $P(z_i > 0 \mid |z_i| = x) = \frac{1}{2}$. Using the result above we can write down the density function for this component:

$$f_s(x|z_i > 0, z_j \leq 0) = f_s(x|z_i \leq 0, z_j > 0) = f_{hn}(x) = \frac{\sqrt{2}}{\sqrt{\pi}} e^{-x^2/2} \quad \text{when } x > 0 \tag{2.20}$$

The fit of this distribution to our simulated data is shown in Figure 2.10 B.

Case $z_i > 0$, $z_j > 0$

In the final case, $z_i > 0$ and $z_j > 0$. The score function then becomes the norm of two standard normal variables:

$$s(i,j|z_i > 0, z_j > 0) = \sqrt{z_i^2 + z_j^2} \tag{2.21}$$

This is a common formulation for Rayleigh random variable with the following PDF:

$$f_R(x) = \frac{x}{\sigma^2} e^{-x^2/(2\sigma^2)} = x e^{-x^2/2} \tag{2.22}$$

Using similar reasoning as in Equations 2.16 to 2.19, we can also show that the conditioning on the sign of z_i and z_j does not change the underlying distribution of the random variable:

$$f_R\left(x|z_i > 0, z_j > 0\right) = \frac{P\left(z_i > 0, z_j > 0|\sqrt{z_i^2 + z_j^2} = x\right) f_R\left(x\right)}{P\left(z_i > 0, z_j > 0\right)} \tag{2.23}$$

$$= \frac{\frac{1}{4} \times f_R\left(x\right)}{\frac{1}{4}} \tag{2.24}$$

$$= f_R\left(x\right) \tag{2.25}$$

The relationship $P\left(z_i > 0, z_j > 0|\sqrt{z_i^2 + z_j^2} = x\right) = \frac{1}{4}$ that was substituted in Equation 2.24 becomes intuitive when the score function is interpreted as a distance metric. This way every value of x is generated from an unit circle of z_i and z_j variable pairs, of which $z_i > 0, z_j > 0$ spans exactly a quarter of. This allows us to write down the density of the final component:

$$f_s(x|z_i > 0, z_j > 0) = f_R(x) = x e^{-x^2/2} \quad \text{when } x > 0 \tag{2.26}$$

The fit of this density function is illustrated in Figure 2.10 C.

PDF of the whole mixture

To obtain the PDF of the whole mixture, we simply multiply the above derived conditional

PDF by their mixing coefficients:

$$\begin{aligned} f_s(x) = P(z_i \leq 0, z_j \leq 0) \times f_s\left(x | z_i \leq 0, z_j \leq 0\right) \\ + P(z_i > 0, z_j \leq 0) \times f_s\left(x | z_i > 0, z_j \leq 0\right) \\ + P(z_i \leq 0, z_j > 0) \times f_s\left(x | z_i \leq 0, z_j > 0\right) \\ + P(z_i > 0, z_j > 0) \times f_s\left(x | z_i > 0, z_j > 0\right) \end{aligned} \tag{2.27}$$

i.e.

$$f_s(x) = f(x) = \underbrace{\frac{1}{4}}_{\text{both}<0} \times \underbrace{\delta(x)}_{\text{Point mass}} + \underbrace{\frac{2}{4}}_{\text{one}>0} \times \underbrace{\frac{\sqrt{2}}{\sqrt{\pi}} e^{-x^2/2}}_{\text{Half-normal}} + \underbrace{\frac{1}{4}}_{\text{both}>0} \times \underbrace{x e^{-x^2/2}}_{\text{Rayleigh}} \tag{2.28}$$

Survival function

To compute the P values of this function we need to integrate it for all values as extreme or more extreme than the observed, i.e. to estimate

$$P(s(i,j) \geq x) = \int_x^{\infty} f_s(x) dx \tag{2.29}$$

Necessarily (as the scores are non-negative):

$$P(s(i,j) \geq 0) = 1 \tag{2.30}$$

For any $x > 0$, one of the z_i or z_j must necessarily be non-zero:

$$x = \sqrt{max(0, z_i)^2 + max(0, z_j)^2} > 0 \iff z_i > 0 \vee z_j > 0 \tag{2.31}$$

The above result shows that for any $x > 0$, the Dirac delta component does not contribute anything to the probability density, and thus only the function $g_s(x)$ has to be considered in

the integral:

$$g_s(x) = f_s(x) - \frac{1}{4}\delta(x) = \frac{2}{4} \times \frac{\sqrt{2}}{\sqrt{\pi}} e^{-x^2/2} + \frac{1}{4} \times x e^{-x^2/2} \tag{2.32}$$

$$P(s(i,j) \geq x > 0) = \int_x^\infty g_s(x)\ dx = \frac{1}{4}\left(2 \times \text{erfc}\left(\frac{x}{\sqrt{2}}\right) + e^{-x^2/2}\right) \tag{2.33}$$

where erfc is the complementary error function $\text{erfc}(x) = 1 - \frac{1}{\sqrt{\pi}} \int_{-x}^{x} e^{-t^2} dt$.

Therefore, the final expression for the p-values of CLR algorithm can therefore be written as

$$\text{p-value} = P(s(i,j) \geq x) = \begin{cases} \frac{1}{4}\left(2 \times \text{erfc}\left(\frac{x}{\sqrt{2}}\right) + e^{-x^2/2}\right), & \text{if } x > 0 \\ 1 & \text{if } x = 0 \end{cases} \tag{2.34}$$

The function $g_s(x)$ itself is an un-normalised conditional probability density, given $x > 0$:

$$f_s(x|x > 0) \propto g_s(x) \tag{2.35}$$

The necessary normalisation coefficient γ can therefore be obtained trivially:

$$\int_0^\infty f_s(x|x > 0)dx = \int_0^\infty \gamma g_s(x)dx = 1 \iff \gamma = \frac{4}{3} \tag{2.36}$$

The conditional probability density is therefore:

$$f_s(x|x > 0) = \frac{f_s(x)}{P(x > 0)} = \frac{4g_s(x)}{3} \tag{2.37}$$

$$= \frac{2}{3} \times \frac{\sqrt{2}}{\sqrt{\pi}} e^{-x^2/2} \times \frac{1}{3} \times x e^{-x^2/2} \tag{2.38}$$

The fit of this function is shown empirically in Figure 2.10 D. Similarly, the Figure 2.10 E shows the fit of the survival function to the empirically estimated one.

Now that we can estimate the p-values associated with the *pseudo z-score* statistics, such values

q-value	CLR score	Network size
$1 \times 10^{-}5$	5.91	1,478 edges
$1 \times 10^{-}4$	5.43	2,274 edges
0.001	4.88	4,005 edges
0.01	4.22	8,579 edges
0.05	3.64	16,989 edges
0.1	3.34	24,696 edges

Table 2.1: Conversion table between the estimated CLR scores and their adjusted p-values for our data, as described in section 2.2.3

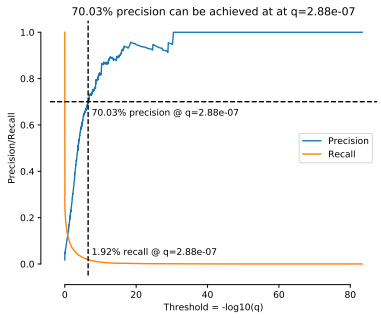

Figure 2.11: Determination of *high-confidence* threshold for detection of novel protein interactions. Threshold has been selected as the lowest threshold at which the precision is at least 70%.

can be used to estimate a sensible threshold for network edges. These estimated p-values, while directly related to the estimated scores, not only have a direct benefit of interpretability, but also allow us to control for the dataset size. Specifically, application of multiple hypothesis testing on the whole dataset (Benjamini/Hochberg) converts the estimated p-values into false discovery rate adjusted p-values (i.e. q-values), which are scaled by the data size. We use the estimated q-values in subsequent sections directly.

Determining a high-confidence interaction set

One of the major goals of this project has always been the identification of novel protein interactions, novel subunits of protein complexes, or even novel protein complexes themselves. This task is directly answerable by comparing the inferred network with the BioGRID one.

Since a novel protein interaction is a substantial claim that requires substantial data, we only consider the predictions with the strongest confidence for further analysis. In order to establish this *high-confidence* set of predicted interactions we again make use of the Precision-Recall

Curve (PRC) curve. As evident in Figure 2.9, the precision of all classifiers (i.e the number of predicted edges that are known versus total number of predicted edges) sharply drops as the recall increases. The latter statistic being just a proxy for the stringency of the threshold, we choose our high-confidence threshold to be the lowest threshold at which the precision is no lower than 70% (Figure 2.11). This constraint selects a threshold with q-value approximately equal to 2.88×10^{-7}.

The 30% of the remaining predictions which were not in BioGRID were then manually inspected and evaluated by conducting a literature review. Results of which are presented in section 2.3.4.

Visualisation & filtering

The prediction network has been estimated for all of the proteins in the data set, so the relationships between proteins can directly be retrieved in the interactive format. For the purposes of visualisation and further analyses, however, we only consider the proteins with at least five non-zero ratios observed in the SNAP dataset. This serves to filter a set of proteins who may get a high interaction score by chance, but in the end are not of particular interest for downstream analyses. Similarly, histone proteins have been removed from the further analyses as their binding profiles are misleading (see Figure 2.7).

The network visualisations in this thesis have been achieved with the help of two software programs. Large, connected networks were visualised via the ForceAtlas algorithm in Gephi. Small or disjoint networks were instead visualised using Cytoscape.

2.2.4 Factor analysis

In the wake of the 20th century C. Spearman was contemplating why some individuals perform better in intellectual tasks than others. He believed each person to be associated with a value, a score, an energy, which he called *general intelligence*, that would solely determine one's intellectuality. In his seminal paper[111], Spearman proposed a model to estimate this

coefficient from a set of cleverly designed tests, which became the first conceptual use of factor analysis as an exploratory technique.

Since Spearman, the idea of summarising complex behaviour via a set of hidden factors (i.e. Factor Analysis (FA)) has spread through the field of psychology. This idea has been proven to be so mesmerising that overreaching conclusions from the analysis were routine. Multiple high-profile theories, including the existence of Spearman's general intelligence factor, have since been questioned[112, 113], but the general approach still remains valuable.

Evidence exists, that nucleosome language is also guided by a small number of factors. Specifically, it has been reported that the majority of variance of the co-occurence patterns of histone PTMs can be explained by a small set of principal components[26]. We wonder if the same observation could be made with respect to the chromatin reader response to nucleosomal combinations, and, if it could, whether the estimated factors could be characterised.

Mechanistically, Factor Analysis (FA) and PCA solve a similar problem. Given that our goal is not only to infer the number of constituents of the histone code, but to characterise them as well, we choose the former method as it provides more flexible tools for data interpretation.

Nevertheless, the added flexibility provides increased concern about the ability to misuse the model. In this section we therefore guide the reader not only through the implementation of the method, but also through both the formal and the intuitive definitions of the model, the caveats associated, and our proposed solutions to these caveats.

Formal definition

The modern definition for the FA model can be found in any recent textbook, for instance [114–116]. We reproduce it here using Barber's notation[116].

Given a dataset of N observable data points, $\mathcal{V} = \left\{\mathbf{v}^1, \ldots, \mathbf{v}^N\right\}$, each of the dimension D we aim to find the structure of $H \leq D$ hidden (latent) factors that approximate the data.

Specifically, we assume that a given vector $\mathbf{v}$ is generated from the following model:

$$\mathbf{v} = \sum_{i}^{H} [\mathbf{f_i}\ h_i] + \boldsymbol{\epsilon} \tag{2.39}$$

Where h_i is the hidden value of the i-th hidden factor, $\mathbf{f_i}$ is a D-dimensional vector of the factor loadings for this hidden factor, and the $\boldsymbol{\epsilon}$ is a D-dimensional Gaussian noise term described further below.

In the matrix form, the same equation can be expressed as:

$$\mathbf{v} = \mathbf{F}\mathbf{h} + \boldsymbol{\epsilon} \tag{2.40}$$

Where $\mathbf{F}$ is the $D \times H$ factor loadings matrix, and $\mathbf{h}$ is the vector of hidden coordinates associated with the visible vector $\mathbf{v}$. Note that [116] adds an intercept term to Equation 2.40, but it is dropped here for simplicity, $\mathbf{v}$ are thus assumed to be zero centred. The last component in the equation, $\boldsymbol{\epsilon}$ is the Gaussian noise term with covariance Ψ:

$$\boldsymbol{\epsilon} \sim \mathcal{N}(\epsilon|\mathbf{0}, \boldsymbol{\Psi}) \tag{2.41}$$

The factor analysis model makes an explicit assumption that the covariance of the noise is diagonal:

$$\boldsymbol{\Psi} = \mathrm{diag}(\psi_1, \psi_2, \ldots, \psi_D) \tag{2.42}$$

A special case of this model is probabilistic PCA, which constrains the noise variance to be the same across all dimensions, i.e. $\boldsymbol{\Psi} = \sigma^2\mathbf{I}$. We see that factor analysis is a more flexible model in this regard as dimensions are allowed to have different levels of noise associated with them.

As [116] notes, correctly, the above-described structure of the model ensures that the observed

values, given the hidden ones, are Gaussian distributed:

$$p(\mathbf{v}|\mathbf{h}) = \mathcal{N}(\mathbf{v}|\mathbf{Fh}, \mathbf{\Psi}) \tag{2.43}$$

The prior distribution on hidden variables is assumed to be standard normal:

$$p(\mathbf{h}) = \mathcal{N}(\mathbf{h}|\mathbf{0}, \mathbf{I}) \tag{2.44}$$

Assumption of a normal prior results in the following marginal distribution of visible data points, which is also normal:

$$p(\mathbf{v}) = \mathcal{N}(\mathbf{v}|\mathbf{0}, \mathbf{FF^T} + \mathbf{\Psi}) \tag{2.45}$$

It is clear from the representation above that the model is defined over the covariance matrix of the data ($\mathbf{\Sigma_D} = \mathbf{FF^T} + \mathbf{\Psi}$) and not the data itself.

Barber provides the mathematical derivation for the maximum-likelihood methods of fitting the parameters $\mathbf{F}$ and $\mathbf{\Psi}$[116, Ch. 21.2] of this model, which will not be discussed here as the implementation of these methods is provided in R package psych[117].

Estimating the number of factors

Unlike PCA, where it is often convenient to extract all possible principal components and then selectively remove them based on explained variance, factor analysis models require the number of hidden factors H to be chosen *a priori*.

It is accepted to estimate this number from the eigenvalues obtained from PCA analysis on the same dataset. The number is chosen either by manual inspection of the plot of eigenvalues in decreasing order (i.e. scree plot method), by comparison to eigenvalues generated from

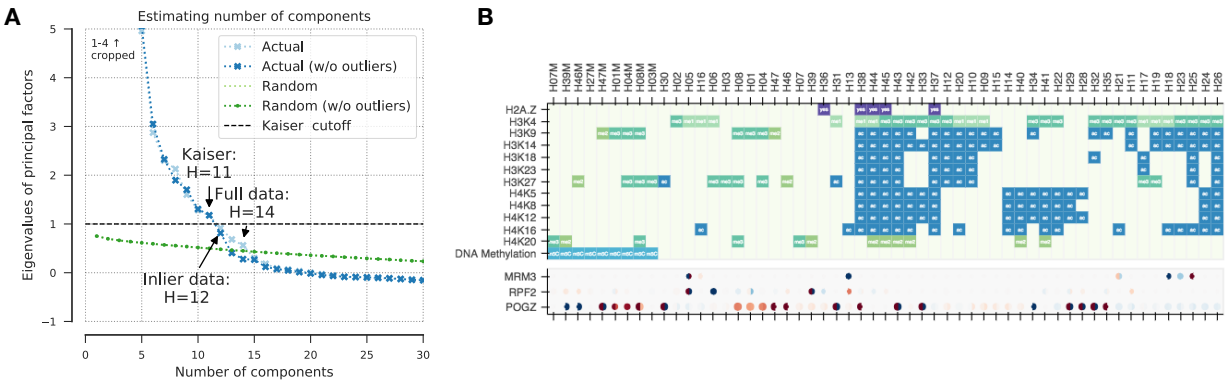

Figure 2.12: Estimation of the number of factors for Factor Analysis (FA) (A) Eigenvalue scree-plot. Kaiser criterion predicts 11 hidden dimensions in the data. Parallel analysis predicts 14 hidden dimensions on the full dataset, but this number drops to 12 after removal of only three outlier proteins. (B) Outlier proteins removed from estimation.

randomly shuffled data (i.e. *parallel analysis* method), or with the help of the Kaiser criterion, which retains the factors as long as their associated eigenvalues are greater than one [118].

Alternatively, the Minimum Average Partials method computes the average partial correlations between variables after removing each of the principal components sequentially. At the correct number of components, the common variation should be removed from the dataset and thus the criterion would reach the minimum[118]. More recently, a Revised Parallel Analysis method[119] and Empirical Kaiser Criterion[120] methods have been proposed as well. Meta studies cite Kaiser criterion as the most commonly used method, as it is often the default option in the software[118, 121]. Simulation studies, however, find parallel and minimum average partial methods more accurate[118, 121]. There is no general solution for this problem, to quote [121] directly: "researchers are advised to use both multiple criteria and reasoned reflection".

In this study we have chosen the number of factors based on multiple criteria described here. The results are summarised in Figure 2.12. Kaiser criterion predicts $H = 11$ hidden dimensions in the data. Parallel analysis predicts 14 hidden components for the full-dataset, but this number drops to 12 once three outlier proteins (RPF2, MRM3, POGZ) have been removed (Figure 2.12 B). These three outlier proteins became apparent from manual inspection of the factor scores (see section 2.2.4) obtained from the model trained with $H = 14$ dimensions. The last dimensions in this model tended to overfit and thus assigned high factor scores to these

three proteins only. Due to the instability of the estimates from parallel analysis to outlier proteins, and the low eigenvalues for the 12th component we have decided to accept the Kaiser estimate of $H = 11$ hidden factors in our data, but still remove the outlier proteins. No obvious over-generalisation to outlier proteins could be detected in this 11 factor model.

Utilising Rotational Invariance

An interesting feature of factor analysis models is their invariance to factor rotation. Recall the final expression of the marginal data likelihood described in Equation 2.45. In this formulation, we can substitute any factor loading matrix $\mathbf{F}$ with a rotated matrix $\mathbf{FR}$ where $\mathbf{R}$ is any rotational matrix (i.e. $\mathbf{RR^T} = \mathbf{I}$) to get back the same model formulation:

$$(\mathbf{FR})(\mathbf{FR})^T + \mathbf{\Psi} = (\mathbf{FR})(\mathbf{R^T F^T}) + \mathbf{\Psi} = \mathbf{F}(\underbrace{\mathbf{RR^T}}_{=\mathbf{I}})\mathbf{F^T} + \mathbf{\Psi} = \mathbf{FF^T} + \mathbf{\Psi} \tag{2.46}$$

Therefore, for every inferred factor loading matrix $\mathbf{F}$, there exists an infinite number of alternative rotated solutions that fit the model as well as the inferred solution does. Such result questions whether the loadings from FA models are interpretable at all since they are, essentially, undefined.

What is important to note, however, is that even though multiple solutions are possible, a single solution is often sufficient. And among multiple solutions it is often convenient to pick the one that is easiest to interpret. Thurstone has provided a set of rules that, when satisfied, produce easy to interpret factor loading matrices $\mathbf{F}$. Multiple algorithms to find rotational matrices $\mathbf{R}$ have since been proposed, aiming to satisfy these, or similar, criteria. A modern review of these methods can be found in [122]. Development of the Gradient Projection Algorithm[123] has allowed an arbitrary criterion to be optimised without the need to develop a bespoke algorithm, which became a de facto standard approach in factor analysis since.

The proposed rotations in the factor analysis field are split into two groups: orthogonal and oblique. Orthogonal rotations, as the name suggests, preserve the orthogonality between factor

loadings. Oblique rotations do not have this requirement and thus produce factors that are correlated. Best practice guides for factor analysis often prefer oblique rotations to orthogonal ones[124, 125], with the assumption that many real-world phenomena are indeed correlated and not orthogonal. Oblique rotations, however, result in solutions that are intricate as the factor loadings matrix is split in two matrices: a matrix of correlations of each factor and visible variables and a standardised partial correlation matrix with other factors conditioned out[124]. We would argue that appropriate interpretation of these two statistics is more complicated than the interpretation of the (single) factor loadings matrix produced by orthogonal rotations. Furthermore, the (standard) PCA methods have already familiarised the modern scientist with projections onto orthogonal axes, which aids the interpretation of the factor scores (section 2.2.4). We therefore consider only orthogonal rotations.

At the time of writing, 7 orthogonal rotations have been implemented in the psych[117] package:

quartimax – one of the first rotation methods, proposed independently by multiple authors, which aims to simplify the factor loadings by maximising a fourth-order function on their values. It is inherently biased towards producing a fairly general first factor, as the optimisation function is simplifying the rows (i.e. the values associated with visible dimensions) of the loading matrix[122, 126].

varimax – conceptually similar to the *quartimax* method, but designed to simplify the columns of the factor loading matrix (i.e. the factors themselves). This reduces the bias towards general factors as compared to the *quartimax* method [122, 126]. It is perhaps the most commonly used orthogonal rotation, often the default one[122].

equamax – rotation that is designed to compromise between row and column complexity minimisation by the *quartimax* and *varimax*[122].

bentlerT – this rotation comes from Bentler's work in determining a scale-free simplicity measure for factor loadings. This measure is based on generalised variance of the loading matrix and is maximal when each factor generates a single variable[127, 128].

geominT – minimises the sum of geometric means of squared factor loadings across the rows (i.e. variable complexity), formulated naively it is minimised when each row of the factor loading matrix has a zero in it, regardless of what other values are. To alleviate this a small positive value is added to the squared factor loadings. Originally designed for oblique rotations, and therefore not necessarily optimal for orthogonal ones.[122]

bifactor – unlike the rotations above, *bifactor* rotation explicitly ignores the structure of the first factor, attempting to achieve the best structure for remaining factors only. In this sense this rotation trades off the interpretability of the first (general) factor for additional freedom to optimise the remaining ones[129].

varimin – a rotation which aims to minimise the objective maximised by *varimax* rotation. Contrary to other rotations described here, it has been designed to produce a complicated factor structure. The author reasons that such representation may be more honest than the traditional *simple* one.[130].

With so many rotations to choose from, as Bartholomew notes, "the user [is left] with a bewildering array of 'solutions' one of which, surely, must show what the analyst desired."[131] In order to distance ourselves from the inherent bias in choosing an appropriate rotation, for this research we have employed a computer-driven approach at selecting the best rotation matrix.

First, we have generated the model using all of the rotations available to us (Figure 2.13 A). Specifically, the loadings from factor rotations were discretised into three bins in the following fashion:

1. loadings that are less than -0.3 were discretised into the negative loading bin,
2. values between -0.3 and 0.3 were assumed to be approximately zero,
3. and values greater than 0.3 were put in the positive bin.

In Figure 2.13 A black squares indicate loadings that are either less than -0.3 or greater than 0.3.

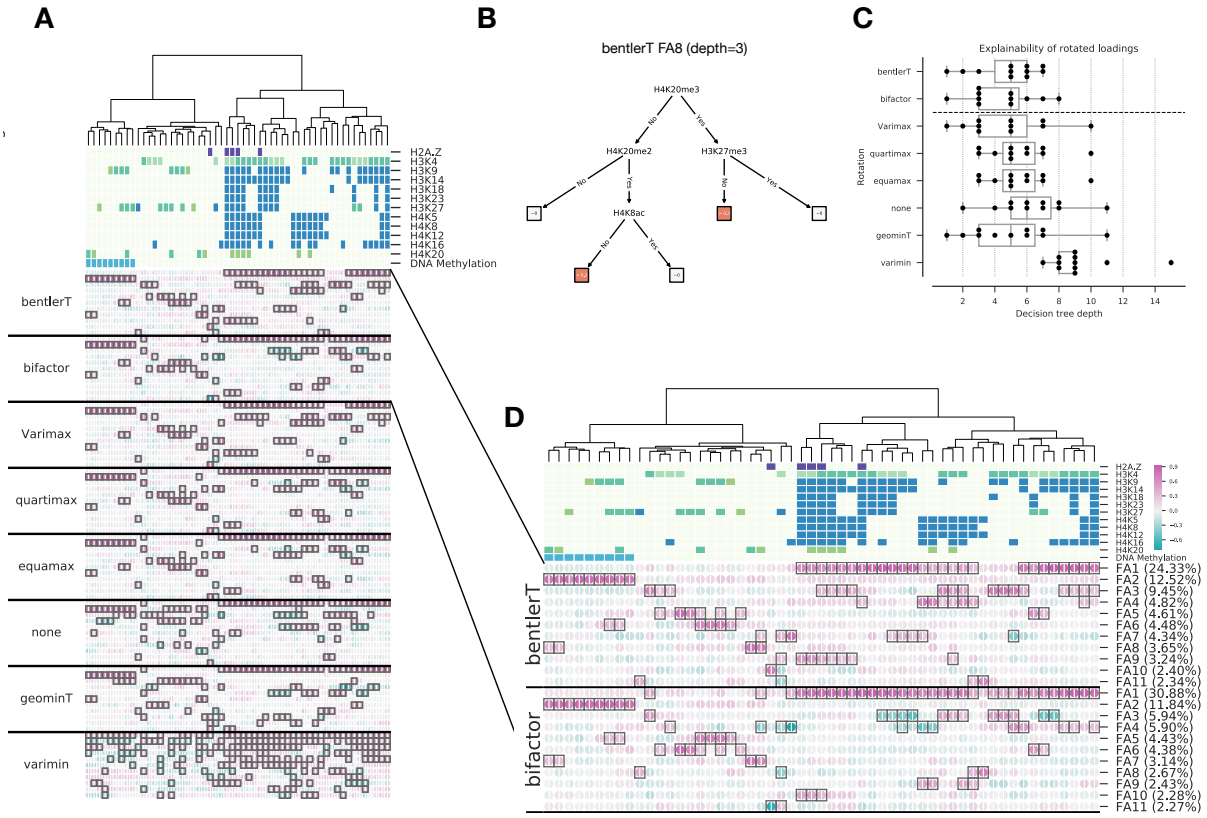

Figure 2.13: Summary of the procedure of choosing the best rotation for the hidden axes. (A) All implemented rotations have been profiled and their loading matrices estimated. Loadings greater than 0.3 (or lower than -0.3) have been deemed important (black squares). (B) Decision tree classifiers have been trained to explain aforementioned important loadings. An example of decision tree for the eight factor in *bentlerT* rotation is shown. (C) The depths of decision trees required to explain the loadings completely have been recorded. Two rotations with lowest maximal depth required have been selected for further analyses. (D) The resulting loadings matrix for the two best rotations identified by the decision tree procedure.

Binarisation of the loadings is common in the FA field as it reduces the complexity. The exact threshold chosen varies by application. Generally, loadings below 0.4 are considered as low and their significance is questioned[125]. For our purposes of identifying the appropriate rotation, the choice of this parameter is fairly arbitrary. We have chosen to err on the side of false positives, rather than false negatives, as we believe this allows for more honest profiling of the explainability of a rotated factor loading, and therefore reduced this threshold to 0.3.

For each of the factors in the loading matrix, we have then trained a set of decision tree classifiers with increasing depth, aiming to predict the aforementioned bins (Figure 2.13 B). Decision tree complexity was artificially controlled by setting the threshold on the maximum depth of the

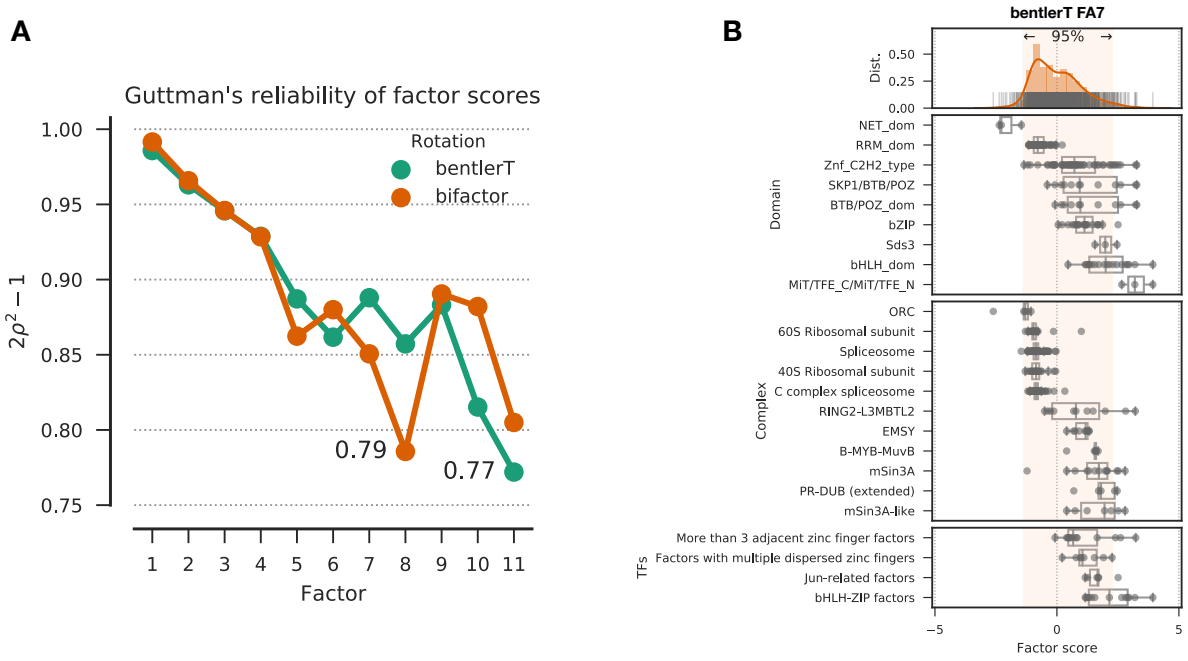

Figure 2.14: Factor score estimation. (A) Guttman's reliability estimates for factor scores for the two best rotations, which describe the minimum correlation between the unknown true factor scores and their estimates. (B) Domains, complexes and transcription factors whose scores are significantly different from the rest for the 7-th factor in the *bentlerT* rotation.

tree. Starting from a trivial tree with only one node, the tree depth has been increased in unit steps until the tree was able to *memorise* the simplified loading matrix. This final depth of the tree was then used to estimate the complexity of resulting factor loading. We found that in some cases decision tree classifiers performed when the output classes were weighted, but in some other cases better performance was achieved with unweighted classification. We therefore considered both settings. The tree sizes required to memorise various rotations are shown in Figure 2.13 C. We have chosen *bentlerT* and *bifactor* rotations for further analyses as they resulted in overall lower maximum tree size than others. The loadings for these two rotations are shown in Figure 2.13 D.

Factor scores – projections onto the hidden axes

It is tempting to think about the model described in the previous sections from a generative point of view. Specifically, a valid request for a researcher is to obtain the values of the hidden variables $\mathbf{h}$ that were used to generate each of the visible data points $\mathbf{v}$, which has the

connections of embedding the variables in a lower-dimensional space. These embeddings are often referred to as *factor scores* in the factor analysis literature.

A well established result in the literature, however, is that such scores are not identifiable[132, 133]. This stems from the fact that the values for hidden components depend on the values for random noise parameters which we do not know, nor can we estimate from the model's parametrisation. Specifically, Guttman has shown that for every factor score estimate, it is possible to construct an adversarial factor score estimate, whose correlation to the original estimate in the worst case (ρ^*) is[132, 134]:

$$\rho^* = 2\rho^2 - 1 \tag{2.47}$$

Where ρ is correlation coefficient between the factor and the original variables. From Equation 2.47 it is clear to see that it becomes possible to obtain two sets of perfectly valid factor scores which are anti-correlated (i.e. $\rho^* \leq 0$) for when $\rho^2 \leq 0.5$.

With this in mind, numerous attempts at approximating the factor scores have been proposed over the last century, a review of which is provided in [132]. One of the first estimators provided is attributed either to Thomson or to Thurstone and has the following form[132, 135]:

$$\hat{\mathbf{h}} = \mathbf{F^T \Sigma_D^{-1} v} \tag{2.48}$$

Where $\mathbf{\Sigma_D}$ is the data covariance matrix, and $\mathbf{F}$ is the factor loading matrix as described before. Alternatively, the estimate can be specified as described below[135]:

$$\hat{\mathbf{h}} = \left(\mathbf{I} + \mathbf{F^T \Psi^{-1} F}\right)^{-1} \mathbf{F^T \Psi^{-1} v} \tag{2.49}$$

While the above expression has been historically derived through regression methods, Equation 2.49 is implicitly computing the expected value of $\mathbf{h}$ given the model parameter estimates

and the observed dataset, and is therefore identical to the E step expressions of the EM algorithm as described in [114, eq. 12.66] and [116, eq. 21.2.33].

As shown in Figure 2.14 A the reliability estimates for these factor scores are 0.77 and 0.79 for *bentlerT* and *bifactor* rotations respectively, which are still acceptable.

To probe the factor scores further we have performed a series of t-tests comparing the mean scores for the members of known protein complexes, proteins containing specific domains and transcription factor families against the remaining data. Only protein complexes, domains and transcription factors containing at least 2 members each were tested. The t-tests have been conducted separately for each group of proteins and factor, the p-values were corrected for multiple hypothesis testing using Benjamini-Hochberg method, with FDR threshold set to 0.01. An example result is shown in Figure 2.14 B where the groups regulated by the 7th axis of the *bentlerT* rotation are shown. Other results are used in the subsequent results sections.

2.2.5 Identification of DNA-binding transcription factor binding sites

The methods described so far have only taken nucleosomal modifications into account. As DNA-binding Transcription Factors (TFs) have proven to be central in a number of analyses (discussed further in section 2.3.3 and section 2.3.6), we have chosen to investigate the effects of the DNA sequence as well.

To identify the locations of possible binding sites for transcription factors on the DNA construct used in the nucleosomal pull-downs (shown in Figure B.6) we have downloaded the core set of transcription factor binding models from the HOCOMOCO (v11) database[136] and scanned them against the provided sequence using FIMO (v5.0.1)[137]. Default settings were used, but the p-value threshold has been decreased to 0.01. We have chosen a less strict p value threshold than the suggested setting of 0.0001 since, due to lack of competition within the short stretch of DNA sequence, the transcription factors are more likely to bind motifs that would perhaps not exhibit such a strong response in *in vivo*.

The transcription factors in HOCOMOCO dataset were then clustered based on their binding

profiles in the SNAP experiment using hierarchical clustering with cosine metric and complete linkage. The factors were divided into nine clusters by cutting the dendrogram at distance of 0.65, and removing any clusters with less than three members. Association between transcription factor families and clusters was tested using hypergeometric test using Benjamini/Hochberg correction for p-values (threshold: 0.01).

2.2.6 Linear modelling of epigenetic reader binding preferences

In section 2.2.4 we have described a method to discern the key regulatory dimensions of this dataset. As we will discover in subsequent section 2.3.5, the results from this method are not always straightforward to interpret. At the same time we are interested in whether the epigenetic reader binding profiles could be summarised in a succinct way (e.g. CBX5 is a reader for H3K9me), and whether such descriptions could be estimated in high-throughput fashion.

In this section we therefore describe a simple and easily interpretable method to summarise the epigenetic protein binding preferences in terms of nucleosomal modifications. Specifically, we assume that the observed binding ratios for the analysed proteins can be explained by a simple linear combination of the nucleosomal modifications present at a given nucleosome:

$$\begin{aligned}\log2(\text{Ratio H/L}) \approx &[\text{H2A.Z}] \times w_{\text{H2A.Z}} + [\text{H3K4me1}] \times w_{\text{H3K4me1}} + [\text{H3K4me3}] \times w_{\text{H3K4me3}} \\ &+ \cdots + [\text{DNA methylation}] \times w_{\text{DNA methylation}}\end{aligned} \tag{2.50}$$

Where the values in square brackets are set to one, when a modification is present, and zero when the modification is absent on the recombinant nucleosome. For instance, this scheme would encode the pull-down H01M (H3K9me3+DNA methylation) as:

$$\begin{aligned}\log2(\text{Ratio H/L}) \approx\ & 0 \times w_{\text{H2A.Z}} + \cdots + 1 \times w_{\text{H3K9me3}} + \cdots + 1 \times w_{\text{DNA methylation}} \\ =\ & w_{\text{H3K9me3}} + w_{\text{DNA methylation}}\end{aligned} \tag{2.51}$$

We refer to the variables w as *coefficients*. Since the scale between all of the nucleosome-specific variables is the same: between zero and one, the coefficients are directly comparable with larger values contributing to the observed ratios more. Since, due to the design of the SNAP experiment we never observe H4K5ac without H4K12ac we merge these predictors into one H4K5ac/H4K12ac as otherwise the design matrix will not be full rank.

We also add a term to *forward_bias* which, as the name suggests, is designed to capture the bias in the forward experiments. The associated variable is set to one, when forward ratios are estimated, and zero otherwise. Along with this, the sign of the reverse ratios has been flipped to match the range of the forward ones.

In order to encourage sparsity of the parameter estimates, we have included a *LASSO* penalty on the parameters (as implemented in scikit-learn). We have not included an intercept term in this model.

The model has been trained on the filtered set of results from the network analysis (see section 2.2.3), predicting the ratios for each protein separately. We have used Leave One Group Out (LOGO) cross-validation, testing predictions on the data from one pull-down (both forward and reverse ratios), while training the model on the rest. The *LASSO* regularisation parameter has been chosen by pooling the out-of-sample predictions across all of the regulatory folds and computing their r^2 score with respect to the true observed values. The best parameter has been chosen by maximising a weighted average of these r^2 scores using the following weights for proteins:

$$\text{weight}(p) = \text{n_significant}(p) + \frac{\text{n_nonzero}(p)}{D} \tag{2.52}$$

Where n_significant(p) computes the number of experiments the observed ratio for protein is

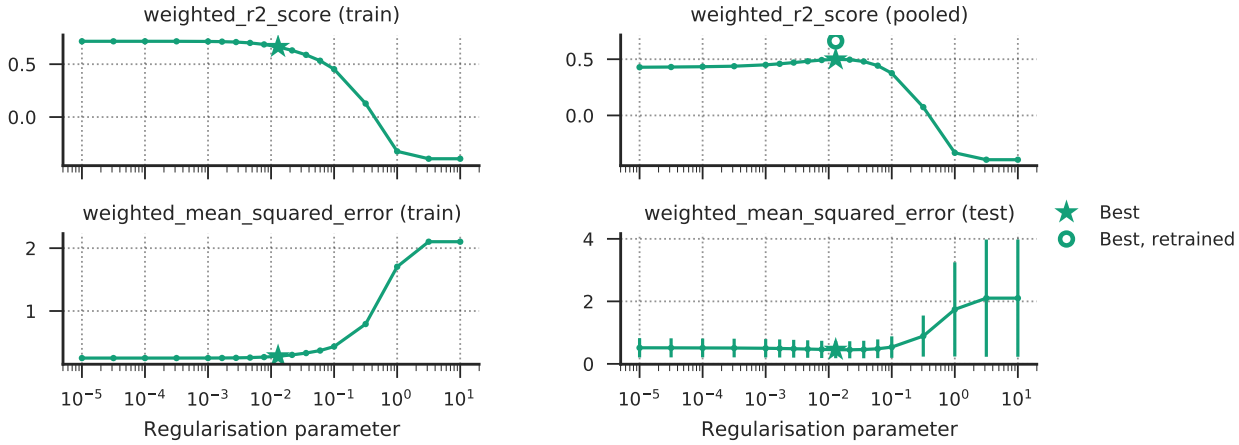

Figure 2.15: Summary statistics for the training of linear model predictions. The best model has been selected by maximising the pooled weighted R^2 statistic from LOGO cross-validation which has repeatedly left out one of the pull-down experiments from the training set (star). The model has then been retrained using the whole dataset and the statistic has been recomputed (circle).

significantly enriched (as described in section 2.2.2), n_nonzero(p) corresponds to number of non-zero data points for a protein, and $D = 110$ (i.e. the total number of data points per protein). The first term in the equation is intended to give more weight for proteins strongly regulated by a multitude of nucleosomes. The second term acts as a form of smoothing to remove the possibility of zero weights.

Figure 2.15 shows the results of the model training procedure. It is clear that the regularisation parameter chosen by the weighted R^2 score statistic is slightly lower than what it would have been had a mean squared error on the test dataset been used instead. This is a more accurate threshold, as the threshold selected by the weighted mean square error tends to over-regularise the models to give up making predictions (this can be seen via the drop in the R^2 score statistics for the training dataset coupled with the increase in the weighted mean square error).

Once the regularisation parameter has been set, we have retrained the model using the complete dataset. In order to select the proteins whose behaviour is approximated well by this model individual R^2 scores between the predicted and observed values have been computed again . The distribution of these scores has been plotted on a histogram Figure 2.16. Note that while the distribution in the left panel of Figure 2.16 appears to be trimodal, indicating the existence of an *intermediate* set of protein prediction accuracy, this effect is reduced upon the removal

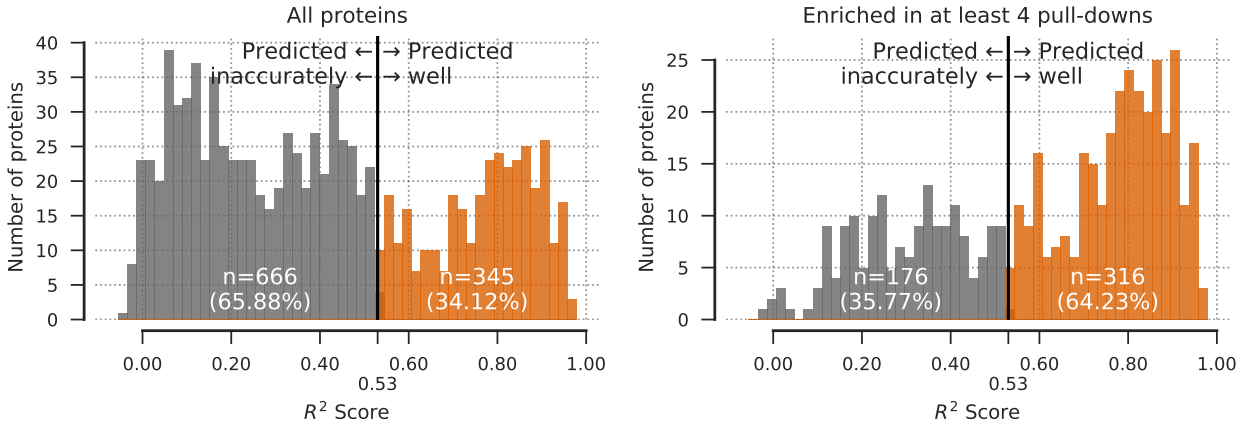

Figure 2.16: Histogram of the distribution of the scores from linear model analysis. The left panel displays model's performance for the whole dataset, whereas the right panel displays only proteins significantly enriched in at least 4 pull-downs. Decision line has been estimated by fitting a mixture of two Gaussian distributions to the estimated scores as described in text.

of proteins enriched in fewer than 4 experiments. In order to simplify the further discussion we have therefore only classified the proteins into two classes only. The decision line between these two classes ($R^2 = 0.53$) has been selected by fitting a two Gaussian mixtures to the data. According to this criterion 34.12% of the complete dataset is predicted accurately (Figure 2.16 left panel). This ratio is further pronounced in Figure 2.16, which removes the proteins that were enriched in fewer than 4 pull-downs. We discuss the results from the 316 proteins that are well predicted in the corresponding results section (section 2.3.7).

2.3 Results

2.3.1 Modified nucleosomes profile major epigenetic processes

In section 2.2.2 we describe a method for detecting the proteins that are significantly enriched on the modified nucleosomes. To investigate what biological pathways these enriched proteins are involved in, as well as how these pathways differ from nucleosome to nucleosome, we have used clusterProfiler[138] to perform GO biological process enrichment analysis on the significantly enriched protein sets. The parameters for selection of the enriched terms were chosen in an *ad-hoc* fashion, relying on the default parameters of clusterProfiler. All human proteins were

Figure 2.17: Gene Ontology (GO) enrichment analysis on the protein sets significantly regulated by the modified nucleosomes. Processes that are regulated can be split into four groups. The majority of nucleosomes regulate DNA and histone modification as well G0 and G1 transition processes. Other, more specialised processes are regulated by a subset of nucleosomes queried and form individual clusters.

used as the background set. For each of the nucleosomes, five enriched GO terms which contain the highest number of cluster proteins were selected. The Benjamini/Hochberg adjusted p values of these categories were visualised in Figure 2.17. In this figure, it can be seen that nucleosomes are split into four groups by the pathways they regulate. The topmost group, containing pathways enriched in most nucleosomes corresponds to processes that play a part in writing or removing modifications on the DNA, histones and chromatin. Cell-cycle regulation, specifically G0 to G1 transition is also enriched in most nucleosomes.

Interestingly, two complementary processes: histone deacetylation and histone acetylation are associated with different nucleosome groups. The former falls in the topmost group, enriched in most nucleosomes, while the latter is not enriched in the H3K9me2/3/H3K27me2/3 and H4K20me3 nucleosomes (second group). Along with histone acetylation, chromatin remodelling and Double-Strand Break (DSB) repair processes are found to be regulated by similar nucleosomes. DNA replication-dependent processes as well as proteins involved in interleukin-7 signalling and rDNA silencing form their own cluster as they are only enriched in a small subset of nucleosomes. Unlike DNA-replication dependent processes, DNA replication initiation processes form another cluster, which is distinct from the previous group. A subset of this cluster also involves DNA repair regulation, however nucleosome regulation for these processes is different.

It is clear that most chromatin processes are captured by our set of nucleosomes. At this point we cannot reason anything about why the pathways enriched differ from nucleosome to nucleosome, but we aim to answer these questions in the subsequent sections.

2.3.2 Network interaction scores follow the expected distribution

In order to be able to study the response of epigenetic proteins to nucleosome modifications, it is first necessary to characterise the proteins that are observed in the data.

In section 2.2.3 we have described a procedure for network inference from the processed dataset. Specifically, we have shown that the CLR algorithm performed best for our data (section 2.2.3).

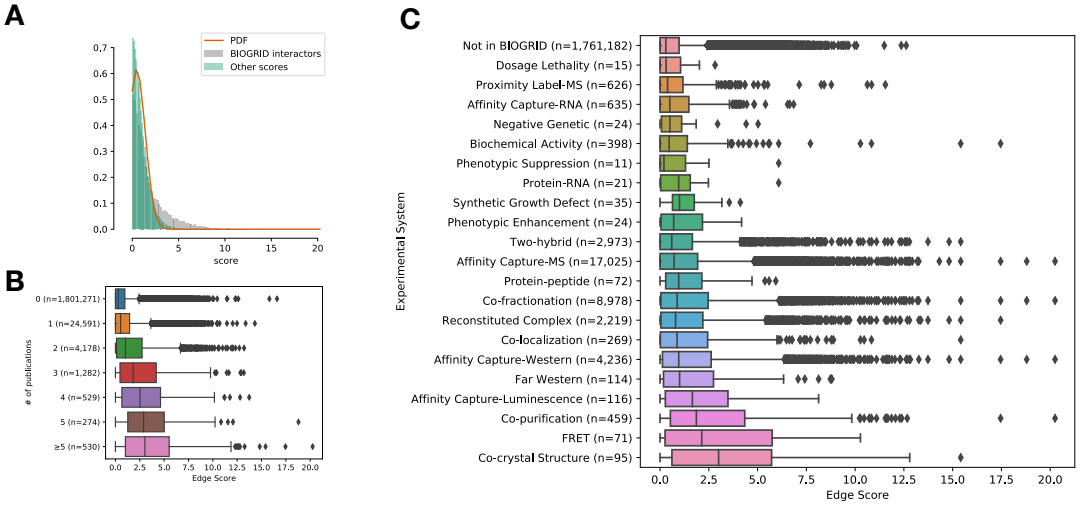

Figure 2.18: Distribution of scores obtained from CLR algorithm. The scores on the x axis are the raw *pseudo-z* scores obtained from CLR algorithm. In order to convert to q-values see Table 2.1. (A) Distribution of the scores obtained, compared with the PDF derived in section 2.2.3. (B) Relationship between predicted interaction score and number of publications associated with interaction. (C) Relationship between predicted interaction score and the experimental method used to establish this interaction.

We have improved this method by deriving the distribution for its performance under the null hypothesis. Figure 2.18 A shows that this derived PDF fits the distribution of scores obtained for the interactions that are not in BioGRID well. The interactions that are in the BioGRID database, on the other hand, have noticeably higher scores on average.

Similarly, Figure 2.18 B shows that interactions that were reproduced in multiple publications have higher scores predicted from our data. Splitting the predicted scores by the experimental method they have been determined with (Figure 2.18 C) shows that interactions that have been shown in co-crystal structure (an undeniable indicator of complex formation) have the highest median predicted score from our model. FRET, co-purification, affinity-capture luminescence, far-western, and affinity-capture western are the next five experimental methods with highest scores. Interestingly, dosage lethality and proximity label-MS experiments are the two worst performers according to the same chart. Both of these methods, however, are stronger indicators of a functional, rather than direct, interaction.

This data is noted here to show that the network estimation method behaves as we would expect it to behave: assigning higher scores to stronger predictions, as well as randomly-distributed

scores for false predictions.

2.3.3 Estimated networks recover major regulatory blocks of chromatin

Drawing the estimated chromatin interaction network requires one to choose a minimal confidence score required for an edge between two nodes A and B to be drawn. The connectivity of the estimated network depends on the threshold that we set for the interaction confidence. We have found that drawing an edge between nodes if their interaction score $q \leq 0.05$ produces a network that is too dense (Figure 2.19 A). Increasing the stringency requirement (and thus requiring lower q values), reduces the number of edges between network nodes until eventually the network becomes completely disconnected. We have found the threshold of $q \leq 0.001$ to provide a good trade-off between sparsity and connectivity for all practical purposes.

Plotting this network, as we have done in Figure 2.19 B, reveals distinct regulatory subunits of the chromatin and their interconnections. In this figure we have manually annotated the locations of protein complexes associated with the chromatin processes based on the labels of the nodes (not shown).

At the top of the figure we see the members of heterochromatic HUSH and PRC1 and PRC2 protein complexes, and their associated subunits. NuRD chromatin remodeler seemingly connects the hetechromatic branch of the network to SWI/SNF chromatin remodellers. The area of the network where the SWI/SNF complex resides is itself connected to an area containing the H2A.Z chaperone SRCAP and the histone acetyltransferase (HAT) NuA4. TFIID and histone acetyltransferases (HATs) which share subunits with this protein complex (e.g. ATAC, SAGA and PCAF) are further connected together. Interestingly, repressive complexes such as mSin3A and EMSY are are strongly connected to these HATs indicating that similar nucleosomes regulate their behaviour.

On the opposite side of the network projection we see the Integrator and Mediator complexes which are associated with transcription processes. COMPASS-like MLL3,4 integrates nearby,

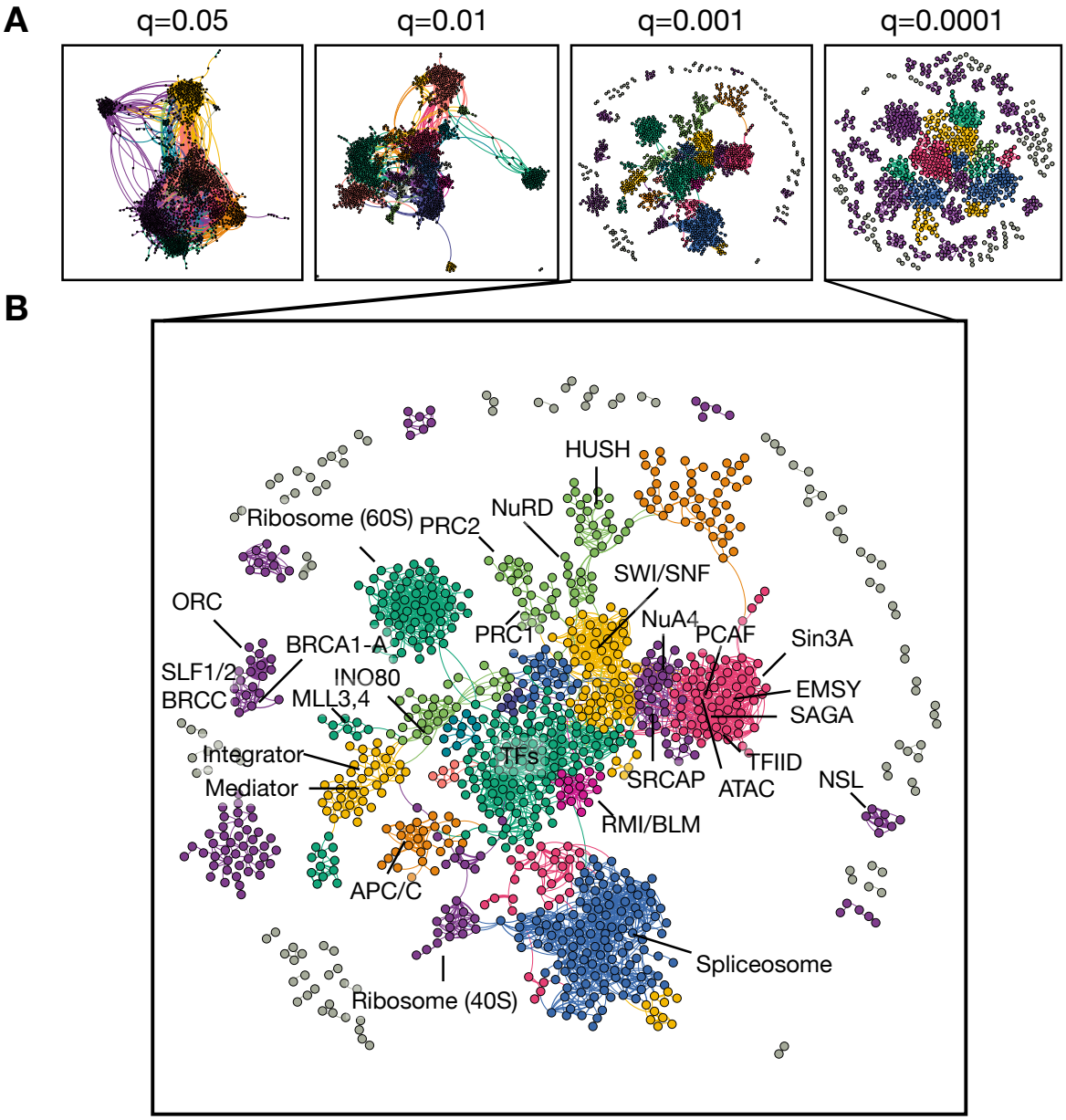

Figure 2.19: Chromatin regulatory network identified from the network analysis. (A) Effect of the network threshold on the network structure. Higher thresholds result in a sparser network, breaking down inter-complex interactions, but keeping the strongest intra-complex ones. (B) Protein complexes recovered by the network annotated onto the $q = 0.001$ network.

and INO80 chromatin remodeler seemingly connects this branch to the main network cluster. At the centre of this main network component lie DNA-binding Transcription Factors (TFs), which constitute a link to the other parts of the network.

We could not fit all of the names of the recovered protein complexes or other regulators into a single figure without making it illegible. It is nevertheless still clear from this annotation that the estimated network recovers the majority of the processes associated with chromatin. This visualisation also provides us with a global overview of the dataset, which we will revisit in section 2.3.5.

2.3.4 High-confidence network predictions identify novel subunits of protein complexes

In the previous sections we have explored the general structure of the inferred network. As already shown in section 2.2.3 and Figure 2.18, the higher network scores are associated with fewer predictions of edges that are not yet known in the literature, and stronger evidence of the ones that are. In this section we therefore explore a network obtained at a high-confidence threshold, at which only 30% of predicted edges are not in the BioGRID database to search for novel interactions between epigenetic proteins.

The complete set of such interactions is shown in Figure 2.20. We have classified the predictions into two categories: known predictions (i.e. the ones that already appear in the BioGRID database, blue edges) and novel ones (red edges). Due to the stringency of the threshold, we necessarily have very few predictions, known or novel ones, what provides a misleading picture when taken on their own. In order to amend this we have also included the interactions in the database that are known, but not predicted by the model at this threshold (grey).

While we use the term "novel" prediction, we would like to stress that the predictions are novel, given our data, and not necessarily novel in the literature. There are a few scenarios which would result in the model marking a known prediction as novel: any protein interaction predictions already identified in model organisms, but not yet confirmed in humans would

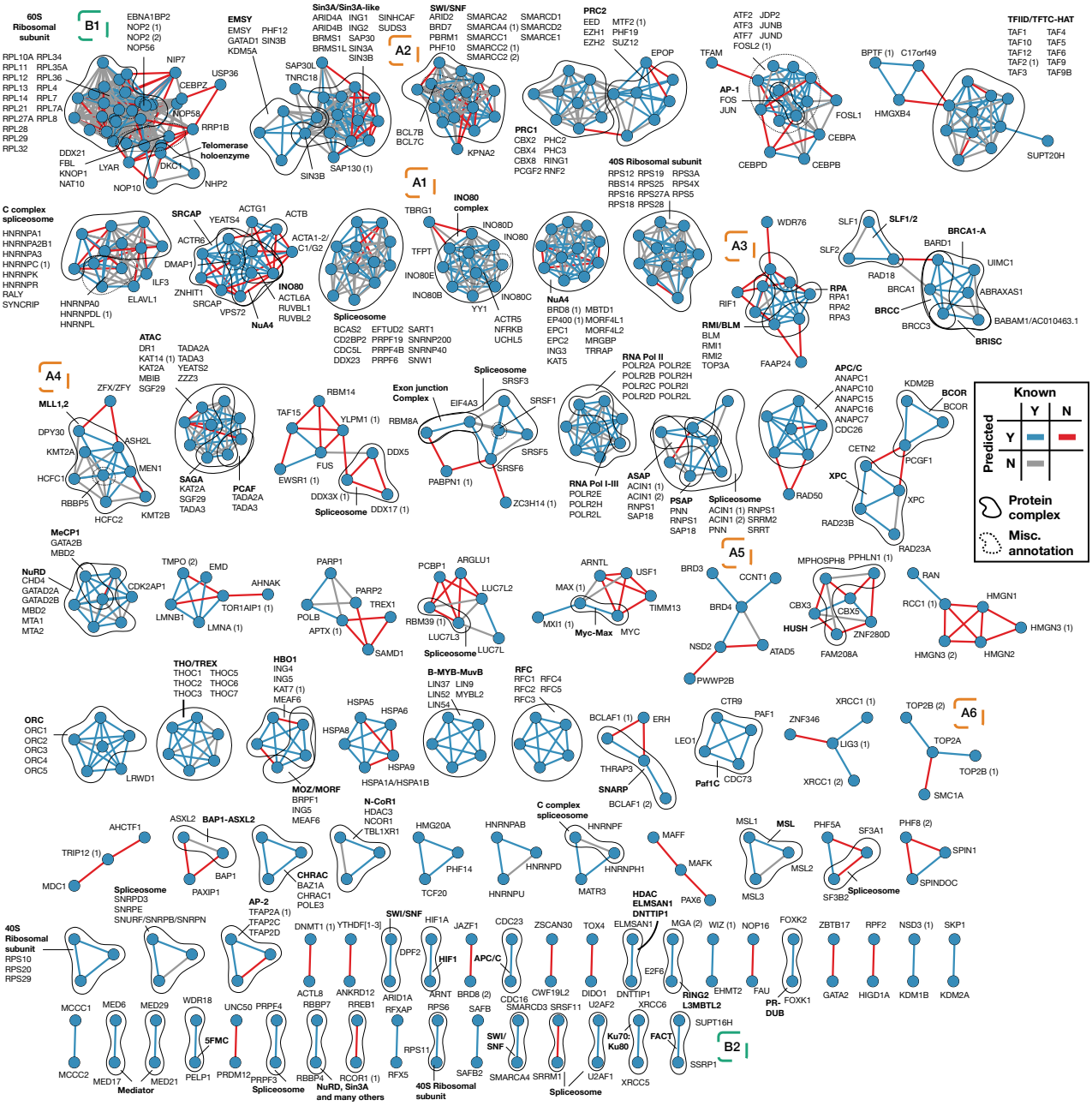

Figure 2.20: High-confidence predictions from the network analysis . Circles correspond to proteins, whose names are annotated nearby. Predicted interactions that have previously been reported in BioGRID are marked in blue, novel interactions are highlighted in red. Known interactions that were not predicted at this threshold are gray. Solid contours indicate protein complexes, while dotted contours group other annotations.

appear as novel in our data, since we only take the human interaction subset from BioGRID and attempt no cross-species matching. Any protein interactions in the literature that have not yet been uploaded to BioGRID would also be marked as "novel".

Finally, any equally strong correlations between a group of proteins would result in an edge between every combination of these proteins from our model. In certain cases, the protein interaction databases would not create a full set of edges between all of the proteins in the group, and would only create an edge between proteins that directly contact each other. In those cases the edges which are missing in the database would appear as novel.

Interestingly, upon closer inspection of the cases where our model completes the clique, we saw instances of proteins completing the clique that are actually not known to interact with the complex in the literature. In such cases the incomplete edges almost exclusively come from large scale affinity pull-down datasets where a member of protein complex was used as a bait. These studies are generally set up to confirm a particular hypothesis an author may have about the protein complex, whose results get published in the main article, while the rest of the data is archived somewhere in the supplementary data. This supplementary data slowly makes its way into the BioGRID database and through it to our model. While a single affinity pull-down study does not necessarily provide sufficient evidence for a truly novel interaction, when combined with predictions from our data, as well as other data sources, it serves as a strong indicator of a truly novel interaction.

The pattern described in the previous paragraph would appear approximately as a vertex poking out of a complex-associated clique in Figure 2.20. This peak would have one or two predicted known interactions in between the unassociated vertex and the complex, as well as a set of novel interactions linking it to the rest of a known structure. While reviewing this data have therefore paid the highest attention to these network motifs. We describe the most noteworthy predictions below.

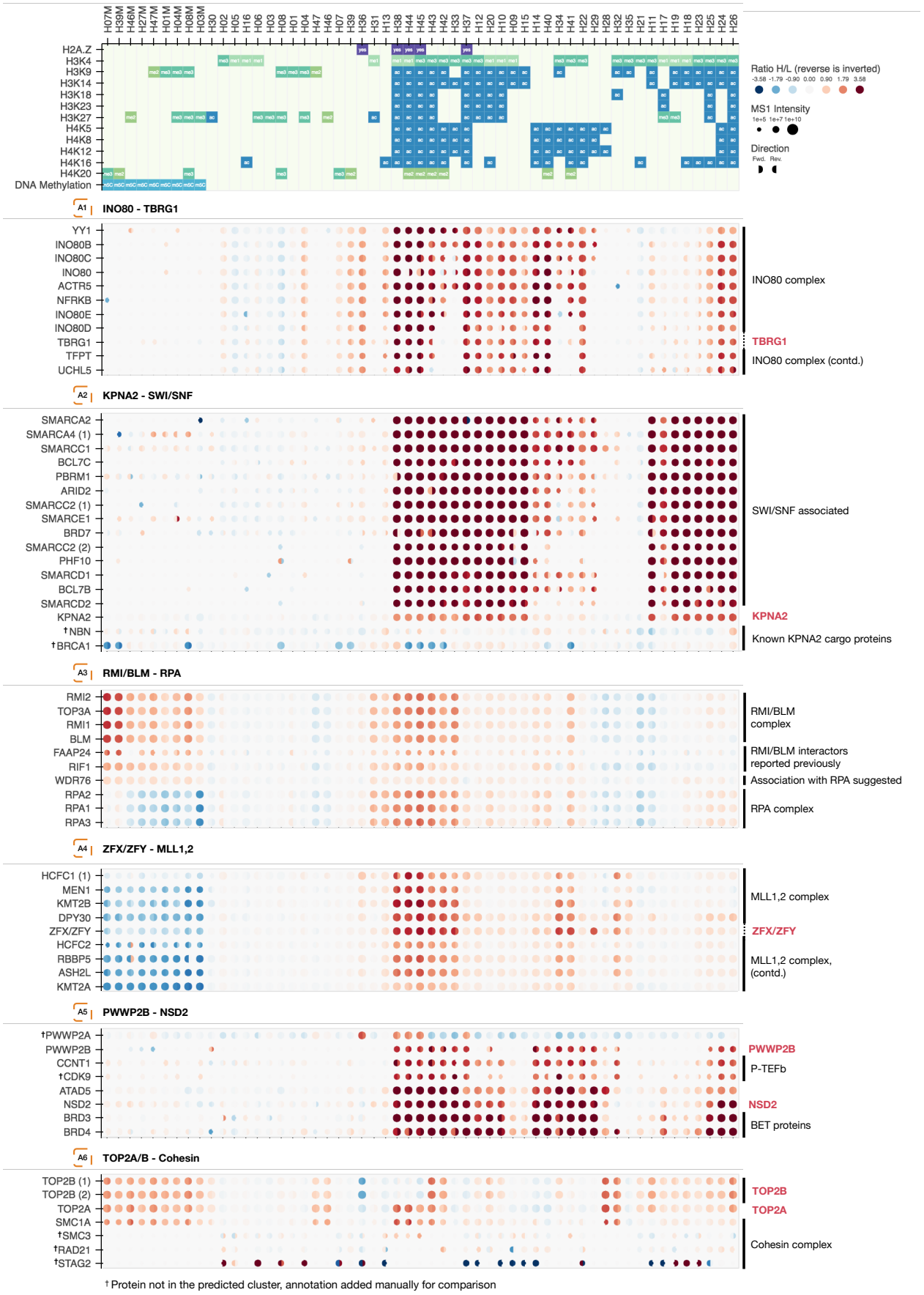

Figure 2.21: Heatmap representation of model's predictions of substantial interest. Proteins marked with "†" correspond to proteins added manually for illustrative purpose.

Predictions of substantial interest

TBRG1 is predicted to be a subunit of INO80 The first cluster we would like to highlight is the cluster formed by the members of the INO80 complex (Figure 2.20 A1). The INO80 complex is a chromatin remodeler responsible for exchanging the H2A.Z histone variant against canonical H2A and for nucleosome sliding[51]. Our data show that Transforming growth factor beta regulator 1 (TBRG1) behaves similarly, if not identically to the other subunits INO80 complex (Figure 2.21 A1). TBRG1 has been discovered as an activator for TP53, a cell-cycle regulator, as well as a protein required in maintenance of chromosomal stability[139].The interaction between TBRG1 and INO80 has previously been suggested by two large-scale interactome studies uploaded to BioGRID[75, 140]. The interaction has also been suggested by Pardo et al. while studying the TBRG1 in another context[141].

KPNA2 is predicted to interact with SWI/SNF Another interesting prediction we would like to highlight is the interaction between the importin subunit alpha-1 (KPNA2) and SWI/SNF protein complex (Figure 2.20 A2). It is clear that KPNA2 shows weaker response to the nucleosomes than other members of SWI/SNF complex (Figure 2.21 A2), suggesting a weak, transient interaction.

This is a very unusual prediction: one would not expect KPNA2 to interact with nucleosomes. Instead, like other alpha importins, KPNA2 should dimerise with beta importins to bring cargo proteins into the nucleus together[142]. While KPNA2 may bind the SWI/SNF complex via the Nuclear Localisation Signal (NLS) motif in the protein to import it into nucleus, we could not find a study describing this. Instead, another protein, KPNA6 has been shown to import a SWI/SNF protein SMARCC1 instead[143]. Nevertheless, two affinity pull-down studies [144, Supplementary Table 1] and [145, Supplementary figure 1B] show interaction between KPNA2 and SWI/SNF.

KPNA2 has been extensively studied in the cancer setting, as unusually high amount of this protein is a marker for poor prognosis[142]. In this field it has been shown to import NBN and BRCA1 DNA repair factors to the nucleus[142, 146]. If a faulty cargo-release was to blame,

we would expect to see a similar correlation between KPNA2 and these DNA-repair factors, but this is not the case (Figure 2.21 A2). Given the SWI/SNF involvement in the activation of DNA repair[147], we could instead speculate that KPNA2 not only chaperones the DNA repair factors into the chromatin, but also targets them to SWI/SNF-bound loci.

WDR76 - RMI/BLM - RPA The third prediction cluster we would like to highlight, Figure 2.20 A3, consists of the core subunits of the RMI/BLM complex and RPA complexes as well as three adapter proteins: Fanconi anemia core complex-associated protein 24 (FAAP24), Telomere-associated protein RIF1 (RIF1), and WD repeat-containing protein 76 (WDR76). Two of the aforementioned adapter proteins, FAAP24 and RIF1 have already been reported to interact with RMI/BLM[148, 149]. WDR76, a reported 5-hydroxymethylcytosine (5-hmC) reader[150], has been shown to associate with the RPA complex at DNA damage sites[151], yet our data hints towards a stronger association with the RMI/BLM instead (see Figure 2.21 A3).

But perhaps the most interesting finding in this cluster is not the predictions of the novel adaptor proteins, but the actual behaviour of the RMI/BLM and RPA complexes themselves. At the core of RMI/BLM lies the RecQ DNA helicase BLM, mutations in which cause a rare genetic disorder called Bloom syndrome. This helicase is known to be involved in homologous recombination DNA repair as well as restarting stalled DNA replication processes[152]. The other core RMI/BLM proteins, RMI1 and RMI2, share sequence homology with the members of RPA complex[152], and functional links between the function of these two complexes have previously been reported[153, 154].

Rather interestingly, at least in some cases these two protein complexes do not cooperate, but instead compete for binding. For instance, direct competition between the two complexes has been shown when binding single-stranded DNA catenanes[155]. We see a similar competition in Figure 2.21 A3, where the presence of DNA methylation switches the complexes into two mutually exclusive binding profiles: RMI/BLM complex binds nucleosomes with methylated DNA very strongly, while the RPA complex is excluded. In absence of DNA methylation the two complexes behave very similarly. Since the mutual information score is invariant to the

sign of correlation, the interactions between the complexes are predicted strongly. As far as we are aware, this has not yet been shown in the literature.

ZFX/ZFY - MLL1,2 The fourth cluster of interest, Figure 2.20 A4, recovers the key subunits of COMPASS-like MLL1,2 complex, including the catalytic subunits KMT2A and KMT2B, as well as a zinc finger protein that is specific to either the X or the Y chromosome. While the mass spectrometer was not able to distinguish Zinc finger X-chromosomal protein (ZFX) from Zinc finger Y-chromosomal protein (ZFY), we could assume that in the female HeLa cells we see the former.

We could not find any evidence for the interaction between ZFX and the COMPASS-like MLL1,2 complex - a well known H3K4 methyltransferase. In a recent study, however, ZFX has been shown to bind at +240bp of CpG promoters in cancer cells, just upstream of the peak of H3K4me3[156]. The same study has also reported downregulation of a subset of ZFX target genes upon siRNA-mediated knockdown, but provided no mechanism. We could speculate from our data that this mechanism is related to ZFX-dependent targetting of COMPASS-like MLL1,2 H3K4 methyltransferase activity, but this cannot be proven conclusively without additional experiments.

PWWP2B-NSD2 association At the centre of the fifth cluster of interest (Figure 2.20 A5), lies Bromodomain Containing Protein 4 (BRD4). This protein is a key epigenetic regulator responsible for releasing RNA Polymerase II from the paused state. This release is achieved via a direct recruitment of the P-TEFb complex via the C-terminal domain of BRD4[157]. One of the predicted interactors in this cluster, CCNT1, is a subunit of P-TEFb. BRD4 is a well-known example of the Bromodomain and Extra-Terminal (BET) family of proteins, a family which also contains the protein BRD3. This protein is highly homologous to BRD4, but lacks the C-terminal domain responsible for recruitment of P-TEFb[29]. Due to the large degree of homology between the two proteins it is hard to tell whether the interaction predicted by our model is due to true physical interaction, or just an indication of highly-specific targeting to nucleosomes shared by the family.

BRD4 can also control RNA Polymerase II via a mechanism independent of P-TEFb. This mechanism utilises the Extra-Terminal (ET) domain conserved in the whole family to recruit multiple proteins, including ATAD5[158]. The study which reported this interaction has also detected an interaction between BRD4 and NSD2 in their experiments, as well as the interaction between BRD4 and BRD3, however the authors chose not to discuss it in the main text.

The predicted NSD2-PWWP2B interaction has never been reported so far, nor has PWWP2B been reported to interact with BRD4. NSD2 is a histone lysine methyltransferase responsible for transferring two methyl groups to lysine 36 of histone H3 (H3K36me2)[159]. Trimethylation of lysine 36 is a landmark of actively transcribed gene bodies. H3K36me3 levels at these gene bodies is often sharply reduced in cancer cells, and instead an increase of H3K36me2 is observed[159]. NSD2 can cause the spreading of this histone modification itself by binding to H3K36me2 via one of its PWWP domains[159]. The PWWP2B protein also contains a PWWP domain, which is believed to bind H3K36me3 due to the similarity to DNMT3A and DNMT3B PWWP domains[160]. In a study of PWWP2A, a protein which contains a homologous PWWP domain to PWWP2B, the authors found that nucleosomes bound by it were no more enriched in H3K36me3 than the control[161, p. 109]. The authors have therefore considered the PWWP domain to not be the primary mechanism of PWWP2A localisation, showing that it is recruited to H2A.Z-containing nucleosomes instead via an alternative mechanism[162]. This recruitment to H2A.Z is PWWP2A specific and not applicable to PWWP2B (Figure 2.21 A5).

It is clear that the recruitment of these factors is not mediated via the H3K36 methylation in our experiments, as we did not test any nucleosomes with this residue modified. The recruitment of PWWP2B to the nucleosome therefore has to be indirect. In support of this theory, pull-downs in HEK293 T-Rex cells using BRD4 as a bait have previously reported PWWP2B interaction[163].

TOP2A/B-SMC1A The sixth cluster of interest, Figure 2.20 A6, contains one of the subunits of the Cohesin complex, SMC1A. Interaction between Topoisomerase 2B (TOP2B) and Cohesin has been reported recently[164]. While we do recover only a single Cohesin subunit in

this cluster due to poor data quality of other subunits, our data suggests a stronger association between TOP2A and SMC1A than TOP2B (Figure 2.21 A6). From Figure 2.21 A6 it is clear that the differences between two topoisomerases are only evident at H2A.Z containing nucleosomes where TOP2B is excluded while TOP2A and SMC1A remains bound. Interestingly, this regulation by H2A.Z together with histone acetylation is predicted as one of the key regulatory dimensions by Factor Analysis (FA) in subsequent section (section 2.3.5, see factor 9).

A link between H2A.Z and Cohesin has previously been demonstrated[165], where H2A.Z was shown to be required for maintenance of appropriate Cohesin levels on chromosome arms. Curiously, the authors have compared the defect caused by lack of H2A.Z to that of topoisomerase II dysfunction, supporting our observation that there might be a physical link between Cohesin and topoisomerases.

A note on sensitivity of SNAP data

Lastly, we would like to illustrate the predictive power of large scale SNAP datasets by example. Figure 2.22 illustrates the clusters B1 and B2 in Figure 2.20. The former cluster corresponds to the joint cluster of the 60S Ribosomal subunit and the Telomerase holoenzyme forming the largest high-confidence prediction cluster in the data set. The latter cluster, cluster B2, highlights the FACT complex consisting of the two proteins SSRP1 and SUPT16H whose interaction score is the highest in the dataset.

As seen in Figure 2.22, both complexes barely show any forward or reverse response to the nucleosomes. Such low responses are indicative of the proteins or protein complexes not interacting with the PTMs tested in our experiments, but interacting with the nucleosomes non-specifically. Proteins that interact with RNA, such as the ribosomal proteins, often exhibit such weak responses, which could be hypothesised to occur due to cross-reactivity to the DNA. A subset of these proteins, including the ribosome, may come from cytoplasmic contamination. Some known nuclear proteins, including the aforementioned FACT, and some TFs also exhibit weak responses to our nucleosomes, perhaps due to similar non-specific interactions with the nucleosomes or the DNA. Nevertheless, even such non-specific interactions with the nucleosomes

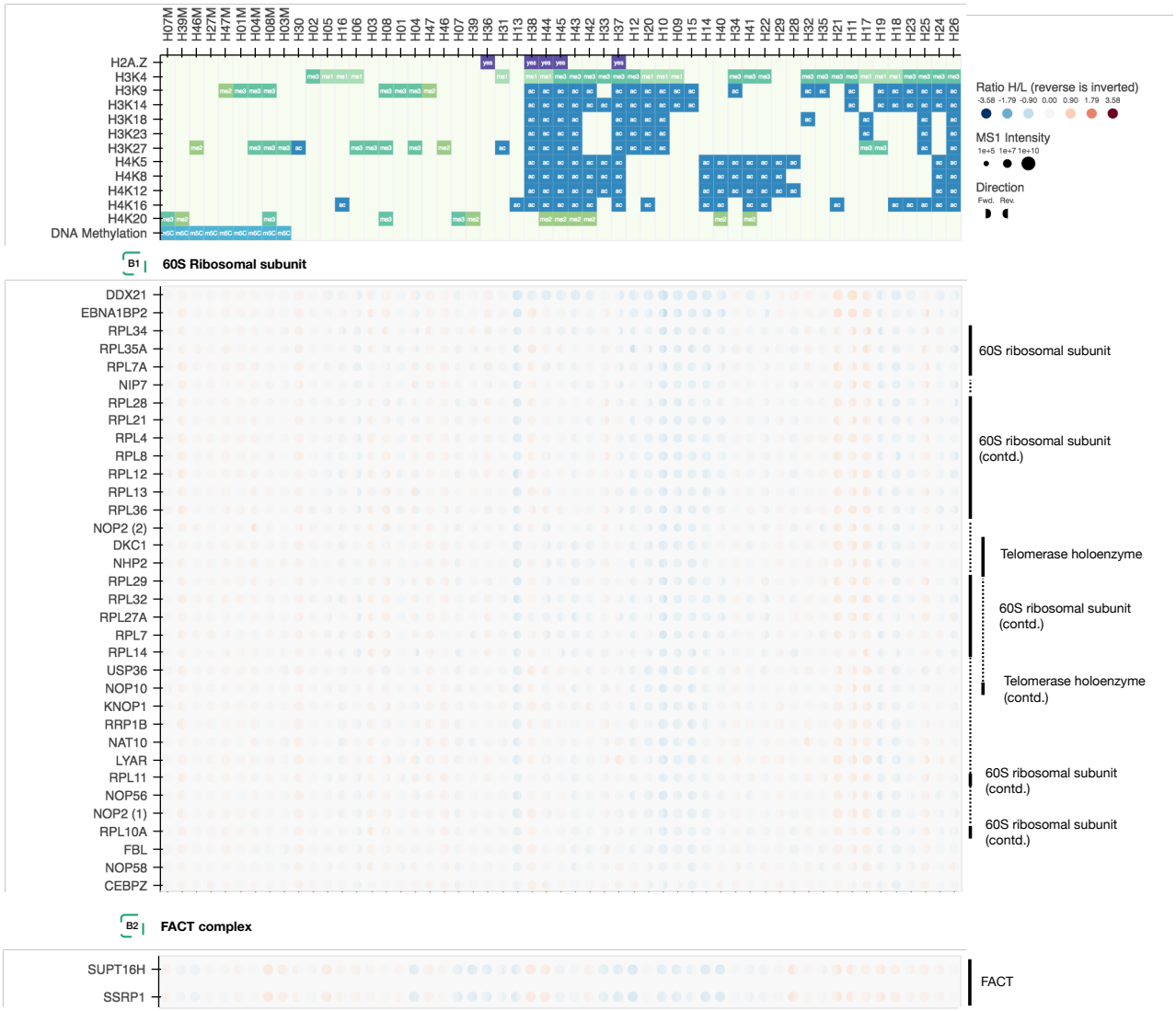

Figure 2.22: Heatmaps for B1 and B2 high-confidence prediction clusters in Figure 2.20. Note almost no response to the modified nucleosomes, yet a strong correlation across the proteins.

are subject to small perturbations, which are detectable, but may primarily be attributed to noise. All subunits of chromatin complexes that are largely indissoluble are subjected to the same (albeit weak) perturbation simultaneously, which clearly causes a similar binding pattern across the set of profiled nucleosomes. As seen in Figure 2.22 this property holds for both large (i.e. 60S Ribosomal subunit) and small (i.e. FACT) protein complexes. Our method is sensitive enough to detect such fluctuations.

Other examples of such associations include the associations that we detect occur between the spliceosome, RAD50 and APC/C complexes as well as RNA Polymerase II subunits.

While we find these predictions interesting, and are confident in them, it is hard to speculate about the nature of this low-level regulation given the data that we have. We therefore do not put strong emphasis on these predictions here, however note that they might still be of interest to the reader as they demonstrate the power of the approach to predict protein complexes.

2.3.5 Eleven regulatory dimensions, two complementary viewpoints

So far we have only considered the dependencies across the proteins identified in the sample. It is no less important to study the relationships between differentially modified nucleosomes as well. We have already seen in section 2.3.1 that major regulatory pathways are shared by the different modified nucleosomes. In this section we extend this analysis and explicitly describe the constitutients of the regulatory space described by these modified nucleosomes.

Specifically, we use Factor Analysis (FA), a method described in section 2.2.4, to identify 11 regulatory axes – hidden dimensions each protein is embedded in – that are sufficient to describe the behaviour of our data. As discussed in the accompanying methods section, these dimensions are not unique and can be rotated to provide a different viewpoint of the data. We have carefully selected two of the available viewpoints that we believe are complimentary to each other, which are described below.

Figure 2.23 A shows the clustering of loadings produced by the two rotations. The values plotted correspond to the correlation between the hidden dimensions (i.e. regulatory axes) and visible dimensions (i.e. response to modified nucleosome) and indicate which nucleosomes are influenced by which dimension. While some of the regulatory axes (e.g. *bentlerT's* FA7 and *bifactor's* FA4) are inverted, it is clear that both rotations result in very similar representations of the regulatory subspace as the regulatory axes cluster in pairs. At the same time, there are a few important distinctions between the two methods. Since *bifactor* rotation aims to create the first factor as general as possible[129] it captures the whole set of nucleosomes that are acetylated, or have H3K4me3 signal present. *bentlerT* rotation, which avoids general factors altogether mostly captures nucleosomes associated with H3K14ac more strongly than other acetylations (i.e. the loading for such nucleosomes is higher, see Figure 2.23 for data

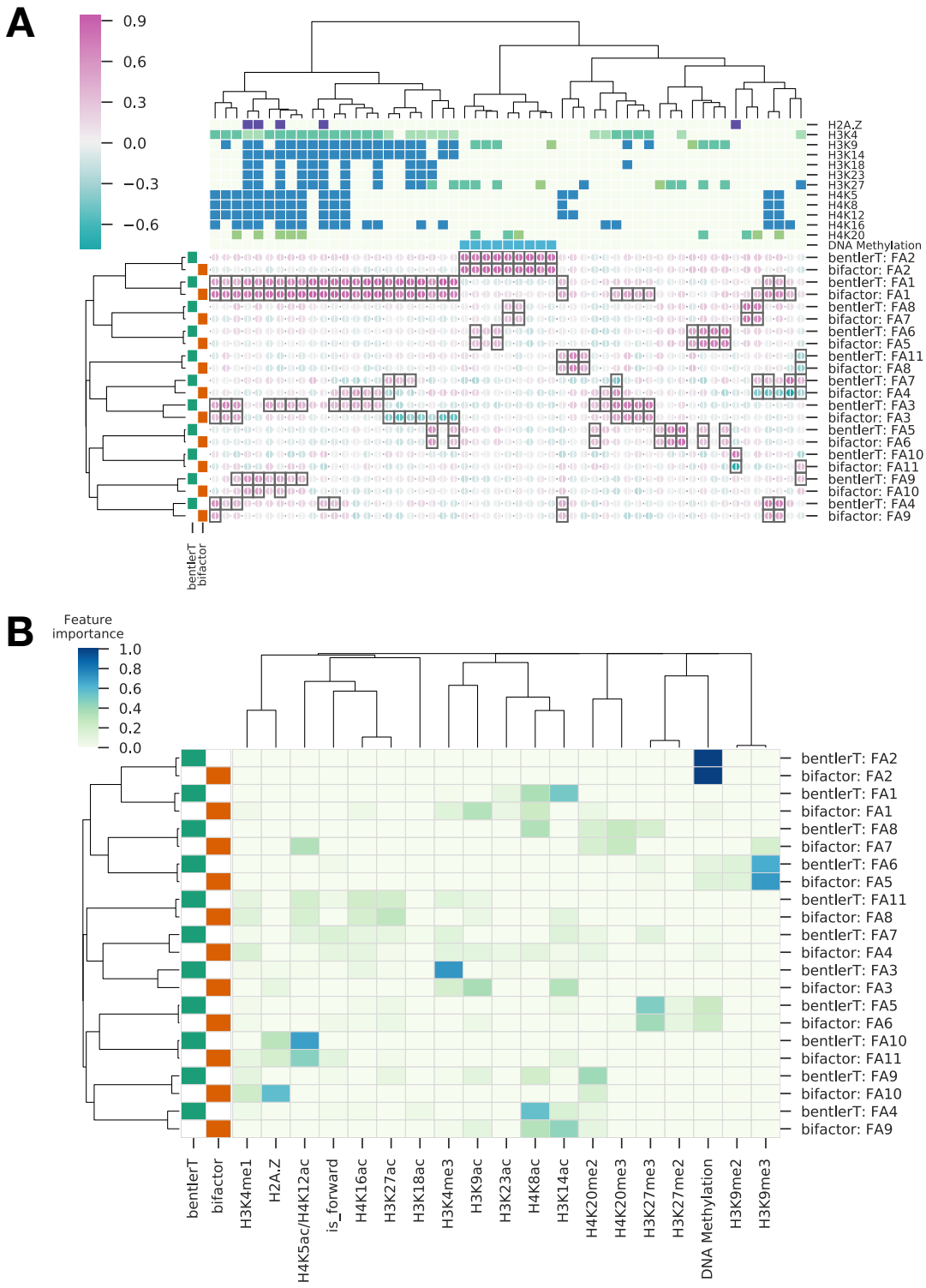

Figure 2.23: Comparison between *bentlerT* and *bifactor* rotations. The rows of this heatmap are clustered by absolute correlation. (A) Clustering of the identified axes using absolute correlation. Axes clearly group in pairs. (B) Feature importances as reported by the decision trees trained to specify the axes (see section 2.2.4). The loading patterns are further summarised in Table 2.2.

bentlerT rotation			*bifactor* rotation		
Factor	Explained variance	Key regulators	Factor	Explained variance	Key regulators
FA1	24.33%	**H3K14ac**, H4K8ac, H3K23ac	FA1	30.88%	**Acetylation**, H3K4me3
FA2	12.52%	meDNA	FA2	11.84%	meDNA
FA3	9.45%	H3K4me3	FA3	5.94%	H3K4me3 (+), H3K9acK14ac (–)
FA4	4.82%	H4 acetylation	FA9	2.43%	H4 acetylation
FA5	4.61%	H3K27me3/2	FA6	4.38%	H3K27me3/2
FA6	4.48%	H3K9me3/2	FA5	4.43%	H3K9me3/2
FA7†	4.34%	**H4K16ac (+)**, H3K4me1K27ac (+), H4K20me2 (+), H3K4me1K9acK14ac (+), H3K9acK14ac (+), H4ac (+), **H3K4me3+H4K16ac (–)**	FA4†	5.90%	**H4K16ac+H3K4me3/1 (+)**, H3K4me+H3K9acK14ac (+); H3K9acK14ac (–), H4K16ac (–), H4ac (–), H4K20ac (–), H3K4me1K27ac (–)
FA8	3.65%	H4K20me2/3	FA7	3.14%	H4K20me2/3
FA9†	3.24%	H2A.Z/H4K20me2 with acetylation and H3K4me, H3K4me1K27ac	FA10†	2.28%	H2A.Z with acetylation and H3K4me1/H4K20me2
FA10	2.40%	H2A.Z	FA11	2.27%	H2A.Z (–)
FA11†	2.34%	**H4K5acK12ac** (+), H3K27ac (+), H3K4me1H3K27ac (–)	FA8†	2.67%	**H4K5acK12ac** (+), H3K27ac (+), H3K4me1H3K27ac (–)

Table 2.2: Readouts of loaded regulators for the 11 regulatory axes predicted by FA. The regulators are sorted by expected variance of *bentlerT* rotation on the left, and by the closest *bifactor* equivalent (as seen in Figure 2.23) on the right. "(+)" and "(–)" indicate diverging dimensions where positive values on the axis are correlated with modifications indicated by (+), and negative values are associated with the (–) marked modifications. When these indicators are omitted, positive values are assumed to be associated with the specified mark, and negative values with its absence. In cases where multiple regulators are present, the strongest one is written in bold text. Dimensions indicated with † share complex patterns that are hard to describe from the loading alone.

and Table 2.2 for summary). The two rotations converge at the third component, FA3, where *bentlerT* re-captures H3K4me3 associated nucleosomes, while *bifactor* re-captures the specificity of H3K14ac by creating a divergent dimension.

As shown in Table 2.2, most of the regulatory dimensions can be identified directly from the loading matrix. In some cases, specifically FA7, FA9 and FA11 of *bentlerT* rotation, and their equivalent FA4, FA10, FA8 in *bifactor* result load multiple nucleosomes that have few marks in common. To investigate these more complicated dimensions, we have tested the estimated factor scores for association with annotations from UniProt, Interpro and HOCOMOCO databases.

Figure 2.24 displays the result. From this figure it is clear that *bentlerT*:FA7, and *bifactor*:FA4 dimensions split the Spliceosome and the ribosome (RNA-associated protein complexes), from DNA-binding proteins such as transcription factors. The *bentlerT*:FA9 – *bifactor*:FA10 pair separates the Mediator, Integrator complexes from the ATAC, MLL, NSL and PRC1 complexes. The last pair, *bentlerT*:FA11, *bifactor*:FA8, separates Fos/Jun, HMGN, bZIP and Skn-1 related TFs, as well as RMI/BLMand TFIID from MLL3,4, CHRAC, NuA4, SRCAP and Topoisomerase II. This regulation of DNA and RNA binding factors by the by FA7-FA4 and FA11-FA8 pairs is particularly interesting and we believe the complex pattern of modifications it is associated with could be a proxy for the accessibility of nucleosomal DNA. We discuss this further in section 2.3.6.

Summarising all of the information above we are able to divide the interaction network of epigenetic readers into distinct modules that are controlled by these regulatory axes. The visual summary of these results is shown in Figure 2.25, and described below. The axes are listed in the order obtained from the *bentlerT* rotation, but the differences of the equivalents obtained using the *bifactor* rotation are noted where appropriate.

Factor 1 captures active euchromatin associated with histone acetylation, especially with H3K14ac. Strongly associated with the SWI/SNF chromatin remodeler complex, while weakly negatively associated with the spliceosome. The *bifactor* rotation combines the effects of this factor in *bentlerT* with partial effects of the third factor, making it less specialised to SWI/SNF-

Figure 2.24: Protein domains, complexes or transcription factor families which have greater mean factor score than the rest (as identified by student's t test), for the three complex axes in FA. The plotted value is the Student's t statistic.

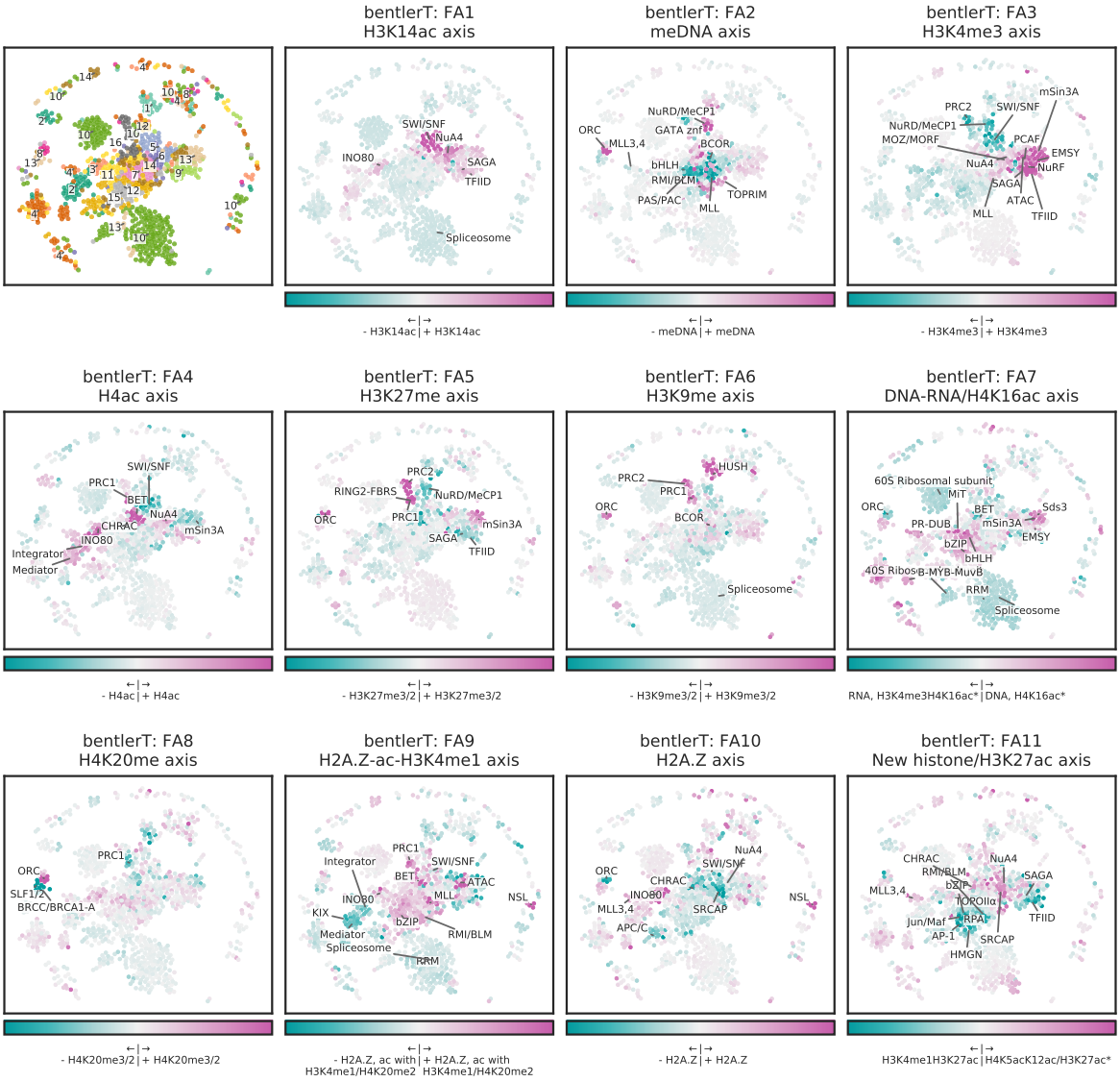

Figure 2.25: Summary figure for the results of FA. Dots, representing proteins, are coloured based on the factor score for corresponding dimension. The top-left panel is divided into protein groups obtained from hierarchically clustering the data (Figure B.5 A). Annotations were manually curated from the set of significantly regulated protein complexes, domains and transcription factors, available in Figures B.1, B.2 and B.3. The equivalent figure for *bifactor* rotation can be found in the appendix (Figure B.4).

associated H3K14ac, and adding more emphasis to other acetylations and trimethylation of lysine 4 on histone H3 (H3K4me3).

Factor 2 captures regulation by DNA methylation. It divides proteins that associate with DNA methylation, such as proteins containing methyl-CpG binding domains, from the proteins repelled by the DNA methylation such as CXXC-type zinc finger-containing proteins.

Factor 3 captures the regulation by H3K4me3. It splits the proteins associating to H3K4me3 from proteins that associate with H3K4me1 or no methylation on lysine four of the histone H3. The *bifactor* rotation makes this split a bit more pronounced and splits H3K9acK14ac associated proteins from H3K4me3-associated ones to compensate for adding H3K4me3 association to the first factor already. The TFIID, EMSY, PCAF, SAGA, SIN3A complexes positively associate with this dimension, while SWI/SNF complexes associate negatively. This negative association is more pronounced in *bifactor*.

Factor 4 captures regulation by histone H4 acetylation. Positively associated with bromodomain containing proteins, especially members of the BET family. Positively associates with CHRAC, INO80, Integrator, Mediator complexes. Negatively associated with SWI/SNF complexes, likely due to the previous factors explaining their behaviour better.

Factor 5 captures regulation by methylation on lysine 27 of histone H3 (H3K27me3/2). Polycomb group complexes PRC1 and PRC2 are primarily regulated via this dimension. ORC and SIN3A are also regulated positively. While NuRD, TFIID complexes are regulated negatively. Response to me3 appears to be stronger than to me2.

Factor 6 captures regulation by methylation on lysine 9 of histone H3 (H3K9me3/2). This dimension regulates the CAF-1, HUSH complexes as well as ORC complexes positively via chromodomain containing proteins. Response to me3, again, appears to be stronger. There are few protein families regulated negatively.

Factor 7 captures a set of nucleosomes that are hard to summarise via a common histone modification. The divergence between H4K16ac and H4K16ac+H3K4me3/1 is common to both rotations. In the *bentlerT* rotation, proteins associating with H4K16ac have positive scores from this factor, and proteins associating with H3K4me3+H4K16ac - have negative scores instead. The corresponding factor 4 in the *bifactor* rotation flips the sign of this relationship. From the analysis in Figure 2.24, this dimension seemingly separates the DNA-associated transcription factors from the RNA-associating spliceosome and ribosome subunits. We further explore this relationship between section 2.3.6.

Factor 8 captures the regulation by H4K20me2/3 containing nucleosomes. ORC proteins associate with this dimension positively, while the BRCA1-A, BRCC and SLF1/2 complexes associate negatively. Response to me2 and me3 appears to be equal.

Factor 9 is another factor with a complicated loading which captures a set of nucleosomes that contain H2A.Z/H4K20me2 together with fully acetylated histones in combination with H3K4me1, similar to the TOP2A and Cohesin subunit SMC1A binding pattern already observed in Figure 2.21 A6. This complex set of modifications captured by the regulatory axes correlates with the PRC1 , MLL1,2, NSL, ATAC and RMI/BLM complexes positively, and with Integrator, Mediator complexes negatively. The positive association with INO80 and negative association with Spliceosome is specific to *bentlerT* rotation, but its significance is lost in the corresponding *bifactor* factor 10. Positive association with bZIP-domain containing proteins and BET proteins is also significant in both rotations.

Factor 10 captures regulation by H2A.Z-only containing nucleosome. It is interesting that this dimension only loads on the nucleosome containing H2A.Z on its own, but not the other H2A.Z containing nucleosomes, which are captured by factor 9. Notably the corresponding factor 11 in the *bifactor* rotation also loads H3K4me1K27ac in the opposite direction. The INO80 and NSL complexes associate positively with this axis, while NuA4 and SRCAP, ORC, CHRAC and APC/C associate negatively.

Factor 11 is a complex dimension which captures regulation by histone H4 acetylation without H4K16ac, especially H4K5acK12ac, and also splits the effect of H3K27ac (positive) from H3K4me1K27ac (negative). Positively associates with Topoisomerase II, and the SRCAP and CHRAC complexes, but negatively associates with the TFIID, RPA and RMI/BLM complexes. Associates with Fos- and Jun- related transcription factors. May serve as a proxy to linker DNA accessibility (see section 2.3.6).

2.3.6 Factors with complex loadings regulate DNA-binding transcription factor response

In the section 2.3.5 we have explored the key regulatory dimensions of the SNAP dataset. The factor 7 within these dimensions has proven to be the most surprising as it clearly separated the DNA-binding transcription factors from the rest of the dataset. In this section we explore this relationship further.

We have downloaded the complete list of DNA-binding transcription factors from HOCOMOCO and have clustered their profiles, as described in section 2.2.5 (Figure 2.26 A). We have found Basic Helix-Span-Helix (bHSH) and Basic Leucine Zipper (bZIP) factors to be significantly over-represented in clusters 2 and 6, respectively (Figure 2.26 B and C, hypergeometric test, adjusted p-value threshold of 0.01).

In order to determine if this clustering is DNA sequence motif-dependent we have scanned the dinucleosomal DNA sequence for TF binding site motifs (as described in section 2.2.5). Most of the binding sites found occurred within the segments of DNA where we would expect the nucleosomes to be placed (Figure 2.26 D, Figure B.7 A).

Only cluster 6 contains seemingly strong motifs in the centre of the linker DNA (see Figure B.7 A), with other clusters sharing a pattern less pronounced. Interestingly, histone H1 (HIST1H1C, HIST1H1[D,E]) and TFAM show a very similar binding pattern to the proteins in cluster 6 (see Figure 2.28), in fact, the interaction is strong enough to be in the high-confidence set (see section 2.3.4), had the histone proteins have not been excluded from it (as described

section 2.2.3). The behavioural similarity of these factors to the linker histone H1 further suggests the involvement of the linker DNA. TFAM is a mitochondrial DNA binding protein that likely appears in our dataset as a contaminant. The similarity of its binding patterns to those of histone H1 is still still intriguing as it is known to package the DNA in mitochondria[166].

Four regulatory axes control the DNA binding transcription factors

Plotting the median factor scores across each of the clusters, we observe that factors FA2, FA7, FA9, and FA11 are the strongest regulators of the median behaviour of the clusters (Figure 2.27 A). Unsurprisingly, proteins in all clusters have a strong positive score associated with FA7, which we previously described to separate DNA binding proteins from the RNA binding ones. The separation of transcription factors is even more pronounced, when plotted in a rank-based plot (Figure 2.27 B). Note, that while this is true for transcription factors in general, some notable TFs, such as TBP are not enriched in this dimension. From Figure 2.26 A, it is clear that TBP binding pattern is dominated by H3K4me3 nucleosomes (as is the binding pattern of TFIID proteins), and therefore it is not primarily recruited by the DNA interactions.

As certain DNA binding factors are excluded by DNA methylation, while others are recruited by it, we see the response to FA2 differ across the clusters. FA9 and FA11, were the only two other factors whose loadings were complicated (see Table 2.2) are the only remaining factors to vary across the clusters. A trend line separating the TF clusters can be seen clearly when the scores for these two factors are plotted on a two-dimensional grid (Figure 2.27 C). In this plot the cluster 6 is a clear outlier with the strongest negative association with factor 11. We have previously remarked that this cluster is apparently the only cluster which contains a strong binding motif within the linker DNA. Specifically, FOS and JUN and related proteins bind to the TGAGTCA motif, a variant of which, TGA**T**TCA is present on the linker DNA sequence (bp 180-186, see Figure B.7 A, Figure B.6). This, together with the observation that these proteins behave similarly to the histone H1 (Figure 2.28), lead us to suggest that factor 11 may be indicative of the linker DNA accessibility. This observation is further strengthened by observing that members of the cluster 2, the cluster exhibiting the second strongest negative

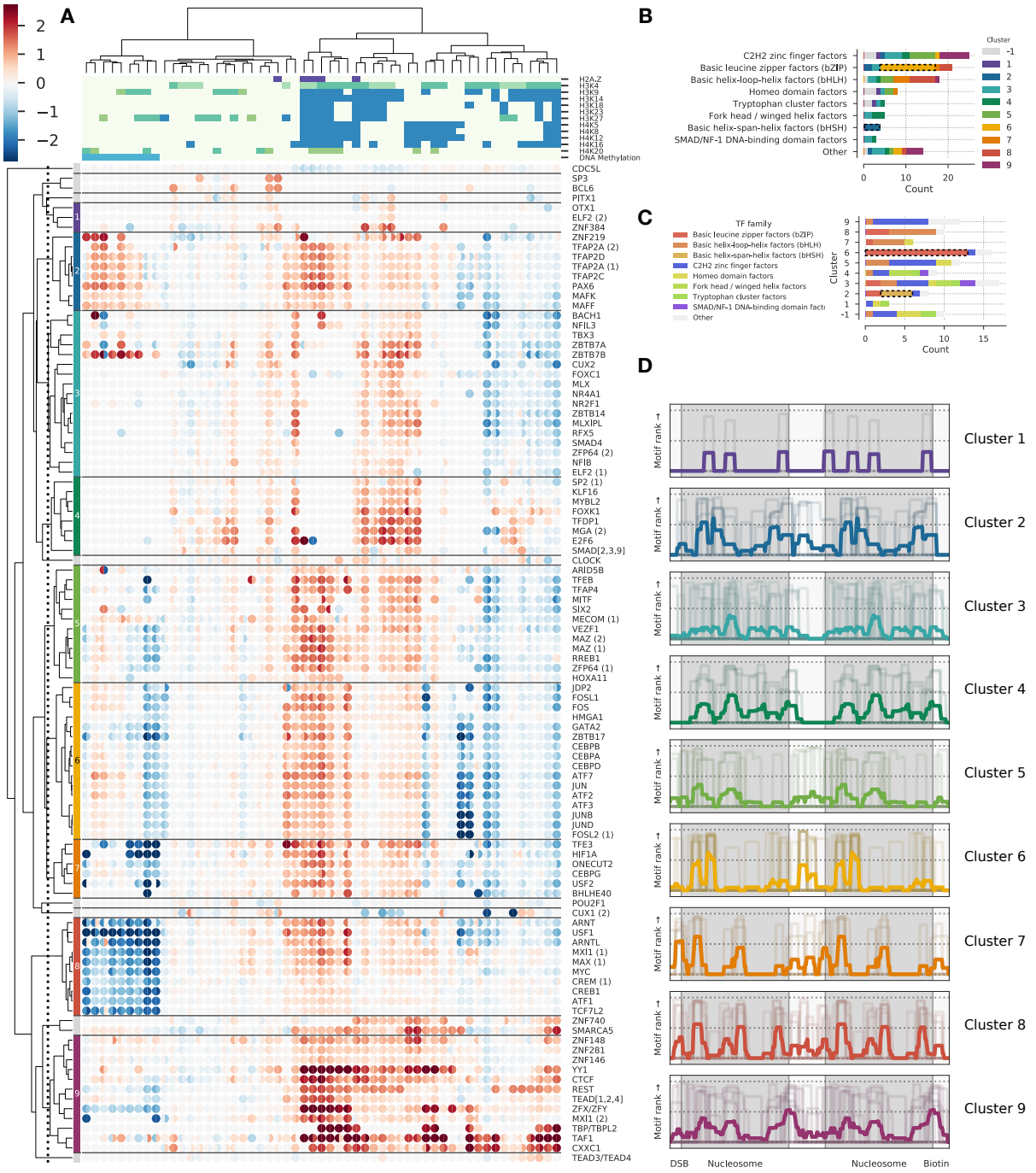

Figure 2.26: (A) Clustering of all transcription factors observed in the SNAP dataset (metric: cosine, method: complete). (B) Distribution of transcription factor families across the resulting clusters. (C) Distribution of clusters across transcription families. Families statistically overrepresented in clusters are highlighted by dotted lines (hypergeometric test, B/H adjusted p-value < 0.01) (D) Mean distribution of predicted binding sites for cluster members across the di-nucleosomal DNA stretch. The values plotted on the y axis correspond to rank-normalised locations of binding sites for a protein in the cluster (as measured by the p-value). Each dim line corresponds to individual protein in the dataset, whereas highlighted line corresponds to the mean behaviour. See Figure B.7 A for the same data visualised in a heatmap format.

response to factor 11 on average (Figure 2.27 A), also exhibit a potential binding site on the linker DNA, albeit with a lower score (Figure 2.26 D).

Relationship between CHD- family of chromatin remodelers and the regulatory axes that control TF binding

Interestingly, the preference for H3K4me3 and dislike of H4K16ac (the inverse pattern observed in the loadings of FA7) has previously been established as the binding pattern of CHD1[49]. This is also evident in Figure 2.27 B, where the CHD1 clearly lies on the opposite side of the factor 7 from the TFs. Rather surprisingly, another CHD- chromatin remodeler, CHD2, shows a similar relationship within the score distribution of the factor FA11 instead (Figure 2.27 D).

These two observations have lead us to investigate the behaviour of other CHD chromatin remodelers as well. We have therefore computed the median correlation of other CHD remodelers with the TF binding clusters (Figure 2.27 E). CHD1 has shown a negative, albeit weak, correlation with most clusters, except clusters 8 and 9, where the correlation was positive. CHD2 has shown a negative correlation with cluster 6 (Median Spearman's $\rho = -0.59$), another CHD-protein, CHD7 has shown a strong positive correlation with the cluster members ($\rho = 0.77$), which can be seen in Figure 2.28. Similarly, all CHD proteins have associated with specific clusters in seemingly combinatorial pattern (Figure B.7 B) pointing towards a relationship between CHD remodellers and TFs in their modes of nucleosome binding.

It is important to note that this evidence is correlative and therefore insufficient to define causation. It is unlikely that any ATP-dependent remodelling occurs in our experiments, as the amount of adenosine triposphate (ATP) in the nuclear extracts is negligible at most. ATP-independent nucleosome structure modification has previously been shown for FACT[167], but not for CHD remodelers.

Therefore this point we can only highlight that the factors with complex loading patterns control the binding of TFs, histone H1 and CHD- remodelers. But taken taken into consideration that the binding sites of the identified TFs cluster at the DNA entry and exit sites as well as the dyad

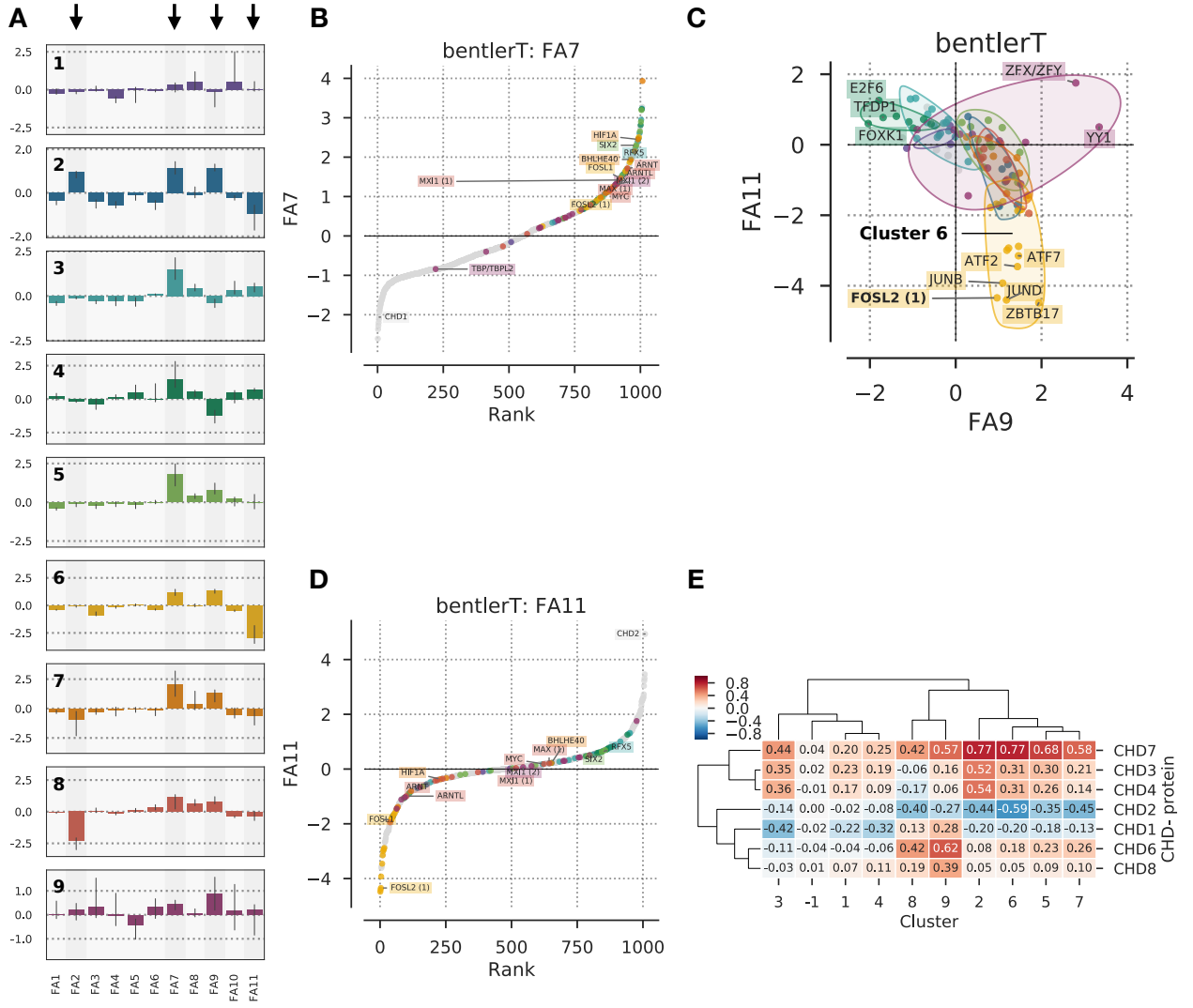

Figure 2.27: (A) Median factor scores from the bentlerT factor model (see section 2.3.5), for DNA binding proteins in each of the clusters. Four most important regulatory TF axes FA2, FA7, FA9 and FA11 are indicated by arrows. (B) Rank plot of axis FA7, which is consistently enriched in all of the 9 clusters, showing separation between CHD1 and transcription factors. (C) FA9 and FA11 axes together separate the clusters. FA11 axis influences the members of the cluster 6 the most. (D) Similarly to CHD1-TF separation seen in (B), CHD2 lies on the opposite side of FA11 from the members of cluster 6. (E) Median Spearman correlation between the transcription factor clusters and members of the CHD family of chromatin remodelers.

Figure 2.28: Heatmap of members of the TF cluster 6 along of with their high confidence interactors HIST1H1C, HIST1H1[D,E] and TFAM, together with a subset of CHD proteins. Similar heatmaps for other clusters are available in Figure B.7 B.

axis of the nucleosome (Figure 2.29), or the linker DNA between nucleosomes (similar to histone H1), and their binding is somehow related to binding patterns of CHD chromatin remodellers, these results likely indicate that these axes control DNA accessibility associated with the di-nucleosomes. Further research is therefore required to establish the exact mechanisms behind these observations.

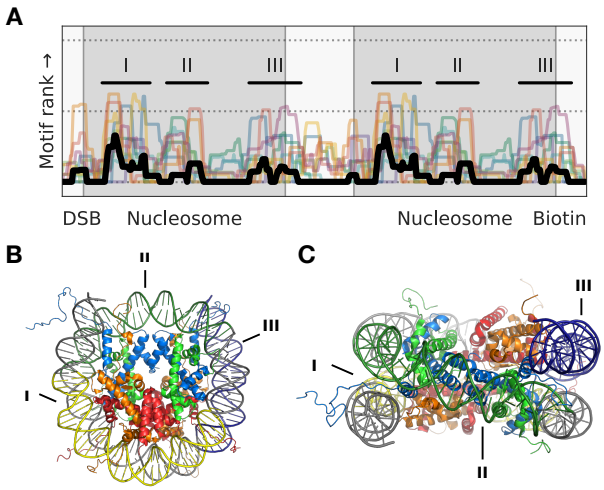

Figure 2.29: Visualisation of the TF hotspots on the nucleosomal core particle. (A) Three hotspots of TF motifs on the nucleosomes used in SNAP experiment. (B) and (C) localisation of these hotspots on the 3D structure of the nucleosomal core particle (PDB identifier 1EQZ). DNA within the motifs I, II and III is colour coded in yellow, green and blue respectively. Histones H2A and H2B are coloured orange and red, while histones H3 and H4 are blue and green. We kindly thank Dr. Andrey Tvardovskiy for help with the 3D structure visualisation.

2.3.7 Linear models recover epigenetic reader binding profiles

In the previous results sections we have described the key regulatory axes for the whole dataset that succinctly defined the regulation of the chromatin interaction network that we established. Assigning the factor scores to the epigenetic proteins allowed us to break down this network into distinct subunits that respond to these regulatory determinants. While some of the estimated regulatory axes undoubtedly represent known chromatin regulatory elements, others are not trivial to interpret as they correspond to a complex regulation by multiple nucleosomes.

Insights into chromatin biology suggest that chromatin readers contain very specialised protein domains capable of recognising individual nucleosomal modifications (see section 1.1.3 in the introduction), which can be combined to create a combinatorial response. It is therefore natural to describe the epigenetic readers in terms of the nucleosome modifications they recognise. For instance, HP1 proteins are commonly referred to as readers for H3K9me2/3. Conceptually, such definitions establish a recognition motif for epigenetic readers, similar to the DNA recognition motifs that TFs associate with.

In this section we are therefore interested in inferring such recognition motifs automatically.

We intend to develop a model that could determine that CBX5 is a reader for H3K9me2/3 or that SPIN1 is a reader for H3K4me3, as well as identifying readers that recognise multiple nucleosomal modifications.

We base this model on a *LASSO* linear regression framework (see section 2.2.6), assuming that the ratio of the chromatin reader binding to the nucleosome can be directly explained by the modifications. In this sense, the estimated coefficients from the linear model correspond to the $\log_2$ fold-changes in the chromatin reader abundance associated with the modification. Searching for proteins with positive coefficient associated with H3K27me3, for instance, and a negative coefficient associated with DNA methylation would therefore produce a list of proteins that are recruited by H3K27me3, but excluded by the presence of DNA methylation.

In this model, the influence of combinatorial combinations of PTMs is assumed to be additive. We have already illustrated the intuition behind such an additive response in Figure 2.2, where UHRF1 binding to H3K9me3 was enhanced by the presence of methylation on the DNA. Given the histone code hypothesis, in its literal sense, such additive responses should not be common as we would expect a combination of the histone modifications to produce a distinct response that could not be explained by simple combination of the two responses[26]. Nevertheless, additive response is the simplest possible hypothesis about combinatorial regulation of nucleosomal PTMs and therefore should be explored first.

Surprisingly, 316/492 (64.23%) of protein binding profiles in the analysed dataset of proteins enriched in at least four pull-downs can be approximated by this simple model (see section 2.2.6).

Looking into the representative subset of such predictions (Figure 2.30), it is clear that known epigenetic reader interactions with histone PTMs are recovered by the model. For instance, the model correctly predicts the interaction between CBX5 and H3K9me3/2[168], between SPIN1 and H3K4me3[169], and between MECP2 and DNA methylation[170].

The binding profile of other proteins is also succinctly summarised by the estimated binding motifs from the model (Figure 2.30 B). For instance, MBD4 is clearly recruited to nucleosomes containing methylated DNA, but excluded from nucleosomes containing H2A.Z. SUZ12, the

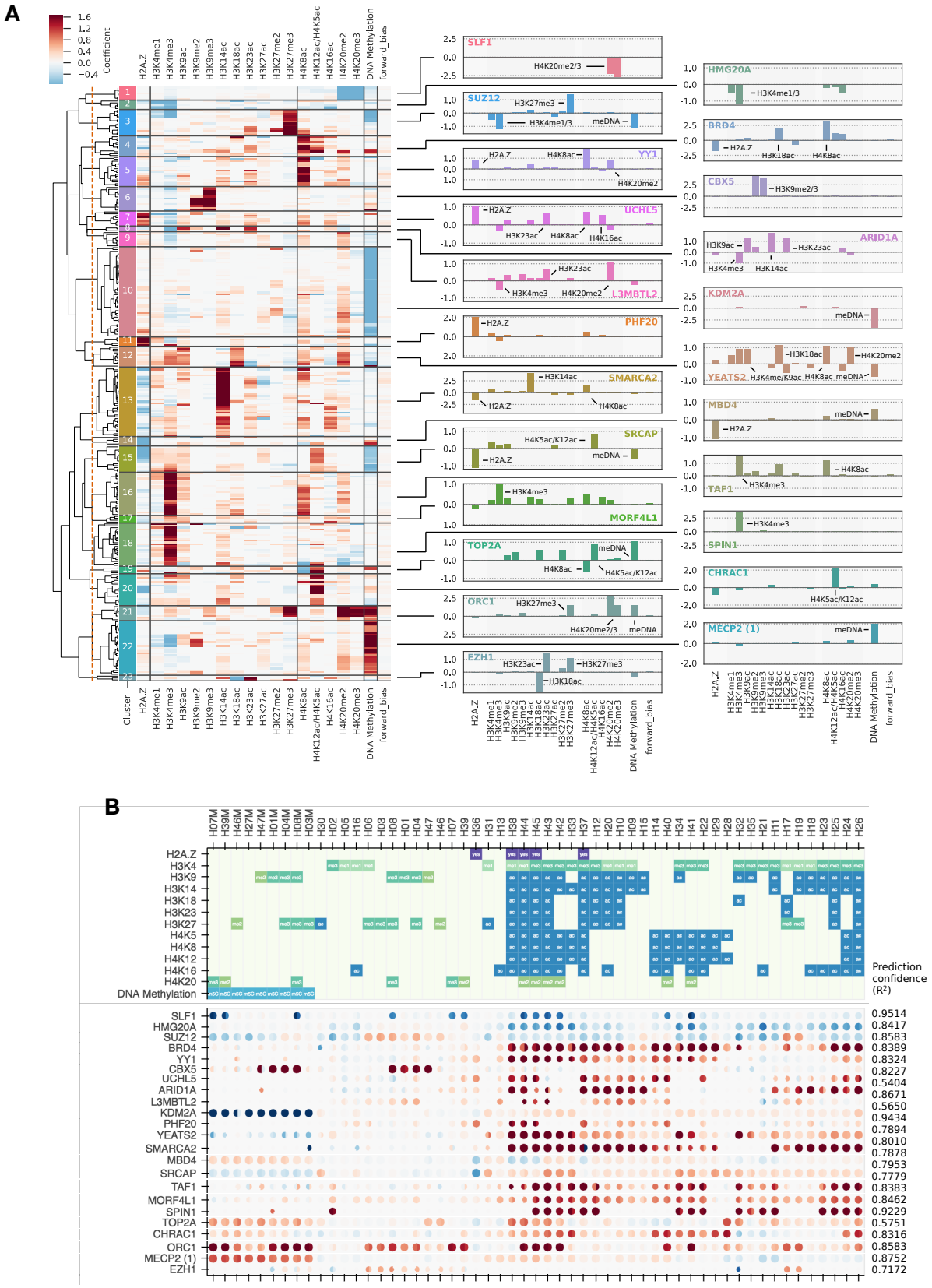

Figure 2.30: Epigenetic reader motifs estimated from the SNAP data. All 316 well-predicted proteins in the right panel of Figure 2.16 are shown in the clustered heatmap in (A). For each of the sub-clusters in this heatmap, a bar plot of estimated coefficients is shown on the right. The raw binding ratios for each of the representative proteins are shown in (B).

DNA/RNA binding subunit of PRC2 complex, is clearly recruited to H3K27me3 as predicted from the model (and the literature[5]), but from Figure 2.30 B it is clear that this recruitment is reduced at regions containing DNA methylation, which is clearly indicated in the model's output. Similarly, the model clearly highlights the SUZ12 apparent exclusion from regions marked with H3K4me1/3.

The response to DNA methylation and H3K4me1/3 is less pronounced in another subunit of this complex – EZH1. EZH1 is the alternative, weaker, catalytic subunit of PRC2 (as compared to EZH2)[171]. Interestingly, the model predicts an antagonistic response between H3K18ac and H3K23ac for this protein (Figure 2.30 A). Looking into the binding profiles described in the Figure 2.30 B, it is clear that this response from the model comes from the attempt to explain the apparent exclusion of EZH1 by the nucleosome H32 (H3K4me3K9acK18ac, H4 unmodified). The positive coefficient associated with H3K23ac cancels the effect of negative score on H3K18ac in all nucleosomes but H32. In this particular case, it is also clear that EZH1 follows a similar pattern to SUZ12 in most nucleosomes, but the intensity of the signal is much lower (Figure 2.30 B, as indicated by the sizes of the circles in the figure). This lower intensity likely causes the signal to disappear in certain nucleosomes creating a sparse pattern that we see in Figure 2.30 B. While the model provides a faithful representation of the data available to us it highlights that while the coefficients are fairly indicative of the overall behaviour of the protein in a general case, they should be interpreted within the context of the study design.

Coefficient relationship with protein domains

To profile the model's behaviour further we have asked whether any protein domains were significantly associated with the coefficients for a given modification. Specifically we have performed a set of t-tests testing that on average proteins containing the domain have higher (or lower) coefficients than the proteins which do not contain this domain. Figure 2.31 provides the results of this procedure.

Unsurprisingly, we see a strong association between HP1-associated chromo domains and H3K9me2/3. Similarly, Polycomb-associated domains show association with H3K27me3, while SWI/SNF as-

sociated domains have high H3K14ac association coefficients, as expected from the literature[5, 172]. While some of these domains are not direct binders of these modifications, we cannot distinguish direct binding of proteins from binding through association with another protein that might contain a direct binding domain.

Perhaps more mechanistically, the model identifies a negative correlation between Ankyryn repeat domains and H4K20me2/3, which is also established[34]. Methyl-CpG binding domains are predicted to associate strongly with DNA methylation. Unsurprisingly, bromodomains associate positively only with H3K14ac and H4K16ac, but rather surprisingly they do not associate with other acetylated residues and even show negative association with H3K27ac. BET subfamily of bromodomain-containing proteins, which contains an extra N-terminal domain (*NET_dom* in the figure) also shows this negative association with H3K27ac, albeit it associates positively with other acetylated residues H3K18ac, H4K8ac and H4K5ac/H4K12ac. The unexpected negative associations between bromodomains and H3K27ac are most likely due to the combination of the effects specific to study design (e.g. similar to the ones observed for EZH1), the bromodomain-associated preference to bind multiply acetylated residues, and the assumptions of the model.

In our study, H3K27ac only occurs either on its own, or in combination with at least four other acetylated factors (see Figure 2.1). Bromodomains are well known to target multiple nearby acetylated residues[29]. The imposed LASSO regularisation term, which aims to minimise non-zero predictors for a given protein is incapable to provide an accurate picture of such interactions, artificially associating only one of the acetylated residues with the target protein, and adjusting for the combinatorial effect by assigning the negative weights for the others. While this still approximates the protein binding patterns well (as indicated by high R^2 confidence scores), the coefficients cannot be interpreted directly.

Taken together, we show that the behaviour of a significant amount of chromatin binding proteins can be summarised by the linear combination of their responses to individual modifications. These responses therefore provide an unique modification signature (or motif) that defines the epigenetic reader. We show that the coefficients estimated by the model capture

the current knowledge of the literature, as well as accurately describe the binding profiles for the epigenetic readers which provides another useful view to the dataset.

We describe a few cases where the interpretation of the model's predictions may be misleading, suggesting that the model may be overfitting and would therefore require optimisation. In order to investigate the scale and significance of these shortcomings we therefore develop a similar model based on Chromatin immunoprecipitation sequencing (ChIP-seq) in chapter 3.

In such cases a different model may be more appropriate. For instance, the regulatory axes in FA are able to capture the behaviour of bromodomains more accurately as they associate only with the general acetylation and H4 acetylation factors in the *bentlerT* rotation (FA1 and FA4, see Figure B.2).

The linear model also captures only the main constituents of the complex modification signatures that are recognised by the readers. The model cannot therefore disentangle hierarchical binding events where certain modifications might influence binding to other modifications. Different models will be required to understand such complicated binding patterns.

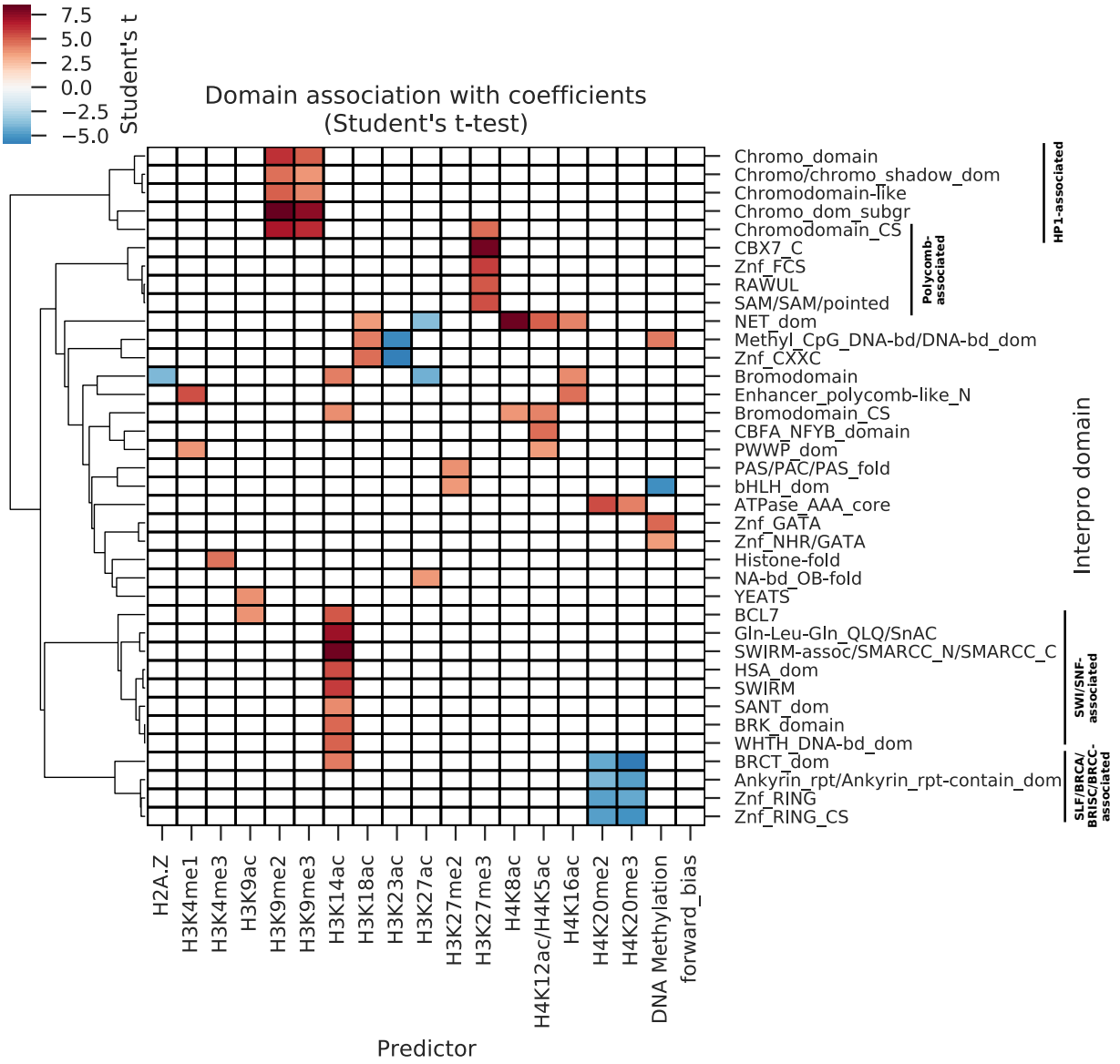

Figure 2.31: Association between predicted epigenetic reader motifs and Interpro domains. Proteins that possess these domains contain significantly higher (or lower) average coefficient values compared to proteins that do not possess this domain. (Student's t-test, Benjamini/Hochberg correction, FDR threshold 0.01)

2.4 Discussion

Epigenetic processes are strongly associated with nucleosome modifications. These modifications are thought to form a language, or a code, which can be read by epigenetic reader proteins that then further act, together or individually, to influence downstream regulatory events. While a lot is known about this regulation on a local (single Post-Translational Modification (PTM), single protein) scale, the global picture on how the combinatorial language of the nucleosome is translated still largely unclear.

SILAC nucleosome affinity purification (SNAP) experiments can directly measure the change in protein abundance bound to a modified nucleosome versus an unmodified one and therefore provide a very direct view of such regulatory interactions. In this chapter we have explored the data from 55 such SILAC nucleosome affinity purification (SNAP) experiments that have profiled a representative library of modified di-nucleosomes and the protein response associated with them. We have studied the relationships between epigenetic proteins that we observed in this data, as well as their response to the nucleosomal modifications in order to describe this epigenetic regulation and to provide new insights into the regulation of chromatin by DNA and histone modifications.

Specifically, we have established a data pre-processing method for a large scale combined analysis of multiple Stable Isotope Labelling with Amino acids in cell Culture (SILAC) affinity purifications, which reduces the biases in the data and augments it with information from public databases, and developed an interactive interface to display the results. We have then used the Context-Likelihood of Relatedness (CLR) algorithm, which we have further improved, to infer the regulatory network of epigenetic factors so that we can study the relationships between them. We have then used a Factor Analysis (FA) approach to elucidate the building blocks of the nucleosomal code captured by this experiment. Finally, we have designed a linear model which attempts to succinctly summarise the behaviour of epigenetic proteins in terms of linear combinations of responses to individual PTMs.

We showed that the network analysis recovers the key regulators of the epigenetic response

in the cells and relationships between them. While we did not identify new complexes, we confirmed many known interactions and identified novel ones. We have highlighted six high-confidence predictions of novel interactions that the network has identified to illustrate how the dataset can be used. The Factor Analysis approach has identified 11 key regulatory dimensions present in our dataset, eight of which we were able to directly link to the known constituents of the histone code. We have provided evidence that the remaining three axes act as a proxy to the accessibility of the nucleosomal DNA. Finally, we have shown that the response of 316 out of 492 proteins that were significantly enriched in at least 4 pull-downs could be explained by the linear combination of responses to individual modifications.

Collectively, these observations provide an unprecedented view of the nucleosome language on a global scale. The epigenetic reader interaction network provides a mechanistic link to the relationships between nucleosomal modifications, where a path could be traced between the reader and the writer proteins. The predicted novel subunits of chromatin complexes highlight the presence of potential missing links in chromatin biochemistry that need to be explored. The 11 regulatory dimensions described in our dataset highlight that reader responses to nucleosomal modifications is also low dimensional, reflecting the distribution of such modifications. Similarly these analyses identify the constituents of the epigenetic readouts and highlight the importance of DNA accessibility control. Finally, the linear model analysis provides a succinct description of the binding preferences for individual epigenetic readers. This model also shows that a significant proportion of epigenetic readers respond to nucleosome modifications in an additive fashion.

Together, the results described in this section establish an integrated knowledge resource about epigenetic reader proteins. We provide the information about their interactions, binding preferences as well as the building blocks of the nucleosomal code. The scale and scope of this resource is perhaps closest to the EpiFactors[39] database, which manually curated the literature to collect information about epigenetic proteins. Our approach, however, has the advantage of being completely data and experiment driven.

It has to be noted, however, that while we were able to recover the key regulatory dimensions

of chromatin, for instance the dimensions described in the factor analysis section, these findings are necessarily biased towards the original study design described in section 2.1.1 of the introduction. For instance, as evident from the bifactor rotation two factors – Factor 1, which corresponds to lysine acetylation, and Factor 2, which corresponds to the response of DNA methylation – describe 42.72% of variance observed our dataset (see section 2.2.4 and 2.3.5). While the low inherent dimensionality of histone PTM co-occurrence has previously been reported in Chromatin immunoprecipitation sequencing (ChIP-seq) datasets[72, 173], our results have to be interpreted in context of the nucleosome library analysed. Acetylation and DNA methylation each describe distinct groups of nucleosomes in our experiment (see Figure 2.1) and therefore should be expected to explain a large proportion of variance by design.

Instead, it is perhaps more interesting that eight of the eleven estimated regulatory dimensions identified by the factor analysis, can be easily characterised in terms of few nucleosomal modifications which have previously been described as chromatin landmarks. For instance, acetylation, DNA methylation, H3K9me2/3 or H3K27me2/3 are each represented by a factor. Together this suggests that linear response to regulatory signals is common among epigenetic proteins, which is generally not how scholars interpret the histone code hypothesis[26]. Interestingly, the response to factors shows seemingly hierarchical patterns, for instance proteins that respond to H3K14ac (factor 1) also respond to factor 3, which regulates H3K4me3 (Figure 2.25). This would suggest that acetylation, or in this case H3K14ac, establishes the context (active transcription), whereas H3K4me3 in this case serves as a further specifier (i.e. active enhancers vs promoters), which is consistent with our interpretation of histone code as a language and the literature[3]. Nevertheless, we also discover three factors with complex loadings, spanning non trivial combinations of histone modifications, out of which, factor 7 is perhaps closest in spirit to the histone code hypothesis, as it encodes a directional switch to the response to H4K16ac, when it occurs together with H3K4me1/3. We have suggested in this thesis that such factors may provide a proxy to DNA accessibility.

Given the result above, it is therefore not surprising that linear models which attempt to describe the protein binding behaviour, based on the histone PTM signature of the nucleosome, can predict the epigenetic reader behaviour well. Unfortunately, the predictions from the linear

model are also biased by the study design. We have previously illustrated how artefacts, such as H3K18ac/H3K23ac complementarity in terms of EZH1 binding could be introduced by the experiment design (see section 2.3.7). Perhaps less intuitively, since all subunits of large protein complexes behave nearly identically (see Figure 2.22 for instance) and we have benchmarked our analyses on proteins individually, larger protein complexes necessarily pose more weight than individual proteins and thus dominate the performance metrics. In order to truly *crack* the nucleosome code it would be desirable to reduce the aforementioned biases, by including a more combinations of histone PTMs on the nucleosomes, ideally with sufficient randomisation, as well as adjusting the data metrics for protein complexes.

Unfortunately, given the large space of potential combinations of nucleosomal modifications such a study is infeasible given the current experimental set-up. The current throughput could perhaps be increased by the use of other isotopic labelling reagents instead of SILAC. For instance, Tandem Mass Tag (TMT) labelling could allow up to 12 samples to be analysed simultaneously by the mass spectrometer, but with a reduced sensitivity. But even then, the nucleosome purifications themselves would provide the bottleneck. Tom Muir's group has pioneered the use of DNA barcoding to allow multiple nucleosomes to be analysed in a single experiment[174], but this approach is severely limited in the number of proteins that can be analysed in addition to the modified nucleosomes that need to be prepared and therefore unsuitable for our experiments.

At the same time, a completely randomised study design will necessarily result in a wasted effort as the majority of the nucleosomes studied will have no biological meaning, given that only a small subspace of chromatin marks that is utilised in vivo[26, 72, 173]. A solution that is somewhere in the middle is therefore desired, where the nucleosomes are designed with a small bias to make them biologically relevant, but with added randomness to aid the analysis.

We suggest that the data from Next-Generation Sequencing (NGS) experiments can provide the necessary blueprint for such further studies as it can be used as an indicator of what chromatin states exist *in vivo*, as well as their relative distribution. Similarly, ChIP-seq datasets of chromatin interacting proteins provide a complementary view of the SNAP data, which can

be used to benchmark the performance of some of the predictions.

In the next chapter we therefore establish a linear model based on ChIP-seq datasets which is conceptually equivalent to the one presented in this chapter. We compare the predictions of the two models and investigate their disagreements to investigate where and how the study design or the model could be improved. We further describe a ChIP-seq-based method to identify the chromatin states and quantify their distribution and relationship with epigenetic readers. This model can be used to guide the design of further SNAP studies, by introducing more nucleosomes indicative of the chromatin states, as well as provide insights into the epigenetic readers with multi-state binding behaviour.

Chapter 3

Next generation sequencing datasets

3.1 Introduction

In the previous chapter we have described an in-depth analysis of the SILAC nucleosome affinity purification (SNAP) dataset that has allowed us to query the nucleosomal language to understand how epigenetic readers interact with it. In this chapter we describe a proof-of-concept method that augments these observations with the insights obtained from Next-Generation Sequencing (NGS) datasets.

While it takes a vast amount of effort to produce a single Chromatin immunoprecipitation sequencing (ChIP-seq) dataset, the broad availability of published data make them essentially free to analyse. This availability makes ChIP-seq data a perfect *in vivo* companion to any *in vitro* experiment, including the SNAP experiment described in the previous chapter.

Specifically, we are interested in whether the chromatin state descriptions provided by the ChIP-seq datasets can be adjusted to focus not on the state at a given chromatin locus, which is what this data directly describes, but rather to give insights on the interactions of chromatin readers with these states. We would like to be able to elucidate a description of the binding preferences (in terms of nucleosomal modifications) of the chromatin reader proteins, like we did with SNAP dataset in section 2.3.7. At the same time we are interested in the frequency

and the structure of co-occurrence of histone Post-Translational Modifications (PTMs) and how does it influence the chromatin reader behaviour *in vivo*. We are particularly interested in how the observations made from this dataset compare to the work presented in the previous chapter and whether they could be used to validate and improve the models described previously.

We propose a linear Learning to rank (LETOR) model, which is similar to the linear model described for SNAP data in section 2.2.6, with the intent to predict the epigenetic reader interactions with the histone PTMs. We compare the predictions from this model with the predictions made from the SNAP dataset and investigate where the models disagree. We also propose a clustering method, which is inspired by the *ChromHMM* state discovery method[175] but is modified not to depend on the chromatin locus. We show that this model can help investigate how such predictions are affected by the *in vivo* distribution of the chromatin states as well as a blueprint for the future SNAP datasets.

In the discussion at the end of this chapter we propose how our chromatin clustering method could be extended to integrate with other Affinity Purification Mass Spectrometry (AP-MS) datasets, which, in theory, could provide a long sought link between the NGS and Mass Spectrometry (MS) datasets.

3.2 Methods

3.2.1 Pipeline for high-throughput ChIP-seq data processing

The ease of use of the NGS technology as well as the reduction in associated costs has sparked an exponential growth in the amount of the data available. While the post-processed data is generally published together with the raw sequencing reads, the advances in sequencing technology and genome assemblies, as well as software best practices have created a fragmented environment where one dataset may be analysed very differently from the other. The useful trend of re-processing the old data with new tools was established only recently. For instance, the latest release of the ENCODE[176] database includes the reprocessed data from an older

Roadmap release[52], but even with this addition, the ENCODE database to this day only contains the data obtained from the associated consortia and therefore lacks the large data corpus that is otherwise available in the Short Reads Archive (SRA). Therefore, the availability of the post-processed data still cannot be relied on, and thus the researchers are still required to process the data themselves.

Data processing requires a suitable pipeline to be selected. The ENCODE data processing pipeline is probably the most up to date, but its major focus is to maximise the information extracted from a given dataset. This approach is not suitable for comparative analyses as datasets with longer sequenced reads or larger sequencing depth would appear to be over-enriched at hard to align loci. In comparative analyses it is therefore necessary to reduce all datasets to the lowest denominator. The pipeline of the Roadmap project[52] has been designed to do just that, as the comparability between the datasets as it was the core priority of the project.

In this section we describe this pipeline and its reimplementation in our analysis. We use the Roadmap pipeline more or less verbatim, with a few minor updates. Our implementation is based on the luigi framework, which allows seamless integration with computing clusters running Open Grid Scheduler (OGS) for maximal throughput. We have also deposited it to GitHub[1] so other researchers can benefit from it too. The schematic for the implemented pipeline is shown in Figure 3.1.

We consider two major sources for ChIP-seq datasets: ENCODE and Short Reads Archive (SRA) as together they capture most, if not all, short read datasets. The latter database is somewhat hard to browse, and we instead recommend the use of the ReMap[177] project for discovery of such datasets as it provides a manually curated and accessible list of such annotations as well as a set of quality control metrics.

The selected datasets are downloaded and aligned to the *hg38* reference genome, using Bowtie 2[178]. We have substituted Bowtie 2 instead of Pash[179] aligner used in original Roadmap publication as we have found Pash processing to take significantly more time as it parallelises

[1] `https://github.com/lukauskas/chip-aligning-pipeline`

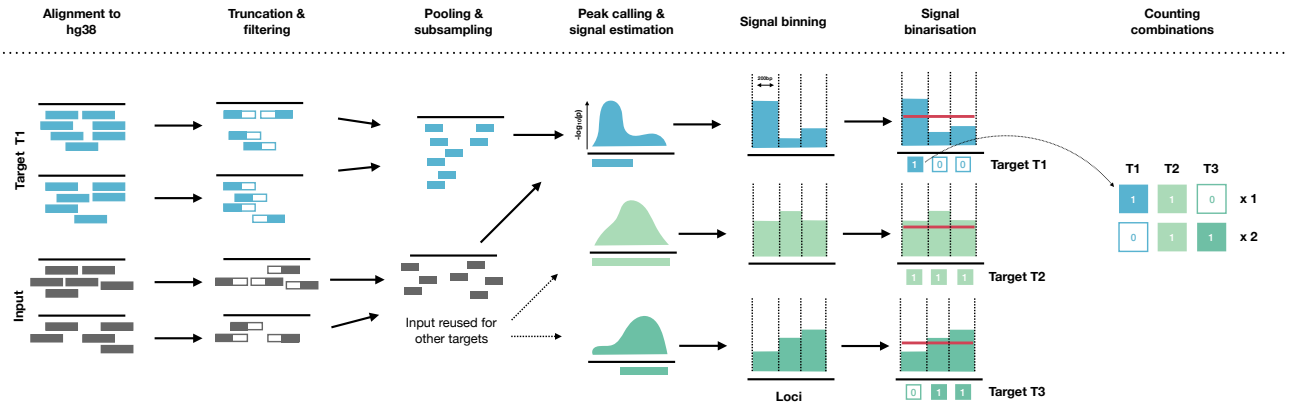

Figure 3.1: Schematic for the ChIP-seq pipeline used in this project. Briefly, the reads from publicly available datasets are aligned to the reference genome. The reads are then truncated to 36bp and samples from multiple data sources are pooled. The pooled samples are subsampled to the same number of reads to control for sequencing depth. MACS peak caller is then used to generate signal tracks representing the enrichment over input. This signal is then binned to 200bp windows across the genome. The binned signal is either considered as is, or thresholded to binary format to indicate presence/absence of signal. The patterns of signals across a genomic locus are grouped by similarity and counted.

poorly. Another popular aligner of short reads, *BWA*[180], has also been implemented in the pipeline, but our tests have shown it to be much slower than Bowtie 2 even when the analysis is run with the `--very-sensitive-local` setting which trades the efficiency for accuracy.

The aligned reads then undergo a filtering step which is identical to that of the Roadmap[52] publication. This step is attempts to control for discrepancy in the lengths of sequencing reads, which necessarily introduce a bias in mappability across the datasets as longer reads can be mapped to more regions of the genome. In this pipeline, the aligned reads are truncated to be 36 bp long (i.e. to the length of the shortest reads in the collection). Then, duplicated reads resulting after this truncation are removed (as they would be removed by the aligner). Finally, the reads that could not have been mapped to the reference genome, had they been 36 bp are removed with the help of genome mappability tracks. Such tracks for *hg19* reference genome were originally provided by the Roadmap project, but were then updated to *hg38* by others[181].

After the reads have been aligned and filtered, the data for replicates (both technical and biological) is pooled. In order to control for differences in library depth that this would achieve, the pooled datasets are randomly subsampled to contain no more than 30,000,000 reads each.

In Roadmap dataset this number varies between 30,000,000 and 50,000,000 reads, depending on the tier of the cell-type.

The aligned reads are then processed by MACS[182] peak caller, for which the fragment length parameter was estimated using phantompeakqualtools package[183] as described in the Roadmap pipeline. MACS is run in *bgdcmp* mode which generates a *signal* track for the genome, where each base pair is associated with a value that can be treated as a negative $\log_{10}(p)$ value of enrichment over background.

The genome is then divided into 200bp wide bins (approximately one nucleosome per bin), and the maximum signal value (i.e. $-log_{10}(p)$ of target versus input) falling in each of the bins is computed. Bins falling to the blacklisted regions of the genome[184][2] are removed. Bins that could not be completely mapped using the mappability criterion described previously, including bins whose 350 bp neighbourhood could not completely be mapped, are removed.

Depending on the application we then either: binarise the values in the bins, assuming that any bins with the signal greater than 2 (i.e. equivalent to p-value of 0.01) contain the target of interest, while others do not; or work with the real signal values, as described in the further sections. In the applications where the genomic windows are thresholded into binary format, we group windows with identical patterns together, computing how many times a particular pattern occurs across the whole genome and only work with such unique combinations of binary vectors.

3.2.2 Learning to rank the epigenetic reader binding preferences

In this section we are again interested in describing the epigenetic reader behaviour, based on the histone PTMs the readers recognise. We have already described how this could be achieved in SNAP datasets in section 2.2.6. In this section, we describe a comparable method developed for ChIP-seq datasets. Overall, we are interested in comparing the predictions from the two models in order to evaluate their performance as well as to gain insights on the differences

[2]See `https://sites.google.com/site/anshulkundaje/projects/blacklists` as well.

between *in vivo* (i.e. ChIP-seq) and *in vitro* (i.e. SNAP) regulation, as well as to obtain information about the importance of nucleosomal modifications that were not profiled in one dataset, or the other.

Unlike in SNAP, post-processed data from ChIP-seq experiments (section 3.2.1) does not directly provide us with a quantifiable readout of the amount of protein associated with a particular nucleosome. It is not accurate to assume that a signal stronger by a factor of ten is equivalent to ten times of the protein bound to the nucleosome. Instead, we can only assume that the regions associated with more reads are more likely to associate with the chromatin factor, without making any assumptions of by how much.

To make the data more similar to the data obtained from SNAP datasets, we treat the 200bp windows across the genome as individual nucleosomes marked with a combination of PTM marks. The exact combination of marks present in these loci are defined by binarisation procedure described in section 3.2.1, where the PTM is assumed to be present on this assumed nucleosome if the signal for the corresponding ChIP-seq track exceeds the threshold, and otherwise it is assumed to be absent. We then group the loci with identical nucleosomal PTM profiles together and compute the mean ChIP-seq signal for the epigenetic reader across these loci, resulting in a dataset shown in Figure 3.2 A.

Averaging of the signal across multiple loci makes the raw values of this signal even less quantitative than they were for individual loci. However, it is still valid to assume that "nucleosomes" associated with higher signal on average are more likely to contain the PTMs important for the recruitment of the epigenetic protein. We therefore treat this data as a ranking of nucleosomes from the most likely to associate with an epigenetic reader protein to the least likely ones, based on this average signal statistic. From this perspective, in order to predict the PTMs responsible for the recruitment (or exclusion) of an epigenetic reader protein, we need to predict this ranking of the nucleosomes from the modifications present on them. This turns our biological problem into a Learning to rank (LETOR) problem, which has been extensively studied in the information retrieval field.

It is important to note that while the work presented in this section should be attributed to

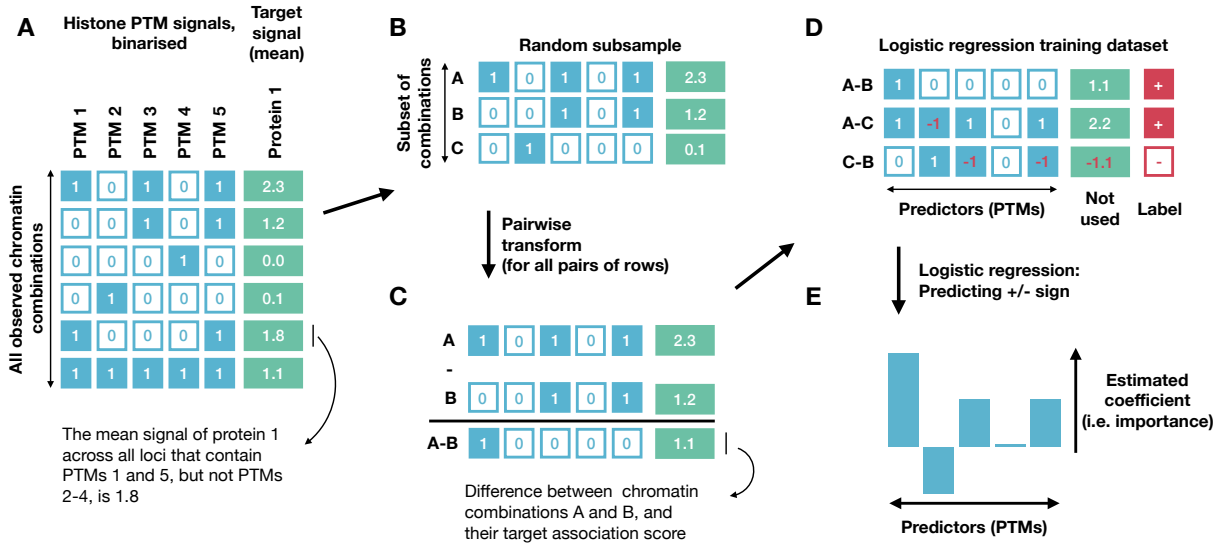

Figure 3.2: Illustration of the LETOR model for chromatin reader binding preferences from ChIP-seq data. (A) The input data format that is processed by the model. The columns of the dataset represent ChIP-seq tracks for histone PTMs analysed, while rows correspond to the unique (i.e. non-duplicated) combinations of the binarised PTMs patterns observed across the genome. The green column corresponds to the average signal value of the target protein across all loci that contain the particular pattern of PTMs associated with this row. (B) A random subsample of the dataset in (A). (C) Illustration of pairwise transform procedure, where the difference between all pairs of the subsampled data rows in (B) is computed. (D) The dataset resulting from the pairwise transformation of the dataset in (B). The pairwise differences between binary combinations of histone PTM signals is used as a set of predictors to the logistic regression classifier which aims to predict the sign of the pairwise difference of the target variable associated with these loci. Note that the absolute value of the difference is not considered, just the sign. (E) Illustration of the estimated coefficients from the logistic regression procedure. In this framework, the sign of the estimated coefficient is indicative of preference vs exclusion, while the absolute value of the coefficient is indicative of the importance of the feature

the thesis author, the work of an outstanding MSc student, Luis Fernando Camarillo Guerrero, advised by Dr. Peter DiMaggio and the author himself, should be acknowledged, as his work[185] has explored the early ideas which helped to shape this method.

Human H1-hESC, H9 and IMR-90 cell lines provide us with 731,623, 956,757 and 844,493 distinct combinations of histone PTMs across all of the 200bp windows in the genome respectively, providing us with nearly a million virtual nucleosomes to analyse per dataset. Out of these combinations 260,749, 282,008 and 196,477 (respectively) occur at least in two loci in the genome and even fewer occur in more locations.

The LETOR procedure we have utilised in this section computes the difference between all of the binary vectors representing nucleosomes in our dataset (resulting a vector of plus or minus

ones and zeroes indicating the difference in PTMs present) and uses it to predict the sign of the difference in target signals associated with the nucleosomes. Such transformation has been introduced by [186] as a way to transform a LETOR problem (in this case, ordinal regression) to a classification problem.

It is easy to see that such pairwise transformation uses $O(n^2)$ of memory, which makes it infeasible given our scale. We have chosen to counter this by randomly sub-sampling 5000 of the distinct nucleosome combinations randomly, weighing the probability of selection by how often this particular nucleosome occurs in genome. The pairwise transformation is then performed only on the nucleosomes in this sub-sampled data set (Figure 3.2 B-C). We repeat this sampling 150 times for each of the targets.

We use a logistic regression classifier to predict the sign of the target obtained via pairwise transformation from the difference in the predictors (PTMs, Figure 3.2 D-E). To maintain a balanced classification problem, we arbitrarily multiply the data rows by minus one (i.e. equivalent to subtracting B-A instead of A-B) so the target sign is positive in exactly half of the transformed data, and negative in the remaining half. The pairs of nucleosomes which have identical target association score and therefore a difference of zero are dropped.

We constrain the training of our classified with a L1 penalty as implemented in scikit-learn to match the Lasso penalty used in the modelling of SNAP data[3]. The best regularisation coefficient is chosen using a 4-fold cross validation from sample of 20 coefficients logarithmically spaced between (10^{-4} and 10^{4}). The best coefficient is used to retrain the data on the complete subsample (ignoring CV folds) and the results of this complete model are then analysed further.

For the analysis we have selected all the targets in the confident prediction set described in section 2.2.6 for which a ChIP-seq dataset from human *H1-hESC*, *H9* or *IMR-90* cell lines has been deposited into ENCODE. These three cell lines have been chosen as they provide the largest coverage of histone PTMs ChIP-seq data as compared to other established human cell lines. We have also included two datasets for BRD4 that have not been deposited to ENCODE,

[3]It is important to note that in scikit-learn implementation of Logistic regression, the meaning of the regularisation term is the exact inverse form the Lasso meaning of the regularisation term used in section 2.2.6

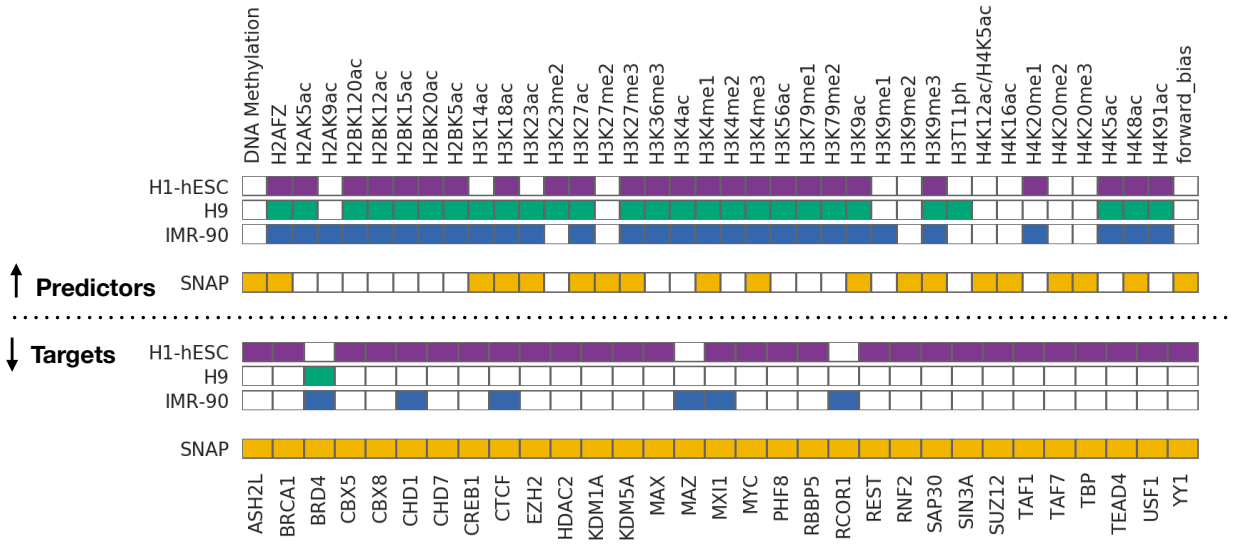

Figure 3.3: Summary of the dataset of ChIP-seq based predictions for protein interactions with histone PTMs. The top panel lists the predictor datasets available across the three cell lines studied. An overlap with predictors available in SNAP data is also provided for convenience. The bottom panel shows the availability of target datasets across the three cell lines. The target set has been chosen by the data availability of proteins in SNAP data (see section 2.3.7). It is clear that data for most of the targets is available in *H1-hESC* cell line only.

but were available from the SRA. This data comes from two studies: [187] and [188]. The latter study also provides a H3K27ac ChIP-seq dataset for IMR-90 cell-line which we have pooled with the equivalent data from ENCODE. The dataset is summarised in Figure 3.3. Accession codes are provided in the appendix (Table C.1).

3.2.3 Clustering chromatin states using mixtures of independent Bernoulli variables

The method described in the previous section attempts to establish the relationship between epigenetic reader proteins and the corresponding histone PTMs they recognise. In this section we instead focus on the relationship between such readers and chromatin states (i.e. groups of PTMs that frequently co-occur together), to determine how the co-occurrence patterns of PTMs impact these predictions. Specifically we are interested in whether chromatin readers interact with a single chromatin state, or whether they interact with multiple chromatin states that may or may not share a common modification pattern. We are also interested in how do

the chromatin states (and their distributions) differ across the datasets, and what impact does their distribution have on the performance of predictive models.

We describe a Bernoulli Mixture Model (BMM) which can simultaneously cluster ChIP-seq data from multiple cell lines, which is achieved by relaxing certain assumptions of the Bernoulli Mixture Models (BMMs) as well as providing a strategy to deal with missing data points. Like in the previous section, we describe our models in a way that is compatible with Mass Spectrometry (MS) data and briefly discuss this extension in the final section of the discussion.

Bernoulli distribution provides a natural way to represent the binarised ChIP-seq signals produced by the analysis pipeline described in section 3.2.1. The general assumption behind the term *chromatin state* is that the PTMs at any given locus are dependent on each other. For instance, presence of H3K4me1 is likely to be an indicator of H3K27ac as well: both of these marks are associated with active enhancers[189]. It is therefore natural to assume that the chromatin loci come from a set of states, which are not known to us in advance, each associated with a subset of chromatin factors. Examples of such states include enhancer, promoter, gene body regions each associated with distinct epigenetic signatures. Nearby regions of chromatin are likely to come from a similar state. Ernst & Kellis, have used a Hidden Markov Model (HMM) to capture such dependencies[175], their method titled ChromHMM is still being used to segment the chromatin today[52].

ChromHMM has inspired the method described in this section. We, however, treat all the loci in the genome independently as we did in the previous section (section 3.2.2). This independence assumption allows the chromatin states inferred from the model to be treated in a way that is more similar to MS datasets which do not have this positioning information. This, by definition, reduces the model to a mixture model formulation (thus the BMM name we use for the model). In other words, we are only interested in what chromatin states are present in the dataset, and not how they depend on genomic location.

From a generative perspective, mixture models, assume that each of the observed data points are associated with a hidden state that we cannot directly observe. The states are not related to each other, and are rather associated with each locus randomly via a process similar to a

dice roll. We refer to the probability distribution associated with this dice roll as the *mixing probabilities*.

Traditionally, one assumes one set of mixing probabilities per dataset. We, on the other hand, extend our model to allow simultaneous analysis of multiple datasets (representing different cell-types in ChIP-seq), each allowed to associate with their own set of mixing probabilities. Using the dice analogy again, our approach is equivalent of using a different (biased) dice for each of the data sources. The states themselves, however, are assumed to be the same across all of the data sources, only their distribution (i.e. number of loci in a particular state) are allowed to differ.

We refer to the probability of observing a histone PTM, given the state, as *emission probability*. While the mixing probabilities may differ from dataset to dataset, the emission probabilities are forced to be the same.

Unlike the LETOR model described in section 3.2.2, we treat the epigenetic proteins in the same way as PTMs in this model by binarising their signal and concatenating to the nucleosome PTM-associated vectors. The data from multiple datasets is pooled (ignoring the locus information), resulting in a dataset that is illustrated in Figure 3.4 A. In cases where no ChIP-seq data is present for a given PTM or epigenetic reader in a dataset, we treat it as missing data-point and probabilistically account for both possibilities of its value.

We refer to the combined dataset as $\mathcal{X} = \{\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_N}\}$ where each $\mathbf{x_i}$ represent a D-dimensional binary vector indicating the presence or absence of D PTMs and chromatin readers, with only some of the values observed (Figure 3.4 A). The N data vectors represent the combined (pooled) data from multiple cell-lines, the auxiliary variable t_i, associated with every point tracks which dataset the data point came from (1 of C, Figure 3.4 A). This information is needed to correctly estimate the mixing probabilities. The mixing probabilities themselves are encoded using the parameter π, while the K hidden states they generate are indicated with a random variable z, which is indexed same way as the loci (Figure 3.4 B). Emission probabilities are indicated via the parameter p, of which there are $K \times D$ values (Figure 3.4 B).

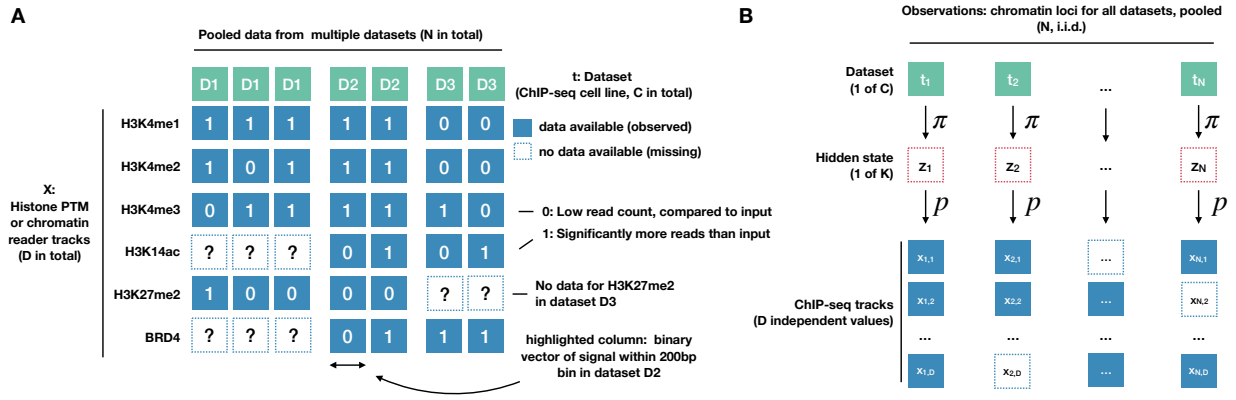

Figure 3.4: BMM model for establishing the interaction between chromatin states and epigenetic readers. (A) Description of the input dataset to the model. Each 200bp locus in the ChIP-seq dataset is treated as a binary vector indicating presence or absence of a histone PTMs (columns, N in total). Loci are treated independently. Cell lines are pooled together (datasets D1, D2 and D3 in the figure – one for each cell line analysed) as if they came from the same dataset. The exact relationships between the binary vectors and the cell lines are tracked via the $\mathbf{t}$ indicator variable, as it uniquely defines the π variable (panel B). The model considers an union of the ChIP-seq tracks available across the datasets. The ChIP-seq tracks that have not been explored in a given cell line are indicated as *missing* (question marks in the figure). (b) Schematic of the BMM model introduced in section 3.2.3. The data is split into C datasets, in this case corresponding to cell lines in the ChIP-seq dataset. Each locus in the genome (i.e. a 200bp bin) is then assumed to come from K hidden states (i.e. chromatin states), each of which associated with a probability π that is dependent on the state and dataset. These states are then assumed to emit D ChIP-seq *tracks* which are modelled as Bernoulli random variables with probability p dependent on the state and track. These tracks represent distinct ChIP-seq targets such as histones or proteins. Most of such *tracks* are observed (i.e. we have the data for), but the model allows for some of these data points to be missing so datasets with different sets of chromatin marks can be combined.

Formal definition

By definition mixture models assume that the data comes from a set of K possible hidden states (in this case: enhancers, promoters, etc.), indicated by the hidden variable Z. Each chromatin locus i (from dataset pooled across the cell lines) is assumed to be associated with a single state z_i. Probability that an i-th data-point comes from a hidden state is defined in terms of categorical distribution, defined by the vector of mixing proportions $\pi = [\pi_1, \pi_2, \ldots, \pi_K]$:

$$P(\mathbf{z_i}|\pi) = \prod_{k=1}^{K} \pi_k^{z_{i,k}} \quad \text{s.t} \sum_{l=1}^{K} \pi_l = 1 \quad \text{for each pooled chromatin locus } i = 1 \ldots N \tag{3.1}$$

Here we made use of 1-of-K encoding of the hidden state membership variable. In this case each $\mathbf{z_i}$ is a binary vector of length K whose all elements are set to zero, except one. The data

point i is assumed to be generated by the state k if and only if $z_{i,k} = 1$. Using this notation $\mathbf{z}$ is a $N \times K$ dimensional matrix containing one such row vector for each of the N observations.

In a classic mixture model setting, only one such vector π would be associated with the complete dataset. But, as we have already described, we pool the data from C datasets (one dataset per cell line). Each dataset c in our model is allowed to have its own set of mixing proportions $\pi_{\mathbf{c}}$. We create an indicator variable $\mathbf{t}$ to mark which dataset a data point was produced from (see Figure 3.4). This variable is also encoded in one-of-C format, turning it into a $N \times C$ matrix, similarly to how $\mathbf{z}$ has been encoded. We modify the equation above to incorporate this:

$$\begin{aligned} P(\mathbf{z_i}|\pi, \mathbf{t_i}) &= \prod_{c=1}^{C} P(\mathbf{z_i}|\pi_{\mathbf{c}})^{t_{i,c}} \\ &= \prod_{c=1}^{C} \left[\prod_{k=1}^{K} \pi_{c,k}^{z_{i,k}} \right]^{t_{i,c}} \qquad \text{s.t} \sum_{l=1}^{K} \pi_{c,l} = 1 \quad \text{for all } c = 1 \dots C \end{aligned} \tag{3.2}$$

Where we have changed the notation slightly to convert π into a $C \times K$ matrix, whose c-th row lists the mixing coefficients for the c-th dataset.

The conditional distribution for $\mathbf{x_i}$, which represents whether signal for a specific histone PTM or chromatin reader (of which there are D) is present at chromatin locus i, is given as:

$$\begin{aligned} P(\mathbf{x_i}|\mathbf{z_i}, \mathbf{p}) &= \prod_{k=1}^{K} P(\mathbf{x_i}|\mathbf{p_k})^{z_{i,k}} \qquad \text{for all } i = 1 \dots N \\ &= \prod_{k=1}^{K} \left[\prod_{d=1}^{D} p_{k,d}^{x_{i,d}} \left(1 - p_{k,d}\right)^{1 - x_{i,d}} \right]^{z_{i,k}} \qquad \text{for all } i = 1 \dots N \end{aligned} \tag{3.3}$$

Here $\mathbf{p}$ is a $K \times D$ matrix of Bernoulli emission probabilities for each of the hidden states and data dimensions. Specifically, $p_{k,d}$ represents the probability of observing signal from the d-th ChIP-seq track (i.e. PTM or epigenetic protein) from the state k.

It should also be noted that it is common for some of the cell lines c to lack some of the D

histone PTMs or chromatin reader proteins. To address this the histone PTMs and chromatin readers are split into observed and hidden (i.e. missing) subsets:

$$\mathcal{X} = \{\mathcal{X}_o, \mathcal{X}_h\} \text{ and equivalently,}$$
$$\mathbf{x_i} = \{\mathbf{x_{o,i}}, \mathbf{x_{h,i}}\} \text{ for all } \mathbf{x_i} \in \mathcal{X} \tag{3.4}$$

We assume that the missing values, $\mathcal{X}_h$ are missing at random, i.e. the probability of observing them is not dependent from their value. This assumption could be made less general by conditioning the missing values on the dataset indicator variable $\mathbf{t}$, but we do not see it necessary.

The hidden variables factorise Equation 3.3 as:

$$P(\mathbf{x_i}|\mathbf{z_i}, \mathbf{p}) = \prod_{k=1}^{K} \left[\prod_{d:x_{i,d} \in \mathcal{X}_o} \left(p_{k,d}^{x_{i,d}} \left(1 - p_{k,d}\right)^{1-x_{i,d}} \right) \times \prod_{e:x_{i,e} \in \mathcal{X}_h} \left(p_{k,e}^{x_{i,e}} \left(1 - p_{k,e}\right)^{1-x_{i,e}} \right) \right]^{z_{i,k}} \tag{3.5}$$

The joint distribution of all of the data points can now be expressed as:

$$\begin{aligned} P(\mathcal{X}, \mathbf{z}|\pi, \mathbf{p}, \mathbf{t}) &= \prod_{i=1}^{N} P(\mathbf{z_i}|\pi, \mathbf{t_i}) \times P(\mathbf{x_i}|\mathbf{z_i}, \mathbf{p}) \\ &= \prod_{i=1}^{N} \prod_{k=1}^{K} \left[P(\mathbf{z_i} = k|\pi, \mathbf{t}) \times \prod_{d:x_{i,d} \in \mathcal{X}_o} P(x_{i,d}|p_{k,d}) \times \prod_{e:x_{i,e} \in \mathcal{X}_h} P(x_{i,e}|p_{k,e}) \right]^{z_{i,k}} \\ &= \prod_{i=1}^{N} \prod_{k=1}^{K} \left[\prod_{c=1}^{C} \pi_{c,k}^{t_{i,c}} \times \prod_{d:x_{i,d} \in \mathcal{X}_o} \left(p_{k,d}^{x_{i,d}} \left(1 - p_{k,d}\right)^{1-x_{i,d}} \right) \times \prod_{e:x_{i,d} \in \mathcal{X}_h} \left(p_{k,e}^{x_{i,e}} \left(1 - p_{k,e}\right)^{1-x_{i,e}} \right) \right]^{z_{i,k}} \end{aligned} \tag{3.6}$$

Where $\mathbf{z_i} = k$ indicates that $z_{i,k} = 1$ in the 1-of-K encoding.

We can marginalise the unobserved variables $\mathcal{X}_h$ and $\mathbf{z}$ from this joint distribution to obtain

the likelihood for the data.

$$
\begin{aligned}
P(\mathcal{X}_o|\pi,\mathbf{p},\mathbf{t}) &= \sum_{\mathbf{z}}\sum_{\mathcal{X}_h} P(\mathcal{X},\mathbf{z}|\pi,\mathbf{p},\mathbf{t}) \\
&= \prod_{i=1}^{N}\sum_{k=1}^{K}\left[P(\mathbf{z_i}=k|\pi,\mathbf{t_i}) \times \prod_{d:x_{i,d}\in\mathcal{X}_o} P(x_{i,d}|p_{k,d}) \times \prod_{e:x_{i,d}\in\mathcal{X}_h} \underbrace{\sum_{y\in\{0,1\}} P(x_{i,e}=y|p_{k,e})}_{=1}\right] \\
&= \prod_{i=1}^{N}\sum_{k=1}^{K}\left[P(\mathbf{z_i}=k|\pi,\mathbf{t_i}) \times \prod_{d:x_{i,d}\in\mathcal{X}_o} P(x_{i,d}|p_{k,d})\right] \\
&= \prod_{i=1}^{N}\sum_{k=1}^{K}\left[\prod_{c=1}^{C}\pi_{c,k}^{t_{i,c}} \times \prod_{d:x_{i,d}\in\mathcal{X}_o}\left(p_{k,d}^{x_{i,d}}\left(1-p_{k,d}\right)^{1-x_{i,d}}\right)\right]
\end{aligned}
\tag{3.7}
$$

In order to simplify the equations, as well as to provide a performance improvement in the implementation, we define a *support* variable $s_{i,k}$, similarly to how it is defined in [190]:

$$
\begin{aligned}
s_{i,k} &\equiv P(\mathbf{z_i}=k|\pi,\mathbf{t_i}) \times \prod_{d:x_{i,d}\in\mathcal{X}_o} P(x_{i,d}|p_{k,d}) \\
&= \prod_{c=1}^{C}\pi_{c,k}^{t_{i,c}} \times \prod_{d:x_{i,d}\in\mathcal{X}_o}\left(p_{k,d}^{x_{i,d}}\left(1-p_{k,d}\right)^{1-x_{i,d}}\right)
\end{aligned}
\tag{3.8}
$$

Using this notation the joint distribution and the data likelihood simplifies to the following equations:

$$
P(\mathcal{X},\mathbf{z}|\pi,\mathbf{p},\mathbf{t}) = \prod_{i=1}^{N}\prod_{k=1}^{K}\left[s_{i,k} \times \prod_{e:x_{i,d}\in\mathcal{X}_h}\left(p_{k,e}^{x_{i,e}}\left(1-p_{k,e}\right)^{1-x_{i,e}}\right)\right]^{z_{i,k}} \tag{3.9}
$$

$$
P(\mathcal{X}_o|\pi,\mathbf{p},\mathbf{t}) = \prod_{i=1}^{N}\sum_{k=1}^{K} s_{i,k} \tag{3.10}
$$

Training of the model

We cannot optimise the observed data likelihood (Equation 3.10) directly as the summation over K hidden states will remain in the logarithm. Instead, as described extensively in the textbooks, for instance [114–116, 191], Expectation Maximisation (EM) algorithm can be used to iteratively optimise the logarithm of complete data likelihood (Equation 3.9) instead. While the optimisation algorithm for conventional BMMs has been described previously in any of the publications cited above, for convenience, we reiterate this procedure here describing the extension for missing data and multiple datasets. The expression for the complete data log-likelihood l_c is:

$$\begin{aligned}
l_c &\equiv \log P(\mathcal{X}, \mathbf{z}|\pi, \mathbf{p}, \mathbf{t}) \\
&= \sum_{i=1}^{N}\sum_{k=1}^{K} z_{i,k}\left[\sum_{c=1}^{C}\left(t_{i,c}\times\log(\pi_{c,k})\right) + \sum_{d=1}^{D}\left(x_{i,d}\log(p_{k,d}) + (1-x_{i,d})\times\log(1-p_{k,d})\right)\right] \\
&= \sum_{i=1}^{N}\sum_{k=1}^{K} z_{i,k}\left[\log(s_{i,k}) + \sum_{e:x_{i,e}\in\mathcal{X}_h}\left(x_{i,e}\log(p_{k,e}) + (1-x_{i,e})\times\log(1-p_{k,e})\right)\right] \\
&= \sum_{i=1}^{N}\sum_{k=1}^{K}\left[z_{i,k}\times\log(s_{i,k}) + \sum_{e:x_{i,e}\in\mathcal{X}_h}\left(z_{i,k}\times x_{i,e}\times\log(p_{k,e}) + z_{i,k}\times(1-x_{i,e})\times\log(1-p_{k,e})\right)\right]
\end{aligned} \tag{3.11}$$

Since we do not know the values of $\mathbf{z}$, nor do we know the values of $\mathcal{X}_h$, the l_c cannot be computed. Instead, we rely on the expected value for this log-likelihood:

$$\begin{aligned}
&\mathbb{E}[l_c] \\
&= \sum_{i=1}^{N}\sum_{k=1}^{K}\left[\mathbb{E}[z_{i,k}]\times\log(s_{i,k}) + \sum_{e:x_{i,e}\in\mathcal{X}_h}\mathbb{E}[z_{i,k}\times x_{i,e}]\times\log(p_{k,e}) + \mathbb{E}[z_{i,k}\times(1-x_{i,e})]\times\log(1-p_{k,e})\right]
\end{aligned} \tag{3.12}$$

This expectation operator $\mathbb{E}$ is bound upon the distribution of hidden data, given the observed one. In the Expectation Maximisation (EM) algorithm we would compute this probability and

the resulting expectation using parameters in the previous iteration (E-step), and then aim to obtain the new set of parameters in the next iteration (M-step) via optimisation. We therefore substitute parameters with their *dashed* versions, e.g. $\mathbf{p} \rightarrow \mathbf{p}'$, to indicate the parameters obtained in the *previous* iteration. The probability that is necessary to obtain the expectation can then be expressed as:

$$P(\mathcal{X}_h, \mathbf{z}|\mathcal{X}_o, \pi', \mathbf{p}', \mathbf{t}) = \frac{P(\mathcal{X}_h, \mathcal{X}_o, \mathbf{z}|\pi', \mathbf{p}', \mathbf{t})}{P(\mathcal{X}_o|\pi', \mathbf{p}', \mathbf{t})} = \frac{P(\mathcal{X}, \mathbf{z}|\pi', \mathbf{p}', \mathbf{t})}{\sum_{\mathcal{X}_h} \sum_{\mathbf{z}} P(\mathcal{X}_h, \mathcal{X}_o, \mathbf{z}|\pi', \mathbf{p}', \mathbf{t})} \tag{3.13}$$

The numerator for this equation, as well as the denominator have already been determined in the Equations 3.9 and 3.10 respectively:

$$P(\mathcal{X}_h, \mathbf{z}|\mathcal{X}_o, \pi', \mathbf{p}', \mathbf{t}) = \frac{\prod_{i=1}^{N} \prod_{k=1}^{K} \left[{s'}_{i,k} \times \left(\prod_{e \notin \Omega_i} {p'}_{k,e}^{x_{i,e}} \left(1 - {p'}_{k,e}\right)^{1-x_{i,e}} \right) \right]^{z_{i,k}}}{\prod_{i=1}^{N} \sum_{l=1}^{K} {s'}_{i,l}} \tag{3.14}$$

We note that only three expectation terms are needed for the expected log-likelihood defined in Equation 3.12, which can be expressed as:

$$\mathbb{E}[z_{i,k}] = 1 \times P(\mathbf{z_i} = k|\mathbf{x_{o,i}}, \pi, \mathbf{p}, \mathbf{t_i} = c) + 0 \tag{3.15}$$

$$\mathbb{E}[z_{i,k} \times x_{i,e}] = (1 \times 1) \times P(\mathbf{z_i} = k, x_{i,e} = 1|\mathbf{x_{o,i}}, \pi', \mathbf{p}', \mathbf{t_i} = c) + 0 \tag{3.16}$$

$$\mathbb{E}[z_{i,k} \times (1 - x_{i,e})] = (1 \times 1) \times P(\mathbf{z_i} = k, x_{i,e} = 0|\mathbf{x_{o,i}}, \pi', \mathbf{p}', \mathbf{t_i} = c) + 0 \tag{3.17}$$

The probabilities required for this likelihood can be computed by observing that Equation 3.14 factorises over k, i and e:

$$\begin{aligned} P(\mathcal{X}_h, \mathbf{z}|\mathcal{X}_o, \pi', \mathbf{p}', \mathbf{t}) &= \prod_{i=1}^{N} \prod_{k=1}^{K} P(\mathbf{x_{h,i}}, \mathbf{z_i} = k|\mathbf{x_{o,i}}, \pi', \mathbf{p}', \mathbf{t})^{z_{i,k}} \\ &= \prod_{i=1}^{N} \prod_{k=1}^{K} \left[\frac{{s'}_{i,k}}{\sum_{l=1}^{K} {s'}_{i,l}} \left(\prod_{e \notin \Omega_i} {p'}_{k,e}^{x_{i,e}} \left(1 - {p'}_{k,e}\right)^{1-x_{i,e}} \right) \right]^{z_{i,k}} \end{aligned} \tag{3.18}$$

Note that we used the observation that product $\prod_{k=1}^{K} [\cdot]^{z_{i,k}}$ will be equal to 1 in all but one of the K iterations (due to one-of-k encoding of $\mathbf{z}$), to be able to move the normalisation constant $\sum_{l=1}^{K} {s'}_{i,l}$ into it.

Marginalising all of the missing datapoints associated with entry i ($\mathbf{x_{h,i}}$) across each of the terms, we obtain the expression for the responsibility of state k generating observation i:

$$\begin{aligned} P(\mathbf{z_i} = k|\mathbf{x_{o,i}}, \pi', \mathbf{p}', \mathbf{t_i} = c) &= \sum_{\mathbf{x_{h,i}}} P(\mathbf{x_{h,i}}, \mathbf{z_i} = k|\mathbf{x_{o,i}}, \pi', \mathbf{p}', \mathbf{t_i} = c) \\ &= \frac{{s'}_{i,k}}{\sum_{l=1}^{K} {s'}_{i,l}} \times \prod_{e:x_{i,e} \in \mathcal{X}_h} \sum_{x_{i,e} \in \{0,1\}} \left({p'}_{k,e}^{x_{i,e}} \left(1 - {p'}_{k,e}\right)^{1-x_{i,e}} \right) \\ &= \frac{{s'}_{i,k}}{\sum_{l=1}^{K} {s'}_{i,l}} \end{aligned} \tag{3.19}$$

We can similarly marginalise all but one missing variable in $\mathbf{x}_{h,i}$ to obtain:

$$\begin{aligned} &P(\mathbf{z_i} = k, x_{i,e}|\mathbf{x_{o,i}}, \pi', \mathbf{p}', \mathbf{t_i} = c) \\ &= \frac{{s'}_{i,k}}{\sum_{l=1}^{K} {s'}_{i,l}} \times \left({p'}_{k,e}^{x_{i,e}} \left(1 - {p'}_{k,e}\right)^{1-x_{i,e}} \right) \times \prod_{f:x_{i,f} \in \mathcal{X}_h, f \neq e} \sum_{x_{i,f} \in \{0,1\}} \left({p'}_{k,f}^{x_{i,f}} \left(1 - {p'}_{k,f}\right)^{1-x_{i,f}} \right) \\ &= \frac{{s'}_{i,k}}{\sum_{l=1}^{K} {s'}_{i,l}} \times \left({p'}_{k,e}^{x_{i,e}} \left(1 - {p'}_{k,e}\right)^{1-x_{i,e}} \right) \end{aligned} \tag{3.20}$$

Therefore the three expected values required for Equation 3.12 are:

$$\begin{aligned} \mathbb{E}[z_{i,k}] &= 1 \times P(\mathbf{z_i} = k|\mathbf{x_{o,i}}, \pi, \mathbf{p}, \mathbf{t_i} = c) + 0 \\ &= \frac{{s'}_{i,k}}{\sum_{l=1}^{K} {s'}_{i,l}} \end{aligned} \tag{3.21}$$

$$\begin{aligned} \mathbb{E}[z_{i,k} \times x_{i,e}] &= (1 \times 1) \times P(\mathbf{z_i} = k, x_{i,e} = 1|\mathbf{x_{o,i}}, \pi', \mathbf{p}', \mathbf{t_i} = c) + 0 \\ &= {p'}_{k,e} \times \frac{{s'}_{i,k}}{\sum_{l=1}^{K} {s'}_{i,l}} \end{aligned} \tag{3.22}$$

$$\begin{aligned} \mathbb{E}[z_{i,k} \times (1 - x_{i,e})] &= (1 \times 1) \times P(\mathbf{z_i} = k, x_{i,e} = 0|\mathbf{x_{o,i}}, \pi', \mathbf{p}', \mathbf{t_i} = c) + 0 \\ &= (1 - p'_{k,e}) \times \frac{{s'}_{i,k}}{\sum_{l=1}^{K} {s'}_{i,l}} \end{aligned} \tag{3.23}$$

Defining two other variables for simplicity (the former of which is also inspired by the notes in

[190]):

$$z^*_{i,k} \equiv \mathbb{E}[z_{i,k}] \tag{3.24}$$

$$x^*_{i,k,d} \equiv \begin{cases} x_{i,d} & \text{if } x_{i,d} \in \mathcal{X}_o \\ p'_{k,d} & \text{otherwise} \end{cases} \tag{3.25}$$

The auxiliary function Q, i.e. the expected complete log likelihood of the data given the previous set of parameters, can then be expressed as:

$$Q(\theta, \theta') \equiv \mathbb{E}[l_c] = \sum_i^N \sum_c^C t_{i,c} \left(\sum_k z^*{}_{i,k} \log \pi_{k,c} + \sum_d^D \left(z^*_{i,k}\ x^*_{i,k,d} \log p_{k,d} + z^*_{i,k}\ (1 - x^*_{i,k,d}) \log(1 - p_{k,d}) \right) \right) \tag{3.26}$$

The M-step updates are then obtained by maximising Q with respect to π and $\mathbf{p}$:

$$\pi_{k,c} \leftarrow \frac{\sum_{i=1}^N t_{i,c}\ z^*_{i,k}}{\sum_{j=1}^N t_{j,c}} \tag{3.27}$$

$$p_{k,d} \leftarrow \frac{\sum_{i=1}^N z^*_{i,k}\ x^*_{i,k,d}}{\sum_{j=1}^N z^*_{j,k}} \tag{3.28}$$

Computing and maximising the auxiliary function iteratively ultimately results in the maximisation of the observed data likelihood. The proof for this is beyond the scope of this thesis, but is defined in any of the classic textbooks, e.g. [114–116, 191].

Maximum a posteriori approach

In the previous section we have derived, and optimised BMM model using a frequentist formulation i.e. optimising the parameters using the likelihood of the data directly. To prevent the zero-probability problem, as well as to give extra control on the model, it is beneficial to apply a Bayesian prior on the model's parameters π and $\mathbf{p}$. The use of conjugate priors allows

to optimise the parameters in a maximum aposteriori approach using the EM directly. The only change necessary involves the maximisation of $Q(\theta, \theta') + \log(p(\theta))$ (where θ represents the collection of model's parameters) in the M step, instead of the Q function directly[115].

The conjugate priors for the parameters π and $\mathbf{p}$ come from Dirichlet and Beta probability families respectively:

$$P(\pi_{\mathbf{c}}|\gamma) = \text{Dirichlet}(\pi_{\mathbf{c}}|\gamma) \propto \prod_{k=1}^{K} \pi_{c,k}^{\gamma_k - 1} \tag{3.29}$$

$$P(p_{k,d}|\alpha, \beta) = \text{Beta}(p_{k,d}|\alpha, \beta) \propto p_{k,d}^{\alpha-1} \times (1 - p_{k,d})^{\beta-1} \tag{3.30}$$

The corresponding M-step update in this case becomes:

$$\pi_{k,c} \leftarrow \frac{\gamma_k - 1 + \sum_{i=1}^{N} t_{i,c}\ z^*_{i,k}}{\left(\sum_{j=1}^{N} t_{j,c}\right) + \left(\sum_{l}^{K} \gamma_l\right) - K} \tag{3.31}$$

$$p_{k,d} \leftarrow \frac{\alpha - 1 + \sum_{i=1}^{N} z^*_{i,k}\ x^*_{i,k,d}}{\left(\sum_{l=1}^{L} \alpha + \beta\right) + \left(\sum_{j=1}^{N} z^*_{j,k}\right) - 2K} \tag{3.32}$$

Intuitively, these priors add a pseudocount of $\gamma_k - 1$ samples coming from the k-th state, as well as a pseudocount of $\alpha - 1$ of positive values observed in each dimension, for each state. Similarly a pseudocount of $\beta - 1$ negative values observed. In other words, this parametrisation is equivalent to Laplace smoothing.

Here we assumed that π parameters for all datasets, equivalently p parameters for all states, share the same set of hyperparameters, but with trivial changes the expressions above can be updated to allow the datasets to have unique priors.

In our experiments we have trained the models by assuming $\gamma_k = 1000$, and $\alpha = \beta = 10$, intuitively adding a pseudocount of 999 loci observed at each state, and 9 positive (as well as 9 negative) examples for each of the chromatin marks at each of the states.

Selecting the number of hidden states

The formal definition of the model above, requires the number of hidden states K to be selected *a priori*. In our application, however, no reasonable guess for this number can be made. Instead, the number of the hidden states can be estimated using Bayesian Information Criterion (BIC) statistic, which is based on the data likelihood:

$$\mathrm{BIC}(K) = -2 \log P(\mathcal{X}_o | K, \hat{\pi}, \hat{\mathbf{p}}) + \nu_K \log(N) \tag{3.33}$$

where ν_K is the number of free parameters in the model with K states.

Another approach, proposed by Biernacki et al.[192] suggests a BIC-like statistic defined on the complete log likelihood instead:

$$\mathrm{ICL}(K) = -2 \log P(\mathcal{X}_o, \hat{\mathcal{X}}_h, \hat{\mathbf{z}} | K, \hat{\pi}, \hat{\mathbf{p}}) + \nu_K \log(N) \tag{3.34}$$

Where ICL stands for Integrated Complete Likelihood and $\hat{\mathcal{X}}_h$ and $\hat{z}$ are the maximum-likelihood estimates for the hidden variables:

$$\hat{z}_i = \mathrm{argmax}_{k=1}^{K} \; P(\mathbf{z_i} = k | \mathbf{x_{o,i}}, \hat{\pi}, \hat{\mathbf{p}}, \mathbf{t_i} = c) = \mathrm{argmax}_{k=1}^{K} \; \hat{z}^*_{i,k} = \mathrm{argmax}_{k=1}^{K} \hat{s}_{i,k} \tag{3.35}$$

$$\hat{x}_{i,e} = \mathrm{argmax}_{x_{i,e}} P(x_{i,e} | \hat{\mathbf{p}}, \hat{\mathbf{z}}_\mathbf{i} = k) = \begin{cases} 1 & \text{if } p_{k,e} \geq 0.5 \\ 0 & \text{otherwise} \end{cases} \tag{3.36}$$

This latter approach is becoming more popular nowadays. This method penalises the maximum likelihood by the entropy of information in the hidden states, which in turn penalises the models whose hidden dimensions are similar, and therefore selects for models with fewer states, see [193, 194] for instance. We observe this in our experiments as well (see Figure B.10) and therefore use this method.

Model training for the ChIP-seq data analysis

We have implemented the model described above in Python, with key computations implemented in C for efficiency. The implementation is freely available on GitHub[4].

The raw ChIP-seq data for the model has been pre-processed as described in section 3.2.1. The model has been trained on the data available from the same three cell lines used in section 3.2.1. We have included all the histone modifications for which the ChIP-seq data was available into the model, as well as BRD4, CBX5, CHD1, CHD7, CTCF, HDAC2, KDM1A, YY1 and POLR2A chromatin-interacting proteins. We have also included the S5-phosphorylated (i.e. initiated) POLR2A for reference.

The models have been trained for the three cell lines in combination, as well as each of these cell-lines individually. As described earlier, we chose strict priors for the mixing and emission probabilities ($\gamma = 1000$, $\alpha = \beta = 10$) to regulate the model. In order to reduce the runtime the models were only trained using the chromatin modification patterns that are present in at least 30 loci of the genome, discarding the rest. The models have been trained 10 times, each initialised differently using the random prototypes technique[195, Eq. 19]. The models were trained until convergence, or up to 20,000 iterations of the EM algorithm, which ever happened earlier. The convergence criterion was triggered when log-likelihood of the data decreased by less than 0.01 in a single iteration.

3.3 Results

3.3.1 ChIP-seq based models provide a complementary view of epigenetic reader binding profiles

At the beginning of this chapter (section 3.2.2), we have discussed a LETOR model developed for the ChIP-seq data, similar to the *LASSO* model described for SNAP dataset in section 2.3.7.

[4] `https://github.com/lukauskas/bernoulli-mixture-model`

Both of these models aim to characterise the epigenetic reader binding preferences in terms of linear combination of nucleosomal modifications they interact with.

These models, however, fundamentally differ in terms of the context which is being described. The LETOR model, described for the ChIP-seq data estimates the relationships observed in *in vivo*, whereas the *LASSO* model necessarily describes only the results obtained in the particular *in vitro* study. Given the success of the *LASSO* model in describing the binding profiles of hundreds of proteins *in vitro* (section 2.3.7), we are interested in whether these results translate to *in vivo* studies.

To answer these questions, we have trained the LETOR model on the data for 31 epigenetic reader proteins, in three cell lines (see Figure 3.3) and compared the model's results with those obtained from SNAP dataset. To quantify the level of agreement between the two models, we have computed the correlation between the estimated interaction coefficients predicted by these models. Since the large degree of predictors for protein interactions do not overlap across the datasets (e.g. we have no H3K9me2 ChIP-seq data), we have only computed the correlation across the predictor pairs that overlap.

This procedure has shown that the predictions between the models to match for six of the assayed proteins: PHF8, CBX5, SIN3A, SAP30, CHD1 and TAF1 which exhibit significant correlation between the two models (Benjamini/Hochberg adjusted one-sided t-test p-value for Pearson correlation less than 0.05, see Table B.2). The estimated coefficient profiles for these proteins (Figure 3.5) are in agreement with the literature. For instance, PHF8 dependence on H3K4me3 has been well described[196]. Similarly, CBX5 is known to interact with H3K9me2 and H3K9me3[168] both of which are predicted from SNAP data, but only the latter from ChIP-seq due to lack of H3K9me2 ChIP-seq studies.

On the other hand, the predictions for the majority of the proteins analysed do not match across the data sources. We have initially assumed this to be due to the low coverage of the random sampling procedure which was necessary to fit the pairwise-transformed dataset into memory (see section 3.2.2). Judging from the heatmap of all predictions (Figure B.8), H3K4me, H3K27me3 and H4K8ac are clearly over-represented as the predictors, which could also suggest

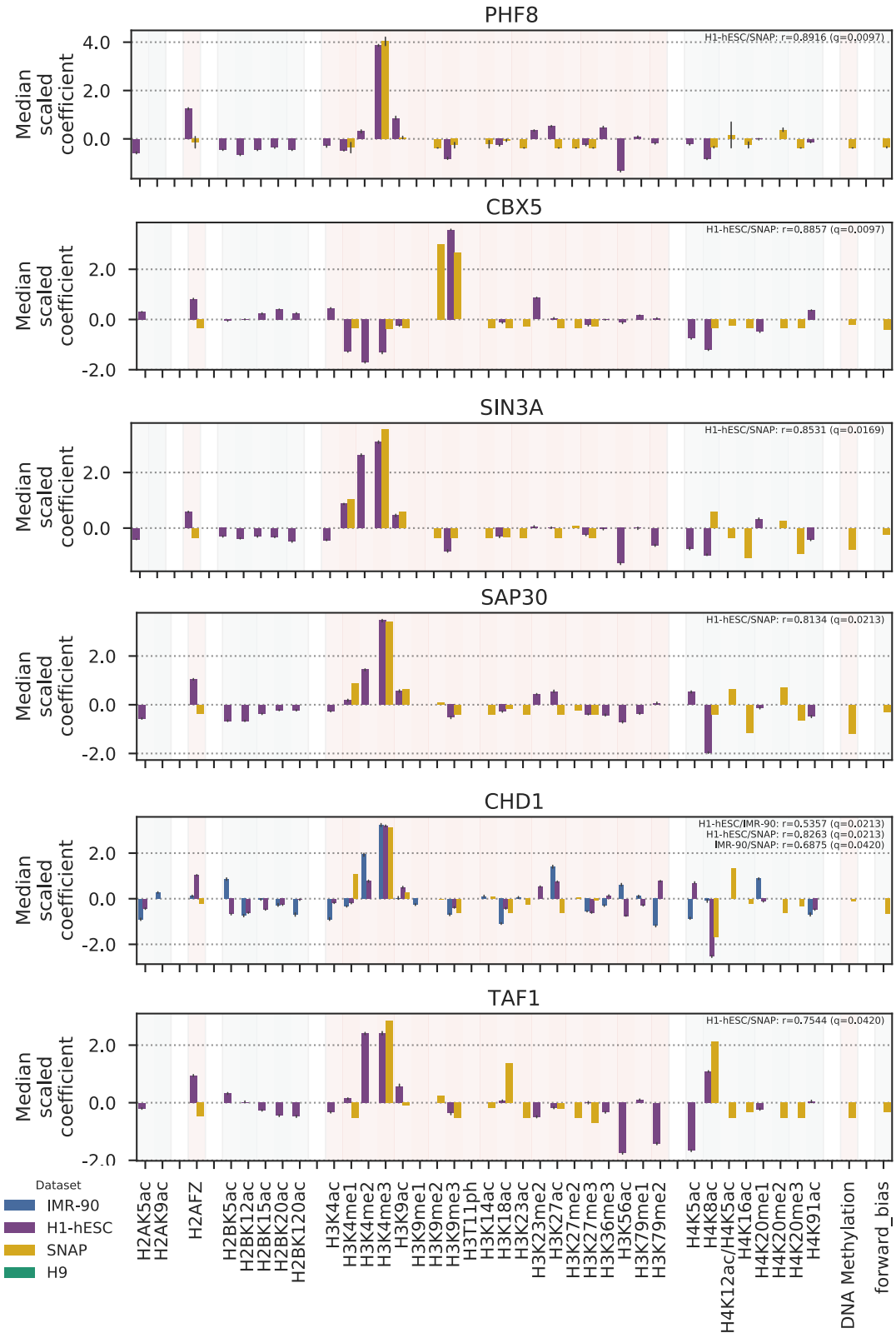

Figure 3.5: A plot of the coefficients estimated from ChIP-seq (see section 3.2.2) and SNAP (see section 2.3.7) data where the two estimates match. Error-bars indicate bootstrapped confidence intervals of coefficients obtained across 150 runs of the model.

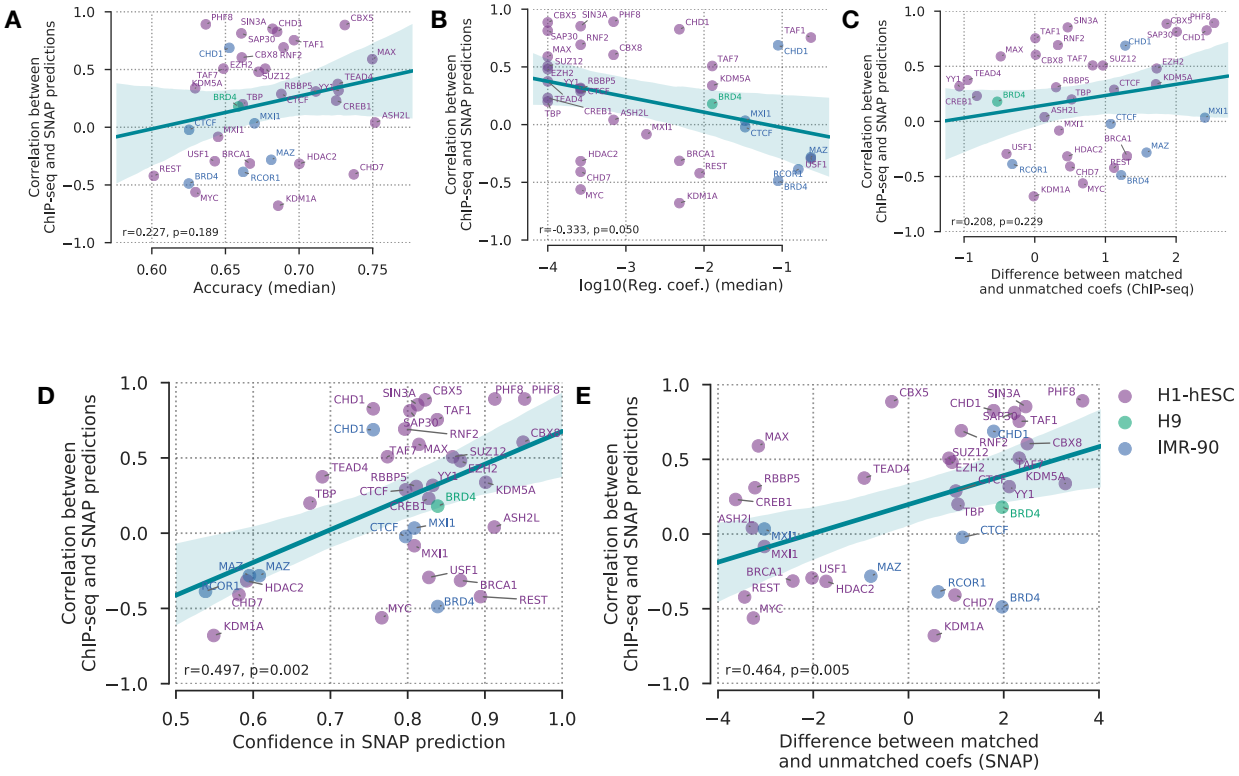

Figure 3.6: Relationship between performance statistics and agreement of the ChIP-seq and SNAP models. The y axis in all of the plots represents the Pearson correlation between the ChIP-seq and SNAP predictions. The X axis shows: the median accuracy of predictions in ChIP-seq model (A), the base 10 logarithm of regularisation coefficient in L1 logistic regression, as obtained during training (note that lower values indicate more regulation, B), difference between the maximal value for ChIP-seq-obtained coefficient that has a counterpart in SNAP data and the maximal coefficient that is unmatched (C), confidence in SNAP prediction, as described in section 2.2.6 (D) and similar statistic as in (C) instead applied to SNAP coefficients. Linear regression line is drawn on all plots, with the bootstrap confidence interval shaded. Pearson correlation coefficient and the associated p-value is shown in the lower-left corner of each plot.

insufficient sampling.

We have therefore visualised the raw ChIP-seq signal of the strongest predictor for each of the targets across the 10,000 highest ChIP-seq peaks associated with this target protein (Figure B.9). We saw only four cases where the strongest predictor did not colocalise with the regions most enriched with the target protein: the majority of CTCF and REST peaks in IMR-90 and H1-hESC datasets respectively were devoid of H3K4me3, while H4K8ac was present at the highest peaks of BRCA1 and KDM5A in the H1-hESC dataset, even though the model predicts a negative association with this predictor.

Instead, we have found that the agreement between the two models is weakly dependent on the confidence in the SNAP prediction (obtained from the r^2 statistic, see section 2.3.7), as shown in Figure 3.6 D. Inspecting this plot, it is clear that the data is split in two groups with lower-confidence predictions from SNAP dataset having the worst scores. No such correlation has been observed with the accuracy of predictions from the ChIP-seq model (Figure 3.6 A). A weaker correlation has been observed with the strength of the regularisation parameter[5] in ChIP-seq model (Figure 3.6 B).

Taken together, these observations indicate that the SNAP dataset captures the *in vivo* behaviour observed in ChIP-seq well, however in some cases it fails to adequately capture the variation within itself (resulting in lower r^2 scores of confidence in the prediction), resulting in the set of coefficients that are not represented in ChIP-seq data. At the same time, it is clear that both models agree on proteins which have simpler epigenetic signatures more (CBX5, for instance, as indicated by stronger regularisation). Which clearly highlights the limitations of the assumption that epigenetic reader interactions with nucleosomal modifications can be explained linearly, which is, of course, expected as the linearly combinable response to combinatorial PTMs is the simplest.

The lack of overlap between the predictors in the two data sources plays a part in this too. For instance ChIP-seq-based predictions for MXI1 (Figure 3.7, top panel), correlate strongly among the cell-lines studied. Neither of these predictions correlate with the observations from SNAP datasets. It is clear, however, that the latter model estimates DNA methylation as the key predictor, a mark that we have not included in the Next-Generation Sequencing (NGS) dataset, even though methods to analyse this modification have been established[197].

To investigate the extent of this observation in a quantitative way, we have computed the difference between the maximal coefficient obtained from the two models that is present in both data sources (i.e. is *matched* by the other source) and the maximal coefficient that is *unmatched* by the other data source. As seen in Figure 3.6 C, there is little dependency between this statistic and the resulting correlations in the ChIP-seq dataset. On the other hand, a

[5]In scikit-learn implementation of logistic regression, lower values indicate stronger regularisation penalty

similar statistic applied to the SNAP data Figure 3.6 E weakly correlates with the measure of agreement between the two data sources. Proteins with unmatched coefficients having higher importance than matched ones in SNAP data (i.e. negative matched-unmatched difference) show lower agreement scores, while proteins where the highest coefficients are matched across the data sources show higher agreement. These observations again indicate that the chromatin states in the SNAP dataset capture the regulation observed in the ChIP-seq dataset, but also provide additional information that is not present in the ChIP-seq datasets. At the same time, these observations could be used to raise concerns about the overfitting of the models as it is clear that the explanations provided by the models do not generalise to alternative contexts.

Interestingly, four predictions: KDM1A, CHD7 in H1-hESC cells as well as RCOR1 and BRD4 in IMR-90 cells are clear outliers in Figure 3.6 E and therefore cannot be explained by the lack of data for their strongest predictors. The estimated coefficients for these proteins are shown in Figure 3.7. Notably the models disagree about the response of CHD7, RCOR1 and KDM1A to H3K4 methylation. The SNAP data for these proteins clearly shows the complex relationship between between H3K4me3 and H4K16ac where the proteins are recruited to H4K16ac, but excluded when the H3K4me3 is also present on the nucleosome (see Figure 3.8, compare H21 with H13, for instance), a pattern that regulates many Transcription Factor (TF) (see section 2.3.5, factor 7). This observation provides a good candidate to explain both the divergence of the models as well as potential shortcomings of the linear assumption. It would be further interesting to explore the distribution of H3K4me3 and H4K16ac *in vivo* and its impact to protein binding, however we do not have this dataset.

Coefficients for such complex dependencies may be approximated by the linear models by overfitting, but in these cases the coefficients might not be interpretable in isolation, nor would they generalise. Thus the negative association with H3K4me3 in the predictions from SNAP data could be explained by adjustment of the positive influences by other marks which are not present in the ChIP-seq datasets. Similarly, SNAP predictions put equal emphasis on each of the nucleosomes profiled, while the ChIP-seq-based predictions weigh them by genomic occurrence. It is not clear how frequent this complex combinatorial state is in the genome.

BRD4 protein is an outlier among these outliers as its binding does not exhibit this response to H4K16ac. BRD4 is well known to recognise acetylated lysines on histone tails, especially on the histone H4[29]. Its association with promoters[198] and super-enhancers, decorated by H3K27ac is also well-established[188]. Predictions from the SNAP experiments clearly identify the link with histone H4 association, but the predictions from ChIP-seq put less emphasis on it. In fact, it is clear that ChIP-seq predictions across the H9 and IMR-90 cell lines are not consistent with each other, though more consistent than either are with SNAP (Figure 3.7). The strongest disagreement between these cell lines is observed between the coefficient values for H3K4me3.

We speculate that perhaps such disagreement could be explained by multiple binding modes associated with this protein, which strongly depend on the *in vivo* context. In this view the binding of BRD4 is dependent on the complex mechanism of the events at a chromatin locus, which adjusts which PTMs are necessary for the recruitment of this factor. Such view is supported by the literature, for instance Shi et al. have shown TWIST1 to mimic the acetylation on histone H4 in order to recruit the BRD4 via interaction with its second bromodomain[199], allowing BRD4 to bind to nucleosomal signatures that would otherwise be unstable due to lack of interaction with this second bromodomain. *In vitro* data obtained from SNAP is unlikely to discern such signatures, as the nuclear extracts provide only the averaged global view of such interactions without the ability to discern local environment of the chromatin (in fact we do not observe TWIST1 at all as it is weakly expressed in HeLa cells).

Our approach to the modelling of the ChIP-seq data has also been designed to capture the global regulation in order to make it comparable to the results from SNAP datasets. In the next section we describe our preliminary work on the alternative approach to the ChIP-seq data analysis which is designed to capture the local regulation by chromatin states observed *in vivo*. We show how this method can be used to analyse studies from multiple datasets simultaneously and to recover both simple and complex regulation of epigenetic factors. We also show that this model can be used to explicitly characterise the differences between cell lines, in terms of which states of epigenetic reader binding are utilised.

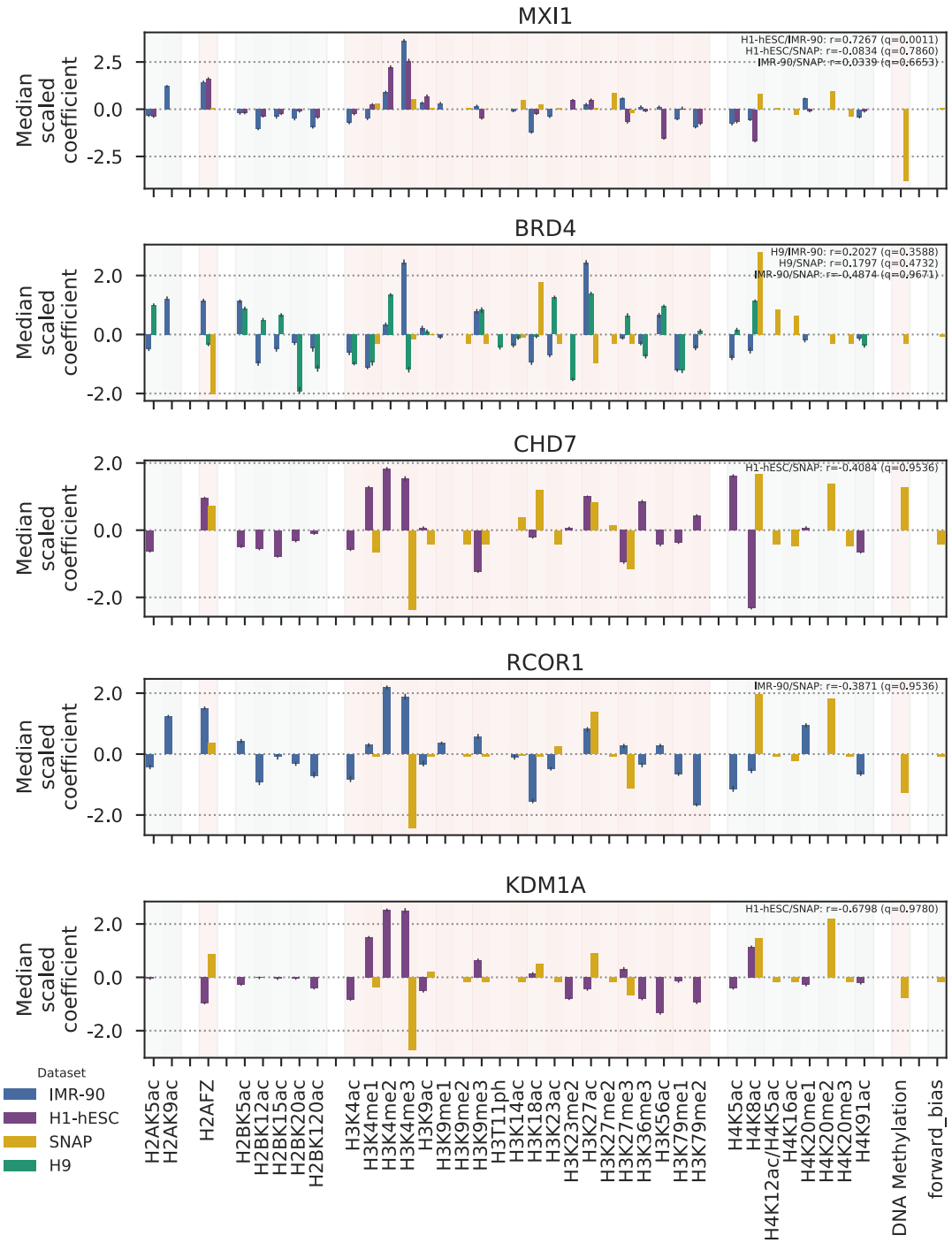

Figure 3.7: Profiles of chromatin interaction coefficients in a selection of cases where predictions from the ChIP-seq model do not agree with the predictions from SNAP dataset. Error-bars indicate bootstrapped confidence intervals of coefficients obtained across 150 runs of the model. See also Figure 3.8 for heatmap of their binding profiles from SNAP data.

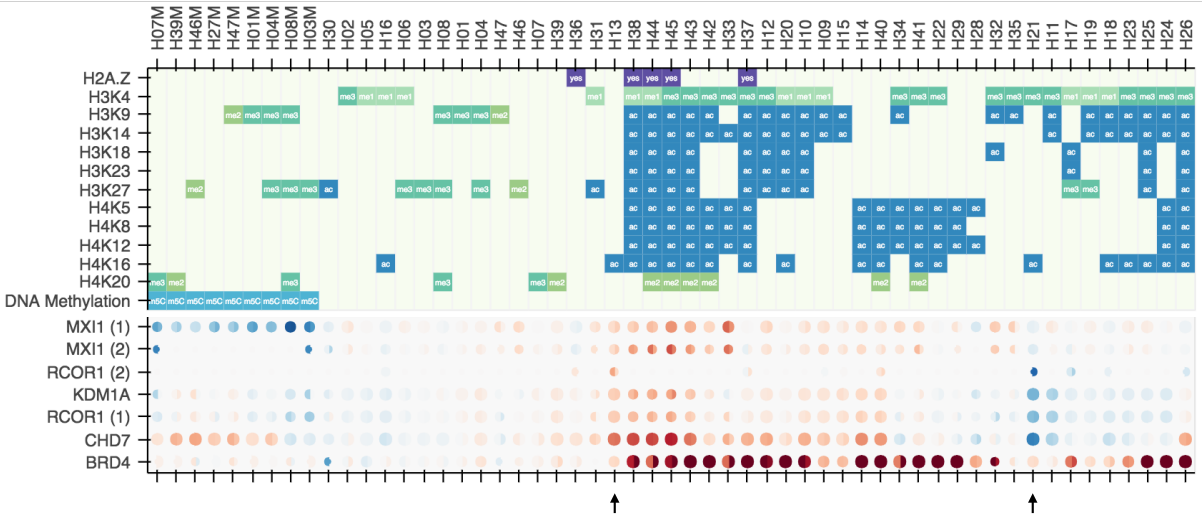

Figure 3.8: Visualisation of SNAP data for the proteins shown in Figure 3.7. Note the difference in the response to H4K16ac when H3K4me3 is present for some of the proteins (indicated by arrows).

3.3.2 Bernoulli mixture models highlight the importance of chromatin-reader associated states

In the previous section we have described the results from the predictive model of protein binding to nucleosomal modifications on ChIP-seq data. We have shown that the predictions from this model correlate to the predictions from SNAP data for some of the proteins analysed. In other cases, we have attempted to identify the factors that influence the discrepancy between these modifications. We have hypothesised that some chromatin readers may respond to multiple chromatin states, which are dependent on the local context of chromatin regulation and may be more (or less) utilised across the cell lines. In this section we describe the preliminary results obtained from the BMM method described in section 3.2.3 to simultaneously analyse ChIP-seq datasets in multiple cell-lines in order to identify the groups (i.e. clusters) of chromatin states and their relationship with chromatin binding proteins.

Specifically, the model assumes a small number of chromatin states that exist but are not observed. The algorithm then attempts to model these states as a mixture of independent Bernoulli random variables and recovers the number of such states (using Integrated Complete Likelihood (ICL) criterion), their prior probabilities (i.e. mixing probabilities) as well as the probability that the state *emits* the signal of ChIP-seq track from the provided data (see

section 3.2.3).

In this section we use a subset of the ChIP-seq data analysed in the previous section containing all of the histone PTMs and the epigenetic proteins BRD4, CBX5, CHD1, CHD7, CTCF, HDAC2, KDM1A and YY1. We also include the ChIP-seq data for POLR2A (general, as well as phosphorylated at serine 5) for reference.

We have profiled the model in three cell-lines for which an extended histone ChIP-seq dataset was available (H1-hESC, H9 and IMR-90). The model has been fit for all three cell lines simultaneously (as described in section 3.2.3), as well as individually. In this section we describe the results from the combined model, however the results for individual analyses are provided in the appendix (Figures B.11, B.12 and B.13). Performance statistics for all four models are summarised in Figure B.10.

The results from training the model are presented in Figure 3.9. As shown in the panels A and B of this figure, the BMM model recovers most of the regulatory states identified by the Roadmap analyses. Even the highly-specialised chromatin states, such as bivalent promoters which are marked with both H3K4me3 and H3K27me3, have been recovered (Figure 3.9, state S26). The estimated states cluster by their function. Some of the states are clear outliers among these clusters, for instance a quiescent-like state S5 appears within the the enhancer-like state cluster. These outliers are likely to be results of either inaccurate labels assigned to these states based on the Roadmap data (as only 6 chromatin marks: H3K4me1, H3K4me3, H3K9me3, H3K36me3, H3K27ac and H3K27me3 were analysed in the Roadmap models), or similar states having different function depending on the genomic context, which we do not take into account.

The model is also capable of identifying known relationships between epigenetic marks and epigenetic proteins. As shown in Figure 3.9 A1, a well-established relationship between the CBX5 and H3K9me3 is clearly recovered by the model[168].

Simultaneous training of the model on multiple ChIP-seq cell-lines allows it to make cross-dataset predictions. For instance, the model predicts a state where H2AK9ac co-occurs with the

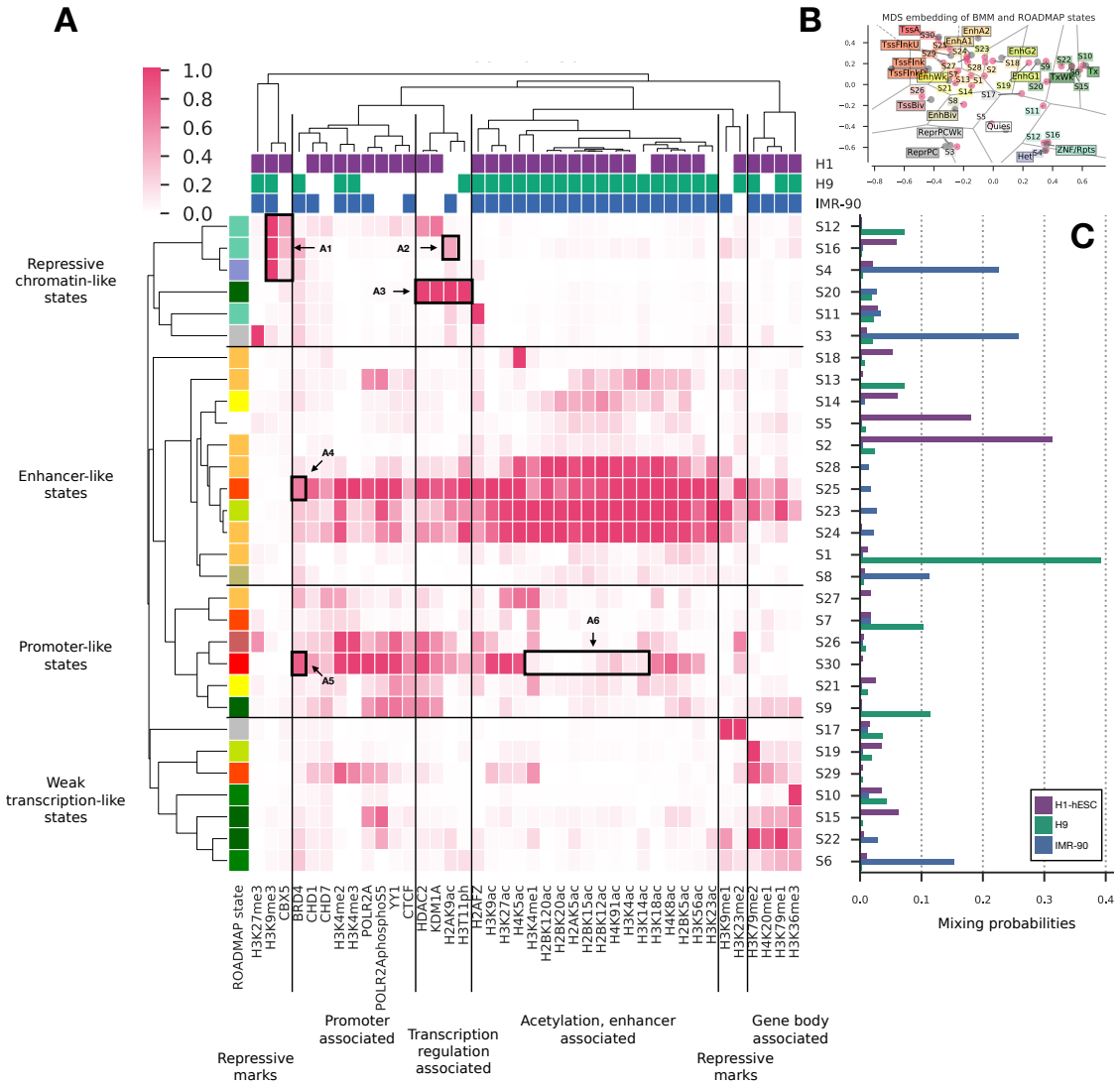

Figure 3.9: A summary of the inferred BMM model as described in section 3.2.3. (A) Visual representation of the emission probabilities in the inferred model. Colour scale represents emission probability. The colour coded section below the column dendrogram indicates availability of the track in cell-type specific dataset. Rows are colour coded by most-similar ROADMAP state, as described in (B). See text in section 3.3.2 for the description of items highlighted by arrows. (B) Multidimensional Scaling (MDS) embedding of model's states along with the states from Roadmap's 18 state model[52]. Nodes are assigned to the colour of closest state in Roadmap dataset prior to the embedding. The same colours are reflected on the left-hand side of the panel (A). (C) Mixing probabilities for each of the three datasets combined, equivalent to the proportion of loci in that state within the particular dataset. See section 3.2.3 for definition.

aforementioned CBX5 and H3K9me3 (Figure 3.9 A2). Similarly, HDAC2, KDM1A, H2AK9ac and H3T11ph are predicted to co-exist in state S20 (Figure 3.9 A3). While these observations are interesting as none of them could have been made without the joint integration of all three datasets, and in fact the link between H3T11ph and demethylation of lysine H3K9me3 has been proposed before[200], we refrain from claiming their significance. It is clear from Figure 3.9 C that states containing strongest out-of-dataset claims are most likely to be present in the cell lines that lack the strongest predicted targets, but are nearly absent in the datasets where these targets exist. For instance, S16 is most prominent in the H1-hESC dataset (which does not contain data for H2AK9ac). Similarly, the state S20 is more common in H9 and IMR-90 datasets, neither of which provide sufficient information about the interplay between the marks highlighted in A3. This behaviour could result from the model discovering that a parameter that is free and therefore assigning a value to it in a fairly random fashion. Additional controls, such as cross-validation based analyses, are needed to determine whether this is an artefact or not.

Another problem with the current formulation of the model is that it is also prone to creating states that are overly-specific to a particular dataset. For instance, the states S1 and S2 which capture a significant subset of H9 and H1-hESC datasets respectively just by themselves. The use of the large prior on the mixing probabilities (see the methods in section 3.2.3) does not seem to counteract this behaviour. While it is true that certain states are more common than others in the dataset, we would have expected a higher degree of generalisation.

Despite the two issues described above, the model provides some insights into the behaviour of the predictive models described previously. In section 3.3.1 we used the protein CBX5 as an example whose binding pattern is predicted consistently across the *in vivo* and *in vitro* models, as well as BRD4 to suggest the existence of complex regulation patterns across the chromatin states. The results from the BMM model provide a clear illustration of the different regulatory patterns for such proteins.

Most of the CBX5 signal is explained by the states highlighted in Figure 3.9 A1. These states are very similar to each other: specifically they are all marked with the presence of H3K9me3.

In this sense CBX5 interacts with chromatin in a very simple fashion, which is solely dependent on H3K9me3. This is clearly represented in the results from linear model analyses. The signal for BRD4, on the other hand, can be emitted by most of the states in the model. The multi-state binding behaviour of BRD4 is also visible from clustered ChIP-seq profile plots (Figure B.14). Since the data for BRD4 was borrowed from sources that are not part of the ENCODE, we do not discount the probability that this behaviour could be linked to the biases associated with these datasets. While further work on the data processing pipeline (which is described in section 3.2.1) is necessary to identify if and how this bias should be corrected, we can use this all-over distribution of BRD4 to illustrate certain aspects of BMM models and highlight their strengths.

According to the model, the most likely mark to co-occur with BRD4 is H3K9me3 in the H1-hESC and IMR-90 cells, and H2BK5ac in the cell line H9 (Figure 3.10 A, computed via Bayes' theorem). The association with H3K9me3 comes directly from the states S16 and S4 which are among the most likely states to observe in H1-ESC and IMR-90 datasets, given BRD4 signal (Figure 3.10 B). In the H9 cell line, on the other hand, the most probable state to be observed, given BRD4 signal is S1, by a wide margin. As seen in Figure 3.9, the prior probability of this state in H9 dataset is approaching 40% and is therefore nearly four times larger than the mixing coefficient for the second *a priori* most popular state S9. It is clear that in this case the *a priori* mixing probabilities dominate the *a posteriori* probabilities of association with epigenetic reader proteins. This necessarily means that analysing the genomic data globally tells us more about the dataset than the binding preferences of the readers.

We argue that it is perhaps more relevant to explore the states most likely to emit BRD4 signal instead, specifically S25 and S30 (Figure 3.10 C). These states occur fairly infrequently and therefore do not contribute much to the global statistics. For instance, S25 is present in slightly more than 1.5% of IMR-90 data (see Figure 3.10 D). The states themselves are similar and correspond to the binding environment we would expect to see BRD4 located in. Both of the states correspond to highly-acetylated chromatin (which is expected, given that BRD4 has two bromodomains[29]), but state S30 lacks the signal for enhancer-associated H3K4me1 and a subset of acetylations (Figure 3.9 A4, A5).

Interestingly, a variant of these states can be recovered not only from the combined models but also from models trained individually (Figure 3.10 E). Surprisingly, both of these states can be recovered in the H1-hESC dataset, for which no BRD4 ChIP-seq data exists. The only instance where both of these states could not be recovered was observed in the H9 cell line, for which no equivalent state to S30 could be found. This was to some extent expected as the mixing probability of this state associated with the H9 dataset is extremely low in the combined analysis (Figure 3.10 D) providing further evidence that simultaneous analysis of datasets recovers the states common across the datasets correctly.

Together these results provide a proof of concept for the utility of BMMs models in establishing the epigenetic reader targets on the chromatin. We have shown that simple (e.g. CBX5) and complex (e.g. BRD4) chromatin readers could be distinguished by these models through a simultaneous clustering of multiple ChIP-seq datasets. This distinction allows to split the chromatin readers into subsets that could be explained by the simple linear models (e.g. LETOR model described previously), and the readers that require higher-order models to be developed in order to explain their targeting. At the same time these models allow the exploration of the key differences between the chromatin state distribution across the datasets which enables one to study how these differences affect the chromatin reader binding in different cellular contexts. Finally, the established chromatin states associated with the reader proteins provide a direct blueprint for future *in vitro* studies such as SNAP as the library of recombinant nucleosomes could be sampled directly from the output of the model.

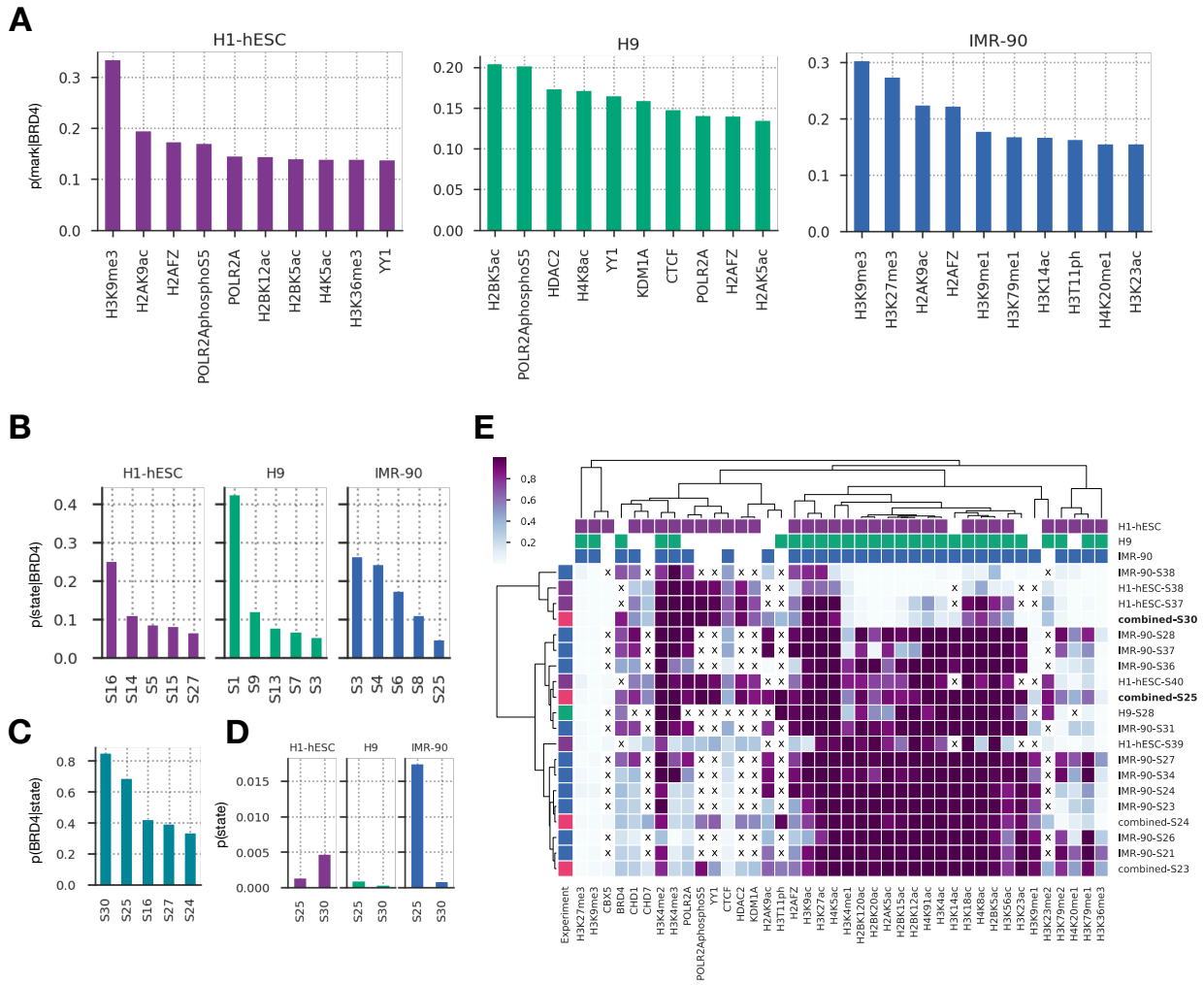

Figure 3.10: A summary of the BMM model results for the BRD4 protein. (A) A posteriori estimates of the probability of observing chromatin marks or proteins, given the presence of BRD4 at that locus. (B) Most likely hidden states given presence of BRD4. Note that for both H9 and IMR-90 datasets, for which we have the BRD4 data, the a posteriori most likely states are also most likely a priori (see Figure 3.9 C). (C) States most likely to associate with BRD4. (D) Mixing proportions for the two most likely BRD4 states. Note that S25 is more common in H9 and IMR-90 datasets, but S30 in H1-hESC. (E) Clustering of states similar to S30 and S25 obtained from models estimated from single cell line data only (Figures B.11, B.12 & B.13). Rows are clustered by cosine metric, using complete linkage. The colour coding represents the experiment: red represents the combined model, while purple, green and blue represent H1-hESC, H9 and IMR-90 models respectively. Note that similar states are recovered from single cell-line models, including H1-hESC which clearly recovers both the state S30 (H1-hESC S37/S38) and the state S25 (H1-hESC S40), even though no BRD4 dataset was present for this cell line. Interestingly no state equivalent to the combined S30 state has been obtained in H9-only model.

3.4 Discussion

Understanding the language of covalent histone modifications is still an active topic of research in epigenetics. In this thesis we primarily focus on how the epigenetic readers interact with this language to elucidate how this language is understood. In chapter 2 we have analysed a large scale SILAC nucleosome affinity purification (SNAP) dataset, which has provided us with novel insights into this language from *in vitro* perspective. In this chapter we describe the preliminary work which augments these results with *in vivo* insights from the Chromatin immunoprecipitation sequencing (ChIP-seq) datasets.

Chromatin immunoprecipitation sequencing (ChIP-seq), due to its popularity and accessibility, provides a perfect dataset for this use case. The local, context-specific regulation, which is not captured in SILAC nucleosome affinity purification (SNAP) data as it is available only *in vivo* is readily captured within the results of these analyses. In this chapter we have described the data preprocessing necessary to be able to compare multiple ChIP-seq datasets from multiple sources. We have then described a linear Learning to rank (LETOR) model which queries the ChIP-seq data in a way that is comparable with the SNAP results and have used it to investigate where the results predicted from *in vitro* studies match with the behaviour observed *in vivo* and where do they not. We have further described a Bernoulli Mixture Model (BMM) approach which is designed to provide the insights into the local distribution of chromatin states across the genome and their relationship to chromatin readers proteins.

We have shown that only a handful of predictions made by linear models match between SNAP and ChIP-seq datasets. While these differences could potentially be attributed to the mismatch between the cell lines used in SNAP and ChIP-seq experiments, under the histone code hypothesis we would expect the regulation by histone Post-Translational Modification (PTM) to be conserved across multiple cell types, just like the genetic code is. Indeed, performance statistics indicate that predictions that both models agree on correspond to the epigenetic readers that interact with a small number of PTMs in a relatively simple pattern. BMM models have shown that such proteins interact with few nucleosomal states, which are marked with similar modifications. Other epigenetic readers, such as BRD4, seemingly interact with more

epigenetic states. The interacting states remain consistent across the datasets (cell types), however, their overall distribution (in terms of percentages) varies, which therefore changes an estimated *average* binding profile of the protein. This observation seemingly suggests that perhaps the role of epigenetic reader proteins in epigenetic regulation is overestimated. Perhaps the writer/eraser proteins and the modification cross-talk that they establish is more important. In other words, the specific effects of combinatorial histone modifications may primarily be encoded by preventing the combinations from co-occurring, rather than by producing a distinct *readout*. Such mechanisms have been discovered previously, for instance H3K36me3 prevents H3K27me2/3[44], and a similar hypothesis has been raised before[26], but our data only allows us to speculate about this.

We have also shown that the estimated binding profiles are likely to mismatch in cases where the important regulators of the proteins are not covered by one of the two datasets. This suggests overfitting, as well as highlights the need to match the signals used for the predictions as much as possible by including data from other Next-Generation Sequencing (NGS) datasets (e.g. DNA methylation) or by filling the gaps from Mass Spectrometry (MS) datasets (see section 3.4). We suggest that epigenetic readers which exhibit complex, hierarchical binding to the nucleosomes can be captured by BMM models instead. Not only do these models extract the epigenetic reader interactions with chromatin states, but they also provide information on how this regulation varies between datasets.

In the context of elucidating the language of epigenetic modifications, these preliminary results serve to highlight the utility and the necessity of data integration cross experiments. These models clearly put this data into context as they identify the strengths and weaknesses of the models developed in chapter 2, as well as provide guidance on where the future research should focus on. Specifically, these models could be used to alleviate the over-fitting of the models both from the computational as well as the experiment-design points of view. These models also serve to provide insights on how the nucleosomal modifications work in combination *in vivo*, which will facilitate the design of non-linear predictive models as well as the design of novel experimental studies.

We have also identified ways to improve the ChIP-seq methods described in this chapter. In section 3.3.2 we have remarked that the seeming association of the BRD4 with most of the states predicted in the model could be partially due to the use of non-ENCODE dataset. It would therefore be interesting to compare the input (control) samples from ENCODE and non-ENCODE sources to see what the extent of this bias is and how can it be corrected.

Similarly, the comparisons between SNAP and ChIP-seq based linear modelling procedures revealed the damaging role of having a vastly different set of predictors to these models. In the future experiments, inclusion of the NGS datasets on DNA methylation would provide a more accurate benchmark between the two models, similarly, the integration with the data from MS studies could fill in the gaps too. At the same time in order for the model predictions to be biologically meaningful, the models have to be able to generalise and, to some extent, "know what they do not know" and therefore further work is needed in this area as well.

The BMMs developed in this chapter have allowed modelling of the multiple ChIP-seq datasets simultaneously. These models are powerful enough to make cross-dataset predictions of cross talk between histone PTMs, but in this chapter we did not focus on them as we provided no framework to assess their accuracy. Future work is therefore needed to establish a cross-validation platform where the accuracy of cross-dataset predictions in the model could be quantified. Such framework could be designed either in terms of a simulation study (generating synthetic datasets based on known state distributions), or from actual ChIP-seq datasets for which the cross-dataset predictions are included artificially by removal of a subset of data tracks.

Finally, from a more general viewpoint, we view these methods to be tightly integrated into the *in vitro* experimental procedures. These approaches can be used to suggest nucleosomal states that need to be used in *in vitro* experiments such as SNAP, to adequately capture the *in vivo* chromatin environment of an epigenetic reader. The results from such additional *in vitro* experiments can then be cross-checked with this data again (as we have done in this chapter), highlighting potential issues with the methods used or nucleosomes studied, at which point the ChIP-seq data can be used again to suggest further experiments, creating a feedback loop. We

believe that such feedback loops will prove to be crucial in the understanding of the chromatin regulation at the nucleosomal level.

Extension to bottom-up affinity pull-down mass spectrometry studies on histones

We have only focused on the application of this model on NGS data in this thesis, but, as briefly mentioned throughout this chapter, the BMM model formulated earlier could in theory be compatible with the data from MS. Which would simultaneously allow us to fill in the gaps in ChIP-seq datasets, and simultaneously re-establish the cross-talk between PTMs lost to proteolytic digestion and place them on the genome. The ideas discussed here are purely hypothetical, but they align towards our future long-term goal of unified integration of the NGS and MS data.

Mass Spectrometry (MS) allows the measurement of the ratio of co-occurrence between nearby histone PTMs that occur on the same peptide. In bottom-up mass spectrometry (see section 1.2.1), such readouts are often limited to one or two modifications per histone tail. For instance, we can accurately quantify the presence of all of the the possible combinations of modifications between the lysines 9 and 14 on the histone H3, but we could not tell how often these modifications co-occur with, say methylation on the lysine 4 on the same histone (see Figure 1.7). ChIP-seq data, on the other hand, provides information that could help bridge this gap across the peptide boundaries. The information about how often the modifications co-occur on at chromatin loci, which is provided by ChIP-seq, could be used to statistically *re-join* the proteolytically cleaved peptides and therefore provide us with information of the co-occurrence of the PTMs across the whole nucleosome. Similarly, same data could be used to place MS-derived PTMs onto the genome, as well as fill in the missing data in ChIP-seq.

There are two major issues to the integration of this data to the current definition of the BMM model: (1) the PTMs are not distributed as independent Bernoulli variables in the MS datasets, as they resemble categorical variables more; (2) mass spectrometry does not directly count the data in a discrete fashion (i.e. like we count ChIP-seq loci), but rather provides us with the read-out of the proportion of the data directly.

To expand on the first issue, in the mass spectrometer we would never observe both H3K4me2 and H3K4me3 on the same lysine as a residue can be modified at most once. On the other hand, this scenario is quite common in ChIP-seq as the experiments are independent. It would therefore be more accurate to treat the data from the MS studies as a categorical random variable. Using a one-of-k encoding for the variables again, and similar notation as before, the emission probabilities for such random variables would be expressed as

$$P(\mathbf{x_i}|\mathbf{z_i}, \mathbf{p}) = \prod_{k=1}^{K} \left[\prod_{d=1}^{D} p_{k,d} \right]^{z_{i,k}} \quad \text{s.t} \sum_{d=1}^{D} p_{k,d} = 1 \tag{3.37}$$

Comparing the equation Equation 3.37 above with previously defined Equation 3.3 for ChIP-seq data, we see that both equations are parametrised in a very similar way. The key difference between the equations is how the sum-to-one constraint is handled. Categorical (MS-associated) parameter introduces a Lagrange multiplier, while the independent Bernoulli constraint explicitly adds the $1 - p_{k,d}$ expression to the product.

Since the data from MS experiments reflects the true distribution in the cells more accurately, it is reasonable to assume that the emission probabilities from the independent Bernoulli formulation would be related to the similar probabilities from the full categorical model. Such dependency could be modelled probabilistically, but the extra link between variables would complicate the model estimation significantly. Instead, due to a similar formulation, we suggest that the categorical quantities from the MS experiments could be approximated with the independent Bernoulli distribution. This would incur an extra penalty to the likelihood equivalent to $1 - p_{k,d}$ making it necessarily lower, yet easy to estimate.

This Bernoulli approximation provides a direct way to handle the discrepancy between the tracks measured in the ChIP-seq and the modifications quantified in MS data using the missing value mechanism described earlier. This way, each of the MS peptides, would produce a partial vector of the modifications observed in this peptide, but have missing data values for the residues on the related peptides. The data from ChIP-seq would then serve to reconnect the relationships broken down by trypsin digestion.

The second issue, concerns the discrepancy between the discrete loci measured in ChIP-seq that can be counted, and the relative read out of the proportion of co-modified peptides in MS that is not discrete. It could be alleviated by observing that the necessary statistics computed for the BMM model have to be computed only for the unique combinations of histone PTMs as they solely depend on the observed PTM pattern at the locus. The complete likelihood is then computed by multiplying the results obtained from the unique combinations of PTMs by the number of times they occur in the genome. In MS, the relative proportions of the PTMs that were estimated correspond to such counts and are constrained to sum-to-one, which can also easily be achieved for the ChIP-seq model without changing its formulation. Probabilistically, this is equivalent to adding a prior probability of observing an unique PTM combination, which is equal to the proportion of the genome covered by that modification pattern.

So far we have only described described a blueprint that combines the two datasets, we have not yet addressed the main goal of these models: the goal to discern the different chromatin states an epigenetic reader might target. We have already integrated epigenetic readers as independent Bernoulli variables in the ChIP-seq-based models, but the bulk-chromatin bottom-up MS based read-outs provide no such information. Similarly, we have already learned in this chapter that the distribution of epigenetic states varies vastly across the ChIP-seq datasets in different cell lines, even when these datasets were produced using similar experimental techniques. This variation is likely to be even more pronounced between ChIP-seq and MS-based datasets.

We suggest that it is more sensible to compare the ChIP-seq datasets with the data from Affinity Purification Mass Spectrometry (AP-MS) experiments rather than bulk-chromatin MS samples. Affinity Purification Mass Spectrometry (AP-MS) experiments immobilise a chromatin reader to beads and use this immobilisation to purify the nucleosomes associated with this reader (i.e. the reverse experiment to SNAP which immobilises the nucleosomes and purifies chromatin readers). The results from these experiments could be fit into the BMM compatible format by setting the variable corresponding to the analysed epigenetic reader to 1 in all of the partial nucleosome vectors derived from this experiment.

Antibody based immunoprecipitation has been successfully applied to this problem in the

past[201], but recombinant-protein based approaches, such as Matrix-Assisted Reader Chromatin Capture (MARCC)[202] can be more flexible as they allow the possibility of mutagenesis studies. Briefly, in the MARCC approach one expresses the epigenetic reader protein (or its recognition domain) recombinantly with the addition of an affinity tag (i.e. HaloTag) in bacterial cells. A large amount of mononucleosomes is then purified from mammalian cells via digestion with micrococcal nuclease (i.e. MNase) enzyme. The purified protein is immobilised via the affinity tag and subjected to purified nucleosomes. Target-interacting nucleosomes are then enriched via the target-attached beads and subjected to Mass Spectrometry based analyses to identify what modifications are found in the nucleosomes enriched by the target.

While we have no data from such experiments yet, we have, however, optimised the first step of this procedure – the nucleosome extraction procedure for large scale nucleosome purification. This procedure involves growing a large amount of HeLa cells, extracting their nuclei, and purifying nucleosomes from these nuclei via MNase digestion (see section B.2). The last step is sensitive to experimental conditions and the batch of the enzyme and therefore needs to be optimised.

During this optimisation we have observed that longer MNase digestion times lead to fewer soluble mononucleosomes purified as the nucleosomes started to precipitate (Figure 3.11 B and C, soluble fraction S1). It is widely accepted that MNase starts release of the mononucleosomes at open chromatin regions and then slowly moves towards heterochromatic regions[203]. As further nucleosomes are released the already solubilised nucleosomes start to precipitate, but can then be recovered by overnight incubation in a buffer which strips the solution of Mg2+ and K+ ions (Figure 3.11 B and C, recovered fraction S2). The rest of the chromatin can be recovered using a large quantity of NaCl (Figure 3.11 B and C, high-salt recovered fraction S3).

The most likely explanation of this phenomenon comes from the biophysical properties of nucleosomes. It has previously been shown that magnesium cations chelate the negative charge of the DNA and cause nucleosome aggregation and precipitation[204, 205]. This effect is less pronounced in presence of acetylation on histone tails[204]. This result is particularly important as it indicates that S1 soluble fraction may be over-enriched for acetylated nucleosomal

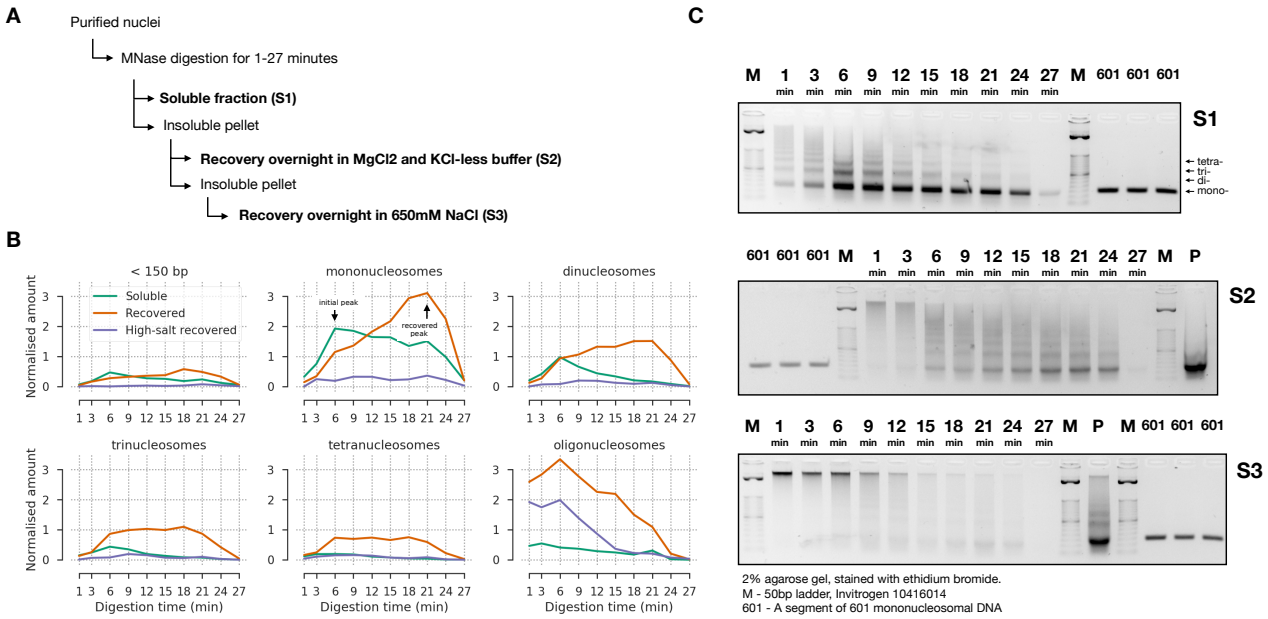

Figure 3.11: Quantification of mononucleosome yield from incremental digestions using MARCC protocol. Panel (A) summarises the experimental procedure, based on the protocol section B.2. (B) contains a quantitative representation of the results in (C). S1 represents the soluble nucleosome fraction, S2 represents the nucleosome fraction that can be recovered after overnight incubation in absence of MgCl2, S3 represents the nucleosome fraction recovered with 650mM NaCL. (see section B.2). The data-point at 27 minute mark is an outlier due to sample handling problems resulting from significant increase in sample viscosity with increased release of DNA.

residues. Alternatively, as the MNase progresses, more and more heterochromatic proteins are being released into the solution. It is possible that these heterochromatic proteins, re-bind to free-floating nucleosome core particles and precipitate them out of the solution too, via a process similar to the documented behaviour of HP1[19] protein, which causes a phase-separation behaviour in cells.

We have already shown that the differences in chromatin state distributions may impact the downstream results significantly, especially when the data is analysed on a global scale. In order to successfully model the data from ChIP-seq and MS experiments in combination, these biases need to be understood completely and accounted for so the epigenetic proteins studied are subjected to their natural environment. That is, euchromatic proteins should be studied in terms of euchromatic nucleosomes, while the heterochromatic proteins should be studied in terms of heterochromatic nucleosomes as otherwise misleading conclusions could be made.

While we have not obtained any further data for these experiments in time for the submission of this thesis, we have obtained a small collaboration grant to investigate this exact problem of combined analysis of ChIP-seq and MS datasets that we intend to pursue in the very near future. We intend to combine these datasets to obtain a complete view of the nucleosomal modifications at the given locus of the genome. This can be achieved by simultaneously using the MS data to fill-in the gaps in ChIP-seq datasets resulting from only a subset of modifications sampled in these experiments, as well as using the modifications sampled in this NGS dataset to virtually re-link the peptides that were broken up by proteolytic digestion. This will empower analyses similar to the ones described in this chapter to query the full scale of nucleosomal modifications, which will necessarily further our understanding into their language. We believe that the extension to BMM described in this section will provide a reasonable approximation of this information and therefore will establish a foundation for further development of the method.

Chapter 4

Conclusion

This thesis has aimed to provide a data-driven approach to elucidate the mechanisms by which epigenetic readers translate the language of nucleosome modifications.

We were interested in identifying how epigenetic readers interact with each other and with chromatin, what constitutes the core components of the nucleosome code, as well as which readers are responsible for interpreting these constituents.

We have assumed the view on covalent histone code hypothesis described in section 1.1.3, which was heavily inspired by the Rando's 2012 opinion paper[26]. Here we view the histone code as a language, in which histone Post-Translational Modifications (PTMs) serve as a words forming "sentences" across the nucleosome core particle. We have therefore sought for similarities between PTMs to identify the constituents that define the topics of this language, as well as attempted to identify the keyword PTMs and how they are interpreted by the epigenetic readers.

We have based our findings on a novel SILAC nucleosome affinity purification (SNAP) study which queries the regulatory subspace of nucleosomes to an unprecedented degree and used the information available in the literature as well as data from Next-Generation Sequencing (NGS) experiments to provide context and an alternative view of the findings presented.

4.1 Summary of Thesis Achievements

We have achieved this by developing a bespoke method to process large-scale SILAC nucleosome affinity purification (SNAP) data which accounts for the biases unique to this dataset. We have further integrated this Mass Spectrometry (MS) dataset with the data from multiple databases including UniProt, Interpro, BioGRID, EpiFactors and Complex Portal to provide a holistic view of novel interactions in the context of the published literature.

We have identified the interactions between epigenetic factors by profiling a number of network inference algorithms, out of which the Context-Likelihood of Relatedness (CLR) algorithm performed the best. We have further improved this algorithm by deriving a closed form expression for the p-value of the interaction probability. This derivation of the p-values allows automatic adjustment for the confidence threshold to network size via the multiple hypothesis testing correction procedure.

The inferred protein-protein interaction network has allowed us to query it for novel protein-protein interactions. We have found that in many cases the evidence for such predictions has already been suggested in the literature, albeit hidden in the supplementary material of the associated papers (see section 2.3.4). In terms of understanding the epigenetic language, this latter observation highlights that there potentially are significant gains to achieve from integration of published data. We have therefore compiled the evidence presented in the literature with the observations from our data to provide a strong case for the importance of our predictions.

We have further used this network to establish a curated list of epigenetic protein complexes observed in the SNAP dataset, including protein complexes that were not included in the EpiFactors database or the Complex Portal and have compiled all this information into a web based interactive visualisation of the dataset that provides a convenient way to query the protein behaviour and their relationships to others.

We have further studied the language of nucleosome modifications by applying a Factor Analysis (FA) method which has allowed us to identify 11 constituents of the nucleosome code which control the behaviour of epigenetic proteins observed in our data. Interestingly, we have shown

that only two of such factors are sufficient to explain up to 42.72% of variance in our dataset. This was an expected result, given our current understanding of the complexity of histone code[26], and similar observations have previously been made on ChIP-seq data by others[72, 173]. However, as we have discussed in section 2.4, such result may also be explained by our nucleosomal library design.

Perhaps more interestingly, therefore, we have shown that eight of these factors are relatively simple in their composition, and generally correspond to presence or absence of known chromatin landmarks such as lysine acetylation or H3K27me2/3 (section 2.3.5). We only identify three factors with non-trivial combinatorial loadings, one of which is factor 7 which captures the antagonistic response to H4K16ac+H3K4me1/3 as compared to H4K16ac on its own. This factor is the closest to the spirit of the histone code hypothesis, as it encodes a switch in response to H4K16ac by the addition of H3K4me1/3. We have also found that DNA binding transcription factors respond exclusively to these three complex regulatory dimensions (as well as the DNA methylation factor, see section 2.3.6), which may suggest that control of DNA accessibility is one of the potential mechanisms of reading the nucleosomal code.

We have illustrated how the estimated regulatory network of chromatin factors is controlled by these constituents by showing that network subunits respond to different combinations of such factors which provides an insight on how the language of nucleosomal modifications is interpreted. For instance, from Figure 2.25 it is clear that the proteins which respond to lysine acetylations, especially H3K14ac, can further be split by their response to H3K4me3. According to our language-based view of histone PTMs, this would correspond to H3K14ac acting as a keyword to establish the interpretation context, while H3K4me3 providing the necessary clarification. This is compatible with the generally established view in the literature, where a general association of lysine acetylation with active transcription is assumed, while H3K4me1/3 is often used to separate active enhancers from promoters[3].

We have shown that models which assume a simple linear additive response to nucleosomal modifications can explain a significant number of epigenetic reader binding profiles in the *in vitro* SNAP dataset. We have also developed a similar linear model for Chromatin immuno-

precipitation sequencing (ChIP-seq) datasets and compared the results from the two models to identify their strengths and weaknesses. We have shown that both models approximate the protein behaviour well on the datasets they have been trained on, but the *in vitro* and *in vivo* based models diverge for epigenetic readers that exhibit a complex binding behaviour. These observations provide an interesting viewpoint to the rhetorical question on whether a linear response to PTMs should be considered histone code (proposed in [26]). Our results suggest that in a given setting, the protein response to the nucleosomal modifications could be approximated by linear combination of individual effects, however, these approximations only generalise for the simplest of epigenetic readers. Predictions for protein showing complex binding behaviour, seemingly depend on the chromatin context (state), and their distribution, which is dependent on the presence of unknown co-factors and cell context. This would seemingly suggest that given the context, such as the distribution of modified nucleosomes, the histone code could be interpreted in a linear fashion, however this interpretation is not universal, and therefore the context itself can change it. The idea of context-specific transcription regulation is not a novel one: a classic example is the YY1 protein which switches from repressive role to activating role in presence of E1A adenovirus protein[206]. Similarly, presence of TWIST1 has been shown to play a role in BRD4 targeting[199]. Our data suggests that DNA methylation may cause a switch between the RMI/BLM and RPA complexes (section 2.3.4). The histone code analyses, therefore, might not translate throughout the studies, which may be why the histone code hypothesis is still debated nearly twenty years in. In our case, however, we do not have sufficient data to distinguish statistical overfitting from true context-dependent switch, but this observation nevertheless raises an interesting question on why the contexts are so different to begin with. It is perhaps the writer and eraser proteins, rather than the reader proteins, that play the deciding role on the interpretation of nucleosomal modifications. Since the nucleosomal modification context seems to vary not only across cell types[52], but also within the same cell type during the lifetime of an organism[207] it provides an important avenue to explore.

In order to study these chromatin states we have proposed a Bernoulli Mixture Model (BMM) which allows joint identification of chromatin states in multiple cell lines from ChIP-seq data. We have shown that epigenetic reader protein BRD4 interacts with similar chromatin states

across the datasets, but these states are distributed at different proportions across the cell lines, which changes the interpretation of 'average' behaviour of BRD4, which depends on the cellular context. We have reasoned that the models and the experiment designs should account for the distribution of these states to be able to unify the information from multiple data sources, both *in vivo* and *in vitro*, in order not to generalise to the 'average' behaviour in one of the contexts and be able to determine the potential variations in the response, due to cellular context. Together these observations provide a foundation for further development of higher-order models of epigenetic interaction that could capture complex binding patterns.

Finally, we have made a number of contributions to the open source software, including submitting bug fixes/implementing new features to the *seaborn* and *ggrepel* software packages, as well as reporting a number of issues to the Bokeh library which have since been fixed by their core developers. We have also released the code for the ChIP-seq pipeline and the BMM implemented in this thesis, which may be of interest to the community.

4.2 Applications

The results presented in this thesis are directly applicable to a number of scientific fields.

Scientists who work in the field of epigenetics will directly benefit from the publication of the interactive visualisation of SNAP data as it will allow them to query their protein of interest and visualise the nucleosomes bound by this protein in an unprecedented way. Similarly, the results from our network, factor, and linear modelling analyses, will allow to describe the relationship between nucleosome modifications and epigenetic readers which will move us further towards the goal of understanding the language of nucleosome modifications and how epigenetic information is converted into biological signals.

Scientists in the field of proteomics will benefit from our description of the challenges involved with working with large scale Stable Isotope Labelling with Amino acids in cell Culture (SILAC) datasets, as well as from the preprocessing pipeline we have developed which allows the correction of biases common to the SILAC datasets.

The field of bioinformatics will benefit from the improvement of the CLR algorithm. The successful application of the FA method to query a complex regulatory space will also help to re-establish this method in the field. The comparative investigation of the ChIP-seq datasets, will help to establish the predictive potential of such data that could be used to both design and verify future studies as well.

4.3 Future work

As this thesis is being written, our collaborators are attempting to verify some of the predictions made by the data. The preliminary results from further *in vitro* experiments performed by Dr. Andrey Tvardovskiy (postdoc, Bartke lab) reproduce the correlation between the INO80 and TBRG1 binding to recombinant nucleosomes that we have predicted using the network analysis. We are further investigating how these proteins interact with the nucleosomes based on the results from the factor and linear model analyses. We expect to obtain conclusive results from these studies in the next few months.

While the online interface is currently functional and used internally every day, it needs some further improvements to make it more appealing and easier to use for other scientists. For the final release, we intend to remove some of the options from the control panel that are rarely used to improve usability. Similarly, we have not yet included the data from the factor analyses and the linear model analyses into this online visualisation which we intend to do in the near future. With these changes implemented we believe the online interface will become a major hub for the information about epigenetic protein interaction with the chromatin.

We further intend to move beyond linear models in attempts to elucidate the epigenetic reader response to complex PTMs binding patterns. While, as we describe in this thesis, linear models can capture the regulation of a large number of the chromatin binding proteins *in vitro*, it is perhaps even more interesting to explore the cases where such models fail. We believe that higher order machine-learning methods, for instance decision tree regressors, could provide a better, yet still intuitive, description of the behaviour of chromatin binding proteins. Over-

fitting will be a major challenge to the application of such higher-order regressors, as even the simpler linear models are prone to this (as we discuss in this thesis). We are currently applying for further funding for this direction of our research.

In order to extract the full utility out of the integrated analysis of the NGS data, future work needs to be done to move the models described in this thesis past the proof of concept state. This work, first and foremost, should focus on simulation and cross-validation studies to establish a benchmark for the performance of the models so that the confidence in cross-dataset predictions could be established. This benchmark could then be used to alter the model to target the issues we encountered in our analyses, such as over-generalisation of chromatin states, impact of the missing predictors, and sensitivity to data sources. Finally, as we briefly mention in section 3.4, we believe that BMM-like models may provide a path to combine the results from Affinity Purification Mass Spectrometry (AP-MS) and ChIP-seq studies. We have obtained a small collaboration grant with the Institute of Computational Biology in Munich to investigate this approach further.

From a more general perspective, there are still many open questions about the chromatin that remain in the field. For instance, we are only beginning to understand the relationship between nucleosomal modifications and intrinsic properties of chromatin, such as DNA accessibility, DNA looping, or chromatin context that is established by epigenetic proteins or cofactor availability. We have highlighted this in the previous chapters, as we have shown that even in controlled *in vitro* settings, these effects not only cannot be eliminated, but also may shape the interpretation of the results. Similarly, even with the large-scale experiments, such as the one analysed in chapter 2, we only barely scratch the surface of potential space of the nucleosomal modifications: the question of the significance of the modifications that were not included remains. Finally, answering these questions will still raise the follow-up question of what is the biological significance of these observations, for instance, is there part of the nucleosomal code which allows for transcriptional memory, which parts of this nucleosomal code play a role in differentiation, can euchromatic marks be inherited?

On one hand, this calls for more data to be produced. On the other hand, it is not possible

to capture such a vast research question in a single dataset. Successful data integration of multiple sources will provide the key in answering these questions in the future. In this thesis we integrated the data from large scale SNAP, NGS studies as well as public databases to provide a general view of nucleosomal integration. In order to truly *crack* the nucleosome code, in the future these analyses should be extended to integrate other datasets as well, for instance, the datasets which profile the specificity of individual protein domains (e.g. [29]), the datasets that profile the distribution of histone modifications in multiple biological conditions (e.g. [207]), as well as loss-of-function studies. This is truly not a trivial task that can be solved directly, however, it provides a goal that can be approached incrementally. Collection and standardisation of the results of such datasets is the first logical step.

Finally, integrative data analyses allow for data driven experiment design which is also needed to successfully interrogate the nucleosomal code. In this thesis, for instance, we show that given the data available to us the interaction between INO80 and TBRG1 is highly likely - an observation that we use for further studies into the relationship between these proteins. At the same time, the integration of other datasets can quickly highlight the gaps in the data, such as the need to profile additional combinations of PTMs in the SNAP experiments in order to make the model's predictions more reliable. This creates a feedback loop between the experiment and the data analysis, which, in our opinion, will be the key in deciphering the chromatin regulation.

Bibliography

[1] John M Greally. "A user's guide to the ambiguous word 'epigenetics'". In: *Nature Publishing Group* 19.4 (Jan. 2018), pp. 207–208.

[2] Tom Sexton and Giacomo Cavalli. "The Role of Chromosome Domains in Shaping the Functional Genome". In: *Cell* 160.6 (Mar. 2015), pp. 1049–1059.

[3] Leah A Gates, Charles E Foulds, and Bert W O'Malley. "Histone Marks in the 'Driver's Seat': Functional Roles in Steering the Transcription Cycle." In: *Trends in Biochemical Sciences* 42.12 (Dec. 2017), pp. 977–989.

[4] Danny Reinberg and Lynne D Vales. "Chromatin domains rich in inheritance." In: *Science* 361.6397 (July 2018), pp. 33–34.

[5] Bernd Schuettengruber et al. "Genome Regulation by Polycomb and Trithorax: 70 Years and Counting". In: *Cell* 171.1 (Sept. 2017), pp. 34–57.

[6] Albert Jeltsch and Renata Z Jurkowska. "New concepts in DNA methylation". In: *Trends in Biochemical Sciences* 39.7 (July 2014), pp. 310–318.

[7] Jesse R Dixon, David U Gorkin, and Bing Ren. "Chromatin Domains: The Unit of Chromosome Organization". In: *Molecular Cell* 62.5 (June 2016), pp. 668–680.

[8] Yasuto Murayama et al. "Establishment of DNA-DNA Interactions by the Cohesin Ring". In: *Cell* 172.3 (Jan. 2018), 465–469.e15.

[9] Bas Van Steensel and Andrew S Belmont. "Lamina-Associated Domains: Links with Chromosome Architecture, Heterochromatin, and Gene Repression". In: *Cell* 169.5 (May 2017), pp. 780–791.

[10] Amy R Strom et al. "Phase separation drives heterochromatin domain formation". In: *Nature Publishing Group* 547.7662 (July 2017), pp. 241–245.

[11] Benjamin R Sabari et al. "Coactivator condensation at super-enhancers links phase separation and gene control". In: *Science* 361.6400 (July 2018), eaar3958–17.

[12] Marc Boehning et al. “RNA polymerase II clustering through carboxy- terminal domain phase separation”. In: *Nature Structural & Molecular Biology* (Sept. 2018), pp. 1–16.

[13] Yingming Zhao and Benjamin A Garcia. “Comprehensive Catalog of Currently Documented Histone Modifications”. In: *Cold Spring Harbor Perspectives in Biology* 7.9 (Sept. 2015), a025064–21.

[14] He Huang et al. “Lysine benzoylation is a histone mark regulated by SIRT2”. In: *Nature Communications* (Sept. 2018), pp. 1–11.

[15] Christopher M Weber, Srinivas Ramachandran, and Steven Henikoff. “Nucleosomes Are Context-Specific, H2A.Z-Modulated Barriers to RNA Polymerase”. In: *Molecular Cell* 53.5 (Mar. 2014), pp. 819–830.

[16] Sergei Rudnizky et al. “H2A.Z controls the stability and mobility of nucleosomes to regulate expression of the LH genes”. In: *Nature Communications* 7 (Sept. 2016), pp. 12958–12.

[17] Luis M Soares et al. “Determinants of Histone H3K4 Methylation Patterns”. In: *Molecular Cell* 68.4 (Nov. 2017), 773–784.e7.

[18] Mira C Patel et al. “BRD4 coordinates recruitment of pause release factor P-TEFb and the pausing complex NELF/DSIF to regulate transcription elongation of interferon-stimulated genes.” In: *Molecular and Cellular Biology* 33.12 (June 2013), pp. 2497–2507.

[19] Adam G Larson et al. “Liquid droplet formation by HP1α suggests a role for phase separation in heterochromatin”. In: *Nature Publishing Group* 547.7662 (July 2017), pp. 236–240.

[20] Carlos Rivera et al. “Histone lysine methylation and chromatin replication”. In: *BBA - Gene Regulatory Mechanisms* 1839.12 (Dec. 2014), pp. 1433–1439.

[21] Frank Lyko. “The DNA methyltransferase family: a versatile toolkit for epigenetic regulation”. In: *Nature Publishing Group* 19.2 (Oct. 2017), pp. 81–92.

[22] Magnolia Bostick et al. “UHRF1 plays a role in maintaining DNA methylation in mammalian cells”. In: *Science* 317.5845 (2007), pp. 1760–1764.

[23] Aurélie Teissandier and Déborah Bourc’his. “Gene body DNA methylation conspires with H3K36me3 to preclude aberrant transcription”. In: *The EMBO Journal* 36.11 (June 2017), pp. 1471–1473.

[24] B D Strahl and C D Allis. “The language of covalent histone modifications”. In: *Nature* 403.6765 (2000), pp. 41–45.

[25] Ruihan Zhang, Jochen Erler, and Jörg Langowski. “Histone Acetylation Regulates Chromatin Accessibility: Role of H4K16 in Inter-nucleosome Interaction”. In: *Biophysj* 112.3 (Feb. 2017), pp. 450–459.

[26] Oliver J Rando. “Combinatorial complexity in chromatin structure and function: revisiting the histone code.” In: *Current Opinion in Genetics & Development* 22.2 (Apr. 2012), pp. 148–155.

[27] Chih Long Liu et al. “Single-nucleosome mapping of histone modifications in S. cerevisiae.” In: *PLOS Biology* 3.10 (Oct. 2005), e328.

[28] Catherine A Musselman et al. “Perceiving the epigenetic landscape through histone readers”. In: *Nature Structural & Molecular Biology* 19.12 (Dec. 2012), pp. 1218–1227.

[29] Panagis Filippakopoulos et al. “Histone recognition and large-scale structural analysis of the human bromodomain family.” In: *Cell* 149.1 (Mar. 2012), pp. 214–231.

[30] Marie Jung et al. “Affinity map of bromodomain protein 4 (BRD4) interactions with the histone H4 tail and the small molecule inhibitor JQ1.” In: *Journal of Biological Chemistry* 289.13 (Mar. 2014), pp. 9304–9319.

[31] Alessio Zippo et al. “PIM1-dependent phosphorylation of histone H3 at serine 10 is required for MYC-dependent transcriptional activation and oncogenic transformation.” In: *Nature Cell Biology* 9.8 (Aug. 2007), pp. 932–944.

[32] Alessio Zippo et al. “Histone crosstalk between H3S10ph and H4K16ac generates a histone code that mediates transcription elongation.” In: *Cell* 138.6 (Sept. 2009), pp. 1122–1136.

[33] Shigeki Iwase et al. “ATRX ADD domain links an atypical histone methylation recognition mechanism to human mental-retardation syndrome”. In: *Nature Structural & Molecular Biology* 18.7 (June 2011), pp. 769–776.

[34] Giulia Saredi et al. “H4K20me0 marks post-replicative chromatin and recruits the TONSL–MMS22L DNA repair complex”. In: *Nature* 534.7609 (June 2016), pp. 714–718.

[35] Adrian Drazic et al. “The world of protein acetylation.” In: *Biochimica et biophysica acta* 1864.10 (Oct. 2016), pp. 1372–1401.

[36] Kenneth K Lee and Jerry L Workman. “Histone acetyltransferase complexes: one size doesn’t fit all”. In: *Nature Reviews Molecular Cell Biology* 8.4 (Apr. 2007), pp. 284–295.

[37] Yong Cai et al. “Subunit composition and substrate specificity of a MOF-containing histone acetyltransferase distinct from the male-specific lethal (MSL) complex.” In: *Journal of Biological Chemistry* 285.7 (Feb. 2010), pp. 4268–4272.

[38] Christopher E Berndsen and John M Denu. “Catalysis and substrate selection by histone/protein lysine acetyltransferases”. In: *Current Opinion in Structural Biology* 18.6 (Dec. 2008), pp. 682–689.

[39] Yulia A Medvedeva et al. “EpiFactors: a comprehensive database of human epigenetic factors and complexes.” In: *Database* 2015 (2015), bav067–10.

[40] Edward Seto and Minoru Yoshida. “Erasers of histone acetylation: the histone deacetylase enzymes.” In: *Cold Spring Harbor Perspectives in Biology* 6.4 (Apr. 2014), a018713–a018713.

[41] Sang Beom Seo et al. “KDM2B is a histone H3K79 demethylase and induces transcriptional repression via SIRT1-mediated chromatin silencing”. In: *bioRxiv* (Sept. 2018), pp. 1–37.

[42] Kwangbeom Hyun et al. “Writing, erasing and reading histone lysine methylations.” In: *Experimental & molecular medicine* 49.4 (Apr. 2017), e324–e324.

[43] Aster H Juan et al. “Roles of H3K27me2 and H3K27me3 Examined during Fate Specification of Embryonic Stem Cells.” In: *Cell Reports* 17.5 (Oct. 2016), pp. 1369–1382.

[44] Karin J Ferrari et al. “Polycomb-Dependent H3K27me1 and H3K27me2 Regulate Active Transcription and Enhancer Fidelity”. In: *Molecular Cell* 53.1 (Jan. 2014), pp. 49–62.

[45] Hui Wu et al. “Histone methyltransferase G9a contributes to H3K27 methylation in vivo.” In: *Nature Publishing Group* 21.2 (Feb. 2011), pp. 365–367.

[46] Gundula Streubel et al. “The H3K36me2 Methyltransferase Nsd1 Demarcates PRC2-Mediated H3K27me2 and H3K27me3 Domains in Embryonic Stem Cells.” In: *Molecular Cell* 70.2 (Apr. 2018), 371–379.e5.

[47] Adam F Kebede et al. “Histone propionylation is a mark of active chromatin”. In: *Nature Structural & Molecular Biology* 24.12 (Oct. 2017), pp. 1048–1056.

[48] Johayra Simithy et al. “Characterization of histone acylations links chromatin modifications with metabolism”. In: *Nature Communications* (Oct. 2017), pp. 1–13.

[49] Gregory D Bowman and Michael G Poirier. “Post-translational modifications of histones that influence nucleosome dynamics.” In: *Chemical Reviews* 115.6 (Mar. 2015), pp. 2274–2295.

[50] Alexandra Stützer et al. “Modulations of DNA Contacts by Linker Histones and Post-translational Modifications Determine the Mobility and Modifiability of Nucleosomal H3 Tails”. In: *Molecular Cell* 61.2 (Jan. 2016), pp. 247–259.

[51] Cedric R Clapier et al. “Mechanisms of action and regulation of ATP-dependent chromatin-remodelling complexes.” In: *Nature Reviews Molecular Cell Biology* 18.7 (July 2017), pp. 407–422.

[52] Yan Li et al. “Integrative analysis of 111 reference human epigenomes”. In: *Nature* 518.7539 (Feb. 2015), pp. 317–330.

[53] Jianxing Feng et al. “Identifying ChIP-seq enrichment using MACS.” In: *Nature Protocols* 7.9 (Sept. 2012), pp. 1728–1740.

[54] Vivien Marx. “Finding the right antibody for the job.” In: *Nature Methods* 10.8 (July 2013), pp. 703–707.

[55] Thea a Egelhofer et al. “An assessment of histone-modification antibody quality.” In: *Nature Structural & Molecular Biology* 18.1 (Jan. 2011), pp. 91–93.

[56] Sangita Pal et al. “The Commercial Antibodies Widely Used to Measure H3 K56 Acetylation Are Non-Specific in Human and Drosophila Cells”. In: *PLoS ONE* 11.5 (May 2016), e0155409–20.

[57] Rohan N Shah et al. “Examining the Roles of H3K4 Methylation States with Systematically Characterized Antibodies”. In: *Molecular Cell* (Sept. 2018), pp. 1–24.

[58] Justin S Becker et al. “Genomic and Proteomic Resolution of Heterochromatin and Its Restriction of Alternate Fate Genes.” In: *Molecular Cell* 68.6 (Dec. 2017), 1023–1037.e15.

[59] Terrence S Furey. “ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions.” In: *Nature Reviews Genetics* 13.12 (Dec. 2012), pp. 840–852.

[60] Jainab Khatun, Kevin Ramkissoon, and Morgan C Giddings. “Fragmentation Characteristics of Collision-Induced Dissociation in MALDI TOF/TOF Mass Spectrometry”. In: *Analytical Chemistry* 79.8 (Apr. 2007), pp. 3032–3040.

[61] Benjamin A Garcia et al. “Chemical derivatization of histones for facilitated analysis by mass spectrometry.” In: *Nature Protocols* 2.4 (2007), pp. 933–938.

[62] Pedro Rodriguez-Collazo, Sanford H Leuba, and Jordanka Zlatanova. “Robust methods for purification of histones from cultured mammalian cells with the preservation of their native modifications.” In: *Nucleic Acids Research* 37.11 (June 2009), e81–e81.

[63] Tobias M Maile et al. “Mass spectrometric quantification of histone post-translational modifications by a hybrid chemical labeling method.” In: *Molecular & Cellular Proteomics* 14.4 (Apr. 2015), pp. 1148–1158.

[64] Zuo-Fei Yuan et al. “EpiProfile Quantifies Histone Peptides With Modifications by Extracting Retention Time and Intensity in High-resolution Mass Spectra.” In: *Molecular & Cellular Proteomics* 14.6 (June 2015), pp. 1696–1707.

[65] Zuo-Fei Yuan et al. “EpiProfile 2.0: A Computational Platform for Processing Epi-Proteomics Mass Spectrometry Data”. In: *Journal of Proteome Research* (2018), acs.jproteome.8b00133.

[66] Annie Moradian et al. “The top-down, middle-down, and bottom-up mass spectrometry approaches for characterization of histone variants and their post-translational modifications.” In: *PROTEOMICS* 14.4-5 (Mar. 2014), pp. 489–497.

[67] Xiulan Chen et al. “Quantitative proteomics using SILAC: Principles, applications, and developments.” In: *PROTEOMICS* 15.18 (Sept. 2015), pp. 3175–3192.

[68] Till Bartke et al. “Nucleosome-Interacting Proteins Regulated by DNA and Histone Methylation”. In: *Cell* 143.3 (Oct. 2010), pp. 470–484.

[69] Kyosuke Nakamura et al. “H4K20me0 recognition by BRCA1-BARD1 directs homologous recombination to sister chromatids.” In: *Nature Publishing Group* 21.3 (Mar. 2019), pp. 311–318.

[70] Andrew Flaus. “Principles and practice of nucleosome positioning in vitro”. In: *Frontiers in Life Science* 5.1-2 (June 2011), pp. 5–27.

[71] Artem Barski et al. “High-resolution profiling of histone methylations in the human genome.” In: *Cell* 129.4 (May 2007), pp. 823–837.

[72] Zhibin Wang et al. “Combinatorial patterns of histone acetylations and methylations in the human genome.” In: *Nature Genetics* 40.7 (July 2008), pp. 897–903.

[73] Minjia Tan et al. “Identification of 67 histone marks and histone lysine crotonylation as a new type of histone modification.” In: *Cell* 146.6 (Sept. 2011), pp. 1016–1028.

[74] V Schwammle et al. “Large Scale Analysis of Co-existing Post-translational Modifications in Histone Tails Reveals Global Fine Structure of Cross-talk”. In: *Molecular & Cellular Proteomics* 13.7 (July 2014), pp. 1855–1865.

[75] Marco Y Hein et al. “A human interactome in three quantitative dimensions organized by stoichiometries and abundances.” In: *Cell* 163.3 (Oct. 2015), pp. 712–723.

[76] Michiel Vermeulen et al. “Quantitative interaction proteomics and genome-wide profiling of epigenetic histone marks and their readers.” In: *Cell* 142.6 (Sept. 2010), pp. 967–980.

[77] Rosa Karlic et al. “Histone modification levels are predictive for gene expression.” In: *Proceedings of the National Academy of Sciences of the United States of America* 107.7 (Feb. 2010), pp. 2926–2931.

[78] Xiaojiang Xu et al. “Application of machine learning methods to histone methylation ChIP-Seq data reveals H4R3me2 globally represses gene expression.” In: *BMC Bioinformatics* 11 (July 2010), p. 396.

[79] Jason Ernst and Manolis Kellis. “Discovery and characterization of chromatin states for systematic annotation of the human genome.” In: *Nature Biotechnology* 28.8 (Aug. 2010), pp. 817–825.

[80] Beibei Xin and Remo Rohs. “Relationship between histone modifications and transcription factor binding is protein family specific.” In: *Genome Research* 28.3 (Jan. 2018), pp. 321–333.

[81] Julia Lasserre, Ho Ryun Chung, and Martin Vingron. “Finding associations among histone modifications using sparse partial correlation networks.” In: 9.9 (2013), e1003168.

[82] Juliane Perner et al. “Inference of interactions between chromatin modifiers and histone modifications: from ChIP-Seq data to chromatin-signaling”. In: *Nucleic Acids Research* 42.22 (Nov. 2014), pp. 13689–13695.

[83] Jian Zhou and Olga G Troyanskaya. “Global quantitative modeling of chromatin factor interactions.” In: 10.3 (Mar. 2014), e1003525.

[84] Scott M Lundberg et al. “ChromNet: Learning the human chromatin network from all ENCODE ChIP-seq data”. In: *Genome biology* (May 2016), pp. 1–19.

[85] The UniProt Consortium. “UniProt: the universal protein knowledgebase”. In: *Nucleic Acids Research* 45.D1 (Jan. 2017), pp. D158–D169.

[86] Birgit H M Meldal et al. “The complex portal - an encyclopaedia of macromolecular complexes”. In: *Nucleic Acids Research* 43.D1 (Oct. 2014), pp. D479–D484.

[87] Andrew Chatr-aryamontri et al. “The BioGRID interaction database: 2017 update”. In: *Nucleic Acids Research* 45.D1 (Jan. 2017), pp. D369–D379.

[88] Malte Beringer et al. “EPOP Functionally Links Elongin and Polycomb in Pluripotent Stem Cells.” In: *Molecular Cell* 64.4 (Nov. 2016), pp. 645–658.

[89] Narkhyun Bae et al. “A transcriptional coregulator, SPIN · DOC, attenuates the coactivator activity of Spindlin1.” In: *Journal of Biological Chemistry* 292.51 (Dec. 2017), pp. 20808–20817.

[90] Jiwen Xin et al. “High-performance web services for querying gene and variant annotation”. In: *Genome biology* (June 2016), pp. 1–7.

[91] Pritam Saha et al. “Application of Principal Component Analysis for Outlier Detection in Heterogeneous Traffic Data”. In: *Procedia Computer Science* 83 (2016), pp. 107–114.

[92] Soren Hauberg, Aasa Feragen, and Michael J Black. “Grassmann Averages for Scalable Robust PCA”. In: *2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2014, pp. 3810–3817.

[93] Jürgen Cox and Matthias Mann. “MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification.” In: *Nature Biotechnology* 26.12 (Dec. 2008), pp. 1367–1372.

[94] Alexander M Crowell et al. “Learning and Imputation for Mass-spec Bias Reduction (LIMBR).” In: *Bioinformatics* (Sept. 2018).

[95] Olivier Ledoit and Michael NMI Wolf. “Honey, I Shrunk the Sample Covariance Matrix”. In: *SSRN Electronic Journal* (2003).

[96] Nicole Krämer, Juliane Schäfer, and Anne-Laure Boulesteix. “Regularized estimation of large-scale gene association networks using graphical Gaussian models.” In: *BMC Bioinformatics* 10.1 (Nov. 2009), p. 384.

[97] Alberto de la Fuente et al. “Discovery of meaningful associations in genomic data using partial correlation coefficients.” In: *Bioinformatics* 20.18 (Dec. 2004), pp. 3565–3574.

[98] Bin Zhang and Steve Horvath. “A general framework for weighted gene co-expression network analysis.” In: *Statistical Applications in Genetics and Molecular Biology* 4.1 (2005), Article17.

[99] Jeremiah J Faith et al. “Large-Scale Mapping and Validation of Escherichia coli Transcriptional Regulation from a Compendium of Expression Profiles”. In: *PLOS Biology* 5.1 (Jan. 2007), e8.

[100] Patrick E Meyer, Frédéric Lafitte, and Gianluca Bontempi. "minet: A R/Bioconductor package for inferring large transcriptional networks using mutual information." In: *BMC Bioinformatics* 9.1 (Oct. 2008), p. 461.

[101] Adam A Margolin et al. "ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context." In: *BMC Bioinformatics* 7 Suppl 1.Suppl 1 (Mar. 2006), S7.

[102] Patrick Emmanuel Meyer. "Information-Theoretic Variable Selection and Network Inference from Microarray Data". PhD thesis. Dec. 2008.

[103] Xiaobo Guo et al. "Inferring Nonlinear Gene Regulatory Networks from Gene Expression Data Based on Distance Correlation". In: *PLoS ONE* 9.2 (Feb. 2014), e87446–7.

[104] Gustavo Stolovitzky, Robert J Prill, and Andrea Califano. "Lessons from the DREAM2 Challenges". In: *Annals of the New York Academy of Sciences* 1158.1 (Mar. 2009), pp. 159–195.

[105] Damian Szklarczyk et al. "The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible". In: *Nucleic Acids Research* 45.D1 (Jan. 2017), pp. D362–D368.

[106] Takaya Saito and Marc Rehmsmeier. "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets." In: *PLoS ONE* 10.3 (2015), e0118432–21.

[107] Fabian Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12.Oct (2011), pp. 2825–2830.

[108] Takaya Saito and Marc Rehmsmeier. "Precrec: fast and accurate precision–recall and ROC curve calculations in R". In: *Bioinformatics* 33.1 (2016), pp. 145–147.

[109] Peter Flach and Meelis Kull. "Precision-Recall-Gain Curves: PR Analysis Done Right". In: *Advances in Neural Information Processing Systems 28* 1 (2015), pp. 838–846.

[110] Sang Cheol Kim et al. "Stouffer's Test in a Large Scale Simultaneous Hypothesis Testing". In: *PLoS ONE* 8.5 (May 2013), e63290–11.

[111] C Spearman. ""General Intelligence," Objectively Determined and Measured". In: *The American Journal of Psychology* 15.2 (Apr. 1904), p. 201.

[112] L L Thurstone. "The vectors of mind." In: *Psychological Review* 41.1 (1934), pp. 1–32.

[113] Stanley A Mulaik. "A brief history of the philosophical foundations of exploratory factor analysis". In: *Multivariate Behavioral Research* 22.3 (1987), pp. 267–305.

[114] Christopher M Bishop. *Pattern Recognition and Machine Learning*. Springer Verlag, Aug. 2006.

[115] Kevin P Murphy. *Machine Learning*. A Probabilistic Perspective. MIT Press, Aug. 2012.

[116] D Barber. *Bayesian reasoning and machine learning*. Nov. 2017.

[117] William Revelle. “psych”. In: (2018).

[118] Marley W Watkins. “Exploratory Factor Analysis: A Guide to Best Practice”. In: *Journal of Black Psychology* 19.5 (Apr. 2018), pp. 009579841877180–28.

[119] Samuel B Green et al. “Type I and Type II Error Rates and Overall Accuracy of the Revised Parallel Analysis Method for Determining the Number of Factors”. In: *Educational and Psychological Measurement* 75.3 (Sept. 2014), pp. 428–457.

[120] Johan Braeken and Marcel A L M van Assen. “An empirical Kaiser criterion.” In: *Psychological Methods* 22.3 (Sept. 2017), pp. 450–466.

[121] Robin K Henson and J Kyle Roberts. “Use of Exploratory Factor Analysis in Published Research”. In: *Educational and Psychological Measurement* 66.3 (July 2016), pp. 393–416.

[122] Michael W Browne. “An Overview of Analytic Rotation in Exploratory Factor Analysis”. In: *Multivariate Behavioral Research* 36.1 (June 2010), pp. 111–150.

[123] Coen A Bernaards and Robert I Jennrich. “Gradient Projection Algorithms and Software for Arbitrary Rotation Criteria in Factor Analysis”. In: *Educational and Psychological Measurement* 65.5 (Oct. 2005), pp. 676–696.

[124] Marley W Watkins. “Exploratory Factor Analysis: A Guide to Best Practice”. In: *Journal of Black Psychology* 19.5 (Apr. 2018), pp. 009579841877180–28.

[125] Cadeyrn J Gaskin and Brenda Happell. “On exploratory factor analysis: A review of recent evidence, an assessment of current practice, and recommendations for future use”. In: *International Journal of Nursing Studies* 51.3 (Mar. 2014), pp. 511–521.

[126] R A Darton. “Rotation in Factor Analysis”. In: *The Statistician* 29.3 (1980), p. 167.

[127] Urbano Lorenzo-Seva. “A factor simplicity index”. In: *Psychometrika* 68.1 (2003), pp. 49–60.

[128] P M Bentler. “Factor simplicity index and transformations”. In: *Psychometrika* 42.2 (1977), pp. 277–295.

[129] Robert I Jennrich and Peter M Bentler. “Exploratory Bi-Factor Analysis”. In: *Psychometrika* 76.4 (June 2011), pp. 537–549.

[130] Suitbert Ertel. “Exploratory factor analysis revealing complex structure”. In: *Personality and Individual Differences* 50.2 (Jan. 2011), pp. 196–200.

[131] David J Bartholomew. *History of Factor Analysis: A Statistical Perspective*. 2nd ed. Vol. 28. Chichester, UK: American Cancer Society, 2014.

[132] James W Grice. “Computing and evaluating factor scores.” In: *Psychological Methods* 6.4 (2001), pp. 430–450.

[133] J H Steiger and Peter H Schönemann. "A history of factor indeterminacy". In: *Theory Construction and Data Analysis in the Behavioral Sciences*. Jossey-Bass Incorporated Pub, 1978.

[134] Louis Guttman. "The determinacy of factor score matrices with implications for five other basic problems of common-factor theory". In: *British Journal of Statistical Psychology* 8.2 (Nov. 1955), pp. 65–81.

[135] Ryne Estabrook and Michael Neale. "A Comparison of Factor Score Estimation Methods in the Presence of Missing Data: Reliability and an Application to Nicotine Dependence". In: *Multivariate Behavioral Research* 48.1 (Jan. 2013), pp. 1–27.

[136] Ivan V Kulakovskiy et al. "HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis". In: *Nucleic Acids Research* 46.D1 (Nov. 2017), pp. D252–D259.

[137] Charles E Grant, Timothy L Bailey, and William Stafford Noble. "FIMO: scanning for occurrences of a given motif". In: *Bioinformatics* 27.7 (Feb. 2011), pp. 1017–1018.

[138] Guangchuang Yu et al. "clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters". In: *OMICS: A Journal of Integrative Biology* 16.5 (May 2012), pp. 284–287.

[139] Van S Tompkins et al. "A novel nuclear interactor of ARF and MDM2 (NIAM) that maintains chromosomal stability." In: *Journal of Biological Chemistry* 282.2 (Jan. 2007), pp. 1322–1333.

[140] Thomas Rolland et al. "A proteome-scale map of the human interactome network." In: *Cell* 159.5 (Nov. 2014), pp. 1212–1226.

[141] Mercedes Pardo et al. "Myst2/Kat7 histone acetyltransferase interaction proteomics reveals tumour-suppressor Niam as a novel binding partner in embryonic stem cells". In: *Scientific Reports* (Aug. 2017), pp. 1–14.

[142] Anders Christiansen and Lars Dyrskjøt. "The functional role of the novel biomarker karyopherin α 2 (KPNA2) in cancer". In: *Cancer Letters* 331.1 (Apr. 2013), pp. 18–23.

[143] Stefanie Hügel et al. "Identification of Importin α 7 Specific Transport Cargoes Using a Proteomic Screening Approach". In: *Molecular & Cellular Proteomics* 13.5 (Apr. 2014), pp. 1286–1298.

[144] Anthony B Mak et al. "A Lentiviral Functional Proteomics Approach Identifies Chromatin Remodeling Complexes Important for the Induction of Pluripotency". In: *Molecular & Cellular Proteomics* 9.5 (May 2010), pp. 811–823.

[145] Toshio Tando et al. "Requiem Protein Links RelB/p52 and the Brm-type SWI/SNF Complex in a Noncanonical NF-κB Pathway". In: *Journal of Biological Chemistry* 285.29 (July 2010), pp. 21951–21960.

[146] Shu-Chun Teng et al. "Importin KPNA2, NBS1, DNA Repair and Tumorigenesis". In: *Journal of Molecular Histology* 37.5-7 (June 2006), pp. 293–299.

[147] Stephanie L Smith-Roe et al. “SWI/SNF complexes are required for full activation of the DNA-damage response”. In: *Oncotarget* 6.2 (Dec. 2014), pp. 732–745.

[148] Dongyi Xu et al. “Rif1 provides a new DNA-binding interface for the Bloom syndrome complex to maintain normal replication”. In: *The EMBO Journal* 29.18 (Aug. 2010), pp. 3140–3155.

[149] Chen Ling et al. “Bloom syndrome complex promotes FANCM recruitment to stalled replication forks and facilitates both repair and traverse of DNA interstrand crosslinks”. In: *Cell Discovery* 2.1 (Dec. 2016), pp. 32351–18.

[150] Cornelia G Spruijt et al. “Dynamic Readers for 5-(Hydroxy)Methylcytosine and Its Oxidized Derivatives”. In: *Cell* 152.5 (Feb. 2013), pp. 1146–1159.

[151] Joshua M Gilmore et al. “WDR76 Co-Localizes with Heterochromatin Related Proteins and Rapidly Responds to DNA Damage”. In: *PLoS ONE* 11.6 (June 2016), e0155492–16.

[152] Kelly A Hoadley et al. “Structure and Cellular Roles of the RMI Core Complex from the Bloom Syndrome Dissolvasome”. In: *Structure/Folding and Design* 18.9 (Sept. 2010), pp. 1149–1158.

[153] Yilun Liu and Stephen C West. “More complexity to the Bloom’s syndrome complex.” In: *Genes & Development* 22.20 (Oct. 2008), pp. 2737–2742.

[154] Donguk Kang et al. “Interaction of replication protein A with two acidic peptides from human Bloom syndrome protein”. In: *FEBS Letters* 592.4 (Feb. 2018), pp. 547–558.

[155] Jay Yang et al. “BLM and RMI1 Alleviate RPA Inhibition of TopoIIIα Decatenase Activity”. In: *PLoS ONE* 7.7 (July 2012), e41208–6.

[156] Suhn Kyong Rhie et al. “ZFX acts as a transcriptional activator in multiple types of human tumors by binding downstream of transcription start sites at the majority of CpG island promoters.” In: *Genome Research* 28.3 (Feb. 2018), pp. 310–320.

[157] Moon Kyoo Jang et al. “The bromodomain protein Brd4 is a positive regulatory component of P-TEFb and stimulates RNA polymerase II-dependent transcription.” In: *Molecular Cell* 19.4 (Aug. 2005), pp. 523–534.

[158] Shaila Rahman et al. “The Brd4 extraterminal domain confers transcription activation independent of pTEFb by recruiting multiple proteins, including NSD3.” In: *Molecular and Cellular Biology* 31.13 (July 2011), pp. 2641–2652.

[159] Saumya M Sankaran et al. “A PWWP Domain of Histone-Lysine N-Methyltransferase NSD2 Binds to Dimethylated Lys-36 of Histone H3 and Regulates NSD2 Function at Chromatin.” In: *Journal of Biological Chemistry* 291.16 (Apr. 2016), pp. 8465–8474.

[160] Su Qin and Jinrong Min. “Structure and function of the nucleosome-binding PWWP domain.” In: *Trends in Biochemical Sciences* 39.11 (Nov. 2014), pp. 536–547.

[161] Sebastian Pünzeler. “PWWP2A : a novel H2A . Z nucleosome interactor involved in cell cycle regulation”. PhD thesis. München, 2015.

[162] Sebastian Pünzeler et al. “Multivalent binding of PWWP2A to H2A.Z regulates mitosis and neural crest differentiation.” In: *The EMBO Journal* 36.15 (Aug. 2017), pp. 2263–2279.

[163] Artyom a Alekseyenko et al. “Ectopic protein interactions within BRD4–chromatin complexes drive oncogenic megadomain formation in NUT midline carcinoma”. In: *Proceedings of the National Academy of Sciences* 114.21 (May 2017), E4184–E4192.

[164] Liis Uusküla-Reimand et al. “Topoisomerase II beta interacts with cohesin and CTCF at topological domain borders.” In: *Genome biology* 17.1 (Aug. 2016), p. 182.

[165] Claudia Tapia-Alveal et al. “H2A.Z-dependent regulation of cohesin dynamics on chromosome arms.” In: *Molecular and Cellular Biology* 34.11 (June 2014), pp. 2092–2104.

[166] T I Alam. “Human mitochondrial DNA is packaged with TFAM”. In: *Nucleic Acids Research* 31.6 (2003), pp. 1640–1645.

[167] Maria E Valieva et al. “Large-scale ATP-independent nucleosome unfolding by a histone chaperone”. In: *Nature Publishing Group* 23.12 (Nov. 2016), pp. 1111–1116.

[168] Joel C Eissenberg and Sarah C R Elgin. “HP1a: a structural chromosomal protein regulating transcription.” In: *Trends in genetics : TIG* 30.3 (Mar. 2014), pp. 103–110.

[169] Weixiang Wang et al. “Nucleolar protein Spindlin1 recognizes H3K4 methylation and stimulates the expression of rRNA genes.” In: *EMBO reports* 12.11 (Oct. 2011), pp. 1160–1166.

[170] X S NAN, R R MEEHAN, and A BIRD. “Dissection of the Methyl-Cpg Binding Domain From the Chromosomal Protein Mecp2”. In: *Nucleic Acids Research* 21.21 (1993), pp. 4886–4892.

[171] Raphaël Margueron et al. “Ezh1 and Ezh2 maintain repressive chromatin through different mechanisms.” In: *Molecular Cell* 32.4 (Nov. 2008), pp. 503–518.

[172] Guillermo Pablo Vicent et al. “Two chromatin remodeling activities cooperate during activation of hormone responsive promoters.” In: *PLoS Genetics* 5.7 (July 2009), e1000567.

[173] Jason T Huff et al. “Reciprocal intronic and exonic histone modification regions in humans.” In: *Nature Publishing Group* 17.12 (Dec. 2010), pp. 1495–1499.

[174] Geoffrey P Dann et al. “ISWI chromatin remodellers sense nucleosome modifications to determine substrate preference”. In: *Nature* 548.7669 (Aug. 2017), pp. 607–611.

[175] Jason Ernst and Manolis Kellis. “ChromHMM: automating chromatin-state discovery and characterization.” In: *Nature Methods* 9.3 (Feb. 2012), pp. 215–216.

[176] ENCODE Project Consortium. "The ENCODE (ENCyclopedia Of DNA Elements) Project." In: *Science* 306.5696 (Oct. 2004), pp. 636–640.

[177] Jeanne Chèneby et al. "ReMap 2018: an updated atlas of regulatory regions from an integrative analysis of DNA-binding ChIP-seq experiments". In: *Nucleic Acids Research* 46.D1 (Nov. 2017), pp. D267–D275.

[178] Ben Langmead et al. "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome." In: *Genome biology* 10.3 (2009), R25.

[179] Cristian Coarfa et al. "Pash 3.0: A versatile software package for read mapping and integrative analysis of genomic and epigenomic variation using massively parallel DNA sequencing." In: *BMC Bioinformatics* 11.1 (Nov. 2010), p. 572.

[180] H Li and R Durbin. "Fast and accurate short read alignment with Burrows-Wheeler transform". In: *Bioinformatics* 25.14 (July 2009), pp. 1754–1760.

[181] Mehran Karimzadeh et al. "Umap and Bismap: quantifying genome and methylome mappability". In: *bioRxiv* (May 2017), pp. 1–15.

[182] Yong Zhang et al. "Model-based analysis of ChIP-Seq (MACS)." In: *Genome biology* 9.9 (2008), R137.

[183] S G Landt et al. "ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia". In: *Genome Research* 22.9 (Sept. 2012), pp. 1813–1831.

[184] ENCODE Project Consortium and The ENCODE Project Consortium. "An integrated encyclopedia of DNA elements in the human genome." In: *Nature Publishing Group* 489.7414 (Sept. 2012), pp. 57–74.

[185] Luis Camarillo Guerrero. "Holistic integration of ChIP-seq data to model epigenetic binding interactions." MA thesis. London, Sept. 2016.

[186] R Herbrich. "Support vector learning for ordinal regression". In: *9th International Conference on Artificial Neural Networks: ICANN '99*. IEE, 1999, pp. 97–102.

[187] Raffaella Di Micco et al. "Control of Embryonic Stem Cell Identity by BRD4-Dependent Transcriptional Elongation of Super-Enhancer-Associated Pluripotency Genes". In: *Cell Reports* 9.1 (Oct. 2014), pp. 234–247.

[188] Nilgun Tasdemir et al. "BRD4 Connects Enhancer Remodeling to Senescence Immune Surveillance." In: *Cancer Discovery* 6.6 (June 2016), pp. 612–629.

[189] Eliezer Calo and Joanna Wysocka. "Modification of enhancer chromatin: what, how, and why?" In: *Molecular Cell* 49.5 (Mar. 2013), pp. 825–837.

[190] Alyson G Wilson. *Clustering Binary Data with Bernoulli Mixture Models*. Sept. 2014. URL: `https://www4.stat.ncsu.edu/~wilson/prelim/Review1.pdf`.

[191] B Marlin. "Missing data problems in machine learning". PhD thesis. Toronto, 2008.

[192] C Biernacki, G Celeux, and G Govaert. "Assessing a mixture model for clustering with the integrated completed likelihood". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22.7 (July 2000), pp. 719–725.

[193] Jennifer Pohle et al. "Selecting the Number of States in Hidden Markov Models: Pragmatic Solutions Illustrated Using Animal Movement". In: *Journal of Agricultural, Biological and Environmental Statistics* 22.3 (June 2017), pp. 270–293.

[194] Marco Bertoletti, Nial Friel, and Riccardo Rastelli. "Choosing the number of clusters in a finite mixture model using an exact Integrated Completed Likelihood criterion". In: *arXiv.org* (Nov. 2014). arXiv: `1411.4257v2 [stat.CO]`.

[195] Alfons Juan, José García-Hernández, and Enrique Vidal. "EM Initialisation for Bernoulli Mixture Learning". In: *Structural, Syntactic, and Statistical Pattern Recognition*. Berlin, Heidelberg: Springer, Berlin, Heidelberg, Aug. 2004, pp. 635–643.

[196] Weijun Feng et al. "PHF8 activates transcription of rRNA genes through H3K4me3 binding and H3K9me1/2 demethylation". In: *Nature Structural & Molecular Biology* 17.4 (Apr. 2010), pp. 445–450.

[197] Aniruddha Chatterjee et al. "Comparison of alignment software for genome-wide bisulphite sequence data". In: *Nucleic Acids Research* 40.10 (Feb. 2012), e79–e79.

[198] Hideyuki Yoshida et al. "Brd4 bridges the transcriptional regulators, Aire and P-TEFb, to promote elongation of peripheral-tissue antigen transcripts in thymic stromal cells." In: *Proceedings of the National Academy of Sciences of the United States of America* 112.32 (Aug. 2015), E4448–57.

[199] Jian Shi et al. "Disrupting the interaction of BRD4 with diacetylated Twist suppresses tumorigenesis in basal-like breast cancer." In: *Cancer Cell* 25.2 (Feb. 2014), pp. 210–225.

[200] Eric Metzger et al. "Phosphorylation of histone H3 at threonine 11 establishes a novel chromatin mark for transcriptional regulation". In: *Nature Cell Biology* 10.1 (Dec. 2007), pp. 53–60.

[201] Gary LeRoy et al. "Proteogenomic characterization and mapping of nucleosomes decoded by Brd and HP1 proteins." In: *Genome biology* 13.8 (Aug. 2012), R68.

[202] Zhangli Su and John M Denu. "MARCC (Matrix-Assisted Reader Chromatin Capture): An Antibody-Free Method to Enrich and Analyze Combinatorial Nucleosome Modifications." In: *Current protocols in molecular biology* 111.July (July 2015), pp. 21.32.1–21.

[203] Jakub Mieczkowski et al. "MNase titration reveals differences between nucleosome occupancy and chromatin accessibility". In: *Nature Communications* 7 (Apr. 2016), pp. 1–11.

[204] M Perry and R Chalkley. "Histone acetylation increases the solubility of chromatin and occurs sequentially over most of the chromatin. A novel model for the biological role of histone acetylation." In: *Journal of Biological Chemistry* 257.13 (July 1982), pp. 7336–7347.

[205] Marta de Frutos et al. “Aggregation of Nucleosomes by Divalent Cations”. In: *Biophysj* 81.2 (Aug. 2001), pp. 1127–1132.

[206] B a Lewis et al. “Adenovirus E1A proteins interact with the cellular YY1 transcription factor.” In: *Journal of Virology* 69.3 (Mar. 1995), pp. 1628–1636.

[207] Andrey Tvardovskiy et al. “Accumulation of histone variant H3.3 with age is associated with profound changes in the histone methylation landscape.” In: *Nucleic Acids Research* 45.16 (Sept. 2017), pp. 9272–9289.

[208] Jan Michael Peters. “The anaphase promoting complex/cyclosome: a machine designed to destroy.” In: *Nature Reviews Molecular Cell Biology* 7.9 (Sept. 2006), pp. 644–656.

[209] S Guelman et al. “The Double-Histone-Acetyltransferase Complex ATAC Is Essential for Mammalian Development”. In: *Molecular and Cellular Biology* 29.5 (Feb. 2009), pp. 1176–1188.

[210] Joonyoung Her et al. “Factors forming the BRCA1-A complex orchestrate BRCA1 recruitment to the sites of DNA damage”. In: *Acta Biochimica et Biophysica Sinica* 48.7 (July 2016), pp. 658–664.

[211] Y S Dong et al. “Regulation of BRCC, a holoenzyme complex containing BRCA1 and BRCA2, by a signalosome-like subunit and its role in DNA repair”. In: *Molecular Cell* 12.5 (Nov. 2003), pp. 1087–1099.

[212] Yujiang Shi et al. “Coordinated histone modifications mediated by a CtBP co-repressor complex.” In: *Nature* 422.6933 (Apr. 2003), pp. 735–738.

[213] Radhika a Varier et al. “Recruitment of the Mammalian Histone-modifying EMSY Complex to Target Genes Is Regulated by ZNF131.” In: *Journal of Biological Chemistry* 291.14 (Apr. 2016), pp. 7313–7324.

[214] Duane D Winkler and Karolin Luger. “The histone chaperone FACT: structural insights and mechanisms for nucleosome reorganization.” In: *Journal of Biological Chemistry* 286.21 (May 2011), pp. 18369–18374.

[215] Zhijiang Yan et al. “A histone-fold complex and FANCM form a conserved DNA-remodeling complex to maintain genome stability.” In: *Molecular Cell* 37.6 (Mar. 2010), pp. 865–878.

[216] I A Tchasovnikarova et al. “Epigenetic silencing by the HUSH complex mediates position-effect variegation in human cells”. In: *Science* 348.6242 (June 2015), pp. 1481–1485.

[217] Monica Rienzo and Amelia Casamassimi. “Integrator complex and transcription regulation: Recent findings and pathophysiology.” In: *Biochimica et biophysica acta* 1859.10 (Oct. 2016), pp. 1269–1280.

[218] Maralice Conacci-Sorrell, Lisa McFerrin, and Robert N Eisenman. “An overview of MYC and its interactome.” In: *Cold Spring Harbor Perspectives in Medicine* 4.1 (Jan. 2014), a014357–a014357.

[219] H H Ng et al. “MBD2 is a transcriptional repressor belonging to the MeCP1 histone deacetylase complex.” In: *Nature Genetics* 23.1 (Sept. 1999), pp. 58–61.

[220] Célia Jeronimo and François Robert. “The Mediator Complex: At the Nexus of RNA Polymerase II Transcription”. In: *Trends in Cell Biology* 27.10 (Oct. 2017), pp. 765–783.

[221] K M Trujillo et al. “Nuclease activities in a complex of human recombination and DNA repair factors Rad50, Mre11, and p95.” In: *Journal of Biological Chemistry* 273.34 (Aug. 1998), pp. 21447–21450.

[222] Arven Saunders et al. “The SIN3A/HDAC Corepressor Complex Functionally Cooperates with NANOG to Promote Pluripotency”. In: *Cell Reports* 18.7 (Feb. 2017), pp. 1713–1726.

[223] Karine Jacquet et al. “The TIP60 Complex Regulates Bivalent Chromatin Recognition by 53BP1 through Direct H4K20me Binding and H2AK15 Acetylation.” In: *Molecular Cell* 62.3 (May 2016), pp. 409–421.

[224] Susanne Bornelöv et al. “The Nucleosome Remodeling and Deacetylation Complex Modulates Chromatin Structure at Sites of Active Transcription to Fine-Tune Gene Expression”. In: *Molecular Cell* 71.1 (July 2018), 56–72.e4.

[225] Julien Vandamme et al. “Interaction proteomics analysis of polycomb proteins defines distinct PRC1 complexes in mammalian cells.” In: *Molecular & Cellular Proteomics* 10.4 (Apr. 2011), pp. M110.002642–23.

[226] Robert K McGinty, Ryan C Henrici, and Song Tan. “Crystal structure of the PRC1 ubiquitylation module bound to the nucleosome”. In: *Nature* 514.7524 (Oct. 2014), pp. 591–596.

[227] Zhiyuan Yang et al. “Recruitment of P-TEFb for stimulation of transcriptional elongation by the bromodomain protein Brd4.” In: *Molecular Cell* 19.4 (Aug. 2005), pp. 535–545.

[228] Kelly A Hoadley et al. “Structure and cellular roles of the RMI core complex from the bloom syndrome dissolvasome.” In: *Structure* 18.9 (Sept. 2010), pp. 1149–1158.

[229] Markus Räschle et al. “DNA repair. Proteomics reveals dynamic assembly of repair complexes during bypass of DNA cross-links.” In: *Science* 348.6234 (May 2015), pp. 1253671–1253671.

[230] Robert K Louder et al. “Structure of promoter-bound TFIID and model of human pre-initiation complex assembly”. In: *Nature* 531.7596 (Mar. 2016), pp. 604–609.

[231] M Brand et al. “Identification of TATA-binding protein-free TAF(II)-containing complex subunits suggests a role in nucleosome acetylation and signal transduction”. In: *Journal of Biological Chemistry* 274.26 (1999), pp. 18285–18289.

[232] Ze Li et al. “Structural insights into the YAP and TEAD complex.” In: *Genes & Development* 24.3 (Feb. 2010), pp. 235–240.

[233] Anamika Patel et al. “On the Mechanism of Multiple Lysine Methylation by the Human Mixed Lineage Leukemia Protein-1 (MLL1) Core Complex”. In: *Journal of Biological Chemistry* 284.36 (2009), pp. 24242–24256.

[234] Bin Wang et al. “Abraxas and RAP80 form a BRCA1 protein complex required for the DNA damage response”. In: *Science* 316.5828 (2007), pp. 1194–1198.

[235] Ramasubramanian Sundaramoorthy et al. “Structural reorganization of the chromatin remodeling enzyme Chd1 upon engagement with nucleosomes.” In: *eLife* 6 (Mar. 2017), p. 1405.

[236] MuhChyi Chai et al. “Chromatin remodeler CHD7 regulates the stem cell identity of human neural progenitors”. In: *Genes & Development* 32.2 (Feb. 2018), pp. 165–180.

[237] N Antony et al. “Creb1 regulates late stage mammalian lung development via respiratory epithelial and mesenchymal-independent mechanisms”. In: *Scientific Reports* 6.1 (May 2016), p. 502.

[238] Chin-Tong Ong and Victor G Corces. “CTCF: An architectural protein bridging genome topology and function”. In: *Nature Reviews Genetics* 15.4 (Jan. 2014), pp. 234–246.

[239] Chen Ling et al. “Bloom syndrome complex promotes FANCM recruitment to stalled replication forks and facilitates both repair and traverse of DNA interstrand crosslinks.” In: *Cell Discovery* 2.1 (2016), p. 16047.

[240] Wilfried Ellmeier and Christian Seiser. “Histone deacetylase function in $CD4^+$T cells”. In: *Nature Reviews Immunology* 40 (Jan. 2018), p. 897.

[241] Yu Jiang Shi et al. “Regulation of LSD1 histone demethylase activity by its associated factors.” In: *Molecular Cell* 19.6 (Sept. 2005), pp. 857–864.

[242] Robert J Klose and Yi Zhang. “Regulation of histone methylation by demethylimination and demethylation”. In: *Nature Reviews Molecular Cell Biology* 8.4 (Mar. 2007), pp. 307–318.

[243] Yu-ichi Tsukada et al. “Histone demethylation by a family of JmjC domain-containing proteins”. In: *Nature* 439.7078 (Feb. 2006), pp. 811–816.

[244] David Frescas et al. “JHDM1B/FBXL10 is a nucleolar protein that represses transcription of ribosomal RNA genes”. In: *Nature* 450.7167 (Nov. 2007), pp. 309–313.

[245] Jacky Guy et al. “Reversal of neurological defects in a mouse model of Rett syndrome”. In: *Science* 315.5815 (2007), pp. 1143–1147.

[246] X S NAN et al. “Transcriptional repression by the methyl-CpG-binding protein MeCP2 involves a histone deacetylase complex”. In: *Nature* 393.6683 (1998), pp. 386–389.

[247] Joanna Rorbach et al. “MRM2 and MRM3 are involved in biogenesis of the large subunit of the mitochondrial ribosome.” In: *Molecular biology of the cell* 25.17 (Sept. 2014), pp. 2542–2555.

[248] Antonis S Zervos, Jenő Gyuris, and Roger Brent. “Mxi1, a protein that specifically interacts with Max to bind Myc-Max recognition sites”. In: *Cell* 72.2 (1993), pp. 223–232.

[249] Wen Liu et al. “PHF8 mediates histone H4 lysine 20 demethylation events involved in cell cycle progression”. In: *Nature* 466.7305 (July 2010), pp. 508–512.

[250] Ryu-Suke Nozawa et al. “Human POGZ modulates dissociation of HP1α from mitotic chromosome arms through Aurora B activation”. In: *Nature Cell Biology* 12.7 (June 2010), pp. 719–727.

[251] Lezanne Ooi and Ian C Wood. “Chromatin crosstalk in development and disease: lessons from REST”. In: *Nature Reviews Genetics* 8.7 (July 2007), pp. 544–554.

[252] J Zhang et al. “Assembly factors Rpf2 and Rrs1 recruit 5S rRNA and ribosomal proteins rpL5 and rpL11 into nascent ribosomes”. In: *Genes & Development* 21.20 (Oct. 2007), pp. 2580–2592.

[253] Keijo M Viiri et al. “DNA-binding and -bending activities of SAP30L and SAP30 are mediated by a zinc-dependent module and monophosphoinositides.” In: *Molecular and Cellular Biology* 29.2 (Jan. 2009), pp. 342–356.

[254] Patrick McDonel, Ita Costello, and Brian Hendrich. “Keeping things quiet: roles of NuRD and Sin3 co-repressor complexes during mammalian development.” In: *The international journal of biochemistry & cell biology* 41.1 (Jan. 2009), pp. 108–116.

[255] Kobe C Yuen and Jennifer L Gerton. “Taking cohesin and condensin in context.” In: *PLoS Genetics* 14.1 (Jan. 2018), e1007118.

[256] Andrew J Bannister and Tony Kouzarides. “Regulation of chromatin by histone modifications”. In: *Cell Research* 21.3 (Feb. 2011), pp. 381–395.

[257] Hui Wang et al. “Crystal structure of a TAF1-TAF7 complex in human transcription factor IID reveals a promoter binding module”. In: *Cell Research* 24.12 (Jan. 2014), pp. 1433–1444.

[258] L S Cox and D P Lane. “Tumour suppressors, kinases and clamps: how p53 regulates the cell cycle in response to DNA damage.” In: *BioEssays : news and reviews in molecular, cellular and developmental biology* 17.6 (June 1995), pp. 501–508.

[259] Su Wu et al. “A YY1–INO80 complex regulates genomic stability through homologous recombination–based repair”. In: *Nature Structural & Molecular Biology* 14.12 (Nov. 2007), pp. 1165–1172.

[260] Rudolf Jaenisch et al. “YY1 Is a Structural Regulator of Enhancer-Promoter Loops”. In: *Cell* 171.7 (Dec. 2017), 1573–1588.e28.

[261] Paul Shannon et al. “Cytoscape: a software environment for integrated models of biomolecular interaction networks.” In: *Genome Research* 13.11 (Nov. 2003), pp. 2498–2504.

[262] Carrie A Davis et al. “The Encyclopedia of DNA elements (ENCODE): data portal update”. In: *Nucleic Acids Research* 46.D1 (Nov. 2017), pp. D794–D801.

[263] Daniel R Zerbino et al. “Ensembl 2018”. In: *Nucleic Acids Research* 46.D1 (Nov. 2017), pp. D754–D761.

[264] Donna Maglott et al. “Entrez Gene: gene-centered information at NCBI.” In: *Nucleic Acids Research* 33.Database issue (Jan. 2005), pp. D54–8.

[265] Robert D Finn et al. “InterPro in 2017—beyond protein family and domain annotations”. In: *Nucleic Acids Research* 45.D1 (Jan. 2017), pp. D190–D199.

[266] Peter Langfelder and Steve Horvath. “WGCNA: an R package for weighted correlation network analysis.” In: *BMC Bioinformatics* 9.1 (Dec. 2008), p. 559.

[267] Yaelim Lee, Daechan Park, and Vishwanath R Iyer. “The ATP-dependent chromatin remodeler Chd1 is recruited by transcription elongation factors and maintains H3K4me3/H3K36me3 domains at actively transcribed and spliced genes”. In: *Nucleic Acids Research* 45.12 (Apr. 2017), pp. 7180–7190.

[268] G R Schnitzler. “Isolation of histones and nucleosome cores from mammalian cells.” In: *Current protocols in molecular biology* Chapter 21 (May 2001), Unit 21.5.

[269] Arash Nabbi and Karl Riabowol. “Rapid Isolation of Nuclei from Cells In Vitro.” In: *Cold Spring Harbor Protocols* 2015.8 (Aug. 2015), pp. 769–772.

[270] Zhangli Su et al. “ChIP-less analysis of chromatin states.” In: *Epigenetics & Chromatin* 7.1 (2014), p. 7.

[271] null RuthenLab org null. “Micrococcal nuclease preparation of native mononucleosomes”. In: (2010), pp. 1–5.

[272] Pavlos Fanis et al. “Five friends of methylated chromatin target of protein-arginine-methyltransferase[prmt]-1 (chtop), a complex linking arginine methylation to desumoylation.” In: *Molecular & Cellular Proteomics* 11.11 (Nov. 2012), pp. 1263–1273.

[273] T Williams and R Tjian. “Characterization of a dimerization motif in AP-2 and its function in heterologous DNA-binding proteins.” In: *Science* 251.4997 (Mar. 1991), pp. 1067–1071.

[274] Holger Stark et al. “Systematic Analysis of Human Protein Complexes Identifies Chromosome Segregation Proteins”. In: *Science* 328.5978 (Apr. 2010), pp. 593–599.

[275] Jörg Mansfeld et al. “APC15 drives the turnover of MCC-CDC20 to make the spindle assembly checkpoint responsive to kinetochore attachment.” In: *Nature Cell Biology* 13.10 (Sept. 2011), pp. 1234–1243.

[276] Andrea Giovanni Murachelli et al. “The structure of the ASAP core complex reveals the existence of a Pinin-containing PSAP complex.” In: *Nature Structural & Molecular Biology* 19.4 (Mar. 2012), pp. 378–386.

[277] Kurt Engeland. “Cell cycle arrest through indirect transcriptional repression by p53: I have a DREAM.” In: *Cell Death and Differentiation* 25.1 (Jan. 2018), pp. 114–132.

[278] Salima Daou et al. “The BAP1/ASXL2 Histone H2A Deubiquitinase Complex Regulates Cell Proliferation and Is Disrupted in Cancer.” In: *Journal of Biological Chemistry* 290.48 (Nov. 2015), pp. 28643–28663.

[279] Eleonora Parlanti et al. “Human base excision repair complex is physically associated to DNA replication and cell cycle regulatory proteins.” In: *Nucleic Acids Research* 35.5 (2007), pp. 1569–1577.

[280] Andreas Ruepp et al. “CORUM: the comprehensive resource of mammalian protein complexes–2009.” In: *Nucleic Acids Research* 38.Database issue (Jan. 2010), pp. D497–501.

[281] Susannah Rankin, Nagi G Ayad, and Marc W Kirschner. “Sororin, a substrate of the anaphase-promoting complex, is required for sister chromatid cohesion in vertebrates.” In: *Molecular Cell* 18.2 (Apr. 2005), pp. 185–200.

[282] Hiroki Takai et al. “5-Hydroxymethylcytosine plays a critical role in glioblastomagenesis by recruiting the CHTOP-methylosome complex.” In: *Cell Reports* 9.1 (Oct. 2014), pp. 48–60.

[283] Ali Shilatifard. “The COMPASS Family of Histone H3K4 Methylases: Mechanisms of Regulation in Development and Disease Pathogenesis”. In: *Annual Review of Biochemistry* 81.1 (July 2012), pp. 65–95.

[284] U Kutay et al. “Export of importin alpha from the nucleus is mediated by a specific nuclear transport factor.” In: *Cell* 90.6 (Sept. 1997), pp. 1061–1071.

[285] Erin A White and Michael Glotzer. “Centralspindlin: At the heart of cytokinesis”. In: *Cytoskeleton* 69.11 (Sept. 2012), pp. 882–892.

[286] Christopher Bot et al. “Independent mechanisms recruit the cohesin loader protein NIPBL to sites of DNA damage.” In: *Journal of Cell Science* 130.6 (Mar. 2017), pp. 1134–1146.

[287] Michiel Vermeulen et al. “Quantitative Interaction Proteomics and Genome-wide Profiling of Epigenetic Histone Marks and Their Readers”. In: *Cell* 142.6 (Sept. 2010), pp. 967–980.

[288] Thiyam Ramsing Singh et al. “MHF1-MHF2, a histone-fold-containing protein complex, participates in the Fanconi anemia pathway via FANCM.” In: *Molecular Cell* 37.6 (Mar. 2010), pp. 879–886.

[289] Anderson T Wang and Agata Smogorzewska. “SnapShot: Fanconi anemia and associated proteins.” In: *Cell* 160.1-2 (Jan. 2015), 354–354.e1.

[290] Steven J Deimling, Jonathan B Olsen, and Vincent Tropepe. “The expanding role of the Ehmt2/G9a complex in neurodevelopment”. In: *Neurogenesis* 4.1 (May 2017), e1316888.

[291] T Kamijo et al. “Mitochondrial trifunctional protein deficiency. Catalytic heterogeneity of the mutant enzyme in two patients.” In: *Journal of Clinical Investigation* 93.4 (Apr. 1994), pp. 1740–1747.

[292] Toshimasa Itoh et al. “Structural and functional characterization of a cell cycle associated HDAC1/2 complex reveals the structural basis for complex assembly and nucleosome targeting.” In: *Nucleic Acids Research* 43.4 (Feb. 2015), pp. 2033–2044.

[293] A Kalousi et al. “Casein kinase 1 regulates human hypoxia-inducible factor HIF-1”. In: *Journal of Cell Science* 123.17 (Aug. 2010), pp. 2976–2986.

[294] Taranjit Singh Rai et al. “Human CABIN1 is a functional member of the human HIRA/UBN1/ASF1a histone H3.3 chaperone complex.” In: *Molecular and Cellular Biology* 31.19 (Oct. 2011), pp. 4107–4118.

[295] Doreen Weidensdorfer et al. “Control of c-myc mRNA stability by IGF2BP1-associated cytoplasmic RNPs.” In: *RNA* 15.1 (Jan. 2009), pp. 104–115.

[296] David Valle-García et al. “ATRX binds to atypical chromatin domains at the 3’ exons of zinc finger genes to preserve H3K9me3 enrichment.” In: *Epigenetics* 11.6 (June 2016), pp. 398–414.

[297] Sushma Iyengar and Peggy J Farnham. “KAP1 protein: an enigmatic master regulator of the genome.” In: *Journal of Biological Chemistry* 286.30 (July 2011), pp. 26267–26276.

[298] Ryan W Serra et al. “A KRAS-directed transcriptional silencing pathway that mediates the CpG island methylator phenotype.” In: *eLife* 3.3 (Mar. 2014), e02313–22.

[299] Xiaodan Ding et al. “Mixed Lineage Leukemia 5 (MLL5) Protein Stability Is Cooperatively Regulated by O-GlcNac Transferase (OGT) and Ubiquitin Specific Protease 7 (USP7).” In: *PLoS ONE* 10.12 (2015), e0145023.

[300] Q Feng et al. “Identification and functional characterization of the p66/p68 components of the MeCP1 complex”. In: *Molecular and Cellular Biology* 22.2 (Jan. 2002), pp. 536–546.

[301] Jinsong Zhang et al. “The N-CoR-HDAC3 nuclear receptor corepressor complex inhibits the JNK pathway through the integral subunit GPS2.” In: *Molecular Cell* 9.3 (Mar. 2002), pp. 611–623.

[302] Audrey Dumax-Vorzet, Peristera Roboti, and Stephen High. “OST4 is a subunit of the mammalian oligosaccharyltransferase required for efficient N-glycosylation.” In: *Journal of Cell Science* 126.Pt 12 (June 2013), pp. 2595–2606.

[303] H Irem Baymaz et al. “MBD5 and MBD6 interact with the human PR-DUB complex through their methyl-CpG-binding domain.” In: *PROTEOMICS* 14.19 (Oct. 2014), pp. 2179–2189.

[304] Jeong-heon Lee et al. “Identification and Characterization of a Novel Human PP1 Phosphatase Complex”. In: *Journal of Biological Chemistry* 285.32 (July 2010), pp. 24466–24476.

[305] Judith A Jaehning. “The Paf1 complex: platform or player in RNA polymerase II transcription?” In: *Biochimica et biophysica acta* 1799.5-6 (May 2010), pp. 379–388.

[306] Thomas Wild and Patrick Cramer. "Biogenesis of multisubunit RNA polymerases." In: *Trends in Biochemical Sciences* 37.3 (Mar. 2012), pp. 99–105.

[307] Aaron C Mason et al. "An alternative form of replication protein a prevents viral replication in vitro." In: *Journal of Biological Chemistry* 284.8 (Feb. 2009), pp. 5324–5331.

[308] Maryna Gorelik et al. "Inhibition of SCF ubiquitin ligases by engineered ubiquitin variants that target the Cul1 binding site on the Skp1–F-box interface". In: *Proceedings of the National Academy of Sciences* 113.13 (Mar. 2016), pp. 3527–3532.

[309] Cameron P Bracken et al. "Regulation of Cyclin D1 RNA Stability by SNIP1". In: *Cancer Research* 68.18 (Sept. 2008), pp. 7621–7628.

[310] H Le Hir et al. "The spliceosome deposits multiple proteins 20-24 nucleotides upstream of mRNA exon-exon junctions". In: *The EMBO Journal* 19.24 (Dec. 2000), pp. 6860–6869.

[311] C E Stebbins et al. "Structure of the VHL-ElonginC-ElonginB Complex: Implications for VHL Tumor Suppressor Function". In: *Science* 284.5413 (Apr. 1999), pp. 455–461.

[312] T Kamura et al. "Activation of HIF1alpha ubiquitination by a reconstituted von Hippel-Lindau (VHL) tumor suppressor complex." In: *Proceedings of the National Academy of Sciences* 97.19 (Sept. 2000), pp. 10430–10435.

[313] K Sugasawa et al. "Two human homologs of Rad23 are functionally interchangeable in complex formation and stimulation of XPC repair activity." In: *Molecular and Cellular Biology* 17.12 (Dec. 1997), pp. 6924–6931.

[314] Elisa T Zhang et al. "Architecture of the human XPC DNA repair and stem cell coactivator complex." In: *Proceedings of the National Academy of Sciences of the United States of America* 112.48 (Dec. 2015), pp. 14817–14822.

[315] Annegret Ulke-Lemée et al. "The nuclear PP1 interacting protein ZAP3 (ZAP) is a putative nucleoside kinase that complexes with SAM68, CIA, NF110/45, and HNRNP-G." In: *Biochimica et biophysica acta* 1774.10 (Oct. 2007), pp. 1339–1350.

[316] Yannick Doyon et al. "ING tumor suppressor proteins are critical regulators of chromatin acetylation required for genome expression and perpetuation." In: *Molecular Cell* 21.1 (Jan. 2006), pp. 51–64.

[317] A Kuzmichev et al. "Role of the Sin3-histone deacetylase complex in growth regulation by the candidate tumor suppressor p33ING1". In: *Am Soc Microbiol* ().

[318] Karen T Smith et al. "Human family with sequence similarity 60 member A (FAM60A) protein: a new subunit of the Sin3 deacetylase complex." In: *Molecular & Cellular Proteomics* 11.12 (Dec. 2012), pp. 1815–1828.

Appendix A

Reference

List of acronyms

5-hmC 5-hydroxymethylcytosine.

5-mC 5-methylcytosine.

AP Affinity Purification.

AP-MS Affinity Purification Mass Spectrometry.

ATP adenosine triposphate.

auPRC area under Precision-Recall Curve.

BET Bromodomain and Extra-Terminal.

bHLH-ZIP Basic Helix-Loop-Helix leucine Zipper.

bHSH Basic Helix-Span-Helix.

BIC Bayesian Information Criterion.

BMM Bernoulli Mixture Model.

bZIP Basic Leucine Zipper.

ChIP-seq Chromatin immunoprecipitation sequencing.

CID Collision-Induced Dissociation.

CLR Context-Likelihood of Relatedness.

CTD C-Terminal Domain.

DC Distance Correlation.

DSB Double-Strand Break.

EM Expectation Maximisation.

ET Extra-Terminal.

FA Factor Analysis.

GO Gene Ontology.

HAT histone acetyltransferase.

HDAC histone deacetylase.

HMM Hidden Markov Model.

ICL Integrated Complete Likelihood.

LC Liquid Chromatography.

LETOR Learning to rank.

LOGO Leave One Group Out.

MARCC Matrix-Assisted Reader Chromatin Capture.

MI Mutual Information.

MNase Micrococcal Nuclease.

MS Mass Spectrometry.

NDR Nucleosome Depleted Region.

NGS Next-Generation Sequencing.

NLS Nuclear Localisation Signal.

PCA Principal Component Analysis.

PDF Probability Density Function.

PRC Precision-Recall Curve.

PTM Post-Translational Modification.

ROC Receiver Operating Characteristic.

SILAC Stable Isotope Labelling with Amino acids in cell Culture.

SNAP SILAC nucleosome affinity purification.

spAUC standardised partial Area Under Curve.

TAD Topologically Associated Domain.

TF Transcription Factor.

TGA Trimmed Grassmann Averages.

TMT Tandem Mass Tag.

TSS Transcription Start Site.

List of protein complexes

This section provides a reference to the list of protein complexes referenced by name in this thesis. This list has been compiled by merging the data provided in the EpiFactors[39] and Complex Portal[86] databases as well as through manual literature review. The descriptions here focus on relevance and function. See Table C.2 for details and sources on composition.

60S Ribosomal subunit Larger subnit of the ribosome. 112, 113

AP-1 FOS-JUN *AP-1 transcription factor complex.* Complex consisting of a dimer of AP-1 family of transcription factors, most notably FOS and JUN, that regulate DNA transcription[86]. . 224, 258

APC/C *Anaphase-Promoting Complex / Cyclosome.* Ubiquitin ligase (E3) complex which deposits polyubiquitin chains on its substrates, marking them to be destructed by the proteasome[208]. . 113, 121

ATAC *ATAC complex.* Histone double acetyltransferase complex which acetylates lysines on histones H3 and H4[86, 209] . 25, 102, 117, 121, 225

BAF Members of SWI/SNF family of chromatin remodelling complexes which regulate DNA accessibility in ATP-dependent manner[51]. The exact structure of the complexes is highly context dependent. Annotation from [39]. 30, 31, 102, 108, 109, 117, 120, 131, 226, 229

BCOR Polycomb related complex, also known as *ncPRC1.1*[5]. Contains histone lysine demethylase KDM2B[5]. 225

BHC Histone deacetylation complex[39]. 26

BRCA1-A Protein complex recruited to double strand breaks, responsible for BRCA1 localisation. Necessary for cell cycle progression. Contains deubiquitinase activity for lysine 63 based ubiquitin chains[39, 210]. . 121, 222

BRCC Ubiquitin ligase complex, involved in DNA damage. Ubiquitinates TP53 in vitro[39, 211]. 121, 222

CAF-1 Chromatin assembly complex, plays a part in assembly of histones onto replicating DNA by bringing histones H3 and H4 to the DNA[39]. 120

CHRAC Member of *ISWI* family of chromatin remodellers whose members are involved in establishing regular spacing of the nucleosomes.[51]. 117, 120–122

Cohesin Protein complex which mediates DNA-DNA interactions required to form Topologically Associated Domains (TADs). 17, 111, 112, 121, 229, 231

COMPASS Histone methyltransferase complex, catalyses H3K4me via SETD1A and SETD1B[5, 39] .. 19, 27, 228

COMPASS-like MLL1,2 Histone methylation and acetylation complex, catalyses H3K4me via KMT2A and KMT2B (MLL1, MLL2)[5], as well as acetylation of Histone H4K16 via KAT8[39]. . 19, 27, 110, 121, 222, 225, 226

COMPASS-like MLL3,4 Histone methyltransferase complex, catalyses H3K4me via KMT2C or KMT2D[5]. Contains a H3K27 demethylase KDM6A[5]. . 19, 27, 29, 102, 117, 225, 226

CtBP *C-terminal binding protein (CtBP) complex.* Transcriptional repressor complex, recruited to PXDLS motif-containing transcription factors, facilitates deacetylation of histone tails and methylation of lysine 9 on histone H3[212]. . 228

EMSY *EMSY protein complex.* Protein complex recruited to trimethylated lysine 4 on histone H3. The complex contains a demethylase for this same residue, KDM5A as well as histone deacetylase complex subunit SIN3B. Despite this, expression of EMSY is positively correlated with expression of its target genes[213]. . 102, 120

FACT *FACT: Facilitates chromatin transcription.* Chromatin remodeler complex capable of disrupting the histone-histone and histone-DNA interactions to evict H2A-H2B dimer in order to free access to the nucleosomal DNA[214]. . 112, 113, 125, 229

Fanconi Anemia *Fanconi Anemia complex.* Protein complex named after a genetic disease caused by mutations in the complex members or associated proteins. Involved in DNA repair upon damage or cell-cycle induced errors. Contains an ATPase (FANCM), able to monoubiquitinate proteins downstream its pathway[215]. . 224

HBO1 Histone acetyltransferase complex containing KAT7. Responsible for bulk H4 acetylation in vivo, can acetylate histone H3 as well[39].. 25, 225

HUSH *HUman Silencing Hub*. Protein complex recruited to H3K9me3, facilitates spreading of H3K9me3 via recruitment of SETDB1[216]. . 102, 120

INO80 *INO80 complex*. ATP-dependent chromatin remodeler, member of INO80 family of remodellers along with SRCAP complex. Catalyses eviction of histones H2A.Z and sliding of nucleosomes, involved in DNA repair and transcription activation[51].. 30, 31, 104, 108, 120, 121, 190, 192, 230, 231

Integrator *Integrator complex*. Multiprotein complex, involved in regulation of most RNA Polymerase II transcribed genes. Involved in 3'-end formation of noncoding RNAs, enhancer transcription and transcription elongation[217]. . 102, 117, 120, 121

LSD-CoREST Repressive histone deacetylation complex which interacts with KDM1A (LSD1) lysine demethylase[39]. 26, 28, 224, 225, 228

Mad-Max *Mad-max protein complex*. Dimer of MXD1 (also known as *MAD*) and MAX proteins. Functionally antagonistic with Myc-Max complex, represses transcription at E-box (CACGTG) promoters via interaction with mSin3A complex[218]. . 220, 226

MeCP1 Histone deacetylation complex which potentially links DNA methylation to histone deacetylation[39, 219] . 26

Mediator Seemingly universal co-activator of transcription, involved in seemingly all aspects of it ranging from mRNA processing to chromatin looping[220]. 102, 117, 120, 121

MLL supercomplex A supercomplex generated from a set of highly-overlapping MLL-like complexes in the EpiFactors database[39], specifically complexes CHD8, MLL2/3, MLL4/WBP7. . 117

MOZ/MORF Histone acetyltransferase complex, acetylates histone H3 via KAT6A and KAT6B[39] . 25, 225

MRN *MRE11-RAD50-NBN*. DNA repair complex[221]. . 227

mSin3A Repressive protein complex containing HDAC1 and HDAC2 histone deacetylases. Involved in multiple cellular processes, including maintenance of pluripotency[222]. We use the same notation as [39], but the prefix "m" is sometimes omitted in the literature. . 26, 102, 120, 219, 224, 226, 228

mSin3A-like Protein complex similar to mSin3A. Annotation from [39]. 228

MSL *Male-specific lethal* histone acetyltransferase complex. Responsible for H4K16ac acetyltransferase activity in vivo via KAT8[37]. . 25, 26, 225

Myc-Max *Myc-max protein complex.* Dimer of MYC and MXD1 (MAD) proteins.Functionally antagonistic with Mad-Max complex, stimulates transcription at E-box (CACGTG) containing promoters.[218]. . 219, 226, 227

N-CoR2 Histone deacetylation complex. 26, 228

NSL *Non-specific lethal* histone acetyltransferase complex. Contains KAT8 histone acetyltransferase, which acetylates H4K5ac, H4K8ac and H4K16ac as part of this complex[37]. . 26, 117, 121, 225

NuA4 Histone acetyltransferase complex with catalytic subunit KAT5 (TIP60). Acetylates histone H4 via KAT5, subunits contain domains that can recognise H3K4me1, H3K4me3, H3K36me3 on histone H3, as well as H4K20me2 and diacetylated histone H4[28, 223]. . 25, 102, 117, 121, 225

NuRD *Nucleosome Remodelling and Deacetylation Complex.* Chromatin remodeler protein, capable of reorganisation of chromatin environment at enhancers, including eviction of chromatin bound proteins. See [224] for a recent review. . 26, 102, 120, 224

NuRF *Nucleosome Remodelling Factor.* Member of *ISWI* family of chromatin remodellers whose members are involved in establishing regular spacing of the nucleosomes[51]. Involved in regulation transcription regulation[51].. 31

ORC *Origin of Replication Recognition Complex.* Protein complex which associates with origins of replication, promotes formation of pre-replication complex required for DNA replication[86]. . 120, 121

PCAF *p300/CBP-associated factor (P/CAF). TAFII* protein-containing complex, that associates with EP300 and CREBBP. Acetylates H3K14ac, H3K18ac, H4K5ac, H4K8ac via KAT2B[39]. . 25, 102, 120, 221, 225

PRC1 *Polycomb Repressive Complex 1.* Repressive protein complex which ubiquitinates lysine 119 of histone H2A (H2AK119). Associates with multiple CBX proteins individually, forming separate complexes[39, 225, 226]. Associates with H3K27me[5]. . 19, 102, 117, 120, 121, 222, 228

PRC2 *Polycomb Repressive Complex 2.* Protein complex capable of recognising and catalysing H3K27 methylation, which allows the spread of facultative heterochromatin[5]. . 19, 28, 43, 102, 120, 131, 224, 229

P-TEFb *Positive transcription elongation factor B.* sTranscriptional regulator complex. Regulates transcription elongation by the phosphorylation of the C-Terminal Domain (CTD) of RNA polymerase II.[227].. 19, 24, 25, 110, 111, 222

RMI/BLM DNA repair associated complex, which contains BLM helicase, mutations of which are known to cause Bloom syndrome[228] . 109, 117, 121, 122, 188, 222, 224, 228

RNA Polymerase II *RNA Polymerase II*. Protein complex responsible for DNA transcription to RNA.. 19, 24, 25, 28, 110, 111, 113, 219, 224, 227, 263

RPA *Replication Protein A* complex. Binds non-specifically to single-stranded DNA, involved in DNA replication, repair and recombination[154]. . 109, 122, 188, 231

SAGA *TAFII* and KAT2A-containing acetyltransferase complex. Coordinates acetylation of H3 as well as deubiquitination of histones H2A and H2B[39]. . 25, 102, 120, 221, 225

SLF1/2-RAD18 Replication associated protein complex, involved in maintaining genome stability upon DNA lesions[229] . 121

Spliceosome Large protein complex responsible for DNA splicing. 121

SRCAP Chromatin remodeler complex. Responsible for deposition of H2A.Z[51]. . 30, 31, 102, 117, 121, 122

Telomerase holoenzyme Telomere maintenance complex[86]. 112

TFIID *The general transcription factor IID*. Protein complex consisting of TBP and a number of *TAFII* proteins necessary for Pre-Initiation Complex (PIC) assembly at promoter regions, which is required for initiation of transcription[230].. 19, 26, 102, 117, 120, 122, 123, 229, 230

TFTC-HAT *TATA-binding protein TBP-free TAFII-containing complex (TFTC) with histone acetyltransferase GCN5 (HAT)*. TBP-free *TAFII* protein complex, containing histone acetyltransferase KAT2A (GCN5). Acetylates histones and may initiate transcription[39, 231]. Related to SAGA and PCAF. . 25, 225

USF1-USF2 *USF1-USF2 upstream stimulatory factor complex*. Dimer between USF1 and USF2[86, EBI-9518693]. . 231

YAP1-TEAD1 *YAP-TEAD complex*. Dimer between YAP1 and *TEAD* transcription factors. Annotation from [86, EBI-11173119] suggests only TEAD1 association, but the complex can form with other *TEAD* factors[232]. . 230, 231

List of proteins

This section provides a reference to the list of proteins referenced by name in this thesis. The descriptions in this section are based on a manual summary of the descriptions provided in UniProt[85] as well as relevant literature, citing the sources where appropriate

ASH2L *ASH2 like histone lysine methyltransferase complex subunit.* Core subunit of histone methyltransferase complexes, specifically MLL1. Stabilises the complex assembly and increases catalytic activity[233]. . 242, 255

ATAD5 *ATPase family AAA domain-containing protein 5.* ATP-binding protein, involved in DNA damage response[85]. Interacts with BRD4[158]. . 111

ATRX *Transcriptional regulator ATRX.*ATP-dependent chromatin remodeler. Preferentially interacts with H3K9me3, but this interaction is inhibited by H3K4me3[33] . 24

BARD1 *BRCA1 associated RING domain 1.* See BRCA1. 222

BLM *Bloom syndrome RecQ like helicase.* DNA helicase part of RMI/BLM complex . 109, 220

BRCA1 *Breast cancer type 1 susceptibility protein.* Tumour suppressor, involved in DNA repair processes as well as cell cycle progression. Together with BARD1 forms an ubiquitin ligase[234]. Member of BRCA1-A and BRCC complexes. . 108, 164, 218, 222, 226, 242, 255

BRD3 *Bromodomain-containing protein 3.* A member of Bromodomain and Extra-Terminal (BET) family of proteins, members of which are characterised by the presence of two bromodomains and an extra-terminal domain. Binds acetylated lysines[29] . 110, 111

BRD4 *Bromodomain-containing protein 4.* A member of BET family of proteins, members of which are characterised by the presence of two bromodomains and an extra-terminal domain. Binds acetylated lysines and regulates transcription elongation by recruiting P-TEFb complex[227]. 19, 22, 24, 25, 110, 111, 147, 161, 166, 167, 170, 172–176, 178, 188, 189, 222, 230, 242, 249, 257

CALM[1-3] *Calmodulin-1*; *Calmodulin-2*; *Calmodulin-3.* A collective entry of three proteins with gene names *CALM1*, *CALM2* and *CALM3* used to illustrate protein labelling algorithm in data preprocessing. . 60

CBX5 *Chromobox protein homolog 5* also known as *HP1 alpha.* Associates with heterochromatin, recognises H3K9me2/3 via the chromodomain, interacts with other chromosomal proteins via chromo-shadow domain[168]. . 18, 24, 39, 51, 55, 95, 129, 161, 162, 165, 170, 172–174, 183, 227, 242, 255

CBX8 *Chromobox protein homolog 8.* Ortholog of Drosophila's polycomb, member of PRC1 complex. Suspected to be functionally distinct from other orthologs in the same family[225]. . 242, 255

CCNT1 *Cyclin-T1.* In combination with CDK9 forms P-TEFb complex, which is responsible in transcription elongation regulation. . 110

CDK9 *Cyclin-dependent kinase 9.* Catalytic subunit of P-TEFb complex, phosphorylates CTD of RNA polymerase II allowing transcription elongation to occur[227]. . 222

CHD1 *Chromodomain-helicase-DNA-binding protein 1.*ATP-dependent chromatin remodeler protein, capable of unravelling the DNA from the core histone octamer[235]. Recruited to promoters of active genes via interaction H3K4me3, recruitment is beneficial for transcription elongation[49]. 125, 126, 161, 162, 170, 242, 249, 255, 257

CHD2 *Chromodomain-helicase-DNA-binding protein 2.* ATP-dependent chromatin remodeler protein . 125, 126

CHD7 *Chromodomain-helicase-DNA-binding protein 7.*ATP-dependent chromatin remodeler protein. Required for activation of superenhancers necessary for neuroepithelial and central nervous system development[236]. . 125, 161, 166, 170, 242, 255

CREB1 *Cyclic AMP-responsive element-binding protein 1.*Member of CREB-related transcription factor family, with consensus binding sequence of *ndRTGACGYhW*[136]. Key regulator in mammalian lung development[237]. . 242, 255

CREBBP *CREB-binding protein.* Histone acetyltransferase. Together with EP300 acetylates H3K27ac, but also has other targets[3] . 26, 220

CTCF *Transcriptional repressor CTCF.*Also known as *CCCTC-binding factor*, after the DNA sequence that it recognises, this zinc-finger protein is one of the key regulators of the three-dimensional conformation of chromatin and boundaries of topologically-associated domains[238]. . 17, 161, 164, 170, 242, 255, 257

DNMT1 *DNA (cytosine-5)-methyltransferase 1.* Maintenance DNA methyltransferase, recruited to hemimethylated DNA via the interaction with UHRF1, required for successful re-establishment of DNA methylation post replication[21, 22]. 20, 230

DNMT3A *DNA (cytosine-5)-methyltransferase 3A.* DNA methyltransferase, recruited to regions with unmethylated lysine 4 on histone H3 (H3K4un) via the ADD domain to establish *de novo* DNA methylation. Also contains a PWWP domain, whose phosphorylation allows its to be targeted to heterochromatin[21] . 20, 111

DNMT3B *DNA (cytosine-5)-methyltransferase 3B.* DNA methyltransferase, recruited to regions with unmethylated lysine 4 on histone H3 (H3K4un) via the ADD domain to establish *de novo* DNA methylation[21]. Recruited to H3K36me3-marked gene bodies via its PWWP domain, where it establishes methylation which helps prevent cryptic transcription[23]. . 20, 111

DOT1L *Histone-lysine N-methyltransferase, H3 lysine-79 specific.*Histone methyltransferase, specific for lysine 79 on histone H3[42] . 28

EHMT1 *Histone-lysine N-methyltransferase EHMT1*, also known as *G9a-like protein 1 (GLP).* Enzyme catalysing mono- and di- methylation of histone H3 lysine 9 (H3K9me1/2) at euchromatic regions[13, 42] . 28

EHMT2 *Histone-lysine N-methyltransferase EHMT2*, also known as *G9a*. Enzyme catalysing mono- and di-methylation of histone H3 lysine 9 (H3K9me1/2) at euchromatic regions[13, 42] . 28

EP300 *Histone acetyltransferase p300*. One of the key histone acetyltransferases, acetylates numerous residues on all core histones, including H3K27ac[3, 39], as well as other proteins. Can also catalyse crotonylation[3] . 26, 220, 223

EPOP *Elongin BC and Polycomb repressive complex 2-associated protein*. Recently identified to facilitate the interaction between elongin and PRC2 complexes[88]. . 58

EZH1 *Histone-lysine N-methyltransferase EZH1*.Catalytic subunit of PRC2. Trimethylates lysine 27 on histone H3 (H3K27me3)[42], the catalytic activity is weaker as compared to that of EZH2[171] . 28, 131, 132, 138

EZH2 *Histone-lysine N-methyltransferase EZH2*.Catalytic subunit of PRC2. Trimethylates lysine 27 on histone H3 (H3K27me3)[42] . 28, 131, 224, 242, 255

FAAP24 *Fanconi anemia core complex-associated protein 24*. Fanconi Anemia associated protein, recruited to stalled replication forks that are associated with RMI/BLM[239] . 109

FANCM *Fanconi anemia group M protein.*. ATPase of the Fanconi Anemia complex . 218

FOS *Fos proto-oncogene, AP-1 transcription factor subunit*.Transcription factor with consensus binding sequence vTGAGTCAb[136]. Dimerises with JUN to form AP-1 complex. . 123, 217, 224

H2AFV/H2AFZ *Histone H2A.V;Histone H2A.Z*.Variant of the canonical histone H2A. H2A.Z is known to reduce the barrier to RNA Polymerase II provided by the nucleosome next to the Transcription Start Site (TSS)[15, 16] . 66

HDAC1 *Histone deacetylase 1*.Class I histone deacetylase. Shows low deacetylase activity in isolation, instead assembles into large deacetylation complexes such as mSin3A, NuRD and LSD-CoREST together with HDAC2[240]. . 26, 219, 224

HDAC11 *Histone deacetylase 11*. The only member of class IV histone deacetylases[35]. . 26

HDAC2 *Histone deacetylase 2*.Class I histone deacetylase. Shows low deacetylase activity in isolation, instead assembles into large deacetylation complexes such as mSin3A, NuRD and LSD-CoREST together with HDAC1[240]. . 26, 161, 170, 172, 219, 224, 242, 256

HIST1H1[D,E] *Histone H1.3*, *Histone H1.4*. A variant of linker histone H1. . 122, 127

HIST1H1C *Histone H1.2*. A variant of linker histone H1. . 122, 127

JUN *Transcription factor AP-1*. Transcription factor with consensus binding sequence dvTGAGTCAYh[136]. Dimerises with FOS to form AP-1 complex. . 123, 217, 224

KAT2A *Histone acetyltransferase KAT2A* (also known as GCN5). Histone H3 acetyltransferase[13, 209], member of ATAC, SAGA, TFTC-HAT complexes. . 26, 221

KAT2B *Histone acetyltransferase KAT2B* (also known as PCAF). See PCAF complex. 26, 220

KAT5 *Histone acetyltransferase KAT5* (TIP60). Subunit of NuA4 and related complexes. Preferably acetylates H2AK5, H3K14, H4K5, H4K8, H4K12 and H4K16[39]. 220

KAT6A *Histone acetyltransferase KAT6A* (MOZ). Histone lysine acetyltransferase. Member of MOZ/MORF complex . 26, 219

KAT6B *Histone acetyltransferase KAT6B* (MORF). Histone lysine acetyltransferase. Member of MOZ/MORF complex . 219

KAT7 *lysine acetyltransferase 7.* (HBO1). Histone lysine acetyltransferase. Member of HBO1 complex. Appears in SNAP data twice: as *KAT7 (1)* (UniProt: O95251) and as *KAT7 (2)* (UniProt: E7EUP3) . 219

KAT8 *Histone acetyltransferase KAT8*, also known as *MOF*. Histone H4K16ac specific lysine acetyltransferase when in MSL complex, less specific as part of NSL complex where it also acetylates H4K5ac and H4K8ac[37]. . 24–26, 218–220

KDM1A *Lysine demethylase 1A.KDM1A*, also known as *LSD1* is a histone lysine-specific demethylase. Acts as transcriptional repressor by demethylating mono- or di-methylated lysine 4 on histone H3 as part of LSD-CoREST complex[241, 242]. In other contexts can serve as transcriptional activator by demethylating lysine 9 on histone H3[242]. . 28, 161, 166, 170, 172, 219, 225, 228, 242, 256

KDM2A *Lysine demethylase 2A.KDM2A* (also *JHDM1A*) is a histone lysine-specific demethylase that demethylates H3K36me2 and, to a lesser extent, H3K36me1 on histone H3[242, 243]. Yeast homologue has been shown to interact with H3K4me3[242]. . 51, 55, 225

KDM2B *Lysine demethylase 2B.KDM2B* (also *JHDM1B*) is a histone lysine-specific demethylase paralogous to KDM2A. Believed to demethylate H3K36me2 (similarly to KDM1A), however conflicting evidence exists suggesting targeting to H3K4me3[244]. Member of BCOR complex. Recently discovered to demethylate H3K79me[41]. 27, 217

KDM5A *Lysine demethylase 5A.KDM5A* (also *JARID1A*) is a histone H3 lysine 4 demethylase, capable of removing H3K4me3, H3K4me2, but not H3K4me1.[242]. . 164, 218, 242, 256

KDM6A *Lysine demethylase 6A* (also known as *UTX*). Histone H3 lysine 27 demethylase, member of COMPASS-like MLL3,4 complex, providing the link between H3K27 and H3K4 methylations[42]. . 29, 218

KMT2A *Histone-lysine N-methyltransferase 2A.* Also known as *MLL1.* Catalytic subunit of COMPASS-like MLL1,2, catalyses H3K4 methylation[5]. . 27, 110, 218

KMT2B *Histone-lysine N-methyltransferase 2B.* Also known as *MLL2.* Catalytic subunit of COMPASS-like MLL1,2, catalyses H3K4 methylation[5]. . 27, 110, 218

KMT2C *Histone-lysine N-methyltransferase 2C.* Also known as *MLL3.* Catalytic subunit of COMPASS-like MLL3,4, catalyses H3K4 methylation[5]. . 27, 218

KMT2D *Histone-lysine N-methyltransferase 2D.* Also known as *MLL4.* Catalytic subunit of COMPASS-like MLL3,4, catalyses H3K4 methylation[5]. . 27, 218

KPNA2 *Importin subunit alpha-1.*Part of nuclear import machinery. Known to import NBN and BRCA1 DNA repair factors into the nucleus[142, 146]. Overexpressed in aggressive cancers[142]. Predicted to interact with SWI/SNF in section 2.3.4. . 108, 109

KPNA6 *Importin subunit alpha-7.* Part of the nuclear import machinery. Has been shown to import SMARCC1 into nucleus[143]. . 108

MAX *MYC associated factor X.*A Basic Helix-Loop-Helix leucine Zipper (bHLH-ZIP) transcription factor with the consensus DNA binding sequence *vnSCACGYGK*[136]. Dimerises with either MYC (Myc-Max complex) or with MXD1 (Mad-Max complex), acts as transcriptional activator when in complex with MYC, and as repressor when in complex with MXD1[218]. Appears in SNAP dataset twice, once as *MAX (1)* (UniProt: P61244), once as *MAX (2)* (UniProt: Q6V3B1). . 219, 226, 227, 242, 256

MAZ *MYC associated zinc finger protein.* Appears twice in SNAP dataset, once as as MAZ (1) (UniProt: I3L0M3/I3L4Y2), once as MAZ (2) (UniProt: P56270/I3L2Z5/A0A087WWR2) . 242, 258

MBD4 *Methyl-CpG-binding domain protein 4.* 129

MECP2 *Methyl-CpG-binding protein 2.* Epigenetic reader, recruited to methylated CpG residues on the DNA[170]. Famously involved in Rett syndrome[245]. Interacts with mSin3A[246]. Appears in data from SNAP twice: once as *MECP2 (1)* (UniProt: P51608), once as *MECP2 (2)* (UniProt: B5MCB4) . 129

MRM3 *rRNA methyltransferase 3, mitochondrial.* Responsible for modification of nucleotides of the 16S rRNA A-loop, important protein for mitochondrial translation[247], one of the three proteins removed from factor analysis (see section 2.2.4) . 87

MXD1 *MAX dimerization protein 1* (also known as *MAD*). Competes with MYC for dimerisation with MAX. The resulting Mad-Max complex represses transcription at E-Box (CACGTG) containing promoters[218]. . 219, 220, 226

MXI1 *Max-interacting protein 1* Transcription factor of bHLH-Zip family, with the consensus DNA binding sequence of vCACRYGKvnSbbbv[136]. Interacts with MAX[248]. Appears twice in SNAP dataset: as MXI (1) (UniProt: F6U3F6) and MXI (2) (UniProt: P50539/B1ANN8) . 165, 242, 256, 258

MYC *Myc proto-oncogene protein.*Transcription factor involved in multiple types of cancer. Forms a complex with MAX (Myc-Max), which targets it to E-Box (CACGTG) containing promoters. Acts as a transcription activator[218]. . 24, 25, 220, 226, 242, 256

NBN *Nibrin.* DNA repair factor, part of MRN complex. . 108, 226

NSD1 *Histone-lysine N-methyltransferase, H3 lysine-36 and H4 lysine-20 specific.*One of the four major methyltransferases for H3K36 in mammals[42]. Catalyses H3K36me1 and H3K36me2 but only on nucleosomal substrates (i.e. not histones)[42]. The UniProt name suggests H4K20me activity, however the recent review[42], does not list it. . 28

NSD2 *Histone-lysine N-methyltransferase NSD2.*One of the four major methyltransferases for H3K36 in mammals[42]. Catalyses H3K36me1 and H3K36me2 but only on nucleosomal substrates (i.e. not histones)[42]. . 28, 111

NSD3 *Histone-lysine N-methyltransferase NSD3.*One of the four major methyltransferases for H3K36 in mammals[42]. Catalyses H3K36me1 and H3K36me2 but only on nucleosomal substrates (i.e. not histones)[42]. Appears twice in SNAP dataset: once as NSD3 (1) (UniProt:Q9BZ95), once as NSD3 (2) (UniProt: H0YE68). . 28

PHF8 *PHD finger protein 8.*Histone lysine demethylase. Demethylates H3K9me1 and H3K9me2 in H3K4me3-dependent manner[196]. Similarly, demethylates H4K20me1 in H3K4me2/3-dependent fashion[249]. Demethylates H3K27me2 independently of H3K4me[249]. Appears twice in SNAP dataset: as *PHF8 (1)* (UniProt: Q9UPP1) and as *PHF8 (2)* (UniProt: B0QZZ4). . 29, 162, 242, 256

PIM1 *Serine/threonine-protein kinase pim-1.* Histone kinase. Phosphorylates serine 10 on histone H3 at *FOSL1* enhancer locus, which ultimately leads to transcription elongation of this gene[31, 32]. . 24, 25

POGZ *Pogo transposable element with ZNF domain.* Binds to CBX5 through its zinc-finger, interaction essential for mitotic progression[250]. One of the three proteins removed from factor analysis (see section 2.2.4) . 87

POLR2A *DNA-directed RNA polymerase II subunit RPB1.*Core subunit of RNA Polymerase II complex. . 161, 170, 256, 258

PWWP2A *PWWP domain-containing protein 2A.* PWWP domain containing protein, recruited to H2A.Z-containing nucleosomes[162] . 111, 227

PWWP2B *PWWP domain-containing protein 2B.* Rather uncharacterised PWWP domain containing protein, similar to PWWP2A . 111

RAD50 *DNA repair protein RAD50.* DNA repair factor, member of MRN protein complex. . 65, 113

RBBP5 *Retinoblastoma-binding protein 5.* Core subunit of multiple *COMPASS-like* complexes. . 242, 256

RCOR1 *REST corepressor 1.*Subunit of a number of REST corepressor complexes, including LSD-CoREST and CtBP. Facilitates KDM1A H3K4 demethylase activity for nucleosome substrates[241]. Appears twice in SNAP dataset: as *RCOR1 (1)* (UniProt: J3KN32) and *RCOR1 (2)* (UniProt: Q9UKL0). . 166, 242, 258

REST *RE1-silencing transcription factor.*Transcriptional repressor, binding repressor element 1 (RE1) sequences. Achieves repression via recruitment of mSin3A, LSD-CoREST, CtBP repressive complexes[251]. . 164, 228, 242, 256

RIF1 *Telomere-associated protein RIF1.* DNA-binding protein which interacts with RMI/BLM complex[148] . 109

RMI1 *RecQ-mediated genome instability protein 1.* Component of RMI/BLM protein complex. . 109

RMI2 *RecQ-mediated genome instability protein 2.* Component of RMI/BLM protein complex. . 109

RNF2 *E3 ubiquitin-protein ligase RING2.* Member of PRC1 protein complex, ubiquitinates H2AK119[5] . 242, 256

RPF2 *Ribosome production factor 2 homolog.* Involved in recruiting ribosomal proteins to pre-ribosomes in yeast[252], one of the three proteins removed from factor analysis (see section 2.2.4) . 87

SAP30 *Histone deacetylase complex subunit SAP30.*Subunit of N-CoR2 and mSin3A complexes. Capable of interacting with histones H2A and H2B, as well as binding and bending DNA[253]. . 162, 242, 256

SETD1A *Histone-lysine N-methyltransferase SETD1A.*Histone lysine methyltransferase specific for lysine 4 on histone H3[13], catalytic subunit of COMPASS complex. . 27, 218

SETD1B *Histone-lysine N-methyltransferase SETD1B.*Histone lysine methyltransferase specific for lysine 4 on histone H3[13], catalytic subunit of COMPASS complex. . 27, 218

SETD2 *Histone-lysine N-methyltransferase SETD2.*One of the four major methyltransferases for H3K36 in mammals[42]. Only methyltransferase that can catalyse H3K36me3[42]. . 28

SETDB1 *Histone-lysine N-methyltransferase SETDB1.*Histone H3 lysine 9 methyltransferase, catalysing the deposition of H3K9me3[13] . 28, 219

SIN3A *Paired amphipathic helix protein Sin3a.* Transcriptional repressor, core component of mSin3A, mSin3A-like, N-CoR2 complexes. 162, 242, 256

SIN3B *Paired amphipathic helix protein Sin3b.*Repressor protein, member of protein complexes containing histone deacetylase (HDAC) activity[254] . 218

SIRT6 *NAD-dependent protein deacetylase sirtuin-6.* Class III NAD+ dependent histone deacetylase. Deacetylates lysine H3K9 and H3K56[40] . 26

SIRT7 *NAD-dependent protein deacetylase sirtuin-7.*Class III NAD+ dependent histone deacetylase. Deacetylates lysine H3K18[40] . 26

SMAD[2,3,9] *Mothers against decapentaplegic homolog 2*; *Mothers against decapentaplegic homolog 3*; *Mothers against decapentaplegic homolog 9.* A collective entry of three proteins with gene names *SMAD2*, *SMAD3* and *SMAD9* used to illustrate protein labelling algorithm in data preprocessing. . 60

SMARCA4 *Transcription activator BRG1.*One of the subunits of the SWI/SNF protein complex. Appears in SNAP dataset twice, once as *SMARCA4 (1)* (UniProt:P51532) and another time as *SMARCA4 (2)* (UniProt:A0A0A0MT49) . 59, 60

SMARCC1 *SWI/SNF complex subunit SMARCC1.* One of the subunits of the SWI/SNF protein complex. . 108, 226

SMC1A *Structural maintenance of chromosomes protein 1A.*Hinge subunit of the Cohesin complex[255] . 111, 112, 121

SPIN1 *Spindlin-1.* Chromatin reader. Recognises H3K4me3, regulates expression of rRNA[169] . 129, 229

SPINDOC *Spindlin interactor and repressor of chromatin-binding protein.* Protein which docks to the SPIN1 and disrupts its histone binding ability[89]. . 58

SSRP1 *FACT complex subunit SSRP1.* Dimerises with SUPT16H to form the FACT complex. 112, 229

SUPT16H *FACT complex subunit SPT16.* Dimerises with SSRP1 to form the FACT complex . 112, 229

SUV39H1 *Histone-lysine N-methyltransferase SUV39H1.* Also known as KMT1A. First discovered histone lysine methyltransferase[256]. Trimethylates lysine 9 on histone H3 (H3K9me3), facilitates the spread of heterochromatin[20]. . 19, 28, 229

SUV39H2 *Histone-lysine N-methyltransferase SUV39H2.* Also known as KMT1B. Histone H3 lysine 9 methyltransferase, functionally similar to SUV39H1[20]. . 28

SUZ12 *Polycomb protein SUZ12.* DNA/RNA binding subunit of the PRC2 complex[5] . 65, 129, 131, 242, 256

TAF1 *TATA-box binding protein associated factor 1.*Largest and subunit of the TFIID complex, responsible for promoter binding[257]. Contains acetyltransferase activity, acetylates lysines on histones H3 and H4[39] . 26, 162, 229, 242, 256

TAF7 *TATA-box binding protein associated factor 7.*Subunit of the TFIID complex, stabilises TAF1[257]. . 242, 256

TBP *TATA-box-binding protein.*Core component of TFIID complex. Binds to TATA sequence at promoters, crucial for the complex assembly[230]. In SNAP data indistinguishable from *TATA box-binding protein-like protein 2 (TBPL2).* . 123, 221, 242, 256

TBRG1 *Transforming growth factor beta regulator 1*, also known as *NIAM: nuclear interactor of ARF and MDM2.* Fairly uncharacterised protein, involved in proliferation and chromosomal stability[139]. Predicted to interact with INO80 in our analyses (see section 2.3.4). . 108, 190, 192

TEAD1 *Transcriptional enhancer factor TEF-1.* Regulates YAP1 transcription via dimerisation forming YAP1-TEAD1 complex.[232] In SNAP data indistinguishable from TEAD2 (UniProt: Q15562) and TEAD4 (UniProt: A0A0A0MRF3) and thus has a label *TEAD[1,2,4].* TEAD4 also appears as another entry with TEAD3. . 221, 230

TEAD2 See TEAD1. 230

TEAD3 Functionally similar to TEAD1. In SNAP data appears either together with TEAD4 (label: *TEAD3/TEAD4*, UniProt: Q99594), or separate (label: *TEAD3*, UniProt: H7C4A7). . 230

TEAD4 *Transcriptional enhancer factor TEF-3.* Transcription factor with consensus sequence RMATTCCWdvvhb[136]. Functionally similar to TEAD1. In SNAP data appears either together with TEAD3 (label: *TEAD3/TEAD4*, UniProt: Q15561,H0YGS2,Q53GI4,H0YG62), or together with TEAD1 (label: *TEAD[1,2,4]*, UniProt: A0A0A0MRF3). . 230, 242, 256

TFAM *Transcription factor A, mitochondrial.* Mitochondrial transcription factor that is involved in packaging the mitochondrial DNA[166] . 122, 123, 127

TONSL *Tonsoku-like protein.*Protein involved in homologous recombination based DNA repair and replication fork stability. Recently shown to recognise unmodified H4K20[34] . 24, 25

TOP2A *DNA topoisomerase 2-alpha.* Catalyses Double-Strand Breaks (DSBs) and subsequent rejoining of the DNA to relieve torsion-related problems during multiple cellular processes[164] . 112, 121

TOP2B *DNA topoisomerase 2-beta.* Catalyses DSBs and subsequent rejoining of the DNA to relieve torsion-related problems during multiple cellular processes[164]. Appears in SNAP dataset twice, once as *SMARCC2 (1)* (UniProt:E9PCY5) and another time as *SMARCC2 (2)* (UniProt:Q02880). . 111, 112

TP53 *Cellular tumor antigen p53.* Major tumour suppressor, required for G1 arrest upon DNA damage[258] . 108, 218

TWIST1 *Twist-related protein 1.* Transcription regulation, shown to contain histone H4 mimic domain, which can facilitate recruitment of BRD4[199]. . 167, 188

UHRF1 *E3 ubiquitin-protein ligase UHRF1.* Recruited to hemimethylated DNA during replication, recruits DNMT1 to facilitate maintenance of DNA methylation[22] . 20, 39, 51, 55, 129, 223

USF1 *Upstream stimulatory factor 1.* bHLH-ZIP transcription factor with consensus binding sequence vGTCACGTGvbb[136]. Dimerises with USF2, see USF1-USF2. . 221, 231, 242, 257

USF2 *Upstream stimulatory factor 2.* bHLH-ZIP transcription factor with consensus binding sequence vvGTCACGTGvbvvbbvbb[136]. Dimerises with USF1, see USF1-USF2. . 221, 231

WDR76 *WD repeat-containing protein 76.* 5-hydroxymethylcytosine (5-hmC) reader protein[150], associates with RPA complex[151] . 109

YAP1 *Transcriptional coactivator YAP1.* Transcription regulator which plays a role in organ development. Dimerises with *TEAD* proteins[232]. See YAP1-TEAD1. . 221, 230

YY1 *Transcriptional repressor protein YY1.*Transcription factor with consensus binding sequence MAARATGGCKSM[136]. Component of INO80 complex [259], a versatile protein recently shown to interact with Cohesin to mediate enhancer-promoter looping[260]. . 161, 170, 188, 242, 257

ZFX/ZFY *Zinc finger X-chromosomal protein*, *Zinc finger Y-chromosomal protein.* ZFX acts as a transcriptional activator, binds downstream of promoters, slightly upstream of H3K4me3 peak of the regulated genes[156]. . 110

List of software & databases

This section provides a reference to the list of software and databases used in this thesis

ARACNE *An Algorithm for the Reconstruction of Gene Regulatory Networks (ARACNE).* Implemented in MINET. See MINET for details. 71, 74

BioGRID *BioGRID interaction database.* Curated and up-to-date database of protein-protein interactions[87]. Used to profile our network inference methods. `https://thebiogrid.org/`. 43, 45, 72, 74, 82, 83, 101, 104–106, 108, 186

Bokeh *Bokeh* (version: 0.12.16). Interactive visualisation library, used extensively in the interactive visualisation of SNAP data as described in section 2.2.1. `https://bokeh.pydata.org/en/latest/`. 54, 57, 189

Bowtie 2 *Bowtie 2: Fast and sensitive read alignment* (Version: 2.3.4.1)[178]. Software for alignment of reads to reference genomes used in the ChIP-seq data processing pipeline in section 3.2.1. `http://bowtie-bio.sourceforge.net/bowtie2/index.shtml`. 142, 143

ChromHMM *ChromHMM: automating chromatin state discovery and characterisation.* Hidden Markov Model (HMM) based method of partition chromatin regions into functional states[175]. 149

CLR *Context-Likelihood of Relatedness (CLR) algorithm.* Implemented in MINET. See MINET for details. 71, 75

clusterProfiler *clusterProfiler: statistical analysis and visualization of functional profiles for genes and gene clusters* (Version: 3.8.1). R package for Gene Ontology (GO) and other ontology enrichment and visualisation[138] . `https://bioconductor.org/packages/release/bioc/html/clusterProfiler.html` . 98

Complex Portal *The complex portal - an encyclopaedia of macromolecular complexes.* Manually curated resource for protein complexes in multiple model organisms. Together with EpiFactors formed the initial seed for protein complexes discussed in this thesis. `https://www.ebi.ac.uk/complexportal/home`. 43, 186, 217, 232

Cytoscape *Cytoscape: Network Data Integration, Analysis, and Visualization in a Box* (version: 3.6.1)[261]. Software for network analysis and visualisation. `http://www.cytoscape.org/`. 83

Docker *Docker.* Platform for containerisation of applications. Used extensively for seamless dependency management throughout the thesis and web services described in section 2.2.1. `https://www.docker.com/`. 53

ENCODE *ENCODE: Encyclopedia of DNA Elements.* First, and recently updated and currently most comprehensible database of chromatin-associated datasets from NGS experiments[176, 262]. `https://www.encodeproject.org/` . 141, 142, 147, 148, 173, 178, 234, 255

Ensembl *Ensembl.* Database providing an integration of multiple genomic datasets, including sequence, transcript information, as well as transcription factor binding data, and other annotations.[263]. `http://www.ensembl.org/`. 59

Entrez *Entrez: NCBI's database for gene information*[264]. One of the first gene-oriented databases developed. The unique integer identifiers for each gene are actively used in biological software. `https://www.ncbi.nlm.nih.gov/gene`. 59

EpiFactors *EpiFactors: a comprehensive database of human epigenetic factors and complexes.* Online resource of manually curated epigenetic protein complexes that together with Complex Portal were the initial seeds of the list of protein complexes in this thesis. `http://epifactors.autosome.ru/`. 43, 45, 136, 186, 217, 219, 232

FIMO *FIMO: Find Individual Motif Occurrences* (version 5.0.1). Tool for scanning DNA sequences for motif occurences[137]. `http://meme-suite.org/doc/fimo.html`. 94

Flask *Flask*. Python-based library for creating dynamic websites. Used in section 2.2.1. `http://flask.pocoo.org/`. 54, 57

Gephi *Gephi: The Open Graph Viz Platform* (version 0.9.2). Network visualisation software. `https://gephi.org/` . 83

GitHub *GitHub*. Repository for source code managed via the *git* platform, that we use for sharing the software code used in this thesis. Author's profile can be accessed at `https://github.com/lukauskas/` . 142, 161

HDF5 *HDF5*. Binary data storage format used extensively for interim data storage, as well as a database in the interactive interface described in section 2.2.1. `https://support.hdfgroup.org/HDF5/`. 57

HOCOMOCO *HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis* (version v11). Database of transcription factor models[136]. `http://hocomoco11.autosome.ru/` . 94, 117, 122

Interpro *InterPro: protein sequence analysis & classification*. Database that provides protein sequence classifications into families, domains as well as other features[265]. `https://www.ebi.ac.uk/interpro/` . 59, 117, 134, 186

luigi *luigi: a Python module that helps you build complex pipelines of batch jobs*. (Version: 2.7.5) Python library for describing pipelines. Used in this project for ChIP-seq data analysis. `https://github.com/spotify/luigi` . 142

MACS *MACS: Model-based Analysis of ChIP-Seq*. (Version: 2.1.1.20160309). A popular peak-caller for ChIP-seq data[182]. `https://github.com/taoliu/MACS` . 33, 143, 144

MaxQuant *MaxQuant: analysis of large-scale mass spectrometric data sets*[93]. Software for processing large scale proteomics datasets. Used in the initial step of SNAP data analysis to estimate protein abundances from raw mass spectrometry datasets. `http://www.biochem.mpg.de/5111795/maxquant` . 63

MINET *MINET: Mutual Information NETworks* (version: 3.38.0). R package implementing multiple network algorithms based on Mutual Information (MI)[100] including CLR, ARACNE and MRNET. `https://www.bioconductor.org/packages/release/bioc/html/minet.html`. 75, 76, 231, 232

Mygene.info *Mygene.info: Gene Annotation as a Service*. (last access date: 2018-04-11). REST web service for gene annotations[90]. Used extensively in this project to fetch and update metadata. `http://mygene.info/` . 59, 60

NGINX *NGINX*. Web server and load balancer used as a load balancer and static content server for the interactive interface described in section 2.2.1. `https://www.nginx.com/`. 53, 57

OGS *Open Grid Scheduler*. Software employed in computing clusters, which is responsible for scheduling the user jobs for a fair usage of computing resources available. `http://gridscheduler.sourceforge.net/`. 142

phantompeakqualtools *phantompeakqualtools: informative enrichment and quality measures for ChIP-seq/DNase-seq/FAIRE-seq/MNase-seq data* (version: GitHub:916b09f1b7). Software tool for robust estimation of fragment length in next-generation sequencing.[183]`https://github.com/kundajelab/phantompeakqualtools` . 144

precrec *precrec: Calculate Accurate Precision-Recall and ROC (Receiver Operator Characteristics) Curves* (version: 0.9.1). R package for accurate and fast calculation and visualisation of Receiver Operating Characteristic (ROC) and Precision-Recall Curve (PRC) curves[108]. `https://cran.r-project.org/web/packages/precrec/index.html`. 73

psych *psych: Procedures for Psychological, Psychometric, and Personality Research* (version: 1.8.4). R package implementing factor analysis procedures[117]. `https://cran.r-project.org/web/packages/psych/index.html`. 86, 89

ReMap *ReMap: An integrative ChIP-seq analysis of regulatory regions*. Integrative analysis of hundreds of transcriptional regulators, providing an atlas of available data as well as quality control metrics[177]. 142

Roadmap *Roadmap: integrative analysis of 111 reference human epigenomes*. Integrative, large scale human epigenomics project generating a dataset for 111 human cell types[52]. Data available at `http://www.roadmapepigenomics.org/`, or via the ENCODE database. 41, 42, 45, 142–144, 170, 171

scikit-learn *scikit-learn: Machine Learning in Python* (version: 0.19.1). Python package for efficient implementations of machine learning algorithms[107]. `http://scikit-learn.org/` . 73, 96, 147, 165

SRA *Sequence Read Archive (SRA)*. Repository for short reads from next-generation sequencing experiments. `https://www.ncbi.nlm.nih.gov/sra`. 142, 148, 255

UniProt *UniProt: the universal protein knowledgebase*. As the title suggests, UniProt database aims to provide an universal set of information about proteins[85]. Out of particular importance are manual, literature consensus-based annotations for each of their entries. `https://www.uniprot.org/` . 43, 59, 117, 186, 221

WGCNA *WGCNA: Weighted Correlation Network Analysis* (version: 1.63). R package implementing the weighted correlation network analysis method described in [266]. `https://cran.r-project.org/web/packages/WGCNA/index.html` . 70, 74

Appendix B

Supplementary information

B.1 Supplementary figures and tables

estimator	AUC ROC	AUC PRC	spAUC PRC @ 0.05	spAUC PRC @ 0.10	spAUC PRC @ 0.15	spAUC PRC @ 0.20	spAUC PRC @ 0.25	spAUC PRC @ 0.30
CLR-MI	0.605228	0.076862	0.658458	0.463663	0.350555	0.281489	0.235450	0.202932
CORR	0.638650	0.073457	0.523011	0.378202	0.295152	0.243577	0.208167	0.182395
RAW-DC	0.618546	0.067514	0.481029	0.353238	0.276761	0.227245	0.193513	0.168958
RAW-MI	0.608449	0.062978	0.520145	0.351410	0.265720	0.214453	0.180571	0.156701
CLR-DC	0.620677	0.055201	0.243112	0.217335	0.187779	0.162446	0.142395	0.126754
CORR-LW	0.630356	0.051911	0.382562	0.233534	0.177995	0.147205	0.127405	0.113261
WGCNA	0.630356	0.051911	0.382562	0.233534	0.177995	0.147205	0.127405	0.113261
MRNET-MI	0.600915	0.048258	0.341106	0.223883	0.172441	0.142334	0.122410	0.108104
MRNET-DC	0.613555	0.045968	0.297383	0.191518	0.147507	0.123277	0.107712	0.096756
ARACNE-MI	0.508866	0.029117	0.235051	0.128918	0.092626	0.074251	0.063131	0.055671

Table B.1: Performance statistics for the network inference algorithms tested in section 2.2.3.

Figure B.1: Complete set of complexes regulated by the hidden factors using the t-test procedure described in Figure 2.14. Shown here is the t-statistic of statistically significant results.

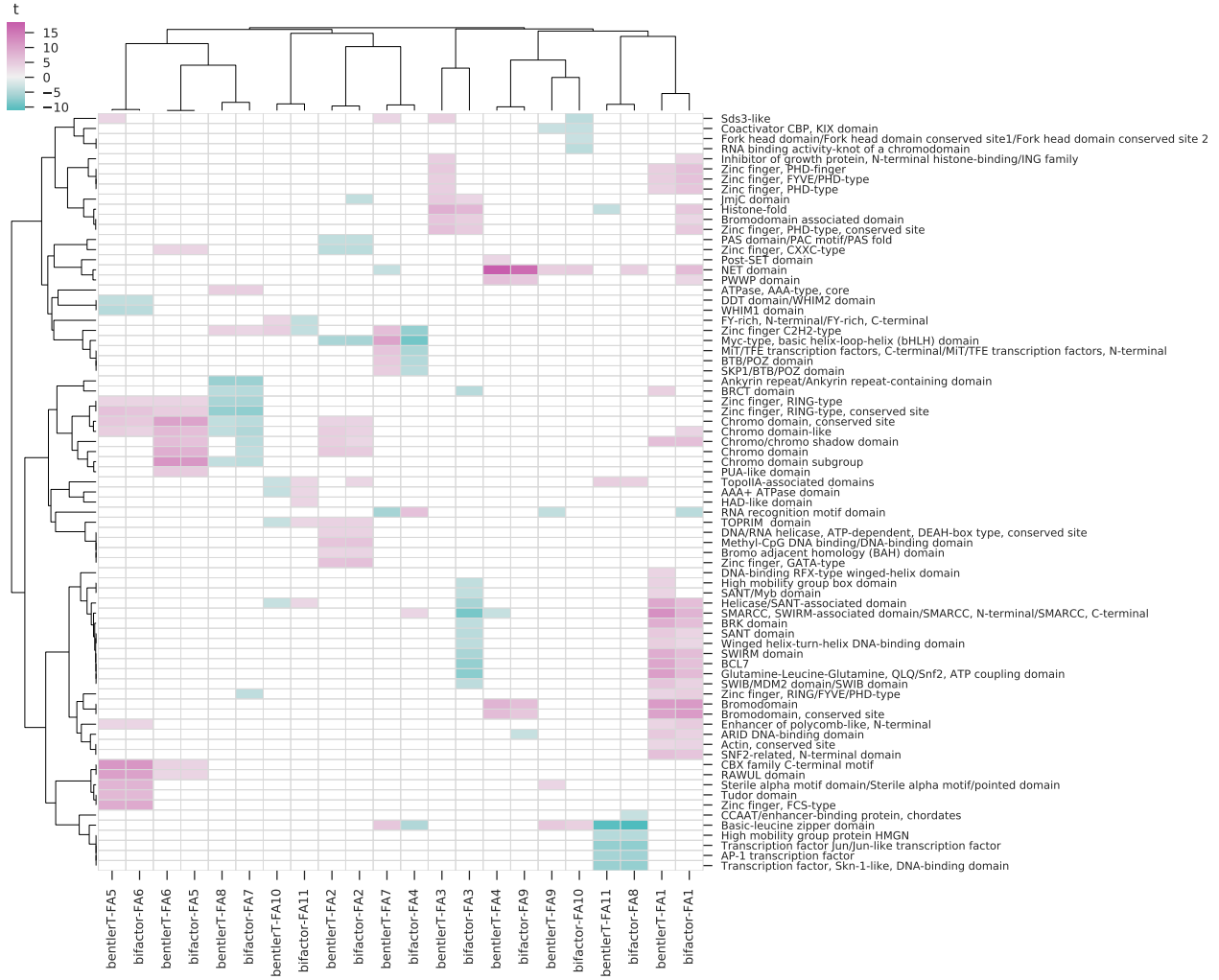

Figure B.2: Complete set of protein domains regulated by the hidden factors as described in Figure 2.14. Shown here is the t-statistic of statistically significant results.

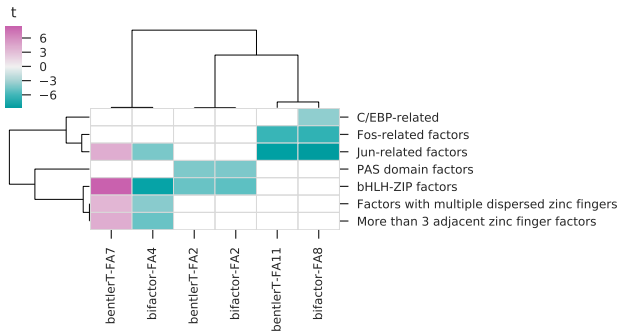

Figure B.3: Complete set of transcription factors regulated by the hidden factors as described in Figure 2.14. Shown here is the t-statistic of statistically significant results.

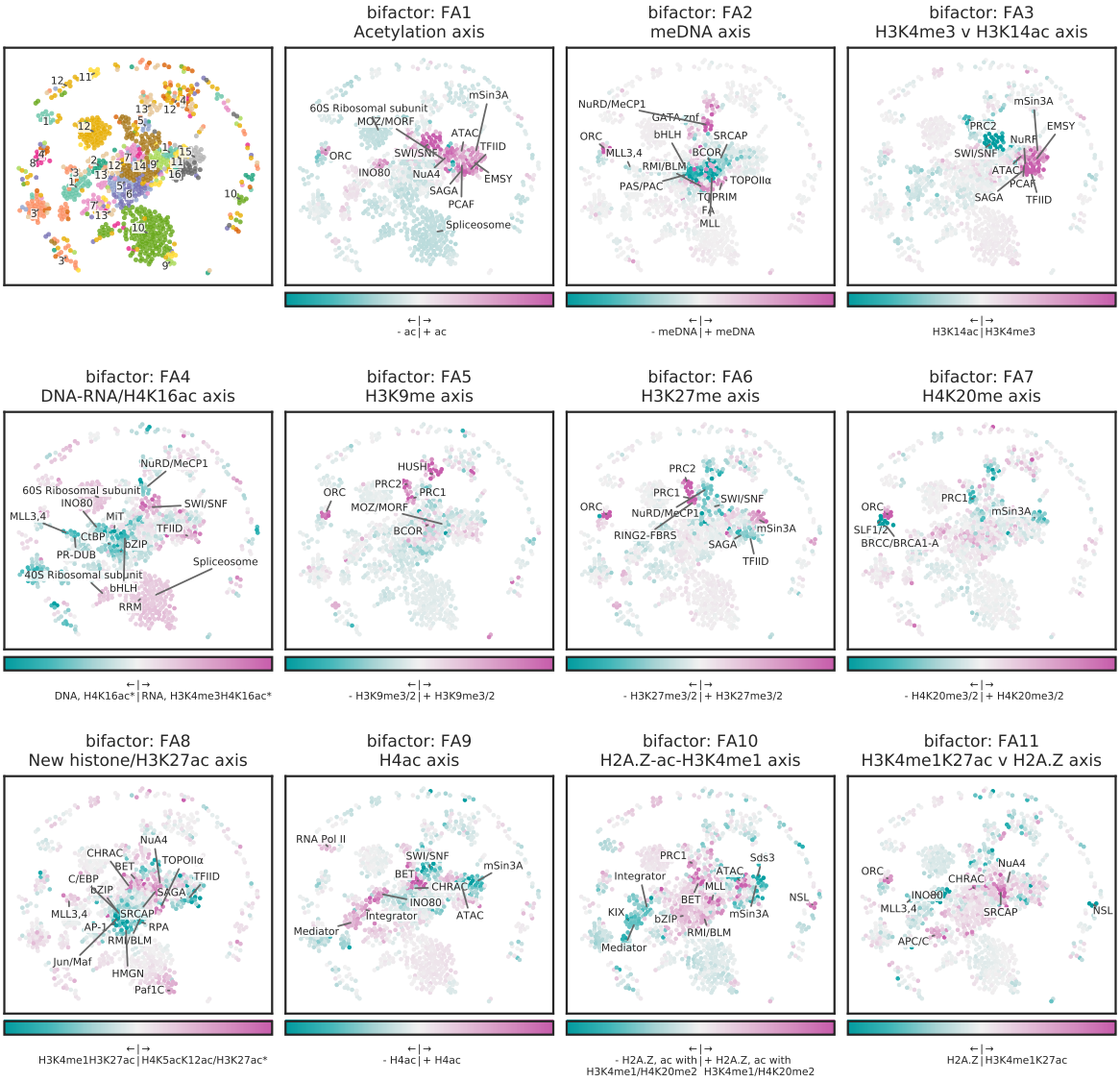

Figure B.4: *bifactor* rotation equivalent to the factor maps of *bentlerT* rotation shown in Figure 2.25. The colour of proteins is determined by the respective factor score. Annotations correspond to the manually curated set of significantly enriched items from Figures B.1, B.2 and B.3. Top left panel corresponds to clusters identified by hierarchical clustering (Figure B.5 B).

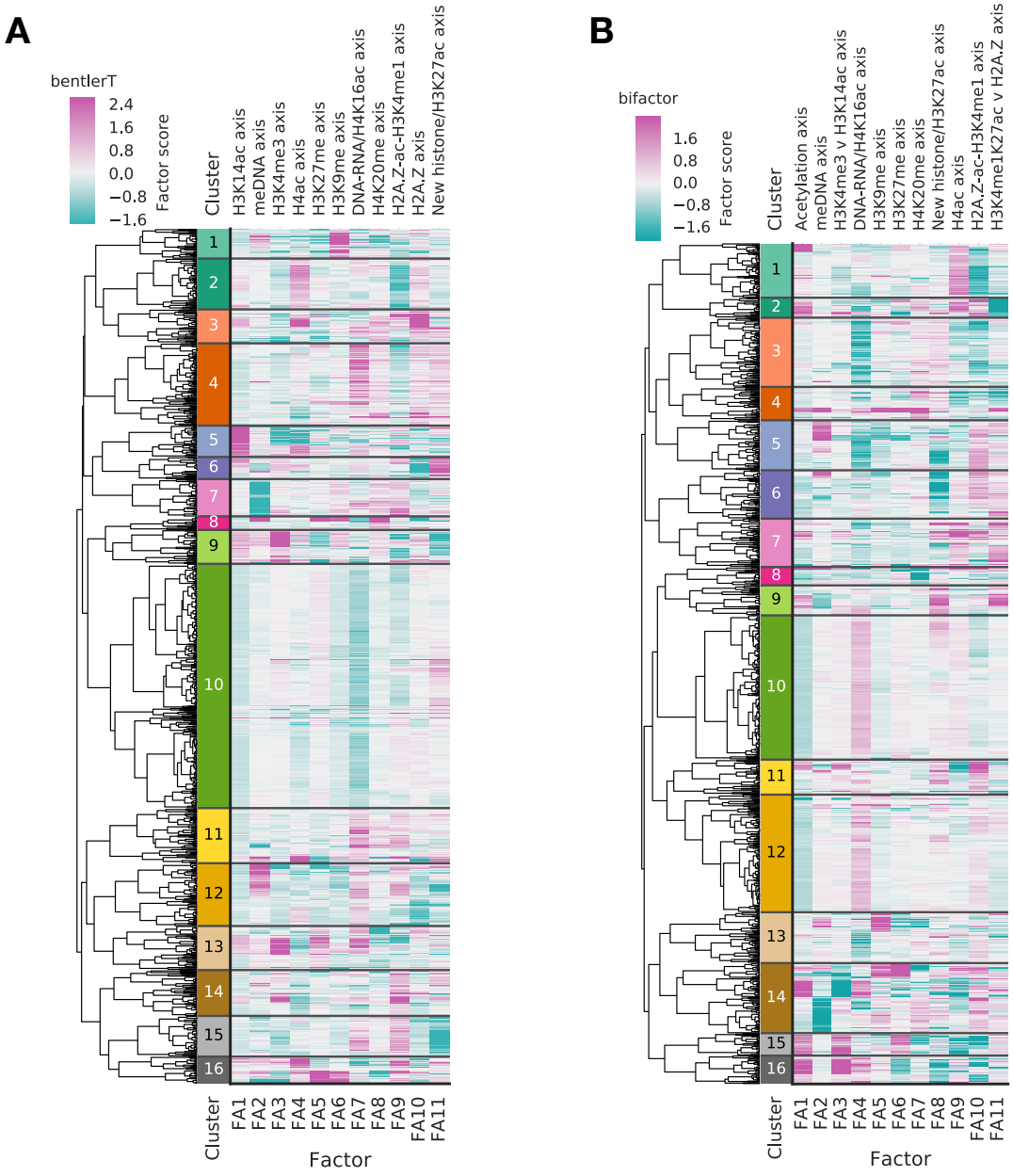

Figure B.5: Factor clusters for the top-left panels of Figure 2.25 (*bentlerT* rotation) and Figure B.4 (*bifactor* rotation).

Figure B.6: DNA sequence used in the production of recombinant di-nucleosomes for SNAP pull-downs. The vector is constructed by concatenation of two core DNA positioning sequences from the original 601 sequence (see [70, Table 3]) with a short 50bp linker. The nucleosomes are expected to span the 601-sequence matching regions, highlighted in red. CpG sites, which are subject to methylation, are underlined.

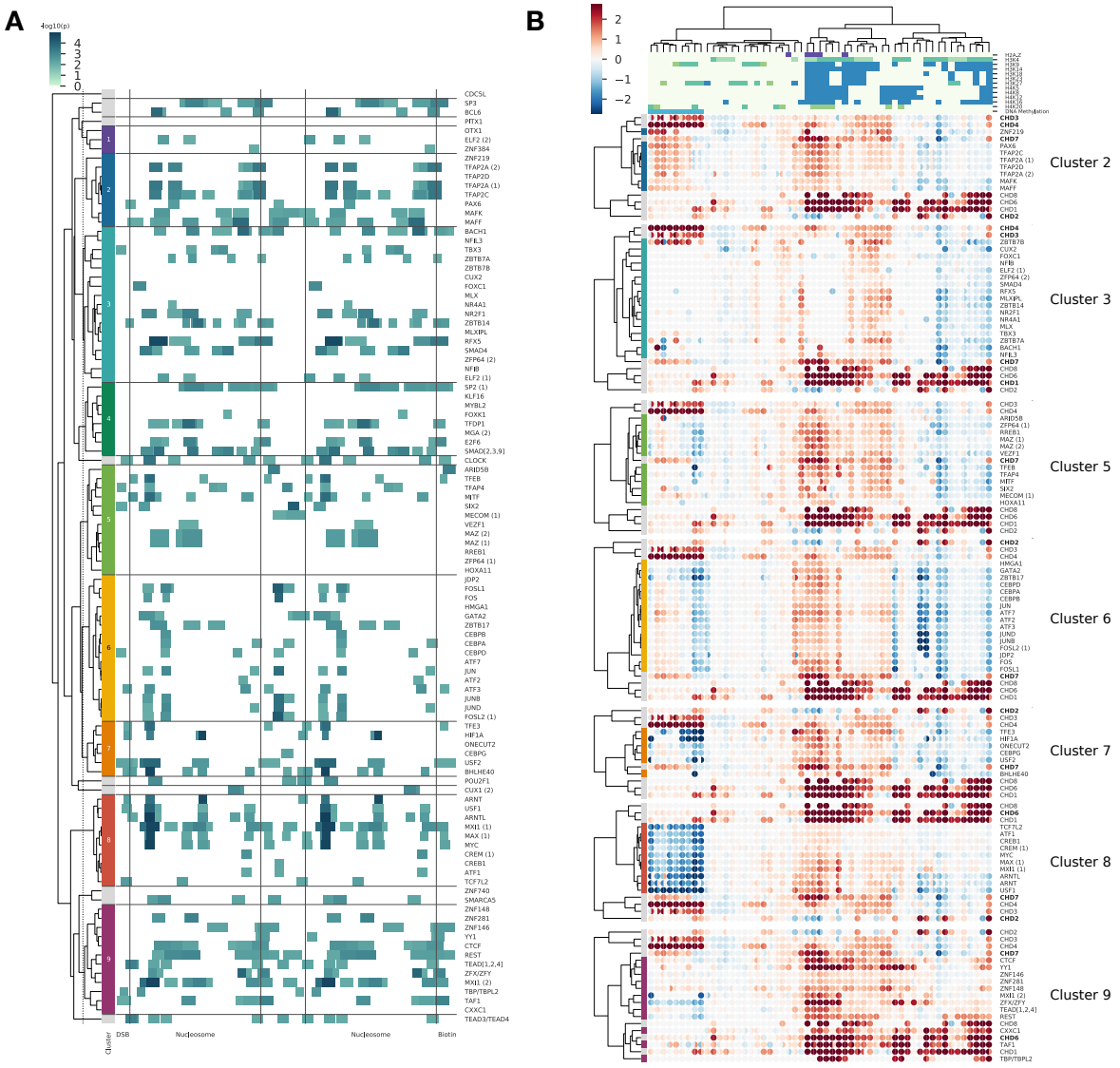

Figure B.7: (A) Distribution of predicted binding sites for all DNA binding transcription factors studied. (B) The heatmap in A, Figure 2.26 augmented with CHD proteins. Clusters 1 and 4 have been skipped as their profiles do not correlate with CHD proteins (see Figure 2.27 E).

target	a	b	n	r	t	p	q
MXI1	H1-hESC	IMR-90	24	0.726662	4.961234	0.000029	0.001128
PHF8	H1-hESC	SNAP	9	0.891595	5.209274	0.000620	0.009653
CBX5	H1-hESC	SNAP	9	0.885675	5.046840	0.000743	0.009653
SIN3A	H1-hESC	SNAP	9	0.853061	4.325298	0.001729	0.016856
SAP30	H1-hESC	SNAP	9	0.813382	3.699318	0.003831	0.021344
CHD1	H1-hESC	SNAP	9	0.826291	3.881382	0.003021	0.021344
		IMR-90	24	0.535657	2.975304	0.003491	0.021344
	IMR-90	SNAP	11	0.687470	2.839954	0.009702	0.042043
TAF1	H1-hESC	SNAP	9	0.754396	3.040656	0.009415	0.042043
RNF2	H1-hESC	SNAP	9	0.691757	2.534467	0.019489	0.076007
CTCF	H1-hESC	IMR-90	24	0.344463	1.721001	0.049643	0.148929
CBX8	H1-hESC	SNAP	9	0.604570	2.008081	0.042302	0.148929
MAX	H1-hESC	SNAP	9	0.589724	1.931964	0.047324	0.148929
SUZ12	H1-hESC	SNAP	9	0.507265	1.557338	0.081674	0.212352
TAF7	H1-hESC	SNAP	9	0.508078	1.560702	0.081281	0.212352
EZH2	H1-hESC	SNAP	9	0.480313	1.448859	0.095324	0.232353
BRD4	H9	IMR-90	25	0.202681	0.992627	0.165610	0.358822
TEAD4	H1-hESC	SNAP	9	0.374651	1.069102	0.160249	0.358822
KDM5A	H1-hESC	SNAP	9	0.338029	0.950277	0.186809	0.383451
RBBP5	H1-hESC	SNAP	9	0.309954	0.862541	0.208479	0.387176
YY1	H1-hESC	SNAP	9	0.317702	0.886490	0.202389	0.387176
CTCF	H1-hESC	SNAP	9	0.287985	0.795645	0.226185	0.400964
CREB1	H1-hESC	SNAP	9	0.231195	0.628719	0.274745	0.465872
TBP	H1-hESC	SNAP	9	0.199597	0.538929	0.303321	0.473181
BRD4	H9	SNAP	11	0.179684	0.547972	0.298519	0.473181
ASH2L	H1-hESC	SNAP	9	0.040818	0.108083	0.458481	0.665309
MXI1	IMR-90	SNAP	11	0.033892	0.101735	0.460599	0.665309
CTCF	IMR-90	SNAP	11	-0.022173	-0.066535	0.525797	0.732360
MXI1	H1-hESC	SNAP	9	-0.083390	-0.221400	0.584449	0.785984
MAZ	IMR-90	SNAP	11	-0.282246	-0.882622	0.799800	0.945218
HDAC2	H1-hESC	SNAP	9	-0.316657	-0.883249	0.796794	0.945218
BRCA1	H1-hESC	SNAP	9	-0.314512	-0.876606	0.795113	0.945218
USF1	H1-hESC	SNAP	9	-0.294145	-0.814255	0.778843	0.945218
REST	H1-hESC	SNAP	9	-0.421453	-1.229598	0.870716	0.953560
CHD7	H1-hESC	SNAP	9	-0.408402	-1.183751	0.862426	0.953560
RCOR1	IMR-90	SNAP	11	-0.387064	-1.259356	0.880209	0.953560
BRD4	IMR-90	SNAP	11	-0.487381	-1.674484	0.935821	0.967132
MYC	H1-hESC	SNAP	9	-0.561933	-1.797346	0.942334	0.967132
KDM1A	H1-hESC	SNAP	9	-0.679814	-2.452490	0.978027	0.978027

Table B.2: Correlation between the ChIP-seq predictions of histone association coefficients across different cell-types and/or SNAP predictions. Key: n – number of overlapping predictors, r – Pearson correlation coefficient, t – t test statistic for Pearson correlation coefficient, p – one-tailed t-test p-value, q – Benjamini/Hochberg adjusted p-value.

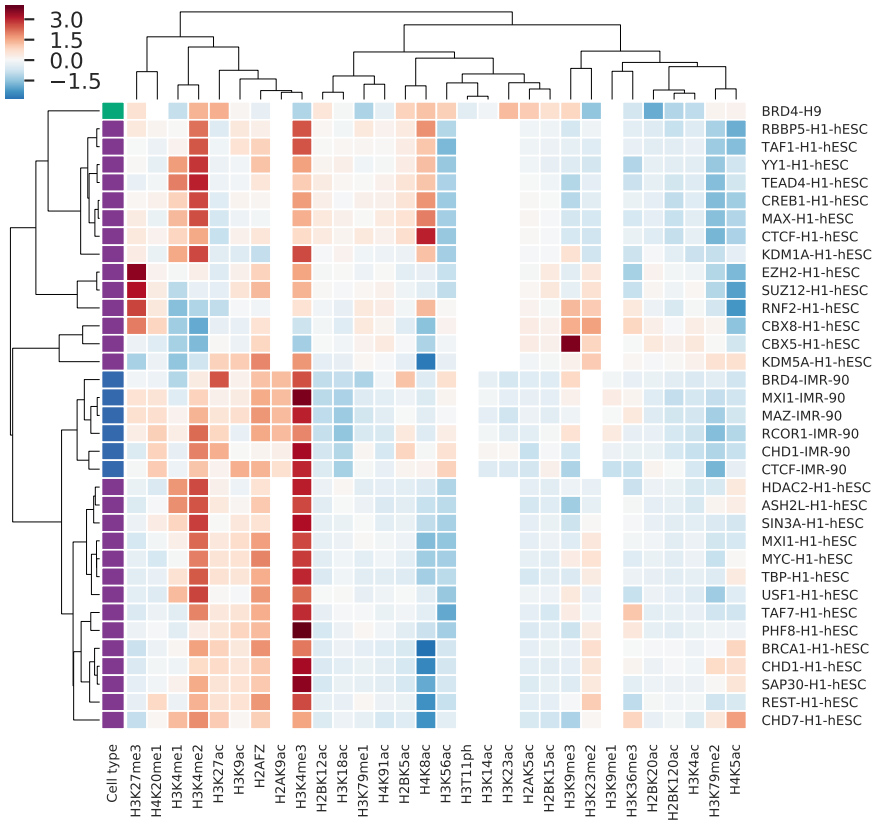

Figure B.8: Complete set of histone-association coefficients estimated from ChIP-seq data.

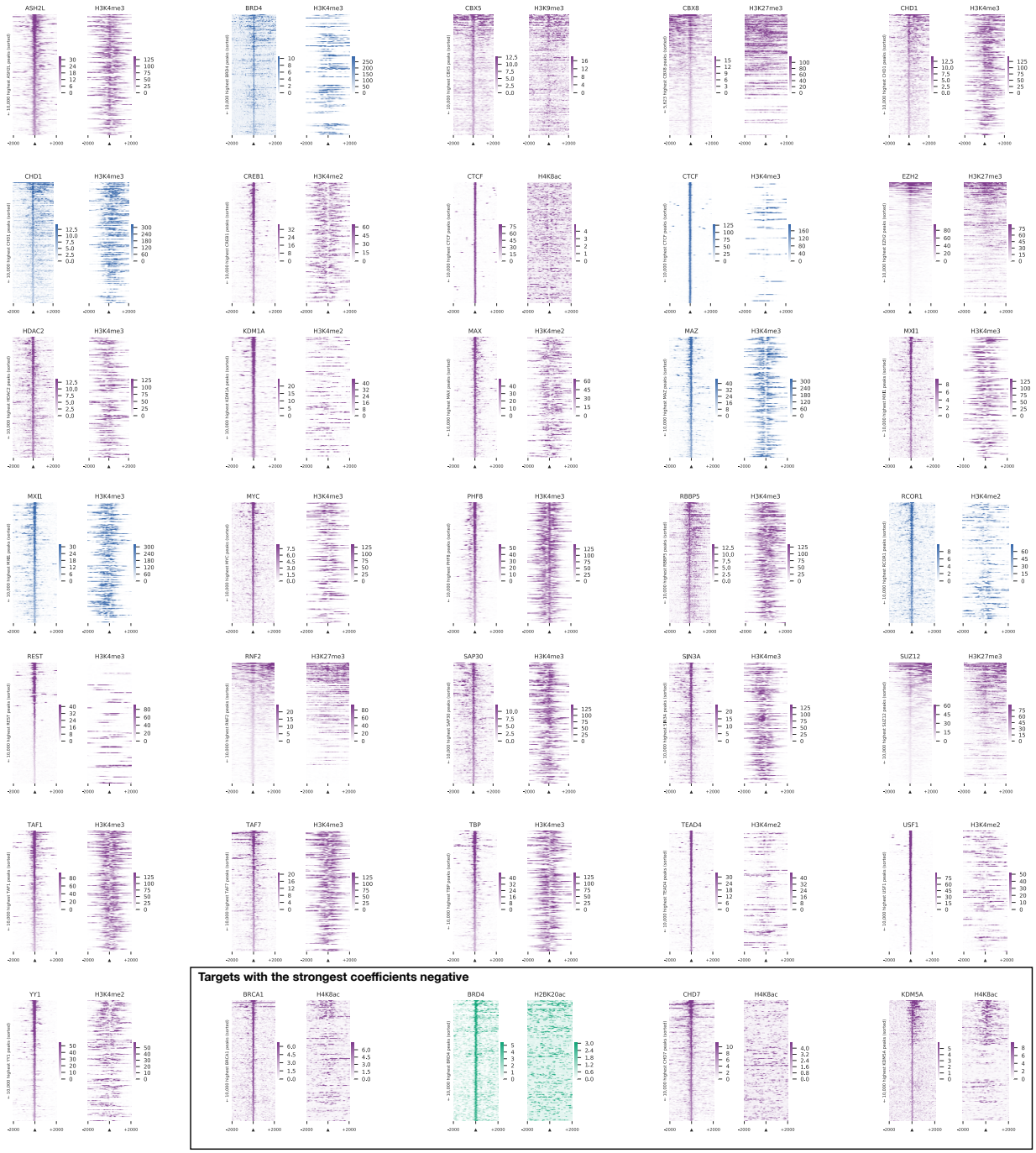

Figure B.9: Heatmap visualisation for the strongest targets predicted for each of the proteins analysed using ChIP-seq prediction model. Top 10,000 highest peaks for each of the targets are visualised along with the profiles of histone modifications in the neighbouring regions for the strongest predictors for these targets, as seen in Figure 3.8. For all targets the strongest predictors are positive, except for the four targets in bottom right corner where these values are negative.

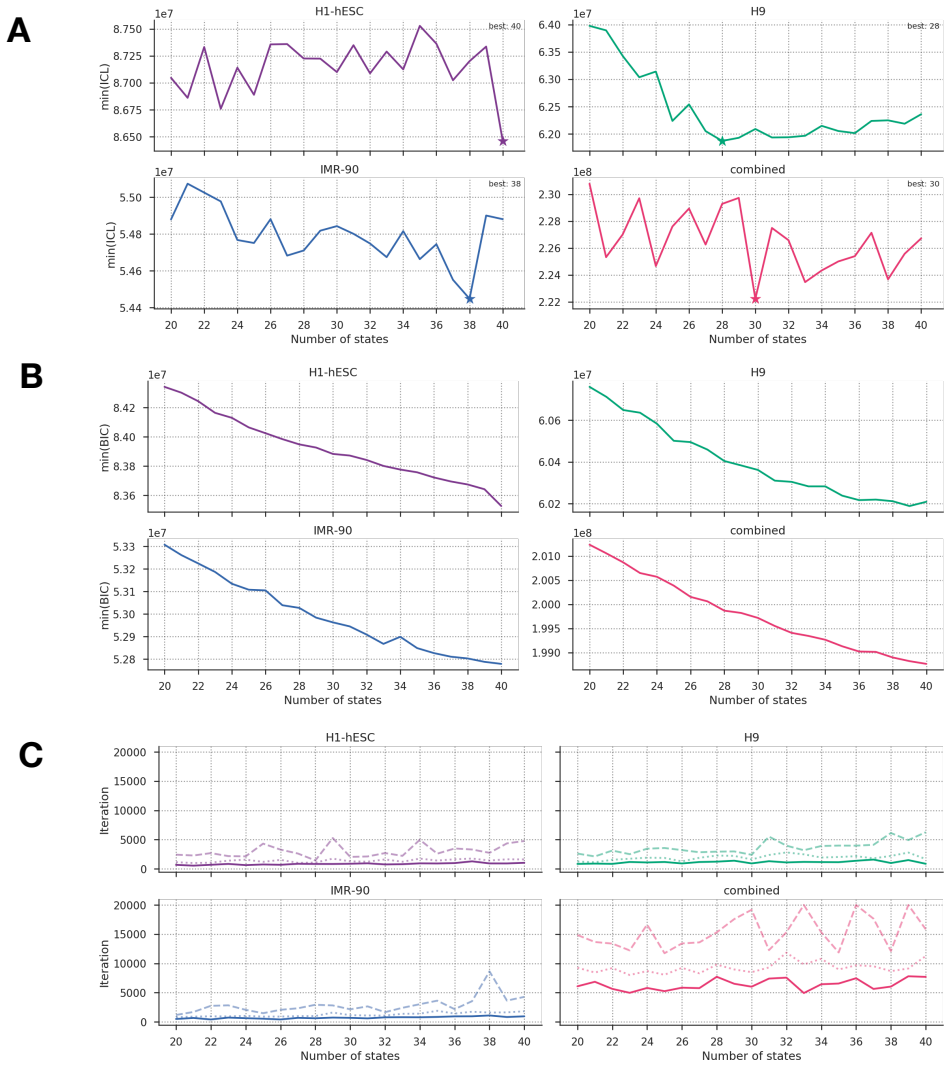

Figure B.10: Main performance indicators for the BMM model training procedure described in section 3.2.3. (A) Integrated Complete Likelihood (ICL) scores for the models evaluated with 20-40 states. The best number of states is highlighted with a star. Four panels represent three single cell type models as well as the combined model. (B) Equivalent plots showing the Bayesian Information Criterion (BIC) statistic for such models. (C) Minimum, median and maximum number of iterations required to fit the models until convergence. Model training has been forced to stop when 20,000 iterations have been reached.

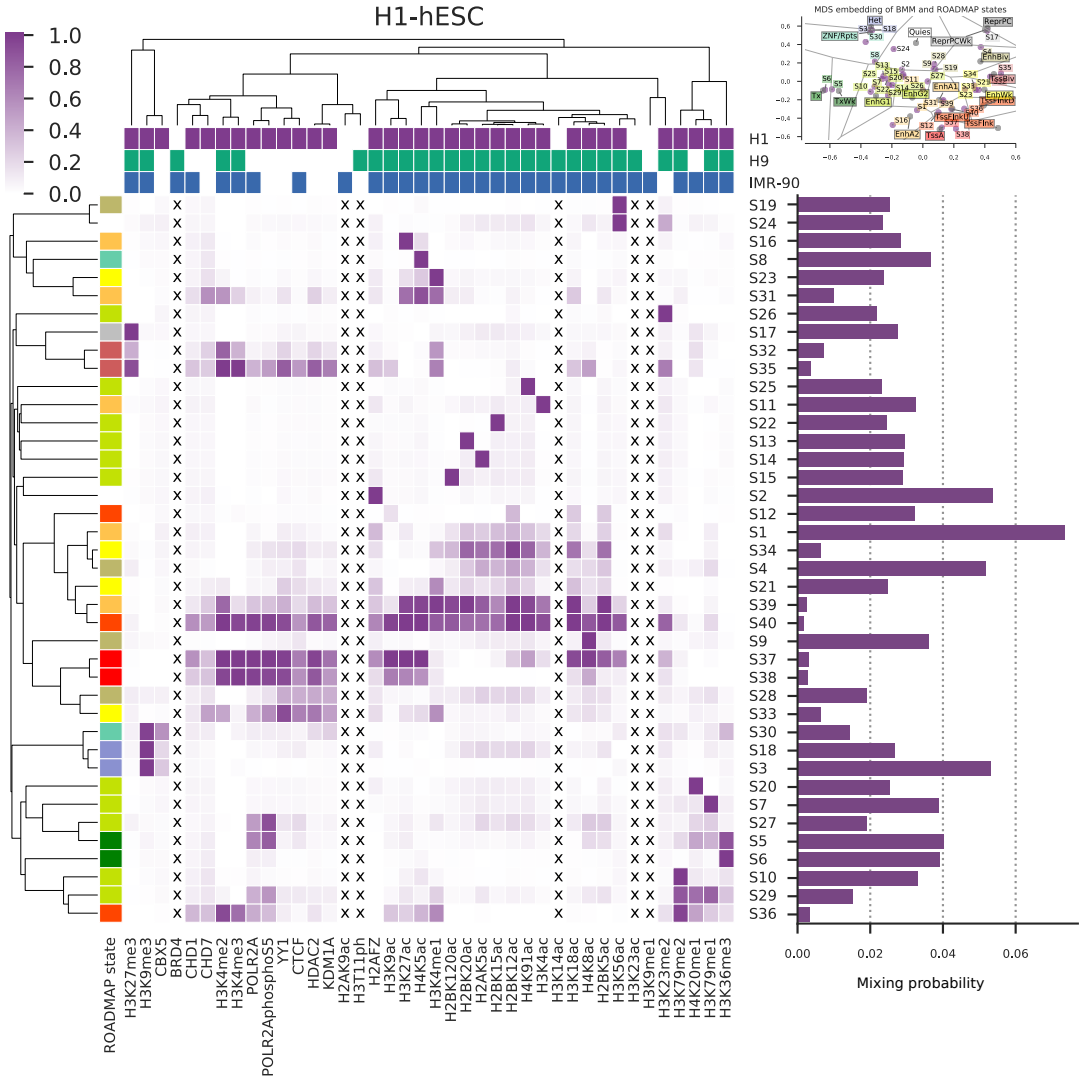

Figure B.11: A summary of the inferred BMM model when trained using the data from H1-hESC cell-line only. See legend Figure 3.9 for the description of the layout. Columns marked with an "X" represent tracks missing in the cell-line specific dataset.

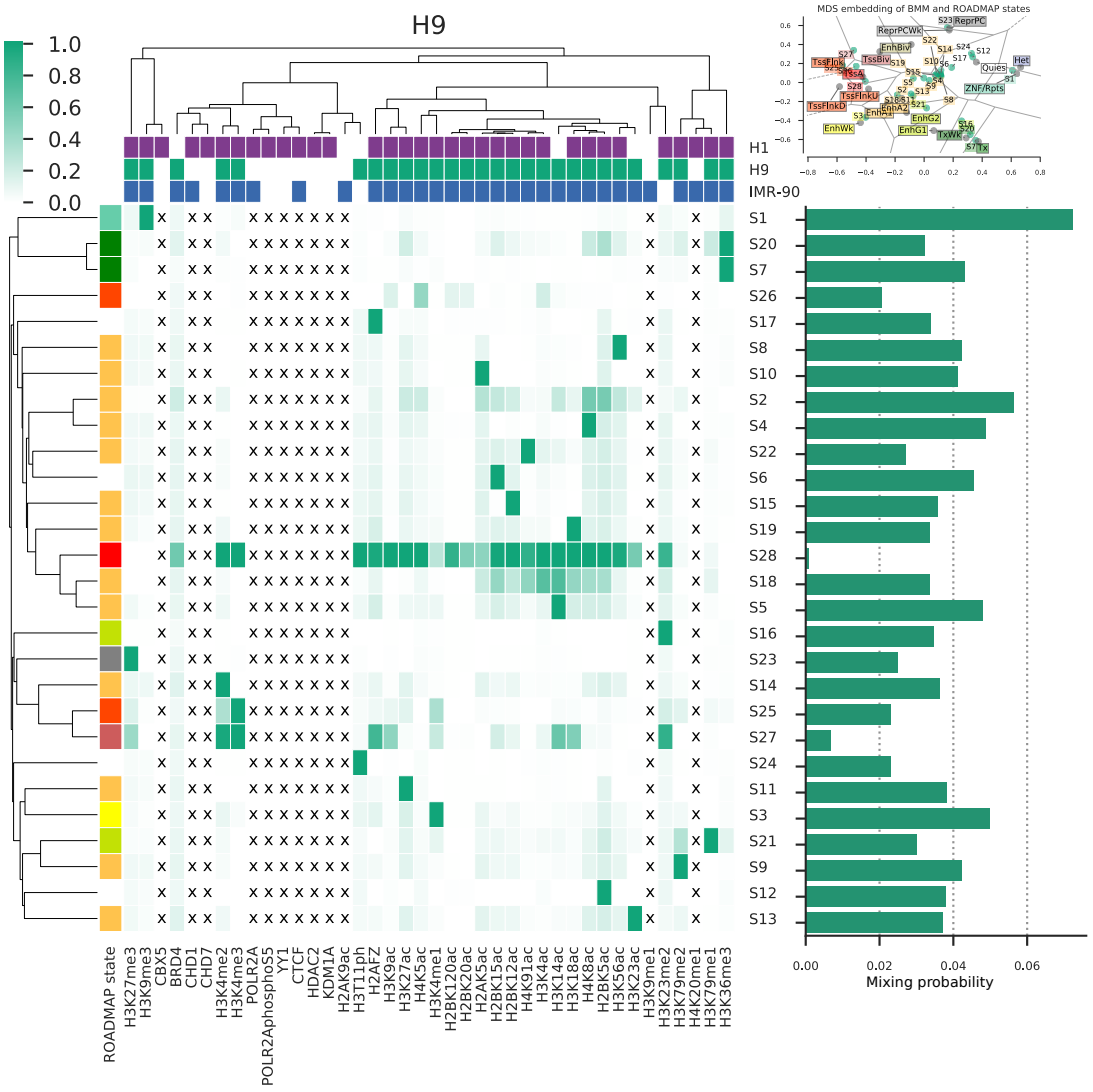

Figure B.12: A summary of the inferred BMM model when trained using the data from H9 cell-line only. See legend Figure 3.9 for the description of the layout. Columns marked with an "X" represent tracks missing in the cell-line specific dataset.

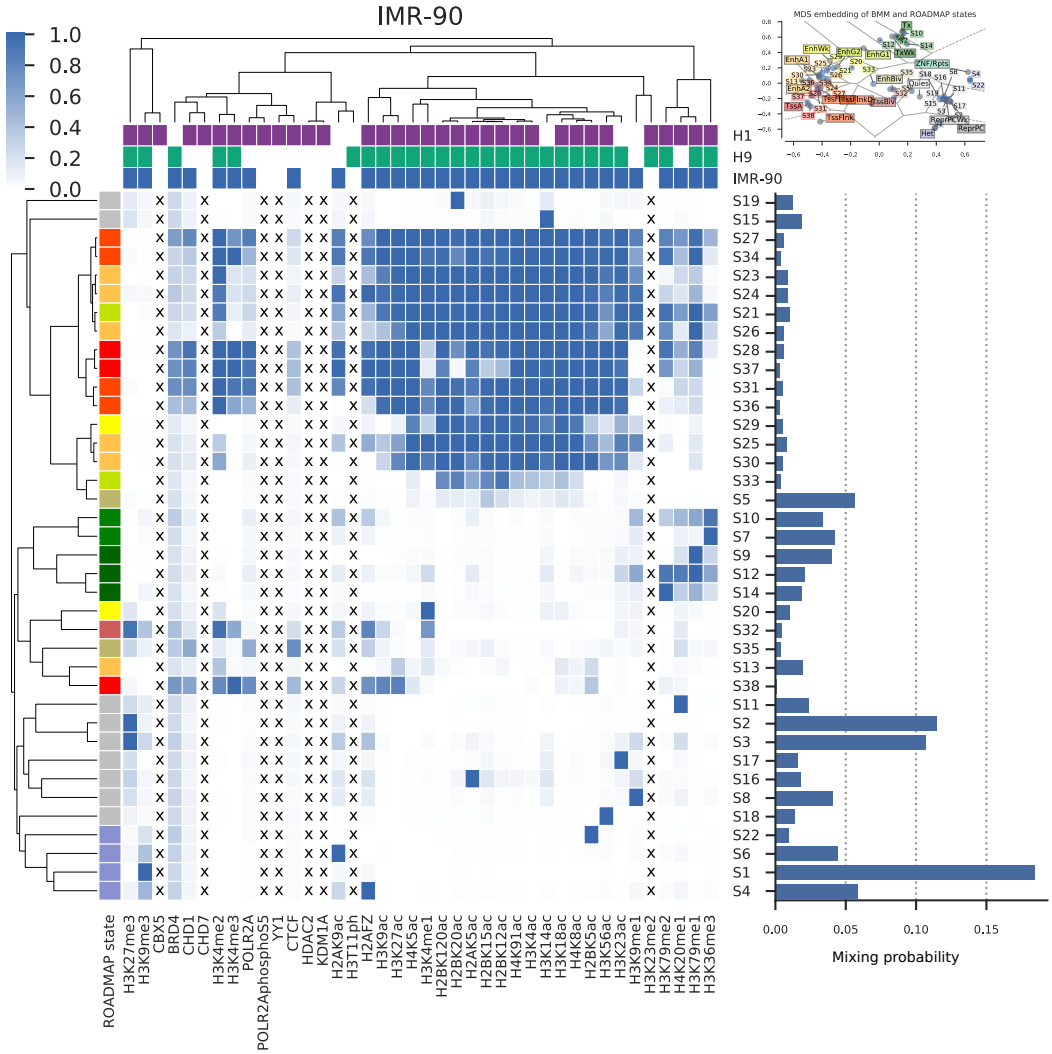

Figure B.13: A summary of the inferred BMM model when trained using the data from IMR-90 cell-line only. See legend Figure 3.9 for the description of the layout. Columns marked with an "X" represent tracks missing in the cell-line specific dataset.

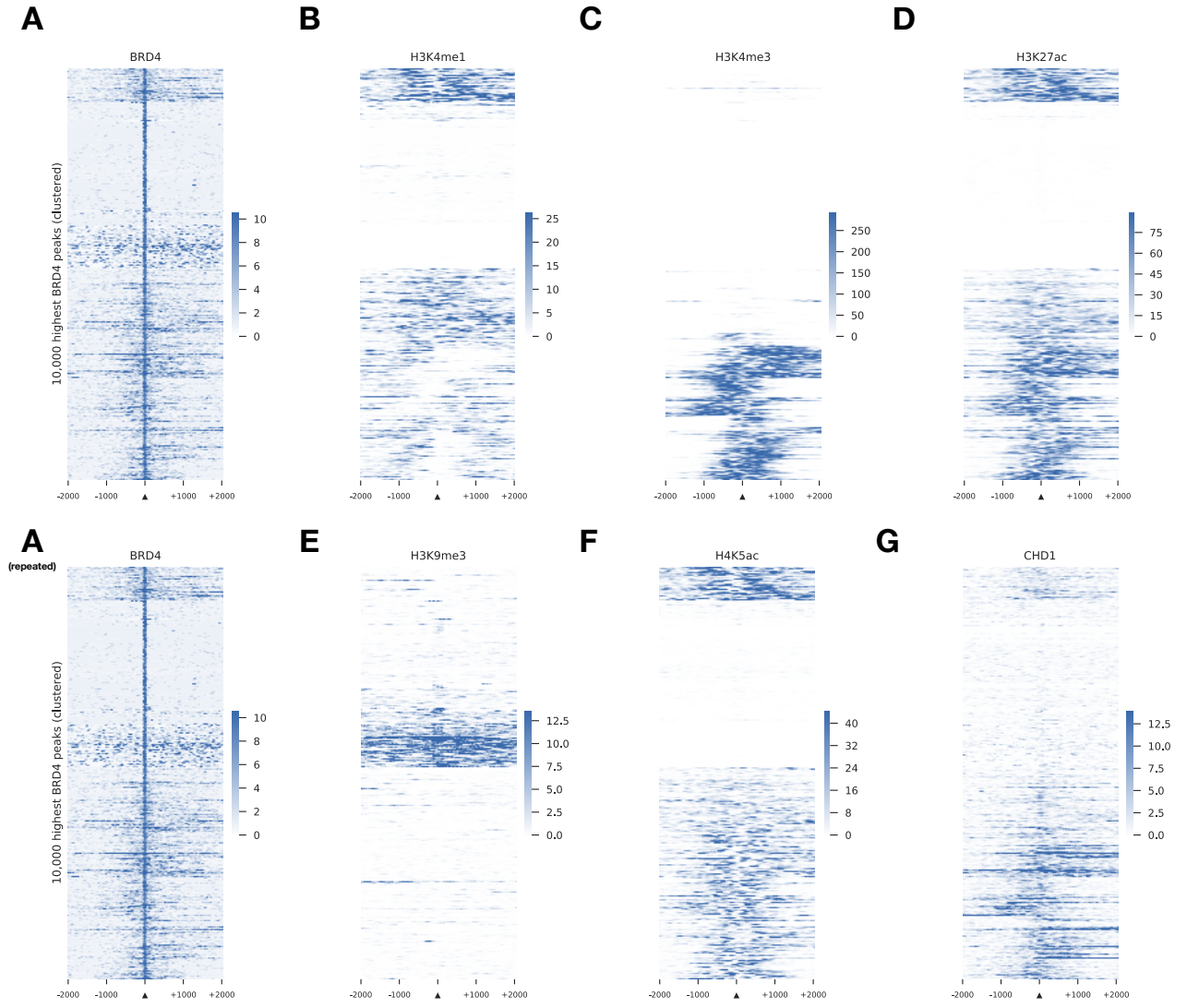

Figure B.14: Profiles of the signal for a selection of ChIP-seq tracks around the top 10,000 highest BRD4 peaks in IMR-90 cell line. The profiles are clustered across all of the visualised dimensions. Note that only a subset of the visualised peaks contain signal for H3K4me1, H3K4me3, H3K27ac or H4K5ac (B, C, D and F). It is not clear what drives the binding across other enriched peaks, a subset of which contains H3K9me3, for instance (E). Interestingly CHD1 localises near BRD4 in most of the visualised peaks. Recruitment of CHD1 by the transcription elongation machinery has been previously reported in [267], this relationship was also evident in correlation between the emission probabilities for the two marks (Figure 3.9 A).

B.2 Large-scale mononucleosome purification from HeLa cells

In this appendix we list the optimised protocol for mononucleosome preparation from HeLa cells used in this thesis.

This protocol is a combination of the protocols published in [268–271] with some minor modifications.

B.2.1 Reagents & buffers

Buffer A	
HEPES pH 7.9 (NaOH)	10 mM
KCl	10mM
MgCl2	1.5mM
Sucrose	340 mM
Glycerol	10% (v/v)

Sucrose cushion	
HEPES pH 7.9 (NaOH)	10 mM
Sucrose	30% (w/v)
MgCl2	1.5 mM

Chromatin recovery buffer	
HEPES pH 7.9 (NaOH)	10 mM
EDTA	0.2 mM

Nuclear storage buffer	
HEPES pH 7.9 (NaOH)	10 mM
KCl	10 mM
MgCl2	1.5 mM
Sucrose	340 mM

Supplements (all buffers, add before use)	
PMSF	0.5mM
DTT	1mM
cOmplete Protease Inhibitor Coctail, EDTA-Free	1X
Sodium Butyrate	10mM

Also:

- Phosphate Buffered Saline (PBS) + 0.1% NP-40
- PBS
- EDTA (0.5M)
- EGTA (0.5M)
- Nuclease micrococcal from Staphylococcus aureus (MNase), Sigma N3755. Other MNases can be used, but the amount added will have to be adjusted. Prepare a 1U/ul solution in 50% (v/v) glycerol, 50mM TRIS Ph8, 0.05 mM CaCl2. Store in −20°C

Prepare all buffers in advance, sterile filter using 0.2um filters and keep at 4°C.

B.2.2 Protocol

Cell culture

1. Grow cells to a density of million cells per ml and harvest by centrifugation at 250 g, using a pre-cooled temperature at 4°C.
2. Wash cells once in PBS, spin down 5 minutes in the centrifuge at 250 g at 4°C.
3. Wash cells once more in PBS augmented with protease inhibitors and sodium butyrate. Collect the cells by centrifugation at 250g for 5 minutes at 4°C.
4. Assess the packed cell volume (PCV). Flash freeze the cells in liquid nitrogen and store at −80°C.

Nuclei isolation (Rapid protocol [269])

1. Prepare enough 2ml low-bind microcentrifuge tubes to fill all the buckets available.
2. Estimate the maximum packed cell volume (PCV) that can be processed in one go, by calculating the combined volume of the 2ml microcentrifuge tubes, divided by 10 (i.e. 6ml for centrifuges with 30 buckets). Thaw an aliquot of cells that corresponds to this PCV

3. Resuspend the cells in PBS (with additives) and divide equally along the microcentrifuge tubes. Compute the PCV per microcentrifuge tube. The subsequent steps below will now refer to this calculated PCV per microcentrifuge tube instead of the total PCV.
4. The subsequent steps have to be performed rapidly
5. Do a pop-spin in the microcentrifuge (10,000 rpm, for approximately 10-15s)
6. Remove supernatant (decanting is usually OK)
7. Suspend cells in 9 PCV volumes of PBS + 0.1% NP-40, mix the cells using P1000 by gently pipetting up and down *exactly* five times. Here the pipette acts as a dounce homogeniser, which break the cells. Solution should become completely solubilised.
8. Pop-spin the nuclei again, remove the supernatant.
9. Repeat the suspension in PBS + 0.1% NP-40 and pipetting up and down, pop-spin and supernatant removal one more time.
10. *Immediately* suspend the nuclei in 6PCV of Buffer A. Mix gently, but thoroughly, to dilute the detergent.

Cleaning the nuclei

This steps removes residual cell debris from the extraction, as well as lysed nuclei and therefore improves reproducibility [271].

1. Apply the suspension to a chilled 7.5ml sucrose cushion (with supplements) in 50ml conical tube. Load 0.5-1 ml PCV per cushion. Agitate the surface tension gently with a stripette to allow some mixing to occur at the top level of the cushion
2. Spin the nuclei down through the cushion for 12 minutes at 1300g at 4°C.
3. Gently resuspend the pelleted nuclei in 1 PCV of Buffer A and transfer to 1.5ml microcentrifuge tube.
4. Rinse the 50ml microcentrifuge tube with another 1 PCV of buffer A, and combine with the previously extracted nuclei.
5. Centrifuge suspension for 5min at 1300g at 4°C.
6. Wash nuclei in 2 PCV Buffer A and centrifuge for 5 min at 1300g, 4°C.
7. *Optionally, if breakpoint is desired:* Suspend the nuclei in no more than 100µl of buffer (approximately 1/4 PCV) A. Add the same amount of nuclei storage buffer, allow to equilibrate for 15min on ice before storing at −80°C.

Estimation of digestion time

Depending on the nuclear preparation, and other environmental conditions the perfect MNase digestion time varies from sample to sample. Therefore, for each sample we first perform an experiment to estimate the best digestion time for that sample, which we then use in the large-scale digest.

1. Dilute the nuclei suspension in buffer A to achieve the concentration of approximately 1.4 ng/ul (+/- 0.2 ng/ul) of DNA as measured by A260 (for measurement take a small aliquot and dilute it 1:50 in 2M NaCl). The final volume should be approximately 2.25 PCV, but it will vary depending on the efficiency of purification. Dilute/reconcentrate if necessary.
2. Prepare 10 low-bind microcentrifuge tubes, and add 7.5µl of 50mM EDTA to each (it will serve to be 10mM EDTA final, once the sample is added).
3. Take 330ul aliquot of the purified nuclei
4. Add CaCl2 to the final concentration of exactly 1mM.
5. Equilibrate the nuclei in 37°C water bath for 5 minutes.
6. Add 1U of MNase per 1mg of DNA (as measured earlier by A260). Make sure to pre-dilute the MNase at least 1:10 in buffer A as the high amount of glycerol in the storage buffer will make pipetting inaccurate. Mix gently with P1000
7. Incubate the samples at 37°C water bath. At regular 3 minute intervals (up to 30 minutes), take 30ul of the digestion reaction and place it in the vial containing EDTA to stop the digestion.
8. Pellet the stopped reactions at maximum speed for 5 minutes at 4°C. Recover supernatant
9. Clean up the DNA using PCR cleanup kit, and run the timed extraction on 2% agarose gel. Estimate the timepoint which produces most mononucleosomes.

Large-scale digestion (S1 fraction)

1. Take the rest of the diluted nuclear stock and divide into aliquots which fit into 1.5ml low-bind vials and pre-incubate them in 37°C water bath as before.
2. Again, add 1U of MNase per 1mg of DNA and let the digestion reaction to proceed for the estimated time.
3. Stop the digestion by adding EDTA to final concentration of 10mM.
4. Pellet the insoluble proteins and at maximum speed for 5 minutes at 4°C.

5. Recover the supernatant, this is the soluble fraction S1. Keep it at 4°C.
6. PCR clean up an aliquot and run on a 2% agarose gel to estimate purity and concentration, similarly, samples can be run on 18% SDS-PAGE gel.

Recovery (optional, S2 fraction)

1. Resuspend pellet in 50ul chromatin recovery buffer.
2. Rotate 1 hour to overnight at 4°C and centrifuge for 15 min at maximum speed of 4°C.
3. This fraction is the S2 fraction of nucleosomes.
4. Run the samples on 2% agarose gel to estimate purity.
5. If desired, this fraction can be pooled with S1.
6. If desired, the remaining chromatin can be resolubilised using a chromatin recovery buffer augmented with 650 mM NaCl (S3 fraction), but this may disrupt the nucleosome.

At this point the purified nucleosomes can be directly dialysed into a suitable buffer using 7K MWCO membrane. If an increase in purity is designed, nucleosomes can be purified even further using sucrose gradient, see [271].

Appendix C

Accession lists

Table C.1: Accessions of datasets used in ChIP-seq experiments. Accessions are prefixed either with "encode" or "sra" tag, specifying the data sources ENCODE and SRA. Non-encode datasets were previously published in [187] and [188]

Cell type	Target	Accessions
H1-hESC	ASH2L	encode:ENCFF311BDI; encode:ENCFF879QBW
	BRCA1	encode:ENCFF000WQT; encode:ENCFF000WQV
	CBX5	encode:ENCFF372AND; encode:ENCFF650WVN
	CBX8	encode:ENCFF529XNL; encode:ENCFF979OQS
	CHD1	encode:ENCFF350FRQ; encode:ENCFF380VMM; encode:ENCFF473ZWP; encode:ENCFF000WRJ; encode:ENCFF000WRL
	CHD7	encode:ENCFF000AVL; encode:ENCFF349SFU; encode:ENCFF580HUJ
	CREB1	encode:ENCFF000ONP; encode:ENCFF000ONS
	CTCF	encode:ENCFF000ONR; encode:ENCFF000OOF; encode:ENCFF000AVS; encode:ENCFF000AVU; encode:ENCFF000RSJ
	EZH2	encode:ENCFF000AVZ; encode:ENCFF000AWC; encode:ENCFF458LKN; encode:ENCFF090FUA; encode:ENCFF225ZAL
	H2AFZ	encode:ENCFF820FVP; encode:ENCFF000AWI; encode:ENCFF000AWJ
	H2AK5ac	encode:ENCFF114DJW; encode:ENCFF799XSQ
	H2BK120ac	encode:ENCFF376UOP; encode:ENCFF169ASV; encode:ENCFF820YCB
	H2BK12ac	encode:ENCFF533SPP; encode:ENCFF085WVC
	H2BK15ac	encode:ENCFF926QHW; encode:ENCFF851KPI
	H2BK20ac	encode:ENCFF891RQT; encode:ENCFF728BZD
	H2BK5ac	encode:ENCFF541IYF; encode:ENCFF520GVN
	H3K18ac	encode:ENCFF155MOK; encode:ENCFF173RMI
	H3K23me2	encode:ENCFF787QFF; encode:ENCFF705ZSE
	H3K27ac	encode:ENCFF481YKF; encode:ENCFF308XJC; encode:ENCFF000AWR; encode:ENCFF000AWV
	H3K27me3	encode:ENCFF539PKS; encode:ENCFF374VEK; encode:ENCFF722IIA; encode:ENCFF137PVL; encode:ENCFF171DUV; encode:ENCFF000AWX; encode:ENCFF000AWY; encode:ENCFF092CMH; encode:ENCFF717YEM; encode:ENCFF478FWQ; encode:ENCFF092MNN; encode:ENCFF908DYY; encode:ENCFF012CJX; encode:ENCFF638SCC

Continued on next page

Cell type	Target	Accessions
	H3K36me3	encode:ENCFF034AQJ; encode:ENCFF759KCC; encode:ENCFF000AXG; encode:ENCFF000AXH; encode:ENCFF534HXO; encode:ENCFF198FZX; encode:ENCFF764IPK; encode:ENCFF458WNH; encode:ENCFF685HBP; encode:ENCFF203VOI; encode:ENCFF929REJ; encode:ENCFF659PKA; encode:ENCFF590KVU; encode:ENCFF037BES
	H3K4ac	encode:ENCFF810EZK
	H3K4me1	encode:ENCFF841YYX; encode:ENCFF713WZR; encode:ENCFF355BBC; encode:ENCFF000AXJ; encode:ENCFF000AXM; encode:ENCFF074BMK; encode:ENCFF719MHS; encode:ENCFF302UMJ; encode:ENCFF332IMT
	H3K4me2	encode:ENCFF388MXW; encode:ENCFF181QBV; encode:ENCFF095UOK; encode:ENCFF000AXQ; encode:ENCFF000AXR
	H3K4me3	encode:ENCFF156FYC; encode:ENCFF654NIP; encode:ENCFF813UAU; encode:ENCFF038JDF; encode:ENCFF438IKZ; encode:ENCFF723SKE; encode:ENCFF000AXS; encode:ENCFF000AXZ; encode:ENCFF193VAJ; encode:ENCFF405HHM; encode:ENCFF296CUY; encode:ENCFF739WMB
	H3K56ac	encode:ENCFF257BUI
	H3K79me1	encode:ENCFF301KZO; encode:ENCFF376WET
	H3K79me2	encode:ENCFF580OHZ; encode:ENCFF519ZRJ; encode:ENCFF000AYD; encode:ENCFF000AYF
	H3K9ac	encode:ENCFF779MAL; encode:ENCFF586LPF; encode:ENCFF000AYG; encode:ENCFF000AYH; encode:ENCFF168NRP; encode:ENCFF181TET; encode:ENCFF888QAU
	H3K9me3	encode:ENCFF546MTH; encode:ENCFF699FAQ; encode:ENCFF298XWU; encode:ENCFF084JSH; encode:ENCFF214DMX; encode:ENCFF557QCN; encode:ENCFF000AWO; encode:ENCFF000AWP; encode:ENCFF421XOC; encode:ENCFF868PQG; encode:ENCFF711AUT; encode:ENCFF303XAI; encode:ENCFF320QCI; encode:ENCFF832DCJ
	H4K20me1	encode:ENCFF387MHG; encode:ENCFF449SLH; encode:ENCFF000AYP; encode:ENCFF000AYQ
	H4K5ac	encode:ENCFF871EJI; encode:ENCFF940AZI
	H4K8ac	encode:ENCFF557LZC; encode:ENCFF783ZRQ
	H4K91ac	encode:ENCFF653DEA
	HDAC2	encode:ENCFF000OPK; encode:ENCFF000OPL; encode:ENCFF000AYV; encode:ENCFF524BES
	Input	encode:ENCFF000WUI; encode:ENCFF202IDR; encode:ENCFF791VTH; encode:ENCFF610XLV; encode:ENCFF451EVC; encode:ENCFF422GYG; encode:ENCFF000WUW; encode:ENCFF634CSE; encode:ENCFF002ECT; encode:ENCFF002ECV; encode:ENCFF051RXP; encode:ENCFF176RJQ; encode:ENCFF000OSP; encode:ENCFF126XLK; encode:ENCFF835IOE; encode:ENCFF094GYG; encode:ENCFF963LOW; encode:ENCFF795IME; encode:ENCFF303YPV; encode:ENCFF740ECI; encode:ENCFF145GCJ; encode:ENCFF659UFR; encode:ENCFF164ZWM; encode:ENCFF809ZPN; encode:ENCFF508ZFI; encode:ENCFF763DAQ; encode:ENCFF805KPU; encode:ENCFF036EGF; encode:ENCFF191QXK; encode:ENCFF663FLW; encode:ENCFF689EIT; encode:ENCFF635VQA; encode:ENCFF064XIS; encode:ENCFF472QMD; encode:ENCFF138KKQ; encode:ENCFF368CYL; encode:ENCFF393KUX; encode:ENCFF352RHK; encode:ENCFF003YFN; encode:ENCFF000OSS; encode:ENCFF000OST; encode:ENCFF952GNX
	KDM1A	encode:ENCFF503DAG; encode:ENCFF220HNB; encode:ENCFF935HUK
	KDM5A	encode:ENCFF281HTF; encode:ENCFF769WTD
	MAX	encode:ENCFF000OQD; encode:ENCFF000OQI
	MXI1	encode:ENCFF000WUC; encode:ENCFF000WUE
	MYC	encode:ENCFF000WSD; encode:ENCFF000WSF; encode:ENCFF000RSE
	PHF8	encode:ENCFF000AZT; encode:ENCFF346JZZ
	POLR2A	encode:ENCFF000RSM; encode:ENCFF000ORP; encode:ENCFF000ORR
	POLR2AphosphoS5	encode:ENCFF000ORG; encode:ENCFF000ORK
	RBBP5	encode:ENCFF279SFM; encode:ENCFF720HOL
	REST	encode:ENCFF079NVQ; encode:ENCFF794IGW; encode:ENCFF000OQQ; encode:ENCFF000OQY
	RNF2	encode:ENCFF517QCO; encode:ENCFF420XPW
	SAP30	encode:ENCFF622UBZ; encode:ENCFF270OEH
	SIN3A	encode:ENCFF000WVE; encode:ENCFF000WVG; encode:ENCFF000OTG; encode:ENCFF000OTJ
	SUZ12	encode:ENCFF523HNT; encode:ENCFF252NFT; encode:ENCFF061EJH; encode:ENCFF003KFS
	TAF1	encode:ENCFF000OVB; encode:ENCFF000OVC
	TAF7	encode:ENCFF000OVI; encode:ENCFF000OVV
	TBP	encode:ENCFF000WWI; encode:ENCFF000WWK
	TEAD4	encode:ENCFF000OWA; encode:ENCFF000OWD

Continued on next page

Cell type	Target	Accessions
	USF1	encode:ENCFF000OWE; encode:ENCFF000OWL
	YY1	encode:ENCFF000OWO; encode:ENCFF000OWS
H9	BRD4	sra:SRR1537736; sra:SRR1537737
	H2AFZ	encode:ENCFF063AKT
	H2AK5ac	encode:ENCFF646THG
	H2BK120ac	encode:ENCFF116XCI
	H2BK12ac	encode:ENCFF330JGJ
	H2BK15ac	encode:ENCFF123QHZ; encode:ENCFF062UVB
	H2BK20ac	encode:ENCFF252DLP
	H2BK5ac	encode:ENCFF060GNH
	H3K14ac	encode:ENCFF516VDL; encode:ENCFF636LTX
	H3K18ac	encode:ENCFF598DSH
	H3K23ac	encode:ENCFF873QZI
	H3K23me2	encode:ENCFF187MRW
	H3K27ac	encode:ENCFF081KCW
	H3K27me3	encode:ENCFF212TLT; encode:ENCFF884SZD; encode:ENCFF942UDZ; encode:ENCFF007RAN
	H3K36me3	encode:ENCFF262NPS; encode:ENCFF347SIP; encode:ENCFF053JOS
	H3K4ac	encode:ENCFF495ENG; encode:ENCFF761XAE; encode:ENCFF641NXU; encode:ENCFF642ITQ
	H3K4me1	encode:ENCFF210BMG; encode:ENCFF579CSL; encode:ENCFF625FYD; encode:ENCFF626XCW
	H3K4me2	encode:ENCFF394TPE
	H3K4me3	encode:ENCFF909NXO; encode:ENCFF744CHR; encode:ENCFF562FPJ
	H3K56ac	encode:ENCFF171BHT
	H3K79me1	encode:ENCFF781USD; encode:ENCFF855BRY; encode:ENCFF906NFS; encode:ENCFF756KNI
	H3K79me2	encode:ENCFF892BHU; encode:ENCFF402CTD
	H3K9ac	encode:ENCFF278SGU
	H3K9me3	encode:ENCFF817YWK; encode:ENCFF234MWJ; encode:ENCFF318GUA; encode:ENCFF029RJH
	H3T11ph	encode:ENCFF828GGJ
	H4K5ac	encode:ENCFF439PIF
	H4K8ac	encode:ENCFF780KZJ
	H4K91ac	encode:ENCFF011LRK; encode:ENCFF140WCY; encode:ENCFF697DCZ
	Input	encode:ENCFF533UCC; encode:ENCFF229SZZ; encode:ENCFF359BBF; encode:ENCFF487VQV; encode:ENCFF416GCS; sra:SRR1537734
IMR-90	BRD4	sra:SRR2748697; sra:SRR3287546
	CHD1	encode:ENCFF000YBW; encode:ENCFF000YBU
	CTCF	encode:ENCFF000YCK; encode:ENCFF000YCM
	H2AFZ	encode:ENCFF305XOA; encode:ENCFF191LBR
	H2AK5ac	encode:ENCFF146YJK; encode:ENCFF137ZYY; encode:ENCFF329HCY
	H2AK9ac	encode:ENCFF626SHH; encode:ENCFF689TFP
	H2BK120ac	encode:ENCFF018XQO; encode:ENCFF230GVA; encode:ENCFF993MEW; encode:ENCFF473KQM
	H2BK12ac	encode:ENCFF127RDZ; encode:ENCFF569PIP; encode:ENCFF789BOJ
	H2BK15ac	encode:ENCFF318ABB; encode:ENCFF758EPO; encode:ENCFF730GKX; encode:ENCFF049QNI
	H2BK20ac	encode:ENCFF047TNQ; encode:ENCFF912FWW; encode:ENCFF526DHU; encode:ENCFF393CTC
	H2BK5ac	encode:ENCFF512FJB; encode:ENCFF165DQP; encode:ENCFF733WHI
	H3K14ac	encode:ENCFF622CVD; encode:ENCFF515BXQ; encode:ENCFF349FXU; encode:ENCFF426BVN
	H3K18ac	encode:ENCFF635ZKA; encode:ENCFF146RUB; encode:ENCFF974SJG; encode:ENCFF803DFS
	H3K23ac	encode:ENCFF393UFA; encode:ENCFF348KZV; encode:ENCFF179XQK; encode:ENCFF374WWV
	H3K27ac	encode:ENCFF419OFF; encode:ENCFF734UGN; encode:ENCFF704RXG; encode:ENCFF200XEH; encode:ENCFF437DDC; sra:SRR2748694; sra:SRR3287543
	H3K27me3	encode:ENCFF930UWF; encode:ENCFF486NAQ; encode:ENCFF900TAE; encode:ENCFF711HPS; encode:ENCFF756OBX; encode:ENCFF615XDP
	H3K36me3	encode:ENCFF847BTL; encode:ENCFF252MOR; encode:ENCFF832NKC; encode:ENCFF881MCC; encode:ENCFF437GEN
	H3K4ac	encode:ENCFF942HBH; encode:ENCFF169WDC; encode:ENCFF064QXB; encode:ENCFF515VZX
	H3K4me1	encode:ENCFF610HSE; encode:ENCFF555JXQ; encode:ENCFF387LHN; encode:ENCFF350IDI; encode:ENCFF123RFO; encode:ENCFF553HYG; encode:ENCFF924XPW
	H3K4me2	encode:ENCFF103PXQ; encode:ENCFF561LGA; encode:ENCFF649IXK; encode:ENCFF737MCM

Continued on next page

Cell type	Target	Accessions
	H3K4me3	encode:ENCFF214ISW; encode:ENCFF591EQV; encode:ENCFF742UXQ; encode:ENCFF172LYA
	H3K56ac	encode:ENCFF434BAM; encode:ENCFF980FLI; encode:ENCFF486FRM
	H3K79me1	encode:ENCFF682GXA; encode:ENCFF451QWY; encode:ENCFF419MLG; encode:ENCFF435IRD; encode:ENCFF298ZOC; encode:ENCFF394HYQ; encode:ENCFF187NVD
	H3K79me2	encode:ENCFF889AII; encode:ENCFF195TDA; encode:ENCFF500BZD; encode:ENCFF548SYP
	H3K9ac	encode:ENCFF199FDT; encode:ENCFF545QSC; encode:ENCFF586BPJ
	H3K9me1	encode:ENCFF877GMR; encode:ENCFF499EGR; encode:ENCFF691TDE; encode:ENCFF612FSU; encode:ENCFF880GRB
	H3K9me3	encode:ENCFF865DWP; encode:ENCFF154QZW; encode:ENCFF283STX; encode:ENCFF800GRJ; encode:ENCFF042EWI; encode:ENCFF535LUF
	H4K20me1	encode:ENCFF469QWN; encode:ENCFF271CRL; encode:ENCFF276ERF; encode:ENCFF465DNJ
	H4K5ac	encode:ENCFF378DBQ; encode:ENCFF842CLZ; encode:ENCFF404SHQ; encode:ENCFF345AUK
	H4K8ac	encode:ENCFF684TXQ; encode:ENCFF061JXC; encode:ENCFF503COJ; encode:ENCFF338KSW; encode:ENCFF296EFK
	H4K91ac	encode:ENCFF071AVO; encode:ENCFF312XAZ; encode:ENCFF863SGB
	Input	encode:ENCFF484LKZ; encode:ENCFF389FEH; encode:ENCFF769XXQ; encode:ENCFF956KYS; encode:ENCFF855STX; encode:ENCFF740CFC; encode:ENCFF272HWA; encode:ENCFF282SBH; encode:ENCFF421HCG; encode:ENCFF679UAT; encode:ENCFF584LXI; encode:ENCFF492DKQ; encode:ENCFF000YCQ; encode:ENCFF000YCS; encode:ENCFF838NYU; sra:SRR2748700; sra:SRR3287549
	MAZ	encode:ENCFF000YDG; encode:ENCFF000YDI
	MXI1	encode:ENCFF000YDP; encode:ENCFF000YDR
	POLR2A	encode:ENCFF000YDX; encode:ENCFF000YDZ
	RCOR1	encode:ENCFF000YCB; encode:ENCFF000YCD

Table C.2: List of protein complexes and their members, as well as the sources for annotation

Complex	Members	Members not detected in SNAP	Sources
40S Ribosomal subunit	FAU, RPS10/RPS10-NUDT3, RPS11, RPS12, RPS13, RPS14, RPS15, RPS15A, RPS16, RPS17, RPS18, RPS19, RPS2, RPS20, RPS23, RPS24, RPS25, RPS27A, RPS28, RPS29, RPS3, RPS3A, RPS4X, RPS5, RPS6, RPS7, RPS8, RPS9, RPSA	RPS21, RPS26, RPS27, RPS27L, RPS4Y1, RPS4Y2	[85, search:"40S Ribosomal"]
5FMC	LAS1L, PELP1, SENP3, TEX10, WDR18		[272]
60S Ribosomal subunit	RPL10, RPL10A, RPL11, RPL12, RPL13, RPL13A, RPL14, RPL15, RPL17/RPL17-C18orf32, RPL18, RPL18A, RPL19, RPL21, RPL22, RPL23, RPL23A, RPL24, RPL26/RPL26L1, RPL27, RPL27A, RPL28, RPL29, RPL3, RPL30, RPL31, RPL32, RPL34, RPL35, RPL35A, RPL36, RPL36A/RPL36AL, RPL37A, RPL38, RPL39, RPL4, RPL5, RPL6, RPL7, RPL7A, RPL7L1, RPL8, RPLP0, UBA52	RIOX1, RIOX2, RPL10L, RPL37, RPL39L, RPL3L, RPL41, RPL9; RPL9P7; RPL9P8; RPL9P9	[85, search:"60S Ribosomal"]
ACF	BAZ1A, SMARCA5		[39, ACF][86, EBI-11739466]
AP-1 FOS-JUN	FOS, JUN	Q13469	[86, EBI-11840849,EBI-11890798]
AP-2 dimers	TFAP2A (1), TFAP2A (2), TFAP2C, TFAP2D	TFAP2B	[273]
APC/C	ANAPC1, ANAPC10, ANAPC11, ANAPC13, ANAPC15, ANAPC16, ANAPC2, ANAPC4, ANAPC5, ANAPC7, CDC16, CDC23, CDC26, CDC27		[274][275][208]
ASAP	ACIN1 (1), ACIN1 (2), ACIN1 (3), ACIN1 (4), RNPS1, SAP18		[276]
ATAC	DR1, KAT14 (1), KAT14 (2), KAT2A, MBIP, SGF29, TADA2A, TADA3, WDR5, YEATS2, ZZZ3	Q92831	[39, ATAC][86, EBI-13636711,EBI-13639399]
B-MYB-MuvB	LIN37, LIN52, LIN54, LIN9, MYBL2, RBBP4		[277]

Continued on next page

Complex	Members	Members not detected in SNAP	Sources
B-WICH	BAZ1B, DDX21, DEK, MYBBP1A, SF3B1, SIRT7, SMARCA5	ERCC6_HUMAN, MYO1C_HUMAN, O00159, Q03468	[39, B-WICH][86, EBI-13952597]
BAF	ACTB, ACTL6A, ARID1A, ARID1B, SMARCA2, SMARCA4 (1), SMARCA4 (2), SMARCB1, SMARCC1, SMARCC2 (1), SMARCC2 (2), SMARCD1, SMARCD2, SMARCD3, SMARCE1	ACL6B_HUMAN, DPF3_HUMAN, O94805, O94805, O94805, O94805	[39, BAF][86, EBI-15483057,EBI-16363954,EBI-16364038,EBI-16364117,EBI-16364199,EBI-16364344,EBI-16364423,EBI-16364506]
BAP1/ASXL2	ASXL2, BAP1		[278]
BCOR	BCOR, BCORL1 (1), BCORL1 (2), KDM2B, PCGF1, RING1, RNF2, RYBP, SKP1, USP7, YAF2		[39, BCOR]
BHC	GTF2I, HDAC1, HDAC2, HMG20B, KDM1A, PHF21A, RCOR1 (2), ZMYM2, ZMYM3	GSE1_HUMAN, ZN217_HUMAN	[39, BHC]
BRCA1-A	ABRAXAS1, BABAM1/AC010463.1, BABAM2, BARD1, BRCA1, BRCC3, UIMC1		[39, BRCA1-A]
BRCC	BABAM2, BARD1, BRCA1, BRCA2, BRCC3	RAD51_HUMAN, Q06609	[39, BRCC][86, EBI-12686412]
BRISC	BABAM1/AC010463.1, BABAM2, BRCC3	F175B_HUMAN	[39, BRISC]
Base Excision Repair Complex	POLB, POLE3, XRCC1 (1), XRCC1 (2)	APE1, Cyclin A, DNA ligase 1, POLA, POLD, UNG2	[279]
C complex spliceosome	ALYREF (1), ALYREF (2), AQR, CACTIN, CDC40, CDC5L, CRNKL1, CWC15, CWC22, CWC27, DDX23, DDX41, DDX5, DHX35, DHX38, DHX8, EFTUD2, EIF4A3, FRG1, GPATCH1, HNRNPA1, HNRNPA2B1, HNRNPA3, HNRNPC (1), HNRNPC (2), HNRNPF, HNRNPH1, HNRNPK, HNRNPM (1), HNRNPM (2), HNRNPR, HNRNPU, ISY1/ISY1-RAB43, LSM2, LSM3, MTREX, PABPC1, PLRG1, PNN, PPIE, PPIL1, PPIL3, PPWD1, PRPF19, PRPF4B, PRPF6, PRPF8, RALY, RBM22, RBM8A, RBMX, SART1, SF3A1, SF3A2, SF3A3, SF3B1, SF3B2, SF3B3, SLU7, SNRNP200, SNRNP40, SNRPA1, SNRPB2, SNRPD1, SNRPD2, SNRPD3, SNRPE, SNRPF, SNRPG, SNURF/SNRPB/SNRPN, SNW1, SRRM1, SRRM2, SRSF1, SYF2, SYNCRIP, TFIP11, U2AF1/U2AF1L5, WDR83, XAB2, ZCCHC8	DGCR14, MAGOH	[280, id:1181]
CAF-1	CHAF1A, CHAF1B, RBBP4		[39, CAF-1][86, EBI-1236585]
CDCA5-PDS5-Cohesin	CDCA5, PDS5A, RAD21, SMC3	PDS5B	[281]
CHRAC	BAZ1A, CHRAC1, POLE3, SMARCA5		[39, CHRAC][86, EBI-12551840]
CHTOP-methylosome	CHTOP, ERH	MEP50, PRMT1, PRMT5	[282]
CKM	CDK8, MED12, MED13	P24863, P24863, Q9BWU1	[86, EBI-9817251,EBI-9844402]
CLOCK-BMAL1	ARNTL, CLOCK		[86, EBI-9696921]
COMPASS-like MLL1,2	ASH2L, DPY30, HCFC1 (1), HCFC1 (2), KAT8, KMT2A, KMT2B, MEN1, RBBP5, WDR5		[39, COMPASS-like MLL1,2][283]
COMPASS-like MLL3,4	ASH2L, DPY30, KDM6A, KMT2C, KMT2D, NCOA6, PAGR1, PAXIP1, RBBP5, WDR5		[39, COMPASS-like MLL3,4][283]

Continued on next page

Complex	Members	Members not detected in SNAP	Sources
COMPASS	ASH2L, CXXC1, DPY30, HCFC1 (1), HCFC1 (2), RBBP5, SETD1A, SETD1B, WDR5, WDR82		[39, COMPASS][283]
CRD-mediated mRNA stability complex	DHX9, HNRNPU, IGF2BP1, SYNCRIP, YBX1		[86, EBI-13949164]
CREST-BRG1	HDAC1, SMARCA4 (1), SMARCA4 (2), SP1	CREST_HUMAN, RB_HUMAN	[39, CREST-BRG1]
CRL4-DDB2	CUL4A, DDB1, DDB2, RBX1	Q13620	[86, EBI-11824157,EBI-12509908]
CSE1L-RAN-KPNA2	KPNA2, RAN	CSE1L	[284]
Calprotectin/iNOS-S100A8/A9	S100A8, S100A9	P35228	[86, EBI-10098135,EBI-10098681,EBI-10105915]
Centralspindlin	KIF23, RACGAP1		[285]
Chromosomal passenger	BIRC5, CDCA8, INCENP	Q96GD4	[86, EBI-10729773]
Cohesin loading complex	MAU2, NIPBL		[286]
Cohesin	RAD21, SMC1A, SMC3, STAG2	STAG1, STAG3	[255]
Condensin II	NCAPD3, NCAPH2, SMC2, SMC4	Q86XI2	[86, EBI-12738826]
Condensin I	SMC2, SMC4	Q15003, Q9BPX3, Q15021	[86, EBI-12737208]
CtBP	CTBP2, EHMT1, EHMT2, HDAC1, HDAC2, KDM1A, RCOR1 (1), RCOR1 (2), RREB1	CtBP1, HPC2, KIAA0222, ZNF217	[212]
Cyclin L2-CDK11B	CCNL2, CDK11B		[86, EBI-11580824,EBI-11581656]
DBB1-DBB2	DDB1, DDB2		[86, EBI-11421235]
DSIF complex	SUPT4H1, SUPT5H		[280, id:2860]
E2F1	E2F1, TFDP1	P06400	[86, EBI-10895433,EBI-6879140]
EMSY	EMSY, GATAD1, HDAC1, HDAC2, KDM5A, MORF4L1, MORF4L2, PHF12, RBBP4, RBBP7, SIN3B		[213][287]
Exon junction	EIF4A3, MAGOHB, RBM8A	O15234, P61326, O15234	[86, EBI-12513973,EBI-12514046,EBI-6551702]
FACT	SSRP1, SUPT16H		[39, FACT][86, EBI-11702997]
FANCM-MHF	CENPS/CENPS-CORT, CENPX, FANCM		[215]
Fanconi Anemia	CENPS/CENPS-CORT, CENPX, FAAP24, FANCA, FANCM	FAAP100, FAAP20, FANCB, FANCC, FANCE, FANCF, FANCG, FANCL, HES1	[288][289][215]
G9a/GLP	EHMT1, EHMT2		[290]
HADHA/B	HADHA, HADHB		[291]

Continued on next page

Complex	Members	Members not detected in SNAP	Sources
HBO1	ING4, ING5, JADE2, JADE3, KAT7 (1), KAT7 (2), MEAF6	JADE1_HUMAN, Q6IE81, Q6IE81	[39, HBO1][86, EBI-12519508,EBI-12519687,EBI-12519720,EBI-12519753,EBI-12519786,EBI-12519819]
HDAC-ELMSAN1-DNTTIP1	DNTTIP1, ELMSAN1, HDAC1, HDAC2		[292]
HIF1 heterodimer	ARNT, HIF1A		[293]
HUCA	CABIN1, HIRA, UBN1	ASF1A, H3.3, H4	[294]
HUSH	FAM208A, MPHOSPH8, PPHLN1 (1), PPHLN1 (2), PPHLN1 (3)		[216]
IGF2BP1-mRNP	DHX9, HNRNPU, IGF2BP1, SYNCRIP, YBX1		[295]
INO80	ACTL6A, ACTR5, ACTR8, INO80, INO80B, INO80C, INO80D, INO80E, MCRS1, NFRKB, RUVBL1, RUVBL2, TFPT, UCHL5, YY1		[39, Ino80][86, EBI-12560114]
Importin	KPNA2, KPNA3, KPNA4, KPNB1		[86, EBI-13645294,EBI-13941565,EBI-13941962]
Integrator	INTS1, INTS10, INTS11, INTS12, INTS13, INTS14, INTS2, INTS3, INTS4, INTS5, INTS6, INTS7, INTS8, INTS9		[217]
KAP1-ATRX	ATRX, TRIM28	SETDB1, ZNF274	[296]
KAP1-HP1	CBX3, CBX5, TRIM28	CBX1, KRAB-ZNF	[297]
KAP1-SETDB1-DNMT1-ZNF304	DNMT1 (1), DNMT1 (2), TRIM28	SETDB1, ZNF304	[298]
Ku70:Ku80	XRCC5, XRCC6		[86, EBI-761759]
L3MBTL1	CBX3, L3MBTL1	RB_HUMAN, P06400	[39, L3MBTL1][86, EBI-11790939]
LSD-CoREST	HDAC1, HDAC2, HMG20A, HMG20B, PHF21A, RCOR1 (2), RCOR3, ZMYM2	CTBP1_HUMAN, GSE1_HUMAN, ZN217_HUMAN, ZN516_HUMAN	[39, LSD-CoREST]
MICOS	CHCHD3, IMMT	APOO, APOOL, CHCHD6, MIC13, MINOS1	[280, id:6255]
MLL supercomplex	ASH2L, C17orf49, CHD8, DPY30, E2F6, HCFC1 (1), HCFC1 (2), HCFC2, HSPA1A/HSPA1B, INO80C, KANSL1, KAT8, KDM6A, KMT2A, KMT2B, KMT2C, KMT2D, LAS1L, MAX (1), MAX (2), MCRS1, MEN1, MGA (1), MGA (2), NCOA6, PAGR1, PAXIP1, PELP1, PHF20, PRPF31, RBBP5, RNF2, RUVBL1, RUVBL2, SENP3, TAF1, TAF4, TAF6, TAF7, TAF9, TEX10, WDR5		[39, CHD8, MLL2/3, MLL4/WBP7]
MLL-HCF	ASH2L, DPY30, HCFC1 (1), HCFC1 (2), HCFC2, KMT2A, MEN1, RBBP5, WDR5		[39, MLL-HCF]
MLL5-OGT-USP7	KMT2E, OGT, USP7		[299]

Continued on next page

Complex	Members	Members not detected in SNAP	Sources
MOZ/MORF	BRD1, BRPF1, BRPF3, ING5, KAT6A, MEAF6	Q8WYB5, Q8WYB5, Q8WYB5, KAT6B_HUMAN	[39, MOZ/MORF][86, EBI-12520181,EBI-12521054,EBI-12521131,EBI-12521415,EBI-12521562,EBI-12521618]
MRN	MRE11, NBN, RAD50		[221]
MSL	KAT8, MSL1, MSL2, MSL3		[86, EBI-12556981]
MUS81-EME1	EME1, MUS81		[86, EBI-11893062]
Mad-Max	MAX (1), MXD1		[86, EBI-10713281]
MeCP1	GATAD2B, HDAC1, HDAC2, MBD2, RBBP4, RBBP7		[300][39, MeCP1]
Mediator	MED1, MED10, MED11, MED14, MED15, MED17, MED19, MED20, MED21, MED22, MED26, MED29, MED30, MED31, MED4, MED6, MED7, MED8, MED9	Q6P2C8, Q9ULK4, O75448, Q9Y2X0, Q9BUE0, Q9H204, Q71SY5	[86, EBI-9696567]
Menin-JUND	JUND, MEN1		[86, EBI-11891862]
Menin-associated-HMT	ASH2L, DPY30, KMT2B, MEN1, RBBP5, WDR5		[39, Menin-associated_HMT]
Mitochondrial respiratory chain complex III	UQCRH, UQCRQ	P47985, P47985, P22695, P08574, P00156, P14927, O14957, P31930, Q9UDW1	[86, EBI-1224506]
Mitochondrial respiratory chain complex I	NDUFA2, NDUFA5, NDUFB10, NDUFS1, NDUFS6, NDUFS8	Q86Y39, Q9UI09, O95139, P03915, O95182, O75438, O95299, O43677, P56556, O14561, P03905, O43674, P03897, Q16795, O95178, P17568, P03886, O43676, O43181, Q9Y6M9, P19404, O75251, P03891, P49821, Q9NX14, O95169, O95167, O95298, P03901, O75489, O75306, O95168, O15239, P51970, P03923, O43920, Q9P0J0, P56181	[86, EBI-1245484]
MutSalpha/beta	MSH2, MSH3, MSH6		[86, EBI-10690880,EBI-10691103]
Myc-Max	MAX (1), MYC		[86, EBI-10705742]
N-CoR1	GPS2, HDAC3, NCOR1, SF3A1, SF3B3, SMARCA4 (1), SMARCA4 (2), SMARCB1, SMARCC1, SMARCC2 (1), SMARCC2 (2), SRCAP, TBL1XR1, TRIM28		[280, id:1413][301]
N-CoR2	HDAC1, HDAC2, HDAC3, NCOR1, NCOR2, SAP30, SIN3A		[280, id:1505]
NELF	NELFA, NELFE	NELFB, NELFCD	[280, id:281]
NSL	HCFC1 (1), HCFC1 (2), KANSL1, KANSL2, KANSL3, KAT8, MCRS1, OGT, PHF20, WDR5		[39, NSL][86, EBI-12555779]
NoRC	BAZ2A, SMARCA5		[39, NoRC][86, EBI-11739089]

Continued on next page

Complex	Members	Members not detected in SNAP	Sources
NuA4-related	ACTB, ACTL6A, BRD8 (1), BRD8 (2), DMAP1, EP400 (1), EP400 (2), EPC1, RUVBL1, RUVBL2, SRCAP, TRRAP, VPS72, YEATS4		[39, NuA4-related complex]
NuA4	ACTB, ACTL6A, BRD8 (1), BRD8 (2), DMAP1, EP400 (1), EP400 (2), EPC1, EPC2, ING3, KAT5, MBTD1, MEAF6, MORF4L1, MORF4L2, MRGBP, RUVBL1, RUVBL2, TRRAP, VPS72, YEATS4		[39, NuA4][86, EBI-12737161]
NuRD	CHD3, CHD4, GATAD2A, GATAD2B, HDAC1, HDAC2, KDM1A, MBD2, MBD3, MTA1, MTA2, MTA3, RBBP4, RBBP7		[39, NuRD][86, EBI-12596290,EBI-12676780]
NuRF	BPTF (1), BPTF (2), RBBP4, RBBP7	SMCA1_HUMAN, P28370	[39, NuRF][86, EBI-12514703]
Nuclear cap-binding complex	NCBP1, NCBP2		[86, EBI-16686827]
ORC	ORC1, ORC2, ORC3, ORC4, ORC5	Q9Y5N6	[86, EBI-2889551]
OST	DDOST, RPN1, RPN2, STT3A, STT3B	DAD1, DC2, KCP2, OST4	[302]
ptefb	CCNT1, CDK9		[86, EBI-11164903]
PBAF	ACTB, ACTL6A, ARID1A, ARID1B, ARID2, BRD7, PBRM1, PHF10, SMARCA4 (1), SMARCA4 (2), SMARCB1, SMARCC1, SMARCC2 (1), SMARCC2 (2), SMARCD1, SMARCD2, SMARCD3, SMARCE1	ACL6B_HUMAN, O94805	[39, PBAF][86, EBI-16362609,EBI-16362862]
PCAF	SUPT3H, TADA2A, TADA3, TAF10, TAF12, TAF5L, TAF6L, TAF9, TRRAP	Q92831	[39, PCAF][86, EBI-12741250]
PR-DUB (extended)	ASXL2, BAP1, FOXK1, FOXK2, KDM1B	ASXL1*, MBD5*, MBD6*	[303]
PRC1	CBX2, CBX4, CBX8, COMMD3-BMI1/BMI1, PCGF1, PCGF2, PCGF6, PHC1, PHC2, PHC3, RING1, RNF2, SCML2	CBX6_HUMAN, CBX7_HUMAN, PCGF3_HUMAN, PCGF5_HUMAN, SCMH1_HUMAN	[39, PRC1]
PRC2	AEBP2, EED, EZH1, EZH2, JARID2, MTF2 (1), PHF1, PHF19, RBBP4, RBBP7, SUZ12		[39, PRC2]
PSAP	PNN, RNPS1, SAP18		[276]
PTW/PP1	PPP1CA, PPP1CB, PPP1CC, PPP1R10, TOX4, WDR82		[304]
Paf1C	CDC73, CTR9, LEO1, PAF1, WDR61	Rtf1	[305]
Piccolo NuA4	EPC1, ING3, KAT5		[39, Piccolo_NuA4][86, EBI-12517701]
RFC	RFC1, RFC2, RFC3, RFC4, RFC5		[86, EBI-1170110]
RING2-FBRS	RING1, RNF2, RYBP, YAF2	CSK21_HUMAN, FBRS_HUMAN, FBSL_HUMAN, PCGF3_HUMAN, PCGF5_HUMAN	[39, RING2-FBRS]
RING2-L3MBTL2	CBX3, E2F6, HDAC1, HDAC2, L3MBTL2, MGA (1), MGA (2), PCGF6, RING1, RNF2, RYBP, TFDP1, WDR5, YAF2		[39, RING2-L3MBTL2]
RMI/BLM	BLM, RMI1, RMI2, TOP3A		[153]
RNA Pol III	POLR1C, POLR2E, POLR2F, POLR2H, POLR2K, POLR2L, POLR3D, POLR3E/LOC101060521	AC19, C11, C128, C160, C17, C25, C31, C34, C82	[306]
RNA Pol II	POLR2A, POLR2B, POLR2C, POLR2D, POLR2E, POLR2F, POLR2H, POLR2I, POLR2J/POLR2J[2,3], POLR2K, POLR2L		[306]
RNA Pol I	CD3EAP, POLR1A, POLR1B, POLR1C, POLR1E (1), POLR1E (2), POLR2E, POLR2F, POLR2H, POLR2K, POLR2L, TWISTNB, ZNRD1	A14, AC19	[306]

Continued on next page

Complex	Members	Members not detected in SNAP	Sources
RNA exosome	EXOSC1, EXOSC10, EXOSC2 (1), EXOSC2 (2), EXOSC3, EXOSC4, EXOSC5, EXOSC6, EXOSC8, EXOSC9	Q8TF46, Q15024, Q8TF46, Q15024, Q9Y2L1, Q15024, Q9Y2L1, Q15024, Q15024, EXOS7_HUMAN	[39, RNA exosome][86, EBI-11821717,EBI-12499861,EBI-12499987,EBI-12500122,EBI-12501647]
RPA	RPA1, RPA2, RPA3		[307]
RSF	RSF1, SMARCA5		[39, RSF][86, EBI-11786636]
SAGA	ATXN7, ATXN7L3, ENY2, KAT2A, SGF29, SUPT20H, SUPT3H, SUPT7L, TADA1, TADA2B, TADA3, TAF10, TAF12, TAF2 (1), TAF5L, TAF6L, TAF8, TAF9, TRRAP	UBP22_HUMAN, Q9UPT9	[39, SAGA,STAGA][86, EBI-12599231]
SCF	CUL1, FBXL6, FBXO11, FBXO28, FBXO38, RBX1, SKP1		[308][280][85]
SCL	KDM1A, RCOR1 (2), SFMBT1		[39, SCL]
SLF1/2-RAD18	RAD18, SLF1, SLF2		[229]
SLX4-TERF2	SLX4, TERF2		[86, EBI-11890315]
SMAD2-4	SMAD4, SMAD[2,3,9]		[86, EBI-10021614,EBI-9692430,EBI-9826300]
SNARP	BCLAF1 (1), BCLAF1 (2), PNN, SCD, SNIP1, THRAP3		[309]
SOSS1-2	INIP, INTS3	Q9BQ15, Q96AH0	[86, EBI-11880520,EBI-12505202]
SRCAP	ACTL6A, ACTR6, DMAP1, RUVBL1, RUVBL2, SRCAP, VPS72, YEATS4, ZNHIT1		[39, SRCAP][86, EBI-12735902]
SUMO1	SUMO1, UBE2I		[86, EBI-9347978]
SUMO2	SUMO2, UBE2I		[86, EBI-9350155]
SWI/SNF BRG1/BRM	ACTB, ACTL6A, ARID1A, ARID1B, BRD7, BRD9, DPF2, HDAC1, HDAC2, PBRM1, PHF10, RBBP4, SIN3A, SMARCA2, SMARCA4 (1), SMARCA4 (2), SMARCB1, SMARCC1, SMARCC2 (1), SMARCC2 (2), SMARCD1, SMARCD2, SMARCD3, SMARCE1	ACL6B_HUMAN, DPF1_HUMAN, DPF3_HUMAN, ACL6B_HUMAN, ACL6B_HUMAN, ACL6B_HUMAN	[39, SWI/SNF BRM-BRG1, SWI/SNF_Brg1(I), SWI/SNF_Brg1(II), SWI/SNF_Brm]
SWI/SNF-like EPAFA/B	ACTB, ACTL6A, ARID1B, MLLT1, SMARCA4 (1), SMARCA4 (2), SMARCB1, SMARCC1, SMARCC2 (1), SMARCC2 (2), SMARCD1, SMARCD2, SMARCE1		[39, SWI/SNF-like EPAFB,SWI/SNF-like_EPAFa]
SWR	BRD8 (1), BRD8 (2), EP400 (1), EP400 (2), ING3, KAT5, RUVBL1, RUVBL2, TRRAP	AN32E_HUMAN	[39, SWR]
Shelterin	ACD, POT1, TERF1, TERF2, TERF2IP, TINF2 (1)		[86, EBI-10887677]

Continued on next page

Complex	Members	Members not detected in SNAP	Sources
Spliceosome	ACIN1 (1), ACIN1 (2), ACIN1 (3), ACIN1 (4), ALYREF (1), ALYREF (2), AQR, BCAS2, BUD13, BUD31, CACTIN, CD2BP2, CDC40, CDC5L, CDK12, CIRBP, CRNKL1, CTNNBL1, CWC15, CWC22, DDX17 (1), DDX17 (2), DDX23, DDX39B, DDX3X (1), DDX3X (2), DDX41, DDX46 (1), DDX46 (2), DDX5, DHX15, DHX16, DHX38, DHX8, DHX9, DNAJC8, EFTUD2, EIF4A3, ESS2, HTATSF1, IGF2BP3, IK, ISY1/ISY1-RAB43, LSM2, LSM3, LSM4, LSM6, LSM7, LUC7L3, MFAP1, MTREX, NCBP1, NCBP2, NUDT21, PABPC1, PAXBP1, PLRG1, PPIE, PPIH, PPIL1, PPIL3, PPWD1, PRPF19, PRPF3, PRPF31, PRPF4, PRPF40A, PRPF4B, PRPF6, PRPF8, PUF60, RBM15, RBM17, RBM22, RBM25, RBM39 (1), RBM39 (2), RBM8A, RNPS1, SART1, SF1 (1), SF1 (2), SF1 (3), SF3A1, SF3A2, SF3A3, SF3B1, SF3B2, SF3B3, SF3B4, SF3B6, SLU7, SMNDC1, SMU1, SNRNP200, SNRNP27, SNRNP40, SNRNP70, SNRPA, SNRPA1, SNRPB2, SNRPC, SNRPD1, SNRPD2, SNRPD3, SNRPE, SNRPF, SNRPG, SNU13, SNURF/SNRPB/SNRPN, SNW1, SPEN, SRRM1, SRRM2, SRRT, SRSF1, SRSF11, SRSF2, SRSF3, SRSF4, SRSF5, SRSF6, SRSF7, SRSF9, SYF2, TCERG1, TFIP11, THOC1, THOC2, THOC3, THOC5, THOC6, THOC7, TRA2B, U2AF1/U2AF1L5, U2AF2, U2SURP, USP39, VIRMA, WBP11, WTAP, XAB2, ZNF207	C19orf43, CFAP20, CPSF6, FNBP3, MAGOH, MOV10, PPIL2, PPM1G, PRPF18, TXNL4A	[280, id:351]
Splicing-dependent exon junction complex	ALYREF (1), ALYREF (2), DEK, RBM8A, RNPS1, SRRM1		[310]
TFIID	TAF1, TAF10, TAF11, TAF12, TAF13, TAF2 (1), TAF3, TAF4, TAF4B, TAF5, TAF6, TAF7, TAF8, TAF9, TBP/TBPL2		[86, EBI-12654601,EBI-12684048]
TFTC-HAT	ATXN7, ATXN7L3, ENY2, KAT2A, SUPT3H, SUPT7L, TADA2B, TADA3, TAF10, TAF12, TAF2 (1), TAF2 (2), TAF4, TAF5, TAF5L, TAF6, TAF6L, TAF7, TAF9, TAF9B, TRRAP	Q9UPT9	[39, TFTC-HAT][86, EBI-12600215]
THO	THOC1, THOC2, THOC5, THOC6, THOC7		[280, id:774]
TREX	ALYREF (1), ALYREF (2), DDX39B, THOC1, THOC2, THOC3, THOC5, THOC6, THOC7		[280, id:770]
Telomerase holoenzyme	DKC1, GAR1, NHP2, NOP10	URS00004A7003_9606, O14746	[86, EBI-11175861]
USF1-USF2	USF1, USF2		[86, EBI-9518693]
VCB	ELOB, ELOC, VHL		[311]
VHL SCF-like	ELOB, ELOC, RBX1, VHL	CUL2*	[312]
WICH	BAZ1B, SMARCA5		[86, EBI-12523587]
XPC	CETN2, RAD23A, RAD23B, XPC		[313][314]
YAP1-TEAD1	TEAD[1,2,4], YAP1		[86, EBI-11173119]
ZAP3	ILF2, ILF3, KHDRBS1, NCOA5, RBMX, YLPM1 (1), YLPM1 (2), YLPM1 (3)		[315]
core HDAC	HDAC1, HDAC2, RBBP4, RBBP7		[39, core HDAC]
mCRD-poly(A)-bridging complex	HNRNPD, PABPC1, SYNCRIP	O75534, Q9H074	[86, EBI-13948957]
mSin3A-like	HDAC1, HDAC2, ING1, ING2, RBBP4, RBBP7, SAP30, SIN3A		[316][317][39, mSin3A-like complex]
mSin3A	ARID4A, ARID4B, BRMS1, BRMS1L, HDAC1, HDAC2, RBBP4, RBBP7, SAP18, SAP30, SIN3A, SIN3B, SINHCAF, SUDS3		[39, mSin3A][318]

Appendix D

Permissions to reproduce figures

ELSEVIER LICENSE
TERMS AND CONDITIONS

Sep 30, 2018

This Agreement between Saulius Lukauskas ("You") and Elsevier ("Elsevier") consists of your license details and the terms and conditions provided by Elsevier and Copyright Clearance Center.

License Number	4430771409836
License date	Sep 16, 2018
Licensed Content Publisher	Elsevier
Licensed Content Publication	Cell
Licensed Content Title	The Role of Chromosome Domains in Shaping the Functional Genome
Licensed Content Author	Tom Sexton,Giacomo Cavalli
Licensed Content Date	Mar 12, 2015
Licensed Content Volume	160
Licensed Content Issue	6
Licensed Content Pages	11
Start Page	1049
End Page	1059
Type of Use	reuse in a thesis/dissertation
Portion	figures/tables/illustrations
Number of figures/tables/illustrations	1
Format	both print and electronic
Are you the author of this Elsevier article?	No
Will you be translating?	No
Original figure numbers	Figure 1B
Title of your thesis/dissertation	Mining of Next Generation Sequencing (NGS) and Affinity-Purification Mass Spectrometry (AP-MS) data to identify chromatin targets of epigenetic regulators
Expected completion date	Oct 2018
Estimated size (number of pages)	160
Requestor Location	Saulius Lukauskas 432M Bone Building Imperial College London, SW7 2AZ United Kingdom Attn: Saulius Lukauskas
Publisher Tax ID	GB 494 6272 12
Total	0.00 GBP

ELSEVIER LICENSE
TERMS AND CONDITIONS

Sep 30, 2018

This Agreement between Saulius Lukauskas ("You") and Elsevier ("Elsevier") consists of your license details and the terms and conditions provided by Elsevier and Copyright Clearance Center.

License Number	4431290981296
License date	Sep 17, 2018
Licensed Content Publisher	Elsevier
Licensed Content Publication	Cell
Licensed Content Title	Nucleosome-Interacting Proteins Regulated by DNA and Histone Methylation
Licensed Content Author	Till Bartke,Michiel Vermeulen,Blerta Xhemalce,Samuel C. Robson,Matthias Mann,Tony Kouzarides
Licensed Content Date	Oct 29, 2010
Licensed Content Volume	143
Licensed Content Issue	3
Licensed Content Pages	15
Start Page	470
End Page	484
Type of Use	reuse in a thesis/dissertation
Portion	figures/tables/illustrations
Number of figures/tables/illustrations	1
Format	both print and electronic
Are you the author of this Elsevier article?	No
Will you be translating?	No
Original figure numbers	Figure 2
Title of your thesis/dissertation	Mining of Next Generation Sequencing (NGS) and Affinity-Purification Mass Spectrometry (AP-MS) data to identify chromatin targets of epigenetic regulators
Expected completion date	Oct 2018
Estimated size (number of pages)	160
Requestor Location	Saulius Lukauskas 432M Bone Building Imperial College London, SW7 2AZ United Kingdom Attn: Saulius Lukauskas
Publisher Tax ID	GB 494 6272 12
Total	0.00 GBP

Terms and Conditions

INTRODUCTION

1. The publisher for this copyrighted material is Elsevier. By clicking "accept" in connection with completing this licensing transaction, you agree that the following terms and conditions apply to this transaction (along with the Billing and Payment terms and conditions established by Copyright Clearance Center, Inc. ("CCC"), at the time that you opened your Rightslink account and that are available at any time at http://myaccount.copyright.com).

GENERAL TERMS

2. Elsevier hereby grants you permission to reproduce the aforementioned material subject to the terms and conditions indicated.
3. Acknowledgement: If any part of the material to be used (for example, figures) has appeared in our publication with credit or acknowledgement to another source, permission must also be sought from that source. If such permission is not obtained then that material may not be included in your publication/copies. Suitable acknowledgement to the source must be made, either as a footnote or in a reference list at the end of your publication, as follows:
"Reprinted from Publication title, Vol /edition number, Author(s), Title of article / title of chapter, Pages No., Copyright (Year), with permission from Elsevier [OR APPLICABLE SOCIETY COPYRIGHT OWNER]." Also Lancet special credit - "Reprinted from The Lancet, Vol. number, Author(s), Title of article, Pages No., Copyright (Year), with permission from Elsevier."
4. Reproduction of this material is confined to the purpose and/or media for which permission is hereby given.
5. Altering/Modifying Material: Not Permitted. However figures and illustrations may be altered/adapted minimally to serve your work. Any other abbreviations, additions, deletions and/or any other alterations shall be made only with prior written authorization of Elsevier Ltd. (Please contact Elsevier at permissions@elsevier.com). No modifications can be made to any Lancet figures/tables and they must be reproduced in full.
6. If the permission fee for the requested use of our material is waived in this instance, please be advised that your future requests for Elsevier materials may attract a fee.
7. Reservation of Rights: Publisher reserves all rights not specifically granted in the combination of (i) the license details provided by you and accepted in the course of this licensing transaction, (ii) these terms and conditions and (iii) CCC's Billing and Payment terms and conditions.
8. License Contingent Upon Payment: While you may exercise the rights licensed immediately upon issuance of the license at the end of the licensing process for the transaction, provided that you have disclosed complete and accurate details of your proposed use, no license is finally effective unless and until full payment is received from you (either by publisher or by CCC) as provided in CCC's Billing and Payment terms and conditions. If full payment is not received on a timely basis, then any license preliminarily granted shall be deemed automatically revoked and shall be void as if never granted. Further, in the event that you breach any of these terms and conditions or any of CCC's Billing and Payment terms and conditions, the license is automatically revoked and shall be void as if never granted. Use of materials as described in a revoked license, as well as any use of the materials beyond the scope of an unrevoked license, may constitute copyright infringement and publisher reserves the right to take any and all action to protect its copyright in the materials.
9. Warranties: Publisher makes no representations or warranties with respect to the licensed material.
10. Indemnity: You hereby indemnify and agree to hold harmless publisher and CCC, and their respective officers, directors, employees and agents, from and against any and all claims arising out of your use of the licensed material other than as specifically authorized pursuant to this license.
11. No Transfer of License: This license is personal to you and may not be sublicensed, assigned, or transferred by you to any other person without publisher's written permission.
12. No Amendment Except in Writing: This license may not be amended except in a writing signed by both parties (or, in the case of publisher, by CCC on publisher's behalf).
13. Objection to Contrary Terms: Publisher hereby objects to any terms contained in any purchase order, acknowledgment, check endorsement or other writing prepared by you, which terms are inconsistent with these terms and conditions or CCC's Billing and Payment terms and conditions. These terms and conditions, together with CCC's Billing and Payment terms and conditions (which are incorporated herein), comprise the entire agreement between you and publisher (and CCC) concerning this licensing transaction. In the event of any conflict between your obligations established by these terms and conditions and those established by CCC's Billing and Payment terms and conditions, these terms and conditions shall control.
14. Revocation: Elsevier or Copyright Clearance Center may deny the permissions described in this License at their sole discretion, for any reason or no reason, with a full refund payable to you. Notice of such denial will be made using the contact information provided by you. Failure to receive such notice will not alter or invalidate the denial. In no event will Elsevier or Copyright Clearance Center be responsible or liable for any costs, expenses or damage incurred by you as a result of a denial of your permission request, other than a refund of the amount(s) paid by you to Elsevier and/or Copyright Clearance Center for denied permissions.

LIMITED LICENSE

The following terms and conditions apply only to specific license types:

15. **Translation**: This permission is granted for non-exclusive world **English** rights only unless your license was granted for translation rights. If you licensed translation rights you may only translate this content into the languages you requested. A professional translator must perform all translations and reproduce the content word for word preserving the integrity of the article.
16. **Posting licensed content on any Website**: The following terms and conditions apply as follows: Licensing material from an Elsevier journal: All content posted to the web site must maintain the copyright information line on the bottom of each image; A hyper-text must be included to the Homepage of the journal from which you are licensing at http://www.sciencedirect.com/science/journal/xxxxx or the Elsevier homepage for books at http://www.elsevier.com; Central Storage: This license does not include permission for a scanned version of the material to be stored in a central repository such as that provided by Heron/XanEdu.
Licensing material from an Elsevier book: A hyper-text link must be included to the Elsevier homepage at http://www.elsevier.com . All content posted to the web site must maintain the copyright information line on the bottom of each image.

Posting licensed content on Electronic reserve: In addition to the above the following clauses are applicable: The web site must be password-protected and made available only to bona fide students registered on a relevant course. This permission is granted for 1 year only. You may obtain a new license for future website posting.
17. **For journal authors:** the following clauses are applicable in addition to the above:
Preprints:
A preprint is an author's own write-up of research results and analysis, it has not been peer-reviewed, nor has it had any other value added to it by a publisher (such as formatting, copyright, technical enhancement etc.).
Authors can share their preprints anywhere at any time. Preprints should not be added to or enhanced in any way in order to appear more like, or to substitute for, the final versions of articles however authors can update their preprints on arXiv or RePEc with their Accepted Author Manuscript (see below).
If accepted for publication, we encourage authors to link from the preprint to their formal publication via its DOI. Millions of researchers have access to the formal publications on ScienceDirect, and so links will help users to find, access, cite and use the best available version. Please note that Cell Press, The Lancet and some society-owned have different preprint policies. Information on these policies is available on the journal homepage.
Accepted Author Manuscripts: An accepted author manuscript is the manuscript of an article that has been accepted for publication and which typically includes author-incorporated changes suggested during submission, peer review and editor-author communications.
Authors can share their accepted author manuscript:

- immediately
  - via their non-commercial person homepage or blog
  - by updating a preprint in arXiv or RePEc with the accepted manuscript
  - via their research institute or institutional repository for internal institutional uses or as part of an invitation-only research collaboration work-group
  - directly by providing copies to their students or to research collaborators for their personal use
  - for private scholarly sharing as part of an invitation-only work group on commercial sites with which Elsevier has an agreement
- After the embargo period
  - via non-commercial hosting platforms such as their institutional repository
  - via commercial sites with which Elsevier has an agreement

In all cases accepted manuscripts should:

- link to the formal publication via its DOI
- bear a CC-BY-NC-ND license - this is easy to do
- if aggregated with other manuscripts, for example in a repository or other site, be shared in alignment with our hosting policy not be added to or enhanced in any way to appear more like, or to substitute for, the published journal article.

Published journal article (JPA): A published journal article (PJA) is the definitive final record of published research that appears or will appear in the journal and embodies all value-adding publishing activities including peer review co-ordination, copy-editing, formatting, (if relevant) pagination and online enrichment.
Policies for sharing publishing journal articles differ for subscription and gold open access articles:
Subscription Articles: If you are an author, please share a link to your article rather than the full-text. Millions of researchers have access to the formal publications on ScienceDirect, and so links will help your users to find, access, cite, and use the best available version.
Theses and dissertations which contain embedded PJAs as part of the formal submission can be posted publicly by the awarding institution with DOI links back to the formal publications on ScienceDirect.
If you are affiliated with a library that subscribes to ScienceDirect you have additional private sharing rights for others' research accessed under that agreement. This includes use for classroom teaching and internal training at the institution (including use in course packs and courseware programs), and inclusion of the article for grant funding purposes.

Gold Open Access Articles: May be shared according to the author-selected end-user license and should contain a CrossMark logo, the end user license, and a DOI link to the formal publication on ScienceDirect.
Please refer to Elsevier's posting policy for further information.
18. **For book authors** the following clauses are applicable in addition to the above: Authors are permitted to place a brief summary of their work online only. You are not allowed to download and post the published electronic version of your chapter, nor may you scan the printed edition to create an electronic version. **Posting to a repository:** Authors are permitted to post a summary of their chapter only in their institution's repository.
19. **Thesis/Dissertation**: If your license is for use in a thesis/dissertation your thesis may be submitted to your institution in either print or electronic form. Should your thesis be published commercially, please reapply for permission. These requirements include permission for the Library and Archives of Canada to supply single copies, on demand, of the complete thesis and include permission for Proquest/UMI to supply single copies, on demand, of the complete thesis. Should your thesis be published commercially, please reapply for permission. Theses and dissertations which contain embedded PJAs as part of the formal submission can be posted publicly by the awarding institution with DOI links back to the formal publications on ScienceDirect.

Elsevier Open Access Terms and Conditions
You can publish open access with Elsevier in hundreds of open access journals or in nearly 2000 established subscription journals that support open access publishing. Permitted third party re-use of these open access articles is defined by the author's choice of Creative Commons user license. See our open access license policy for more information.
Terms & Conditions applicable to all Open Access articles published with Elsevier:
Any reuse of the article must not represent the author as endorsing the adaptation of the article nor should the article be modified in such a way as to damage the author's honour or reputation. If any changes have been made, such changes must be clearly indicated.
The author(s) must be appropriately credited and we ask that you include the end user license and a DOI link to the formal publication on ScienceDirect.
If any part of the material to be used (for example, figures) has appeared in our publication with credit or acknowledgement to another source it is the responsibility of the user to ensure their reuse complies with the terms and conditions determined by the rights holder.
Additional Terms & Conditions applicable to each Creative Commons user license:
CC BY: The CC-BY license allows users to copy, to create extracts, abstracts and new works from the Article, to alter and revise the Article and to make commercial use of the Article (including reuse and/or resale of the Article by commercial entities), provided the user gives appropriate credit (with a link to the formal publication through the relevant DOI), provides a link to the license, indicates if changes were made and the licensor is not represented as endorsing the use made of the work. The full details of the license are available at http://creativecommons.org/licenses/by/4.0.
CC BY NC SA: The CC BY-NC-SA license allows users to copy, to create extracts, abstracts and new works from the Article, to alter and revise the Article, provided this is not done for commercial purposes, and that the user gives appropriate credit (with a link to the formal publication through the relevant DOI), provides a link to the license, indicates if changes were made and the licensor is not represented as endorsing the use made of the work. Further, any new works must be made available on the same conditions. The full details of the license are available at http://creativecommons.org/licenses/by-nc-sa/4.0.
CC BY NC ND: The CC BY-NC-ND license allows users to copy and distribute the Article, provided this is not done for commercial purposes and further does not permit distribution of the Article if it is changed or edited in any way, and provided the user gives appropriate credit (with a link to the formal publication through the relevant DOI), provides a link to the license, and that the licensor is not represented as endorsing the use made of the work. The full details of the license are available at http://creativecommons.org/licenses/by-nc-nd/4.0.
Any commercial reuse of Open Access articles published with a CC BY NC SA or CC BY NC ND license requires permission from Elsevier and will be subject to a fee.
Commercial reuse includes:

- Associating advertising with the full text of the Article
- Charging fees for document delivery or access
- Article aggregation
- Systematic distribution via e-mail lists or share buttons

Posting or linking by commercial companies for use by customers of those companies.

20. **Other Conditions**:

v1.9

SPRINGER NATURE LICENSE
TERMS AND CONDITIONS

Sep 30, 2018

This Agreement between Saulius Lukauskas ("You") and Springer Nature ("Springer Nature") consists of your license details and the terms and conditions provided by Springer Nature and Copyright Clearance Center.

License Number	4432750184032
License date	Sep 19, 2018
Licensed Content Publisher	Springer Nature
Licensed Content Publication	Experimental & Molecular Medicine
Licensed Content Title	Writing, erasing and reading histone lysine methylations
Licensed Content Author	Kwangbeom Hyun, Jongcheol Jeon, Kihyun Park, Jaehoon Kim
Licensed Content Date	Apr 28, 2017
Licensed Content Volume	49
Licensed Content Issue	4
Type of Use	Thesis/Dissertation
Requestor type	academic/university or research institute
Format	print and electronic
Portion	figures/tables/illustrations
Number of figures/tables/illustrations	1
High-res required	no
Will you be translating?	no
Circulation/distribution	<501
Author of this Springer Nature content	no
Title	Mining of Next Generation Sequencing (NGS) and Affinity-Purification Mass Spectrometry (AP-MS) data to identify chromatin targets of epigenetic regulators
Instructor name	n/a
Institution name	n/a
Expected presentation date	Oct 2018
Portions	Figure 1
Requestor Location	Saulius Lukauskas 432M Bone Building Imperial College London, SW7 2AZ United Kingdom Attn: Saulius Lukauskas
Billing Type	Invoice
Billing Address	Saulius Lukauskas 432M Bone Building Imperial College London, United Kingdom SW7 2AZ Attn: Saulius Lukauskas
Total	0.00 EUR
Terms and Conditions	

Springer Nature Terms and Conditions for RightsLink Permissions

Springer Nature Customer Service Centre GmbH (the Licensor) hereby grants you a non-exclusive, world-wide licence to reproduce the material and for the purpose and requirements specified in the attached copy of your order form, and for no other use, subject to the conditions below:

1. The Licensor warrants that it has, to the best of its knowledge, the rights to license reuse of this material. However, you should ensure that the material you are requesting is original to the Licensor and does not carry the copyright of another entity (as credited in the published version).

   If the credit line on any part of the material you have requested indicates that it was reprinted or adapted with permission from another source, then you should also seek permission from that source to reuse the material.

2. Where **print only** permission has been granted for a fee, separate permission must be obtained for any additional electronic re-use.

3. Permission granted **free of charge** for material in print is also usually granted for any electronic version of that work, provided that the material is incidental to your work as a whole and that the electronic version is essentially equivalent to, or substitutes for, the print version.

4. A licence for 'post on a website' is valid for 12 months from the licence date. This licence does not cover use of full text articles on websites.

5. Where **'reuse in a dissertation/thesis'** has been selected the following terms apply: Print rights of the final author's accepted manuscript (for clarity, NOT the published version) for up to 100 copies, electronic rights for use only on a personal website or institutional repository as defined by the Sherpa guideline (www.sherpa.ac.uk/romeo/).

6. Permission granted for books and journals is granted for the lifetime of the first edition and does not apply to second and subsequent editions (except where the first edition permission was granted free of charge or for signatories to the STM Permissions Guidelines http://www.stm-assoc.org/copyright-legal-affairs/permissions/permissions-guidelines/), and does not apply for editions in other languages unless additional translation rights have been granted separately in the licence.

7. Rights for additional components such as custom editions and derivatives require additional permission and may be subject to an additional fee. Please apply to Journalpermissions@springernature.com/bookpermissions@springernature.com for these rights.

8. The Licensor's permission must be acknowledged next to the licensed material in print. In electronic form, this acknowledgement must be visible at the same time as the figures/tables/illustrations or abstract, and must be hyperlinked to the journal/book's homepage. Our required acknowledgement format is in the Appendix below.

9. Use of the material for incidental promotional use, minor editing privileges (this does not include cropping, adapting, omitting material or any other changes that affect the meaning, intention or moral rights of the author) and copies for the disabled are permitted under this licence.

10. Minor adaptations of single figures (changes of format, colour and style) do not require the Licensor's approval. However, the adaptation should be credited as shown in Appendix below.

Appendix — Acknowledgements:

For Journal Content:
Reprinted by permission from [**the Licensor**]: [**Journal Publisher** (e.g. Nature/Springer/Palgrave)] [**JOURNAL NAME**] [**REFERENCE CITATION**(Article name, Author(s) Name), [**COPYRIGHT**] (year of publication)

For **Advance Online Publication papers:**
Reprinted by permission from [**the Licensor**]: [**Journal Publisher** (e.g. Nature/Springer/Palgrave)] [**JOURNAL NAME**] [**REFERENCE CITATION**(Article name, Author(s) Name), [**COPYRIGHT**] (year of publication), advance online publication, day month year (doi: 10.1038/sj.[JOURNAL ACRONYM].)

For Adaptations/Translations:
Adapted/Translated by permission from [**the Licensor**]: [**Journal Publisher** (e.g. Nature/Springer/Palgrave)] [**JOURNAL NAME**] [**REFERENCE CITATION**(Article name, Author(s) Name), [**COPYRIGHT**] (year of publication)

Note: For any republication from the British Journal of Cancer, the following credit line style applies:

Reprinted/adapted/translated by permission from [**the Licensor**]: on behalf of Cancer Research UK: : [**Journal Publisher** (e.g. Nature/Springer/Palgrave)] [**JOURNAL NAME**] [**REFERENCE CITATION** (Article name, Author(s) Name), [**COPYRIGHT**] (year of publication)

For **Advance Online Publication** papers:
Reprinted by permission from The [**the Licensor**]: on behalf of Cancer Research UK: [**Journal Publisher** (e.g. Nature/Springer/Palgrave)] [**JOURNAL NAME**] [**REFERENCE CITATION** (Article name, Author(s) Name), [**COPYRIGHT**] (year of publication), advance online publication, day month year (doi: 10.1038/sj.[JOURNAL ACRONYM])

For Book content:
Reprinted/adapted by permission from [**the Licensor**]: [**Book Publisher** (e.g. Palgrave Macmillan, Springer etc) [**Book Title**] by [**Book author**(s)] [**COPYRIGHT**] (year of publication)

Other Conditions:

Version 1.1