Dissertation zur Erlangung des Doktorgrades
der Fakultät für Chemie und Pharmazie
der Ludwig-Maximilians-Universität München

Scalable Quantitative Interaction Proteomics of Regulatory DNA Elements

von
Thanatip Viturawong
aus Bangkok, Thailand

2014

Erklärung

Diese Dissertation wurde im Sinne von § 7 der Promotionsordnung vom 28. November 2011 von Herrn Prof. Dr. Matthias Mann betreut.

Eidesstattliche Versicherung

Diese Dissertation wurde eigenständig und ohne unerlaubte Hilfe erarbeitet.

Martinsried, am 15. Mai 2014

.....

Dissertation eingereicht am 15.05.2014

1. Gutachter Prof. Dr. Matthias Mann

2. Gutachter Prof. Achim Tresch, PhD

Mündliche Prüfung am 23.06.2014

Table of Contents

For we know in part, and we prophesy in part.

But when that which is perfect is come,
then that which is in part shall be done away.

When I was a child, I spake as a child, I understood as a child, I thought as a child:
but when I became a man, I put away childish things.

For now we see through a glass, darkly; but then face to face:
now I know in part; but then shall I know even as also I am known.

I Corinthians 13:9–12

To my parents

leaaleatallgggelael
llgalleggtelgllaggee
agggeeelgaaelgegel
llegggeeeleatatelgga
gataaeatgallageegaaglllataeggeal
lleaeegtaaagteeteeteggagtgteeeteeta
llgatgataageeageggetaeeallallgall
aeaeaaggaaeggtgatallaaeatagall
ggeaeallaetellgtaggtgtggaateaell
allaegtggeegaageellatggeaaaaaiegt
gggeaatgalleaggtaatgetaaaagtea
lageaeatacateeea

llgaeeatealleaeget
eatallaaallaatgegte
gataallaegaaagtgga
eaetateeeateaaegtg
llgtleaeaeatgaaeaaallagtagagtge
ellleageeellgtegtegglgatgtgtgtaa
lageglleatatgaateaaetetataaaat
allaetgatgegtaaagagatetgaaatell
allggtetatgteactgaaaetateeaaatat
ellatgtegatactgaaegtallgaegtatae

aaggtggteellaellaat
agaatatagtgteagtta
gtataaeeetaatgagga
atetaatallagtataae
atgallageegaagttataeggeatalteace
gtaaagteeteeteggagtgteeteellallga
lgataageageggetaeeallallgallaaea
eaaggaaeegagtgatallaaeatagalteggg
eaeallaetellgtaggtgtggaateaellaat
laegtggeegaageellatggeaaaaategatg
ggeaatgalteaggtaatgetaaaagteeat
ageaeatacateeeaaegtggtatgegtaea
allgaeeatealleaegetaaaatgetgaee
aeatallaaallaatgegtetgeaetaeteaa

Summary

Protein-DNA interaction is central to the understanding of transcriptional regulation. At present, chromatin immunoprecipitation coupled to parallel sequencing (ChIP-seq) is a widely used and scalable technique to identify target DNA sequences of transcription factors of interest. Stable isotope labeling of amino acid in cell culture (SILAC) has been used with protein affinity purification and high-resolution mass spectrometry (AP-MS) to give a complementary perspective of protein interactions at specific DNA sequences. However, large-scale SILAC AP-MS screens for protein-DNA interactions and their quantitative analysis were limited by issues such as cross-batch comparability and variation within the SILAC duplicates.

The work in this thesis introduced several improvements in the workflow of SILAC AP-MS for interactions of proteins and long (> 200 bp) DNA sequences. Specifically, we implemented high-throughput bait generation based on parallel cloning. In addition, we devised a computational processing procedure capable of de-noising SILAC AP-MS data by automatically identifying and removing batch-wise systematic errors. These treatments relieve the bottlenecks in scalable DNA SILAC AP-MS and allow for high-precision quantitative comparisons across experiments as well as experiment batches.

We applied scalable DNA SILAC AP-MS to study protein interactions to highly conserved non-coding elements. Known to possess tightly spatiotemporal-controlled transcriptional regulatory activity, these elements are thought to serve important biological functions. Their origin of conservation is a topic of great interest; however, experimental data are still needed to test the existing hypotheses. We produced an interactome for 190 ultraconserved elements (UCEs) – the most extremely conserved subset of the highly conserved non-coding elements – using the scalable DNA SILAC AP-MS approach we developed. The interaction profile supports a "multiple binding constraints" hypothesis, wherein overlapping functional transcription factor binding sites give rise to higher evolutionary pressure that keeps each nucleotide conserved. We also generated a scanning differential interactome of an ultraconserved

enhancer at five nucleotide-resolution, where we observed the consequences of mutation on the changes in protein interactions.

We cross-validated our SILAC AP-MS interactome with existing ChIP-seq data for transcription factor and chromatin signatures. We found that the interactions of proteins with our DNA affinity baits, where initial epigenetic priming was absent, nevertheless reflected the cellular epigenetic modifications at the corresponding genomic locus. This analysis, carried out over hundreds of DNA sequence-genomic locus pairs, strongly demonstrated the contribution of genetic information in establishing epigenetic states of UCEs.

In summary, we have used scalable DNA SILAC AP-MS enabled by improvements developed by this work, to produce a functionally cross-validated UCE interactome and shed light on the question of the origin of UCE conservation.

1 Introduction

DNA, RNA and protein are complex and extremely diverse biopolymers that together constitute the molecular building blocks of the cell. The flow of information between these polymers became clear as early as 1958, when Francis Crick, co-discoverer of DNA structure, coined the phrase "Central Dogma of molecular biology" which sums up as "DNA makes RNA makes protein". Protein performs myriad structural, metabolic and regulatory functions inside the cell, whereas DNA is the inheritable genetic material, carrying the information required to produce protein via RNA intermediates.

The steps of gene expressions corresponding to the central dogma have been described and their general mechanisms studied in great detail. In eukaryotic cells, the DNA encoding a protein gene is preceded by a promoter sequence which marks the transcription initiation site. An RNA polymerase complex binds to the promoter and synthesizes pre-messenger RNA (pre-mRNA) based on the DNA template. During and after transcription, the pre-mRNA is modified at both ends. The 5' end receives a methylated guanine "cap", and a stretch of around 200 adenines is added to the 3' end, forming the "poly(A)" tail. The pre-mRNA containing both introns and exons is spliced to the smaller mRNA, where the introns are removed. The mRNA is exported from the nucleus and may be specifically targeted to subcellular locations (or even exported across cells), and is eventually translated by the ribosome.

Gene regulation is the complex and dynamic process which controls production of proteins at the appropriate place and time, and may be divided temporally into transcriptional, post-transcriptional, and post-translational regulation. Conceptually, for regulation to be specific to cellular conditions while ensuring stability, a mechanism that provides both orthogonality and redundancy must be established. Both are achieved by specific physical interactions between the regulating molecule and their target DNA or RNA sequence motifs that are directly coupled to the synthesis or localization process ("*trans*" and "*cis*" elements, respectively). Such *cis* regulatory elements include transcription factor binding sites (TFBS) and mRNA localization motifs. Proteins fulfill the function of *trans* entities in a vast majority of processes, although there is also

growing evidence for RNA fulfilling this role. This thesis develops and uses proteomic technology to elucidate protein-DNA interactions at transcriptional regulatory elements that have been revealed through comparative genomics.

1.1 Protein-DNA interactions in transcription regulation

1.1.1 General interactions in RNA Pol-II mediated transcription

The RNA polymerase II complex is responsible for transcription of mRNA in eukaryotic cells. RNA Pol II-mediated transcription is initiated from RNA Pol II-specific promoters. These promoter sequences are diverse but share common characteristics of the "basal DNA elements", namely: the initiator sequence $\mathbf{Y}_2\mathbf{CAY}_5$ (the **A** nucleotide of which becomes the first base of the mRNA), and the so-called "TATA box" element $\mathbf{TATAAW}_3$ flanked by a GC-rich sequence approximately 25 bp upstream of the initiator sequence. Promoters without the TATA box tend to contain a "downstream promoter element" **AGAC**, located approximately 30 bp from the initiation site.

The RNA Pol II complex consists of several general transcription factors TFII(*X*), which are altogether required for promoter-targeted transcription initiation. First, TFIID subunits, consisting of the TATA-binding proteins (TBP) and TBP-associated factors (TAFs), are required to direct the complex to either the TATA box or the downstream promoter element. The TBP directly complexes with the minor groove of the TATA box, introducing a near-perpendicular bend on the DNA towards the major groove, which in turn brings the transcription factors and RNA Pol II into closer proximity. Then, TFIIA, TFIIB, and TFIIF are recruited to the promoter in the order specified. Subsequent binding of TFIIE and the ATP-dependent DNA-helicase TFIIH then causes the promoter DNA duplex to melt, allowing the coding strand to be used as template for transcription. In summary, TBP is the factor that binds directly to the promoter DNA in a sequence-specific manner; further protein-protein interactions between the general transcription factors and RNA Pol II then contribute to the specific localization of the initiation complex.

1.1.2 Specific protein-DNA interactions in transcription regulation

Although the basal promoter elements are critical in precise determination of the initiation site, they are themselves insufficient for high level transcription *in vivo*. A classical scanning mutagenesis study of the ß-globin promoter demonstrated that physiological levels of transcription *in vivo* require interactions at other, more distal elements in addition to the basal elements [1]. Indeed, most promoter sequences possess multiple copies of such non-basal elements that are targeted specifically by transcription factors known as "activators". Several families of activators exist, many of which having partially redundant DNA sequence specificities. Non-basal elements are also targeted by "repressor" proteins, whose binding impedes transcription initiation directly or indirectly. Depending on cellular context, many transcription factors can act as both repressors and activators, and their nuclear localization and DNA binding capability can be modulated by post-translational modification. As a result, the transcriptional output of any given promoter depends on the relative expression levels of relevant transcription factors and their context-dependent mode of action with respect to the promoter. The multiple copies of non-basal elements targeting the same protein can result in a degree of redundancy, where no single non-basal element is critical to maintain the promoter activity.

In addition to interactions at the promoter, sequences known as enhancers also interact with transcriptional factors through conventional TFBSs. By offsetting the local concentration of interacting transcription factors, enhancers are able to positively or negatively influence transcription initiation of the promoters in physical proximity. Because of higher-order structural organization in the nucleus where different parts of the chromosome are brought together, enhancers can act over great distances along the genomic coordinate. Indeed, some enhancers are known to act on promoters several hundred kilobases away.

The way the information on an enhancer is interpreted can be described in the "enhanceosome" model and the "billboard" model. In the enhanceosome model, transcription factors bind co-operatively on the DNA through protein-protein interactions. This requires that the transcription factor binding sites are present

with strict relative positioning and orientation. The interferon-ß locus is the hallmark example of an enhanceosome, where an array of TFBSs are placed next to each other, reflecting the exact topology of protein-protein interactions [2]. The billboard model, in contrast, proposes that the TFBSs are independently interpreted by the transcription factors, possibly in a multi-step process. The billboard model is supported, for instance, by the observation that a single enhancer can act both as a repressor and an activator in the same nuclear environment [3]. A continuum of mechanisms with characteristics of both the billboard model and the enhanceosome model may be attributed to different enhancers [4].

1.1.3 Biochemistry of specific protein-DNA interaction

Transcription factor binding sites are typically 4-20 nucleotides in length and can be of varying degree of specificity. The first proposed mechanism accounting for sequence specificity in protein-DNA interaction was the "direct readout" model: Base-specific hydrogen bonds and non-polar interactions would be formed between the major groove of the DNA and a series of amino acid side chains that provide the complementary chemical groups for interaction. Although consistent with over a thousand protein-DNA complex structures, there is no one-to-one correspondence between the DNA and a "complementary" protein sequence. Further structures have revealed that sequence specificity is generally achieved through combination of base-specific interaction in both the major and the minor grooves, as well as shape-recognition mechanisms that differentiate between size of DNA grooves or the form of the duplex [5].

The different combinations of mechanisms are reflected by the various structures of DNA binding domains that exist, and to some extent, the corresponding family of transcription factor binding sites. The TRANSFAC database, where binding motifs for over a thousand transcription factors are deposited, classifies their motifs as follows: (1) basic domains, (2) zinc-coordinating domains, (3) helix-turn-helix domains, and (4) beta-scaffold domains [6]. Closely related domains often have sequence specificity with similar characteristics, e.g. the homeodomains bind AT-rich sequences, whereas

the Kruppel like factors contain zinc finger domains that recognize GC-rich sequences.

Transcription factors often complex to DNA cooperatively [7-9]. In the special case of homo-/hetero-dimers, the binding sites often consist of two palindromic or near-palindromic "half-sites" [10]. These half-sites may be separated by a few non-specific nucleotides, as is the case with the family of STAT transcription factors [11]. Physical interaction between transcription factors may also result in modulation of sequence specificity. For example, the Hox family of transcription factors bind to DNA through their highly-conserved homeodomains, which can achieve exquisite specificities upon dimerization [12].

1.1.4 Representation of transcription factor binding motifs

Comparison of promoters and enhancers revealed that TFBSs generally vary in sequence, even for a single transcription factor. The most concise representation of a set of TFBSs is the "consensus sequence", where the known sites are aligned and the most common nucleotide/nucleotide combination is given. For example **TTCYWNDGAA** is a consensus binding sequence for the transcription factor Stat6. While this representation is simple, it cannot quantitatively "score" the fit of a sequence to the motif. A more popular representation of a TFBS is therefore the position weight matrix (PWM), which summarizes the frequency or probability of each position in the TFBS being a specific nucleotide. The PWM allows calculation of the log-likelihood – the sum of log-probability across all positions given the specific sequence – which can be treated as a simple score for how well a sequence fits the ensemble of known TFBSs for a given transcription factor. Another advantage of the PWM is that it allows for the calculation of the "information content", which is related to the degree of degeneracy of the set of TFBSs. As an example, Figure 1 illustrates the PWM of Stat6 as curated in the JASPAR database [13].

Because log-likelihoods are calculated from the product of probabilities for each position, this interpretation of the PWM assumes independence of nucleotide identity between each individual position of the TFBS. This assumption has

A	257	665	152	0	13	122	70	664	328	1153	26	1852	1849	621	345
C	663	328	566	0	0	1680	1204	21	534	3	0	0	0	540	490
G	482	372	436	0	0	0	0	219	898	416	1796	0	3	398	235
T	450	487	698	1852	1839	50	578	948	92	280	30	0	0	293	782
	N	N	N	T	T	C	Y	W	N	D	G	A	A	N	N

Figure 1: Representation of Stat6 binding sites as a position weight matrix with the corresponding consensus sequence.

The nucleotide identities corresponding to the consensus are highlighted. Nonspecific bases are faded in grey, and degenerate bases are highlighted in yellow.

been shown to be invalid for some transcription factors, and representations that take into account base dependency can perform better in assessing a fit of a sequence as a TFBS. Examples of such representations include those employing Hidden Markov Models [14] or simply displaying all non-degenerate motifs [15]. Although more accurate, these models are not easy to visualize and summarize, and the PWM representation remains a popular choice owing to its simplicity and relative usefulness.

1.1.5 Epigenetic mechanisms

DNA *in vivo* is packaged into nucleosomes, which themselves are part of higher-order structural organization of the chromatin. The nucleosome contains a "core particle" that consists of two copies each of histones H2A, H2B, H3 and H4 [16]. A DNA duplex forms a 146 bp left-handed superhelix around the octamer, where basic, amide and hydroxyl groups from the histone proteins form a network of hydrogen bonds with the phosphate backbone of the DNA. Although much of the histone-DNA affinity is derived from non-base specific contacts, nucleosomes do nevertheless possess weak sequence specificity [17]. Although the exact sequence preference of the nucleosome is complex, the most prominent predictor of intrinsic nucleosome affinity of a DNA sequence is its GC content [18]. A nucleosome positioned at a promoter can impede transcription, for example, by burying the TATA box inside the DNA-histone interface and preventing its access by the transcriptional machinery. Nucleotide-resolution re-positioning of a nucleosome around regulatory elements can significantly modulate

transcriptional regulation, through mechanisms that often involve a complex interplay between nucleosomes and other transcription factors [19, 20]. In addition to the core particle, a "linker" subunit histone H1 may also be found on nucleosomes. The presence of histone H1 on nucleosomes results in chromatin compaction [21], and relative proportions of histone H1 are understood to account at least partially for the higher structural organization of the chromatin [22].

A substantial proportion of amino acids in the histone proteins are in the relatively unstructured "tail" domains [23], and are subjected to extensive post-translational modifications including acetylation, phosphorylation, methylation, ubiquitinylation, sumoylation and biotinylation [24-29]. These modifications can enhance or repress transcriptional activation through different mechanisms. For instance, lysine acetylation results in neutralization of positive charges that interact with the DNA phosphate backbone, effectively loosening the electrostatic histone-DNA contact. Modified histone residues generally serve as docking sites for modification-specific protein "readers", which in turn are coupled to or recruit chromatin remodeling complexes. This results in compaction or loosening of the chromatin structure, thus modulating transcription activity via chromatin accessibility, or, alternatively, recruitment of proteins that catalyze further histone modification ("writers") [30, 31]. In this way, readers and writers combinatorially influence the chromatin environment [32].

Other aspects of epigenetic controls include: different turnover rates of histone modifications [33]; spread of modifications and chromatin states into neighboring regions and restriction of this process by insulator sequences [34, 35]; DNA modifications which interact with genetic mechanisms [36] and inheritance of chromatin modification [37]. Furthermore, chromatin itself is organized into chromosome territories where distal parts of the genome are brought into proximity: a mechanism that is exploited by distal enhancers [38, 39].

Importantly, although protein-protein interactions appear to dominate in the epigenetic processes that directly influence transcriptional activity, it has been

recently shown that *in vivo* epigenetic states may be recapitulated by *in vitro* reconstitution of native nuclear lysate to naked DNA [40]. This study, based on flow-cytometric measurements of histone modifications on a regulatory DNA sequence, emphasizes the primordial contribution of underlying DNA sequence to epigenetic mechanisms.

1.2 Highly conserved non-coding DNA sequences

As of 2013, the number of completely sequenced eukaryotic genomes was approaching 200 [41]. This increasing wealth of complete genome sequences has enabled extensive comparison between the genome of different species, revealing DNA elements that are conserved between them. In closely-related species, DNA conservation can be attributed to the small amount of time since divergence between their ancestors. Over a larger evolutionary distance, DNA sequence conservation is generally accepted to implicate a biological function [42]. A conserved DNA element may encode a protein or RNA gene, or it may be designated "non-coding", if no evidence of a corresponding gene product has been detected. Conserved non-coding DNA sequences are thus particularly interesting in the context of gene regulation, as they may be potential regulatory elements.

Many classes of conserved non-coding DNA elements have been tabulated. A family of human sequences known as ultraconserved elements (UCEs) was first described in 2004, under very stringent conservation criteria of 100% mouse-human sequence identity over 200 bp [43]. Other classes of lesser but still statistically significantly conserved elements include ultraconserved regions (95% identity, ≥ 50 bp) [44] and long conserved noncoding elements (significant conservation ≥ 500 bp) [45]. All are known under the umbrella term of highly-conserved non-coding elements (HNCEs). Sets of HNCEs that were identified in contexts of different reference species and varying evolutionary depth possess common characteristics despite their sequence diversity. First, HNCEs are found in proximity to similar sets of genes in the genome: namely, those that are related to development [46, 47]. Interestingly, this localization preference holds true across clades even though HNCEs particular to the clade bear no sequence resemblance to each other [48]. Second, HNCE sequences are generally significantly more AT-rich than the rest of the genome, and are often flanked by GC-rich sequences [49]. Closer bioinformatic analysis of HNCEs has consistently revealed large density of TFBSs and higher frequency of overlapping TFBSs.

These sequence and localization attributes of HNCEs as well as experimental data suggest that they may be regulatory elements. Chromatin modification

datasets now reveal that genes close to HNCEs (even though not necessarily the HNCEs themselves) are particularly enriched in both H3K27Me3 and H3K4Me3 marks – modifications that are related to heterochromatin formation and transcription activation respectively that can be found as co-existing "bivalent" marks [50]. HNCEs can function as an enhancer driving reporter genes in mouse and zebrafish out of their native genomic context. The resulting reporter expression can be extremely localized and temporally restricted during embryonic development [51]. This directly demonstrates that HNCEs are capable of regulating transcription. These observations, combined with the peculiar localization around developmental genes and the significant increase in conservation of enhancers active during gastrulation, associate HNCEs with the regulation of genes involved in body plan development. Experimental data supporting this hypothesis, however, are still needed [52].

A central question surrounding HNCEs, particularly UCEs, is the origin of their extreme conservation. The regulatory activity of HNCEs by itself is not sufficient explanation, since non-conserved enhancers also exist. The overlapped TFBSs on HNCEs may contribute to multiple constraints against loss-of-function mutation. As an enhancer may be used multiple times and their logic re-interpreted in a context-dependent manner, these constraints need not manifest themselves simultaneously. However, it was argued that degeneracy of TFBSs would require an extremely dense overlapping of functional sites that had yet to be observed. Alternatively (although somewhat less parsimoniously), these elements may have additional functions to that of an enhancer. Functions such as splicing control, nonsense-mediated decay regulation, homologous recombination, and structural maintenance of chromosomes have been proposed [52]. Even more confounding is the observation that separate deletion of four UCEs from mice revealed no obvious deleterious consequence to the animals or their progeny [53]. Overall, the challenge of explaining HNCE conservation and function remains open.

1.3 Methods for study of protein-DNA interactions

Biological questions involving transcription regulation often concern any of the following three entities: the *cis* regulatory DNA elements, the *trans* interacting proteins, and the target genes being affected. These can be further parameterized by various factors including cell type, signaling state, local and global changes to chromatin structures and higher-order chromosomal organization.

Classically, identification of *cis* regulatory elements and *trans* factors relied on genetic and biochemical techniques, many of which were laborious and designed to assess interactions given *a priori* knowledge of the interaction pair. For example, electrophoretic mobility shift assay (EMSA) [54] and DNA footprinting techniques [55] can assay a DNA sequence for binding to proteins, but does not give information on the identity of the protein itself. The past decade saw tremendous developments in both genomic and proteomic technologies, enabling sequence-based identification and quantification of proteins and DNA in complex mixtures. Marriage of classical biochemistry with "-omics" techniques – notably, massively parallel DNA sequencing and high-resolution mass spectrometry [56] – now enables unbiased discovery of both regulating DNA and regulating protein entities.

1.3.1 Protein-centric methods for DNA interactions

Sequence specificity of a DNA-binding protein can be determined using the protein as the bait to purify a library of DNA, and subsequently analyzing the recovered DNA sequences. For example, "systematic evolution of ligands by exponential enrichment" (SELEX) purifies DNA molecules with highest affinity to the protein out of a pool of random DNA oligonucleotides flanked by invariant adaptors [57]. The recovered DNA molecules from the first round of purification are amplified and their sequences evolved, e.g. by error-prone PCR. The resulting pool is again purified on the protein of interest. After several rounds of purification, amplification and evolution, the sequence pool converges to the set that defines the sequence specificity of the protein. Traditionally, the sequences

of the resulting DNA are determined by cloning and sequencing. Higher throughput in sequencing may be obtained by concatenating the adapter-stripped oligonucleotides prior to cloning [58], or by using next generation sequencing [59]. Another method for sequence specificity determination hybridizes recombinant transcription factors to a microarray of all possible k-mer DNA oligonucleotides ($k \geq 8$) [15]. Known as "protein binding microarray", this method provides a simple, high-resolution and quantitative alternative to SELEX.

Because the above approaches assess protein-DNA interactions in absence of the epigenetic constraints found *in vivo*, they are suitable for determination of biochemical affinity of a protein with DNA. The standard method for monitoring *in vivo* protein-DNA interaction is chromatin immunoprecipitation (ChIP), where interactions at endogenous chromatin are "frozen" by formaldehyde cross-linking. Chromatin is isolated and sheared by sonication to reduce the size of individual DNA fragments to around 200 – 500 bp. The sheared chromatin is then immunoprecipitated with an antibody raised against the protein of interest. DNA is liberated from the recovered chromatin fraction and amplified by PCR. Specific interaction between the factor and a given genomic locus is assessed by quantitative real time PCR with primers targeting the locus of interest (ChIP-PCR). Genome-wide mapping of bound chromatin is possible with next generation sequencing. The latter combination, termed ChIP-seq, was first published independently by at least three groups [60-62] and is now the protein-centric method of choice, as it offers a truly global interaction profile.

A variant on ChIP-seq termed "ChIP-exo" uses 5'-to-3' exonuclease digestion to degrade DNA strands up to the position where the strand is in contact with the protein (and hence protected from digestion). The sequencing reads from the resulting, undigested products can be used to map the position of the transcription factor up to single nucleotide accuracy, a resolution which is far greater than conventional ChIP-seq [63].

1.3.2 DNA-centric methods for protein interactions

ChIP-based methods are now routinely used to identify of DNA targets of a transcription factor of interest. However, development of equivalent technology for the reverse question – identification of protein binders of a given DNA sequence – is more challenging, largely owing to the lack of equivalent biochemistry for amplification and sequencing of protein molecules.

Ab initio identification of protein interactors of a DNA sequence can be done in high-throughput using the yeast one hybrid method. This approach is a variant of the yeast two-hybrid method, a classical genetic screen that re-constitutes a transcriptional activator at a reporter gene by interaction of two proteins of interest. In the yeast one-hybrid method, the bait DNA sequence is placed in front of the promoter driving a selection marker. The reporter strain is used to screen a cDNA expression library of candidate DNA-binding proteins fused to a strong transcriptional activator e.g. Gal4. If the prey-activator fusion binds to the DNA bait, a transcriptional activator complex is recruited to the promoter of the reporter construct and the selection gene is expressed. Appropriate selection conditions then yield colonies, whose transformed cDNA clones encode proteins that interact with the bait sequence [64]. Although high-throughput and unbiased in principle, the method limits the experimental conditions to binding of out-of-context DNA fragment to a fusion protein that is expressed in isolation. Thus, the interaction is assessed without epigenetic constraints and protein-protein interaction contexts are missing.

More recent developments that promise full recapitulation of cellular conditions capture and analyze the native chromatin directly, and may be considered truely complementary to ChIP. These methods include proteomics of isolated chromatin segments (PICh) and insertional chromatin immunoprecipitation (i-ChIP) [65, 66]; both employ mass-spectrometric identification of interacting proteins. The former method uses a complementary DNA oligonucleotide to hybridize and capture the target chromatin fraction, and the latter introduces into the genome a binding site for an exogenous transcription factor as purification handle. A current limitation of DNA-centric chromatin capture is the low signal to noise ratio, owing to the lack of a protein amplification method.

PICh thus requires a staggering amount of material (one billion cells per purification) and was initially demonstrated on telomeric sequences which are present in numerous copies per cell (in contrast with two copies per cell for non-repetitive DNA elements). Use of an orthogonal binding site in iChIP allows for protein-based tandem purification, improving the signal to noise ratio and thus reducing the material needed (100 million cells per purification). iChIP is currently limited by the laborious genome editing step which is required for every target sequence and every variant thereof, nevertheless a limitation which will hopefully be circumvented in future by more robust genome editing technologies [67, 68].

Biochemical affinity purification of proteins coupled to mass-spectrometric analysis (AP-MS) is an attractive approach, as it removes many of the practical limitations mentioned above. In this method, a chemically or enzymatically synthesized DNA bait is conjugated to an affinity handle, allowing immobilization on agarose or sepharose beads. Nuclear lysate is incubated with the DNA-coupled beads, washed, and the bound proteins recovered by specific elution. Under appropriate salt and detergent concentrations and given a suitable nuclear lysate extraction procedure, protein-protein interactions are preserved, enabling identification of both direct DNA binders and proteins that are part of DNA binding complexes. Although use of synthetic DNA has raised questions regarding the missing chromatin context in the experimental conditions, recent evidence suggests that synthetic DNA carrying a genomic regulatory sequence is capable of recruiting histones and mimicking local chromatin environment as found *in vivo* [40]. Use of synthetic DNA in AP-MS results in amplification of interaction signals, as the copy number of DNA used is up to 1,000-fold that of endogenous DNA in a conventional experimental scale. Consequently, only 1% to 10% of the material amount is required, compared to iChIP and PiCH.

AP-MS is the staple method of this study and will therefore be elaborated in greater depth with the principles of MS-based proteomics in the following section.

1.4 Mass spectrometry-based quantitative interaction proteomics

A portmanteau of "proteins" and "genomics", proteomics is the large scale study of proteins. A proteomic experiment identifies and quantifies proteins from complex biological samples, often involving some means of complexity reduction prior to analysis. The first implementation of proteomics was in the pre-human genome era, where samples were fractionated using two-dimensional electrophoresis and protein identities inferred based on the results of amino acid analysis [69], a technology with severe shortcomings that was no match for powerful genomics technologies. Through improved technology, increased computational power and the availability of complete genome sequences, sophisticated means of protein identification and quantification have developed and mass spectrometry (MS) has become the method of choice for proteomic study.

1.4.1 MS-based proteomic workflow

MS-based proteomics may be done either "top-down" or "bottom-up". The former approach submits intact proteins or protein complexes to the mass spectrometer, where they can be iteratively analyzed and fragmented in "tandem mass spectrometry" (see 1.4.3 below). In the more widely implemented bottom-up approach, protein mixtures are pre-processed into peptides which are then analyzed in the mass spectrometer; once all the peptide sequences are identified, proteins are assembled from them, based on a reference sequence database (see 1.4.4 below). This thesis exclusively employs bottom-up analysis and this workflow will be discussed in greater depth.

A typical bottom-up proteomic experiment starts with biochemical isolation of proteins from biological material, such as cells grown in culture or isolated from an organism. Optional enrichment steps may be performed depending on the biological question: For instance, a subcellular fraction may be isolated if only proteins belonging to certain organelles are of interest; or proteins may be affinity purified to study interactions with a specific bait.

Peptides are then generated by treating the proteins with a combination of proteases. Trypsin and/or lysyl-endopeptidase (LysC), which cleave C-terminally to arginine and/or lysine, are routinely used. Peptides may be further enriched for interesting post-translational modifications such as phosphorylation. The final sample is a complex mixture of peptides, which is separated by high-performance liquid chromatography (HPLC) coupled online to a mass spectrometer via an electrospray source (see 1.4.2 below). The peptides eluting from the HPLC column are ionized prior to entering the mass spectrometer. These ions are then mass analyzed, fragmented and their fragment ions analyzed again. The resulting data are processed into peptide sequences and protein identities are inferred using a sequence database.

Proteome coverage in LC-MS studies is constrained by some technical limitations. First, mass spectrometers have a dynamic range that is narrower than the copy number range of proteins being expressed in biological systems. Second, a finite number of fragment mass spectra can be acquired while peptides are being eluted in real time from the HPLC. Hence, the sensitivity, acquisition speed and dynamic range of the mass spectrometer directly influence the "depth" to which a complex protein sample can be covered [70]. This depth can be thought of as the proportion of lowest-abundance proteins that remain undetected. Previously, whole proteome analyses required extensive sample fractionation – such as by gel electrophoresis – with each fraction being analyzed separately to reduce the sample complexity, and thus deepen the proteome coverage. However, recent advances in instrumentation and computational algorithms have made it possible to obtain a comparably deep proteome without the need of fractionation [71-73].

1.4.2 Principles and implementations of mass spectrometry

A mass spectrometer is in essence a mass measuring instrument, consisting of three parts: the ionizer, the analyzer and the detector. Relying on ionization of the sample molecules, the analyzer performs mass- and charge-differentiating perturbations on the ions, and the detector translates a measurement of incidental ions or ion-generated current into mass-over-charge (m/z) ratios.

Several mass-spectrometric technologies have been developed over the past decades with differing ionization, analysis and detection approaches.

The most popular ionization method used for LC-MS is electrospray ionization (ESI). Liquid containing peptides eluting from the HPLC column tip is subjected to a voltage and dispersed into fine aerosol called electrospray. As the solvent evaporates, the charge density of the droplets that carry the peptides increases. Repulsion of like charges within the droplet causes recursive droplet fission, eventually exposing the peptides, which accept the excess charges, to the gas phase. Unlike many other ionization methods, electrospray is very gentle, capable of generating multiply charged ions, and therefore is particularly suitable for analyzing large biomolecules. Electrospray ionization was first used in mass spectrometry almost two decades ago and was recognized with a share of the chemistry Nobel Prize in 2002 [74].

Mass analyzers and detectors may be placed into different groups. A first group resolves ions by recording their flight to the detector, a principle termed Time-of-flight (TOF). The TOF analyzer relates the charge-dependent potential energy of the ions in the electric field with the mass-dependent kinetic energy, which can be measured by the time (hence velocity) the ions take to reach the detector. Quadrupole mass analyzers consist of four parallel hyperbolic electrode rods. Radio-frequency voltages offset by a direct current are applied between each pair of opposing rods, creating an electric field which guides ions of certain m/z in oscillating trajectories along the electrodes, while causing the other ions to collide into them. By manipulating the voltage ratios between the two electrode pairs over time, ions can be swiftly scanned over a desired range of m/z values. For each m/z value the detector the records the signal of the incidenting ion.

Another group of mass analyzers induces all ions to oscillate in a stable path under the influence of an applied electromagnetic field, wherein the oscillation frequency is directly dependent on m/z. The current generated by the oscillating ions is measured and decomposed into their separate m/z contribution by Fourier Transformation (FT). Thus, all ions are detected simultaneously. This principle is employed in the Fourier-transform ion cyclotron resonance (FT-ICR) analyzer, which traps ions in a magnetic field by Lorentz force [75]. The

Orbitrap analyzer uses the quadro-logarithmic electropotential, rather than a magnetic field, to implement the same concept. The Orbitrap consists of a barrel-shaped electrode with an inner, co-axial electrode. The ions rotate around the inner electrode as well as oscillating axially, and the square root of the latter frequency is inversely proportional to m/z [76]. Compared to the other instruments, this group of mass analyzers has a greater m/z resolution that increases with detection time, and is, for instance, particularly suitable for inference of molecular composition by their accurate mass.

1.4.3 Tandem mass spectrometry

Although the m/z deviation obtainable with instruments such as the Orbitrap analyzer is as low as few parts per million, this accurate mass information is still insufficient to infer the peptide sequence, because peptides of differing sequences but identical amino acid composition have identical masses. Further discriminating evidence can be obtained by isolating peptide ions at their m/z, activating them to break covalent bonds. The resulting fragment ions are then re-analyzed in a process called "MS-MS" or "MS2" (in contrast with "MS1" where the precursor ion is detected). Because fragmentation can occur at different covalent bonds, the resulting fragment ions generally include those generated from breaking of the peptide backbone at various positions, especially the peptide bonds (Figure 2). The pattern of m/z values can thus be used to re-assemble (parts of) the original peptide sequence, which in turn is validated for consistency with the accurate mass obtained in MS1 [77, 78].

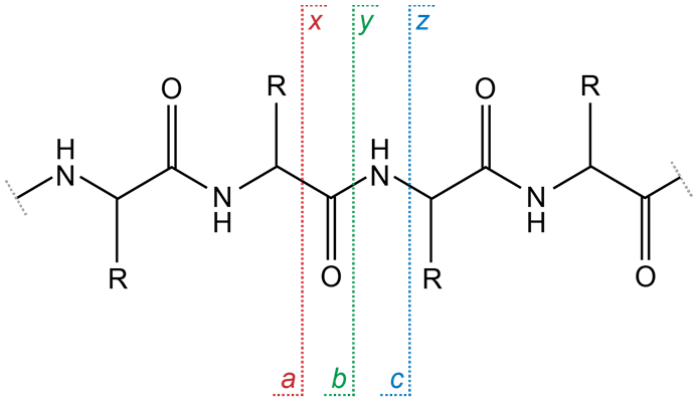

Figure 2: Fragmentation along the peptide backbone. Breakage of different covalent bonds along the peptide backbone during peptide ion fragmentation results in a series of N-terminal ions and corresponding C-terminal ions, which are named according to the Roepstorff-Fohlmann-Biemann nomenclature. In particular, the series of *b*- and *y*- ions are generated by fragmentation at different peptide bonds.

As mentioned above, the online-coupled chromatography that runs over a finite time imposes a practical limit on the number of MS2 scans that can be made, frequently leaving 'sequenceable' precursors unfragmented. This raises the question of how to prioritize ions for MS2 sequencing. When the identities of the peptides of interest are known, their corresponding m/z can be specifically monitored in MS1 and submitted for MS2 sequencing. Multiple reaction monitoring (MRM) is one of the most widely used implementation of this "targeted" approach [79, 80]. More commonly in proteomics, ions are prioritized for fragmentation by their signal intensity in the MS1 acquisition. Typically, between five and ten most intense ions measured in MS1 are submitted for MS2. This "shotgun" approach allows MS-based proteomics to identify peptides without *a priori* knowledge and is the method of choice for hypothesis-generating studies [81].

1.4.4 Peptide and protein identification

While software packages are available for interpretation of MS spectra, the MaxQuant suite is particularly powerful [82]. The implementation of MaxQuant used in this thesis is described in this section.

Mass-spectrometric data consist of MS1 m/z peaks, with many but not all peaks having accompanying MS2 fragmentation spectra. "Features" that are likely to be peptide ions are derived from individual m/z peaks; subsequently, identification of features as peptides is accomplished with reference to a protein sequence database. *In silico* digestion of the protein sequences generates theoretical peptides, which serve as candidates for matching MS1 and MS2 spectra. A peptide identification is declared when both the accurate m/z of the MS1 feature and the MS2 fragmentation pattern are consistent with a theoretical peptide. A score associated with each 'spectrum sequence match' is calculated based on the confidence of the contributing spectral evidence. The proportion of false identifications occurring by chance may be estimated by matching the spectra to a nonsense "decoy" database. A database constructed using all entries from the reference database reversed from C- to N-terminus is commonly used, because of its identical amino acid composition distribution.

Comparison between the decoy scores and the true scores yields a score cutoff that is used to filter the identifications at a desired false discovery rate (FDR).

Although MS1 features that do not have corresponding MS2 spectra do not have direct sequencing evidence, MS2-based identifications of the "equivalent" features from other LC-MS runs may be transferred to the run being analyzed. Features from different runs are assessed for equivalence according to their accurate MS1 m/z and elution time. Because of variation in the chromatography between runs, the retention times of a given peptide in these runs also differ. This can be accounted for by interpolating between the sets of equivalent features that are jointly sequenced across experiments. 'Matching between runs' is especially beneficial in the analysis of complex samples, as it is able to increase peptide identification count by as much as 40%.

Primary interpretation of MS spectra thus results in a list of peptide sequences, from which protein identities need to be inferred. Short peptide sequences (generally < 7 aa) are discarded because they often occur in unrelated proteins. Owing to sequence homology, splice variation, and redundancy in sequence databases, a longer peptide sequence can still be part of several protein sequences. Because such a sequence may not be unambiguously assignable to any one of these proteins, a concept of "protein groups" is introduced. Peptides are assigned to groups of proteins that are defined according to the principle of parsimony ('Ockham's razor): The simplest set of groups that is sufficient to explain all the identified peptides is reported.

1.4.5 Quantitative MS-based proteomics

Peptides are very diverse in their physical properties such as charge, chain length, and hydrophobicity. These properties unequally affect each peptide's digestion and purification yield, behavior in chromatography, and ionization efficiency. As a result, signal intensities of different, equimolar peptide ions in the mass spectrometer are generally not equal. Thus, different strategies that enable quantitative interpretation of MS data have been developed, which may be grouped into label-based methods and label-free methods.

Label-based quantification

In label-based methods, two or more samples are multiplexed and subsequently quantified relative to each other. A different mass label is incorporated into each sample prior to multiplexing. The labels uncouple peptide ion signals originating from the different samples into separate m/z peaks. This is achieved by introducing a defined mass-shift between the labels, typically owing to incorporation of heavy stable isotopes in the labels. Alternatively, the labels can make use of chemical groups that yield different masses upon fragmentation.

Labels are designed so that they are as identical as possible in their physicochemical attributes. Thus, differentially labeled peptides of identical sequence and modifications co-elute from the chromatography and ionize with the same efficiency. The resulting m/z signal intensities are therefore directly comparable. In this way, the ratio of intensities corresponds to the ratio of peptide abundance between samples. Protein abundance ratios are then estimated from the population of corresponding peptide abundance ratios. It follows that the precision of label-based protein quantification improves with the number of quantified peptides attributed to the protein.

Labels may be incorporated metabolically before protein extraction, or afterwards at protein or peptide levels. In the "isotope-coded affinity tag" method (ICAT), cysteine residues on proteins were chemically modified to include a differentially labeled tag which also served as an enrichment handle [83]. However, ICAT quantification is limited to cysteine-containing peptides only; this yields fewer ratio counts, leading to suboptimal quantification precision. A method that bypasses this limitation, termed "dimethyl labeling", incorporates an isotopic variant of dimethyl groups onto all free N-termini and primary amine side-chains [84].

Multiplexed quantification in MS1 increases the complexity and thus reduces the dynamic range of the MS1 spectra. Transferring the quantification peaks to the MS2 spectra, which are much less complex than MS1 spectra, alleviates this problem. The concept is used in "isobaric labeling" methods. Here, each label contains a mass-discriminable "reporter" group, covalently linked to a "balancing group" that adjusts all labels to the same mass. The differentially

labeled peptides are indistinguishable in MS1. Fragmentation of these peptides yields the different reporter groups in the MS2 spectra, where ratios of reporter intensities correspond to ratios of peptide abundance between the samples. Commonly used implementations of isobaric labeling include "tandem mass tag" (TMT) and "isobaric tags for relative and absolute quantification" (iTRAQ) [85]. As mentioned, isobaric labeling enables multiplex quantification without increasing the complexity in the MS1 scan. However, the method ties peptide quantification to MS2-evidenced identification, and is therefore incompatible with quantification by matching. Furthermore, MS1-based quantification can make use of the elution profile, which can be constructed from successive MS1 spectra, to improve quantification precision. To emulate this in MS2-based quantification, successive MS2 scans would have to be performed on the same precursor m/z, resulting in a trade-off between quantification precision and identification depth.

In contrast to chemical labeling, metabolic labeling methods incorporate labels in living cells, allowing samples to be combined even prior to cell lysis. This early mixing advantage means that all downstream handling errors are minimized by parallelization. Formerly, ^{15}N incorporation was used to label cells *in vivo*, but this method resulted in highly complex spectra because the mass shifts between the label counterparts differ wildly between peptides. Stable isotope labeling of amino acids in cell culture (SILAC) is now a widely-used metabolic labeling method [86]. In SILAC, cells are grown in media containing arginine and lysine that have different proportions of ^{13}C and ^{15}N isotopes (Arg0, Arg6, Arg10; Lys0, Lys6, Lys8). Proteins are digested with trypsin or LysC to ensure that almost all resulting peptides are quantifiable, owing to the labeled arginine or lysine at the C-terminus. SILAC has a clear advantage over ^{15}N labeling as every SILAC pair has a specific mass-shift, greatly simplifying the process of identifying label pairs. Metabolic labeling needs to be performed over at least five cell divisions for the labeled proteins to saturate the proteome, and is therefore particularly suitable for cells in culture or small animals. Incorporation over shorter time may be performed as a "pulse" experiment to study proteome dynamics [87].

Intensity ratios derived from label-based methods represent relative abundances of peptides between two or more samples. When one sample is a standard of known amounts, the ratios can then be used to infer absolute amounts in the remaining sample(s). Known as "absolute quantification", this concept has been implemented in a label-based or label-free format (see below) in technologies such as AQUA, PrEST, and iBAQ [88-90].

Label-free quantification

Label-free methods are computational procedures that report quantitative measurements of protein abundances without the use of a mass label. When labeling of biological material is not possible or is cumbersome (such as in clinical samples), label-free methods thus provide an alternative to the above approaches.

An early and simple label-free quantification algorithm was to use the number of MS2 spectra attributed to a given protein as a semi-quantitative measure for that protein's abundance [91]. This "spectral counting" method was improved by weighing each spectrum by the probability of it being acquired given the peptide's physicochemical properties [92]. However, by design, spectra-based methods trade off quantification resolution with identification confidence, as well as being influenced by the chromatographic parameters, which generally vary between experiments. A different approach uses the numbers of peptides identified to estimate protein abundances. The protein abundance index (PAI) is defined for a given protein as the ratio of observed peptide count to the theoretically observable peptide count. Its successor – exponentially modified PAI ($= 10^{PAI-1}$) – is directly proportional to the protein abundance [93, 94].

Quantification based solely on counts of peptides and spectra is discretized by nature. Furthermore, these approaches discard valuable information that is latent in the ion intensity measurements. More accurate label-free quantification method, for instance offered by the MaxQuant suite, takes peptide intensity information into account. In MaxQuant, peptide identifications are first transferred between runs as far as possible (see 1.4.4 above). For each

protein, pairwise sample comparison generates a matrix of median peptide ratios which derived from jointly identified peptides. These ratios form over-determined systems of linear equations that are used to back-calculate the relative protein quantities between samples. Simply known as "label free quantification", this algorithm was first used to quantify the dendritic cell proteomes to a depth of over 6,000 proteins, with superior precision to previous label-free quantification methods [95].

1.4.6 Quantitative interactomics

Mass spectrometry has proven to be a highly sensitive technology for protein identification. An implication of its power in the study of protein interactions is that, when affinity-/immuno-purified proteins are analyzed by mass spectrometry, specific interactors are identified together with several hundred that bind to the beads used in the purification or that bind non-specifically to the bait. Thus, quantitative measures are absolutely essential to identify the specific interactors from the remaining proteins.

Label-free algorithms have been used for quantitative analysis of pulldowns with very good precision [96]. However, label-free quantification readouts in affinity purification are a combination of specific-enrichment, protein expression levels in the lysate, and any contaminants introduced during sample handling. Unfortunately, the contribution of specific enrichment to the quantification cannot be resolved from the other confounding factors in label-free approaches, because these components are mixed into the same quantification "channel". In contrast, this is possible in label-based interaction experiments, where the principle of "label-switching" exposes the specific interactions and confounding factors in different combinations. Furthermore, since the resulting protein sample from pulldowns is low in complexity (typically 500 proteins) and does not suffer from duplication of ion peaks, label-based quantification is particularly attractive for interaction studies.

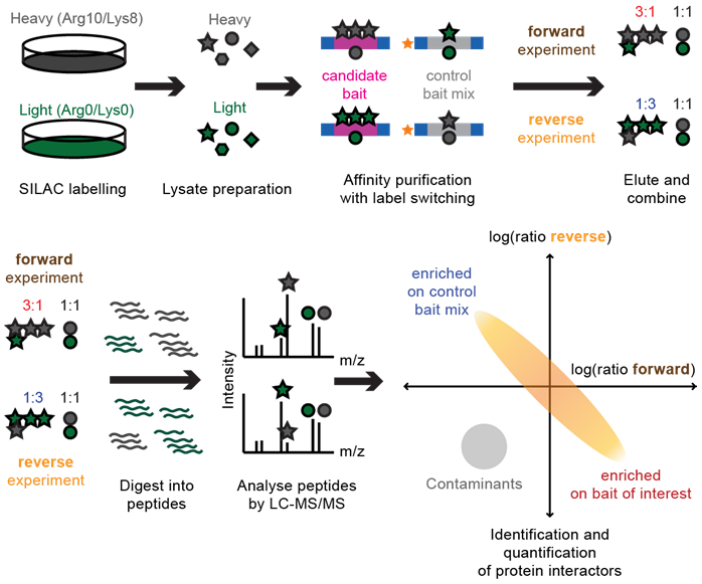

Figure 3: Overview of the forward-reverse SILAC AP-MS experiment design and data interpretation

See text for detailed description.

SILAC-based affinity purification (SILAC AP-MS) is a widely-used label-based approach for studying protein interactions [30, 97-100]. A typical experimental design for SILAC AP-MS is known as the "forward-reverse" setup (Figure 3). Here, interactions are compared between a candidate bait and a control bait, using heavy-labeled and light-labeled lysates. Two sets of affinity purifications are performed: In the first ("forward"), the candidate bait is used to purify the heavy-labeled lysate and the control bait is used to purify the light lysate. In this experiment, a specific interactor with the candidate bait would have a heavy-to-light ratio of greater than 1:1. In the second set of purifications ("reverse"), the lysates are swapped with respect to the bait; here, a specific interactor would have a heavy-to-light ratio of less than 1:1.

Owing to label-switching, specific interactors would therefore have SILAC ratios that are inverse of each other. This is usually visualized in a plot of logarithmized forward and reverse SILAC ratios known as the "forward-reverse plot", where specific interactors lie along the anti-diagonal. Abundance differences of non-specific binders between the heavy and the light lysates would give rise to log SILAC ratios of the same sign, since the heavy-to-right enrichment/depletion is bait-independent for these proteins. Following from these rules, contaminants introduced by manual handling, being always of the light label state, are found in the double negative quadrant in the plot. Thus, the

forward-reverse plot enables intuitive, visual discrimination of specific interaction from background and contaminants. Since the label-switching design subjects both the heavy and light lysates to affinity purification on the specific bait, the forward and reverse experiments also serve as biological duplicates.

1.5 Aims of this study

Comparative genomics has predicted many potential regulatory DNA elements that have been functionally confirmed *in vivo*. The elements of interest in this study were the ultraconserved elements (UCEs) introduced in section 1.2. The primary goal of this thesis was to identify protein-DNA interactions at UCEs. For this question to be addressed in an unbiased manner, a DNA-centric method for protein-DNA interaction was needed. Furthermore, because the exact biological context in which UCEs function is unknown, we focused on the interactions that are intrinsic to the UCE sequences, as opposed to their *in vivo* binding. For the above reasons, we developed a strategy for upscaling the state-of-the-art SILAC AP-MS technology, and used it to discover proteins that bind to or are depleted from specific UCEs.

Secondly, this thesis addresses the curious evolutionary question involving the UCEs: What contributes to their extreme conservation? Although the hypothesis of overlapping TFBSs has been long proposed, it has been argued against largely based on the lack of supporting experimental data [52]. We reasoned that SILAC UCE pulldown experiments could fill in this gap, identifying the motifs that have direct biochemical evidence of binding and assessing the extent to which the superimposition of functionally interacting motifs contributes to the extreme conservation.

Thirdly, this thesis aims to integrate DNA-centric interaction data with complementary protein-centric data recently released by the ENCODE consortium [101], with the intention of critically assessing the relevance of AP-MS data in the chromatin context. Specifically, this work explored the extent to which the DNA sequence and the nuclear proteome together define local epigenetic states in the nucleus. Previously, an exemplary DNA sequence has been demonstrated *in vitro* to recapitulate the native chromatin modifications found in its corresponding locus. Here, we attempt to generalize this observation by comparing chromatin modification ChIP-seq datasets to the AP-MS interaction profiles of the 190 UCEs sequences screened in our interactome.

These aims are critically dependent on quantitative interpretation of SILAC AP-MS data. Although SILAC interactomics is quantitative by nature, actual interpretation of forward-reverse experiments has generally been qualitative. Previous studies focused on interactor calling rather than comparing enrichment factors across baits, and simply excluded false positives as not meaningful. This thesis refined the interpretation of forward-reverse SILAC AP-MS data, incorporating information from the so-called "false positive" hits. For this purpose, we implemented a simple correction procedure that quantitatively decouples expression changes from specific binding, improves enrichment estimates, reduces systematic error, and allows ratios to be used in a truly quantitative manner.

2 Improving large-scale SILAC AP-MS precision by proteome variation uncoupling

Summary

Mass-spectrometric analysis of affinity-purified protein (AP-MS) is a powerful method for unbiased discovery of protein interactions with other biomolecules. An approach using stable isotope labeling in cell culture (SILAC) and the "forward-reverse" label-switching design can be used to discriminate contaminants from specific interactors. However, the enrichment ratios derived from the label-switched experiments often show large variations, preventing the ratios themselves from being used confidently for quantitative interpretation. Here, we introduce an improvement to the processing and interpretation of the SILAC AP-MS data, which corrects for systematic errors introduced by the proteome variation between labeled samples. This simple correction procedure significantly improves quantitative interpretability of label-based AP-MS data that employs label switching, and normalizes systematic differences between batches in large-scale affinity purification screens.

2.1 Introduction

SILAC affinity purification coupled to mass spectrometry (SILAC AP-MS) has been used extensively to discover protein-protein, protein-peptide, and protein-nucleic acid interactions without *a priori* knowledge [30, 97-100]. A typical SILAC AP-MS study compares interaction between a specific bait of interest and a control bait; for instance, a peptide against its post-translationally modified variant or a regulatory DNA oligonucleotide against a point mutation. The principle of SILAC AP-MS has already been described in 1.4.6 above. Briefly, heavy-labeled and light-labeled lysates are affinity-purified with the specific bait and the control bait in different combinations. The resulting data are typically visualized in the "forward-reverse" scatter plot. There, specific enrichments or depletions are found in quadrants wherein the forward and reverse ratios are inversed owing to label switching. Contaminants are found in the double negative quadrant.

Often, studies employing SILAC forward-reverse AP-MS probe interactions against a control bearing small point mutations that were chosen rationally and specifically for every bait: e.g. from conservation, single-nucleotide polymorphism, or known post-translational modification. Because of the small change between the baits, data generated from such studies usually yield a population of specific interactors which are visually separated from the cloud of background binders. In these cases, the actual SILAC ratios were generally not needed to call interactors. Furthermore, quantitative comparisons of ratios between different forward-reverse experiment sets were generally not made.

In contrast to previous studies, the main goals of this thesis are dependent on the ability to perform quantitative cross-comparisons of protein interactions to multiple DNA sequences. However, substantial variation between the forward and reverse ratios often observed in SILAC AP-MS data, reducing the confidence of simply using the average ratios for cross-comparisons. Furthermore, a number of proteins are found in the double positive quadrant that cannot easily be interpreted. This artefact originates from the variation in lysate preparation, and the proteins falling in this quadrant were traditionally considered "false

positives" and simply excluded from further analysis. This rendered SILAC lysates that are vastly different in their observed proteomes owing to variation in lysate preparation incompatible with AP-MS experiments. Furthermore, this qualitative treatment of arbitrarily removing false positives prevents full interpretation of the information in the quantitative data, because it results in many missing quantifications and raises the question of general reproducibility between experiments. Equally problematic is the presence of known transcription factors in the "contaminant" quadrant. Together, these anomalies mean that the traditional interpretation of ratios in this experimental setup may have been suboptimal.

This chapter quantitatively addresses the way lysate variation contributes to the observable ratios in DNA SILAC AP-MS experiments. We offer a simple correction procedure termed "*ΔP*-adjustment", which uncouples this contribution from bait-specific enrichment/depletion. This procedure is applicable to any label-switch experiment where many baits are screened using the same sets of lysate, and where the enrichment can be experimentally uncoupled from the labeling. The resulting, corrected SILAC ratios have significantly less variation between the forward and reverse pulldowns, and now reflect the true magnitude of the random errors in the experiments. We also explored the application of this adjustment procedure in a large-scale, multi-batch screen, and showed that batch-wise adjustment results in further significant error reduction when compared to batch blind adjustment. This observation demonstrates the need for large-scale, multi-batch SILAC AP-MS data to be corrected for batch-to-batch variation, even when the lysates used are equivalent.

2.2 Derivation

2.2.1 Geometrical interpretation of the forward-reverse plot

We recall the intuitive interpretation of the SILAC forward-reverse setup as follows: When the heavy and light proteomes are identical, an a-fold specific enrichment of a given protein results in forward and reverse SILAC ratios (x and y respectively) which are exactly inverse of each other. Since the ratios on the forward-reverse plot are logarithmized, we express them accordingly here. That is,

$$\begin{aligned} \log x &= \log a + error \\ \log y &= -\log a + error \end{aligned} \tag{1}$$

Suppose that, for a different protein, there is no binding preference between the specific and the control baits. Instead, the abundances of this protein in the heavy and light lysates are different, as P_H and P_L respectively. Here, the logarithmized SILAC ratios for both experiments are simply:

$$\log x = \log y = \log P_H - \log P_L + error \tag{2}$$

Now we consider a hypothetical protein, which does have a binding preference for the specific bait but also has an abundance difference between the lysates. We assume that the proteome difference and the specific enrichment components are independent and express the expected SILAC ratios as their product. (This assumption is explored further in the next section.) Working under this assumption, and defining $\Delta P = (\log P_H - \log P_L)$ then,

$$\begin{aligned} \log x &= \Delta P + \log a + error \\ \log y &= \Delta P - \log a + error \end{aligned} \tag{3}$$

These equations form the basis for the visual interpretation of the forward-reverse plot: Enrichment and depletion contribute to the anti-diagonal

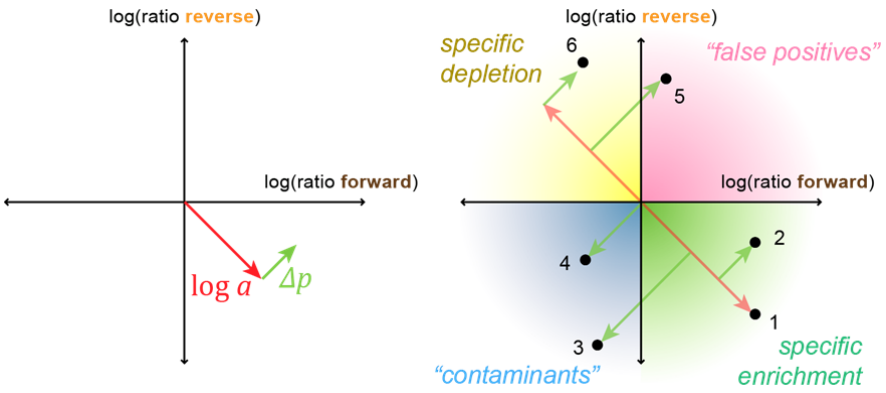

Figure 4: Geometrical representation of the SILAC forward-reverse plot

(**A**) The contribution of specific enrichment to the observed ratio is in opposite signs owing to label switching (red arrow). In contrast, the contribution of the proteome difference between the heavy and light lysates is always in the same direction (green arrow). The result is the offsetting of SILAC ratios from the main anti-diagonal as shown in the right panel. (**B**) Examples of how different proportions of specific enrichment and proteome variation affect data points on the forward-reverse plot. An ideal interactor would have zero abundance difference and a specific binding preference (1), whereas a truly non-specific contaminant would have no enrichment component (4). Usually, however, both components are visible, and ratios are shifted off the main anti-diagonal owing to the relative enrichment in the heavy lysate (2, 5, 6) or in the light lysate (3, 4). This offsetting can push some specific interactors into the "forbidden" contaminant and false positive quadrants. Traditionally, these ratios would be excluded from interpretation. See also **Figure 5** for how the adjustment procedure re-enables quantitative interpretability of these cases.

positioning of the proteins on the plot, and difference in protein levels in the two lysates moves the proteins along the diagonal.

For sets of pulldowns using the same lysates, ΔP can be therefore be deduced from the expressions for both ratios by eliminating the enrichment variable a and averaging over all n baits:

$$\log x + \log y = 2\Delta P + error$$

$$\Delta P = \frac{1}{2n}\sum_{i=1}^{n}(\log x_i + \log y_i + error_i)$$

(4)

By definition, the error term averages to zero across all baits, yielding ΔP as simply the average of all SILAC ratios for the protein.

$$\Delta P = \frac{1}{2n}\sum_{i=1}^{n}(\log x_i + \log y_i)$$

(5)

Now known, this confounding systematic error can be subtracted away to give ΔP-adjusted SILAC ratios:

$$x' = xe^{-\Delta P}$$
$$y' = ye^{-\Delta P} \tag{6}$$

Interpretation of ΔP is consistent with contaminant calling ($\Delta P < 0$) and general lysate variation ($|\Delta P| \gg 0$), which is summarized in Figure 4.

This derivation of the ΔP-adjustment procedure is based on the intuitive, geometrical interpretation of the forward-reverse plot, where independence between the proteome component and the enrichment component is assumed. However, it is clearly not possible that this assumption is valid over all conditions, as the maximum amount of recovered proteins is capped by the amount of binding sites available on the baits. Interplay between the bait concentration, the interactor abundance and the specific enrichment is better accounted for by thermodynamic considerations. In the next section, we show by thermodynamic derivation that the assumption of proteome-enrichment independence is valid over the range of conditions typically encountered in standard SILAC AP-MS experiments.

2.2.2 Thermodynamics of protein-DNA SILAC AP-MS

Affinity purification is a reversible complex formation between a "bait" molecule and a "prey" molecule. In the context of protein-DNA interaction, treating each DNA bait and each protein or complex interactor as a single entity, then the interaction between the bait B and prey P

$$B + P \rightleftharpoons B \cdot P$$

reaches the equilibrium concentration of the complex $B \cdot P$ at

$$[B.P] = \frac{[B][P]}{K_d^{(B,P)}} \tag{7}$$

where $K_d^{(B,P)}$ is the association constant between B and P.

Expressed in terms of initial concentrations $[B]_0$ and $[P]_0$, the above equation can be rewritten as:

$$[B \cdot P] = \frac{([B]_0 - [B \cdot P])([P]_0 - [B \cdot P])}{K_d^{(B,P)}} \tag{8}$$

This equation is quadratic with respect to the bound prey concentration $[B \cdot P]$, and whose exact solution cannot linearly and independently depend on $[B]_0$ and $[P]_0$. However, in typical SILAC AP-MS experimental conditions, this expression for the bait-prey complex concentration can be simplified so that it linearly depends on the initial prey concentration.

We typically use 8 µg of DNA bait and nuclear lysate from 10 million cells per affinity purification. Transcription factor copy numbers have been determined to have a range of ~250 to 300,000 copies per nucleus [102]. Under an estimated reaction volume of 200 µl, these conditions translate to a concentration of ~20 pM to 25 nM for each transcription factor and 250 nM of DNA bait. Within these parameter constraints, the DNA bait is in excess of the preys by at least ten fold, and so $[B]_0 \gg [B \cdot P]$. Considering (8) in this context, $[B]_0 - [B \cdot P]$ can be approximated by $[B]_0$ to yield the following simplified equation:

$$[B \cdot P] = \frac{[B]_0([P]_0 - [B \cdot P])}{K_d^{(B,P)}} \tag{9}$$

Rearranging to solve for $[B \cdot P]$, we obtain:

$$[B \cdot P] = \frac{[B]_0[P]_0}{[B]_0 + K_d^{(B,P)}} \tag{10}$$

Defining a function $f(x)$ as

$$f(x) = \left(\frac{[B]_0}{[B]_0 + K_d^{(B,P)}}\right) \tag{11}$$

Then the bait-prey complex concentration at equilibrium can be expressed as finally as:

$$[B \cdot P] = [P]_0 \cdot f\left(K_d^{(B,P)}\right) \tag{12}$$

Here, f transforms the dissociation constant into a scaling factor, which dictates the proportion of the prey molecules that are part of the complex at equilibrium. The value of $f\left(K_d^{(B,P)}\right)$ depends on the initial bait concentration $[B]_0$, which is invariant across the entire screen. Thus, the bait-prey complex concentration at equilibrium depends linearly and independently on only the initial prey concentration and a function of the dissociation constant. In conclusion, the assumption of independence between the proteome and the specific enrichment, which we used for the geometrical derivation of ΔP-adjustment, is valid under typical SILAC AP-MS experimental conditions.

To complete the thermodynamic derivation of the ΔP-adjustment procedure, we now apply (12) to the ratio calculations. Given the initial concentrations of the control bait $[B_0]_0$, the specific bait $[B_1]_0$, the heavy-labeled prey $[P_H]_0$, and the light-labeled prey $[P_L]_0$, then the observable forward ratio x is given by

$$x = \frac{[B_1 \cdot P_H]}{[B_0 \cdot P_L]} = \frac{[P_H]_0 \cdot f\left(K_d^{(B_0,P_H)}\right)}{[P_L]_0 \cdot f\left(K_d^{(B_1,P_L)}\right)} + error \tag{13}$$

The forward-reverse plot displays the SILAC ratios in the logarithmized form.

$$\log x = (\log[P_H]_0 - \log[P_L]_0) + \left\{\log f\left(K_d^{(B_0,P_H)}\right) - \log f\left(K_d^{(B_1,P_L)}\right)\right\} + error \tag{14}$$

In the reverse experiment, the labels are switched with respect to the baits. The expression for the reverse ratio y thus takes this form:

$$\log y = (\log[P_H]_0 - \log[P_L]_0) + \left\{\log f\left(K_d^{(B_1,P_H)}\right) - \log f\left(K_d^{(B_0,P_L)}\right)\right\} + error \tag{15}$$

We now make an important assumption: We expect the cells that give rise to the heavy and the light lysates to be biochemically equivalent, i.e. any difference in their proteomes does not result in differences in binding affinities of proteins. Under this assumption, then

$$K_d^{(B_0,P_H)} = K_d^{(B_0,P_L)} \text{and also } K_d^{(B_1,P_L)} = K_d^{(B_1,P_H)} \tag{16}$$

Defining a, the logarithmized enrichment factor between the two baits and ΔP, the logarithmized protein abundance fold change between the two samples:

$$a = \log f\left(K_d^{(B_0,P_H)}\right) - \log f\left(K_d^{(B_1,P_L)}\right)$$
$$\Delta P = \log[P_H]_0 - \log[P_L]_0 \tag{17}$$

Substituting (16) and (17) into the expressions for $\log x$ and $\log y$ then yields the intuitive expression for the logarithmized forward and reverse SILAC ratios as given in (3).

$$\log x = \Delta P + \log a + error$$
$$\log y = \Delta P - \log a + error$$

In summary, this section shows that the biochemical bases behind the derivation of the ΔP-adjustment procedure agrees with the geometrical interpretation of the forward-reverse plot.

2.3 Results

To benchmark the ability of the ΔP-adjustment procedure to reduce variability between forward and reverse SILAC pulldown experiments, preliminary SILAC DNA pulldowns for the UCE interactome were performed. We used two sets of SILAC nuclear lysate: the lysates obtained from R1/E mouse embryonic stem cells were purified on 47 UCE baits against a universal control bait, and lysates obtained from HeLa cells were purified on 23 UCE baits. Within each set, the pulldowns were parallelized in the 96-well plate format.

2.3.1 Up to 75% of the interactome have systematic, correctable proteome-difference errors.

First, we visualized the distribution of ΔP of all proteins for each SILAC pair of nuclear lysates. Since SILAC ratios are generally log-normally distributed, we could test the null hypothesis of $\Delta P = 0$ with a Student's T test. Correcting for multiple comparisons, we found significant systematic residual abundance difference in over three quarters of all proteins that were quantified in each pair of lysate at 1% FDR.

The spread of ΔP in the HeLa dataset was large, with 95% of the proteins displaying a systematic abundance difference over a range of 8-fold (Figure 5A). In contrast, 95% of proteins in the R1/E dataset showed an abundance difference of a range of 1.4-fold. This lower deviation was consistent with the fact that the R1/E lysates had been pooled from several preparations, normalizing out individual proteome deviations between preparations. Estimation of ΔP and its associated errors may be visualized for a given protein by plotting its forward and reverse ratios across all baits. In case of a well-behaved DNA binder, these ratios form a tight anti-diagonal whose distance from the anti-diagonal through the coordinate is directly proportional to ΔP (Figure 5B). In sum, this initial analysis demonstrates the potential of ΔP-adjustment in improving the precision of both datasets.

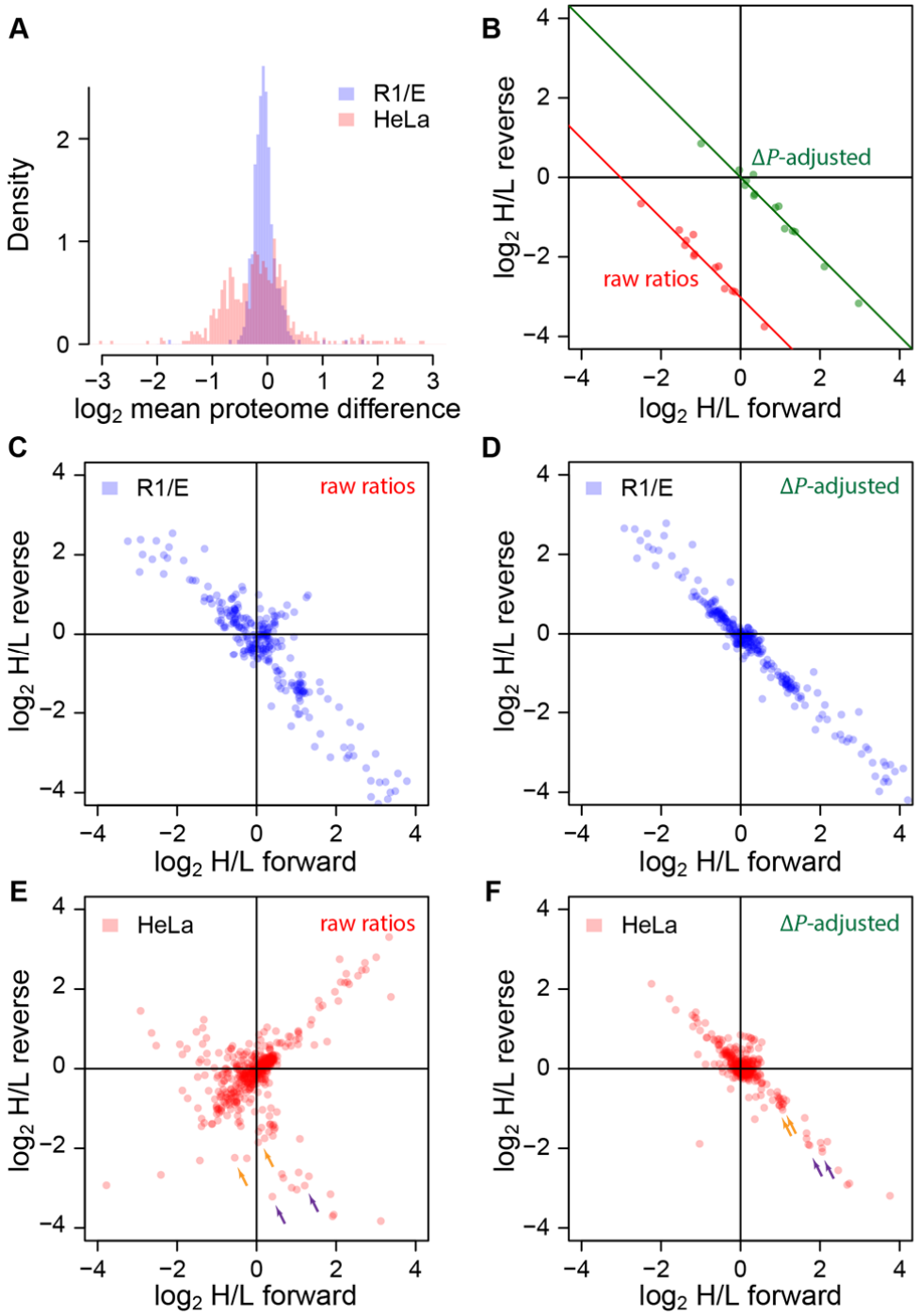

Figure 5: *ΔP*-adjustment significantly improves SILAC AP-MS ratio precision.

A Distribution of estimated proteome difference between heavy and light lysates in the R1/E and HeLa datasets. **B** Forward and reverse SILAC ratios of the protein JUNB as quantified over the 17 baits where they were identified. **C-F** Forward reverse ratios of a random pulldown example prior to adjustment (**C**, **E**) and the corresponding ratios from the sample pulldowns after adjustments (**D**, **F**). Orange arrows indicate JUNB and JUND. Purple arrows indicate ATF1 and ATF7.

2.3.2 Δ*P*-adjustment significantly reduces variability between forward and reverse SILAC AP-MS experiments.

We then performed the Δ*P*-adjustment on both pulldown datasets as described above. The adjustment resulted in a visually appreciable improvement in the ratio reproducibility between the forward and reverse experiments, as seen from the narrowing of data points towards the anti-diagonal where quantitatively reproducible enrichment/de-enrichment are expected. By calculating the root mean squared error (RMSE) of forward versus reverse ratios over all proteins for each experiment, we found that the improvement was highly significant overall: RMSE was reduced by 44% in the R1/E dataset and 79% in the HeLa dataset ($P < 10^{-24}$ in both cases). Thus, even pooled lysates with normalized proteome differences can still benefit from the Δ*P*-adjustment procedure.

One metric used to score specificity of an interactor is to compare their SILAC ratios against the "cloud" of non-specific binders, and calculating significance assuming that the unspecific binder ratios are normally distributed. However, owing to proteome differences, the score originating from the forward ratio can vary greatly from that from the reverse ratio. This is reflected in the irregularly-shaped cloud seen in the uncorrected forward-reverse plots (Figure 5C, E). However, the cloud of nonspecific binders become "regularized" into the anti-diagonal, giving a shape that is ideal for outlier statistics (Figure 5D, F). Overall, the adjustment procedure was able to recover highly-reproducible quantitative interaction data that were obfuscated by proteome differences in the lysate.

2.3.3 Δ*P*-adjustment removes false positives and recovers misclassed interactors

We next inspected the behavior of data points usually interpreted as "contaminants" and "false positives" in the forward-reverse plot. The former correspond to those in the "double negative" quadrant and the latter are those in the "double positive" quadrant. Both sets of data points have prevented a truly quantitative treatment of ratios without a manual filtering step.

After Δ*P*-adjustment, the majority of points both in the "double positive" and the "double negative" quadrant were corrected into the main anti-diagonal where reproducible interactions were found. Importantly, some of these points turned out to be valid, highly reproducible interactors that would have been misclassed as contaminants without the adjustment. For instance, the proteins JUND, JUNB, ATF1 and ATF7 were very far from the main anti-diagonal prior to adjustment, and JUNB would have been classed as a contaminant as it appeared in the lower-left quadrant. After the correction was applied, all four proteins reappeared as interactors that were clearly separated from the central cloud of background binders. Notably, probably owing to similar binding specificity between these proteins, their ratios also became more similar to each other after correction. As expected, the non-DNA binding, actin-associated protein SWF1 remained in the contaminant quadrant: This protein turned out to have irreproducible ratios over all baits.

In conclusion, the adjustment procedure is capable of recovering interactors that would have been missed owing to proteome bias between lysates, while retaining true, irreproducible contaminants in the contaminant quadrant.

2.3.4 Batch-wise Δ*P*-adjustment results in lower ratio variability in multi-batch experiments

The main dataset in this thesis (Chapter 3) was an interactome of 216 DNA baits against the R1/E background. Parallelization of pulldowns on the 96-well format allowed up to 48 AP-MS forward-reverse pairs to be screened, resulting in five batches of pulldowns across several days. Even though nuclear lysates were pooled and equivalent aliquots were used in each batch, the lysates may still be subjected to different conditions on different days, possibly introducing systematic errors for that batch that would be random across batches. This gives rise to the question of whether Δ*P*-adjustment should be applied batch-wise to account for batch-specific handling errors.

Analysis of variance (ANOVA) revealed that almost 60% of the interactome had significantly different Δ*P* values between batches (FDR 0.1%), strongly arguing

in favor of batch-wise adjustment. (Figure 6A, B). Comparing the forward-reverse RMSE between after global adjustment versus after batch-wise adjustment, we found a further reduction of 20% RMSE with batch-wise adjustment ($P < 10^{-15}$, Figure 6C). Thus, batch-wise adjustment should be used for multi-batch data processing even when the originating lysates were equivalent.

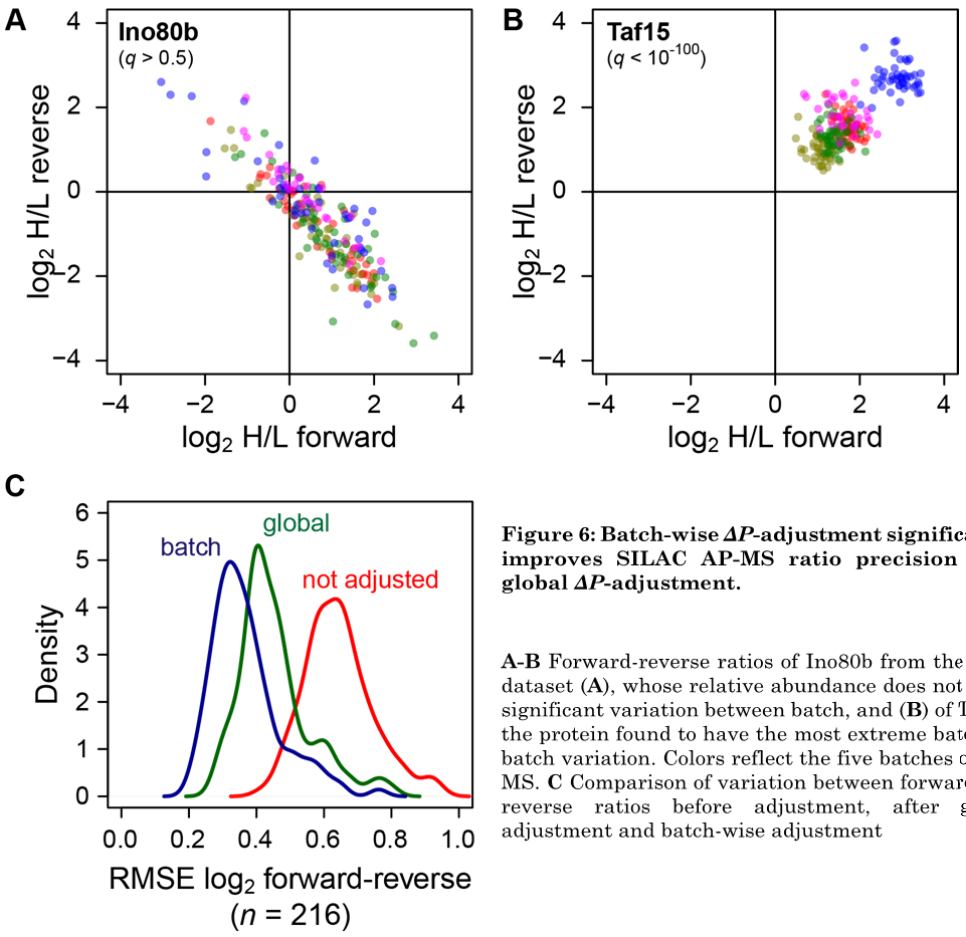

Figure 6: Batch-wise *ΔP*-adjustment significantly improves SILAC AP-MS ratio precision over global *ΔP*-adjustment.

A-B Forward-reverse ratios of Ino80b from the R1/E dataset (**A**), whose relative abundance does not show significant variation between batch, and (**B**) of Taf15, the protein found to have the most extreme batch-to-batch variation. Colors reflect the five batches of AP-MS. **C** Comparison of variation between forward and reverse ratios before adjustment, after global adjustment and batch-wise adjustment

2.4 Discussion

We have introduced a simple "ΔP-adjustment" procedure that uncouples proteome variation in the biological material from the true enrichment signal in forward-reverse SILAC AP-MS. The deconvolution was possible owing to the label switching in the forward-reverse experiments, exposing the proteome variation in different combinations with the specific enrichment. In addition, the range of bait and prey concentrations used in routine SILAC AP-MS purifications is such that the proteome variation is essentially independent from the specific enrichment. This assumption is generally true in affinity purifications where the bait is in large excess of the prey, a condition that is needed for analytical affinity purification experiments where observation of differential binding is the objective.

The procedure significantly de-noised AP-MS datasets, automatically corrected for the otherwise uninterpretable false positives, recovered interactors that would otherwise be missed owing to proteome variation, and normalized batch-to-batch variations. In summary, we have shown that ΔP-adjustment procedure is a highly beneficial and often necessary preprocessing step for large-scale SILAC AP-MS datasets that allows high-confidence, cross-batch, quantitative interpretation.

An obvious limitation of the ΔP-adjustment algorithm is that it is unable to adjust ratios of proteins that are only quantified in one forward-reverse pair, as no further information is available to estimate ΔP in such cases. We have, however, found that such cases are relatively rare owing to matching of peptide identifications across a large number of experiments.

DNA pulldowns routinely use crude nuclear lysate, whose slight variation in preparation can introduce large apparent proteome variations. Steps such as dounce homogenization and lysate clearing of lipids and cytoskeleton are primary sources of these lysate proteome variations. Importantly, the ΔP-adjustment procedure now relaxes the requirement for perfectly correlating heavy and light proteomes for AP-MS, as demonstrated by its ability to collapse the variation of up to 800% for proteins in the HeLa dataset, restoring its

interpretability. While we do not wish to imply that crude nuclear lysate can be prepared with less care, the adjustment does allow the lysates to be used in batch AP-MS despite their variation introduced through the preparation.

An assumption that is built into our procedure is that the proteins in the heavy lysate have equal affinities to the bait as the corresponding proteins in the light lysates. This holds true in the case where all experimental conditions (except for the label state) are identical in the cells. Although not explored in this chapter, it would also be possible to extend the adjustment procedure to include cases where the differentially labeled cells were subjected to different stimuli. With a careful experimental design, it would be possible to use a similar processing to systematically study interactions as a function of different cellular conditions.

3 Interactome of Ultraconserved Elements

Summary

Ultraconserved elements (UCEs) have been subject of great interest owing to their extreme sequence identity and their seemingly cryptic and largely uncharacterized functions. Although in vivo studies of UCE sequences have demonstrated regulatory activity, protein interactors at UCEs have not been systematically identified. Here we combined high-throughput affinity purification, high-resolution mass spectrometry and SILAC quantification to map intrinsic protein interactions for 193 UCE sequences. The interactome contains over 400 proteins, including transcription factors with known developmental roles. We demonstrate based on our data that UCEs consist of strongly conserved overlapping binding sites. We also generated a fine-resolution interactome of a UCE, confirming the hub-like nature of the element. The intrinsic interactions mapped here are reflected in open chromatin as indicated by comparison with existing ChIP data. Our study argues for a strong contribution of protein-DNA interactions to UCE conservation and provides a basis for further functional characterization of UCEs.

3.1 Introduction

Transcriptional regulation is determined by complex interactions of DNA, transcription factors (TFs), and chromatin states. Transcriptional regulatory elements capable of modulating gene expression have been of much interest due to their role in development and disease [103, 104]. Conservation analysis, chromatin modification state analysis and *in vivo* reporter assays have been used to identify several hundreds of such transcriptional enhancers [51, 105, 106]. Among these, ultraconserved elements (UCEs) – DNA elements defined by their 100% sequence identity over 200 bp between human and mouse genomes – have been identified as tissue- and stage-specific enhancers [43, 51, 106]. UCE sequences were predicted to be enriched in binding sites for development-associated TFs, suggesting important developmental regulatory roles. However, relatively few phenotypic alterations have been associated with loss or mutation of UCEs [107-109], and while several hypotheses have been proposed [110], little has been attempted experimentally to account for the ultraconservation of these loci. Similarly, although regulatory potential of UCEs have been demonstrated through embryonic reporter assays, the function and mechanism of these regulatory elements largely remain to be explored.

One starting point to enhancer characterization is through interactor mapping. Recently, chromatin immunoprecipitation (ChIP) has mapped out interaction of the genome to several TFs in great detail [101]. ChIP is protein-centric, i.e. they map out target DNA sequences bound to pre-chosen TFs, limiting the diversity of interaction profiles to *a priori* knowledge. Furthermore, ChIP data reflect an end point of gene regulation, incorporating aspects such as chromatin homeostasis and long-range interactions, rendering the contribution of the underlying DNA sequence difficult to determine. Evidence from a small number of genomic loci as well as whole-chromosome analysis has demonstrated the genetic contribution to establishment of epigenetic states [40, 111]. Thus, DNA-centric study of intrinsic interactions between DNA sequences and DNA-binding nuclear proteins in absence of initial epigenetic priming is valuable to understanding the genetic contribution to transcriptional regulation, which is especially important for dissecting per-nucleotide conservation of UCEs.

Past studies have employed a DNA-centric approach to identify potential binders of small numbers of DNA sequences [65, 98, 99, 112, 113]. Here we have developed a high-throughput platform to screen unbiased interaction profiles for hundreds of DNA sequences, based on our previously described pulldown method using high-resolution mass spectrometry and SILAC quantification [99]. We applied this technology to obtain an interaction map for 193 UCEs, including over half of all non-exonic UCEs in the genome. We found non-exonic UCE sequences to bind TFs and chromatin remodelers with known roles in developmental regulation, whereas proteins that promote chromatin compaction were relatively depleted. We inferred that the protein interactors bind to UCE sequences through densely distributed and often overlapping canonical transcription factor binding sites (TFBSs). Individual DNA bases that are part of overlapping TFBSs were on average more stringently conserved among vertebrates. We also obtained mapped intrinsic interactions of one UCE to five nucleotide resolution and found a high frequency of both gain and loss of binding to occur upon mutation. Finally, comparison of our intrinsic interaction map with existing ChIP-seq data as well as reporter assays linking previous independent observations [114, 115] highlight the functional relevance of these interactions. Overall, our interaction map points towards extremely high information content and complex transcription regulation logic behind many UCEs.

3.2 Results

3.2.1 The UCE interactome

We obtained the interaction map for 129 of 256 non-exonic (nx), 36 of 114 putative-exonic (px), 28 of 111 exonic (ex) UCEs as well as 21 human and 3 mouse random genomic loci by affinity purification, high-resolution mass spectrometry and SILAC quantification in high throughput [96]. We used Topoisomerase-assisted cloning to insert bait sequences amplified from human or mouse genomic DNA into a universal vector backbone. This backbone enabled us to amplify the baits by parallel PCR, where one primer was labeled with desthiobiotin to allow streptavidin capture and specific elution of protein-DNA complexes (Figure 7A). Our interaction map was generated in the context of the R1/E mouse embryonic stem cell line, in keeping with the proposed relevance of UCEs in gene regulation during development, and exploiting the sequence identity of UCEs between mouse and human genomes.

We performed two experiments for each DNA bait of interest. In one set of pulldowns (called "forward"), we incubated heavy-labeled nuclear extracts with the UCE bait, and unlabeled extracts with the mix of 24 random genomic sequences, to dilute out any binding sites arising by chance. SILAC enabled us to accurately quantify the enrichment of interactors of DNA bait over control [86]. In the "reverse" pulldowns, we switched the SILAC labels with respect to the baits, enabling two-dimensional separation of true interactors from false positives [98] (Figure 7A).

Our screen identified a total of 1,709 proteins across the entire interactome, with an average of 870 proteins per MS run. Of these, 223 (13%) were quantified on all UCE baits, and 660 (39%) were quantified in at least half of the baits. We found 425 proteins with enrichment ratio greater than 1.4 for at least three baits (Figure 7C). These proteins represented 10.3% of the R1/E nuclear proteome which we measured for comparison, and showed a slight bias of 2.8 fold towards high-abundance proteins over the 10,000-fold abundance range ($P < 10^{-16}$, Figure 7D) – arguing that endogenous proteins of most expression levels were accessible from our screen. There was excellent reproducibility of

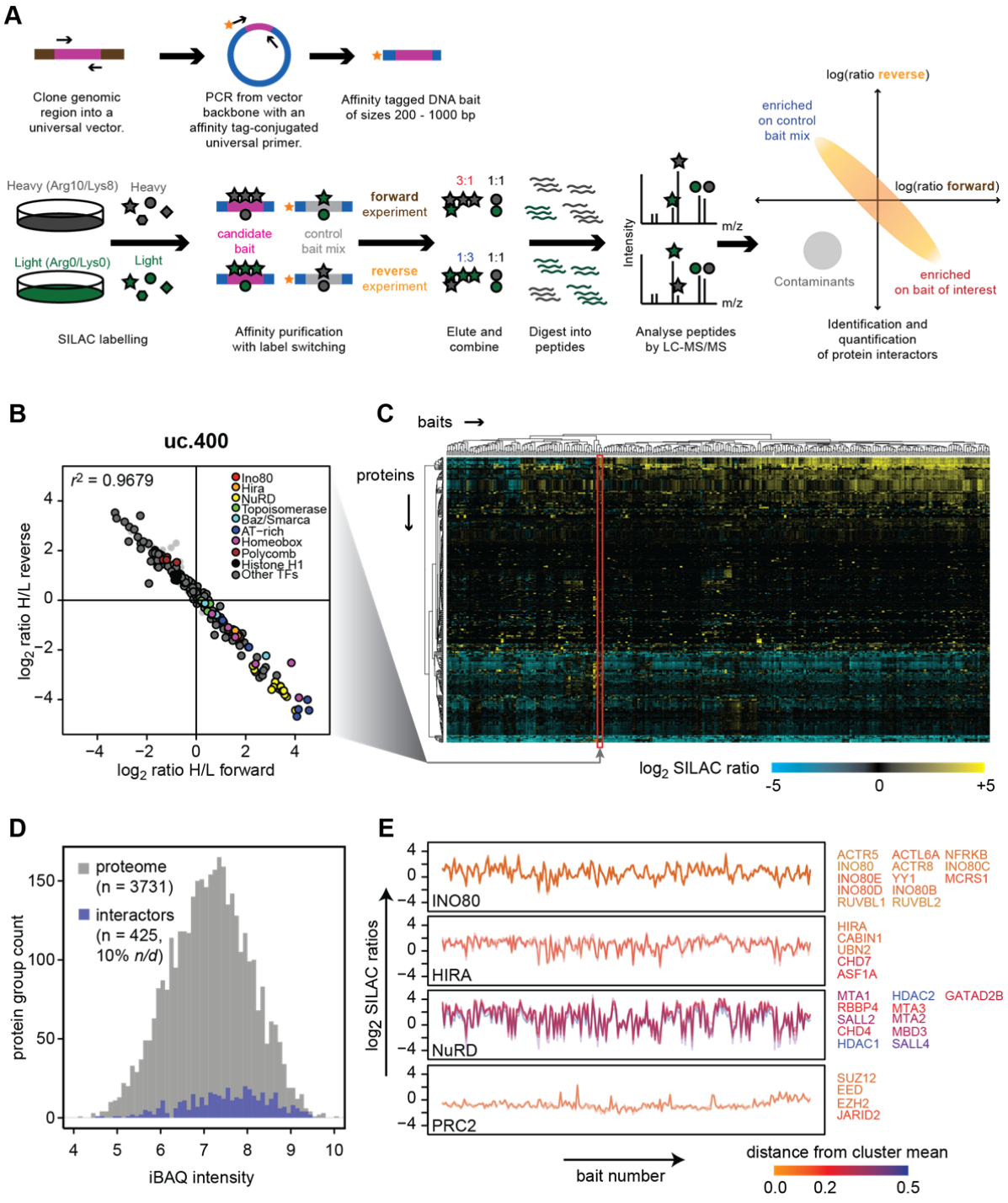

Figure 7: Overview of SILAC affinity purification for protein-DNA interaction screen for UCE sequences.

A Genomic loci of interest spanning between 200 bp and 1,000 bp were amplified using specific PCR primers and cloned into a universal vector. DNA baits were then generated by PCR amplification using affinity tagged primers binding to the flanking sequence on the universal vector backbone. SILAC AP-MS is performed as described in section 1.4.6. **B** Forward-reverse scatter plot of SILAC ratios for uc400 interactors. Colored points indicate proteins belonging to annotated complexes or protein families. **C** Summary interaction map of all 216 DNA baits and 425 interactors. Color bar indicates SILAC ratios of proteins bound to each bait over the mix of 24 random genomic loci. Missing values were filled with zeros for visualization only and not for any analysis. **D** Distribution of iBAQ intensity for R1/E nuclear proteins and for the proteins identified as interactors in our screen. *n/d* denotes the proportion of the 425 identified interactors that were not identified in the nuclear proteome. **E** SILAC ratios of members of complexes inferred *ab initio* from highly correlated interaction profiles. Each color trace represents one protein. Trace colors indicate mean absolute SILAC ratio difference from the average profile of the complex. The protein names are given to the right of each complex profile, colored as in the traces.

SILAC ratios between the forward and reverse pulldowns (Figure 7B, median SILAC ratio $r^2 = 0.91$). Binding profiles of members belonging to the same complex were extremely tightly correlated (Figure 7E), indicating that the proteins bound to the baits as complexes and providing further positive control. In sum, we have generated an unbiased intrinsic protein interactome for UCE sequences that preserves cell-specific protein-protein interactions and takes into account the cell's nuclear context.

3.2.2 Interactors of non-exonic UCEs are enriched for development and chromatin access function

Previous *in silico* sequence analysis of UCEs proposed a role of transcriptional regulatory "hubs" that recruit developmentally functional TFs [110]. Our UCE interactome showed that non-exonic UCE sequences (nxUCE) were more enriched in interactors regardless of SILAC ratio threshold used for interactor calling, followed by possibly-exonic (pxUCE), exonic (exUCE) and random genomic sequences (Figure 8A). Annotation enrichment analysis based on SILAC ratios identified Gene Ontology terms containing the annotations neural, nerve, forebrain, hindbrain, limb and axis as significant classifications for UCE interactors (Figure 8B). Domain enrichment analysis based on Pfam showed homeobox TFs were most significantly enriched at nxUCEs ($P < 10^{-31}$), and to a lesser extent, at pxUCEs ($P < 10^{-12}$) and exUCEs ($P < 0.01$) (Figure 8C). Interestingly, we also found enrichment of leucine zipper family TFs at nxUCEs ($P < 10^{-4}$), a finding not previously predicted from motif analysis based on the JASPAR TF binding motif database.

The TF binding hub proposal demands that the chromatin be accessible for function. Intrinsic open chromatin propensity for UCE sequences could be expected owing to AT-richness predicted to result in in poor nucleosome occupancy [18]. Indeed, in addition to homeobox TFs, nxUCEs also favored binding of several chromatin remodelers and other AT-rich factors including the

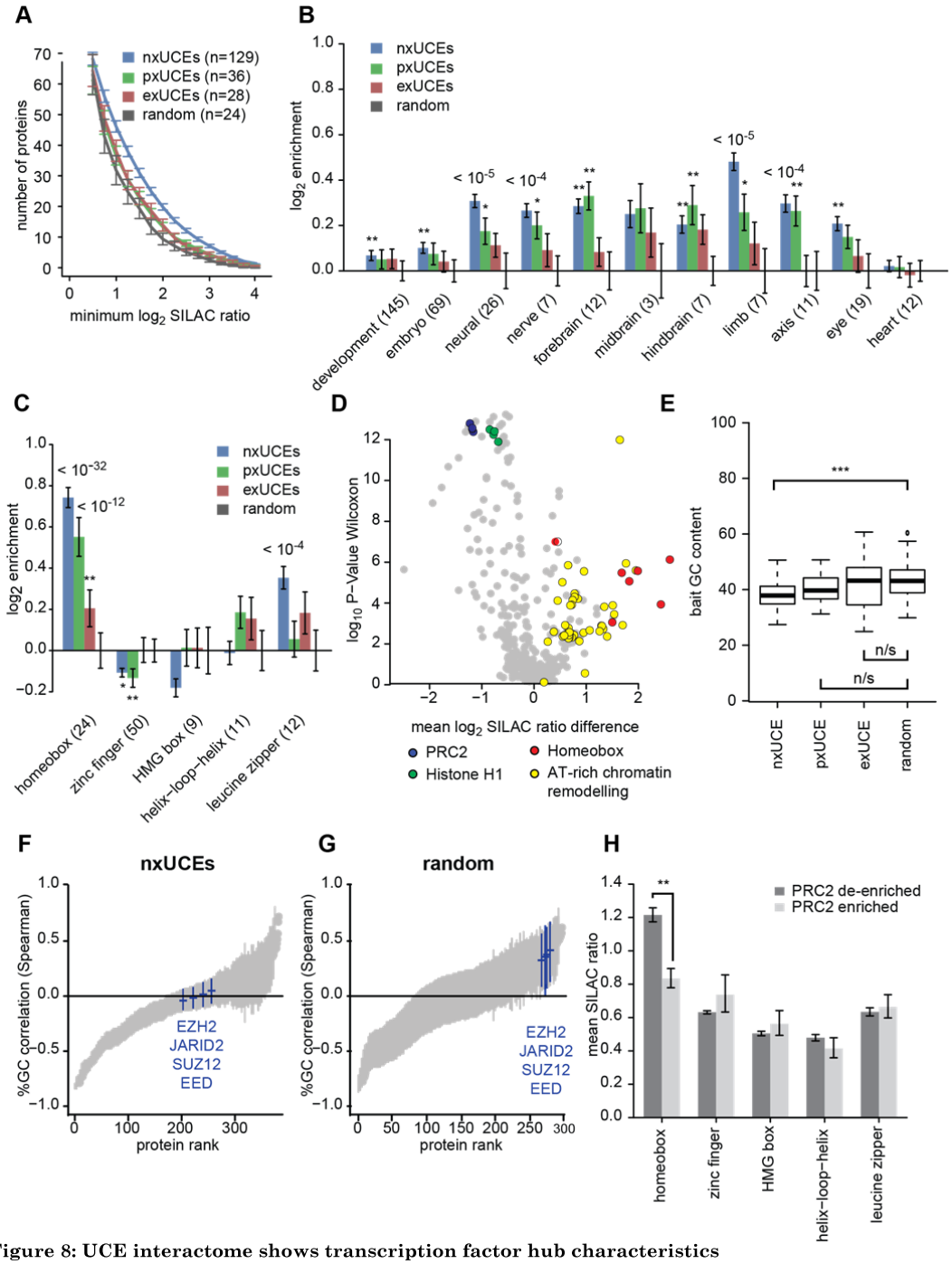

Figure 8: UCE interactome shows transcription factor hub characteristics

A Number of proteins quantified with SILAC ratios greater than those indicated on the *x*-axis, summarized for nxUCEs, pxUCEs, exUCEs and random genomic baits (mean +/- SEM). **B-C** Enrichment of proteins containing GO terms (**B**) or TF classes (**C**) indicated on the x-axis for nxUCEs, pxUCEs, and exUCEs compared to random genomic baits (mean +/- SEM). The numbers of proteins containing indicated GO words are given in parentheses. Significance is indicated (*: P < 0.05, **: P < 0.01, ***: P < 0.001). More significant P values are displayed explicitly. **D** Volcano plot of interactors enriched in nxUCE compared against random genomic loci. Enrichment significance (Wilcoxon rank sum test) is plotted against the mean enrichment. Colored points indicate proteins belonging to the annotated complexes or groups of proteins. **E** GC-content of nxUCE, pxUCE, exUCE and random genomic baits. **F–G** Spearman rank correlation coefficients of SILAC ratio with bait GC content for each interactor given the context of nxUCEs or random genomic loci (estimate +/- 95% CI). Only interactors with at least 20 quantifications were considered. Colored bars indicate proteins belonging to the PRC2 complex. See also **Figure 9**. **H** SILAC ratios TF classes for PRC2-enriched and PRC2-deenriched nxUCE baits (mean +/- SEM).

INO80, NuRD, HIRA, SMARCA/BAZ complexes as well as DNA topoisomerases (Figure 8D). Many of the chromatin remodelers observed in our interactome possess nucleosome shifting or destabilization activity [116-120]. Importantly, although nxUCEs are slightly more AT-rich than random genomic loci (median GC content 37.9% and 43.1%, respectively, Figure 8E), preferential enrichment of nxUCEs for AT-rich binders including homeoboxes generally held significant even when we binned our baits by comparable GC content (Figure 9), indicating that the observed enrichment cannot be explained solely by sequence nucleotide composition.

To further explore possible manifestation of intrinsic open chromatin propensity, we investigated the binding of histone H1 and the PRC2 complex, proteins known to promote heterochromatin formation [21, 27, 121]. Indeed, nxUCEs were relatively depleted in histone H1 and PRC2 complex ($P < 10^{-12}$, Figure 8D), and this effect was equally strong in pxUCEs and exUCEs ($P < 10^{-14}$ and $P < 10^{-6}$ respectively, Figure 9). PRC2 binding is known to depend partially on TFBS density, with absence of TFBS allowing PRC2 to bind to GC-rich regions [122]. Strikingly, we found that PRC2 members were among the interactors with strongest GC preference, but only if random genomic sequences were considered on their own. At nxUCE sequences where interactions were more prevailing, the binding of PRC2 showed no GC preference at all ($P < 0.05$ for SUZ12, $P < 0.01$ for EZH2, EED and JARID2; see also Figure 8F, G and Figure 9), indicating that a different rule than GC content governs binding of PRC2 to nxUCE sequences. Furthermore, we found that the homeobox class of interactors – the class most enriched for nxUCEs – is significantly depleted at PRC2-enriched nxUCE baits over PRC2 de-enriched nxUCE baits ($P < 0.01$, Figure 8H). The differential enrichment became even more significant when the comparison was extended to all the baits ($P < 10^{-5}$). These results demonstrate the inverse relationship between binding of TFs and binding of PRC2 in the context of UCE sequences, and suggest that nxUCE sequences may avoid heterochromatinization in part by exclusion of PRC2 owing to a large population of interactors.

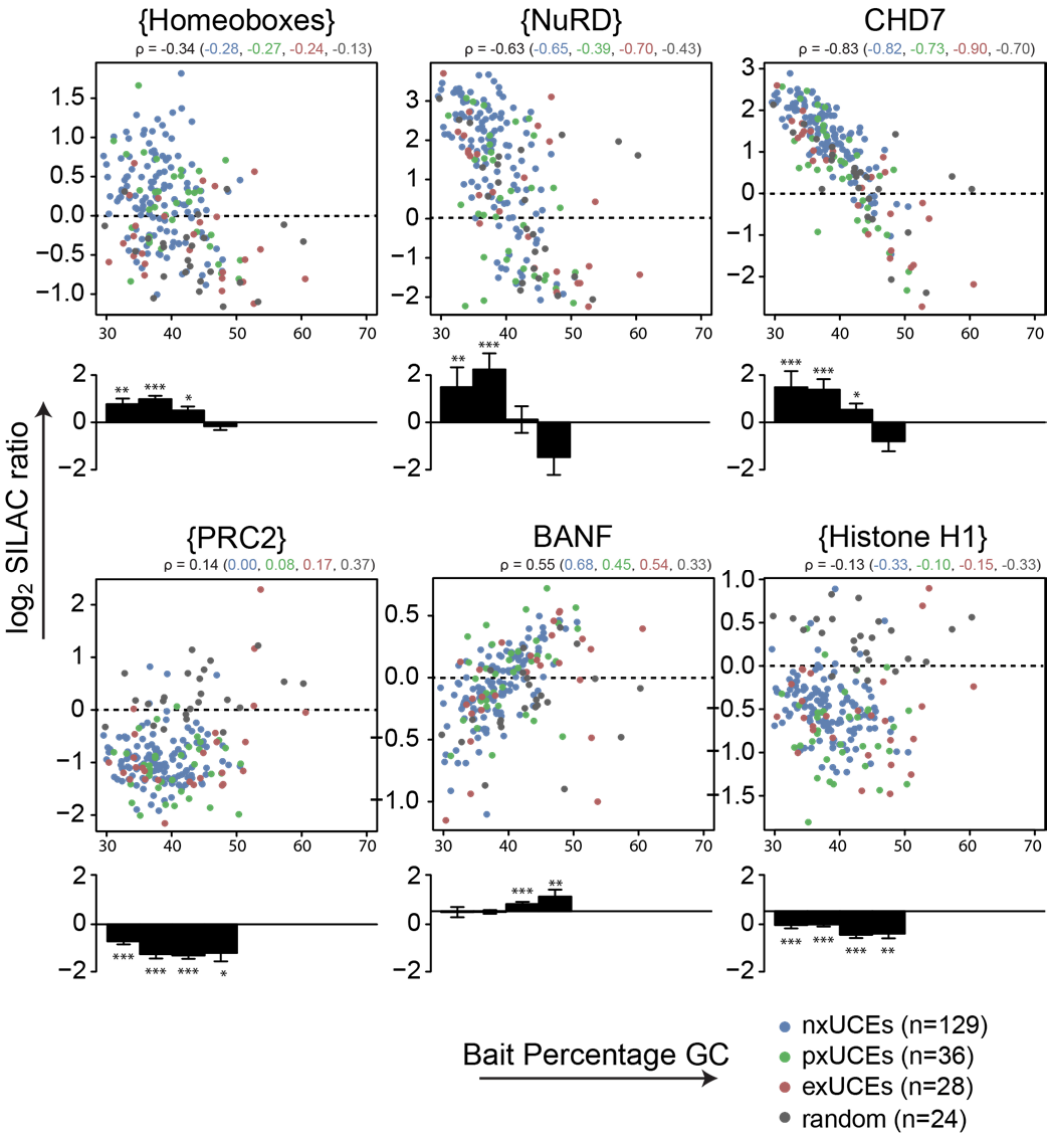

Figure 9: Enrichment/Depletion of AT-/GC-rich binding proteins at UCEs and random genomic loci.

SILAC ratios of proteins or groups of proteins (homeobox transcription factors, NuRD complex, PRC2 complex, and histones H1) that display strong AT-rich (top panels) or GC-rich (bottom panels) sequence preferences are displayed against the bait GC content, colored as nxUCEs, pxUCEs, exUCEs or random genomic loci. The bar plot underneath each scatter plot summarizes the SILAC ratio difference between each nxUCEs and random genomic loci in 5% GC content bins (mean +/- SEM). Significance in ratio means is indicated (*: $P < 0.05$, **: $P < 0.01$, ***:

In conclusion, we have shown that nxUCEs are not only enriched in developmentally relevant TFs, but are also enriched in chromatin destabilization proteins as well as relatively devoid of heterochromatin-promoting proteins. These observations illustrate the inherent biochemical properties of nxUCE sequences appropriate to serve as TF binding hubs.

3.2.3 UCEs are strongly enriched in overlapping TFBSs with conservation bias in overlapped sites.

One proposed explanation for ultraconservation of UCEs is that of high density of functional TFBSs providing multiple constraints accounting for higher evolutionary pressure. High density of TFBSs could result in information compression in the form of overlapping TFBSs, a concept that has been postulated for UCEs and indeed observed in several other instances [110, 123, 124]. Our dataset provided an opportunity to address the multiple-constraint hypothesis directly.

We first used our quantitative UCE interactome to derive binding motifs that are directly relevant to UCEs. We tested for association between differential interactor enrichment and all possible motifs up to 8 nucleotides in length, and found 439 motifs associated with enrichment of 161 interactors at 5% FDR. These included a large number of homeobox, E-box, and leucine zipper, and several other motifs, as well as a number of putative motifs for several factors (see Experimental Procedures). We also correctly found very short motifs for a number of factors. For instance, we identified the CpG dinucleotide as a binding motif for KDM2B ($P < 10^{-17}$), a H3K36 demethylase known to bind to unmethylated CpG at c-*jun* promoter through its CxxC zinc finger [125]. Binding of TFAP2 can be described by presence of a single-nucleotide motif "G", reflecting the GC content as the major influence on the interaction. As a measurement of validity of our motif enrichment, Table 1 compares some of the most significant motifs rediscovered *ab initio* from our dataset to the corresponding known motifs.

To test the overlapping TFBS hypothesis and its relevance for ultraconservation, we mapped the derived motifs to UCE sequences and other sequences, and then compared motif distribution as well as conservation of unmapped bases, singly-mapped bases, and repeatedly-mapped (superimposed) bases (Figure 10A, see also Experimental Procedures). To allow an exhaustive analysis, we included all 481 UCEs, 720 additional enhancers available from the VISTA database of *in vivo* enhancer activity of conserved genomic loci [106] classed by whether they

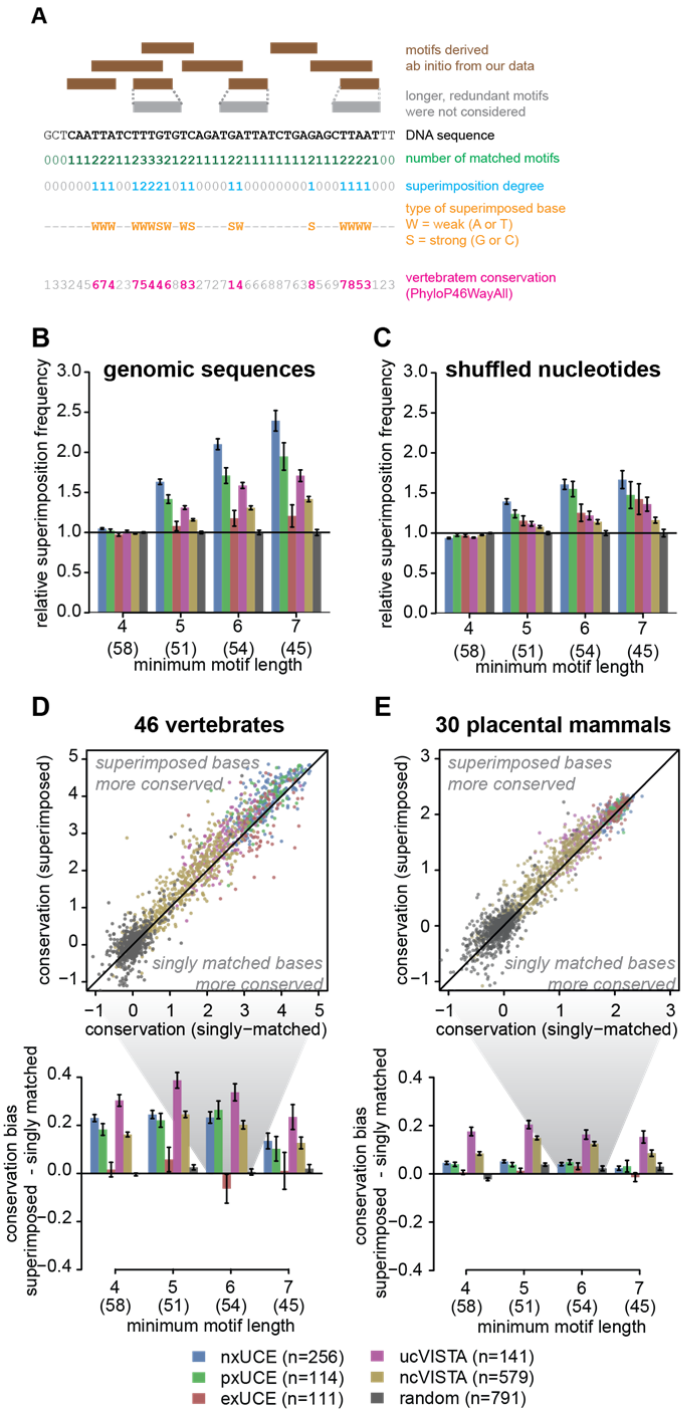

Figure 10: frequency and ultraconservation of overlapping TFBSs at nxUCEs.

A Outline of superimposition analysis. Motifs derived *ab initio* from our analysis were mapped back onto DNA sequences. First, a minimum motif length was decided, and longer motifs containing an existing shorter motif were discarded from mapping to avoid counting redundant motifs. The number of motifs mapped onto each base was counted, and bases were then classed as unmatched, singly-matched or superimposed. Frequencies of each base class and their conservation were then compared. See Experimental Procedures for more information. **B** Relative fraction of superimposed bases given indicated minimum motif length for UCEs and VISTA enhancers, normalized to random genomic loci (mean +/- SEM). Numbers in parentheses indicate the number of motifs considered given each minimum motif length. See text for *P*-values. See also Figure S3. **C** Relative fraction of superimposed bases for shuffled versions of UCEs and VISTA enhancers (mean +/- SEM). Numbers in parentheses indicate the number of motifs considered given each minimum motif length. See text for *P*-values. **D–E** Conservation bias of superimposed bases over singly-matched bases for A and T bases calculated over 46 vertebrates or 30 placental mammals (mean +/- SEM). Outset: Absolute mean conservation scores of singly matched and superimposed

contain UCEs (ucVISTA) or not (ncVISTA), and 791 randomly picked genomic regions.

We found nxUCEs to be most highly enriched for motif superimposition over random genomic loci ($P < 10^{-48}$), followed by pxUCEs ($P < 10^{-11}$, Figure 10B) but

not exUCEs. Similarly, ucVISTA sequences were more enriched for superimposition over random genomic loci than ncVISTA ($P < 10^{-25}$ and $P < 10^{-16}$, respectively) but less enriched than nxUCEs, consistent with UCEs being the most conserved core of ucVISTA enhancers. No superimposition enrichment was observed when we instead used non-enriched motifs taken randomly from the UCEs. Our finding that superimposition degree increases from ncVISTA to ucVISTA and finally nxUCE, and that exUCEs did not show such enrichment, indicate that nxUCEs represent the extreme case of overlapping TFBSs.

To exclude the possibility that AT-richness is solely responsible for the increased motif superimposition at nxUCEs, we shuffled the nucleotides in all sequences used for superimposition analysis to generate synthetic sequences of equivalent GC content. Superimposition enrichment on these sequences was severely abrogated (**Figure 10**c), indicating that AT-richness contributes to but is in itself insufficient to achieve the extent of superimposition observed with nxUCEs by chance. To support this *in silico* finding, we performed pulldowns on random, highly heterogeneous DNA sequences with average GC content of 20% or 40%. Our experiment showed that only some of the proteins that bound preferentially to UCEs also bound preferentially to the synthetic AT-rich bait population (**Figure 11**). Generally, there was insignificant correlation between factor preference for AT-rich sequences and enrichment at nxUCEs (Spearman's $\rho = 0.05$, $P > 0.1$). Notably, factors bound to synthetic GC-rich bait populations were also enriched at nxUCEs, ruling out AT-richness as the sole explanation for motif occurrence and thus superimposition at nxUCEs. Together with the inherent conservation bias for GC nucleotides over AT nucleotides in UCEs but not in random genomic loci (Figure 11), we speculate that GC-rich TFBSs may be under greater selective pressure in AT-rich UCEs in order to preserve certain regulatory function.

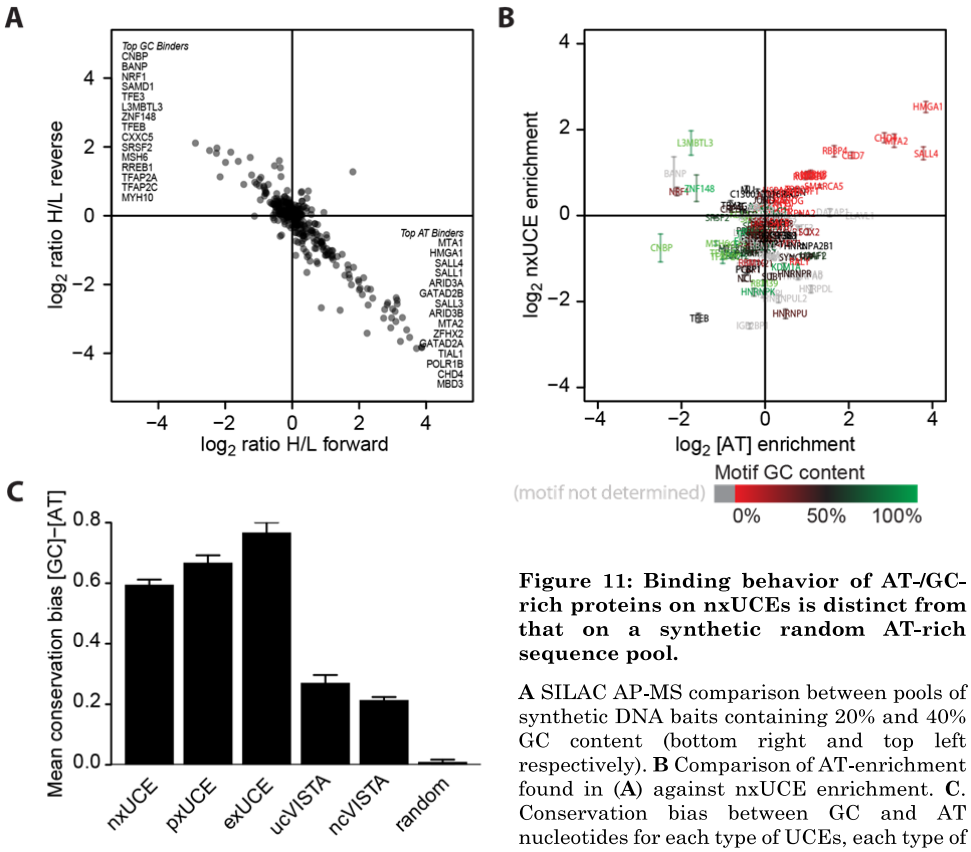

Figure 11: Binding behavior of AT-/GC-rich proteins on nxUCEs is distinct from that on a synthetic random AT-rich sequence pool.

A SILAC AP-MS comparison between pools of synthetic DNA baits containing 20% and 40% GC content (bottom right and top left respectively). **B** Comparison of AT-enrichment found in (**A**) against nxUCE enrichment. **C**. Conservation bias between GC and AT nucleotides for each type of UCEs, each type of

If superimposition of TFBSs also played important biological roles, we would expect DNA bases involved in superimposition to be more deeply conserved. We therefore investigated the extent of DNA base conservation in 46 vertebrates, using an established conservation scoring scheme [126]. For sequences that were putative enhancers, the bases matched by multiple motifs were on average slightly but significantly more conserved than bases mapped only to a single motif ($P < 0.001$). Strikingly, this conservation bias became massively amplified when only AT bases were considered ($P < 10^{-10}$ Figure 10D), consistent with the presence of many AT-rich motifs derived from our data. Conservation bias was also observed in ucVISTA and ncVISTA sequences, concordant with functional overlapping TFBSs reported for loci other than UCEs. The larger difference in VISTA enhancers compared to UCEs can be attributed to the lower conservation baseline for ncVISTA enhancers (Figure 10D). We also found the conservation bias to be reduced when the scoring was restricted to placental mammals (Figure 10E), suggesting early origins of these overlapped sites. In conclusion,

we have shown that nxUCEs represent the extreme case of overlapping, deeply conserved, biochemically functional TFBSs among enhancers.

3.2.4 UCE scanning mutagenesis defines protein binding characteristics and correlates gain of interaction with nucleotide conservation

Although an implication of the multiple-constraint hypothesis is that mutation of nxUCEs causes deleterious consequences, it has been difficult to identify the exact systems that are affected. However, the conservation bias implies that multiple-constraint hypothesis would at least manifest itself in terms of change in protein binding capacity, which in turn could result in regulatory logic alteration at UCEs.

In order to test this hypothesis, we performed a scanning mutagenesis of uc325, a non-exonic UCE that is part of a midbrain/eye development enhancer [106]. Each non-overlapping 5-nucleotide window of uc325 was mutated transitionally – the most frequent mode of nucleotide substitution *in vivo* [127]. Pulldown was performed on the resultant series of baits against the wildtype bait (Figure 12A), and interactors were defined as proteins whose SILAC ratios were in most extreme 5% of all quantified ratios. We discovered 55 interactors for the uc325 set but only 10 for the control set based on a random genomic sequence with comparable GC content. Both gain and loss of interactions were found for uc325, covering the entire span of the bait (Figure 12C). Most of the prominent interaction losses were found in contiguous variants – reflecting binding sites that span more than five nucleotides – whereas interaction gains tend to appear stochastically ($P < 10^{-5}$, Kolomogorov-Smirnov test, Figure 12B). In contrast, only a small region in the control bait appeared to contain prominent interactors (Figure 12B, C). These data indicate that uc325 indeed possesses a hub-like characteristic with numerous and diverse TFBSs as well as latent sites that could be reached within a few transition mutations.

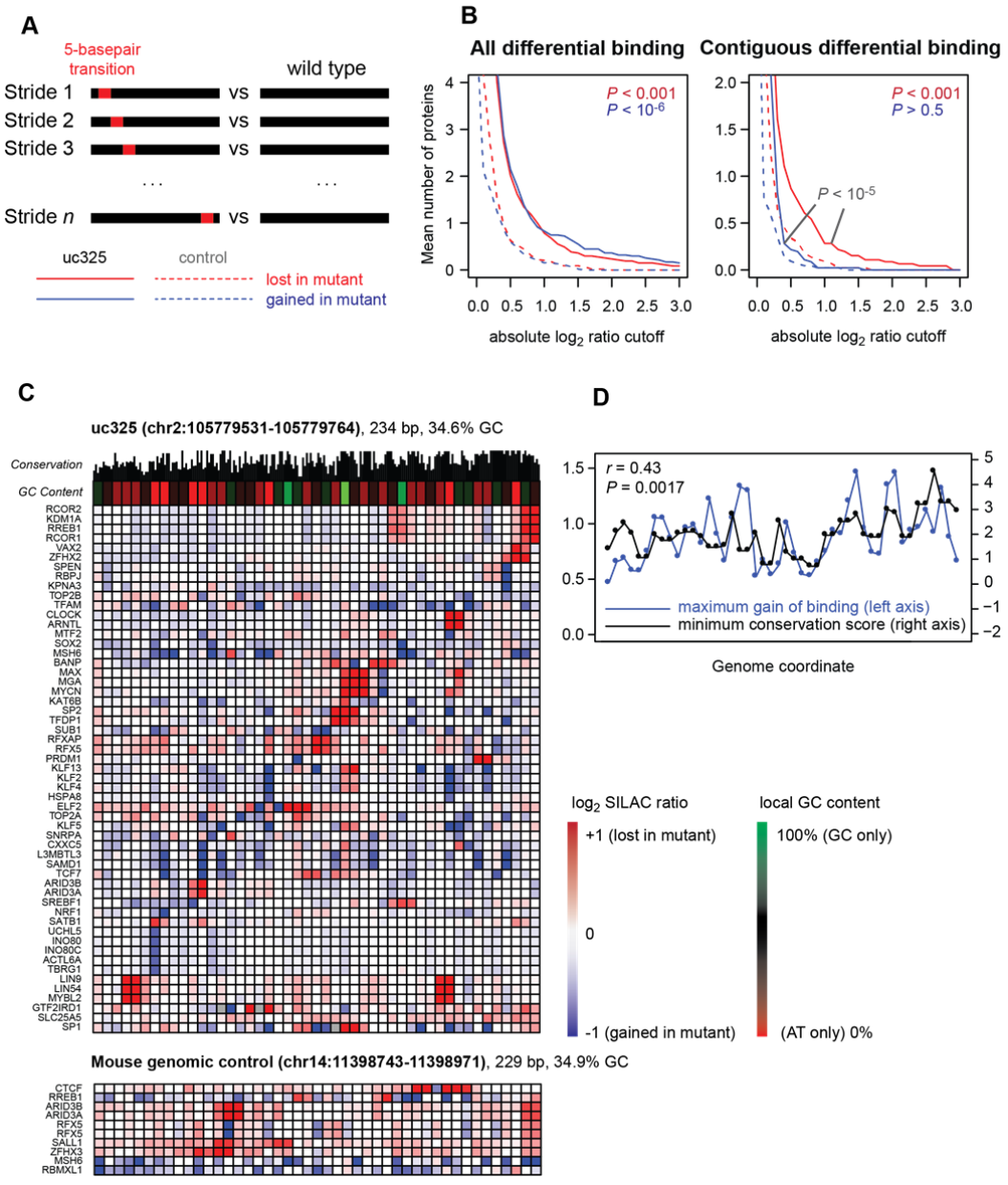

Figure 12: a fine intrinsic interaction map of uc325.

A Non-overlapping 5 nucleotide windows spanning the 234 bases of uc325 were mutated transitionally (i.e. A ↔ G, C ↔ T) and interactors compared to the wildtype in a series of SILAC pulldowns. The same was done for a 229-base random genomic sequence with comparable GC content. **B** Proportion of variant strides that show either loss or gain of binding of at least one protein owing to the mutation for uc325 and control, given a as a function of cutoff ratio used for calling gain/loss. Contiguous differential binding refers to differential binding that spans at least 2 strides. **C** Enrichment of proteins where complete pairwise quantification was achieved for at least 80% of all strides, and where at least one stride showed localized differential enrichment with magnitude exceeding 95% of all quantified ratios. This 95% cutoff corresponded to log2 ratio of 0.59 for uc325 and 0.83 for the control. Interactors were ordered by the location of a prominent differential binding (magnitude exceeding 99% of ratios across all strides). Genome coordinates are based on the mm10 build. Conservation is based on 60-way vertebrate comparison.

D Comparison of maximum magnitude of binding gain/loss of each 5-nucleotide block to the minimum nucleotide conservation smoothened over two successive blocks, giving an effective resolution of 10 nucleotides.

We next investigated whether any relationship exists between uc325 conservation and its scanning mutant interactome. Initially, we had expected the conservation to be correlated to the loss of binding owing to transition mutation, but this turned out not to be the case ($P > 0.5$). Surprisingly, we found conservation of uc325 strides to be significantly correlated with the maximum binding gain owing to mutation ($P = 0.0017$, Figure 12D), whereas such correlation was weaker for the control ($P = 0.027$). When at least two non-correlating proteins were required to be enriched in the mutant, the correlation with conservation remained significant for uc325 ($P = 0.0020$) but not for the control ($P = 0.12$). Interestingly, AT-rich strides tended to give more drastic binding gain upon mutation (correlation with GC content = -0.36, $P = 0.0062$, Figure 12C). We speculate that these AT-rich strides are under selective pressure against developing such TFBSs which could alter the regulatory logic of the UCE. Alternatively, apparent strong gain of binding could be observed if the mutation turned a promiscuous binding site capable of binding several factors weakly into a well-defined, specialized binding site, thereby destroying the 'hub' characteristics which may be required for fine-tuned regulatory function.

3.2.5 Regulatory consequence of the UCE interactome

Evidence for regulatory consequence of UCE interactors could be obtained from perturbation experiments and reporter assays. While it may be difficult to discern the regulatory logic of such complex enhancers without performing very deep perturbation, it should still be possible to address functionality of certain interactions given existing biological knowledge. To demonstrate such a case, we investigated the functionality of the interaction between uc400 and the protein GTF2IRD1.

The 860 bp genomic region containing uc400 possesses forebrain-specific enhancer activity during embryonic day E11.5 (Pennachio et al., 2006). We found that uc400 interacts specifically with the Williams Beuren syndrome protein GTF2IRD1 with a SILAC ratio of around 6:1 in R1/E cells, and also with hGTF2IRD1 in HeLa cells (Figure 13). GTF2IRD1 is known to act as a repressor

via its interaction with the conserved DNA motif containing the core sequence GATTA [115]. Consistently, our motif analysis rediscovered GATTA as a binding motif for GTF2IRD1 (Table 1), which is present in three copies in uc400. GTF2IRD1is expressed ubiquitously with the exclusion of the forebrain during E10.5 [114], a finding in agreement with the forebrain-specific activity of uc400, the role of Gtf2ird1 as a repressor, and our interaction data. Given the degree of corroboration between existing literature and our data, we decided to investigate possible regulatory modulation of uc400 by hGTF2IRD1.

We first confirmed that hGTF2IRD1 bound to uc400 via the GATTA motif, by mutating all occurrences of such motifs to GAGGA. MS-analysis showed hGTF2IRD1 to be the only DNA binding protein bound preferentially to the wildtype uc400 bait compared to the mutant bait (Figure 13B). Interestingly, the data immediately revealed that the mutant uc400 had also gained specific binding of another TF, namely hTEAD1. We then performed reporter assays using wildtype or mutant uc400 as an enhancer driving luciferase reporter, under non-targeting condition or GTF2IRD1-knockdown. Owing to auto-regulation of Gtf2ird1 [128], we also monitored mRNA expression levels together with luciferase reporter activity over a time course (Figure 13D). We found that hGTF2IRD1 knockdown resulted in differential reporter activity modulation of wildtype uc400 relative to the mutant uc400. Because our mutagenesis of uc400 reporter resulted in gain of hTEAD1 binding site (Figure 13B), we also excluded indirect effects of hGTF2IRD1 knockdown on reporter activity through hTEAD1 by showing that its mRNA expression level was only modestly affected throughout the course of the experiment (Figure 13). In conclusion, we have demonstrated regulatory consequence of the interaction between uc400 and the hGTF2IRD1 protein.

To further explore the regulatory relevance of UCE interactors in cellular contexts more globally, we compared our interaction data with existing ChIP-seq data from the ENCODE consortium [101]. We found 12 TFs from our screen with corresponding ChIP-seq data obtained from the H1 human embryonic cell line, giving rise to 31 *cis-trans* interaction pairs relevant to our loci of interest.

ChIP-seq measures if a TF is present at a genomic locus, therefore if there is a signal

Figure 5

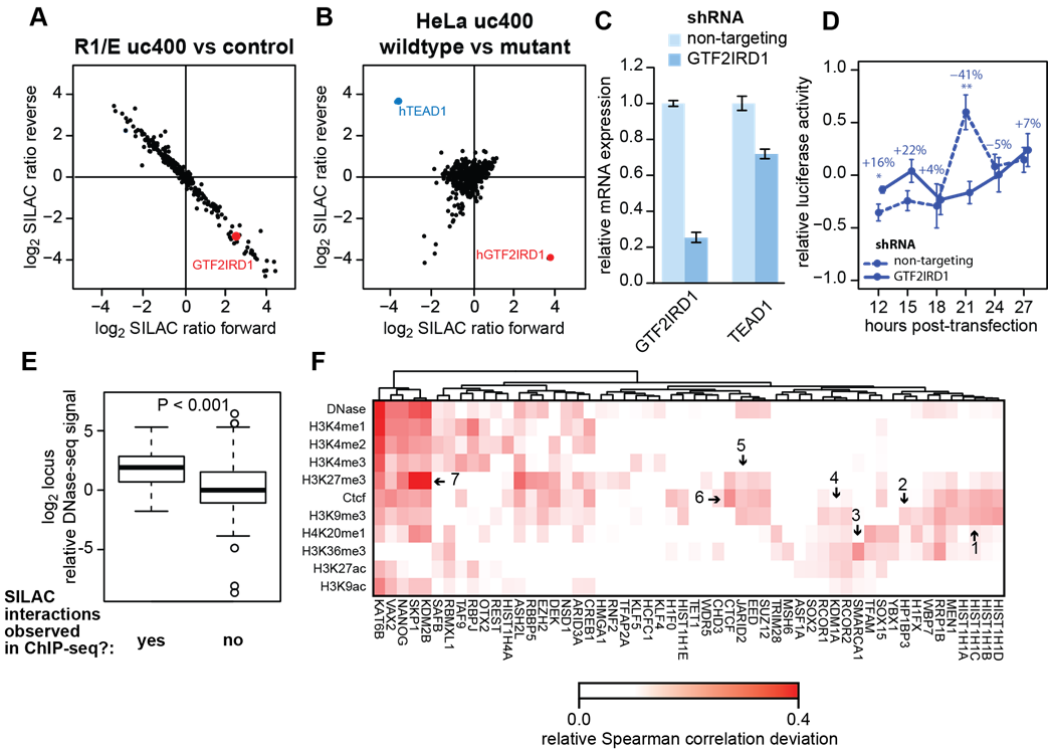

Figure 13: Regulatory consequence of UCE interactions

A Interaction of uc400 with TF GTF2IRD1 in R1/E compared to random genomic loci. **B** Disruption of uc400 interaction with GTF2IRD1 in a GATTA → GAGGA mutant form uc400 compared to wildtype. **C** Relative mRNA expression levels of GTF2IRD1 and TEAD1 at 12 hours post transfection for non-targeting and GTF2IRD1-knockdown conditions (mean +/- SEM). **D** Luciferase reporter activity of wildtype uc400 normalized to that of the GATTA → GAGGA variant at (mean +/- SEM). Significance levels *: $P < 0.05$, **: $P < 0.01$. **E** Distribution of relative DNase-seq signal for the baits containing ChIP-seq interaction congruent to the SILAC interactome, compared to the baits where no ChIP-seq signal was detected.

F Heat map of correlation deviation for all identified chromatin proteins with *in vivo* chromatin states

in ChIP-seq and a pulldown experiment has been performed on the sequence, then we should have also identified the factor by mass spectrometry. This was indeed true in 90% of the cases. Although we do not expect the strength of a ChIP-seq signal to directly correlate with the MS measurements – because of the different nature of the experiments – in 65% of the cases (20 interactions) the SILAC ratios indicated clear enrichment over random genomic sequences. In a few cases, the SILAC ratios loosely correlated with the ChIP-seq scores. We also found a highly significant tendency for loci with congruent interactions to have

more accessible chromatin than the remaining loci, as deduced by DNaseI hypersensitivity signal (Figure 13E). This suggests that open chromatin has an influence on observing intrinsic interactions in the cell. Overall, the available ChIP-seq data validate the relevance of our UCE interactome in a native genomic context.

Regulatory relevance of our interactome in cellular context should also be reflected in cellular chromatin states associated with enhancer and repressor activity. We therefore correlated our SILAC profiles with several histone methylation and acetylation ChIP-seq tracks as well with as the DNaseI hypersensitivity track. Initial analysis of H1-hESC ChIP-/DNase-seq data obtained from ENCODE revealed that, regardless of the track under consideration, proteins whose SILAC ratios most strongly correlated with the ChIP-/DNase-seq signal were those with strong GC-content preference. To correct for this known bias of ChIP-seq datasets [129], we report association between SILAC profiles and ChIP-/DNA-seq profiles in terms of deviation from correlation expected of the interactor's GC preference. We validated our analysis by comparing SILAC profiles to the CTCF ChIP-seq track, and indeed found the SILAC profile of CTCF to be most strongly associated with its own binding in H1-hESCs (Figure 13F, arrow 6).

The analysis recovered several known relationships between intrinsic interactors and cellular chromatin states at corresponding loci. For example, the PRC1 complex was most strongly correlated with the classical Polycomb mark H3K27me3, but also to a lesser extent with the enhancer marks (Figure 13F, arrow 6), a finding in line with the bivalent nature of H3K27 methylation and H3K4 methylation [50, 130]. In contrast, no correlation was observed for the PRC1 complex with H3K27ac, a mark which counteracts Polycomb silencing [131, 132]. Table S4 summarizes the full set of associations between our interaction data and chromatin data along with functional interpretation. These associations indicate that proteins involved in chromatin modification pathways already bind even in initial absence of epigenetic priming. Taken together, our analyses demonstrate the regulatory relevance of our interactome by illustrating

congruence between cell type-specific intrinsic interaction at UCEs and *in cellulo* chromatin modification states.

3.2.6 The UCE interactome is determined by the cellular context

It is conceivable for DNA sequences of high regulatory information density such as UCEs that regulation is cell-type specific. Such variation in regulatory logic should reflect itself in change in interactions. To explore this, we also obtained interaction data for a subsample of UCEs in the HeLa cell context. Comparison between the two datasets revealed that homologous interactors with high sequence identity between mouse and human are more likely to have highly correlated binding. Examples of such homolog pairs include CHD7, TFAP4 and RCOR1 (Figure 14A). However, many highly identical homolog pairs also behave

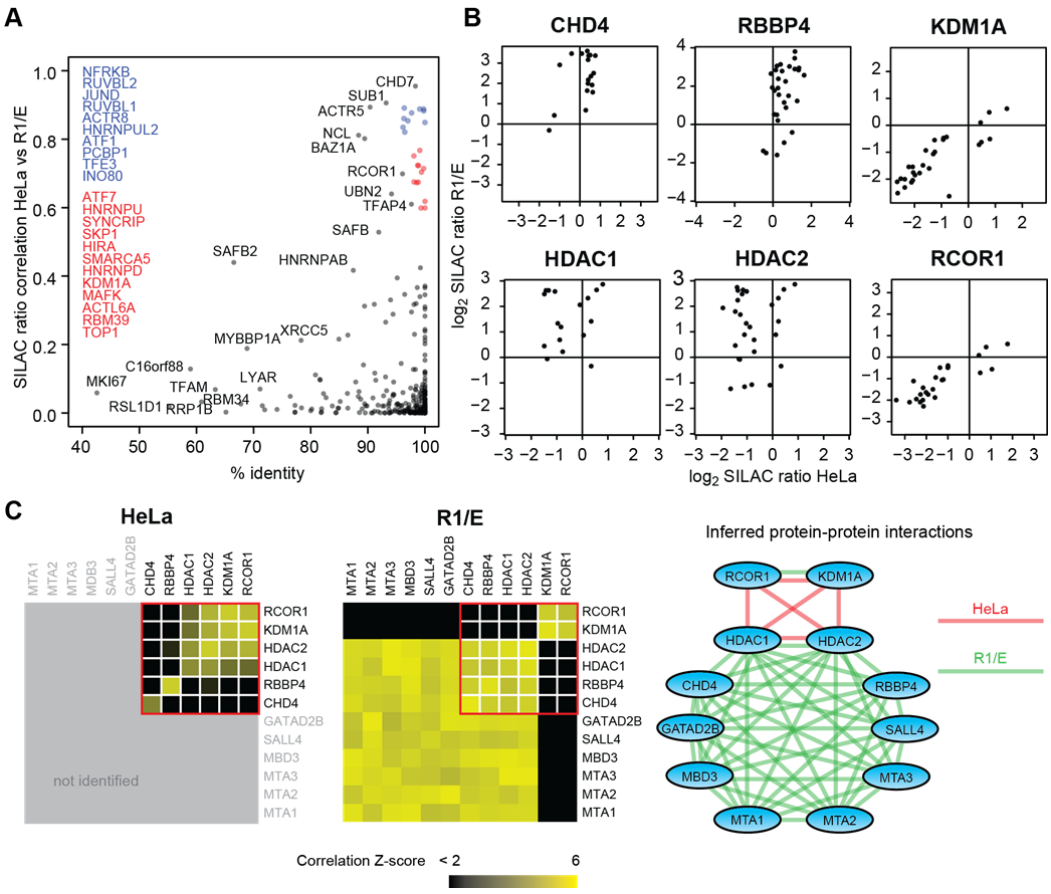

Figure 14: Comparison of UCE interactomes obtained in HeLa and R1/E backgrounds

A Scatterplot showing SILAC ratio correlation between R1/E and HeLa datasets against human-mouse protein sequence identity. Names of the proteins represented by colored points are given on the left.

B Example profiles of proteins with high human-mouse sequence identity. **C** Comparison of protein-protein interactions deduced from profile correlation (see also Figure 1), showing members of the REST co-repressor complex and the NuRD complex, and the switch of complex membership of HDAC1/Hdac1 and HDAC2/Hdac2.

differently between cell lines, indicating effects of cellular context upon intrinsic interaction with our baits (Figure 14A, B). For example, by using profile correlation across baits as a measure for complex organization (Figure 7E), we found that the proteins HDAC2 and HDAC1 bound to our baits in differing contexts: as part of the REST co-repressor complex in the HeLa background, and as part of the NuRD complex in the R1/E background (Figure 14C). Thus, UCE sequences are capable of recruiting different interactors based on the nuclear proteome and protein-protein interactome of the cell.

3.3 Discussion

Despite the comprehensive tabulation of enhancer activities of UCEs, the candidate interactors responsible for regulation have not been systematically characterized. While protein-centric approaches such as ChIP-seq have long allowed for global analysis of interactions of candidate proteins with the genome, a DNA-centric approach is particularly suited to answering this question. We have applied DNA-centric interaction screening to map intrinsic interactions of the sequences of hundreds of ultraconserved elements to obtain two highly information-rich datasets: the UCE interactome and the uc325 differential interactome. The exquisite quantitative accuracy of SILAC, combined with the large scale of the interactome study, allowed us to provide candidate interactors that can be used for follow-up studies of UCE regulatory logic, as well as to quantitatively address interaction tendencies of UCEs as a family of sequences – a question not previously addressable in smaller scale applications of the DNA-centric paradigm.

The analyses demonstrated that the sequences of nxUCEs represent the extreme case when compared to pxUCEs, exUCEs and random genomic sequences in many aspects of protein-DNA interactions. They were most enriched in intrinsic interactors, especially those annotated to be important in tissue specific development; they were most refractory to intrinsically GC-rich binding of the heterochromatin-promoting PRC2 complex); and they were most enriched in deeply conserved, overlapping TFBSs. The latter phenomenon is in the extreme even compared to other non-ultraconserved enhancers in the genome. While the extent to which individual interactions contribute to the regulatory output remains to be determined, we have shown that interactions are recapitulated in cells by ChIP-seq, and as a whole corroborate with observed chromatin states that reflect regulatory consequences. Furthermore, UCEs appear to bind different factors in different cellular background which can be explained in part by rewired protein-protein interaction. All these findings provide strong experimental support to the hypothesis of nxUCEs as highly-constrained transcriptional regulatory modules [43, 110].

If nxUCEs are highly information-dense regulatory circuits, it is conceivable that any mutation would result in regulatory alterations with adverse effects to the organism. This is supported by the conservation bias of overlapping TFBSs inferred from the UCE interactome and the sensitivity of uc325 to mutation with respect to gain and loss of binders. Our observation that mutating hGTF2IRD1 binding sites in uc400 results in gain of Tead1 binding further exemplifies the idea that functional binding sites can be gained spontaneously through mutation of an existing motif. Our finding that fine-resolution conservation of uc325 correlated with the tendency to gain interactors also lends possibility to the concept that UCEs are under selective pressure that not only prevents loss of regulatory function, but also its logical alteration (Figure 4). This is supported by the discovery that while many TFBSs can be functional regardless of their context with neighboring TFBSs, some TFs do indeed have a strict contextual prerequisite [133]. Context dependent binding might provide cell-type specific logic that provides further conservational constraints not yet explored in this study. Still further contribution may come from functional constraints beyond enhancer function [134-136].

We found that pxUCEs and exUCEs were less extreme in their transcriptional regulatory characteristics as indicated by their intrinsic interactions, in line with their possible functional roles beyond transcriptional regulation. We found pxUCEs to behave similarly to nxUCEs in some aspects (Figure 8B and Figure 10D), to exUCEs in others (Figure 8A), and often as an average between nxUCEs and exUCEs (Figure 8B, 2C, and Figure 10B). This raises the possibility that some of the putative exons coinciding with pxUCEs may in fact be functional exons and others may be enhancers.

There remains the general challenge that certain deletions or mutations of UCEs have failed to produce observable deleterious phenotypes [53], which can be interpreted against the high constraint hypothesis. However, this absence of evidence is not surprising, given that almost all ultraconserved enhancers remain to be systematically characterized at the regulatory level, where the context and environment under which they become indispensable need to be determined. Indeed, it is now known that some enhancers contribute to robust

regulation and are indispensable only under certain extreme conditions [137]. Full, systematic *ab initio* functional characterization of regulatory elements, including upstream events, context-dependent regulatory logic, and downstream consequences remains a daunting task. Here we have demonstrated the utility of our approach as a crucial initial step in the process and, complementary to the VISTA enhancer data which tabulated enhancer activity of UCEs, we provide their potential interactors. The use of insertional ChIP where the interaction was queried *in vivo* would be a very attractive follow-up in order to ascertain the exact cell-specificity of interactions [66]. Further integration with data obtained for *in vivo* protein-DNA interactions, protein-protein interactions, long-range DNA interactions, as well as gene expression data, reporter assays and perturbation experiments, will allow deep functional characterization of UCEs with the aim to discover their target genes and functional contexts as well as to decode their exact regulatory logic.

3.4 Experimental Procedures

3.4.1 Stem cell culture and nuclear extract preparation

R1/E cells were SILAC labeled in SILAC DMEM (PAA Laboratories) containing either 73 mg/l Lys-8 HCl and 42 mg/l Arg-10 HCl, or the same concentration of Lsy-0 HCl and Arg-0 HCl. Medium was supplemented with 10% dialyzed FBS (PAA Laboratories), 1x non-essential amino acids (Gibco Life Technologies), 1 mM sodium pyruvate (Gibco Life Technologies), 3 µM CT-99021 (Biomol GmbH), 1 µM PD-0325901 (Biomol GmbH), 50 µM 2-mercaptoethanol (Gibco Life Technologies), 100 u/ml LIF (Millipore GmbH), and penicillin-streptomycin-glutamate. Nuclear extracts were prepared as previously described [138] except for a reduced NP40 concentration of 0.5% to preserve nuclear integrity during cell lysis. Extracts were controlled for presence of Oct4 by western blot.

3.4.2 Cloning and DNA bait generation

UCEs and 24 random mouse and human genomic loci were cloned into pCR8/TOPO/TA (Life Technologies). See Table S3 for genome coordinates of the inserts. Desthiobiotin-conjugated DNA baits of size 200 bp to 1000 bp were generated by PCR using the following primers: forward 5'-desthiobiotin-CAGGCTCCGAATTCGCCCTT-3', reverse 5'-GAAAGCTGGGTCGAATTCGCC-3'. PCR products were concentrated by ethanol precipitation and purified from unincorporated primers on G-50 Sephadex columns (GE Healthcare). Baits for uc325 scanning pulldowns were produced by site-directed mutagenesis PCR.

Baits for random DNA pulldown used to generate data in Figure 11 consisted of 179 bp 5' variable sequence with 20%, 40% or 60% GC content, followed by a constant 3' sequence 5'-AAGGGCGAATTCGGAGCCTG-3'. Baits were synthesized as single stranded DNA by Metabion GmbH. To generate dsDNA bait, 100 pmol ssDNA oligo was annealed with 100 pmol desthiobiotinylated oligo complementary to the constant region (5'-desthiobiotin-CAGGCTCCGAATTCGCCCTT-3'), and extended with 25 units of Klenow exo-fragment (Fermentas), using the provided buffer and supplemented with 25 nmol each of dATP, dCTP, dGTP, and dTTP at 37°C for 1 hour. Pulldowns were

performed using 100 pmol of desthiobiotinylated bait, but otherwise as described in Experimental Procedures.

3.4.3 DNA pulldowns and mass-spectrometric analysis

DNA pulldowns and sample preparation for mass-spectrometric analysis were performed as previously described [98]. Peptides derived from the bound proteins were separated by HPLC over a 140-minute gradient from 2% to 60% acetonitrile, and analyzed in an Orbitrap Elite mass spectrometer (Thermo Fisher Scientific, Germany). Full scan MS spectra were acquired with 120,000 resolution in the Orbitrap analyzer, and up to the 10 most intense ions from each full scan were fragmented with collision induced dissociation and analyzed in the linear ion trap. Mass-spectrometric data were processed with the MaxQuant software version 1.2.6.20 [82]. The complete pulldown dataset from R1/E and the nuclear proteome dataset were searched against the mouse Uniprot database. We mapped Gene Ontology [139] and Pfam [140] annotations to protein groups using the Perseus module in the MaxQuant software suite.

3.4.4 Nuclear proteome of R1/E cells

R1/E nuclear extracts was precipitated in four volume acetone. The pellet of nuclear proteins was resuspended in 8 M urea and proteins digested in solution. Peptides were separated by HPLC over a 240-minute gradient from 2% to 60% acetonitrile, and analyzed in a Q-Exactive mass spectrometer (Thermo Fisher Scientific, Germany) [141]. Five replicates were measured to extend proteome coverage. Mass-spectrometric data were processed with MaxQuant version 1.2.6.20.

3.4.5 Reporter assays

We cloned uc.400 into a modified pGL3/Basic firefly luciferase reporter vector containing a minimum mouse heat shock promoter via the Gateway system as previously described [98]. Primers for amplifying uc.400 were: forward 5'-GCCTCTCTGAAGCGTTCATC-3', reverse 5'-TGGTGTTACGGATCACAACG-3'.

The mutant variant of uc.400 were generated by PCR using mutagenizing primers and subcloned into pCR8/TOPO vector.

Transfection and reporter assays were performed as previously described [98]. Knockdown of hGTF2IRD1 was achieved using shRNA vector generated using pSUPERIOR vector system, following the manufacturer's protocol. The shRNA core half-sequences for GTF2IRD1 and non-targeting construct were CAGAAAGACTAAAGGAAAT and GACTAGAAGGCACAGAGGGAG, respectively. Knockdown was quantified using quantitative real-time PCR and SYBR green system, using the standard $\Delta\Delta C_t$ method and normalizing over GAPDH. Primers used for qPCR were as follows: GAPDH 5'-CAAGGTCATCCATGACAACTTTG-3' and 5'-GTCCACCACCCTGTTGCTGTAG-3'; GTF2IRD1 5'-ATCATCACCAGCCTCGTGTC-3' and 5'-CACCTTCTTGGGGTGCTCT-3'; TEAD1 5'-CATGTCCTCAGCCCAGATCG and 5'-AGGCTCAAACCCTGGAATGG-3'.

3.4.6 Data analysis

Preprocessing

SILAC ratios were corrected to account for residual proteome differences between heavy and light nuclear extracts (see Extended Experimental Procedures for detail). Protein groups were then filtered for having coefficient of determination of SILAC ratios greater than 0.2 across all baits, and for having log2 SILAC ratios exceeding 0.5 in at least three baits. For subsequent analyses, we applied a Gene Ontology annotation filter, requiring the protein groups to contain at least one of these words or their variants as substring of the GO terms: chromatin, DNA, enhancer, genome, helicase, histone, nuclear, promoter, RNA, splicing, transcription, and translation.

Imputation

Where imputation was required, we filled missing logarithmized quantifications with a normal distribution with the mean equal to the minimum SILAC ratio for

each protein, and the standard deviation of 0.5. This number was empirically determined to best simulate the errors of SILAC ratios in the dataset.

Annotation enrichment analysis

We used Pfam annotation to class interactors by domain and imputed SILAC ratios were used to calculate enrichment. For JASPAR prediction [13], we used the standard Position Weight Matrix scoring procedure, normalizing the scores to the maximum value attainable for each motif.

Ab initio motif enrichment

For each k-mer motif where $1 \leq k \leq 8$ (excluding reverse complement redundancies), the median motif occurrence in both orientations was determined. DNA baits were then divided into those having less than or equal to the median occurrence of the motif ("low occurrence"), and those having greater than the median occurrence ("high occurrence"). Wilcoxon rank sum test was then used to calculate significance in difference in imputed SILAC ratios between the "high motif occurrence" and "low motif occurrence" bait sets. We used Benjamini-Hochberg false discovery rate to adjust the P-value for multiple comparisons [142].

Superimposition analysis

We chose a minimum motif length λ, where $4 \leq \lambda \leq 7$. To exclude counting the overlapping of different-length but otherwise redundant motifs, we applied two criteria for keeping a motif: (a) that the motif length was at least λ, and (b) that there existed no shorter motif that was a substring of the motif being considered or its reverse complement. Motifs only significantly associated with de-enrichment of interactors but not enrichment were not considered. Conservation data were obtained from the UCSC Genome Browser (Build hg19). Non-ultraconserved VISTA enhancer coordinates were obtained from the VISTA

database [106]. Conservation data were obtained from the phylop46wayAll and phylop46wayPlecantal tracks of hg19 respectively [126].

ENCODE dataset integration

Broad histone ChIP-seq signal for histone modifications and peaks for TFBSs were obtained from the ENCODE histone ChIP-Seq or DNase-seq tracks mapped to the hg19 build using the UCSC Genome table browser. See Table S3 for the track listing. Only loci corresponding to bait sequences with non-zero signal in both the DNase-/ChIP-seq track and in the control track were considered. For each protein, Spearman correlation coefficient was determined between SILAC ratios logarithmized DNase-/ChIP-seq signal density normalized to control signal density. Correlation coefficient deviation was calculated by subtracting the expected DNase-/ChIP-seq to SILAC ratio correlation given the bait GC content to SILAC ratio correlation, and then normalized to the minimum value.

4 Discussions

4.1 Scalable bait production for DNA SILAC AP-MS

The UCE interactome described in Chapter 3 was derived from several hundreds of SILAC AP-MS experiments. Unlike immunoprecipitation techniques where many commercially available antibodies exist, DNA AP-MS requires sequence-specific DNA baits which vary vastly between studies; as a result, DNA baits are usually prepared from first principle. The actual step of affinity purification has been executed in high-throughput in previous studies [96, 97], and recent developments in sample processing and MS instrumentation have enabled parallelized processing of recovered proteins and minimized the sample analysis time [143]. Due to these improvements, the time-limiting step in large-scale SILAC AP-MS screens is now the bait preparation.

Owing to the prohibitively high cost of chemically synthesizing long DNA oligonucleotides, generation of DNA baits longer than 200 bp, such as those used in this thesis, relies on enzymatic synthesis. Long DNA sequences could be amplified out of genomic DNA, using polymerase chain reaction (PCR) in which one primer carries a chemically conjugated affinity tag. Unfortunately, the efficiency, purity and reproducibility of genomic PCR are highly dependent on both the primers and the target sequence. The issue of reproducibility was particularly problematic as the amount of DNA needed for an AP-MS is of a much larger scale than that that obtainable from a single conventional PCR vessel. Furthermore, a sequence-specific affinity-tagged primer would be needed for amplification of every genomic locus; such a primer is very expensive, costing over 50 times more than conventional untagged primers.

We overcame these limitations by sub-cloning the UCE sequences into an intermediate vector, because vector PCR is much more efficient than genomic PCR. We produced a computer script that requested primer sequences through the Primer3 API for each UCE [144]. Thanks to the relatively low GC content of the UCE sequences, automated primer design was almost always successful in finding a primer pair within 100 bp on either side of the UCE boundary. We had found that genomic amplification using these primers had about 35% failure

rate (defined as no visible product or incorrectly-sized product). Most of these failed amplifications could be combined across batches and iteratively re-attempted under various conditions to achieve a correct product. The yield of each successful genomic PCR, however, varied greatly. We then used topoisomerase-assisted cloning to insert the PCR product into a universal backbone. This process was more than 95% successful, as judged by at least one of two randomly picked clones passing colony-PCR validation. Once subcloned into the intermediate vector, the bait was amplified using a pair of universal primers that bind to the backbone flanking the UCE sequences. This reaction was 100% successful, and the product showed no appreciable yield variation judged by absorbance-based DNA quantification. This improvement of yield and reproducibility resulting from moving from genomic to vector DNA was instrumental in achieving the required throughput. Once the UCE clones were made and validated, it was possible to generate the 200 UCE baits on demand within few days.

In section 3.2.4, we generated a differential interactome of uc325 against a transition mutation control. Instead of sub-cloning each mutation control, which would involve laborious preparation and validation steps, we simply constructed each bait from two rounds of PCR: first, each mutant variant bait was amplified as two "halves", with an overlap region where the mutation occurs; these halves then served as templates to re-assemble the mutagenized bait (Figure 15A). The first PCR was easily validated by visualizing in gel electrophoresis, where the combination of the "left" and "right" products would create a diagnostic pattern when the scanning mutant variants are placed in order (Figure 15B). The corroboration between the data shown on Figure 12 and the expected disruption of binding (given the knowledge of existing binding motifs) validated this differential bait generation method. Nevertheless, it turned out that a few interactors constantly bound either to the wildtype or the mutant, regardless of the underlying mutation. Since these interactions were not sequence-specific, we attributed their indiscriminate enrichment/depletion to the aforementioned additional steps required to generate the mutant baits. These additions may have resulted in different sets of residual proteins associated with the bait prior

to the pulldowns. We therefore excluded the proteins that show this behavior from the analysis.

In summary, a well-chosen combination of high-throughput cloning, automated PCR primer design, universal affinity-tagged PCR primers for bait synthesis, and computational algorithms for systematic artifact exclusion together allowed the several hundreds of DNA baits between 200 and 1000 bp to be prepared in a short time and at a low cost.

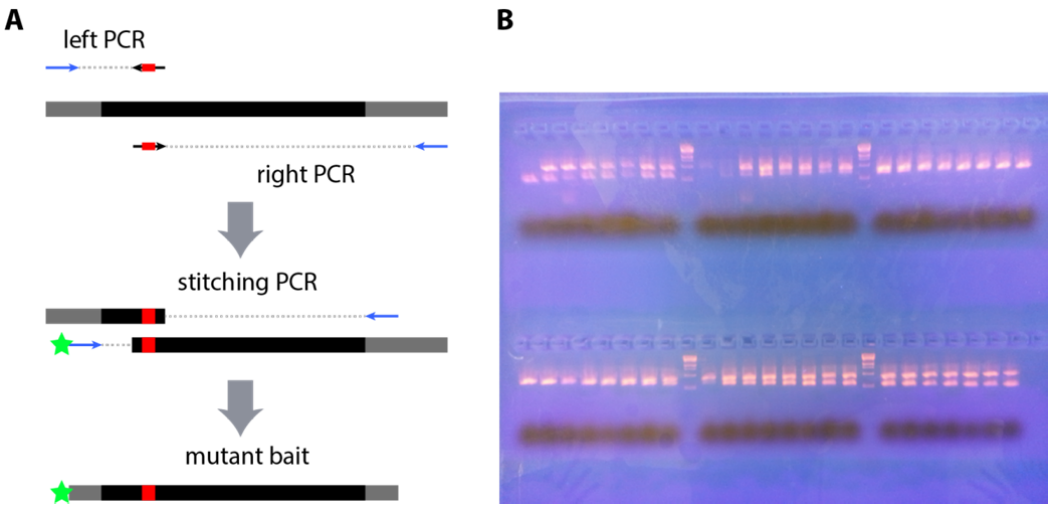

Figure 15: High-throughput production of uc325 scanning variant baits

A PCR scheme for generation of a single variant. The pCR8 vector containing the wildtype uc325 insert serves as template upon which the "left" and "right" mutagenesis PCRs (each one carrying a mutagenic primer) are performed. The resulting products then serve as template for the synthesis of the actual bait, using the same universal, affinity-tagged primers as were used for amplifying other UCE baits.

B Diagnostic gel electrophoresis to validate the identity of the first PCR step. As the reactions are loaded in the order of the position of scanning mutation, the left half grows bigger as the right half grows smaller.

4.2 Quantitative interpretation of SILAC AP-MS data

The first large-scale SILAC AP-MS dataset was generated for protein-protein interactions [97]. The scale-free property of the known protein-protein interaction networks implicates that the vast majority of proteins have a modest number of binding partners, and only few proteins act as "hubs" for a large number of interactors [145]. The identities of protein binders, their affinities and stoichiometries are related to the physical properties of the molecules and, for a substantial number of small, stable protein complexes, only few quantized possibilities exist [146]. However, the scenario for DNA-centric protein-DNA interaction is very different. A DNA sequence is capable of binding a large number of transcription factors, each with a different stoichiometry and affinity, depending on the number and strengths of sites available on the sequence. When protein interactions at a DNA sequence are compared to those at a point mutation variant [98-100], the number of binding sites affected on the DNA sequence is minimal. However, comparison of protein interactions between two completely different DNA sequences typically result in a distribution of enrichment factors that is more continuous than that of protein-protein interactions.

This continuous distribution of enrichment was indeed observed in the UCE interactome. Its quantitative interpretation was clearly of biochemical significance. The most striking evidence for this was in the binding of the transcription factors with known AT-rich or GC-rich sequence preferences, where we found strong correlations between the enrichment folds and the bait nucleotide compositions. Additionally, most of the downstream analyses – gene annotation enrichment, motif discovery, ChIP-seq data integration and interaction correction profiling – required that the SILAC ratios were interpreted quantitatively across baits. Experimentally, this quantitative cross-comparison was made possible by the use of a universal control bait. However, owing to the proteome variations between lysates, errors between forward and reverse SILAC ratios initially prevented the quantitative treatment of these ratios with confidence. Even though we had pooled SILAC nuclear lysates from twenty preparations, in an attempt to reduce the residual heavy-to-light

difference as far as possible, a small residual proteome variation of up to 1.4 fold still remained.

We therefore developed and applied a proteome variation uncoupling procedure (or "*ΔP*-adjustment") to remove this systematic confounding factor from our dataset, taking advantage of the label-switching in the experimental design. We found that this procedure was able to collapse most of the errors observed between the forward and reverse datasets. The resulting SILAC ratios were pushed into distributions that are expected of ideal forward-reverse experiments: a cloud of oppositely-signed forward and corresponding reverse SILAC ratios, and an absence of ratios in the forbidden "same-sign" quadrants. As this ideal could be always be achieved after introducing the adjustment procedure, the need for false positive and contaminant calling, as had been traditionally done, was removed. This enabled indiscriminate, quantitative treatment of all SILAC ratios.

We found better forward-reverse reproducibility if we performed the *ΔP*-adjustment batch-wise rather than globally. There are important implications of this result: Even though the originating lysates of each batch are equivalent, day-to-day handling variation can produce significant batch-wise systematic errors. Parameters such as protein stability in the lysate, its stickiness in the elution process, its degree of non-specific binding owing to random variation in competing DNA, may all contribute to the apparent batch-wise proteome variations. Furthermore, with cross-batch normalization, it would in future no longer be absolutely necessary to calculate the amount of lysate needed prior to the screen, and scale the production accordingly as we have done here ("vertical scaling"); instead, cells may be grown and lysates prepared on demand ("horizontal scaling"). This is true as long as the different lysates are derived from cells that are biochemically equivalent with respect to their protein interaction affinities.

Another implication of the *ΔP*-adjustment concerns other SILAC AP-MS screens where every forward-reverse experiment pair is performed with a different combination of labeled lysates; for instance, owing to genetic background difference or presence of a bait expression vector. In such cases, each forward-

reverse experiment pair has its individual proteome variation that cannot be compared across the board. Interpretation of the SILAC ratios is then limited to classical interactor calling. As mentioned above, this may be acceptable for protein-protein interaction screens of relatively small and stable complexes. However, future SILAC AP-MS studies of protein-DNA interactions under varying cellular conditions should incorporate this consideration into the experimental design.

4.3 Origins of UCE ultraconservation

There are several hypotheses regarding what contributes to ultraconservation of UCEs. Based on experimental data that demonstrated enhancer activity of UCEs, one popular hypothesis is that UCEs contain a high density of overlapping, functional TFBSs, such that no single nucleotide can be mutated without disrupting one or more TFBSs. An argument against the overlapped TFBS hypothesis as the sole contributor to the extreme conservation is often made as follows: Because TFBSs are degenerate, mutation of a single nucleotide belonging to a given TFBS does not always disrupt its ability to bind to the transcription factor. This degeneracy therefore requires the TFBSs to overlap so densely that not a single degenerate base remains. Experimental data supporting or refuting this hypothesis were much needed prior to this study.

We used the UCE interactome to derive the TFBSs for which we found significant evidence of correlation with the binding enrichment. Mapping these motifs back onto the UCE sequences and other sets of control sequences, we found that UCEs were indeed enriched in overlapping TFBSs when compared to other non-conserved enhancers. Importantly, we found that the DNA bases that belonged to overlapping TFBSs were significantly more conserved than those that did not. The latter observation also applied to non-conserved enhancers, so the idea of overlapping TFBSs providing greater evolutionary constraint was clearly plausible. Combined with the significantly higher proportions of overlapping TFBSs for non-exonic UCEs over those for non-conserved enhancers, we conclude that the contribution of overlapping TFBSs to the extreme conservation of non-exonic UCEs is substantial.

Based on our scanning mutagenesis pulldown with one cell line, we saw that indeed many positions, but not all, resulted in multiple disruption of binding owing to transition mutation. It is difficult to assess how strong a disruption there should be before the regulatory function is sufficiently abrogated to result in evolutionary pressure. Furthermore, given the relatively small evolutionary distance between human and mouse, it is more likely that individual TFBSs

keep their sequences despite the degeneracy. This argument, however, would not be applicable to HNCEs defined over larger evolutionary distances.

There is no a priori requirement for a set of sequences, that are identified based solely on a given percentage sequence identity over an arbitrary length, to be under the same kind of evolutionary constraints. We have found, for example, that the enrichment and conservation of overlapping TFBSs did not apply to exonic UCEs. While this does not automatically imply that the underlying coding sequence is functionally critical, it does demonstrate the possibility that the constraints are distributed differently over different processes. Other functional constraints have been proposed, including splicing, nonsense mediated decay, recombination, and structural organization of the chromatin. The degeneracy argument, if true, would imply that UCEs are able to use the remaining "information space" for other functions. However, this raises the question of why different regulatory functions should be compressed into one superimposed locus, given the vastness of genome.

One tentative hypothesis that has not received much attention in the past, but to which our experimental data led us, is that non-exonic UCEs may have evolved into "local minima" in the evolutionary pressure landscape. UCEs, assumed to be critical regulatory elements, may have evolved to a point of such complex functional logic that a mutation would result in a qualitative change in the regulatory logic, rather than its complete disruption. This hypothesis would be consistent with the positioning of UCEs close to genes involved in development, and corroborates many aspects of our dataset. First, the UCE sequences, once shuffled, were still able to yield small but significant enrichments in overlapping TFBSs. We attributed this finding to their AT-richness, which is compatible with the nucleotide composition of many developmentally-regulating TFBSs. Second, we found a significant conservation bias of GC bases over AT bases in UCEs but not in random genomic loci. Since transition mutation inverts the GC content, this observation would be consistent with an evolutionary pressure to prevent formation of even more AT-rich bases. Finally, our scanning screen of uc325 showed that regions of UCEs that were deeply conserved were also those that gained interactions by mutation. This

"qualitative logic alteration" hypothesis would also be compatible with the multiple constraint hypothesis, as they each provide different contributions to the conservation. In this combination, the qualitative logic alteration hypothesis would also alleviate the requirement of the multiple constraint hypothesis for absolutely ubiquitous overlapping TFBSs. Further interaction and reporter experiments combined with sequence and conservation analysis will be needed to test this idea.

4.4 Outlook: the interactome kaleidoscope

It is accepted that the chromatin environment is a major influence on physiological protein-DNA interactions, often blocking interactions that would otherwise take place. In general, DNA-centric methods that are based on affinity purification of nuclear lysate incorporate the context of the nuclear proteome while being uncoupled from pre-existing epigenetic forces. In the AP-MS approach, this unique combination of advantages further synergizes with sensitive and unbiased detection, and the exquisite quantification precision offered by mass spectrometry.

To date, a number of protein-DNA interactomes have been published, coming from both the protein-centric and DNA-centric perspectives. The existing experimental methods for protein-DNA interaction studies cover a spectrum of biochemical to physiological emphases. Protein-centric methods can reveal a biochemical DNA sequence specificity as a motif (SELEX, protein binding microarray), but can also report true *in vivo* binding events (ChIP and ChIP-derivatives). Similarly, DNA-centric methods can probe intrinsic protein binding preference of a given DNA sequence (AP-MS), or the physiological interaction at the native chromatin (iChIP, PiCH). Although the methods that offer *in vivo* perspectives are needed to ultimately validate hypotheses that concern gene regulation, approaches with heavier biochemical emphasis are still of great importance, because they allow for discovery of interactions in regulatory systems where the biological context is unknown. Scalable implementations of both protein-centric and DNA-centric methods now exist, and their results can now be used to validate each other, as we have done in this thesis. Future interactomes will help us better understand how the different biochemical and physiological emphases of each method give rise to the data that they deliver, and how the shortcomings of one dataset may be complemented by the strength of another.

A protein-DNA interactome gives information about the binding that happens in one experimental condition, and thus provides a static snapshot. Better depth of regulatory understanding may be achieved by introducing perturbations to the

system being studied. Deep perturbation of protein-centric studies is becoming common, owing to the development of scalable ChIP-seq. For instance, this approach is being used by various groups to study DNA targets of a transcription factor across different cell types/stimuli [147, 148]. With large-scale DNA AP-MS screens now possible, similar perturbations may be performed in order to study protein binders across different nuclear environments.

Recently, a novel application of next generation sequencing was developed that allows deep characterization of a regulatory DNA element. Known as "massively parallel reporter assay" (MPRA), the approach measures the regulatory activity (by expression of RNA reporter barcodes) of a short enhancer, as well as thousands of its mutational variants. MPRA has especially promising applications in the field of synthetic biology, as it can be used to aid rational design of artificial regulatory elements. Large-scale AP-MS would be the ideal technique that delivers the complementary deeply-perturbed interactomes, which would allow the differential reporter activity to be linked by differential protein binding. Together, these two technologies can offer the community the hope that, one day, the aspiration to reverse-engineer complex, multi-factorial enhancers will be fulfilled.

5 Bibliography

1. Cowie, A. and R.M. Myers, *DNA sequences involved in transcriptional regulation of the mouse beta-globin promoter in murine erythroleukemia cells.* Mol Cell Biol, 1988. **8**(8): p. 3122-8.

2. Maniatis, T., J.V. Falvo, T.H. Kim, T.K. Kim, C.H. Lin, B.S. Parekh and M.G. Wathelet, *Structure and function of the interferon-beta enhanceosome.* Cold Spring Harb Symp Quant Biol, 1998. **63**: p. 609-20.

3. Kulkarni, M.M. and D.N. Arnosti, *Information display by transcriptional enhancers.* Development, 2003. **130**(26): p. 6569-75.

4. Arnosti, D.N. and M.M. Kulkarni, *Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards?* J Cell Biochem, 2005. **94**(5): p. 890-8.

5. Rohs, R., X. Jin, S.M. West, R. Joshi, B. Honig and R.S. Mann, *Origins of specificity in protein-DNA recognition.* Annu Rev Biochem, 2010. **79**: p. 233-69.

6. Matys, V., O.V. Kel-Margoulis, E. Fricke, I. Liebich, S. Land, A. Barre-Dirrie, I. Reuter, D. Chekmenev, M. Krull, K. Hornischer, N. Voss, P. Stegmaier, B. Lewicki-Potapov, H. Saxel, A.E. Kel and E. Wingender, *TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes.* Nucleic Acids Res, 2006. **34**(Database issue): p. D108-10.

7. Joshi, R., J.M. Passner, R. Rohs, R. Jain, A. Sosinsky, M.A. Crickmore, V. Jacob, A.K. Aggarwal, B. Honig and R.S. Mann, *Functional specificity of a Hox protein mediated by the recognition of minor groove structure.* Cell, 2007. **131**(3): p. 530-43.

8. Li, T., Y. Jin, A.K. Vershon and C. Wolberger, *Crystal structure of the MATa1/MATalpha2 homeodomain heterodimer in complex with DNA containing an A-tract.* Nucleic Acids Res, 1998. **26**(24): p. 5707-18.

9. Chen, L., J.N. Glover, P.G. Hogan, A. Rao and S.C. Harrison, *Structure of the DNA-binding domains from NFAT, Fos and Jun bound specifically to DNA.* Nature, 1998. **392**(6671): p. 42-8.

10. Ellenberger, T., D. Fass, M. Arnaud and S.C. Harrison, *Crystal structure of transcription factor E47: E-box recognition by a basic region helix-loop-helix dimer.* Genes Dev, 1994. **8**(8): p. 970-80.

11. Ehret, G.B., P. Reichenbach, U. Schindler, C.M. Horvath, S. Fritz, M. Nabholz and P. Bucher, *DNA binding specificity of different STAT proteins.*

Comparison of in vitro specificity with natural target sites. J Biol Chem, 2001. **276**(9): p. 6675-88.

12. Papadopoulos, D.K., K. Skouloudaki, Y. Adachi, C. Samakovlis and W.J. Gehring, *Dimer formation via the homeodomain is required for function and specificity of Sex combs reduced in Drosophila.* Dev Biol, 2012. **367**(1): p. 78-89.

13. Bryne, J.C., E. Valen, M.H. Tang, T. Marstrand, O. Winther, I. da Piedade, A. Krogh, B. Lenhard and A. Sandelin, *JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update.* Nucleic Acids Res, 2008. **36**(Database issue): p. D102-6.

14. Sinha, S., E. van Nimwegen and E.D. Siggia, *A probabilistic method to detect regulatory modules.* Bioinformatics, 2003. **19 Suppl 1**: p. i292-301.

15. Zhu, C., K.J. Byers, R.P. McCord, Z. Shi, M.F. Berger, D.E. Newburger, K. Saulrieta, Z. Smith, M.V. Shah, M. Radhakrishnan, A.A. Philippakis, Y. Hu, F. De Masi, M. Pacek, A. Rolfs, T. Murthy, J. Labaer and M.L. Bulyk, *High-resolution DNA-binding specificity analysis of yeast transcription factors.* Genome Res, 2009. **19**(4): p. 556-66.

16. Luger, K., A.W. Mader, R.K. Richmond, D.F. Sargent and T.J. Richmond, *Crystal structure of the nucleosome core particle at 2.8 A resolution.* Nature, 1997. **389**(6648): p. 251-60.

17. Segal, E., Y. Fondufe-Mittendorf, L. Chen, A. Thastrom, Y. Field, I.K. Moore, J.P. Wang and J. Widom, *A genomic code for nucleosome positioning.* Nature, 2006. **442**(7104): p. 772-8.

18. Tillo, D. and T.R. Hughes, *G+C content dominates intrinsic nucleosome occupancy.* BMC Bioinformatics, 2009. **10**: p. 442.

19. Martinez-Campa, C., P. Politis, J.L. Moreau, N. Kent, J. Goodall, J. Mellor and C.R. Goding, *Precise nucleosome positioning and the TATA box dictate requirements for the histone H4 tail and the bromodomain factor Bdf1.* Mol Cell, 2004. **15**(1): p. 69-81.

20. Tirosh, I. and N. Barkai, *Two strategies for gene regulation by promoter nucleosomes.* Genome Res, 2008. **18**(7): p. 1084-91.

21. Thoma, F. and T. Koller, *Influence of histone H1 on chromatin structure.* Cell, 1977. **12**(1): p. 101-7.

22. Chakravarthy, S., Y.J. Park, J. Chodaparambil, R.S. Edayathumangalam and K. Luger, *Structure and dynamic properties of nucleosome core particles.* FEBS Lett, 2005. **579**(4): p. 895-8.

23. Karatay, S., S. Katircioglu and T. Erbengi, *Ultrastructural investigations on the degenerative and regenerative processes in chorda tympani in bell's paralysis.* Acta Otolaryngol, 1976. **81**(3-4): p. 304-14.

24. Nathan, D., K. Ingvarsdottir, D.E. Sterner, G.R. Bylebyl, M. Dokmanovic, J.A. Dorsey, K.A. Whelan, M. Krsmanovic, W.S. Lane, P.B. Meluh, E.S. Johnson and S.L. Berger, *Histone sumoylation is a negative regulator in Saccharomyces cerevisiae and shows dynamic interplay with positive-acting histone modifications.* Genes Dev, 2006. **20**(8): p. 966-76.

25. Kobza, K., G. Camporeale, B. Rueckert, A. Kueh, J.B. Griffin, G. Sarath and J. Zempleni, *K4, K9 and K18 in human histone H3 are targets for biotinylation by biotinidase.* FEBS J, 2005. **272**(16): p. 4249-59.

26. Stock, J.K., S. Giadrossi, M. Casanova, E. Brookes, M. Vidal, H. Koseki, N. Brockdorff, A.G. Fisher and A. Pombo, *Ring1-mediated ubiquitination of H2A restrains poised RNA polymerase II at bivalent genes in mouse ES cells.* Nat Cell Biol, 2007. **9**(12): p. 1428-35.

27. Cao, R., L. Wang, H. Wang, L. Xia, H. Erdjument-Bromage, P. Tempst, R.S. Jones and Y. Zhang, *Role of histone H3 lysine 27 methylation in Polycomb-group silencing.* Science, 2002. **298**(5595): p. 1039-43.

28. Sedkov, Y., E. Cho, S. Petruk, L. Cherbas, S.T. Smith, R.S. Jones, P. Cherbas, E. Canaani, J.B. Jaynes and A. Mazo, *Methylation at lysine 4 of histone H3 in ecdysone-dependent development of Drosophila.* Nature, 2003. **426**(6962): p. 78-83.

29. Schiltz, R.L., C.A. Mizzen, A. Vassilev, R.G. Cook, C.D. Allis and Y. Nakatani, *Overlapping but distinct patterns of histone acetylation by the human coactivators p300 and PCAF within nucleosomal substrates.* J Biol Chem, 1999. **274**(3): p. 1189-92.

30. Vermeulen, M., H.C. Eberl, F. Matarese, H. Marks, S. Denissov, F. Butter, K.K. Lee, J.V. Olsen, A.A. Hyman, H.G. Stunnenberg and M. Mann, *Quantitative interaction proteomics and genome-wide profiling of epigenetic histone marks and their readers.* Cell, 2010. **142**(6): p. 967-80.

31. Kouzarides, T., *Chromatin modifications and their function.* Cell, 2007. **128**(4): p. 693-705.

32. Karch, K.R., J.E. Denizio, B.E. Black and B.A. Garcia, *Identification and interrogation of combinatorial histone modifications.* Front Genet, 2013. **4**: p. 264.

33. Zee, B.M., R.S. Levin, P.A. Dimaggio and B.A. Garcia, *Global turnover of histone post-translational modifications and variants in human cells.* Epigenetics Chromatin, 2010. **3**(1): p. 22.

34. Talbert, P.B. and S. Henikoff, *Spreading of silent chromatin: inaction at a distance.* Nat Rev Genet, 2006. **7**(10): p. 793-803.

35. Valenzuela, L. and R.T. Kamakaka, *Chromatin insulators.* Annu Rev Genet, 2006. **40**: p. 107-38.

36. Guibert, S. and M. Weber, *Functions of DNA methylation and hydroxymethylation in mammalian development.* Curr Top Dev Biol, 2013. **104**: p. 47-83.

37. Margueron, R. and D. Reinberg, *Chromatin structure and the inheritance of epigenetic information.* Nat Rev Genet, 2010. **11**(4): p. 285-96.

38. Cremer, T. and M. Cremer, *Chromosome territories.* Cold Spring Harb Perspect Biol, 2010. **2**(3): p. a003889.

39. Jin, F., Y. Li, J.R. Dixon, S. Selvaraj, Z. Ye, A.Y. Lee, C.A. Yen, A.D. Schmitt, C.A. Espinoza and B. Ren, *A high-resolution map of the three-dimensional chromatin interactome in human cells.* Nature, 2013. **503**(7475): p. 290-4.

40. Arbab, M., S. Mahony, H. Cho, J.M. Chick, P.A. Rolfe, J.P. van Hoff, V.W. Morris, S.P. Gygi, R.L. Maas, D.K. Gifford and R.I. Sherwood, *A multi-parametric flow cytometric assay to analyze DNA-protein interactions.* Nucleic Acids Res, 2013. **41**(2): p. e38.

41. Kyrpides, N.C., *Genomes OnLine Database (GOLD 1.0): a monitor of complete and ongoing genome projects world-wide.* Bioinformatics, 1999. **15**(9): p. 773-4.

42. Dubchak, I., M. Brudno, G.G. Loots, L. Pachter, C. Mayor, E.M. Rubin and K.A. Frazer, *Active conservation of noncoding sequences revealed by three-way species comparisons.* Genome Res, 2000. **10**(9): p. 1304-6.

43. Bejerano, G., M. Pheasant, I. Makunin, S. Stephen, W.J. Kent, J.S. Mattick and D. Haussler, *Ultraconserved elements in the human genome.* Science, 2004. **304**(5675): p. 1321-5.

44. Aloni, R. and D. Lancet, *Conservation anchors in the vertebrate genome.* Genome Biol, 2005. **6**(7): p. 115.

45. Sakuraba, Y., T. Kimura, H. Masuya, H. Noguchi, H. Sezutsu, K.R. Takahasi, A. Toyoda, R. Fukumura, T. Murata, Y. Sakaki, M. Yamamura, S. Wakana, T. Noda, T. Shiroishi and Y. Gondo, *Identification and characterization of new long conserved noncoding sequences in vertebrates.* Mamm Genome, 2008. **19**(10-12): p. 703-12.

46. Manzanares, M., H. Wada, N. Itasaki, P.A. Trainor, R. Krumlauf and P.W. Holland, *Conservation and elaboration of Hox gene regulation during evolution of the vertebrate head.* Nature, 2000. **408**(6814): p. 854-7.

47. Baxter, L., A. Jironkin, R. Hickman, J. Moore, C. Barrington, P. Krusche, N.P. Dyer, V. Buchanan-Wollaston, A. Tiskin, J. Beynon, K. Denby and S. Ott, *Conserved noncoding sequences highlight shared components of regulatory networks in dicotyledonous plants.* Plant Cell, 2012. **24**(10): p. 3949-65.

48. Engstrom, P.G., S.J. Ho Sui, O. Drivenes, T.S. Becker and B. Lenhard, *Genomic regulatory blocks underlie extensive microsynteny conservation in insects.* Genome Res, 2007. **17**(12): p. 1898-908.

49. Walter, K., I. Abnizova, G. Elgar and W.R. Gilks, *Striking nucleotide frequency pattern at the borders of highly conserved vertebrate non-coding sequences.* Trends Genet, 2005. **21**(8): p. 436-40.

50. Bernstein, B.E., T.S. Mikkelsen, X. Xie, M. Kamal, D.J. Huebert, J. Cuff, B. Fry, A. Meissner, M. Wernig, K. Plath, R. Jaenisch, A. Wagschal, R. Feil, S.L. Schreiber and E.S. Lander, *A bivalent chromatin structure marks key developmental genes in embryonic stem cells.* Cell, 2006. **125**(2): p. 315-26.

51. Pennacchio, L.A., N. Ahituv, A.M. Moses, S. Prabhakar, M.A. Nobrega, M. Shoukry, S. Minovitsky, I. Dubchak, A. Holt, K.D. Lewis, I. Plajzer-Frick, J. Akiyama, S. De Val, V. Afzal, B.L. Black, O. Couronne, M.B. Eisen, A. Visel and E.M. Rubin, *In vivo enhancer analysis of human conserved non-coding sequences.* Nature, 2006. **444**(7118): p. 499-502.

52. Harmston, N., A. Baresic and B. Lenhard, *The mystery of extreme non-coding conservation.* Philos Trans R Soc Lond B Biol Sci, 2013. **368**(1632): p. 20130021.

53. Ahituv, N., Y. Zhu, A. Visel, A. Holt, V. Afzal, L.A. Pennacchio and E.M. Rubin, *Deletion of ultraconserved elements yields viable mice.* PLoS Biol, 2007. **5**(9): p. e234.

54. Hellman, L.M. and M.G. Fried, *Electrophoretic mobility shift assay (EMSA) for detecting protein-nucleic acid interactions.* Nat Protoc, 2007. **2**(8): p. 1849-61.

55. Galas, D.J. and A. Schmitz, *DNAse footprinting: a simple method for the detection of protein-DNA binding specificity.* Nucleic Acids Res, 1978. **5**(9): p. 3157-70.

56. Brenner, S.*, et al.*, *Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays.* Nat Biotechnol, 2000. **18**(6): p. 630-4.

57. Tuerk, C. and L. Gold, *Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase.* Science, 1990. **249**(4968): p. 505-10.

58. Ogawa, N. and M.D. Biggin, *High-throughput SELEX determination of DNA sequences bound by transcription factors in vitro.* Methods Mol Biol, 2012. **786**: p. 51-63.

59. Schutze, T., B. Wilhelm, N. Greiner, H. Braun, F. Peter, M. Morl, V.A. Erdmann, H. Lehrach, Z. Konthur, M. Menger, P.F. Arndt and J. Glokler, *Probing the SELEX process with next-generation sequencing.* PLoS One, 2011. **6**(12): p. e29604.

60. Mikkelsen, T.S.*, et al.*, *Genome-wide maps of chromatin state in pluripotent and lineage-committed cells.* Nature, 2007. **448**(7153): p. 553-60.

61. Johnson, D.S., A. Mortazavi, R.M. Myers and B. Wold, *Genome-wide mapping of in vivo protein-DNA interactions.* Science, 2007. **316**(5830): p. 1497-502.

62. Barski, A., S. Cuddapah, K. Cui, T.Y. Roh, D.E. Schones, Z. Wang, G. Wei, I. Chepelev and K. Zhao, *High-resolution profiling of histone methylations in the human genome.* Cell, 2007. **129**(4): p. 823-37.

63. Rhee, H.S. and B.F. Pugh, *Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution.* Cell, 2011. **147**(6): p. 1408-19.

64. Ouwerkerk, P.B. and A.H. Meijer, *Yeast one-hybrid screening for DNA-protein interactions.* Curr Protoc Mol Biol, 2001. **Chapter 12**: p. Unit 12 12.

65. Dejardin, J. and R.E. Kingston, *Purification of proteins associated with specific genomic Loci.* Cell, 2009. **136**(1): p. 175-86.

66. Hoshino, A. and H. Fujii, *Insertional chromatin immunoprecipitation: a method for isolating specific genomic regions.* J Biosci Bioeng, 2009. **108**(5): p. 446-9.

67. Jinek, M., K. Chylinski, I. Fonfara, M. Hauer, J.A. Doudna and E. Charpentier, *A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity.* Science, 2012. **337**(6096): p. 816-21.

68. Boch, J., *TALEs of genome targeting.* Nat Biotechnol, 2011. **29**(2): p. 135-6.

69. Wilkins, M.R., C. Pasquali, R.D. Appel, K. Ou, O. Golaz, J.C. Sanchez, J.X. Yan, A.A. Gooley, G. Hughes, I. Humphery-Smith, K.L. Williams and D.F. Hochstrasser, *From proteins to proteomes: large scale protein identification by two-dimensional electrophoresis and amino acid analysis.* Biotechnology (N Y), 1996. **14**(1): p. 61-5.

70. Michalski, A., J. Cox and M. Mann, *More than 100,000 detectable peptide species elute in single shotgun proteomics runs but the majority is inaccessible to data-dependent LC-MS/MS.* J Proteome Res, 2011. **10**(4): p. 1785-93.

71. Deeb, S.J., R.C. D'Souza, J. Cox, M. Schmidt-Supprian and M. Mann, *Super-SILAC allows classification of diffuse large B-cell lymphoma subtypes by their protein expression profiles.* Mol Cell Proteomics, 2012. **11**(5): p. 77-89.

72. Nagaraj, N., N.A. Kulak, J. Cox, N. Neuhauser, K. Mayr, O. Hoerning, O. Vorm and M. Mann, *System-wide perturbation analysis with nearly complete coverage of the yeast proteome by single-shot ultra HPLC runs on a bench top Orbitrap.* Mol Cell Proteomics, 2012. **11**(3): p. M111 013722.

73. Thakur, S.S., T. Geiger, B. Chatterjee, P. Bandilla, F. Frohlich, J. Cox and M. Mann, *Deep and highly sensitive proteome coverage by LC-MS/MS without prefractionation.* Mol Cell Proteomics, 2011. **10**(8): p. M110 003699.

74. Fenn, J.B., M. Mann, C.K. Meng, S.F. Wong and C.M. Whitehouse, *Electrospray ionization for mass spectrometry of large biomolecules.* Science, 1989. **246**(4926): p. 64-71.

75. Marshall, A.G., C.L. Hendrickson and G.S. Jackson, *Fourier transform ion cyclotron resonance mass spectrometry: a primer.* Mass Spectrom Rev, 1998. **17**(1): p. 1-35.

76. Makarov, A., *Electrostatic axially harmonic orbital trapping: a high-performance technique of mass analysis.* Anal Chem, 2000. **72**(6): p. 1156-62.

77. Roepstorff, P. and J. Fohlman, *Proposal for a common nomenclature for sequence ions in mass spectra of peptides.* Biomed Mass Spectrom, 1984. **11**(11): p. 601.

78. Steen, H. and M. Mann, *The ABC's (and XYZ's) of peptide sequencing.* Nat Rev Mol Cell Biol, 2004. **5**(9): p. 699-711.

79. *Method of the Year 2012.* Nat Methods, 2013. **10**(1): p. 1.

80. Keshishian, H., T. Addona, M. Burgess, E. Kuhn and S.A. Carr, *Quantitative, multiplexed assays for low abundance proteins in plasma by targeted mass spectrometry and stable isotope dilution.* Mol Cell Proteomics, 2007. **6**(12): p. 2212-29.

81. Cox, J. and M. Mann, *Quantitative, high-resolution proteomics for data-driven systems biology.* Annu Rev Biochem, 2011. **80**: p. 273-99.

82. Cox, J. and M. Mann, *MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification.* Nat Biotechnol, 2008. **26**(12): p. 1367-72.

83. Gygi, S.P., B. Rist, S.A. Gerber, F. Turecek, M.H. Gelb and R. Aebersold, *Quantitative analysis of complex protein mixtures using isotope-coded affinity tags.* Nat Biotechnol, 1999. **17**(10): p. 994-9.

84. Boersema, P.J., R. Raijmakers, S. Lemeer, S. Mohammed and A.J. Heck, *Multiplex peptide stable isotope dimethyl labeling for quantitative proteomics.* Nat Protoc, 2009. **4**(4): p. 484-94.

85. Wiese, S., K.A. Reidegeld, H.E. Meyer and B. Warscheid, *Protein labeling by iTRAQ: a new tool for quantitative mass spectrometry in proteome research.* Proteomics, 2007. **7**(3): p. 340-50.

86. Ong, S.E., B. Blagoev, I. Kratchmarova, D.B. Kristensen, H. Steen, A. Pandey and M. Mann, *Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics.* Mol Cell Proteomics, 2002. **1**(5): p. 376-86.

87. Schwanhausser, B., M. Gossen, G. Dittmar and M. Selbach, *Global analysis of cellular protein translation by pulsed SILAC.* Proteomics, 2009. **9**(1): p. 205-9.

88. Zeiler, M., W.L. Straube, E. Lundberg, M. Uhlen and M. Mann, *A Protein Epitope Signature Tag (PrEST) library allows SILAC-based absolute quantification and multiplexed determination of protein copy numbers in cell lines.* Mol Cell Proteomics, 2012. **11**(3): p. O111 009613.

89. Gerber, S.A., J. Rush, O. Stemman, M.W. Kirschner and S.P. Gygi, *Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS.* Proc Natl Acad Sci U S A, 2003. **100**(12): p. 6940-5.

90. Arike, L., K. Valgepea, L. Peil, R. Nahku, K. Adamberg and R. Vilu, *Comparison and applications of label-free absolute proteome quantification methods on Escherichia coli.* J Proteomics, 2012. **75**(17): p. 5437-48.

91. Liu, H., R.G. Sadygov and J.R. Yates, 3rd, *A model for random sampling and estimation of relative protein abundance in shotgun proteomics.* Anal Chem, 2004. **76**(14): p. 4193-201.

92. Lu, P., C. Vogel, R. Wang, X. Yao and E.M. Marcotte, *Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation.* Nat Biotechnol, 2007. **25**(1): p. 117-24.

93. Ishihama, Y., Y. Oda, T. Tabata, T. Sato, T. Nagasu, J. Rappsilber and M. Mann, *Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein.* Mol Cell Proteomics, 2005. **4**(9): p. 1265-72.

94. Rappsilber, J., U. Ryder, A.I. Lamond and M. Mann, *Large-scale proteomic analysis of the human spliceosome.* Genome Res, 2002. **12**(8): p. 1231-45.

95. Luber, C.A., J. Cox, H. Lauterbach, B. Fancke, M. Selbach, J. Tschopp, S. Akira, M. Wiegand, H. Hochrein, M. O'Keeffe and M. Mann, *Quantitative proteomics reveals subset-specific viral recognition in dendritic cells.* Immunity, 2010. **32**(2): p. 279-89.

96. Eberl, H.C., C.G. Spruijt, C.D. Kelstrup, M. Vermeulen and M. Mann, *A Map of General and Specialized Chromatin Readers in Mouse Tissues Generated by Label-free Interaction Proteomics.* Mol Cell, 2013. **49**(2): p. 368-78.

97. Hubner, N.C., A.W. Bird, J. Cox, B. Splettstoesser, P. Bandilla, I. Poser, A. Hyman and M. Mann, *Quantitative proteomics combined with BAC TransgeneOmics reveals in vivo protein interactions.* J Cell Biol, 2010. **189**(4): p. 739-54.

98. Butter, F., L. Davison, T. Viturawong, M. Scheibe, M. Vermeulen, J.A. Todd and M. Mann, *Proteome-wide analysis of disease-associated SNPs that show allele-specific transcription factor binding.* PLoS Genet, 2012. **8**(9): p. e1002982.

99. Mittler, G., F. Butter and M. Mann, *A SILAC-based DNA protein interaction screen that identifies candidate binding proteins to functional DNA elements.* Genome Res, 2009. **19**(2): p. 284-93.

100. Scheibe, M., F. Butter, M. Hafner, T. Tuschl and M. Mann, *Quantitative mass spectrometry and PAR-CLIP to identify RNA-protein interactions.* Nucleic Acids Res, 2012. **40**(19): p. 9897-902.

101. Encode Project Consortium, *An integrated encyclopedia of DNA elements in the human genome.* Nature, 2012. **489**(7414): p. 57-74.

102. Simicevic, J., A.W. Schmid, P.A. Gilardoni, B. Zoller, S.K. Raghav, I. Krier, C. Gubelmann, F. Lisacek, F. Naef, M. Moniatte and B. Deplancke, *Absolute quantification of transcription factors during cellular differentiation using multiplexed targeted proteomics.* Nat Methods, 2013. **10**(6): p. 570-6.

103. Spitz, F. and E.E. Furlong, *Transcription factors: from enhancer binding to developmental control.* Nat Rev Genet, 2012. **13**(9): p. 613-26.

104. Williamson, I., R.E. Hill and W.A. Bickmore, *Enhancers: from developmental genetics to the genetics of common human disease.* Dev Cell, 2011. **21**(1): p. 17-9.

105. Berman, B.P., B.D. Pfeiffer, T.R. Laverty, S.L. Salzberg, G.M. Rubin, M.B. Eisen and S.E. Celniker, *Computational identification of developmental enhancers: conservation and function of transcription factor binding-site clusters in Drosophila melanogaster and Drosophila pseudoobscura.* Genome Biol, 2004. **5**(9): p. R61.

106. Visel, A., S. Minovitsky, I. Dubchak and L.A. Pennacchio, *VISTA Enhancer Browser--a database of tissue-specific human enhancers.* Nucleic Acids Res, 2007. **35**(Database issue): p. D88-92.

107. Martinez, F., S. Monfort, M. Rosello, S. Oltra, D. Blesa, R. Quiroga, S. Mayo and C. Orellana, *Enrichment of ultraconserved elements among genomic imbalances causing mental delay and congenital anomalies.* BMC Med Genomics, 2010. **3**: p. 54.

108. Poitras, L., M. Yu, C. Lesage-Pelletier, R.B. Macdonald, J.P. Gagne, G. Hatch, I. Kelly, S.P. Hamilton, J.L. Rubenstein, G.G. Poirier and M. Ekker, *An SNP in an ultraconserved regulatory element affects Dlx5/Dlx6 regulation in the forebrain.* Development, 2010. **137**(18): p. 3089-97.

109. Yang, R., B. Frank, K. Hemminki, C.R. Bartram, B. Wappenschmidt, C. Sutter, M. Kiechle, P. Bugert, R.K. Schmutzler, N. Arnold, B.H. Weber, D. Niederacher, A. Meindl and B. Burwinkel, *SNPs in ultraconserved elements and familial breast cancer risk.* Carcinogenesis, 2008. **29**(2): p. 351-5.

110. Siepel, A., G. Bejerano, J.S. Pedersen, A.S. Hinrichs, M. Hou, K. Rosenbloom, H. Clawson, J. Spieth, L.W. Hillier, S. Richards, G.M.

Weinstock, R.K. Wilson, R.A. Gibbs, W.J. Kent, W. Miller and D. Haussler, *Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes.* Genome Res, 2005. **15**(8): p. 1034-50.

111. Wilson, M.D., N.L. Barbosa-Morais, D. Schmidt, C.M. Conboy, L. Vanes, V.L. Tybulewicz, E.M. Fisher, S. Tavare and D.T. Odom, *Species-specific transcription in mice carrying human chromosome 21.* Science, 2008. **322**(5900): p. 434-8.

112. Mirzaei, H., T.A. Knijnenburg, B. Kim, M. Robinson, P. Picotti, G.W. Carter, S. Li, D.J. Dilworth, J.K. Eng, J.D. Aitchison, I. Shmulevich, T. Galitski, R. Aebersold and J. Ranish, *Systematic measurement of transcription factor-DNA interactions by targeted mass spectrometry identifies candidate gene regulatory proteins.* Proc Natl Acad Sci U S A, 2013.

113. Tacheny, A., S. Michel, M. Dieu, L. Payen, T. Arnould and P. Renard, *Unbiased proteomic analysis of proteins interacting with the HIV-1 5'LTR sequence: role of the transcription factor Meis.* Nucleic Acids Res, 2012. **40**(21): p. e168.

114. Palmer, S.J., E.S. Tay, N. Santucci, T.T. Cuc Bach, J. Hook, F.A. Lemckert, R.V. Jamieson, P.W. Gunnning and E.C. Hardeman, *Expression of Gtf2ird1, the Williams syndrome-associated gene, during mouse development.* Gene Expr Patterns, 2007. **7**(4): p. 396-404.

115. Thompson, P.D., M. Webb, W. Beckett, T. Hinsley, T. Jowitt, A.D. Sharrocks and M. Tassabehji, *GTF2IRD1 regulates transcription by binding an evolutionarily conserved DNA motif 'GUCE'.* FEBS Lett, 2007. **581**(6): p. 1233-42.

116. Aalfs, J.D., G.J. Narlikar and R.E. Kingston, *Functional differences between the human ATP-dependent nucleosome remodeling proteins BRG1 and SNF2H.* J Biol Chem, 2001. **276**(36): p. 34270-8.

117. Jin, C. and G. Felsenfeld, *Nucleosome stability mediated by histone variants H3.3 and H2A.Z.* Genes Dev, 2007. **21**(12): p. 1519-29.

118. Rai, T.S., A. Puri, T. McBryan, J. Hoffman, Y. Tang, N.A. Pchelintsev, J. van Tuyn, R. Marmorstein, D.C. Schultz and P.D. Adams, *Human CABIN1 is a functional member of the human HIRA/UBN1/ASF1a histone H3.3 chaperone complex.* Mol Cell Biol, 2011. **31**(19): p. 4107-18.

119. Udugama, M., A. Sabri and B. Bartholomew, *The INO80 ATP-dependent chromatin remodeling complex is a nucleosome spacing factor.* Mol Cell Biol, 2011. **31**(4): p. 662-73.

120. Xie, W., T. Ling, Y. Zhou, W. Feng, Q. Zhu, H.G. Stunnenberg, I. Grummt and W. Tao, *The chromatin remodeling complex NuRD establishes the poised state of rRNA genes characterized by bivalent histone modifications and altered nucleosome positions.* Proc Natl Acad Sci U S A, 2012. **109**(21): p. 8161-6.

121. Lu, X., S.N. Wontakal, A.V. Emelyanov, P. Morcillo, A.Y. Konev, D.V. Fyodorov and A.I. Skoultchi, *Linker histone H1 is essential for Drosophila development, the establishment of pericentric heterochromatin, and a normal polytene chromosome structure.* Genes Dev, 2009. **23**(4): p. 452-65.

122. Mendenhall, E.M., R.P. Koche, T. Truong, V.W. Zhou, B. Issac, A.S. Chi, M. Ku and B.E. Bernstein, *GC-rich sequence elements recruit PRC2 in mammalian ES cells.* PLoS Genet, 2010. **6**(12): p. e1001244.

123. Hermsen, R., S. Tans and P.R. ten Wolde, *Transcriptional regulation by competing transcription factor modules.* PLoS Comput Biol, 2006. **2**(12): p. e164.

124. Ngondo-Mbongo, R.P., E. Myslinski, J.C. Aster and P. Carbon, *Modulation of gene expression via overlapping binding sites exerted by ZNF143, Notch1 and THAP11.* Nucleic Acids Res, 2013.

125. Koyama-Nasu, R., G. David and N. Tanese, *The F-box protein Fbl10 is a novel transcriptional repressor of c-Jun.* Nat Cell Biol, 2007. **9**(9): p. 1074-80.

126. Meyer, L.R.*, et al.*, *The UCSC Genome Browser database: extensions and updates 2013.* Nucleic Acids Res, 2013. **41**(D1): p. D64-9.

127. Collins, D.W. and T.H. Jukes, *Rates of transition and transversion in coding sequences since the human-rodent divergence.* Genomics, 1994. **20**(3): p. 386-96.

128. Palmer, S.J., N. Santucci, J. Widagdo, S.J. Bontempo, K.M. Taylor, E.S. Tay, J. Hook, F. Lemckert, P.W. Gunning and E.C. Hardeman, *Negative autoregulation of GTF2IRD1 in Williams-Beuren syndrome via a novel DNA binding mechanism.* J Biol Chem, 2010. **285**(7): p. 4715-24.

129. Dohm, J.C., C. Lottaz, T. Borodina and H. Himmelbauer, *Substantial biases in ultra-short read data sets from high-throughput DNA sequencing.* Nucleic Acids Res, 2008. **36**(16): p. e105.

130. Ku, M., R.P. Koche, E. Rheinbay, E.M. Mendenhall, M. Endoh, T.S. Mikkelsen, A. Presser, C. Nusbaum, X. Xie, A.S. Chi, M. Adli, S. Kasif, L.M. Ptaszek, C.A. Cowan, E.S. Lander, H. Koseki and B.E. Bernstein,

Genomewide analysis of PRC1 and PRC2 occupancy identifies two classes of bivalent domains. PLoS Genet, 2008. **4**(10): p. e1000242.

131. Pasini, D., M. Malatesta, H.R. Jung, J. Walfridsson, A. Willer, L. Olsson, J. Skotte, A. Wutz, B. Porse, O.N. Jensen and K. Helin, *Characterization of an antagonistic switch between histone H3 lysine 27 methylation and acetylation in the transcriptional regulation of Polycomb group target genes.* Nucleic Acids Res, 2010. **38**(15): p. 4958-69.

132. Tie, F., R. Banerjee, C.A. Stratton, J. Prasad-Sinha, V. Stepanik, A. Zlobin, M.O. Diaz, P.C. Scacheri and P.J. Harte, *CBP-mediated acetylation of histone H3 lysine 27 antagonizes Drosophila Polycomb silencing.* Development, 2009. **136**(18): p. 3131-41.

133. Smith, R.P., L. Taher, R.P. Patwardhan, M.J. Kim, F. Inoue, J. Shendure, I. Ovcharenko and N. Ahituv, *Massively parallel decoding of mammalian regulatory sequences supports a flexible organizational model.* Nat Genet, 2013.

134. Licastro, D., V.A. Gennarino, F. Petrera, R. Sanges, S. Banfi and E. Stupka, *Promiscuity of enhancer, coding and non-coding transcription functions in ultraconserved elements.* BMC Genomics, 2010. **11**: p. 151.

135. Ni, J.Z., L. Grate, J.P. Donohue, C. Preston, N. Nobida, G. O'Brien, L. Shiue, T.A. Clark, J.E. Blume and M. Ares, Jr., *Ultraconserved elements are associated with homeostatic control of splicing regulators by alternative splicing and nonsense-mediated decay.* Genes Dev, 2007. **21**(6): p. 708-18.

136. Scaruffi, P., *The transcribed-ultraconserved regions: a novel class of long noncoding RNAs involved in cancer susceptibility.* ScientificWorldJournal, 2011. **11**: p. 340-52.

137. Perry, M.W., A.N. Boettiger, J.P. Bothma and M. Levine, *Shadow enhancers foster robustness of Drosophila gastrulation.* Curr Biol, 2010. **20**(17): p. 1562-7.

138. Dignam, J.D., R.M. Lebovitz and R.G. Roeder, *Accurate transcription initiation by RNA polymerase II in a soluble extract from isolated mammalian nuclei.* Nucleic Acids Res, 1983. **11**(5): p. 1475-89.

139. Ashburner, M.*, et al.*, *Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.* Nat Genet, 2000. **25**(1): p. 25-9.

140. Bateman, A., L. Coin, R. Durbin, R.D. Finn, V. Hollich, S. Griffiths-Jones, A. Khanna, M. Marshall, S. Moxon, E.L. Sonnhammer, D.J. Studholme, C.

Yeats and S.R. Eddy, *The Pfam protein families database.* Nucleic Acids Res, 2004. **32**(Database issue): p. D138-41.

141. Michalski, A., E. Damoc, J.P. Hauschild, O. Lange, A. Wieghaus, A. Makarov, N. Nagaraj, J. Cox, M. Mann and S. Horning, *Mass spectrometry-based proteomics using Q Exactive, a high-performance benchtop quadrupole Orbitrap mass spectrometer.* Mol Cell Proteomics, 2011. **10**(9): p. M111 011015.

142. Benjamini, Y. and Y. Hochberg, *Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing.* Journal of the Royal Statistical Society Series B-Methodological, 1995. **57**(1): p. 289-300.

143. Kulak, N.A., G. Pichler, I. Paron, N. Nagaraj and M. Mann, *Minimal, encapsulated proteomic-sample processing applied to copy-number estimation in eukaryotic cells.* Nat Methods, 2014. **11**(3): p. 319-24.

144. Untergasser, A., I. Cutcutache, T. Koressaar, J. Ye, B.C. Faircloth, M. Remm and S.G. Rozen, *Primer3--new capabilities and interfaces.* Nucleic Acids Res, 2012. **40**(15): p. e115.

145. Wuchty, S., *Scale-free behavior in protein domain networks.* Mol Biol Evol, 2001. **18**(9): p. 1694-702.

146. Havugimana, P.C.*, et al.*, *A census of human soluble protein complexes.* Cell, 2012. **150**(5): p. 1068-81.

147. Garber, M.*, et al.*, *A high-throughput chromatin immunoprecipitation approach reveals principles of dynamic gene regulation in mammals.* Mol Cell, 2012. **47**(5): p. 810-22.

148. Aldridge, S., S. Watt, M.A. Quail, T. Rayner, M. Lukk, M.F. Bimson, D. Gaffney and D.T. Odom, *AHT-ChIP-seq: a completely automated robotic protocol for high-throughput chromatin immunoprecipitation.* Genome Biol, 2013. **14**(11): p. R124.

149. Viturawong, T., F. Meissner, F. Butter and M. Mann, *A DNA-centric protein interaction map of ultraconserved elements reveals contribution of transcription factor binding hubs to conservation.* Cell Rep, 2013. **5**(2): p. 531-45.

Acknowledgements

I am grateful to Prof Dr Matthias Mann for the unique opportunity to conduct curiosity-driven research, and for his generous support throughout the process without which the thesis would not have come into existence.

Deep gratitude goes to Falk Butter for his patience and indispensable mentoring. I owe to him my fascination in the world of gene regulation, my technical skills, my scientific mindset and my own personal growth.

I am thankful to my Thesis Advisory Committee who made sure that everything was kept on track, and to the coordinators of the IMPRS programme who have made utmost effort to make us international students feel at home.

I thank Korbinian Mayr, Igor Paron, Gabi Sowa and Annette Michalski for their unfaltering support on the LC-MS technicalities, Jürgen Cox for his awe-inspiring bioinformatic wisdom, Bianca Splettstößer and Dennis Kappei for their learned wet lab expertise. Without the faithful columns, robust mass spectrometry, sophisticated computational algorithms, and optimized cell culture techniques, the experiments would have been a much harder quest to complete.

I extend my gratitude to all the members of the Mann department. Particularly, to Felix Meissner, Marco Hein, Daniel Hornburg, and Charo Robles, whose scientific outlook were exceptionally inspiring; to Alison Dalfovo and Theresa Schneider who made administration much less scary; to the *Sen* office for their support and friendship; finally, to the beloved members of the former *Evil* office, recalling the dark humour, blatantly incorrect wall comics, occasional submission shots, strokes of Kloppe and suggestive (mis)spellings: I only have to say **NA NA NA NA NA NA**.

An inspired scientist is a productive scientist. Thus I fondly acknowledge the friends at the ESME and the SMILE music ensembles for much of my inspiration outside the institute walls, as well as for their warm companionship. To Frances Hughes, Joseph Noelliste and Betsy Riley: thank you for your much-needed moral support in the days leading up to the defense. To Silvia Reddehase, Kornkamol Permachariyawong, Noppaodol Mekareeya, Anja Konschak, Camille Jourdain, Christian Eberl, Marlis Zeiler, Joanna McCarter, Dirk Walther, Annette Hellbach, Manuel Lehm, Kerstin Kinkelin, Leonie Mönkemeyer, Kirti Sharma, Kamila Jozwik, and Dan Rocapriore: thank you all for having brought positive touches to my life and given me the space to be over the last four years. I prostrate before Ajahn Brahmavamso, Niramala, and Thich Nhat Hanh for teaching me kindness and patience: ancient spiritual technology that brings more happiness than science can ever promise.

Finally, I thank my parents – to whom this study is dedicated – for showing me the very essence of unconditional love, and for an extremely rare and precious gift: birth as a human being.

Curriculum Vitae

Name:	Thanatip Viturawong
Country of origin:	Thailand
Date of birth:	16 November 1986
Address:	Guardinistraße 23, 81375 München
E-mail address:	viturawo@biochem.mpg.de

Education

2009 – 2014 **Doctoral thesis**
Max Planck Institute of Biochemistry
Munich, Germany
Thesis Title: *Scalable Quantitative Interaction Proteomics of Regulatory DNA Elements*

2008 – 2009 **Master of Natural Sciences**
University of Cambridge
Cambridge, United Kingdom
First class honours
Thesis Title: *Investigation of STAT factor binding in T helper cells*

2005 – 2008 **Bachelor of Arts**
University of Cambridge
Cambridge, United Kingdom
First class honours
Thesis Title: *The Role of Sch9 in TOR signaling*

2001 – 2005 **Harrow School**
Middlesex, United Kingdom
A Levels:
Triple Mathematics including Decision Mathematics and Statistics (AAA), Chemistry (A), Biology (A), Physics (A), Sixth Term Examination Papers in Mathematics (first class)

GCSEs:
Mathematics (A*), Biology (A*), Chemistry (A*), Physics (A*) History (A*), Geography (A*), English (A*), English Literature (A*), Religious Studies (A), Thai as First Language (A*)

Publications

Viturawong, T., F. Meissner, F. Butter and M. Mann, *A DNA-centric protein interaction map of ultraconserved elements reveals contribution of transcription factor binding hubs to conservation.* Cell Rep, 2013. **5**(2): p. 531-45.

Butter, F., L. Davison, T. Viturawong, M. Scheibe, M. Vermeulen, J.A. Todd and M. Mann, *Proteome-wide analysis of disease-associated SNPs that show allele-specific transcription factor binding.* PLoS Genet, 2012. **8**(9): p. e1002982.