

Evolutionary, structural and functional features of cellular signalling networks

Benjamin Lang

Homerton College
University of Cambridge



Dissertation submitted for the degree
of Doctor of Philosophy

September 2012

Advisor:

Dr. M. Madan Babu

Division of Structural Studies
Medical Research Council
Laboratory of Molecular Biology

Declaration of originality

I hereby declare that this dissertation is the result of my own work, and neither includes work submitted for the award of a previous degree nor work done in collaboration, except where specifically indicated in the text. I also declare that the text presented does not exceed a length of 60,000 words, excluding appendices, bibliography and figures.

Benjamin Lang

Cambridge, 27th September, 2012

Evolutionary, structural and functional features of cellular signalling networks

Benjamin Lang

Summary

The post-translational modification of proteins is a fundamental means of biological information processing, with important functions in development, homeostasis and disease. Post-translational modifications (PTMs) can dynamically diversify the proteome in response to intracellular and extracellular signals. Since thousands of modified residues as well as entirely new modification types have recently been discovered in proteins, elucidating their biological functions and identifying the protein components of these PTM systems is a fundamental problem.

Chapter 1 gives an overview of the types and known biological functions of different PTMs, as well as experimental methods used to detect them. Intrinsic disorder in proteins is introduced as a structural feature which may influence local evolutionary rates. Several examples of complex PTM signalling systems are then described.

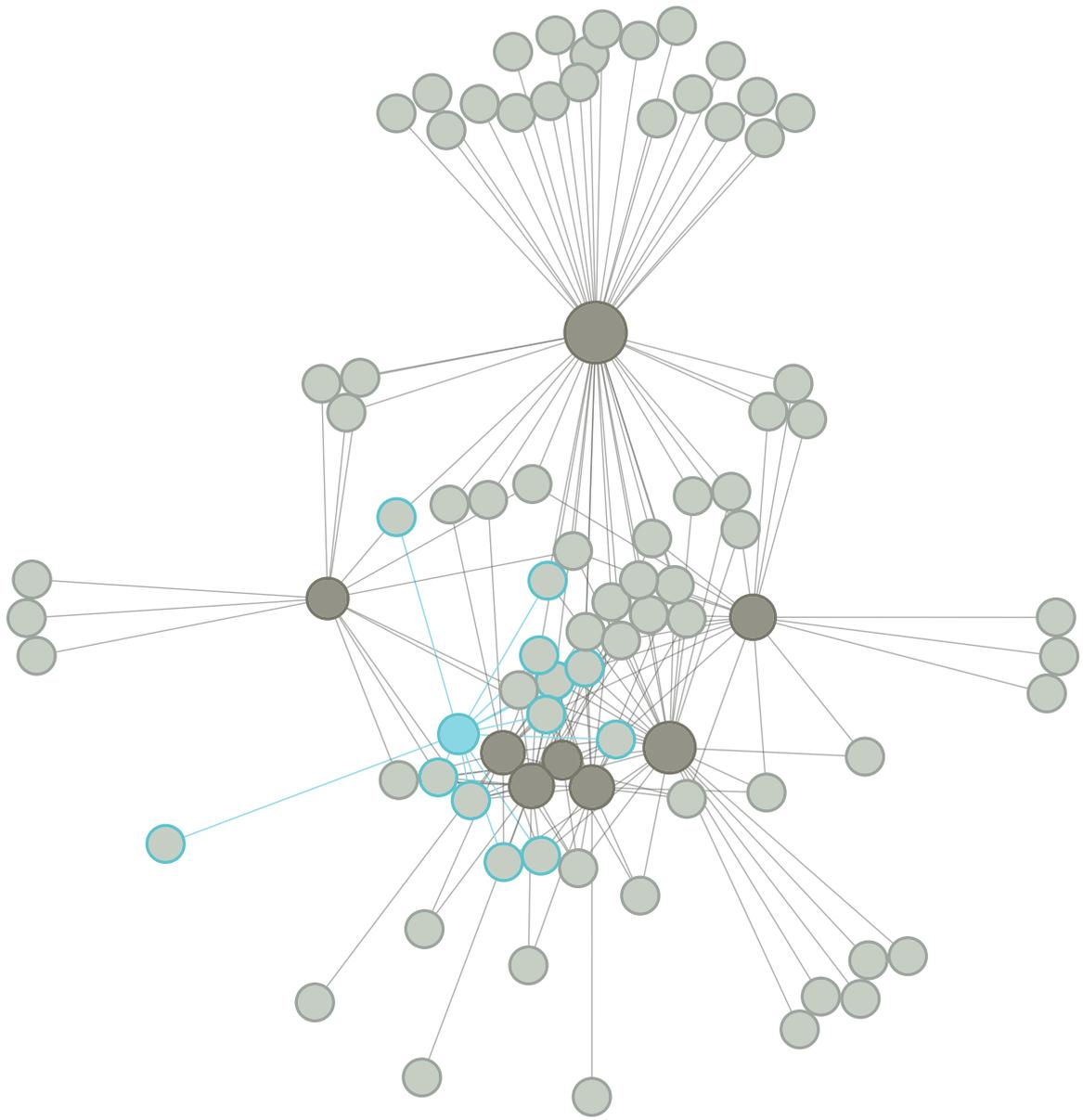
Chapter 2 presents a study of the evolution of modified amino acids in human proteins. By analysing sequence, polymorphism and mutation data at the species, population and individual levels, we observed significant evolutionary constraints on all PTM types for which extensive data was available, as well as overrepresentation of amino acids which mimic modified residues at equivalent positions.

Chapter 3 applies a framework for the identification of important components of PTM signalling systems to lysine acetylation. The proteins of this system were found to be similarly conserved as essential genes. Their evolutionary histories suggested a conserved origin in chromatin regulation, followed by functional diversification.

Chapter 4 extends the scope to signalling via transcriptional regulation, and presents a comprehensive overview of the interactome of the stem cell transcription factor Oct4. The results presented here facilitated characterisation of a novel post-translational modifier of Oct4, the glycosyltransferase Ogt.

Chapter 5 highlights my key findings from applying evolutionary and data integration approaches to signalling networks, and outlines their implications for the study of novel signalling systems and for their engineering in synthetic biology.

This dissertation therefore illuminates evolutionary, structural and functional principles of cellular signalling networks across species and within populations.



Acknowledgements & thanks

I would like to thank my mentor and friend Madan for first inspiring me to go into biological research as a visiting undergraduate student in 2006. Madan's great energy, warmth and insight have always been a source of inspiration for me.

The LMB added to this with its great diversity of research, as well as the exceptional speakers who visit regularly. Quantitative computational research while embedded in a hands-on biological institute is a fantastic experience.

My collaborations with Veronica van Heyningen and Patricia Yeyati in Edinburgh, Ana Pombo and Andre Möller in London, Jyoti Choudhary and Mercedes Pardo at the Sanger Institute, and with Andrew Travers and Harvey McMahon within the LMB have provided a constant stream of great new challenges which have allowed me to get an insight into many aspects of biology beyond my main projects.

I would also like to thank the members of our research group in Theoretical and Computational Biology, which includes the groups of Madan, Sarah Teichmann and Cyrus Chothia. The atmosphere has always been one of friendship, fun, enthusiasm and helpfulness.

I would like to particularly thank Guilhem, Marija, Sreeni, Tina, Andrew, Charles, Melis, Sven, Robin, AJ, Kai, Rekin's, Natalia, Marc, Valentina, Liora, Evia, Joe, Daniel, Filipa, Jörg, Sarath, Yod, Art, Subho, Jing, Muxin, Derek, Jake, Terry, Siarhei, Michael, Emmanuel, my second supervisor Daniela, Sarah, Cyrus, Fabrizio and of course anyone I should have missed for all their help, ideas and for making it all such a brilliant environment full of energy.

I would also like to thank the Herchel Smith Fund for my Research Studentship, which gave me a very generous amount of time and support to carry out my research, and which allowed me to visit several institutes in the USA.

My college, Homerton, provided a beautiful, scenic and even slightly monastic (being so far from town) backdrop and home for my PhD. I have many fantastic memories here, and I have met many brilliant people who I am sure will be my friends for life. I can only recommend it, and I will miss it greatly.

My brilliant and loving family seems almost too ever-present to mention here, but: *Danke* Marie-Luise, Albrecht and Christian for always being there for me!

Finally, my thanks also go to the University's Botanic Garden and modern laptop batteries for a very enjoyable daytime writing experience, which I could quite well get used to.

Benjamin Lang

Cambridge, 27th September, 2012

Table of contents

1. Introduction	1
1.1. Post-translational modifications (PTMs)	1
1.1.1. History and methods of PTM identification.....	4
1.1.2. Architecture of PTM signalling systems	7
1.1.3. Biological functions of individual PTM types	9
1.1.4. Disruption of PTM signalling in infection and disease	14
1.1.5. Chromatin regulation by PTMs.....	15
1.1.6. Additional examples of complex combinatorial PTM signalling	17
1.1.7. Structural contexts of PTM sites.....	18
1.2. Intrinsic disorder in proteins	18
1.2.1. Biological functions of intrinsically disordered regions	19
1.2.2. Intrinsic disorder and PTMs	21
1.2.3. Evolution of disordered regions	21
1.2.4. Protein-protein interaction networks.....	22
1.3. Motivation and overview of this dissertation	24
2. Evolution of post-translationally modified residues.....	26
2.1. Introduction.....	26
2.1.1. Initial motivation	27
2.1.2. Methodological considerations	29
2.2. Methods.....	33
2.2.1. Obtaining PTM sites.....	33
2.2.2. Predicting structural properties.....	34
2.2.3. Obtaining homology clusters	36
2.2.4. Removal of paralogs	37
2.2.5. Alignment of sequences.....	38
2.2.6. Phylogenetic reconstruction	39

2.2.7.	Assessing site conservation using variation scores.....	40
2.2.8.	Assessing evolutionary patterns at PTM sites	42
2.2.9.	Investigating the population and individual levels	44
2.3.	Results	45
2.3.1.	PTM sites are more conserved than structurally similar control residues between species.....	47
2.3.2.	Many PTM sites have ancient evolutionary origins.....	57
2.3.3.	Mutations at PTM sites may mimic or avoid the modified state.....	61
2.3.4.	PTM sites are more constrained than control residues at the population and somatic levels	65
2.4.	Discussion	68
2.4.1.	Quantifying PTM site conservation using conservation scores	68
2.4.2.	Analysis of PTM site evolution using conservation profiles	72
2.4.3.	Analysis of substitution patterns at PTM sites.....	73
2.4.4.	Constraints on PTM sites at the population and somatic levels.....	76
2.4.5.	Summary of key points and perspective.....	77
3.	The human lysine acetylation system.....	79
3.1.	Introduction.....	79
3.1.1.	Mechanisms and functional domains	79
3.1.2.	Biological functions.....	81
3.1.3.	Motivation and analysis outline.....	83
3.2.	Methods.....	84
3.2.1.	Identification of proteins with functions in lysine acetylation.....	84
3.2.2.	Analysis of evolutionary conservation.....	84
3.2.3.	Subcellular localisation.....	85
3.2.4.	Protein-protein interaction network.....	85
3.2.5.	Expression profiles and co-expression network	86
3.2.6.	Candidate ranking	86

3.3. Results	87
3.3.1. Identification of 100 lysine acetylation-related candidate proteins....	87
3.3.2. Nearly all candidate proteins are highly conserved in vertebrates.....	89
3.3.3. Many substrates are exclusively located in the cytoplasm.....	93
3.3.4. Functional classes tend to share physical interactions.....	94
3.3.5. Many candidates show significantly similar expression profiles.....	96
3.3.6. Candidate ranking	99
3.4. Discussion.....	102
4. The interaction network of Oct4	105
4.1. Introduction	105
4.2. Methods	107
4.2.1. Identification of Oct4 interactors	107
4.2.2. Integrating known protein-protein interactions	108
4.2.3. Expression at various differentiation stages.....	108
4.2.4. Transcriptional regulation	108
4.2.5. Sequence identity between mouse and human	109
4.2.6. Developmental, disease and cancer phenotypes	109
4.2.7. Misexpression in cancer types	109
4.2.8. Functional enrichment according to PANTHER annotation	110
4.2.9. Overrepresented domains	110
4.3. Results	110
4.3.1. Identification of Oct4 interactors.....	110
4.3.2. Integrating known protein-protein interactions	113
4.3.3. Expression changes in differentiated cells	114
4.3.4. Regulation by stem cell transcription factors	115
4.3.5. Sequence identity between mouse and human	117
4.3.6. Developmental, disease and cancer phenotypes	118
4.3.7. Misexpression in cancer types.....	121

4.3.8.	Functional enrichment according to PANTHER annotation	122
4.3.9.	Overrepresented domains	123
4.4.	Discussion	124
5.	Conclusions and implications	128
5.1.	Key findings	128
5.1.1.	Evolution of post-translationally modified residues	128
5.1.2.	The human lysine acetylation system	129
5.1.3.	The interaction network of Oct4	131
5.2.	Future directions	131
5.2.1.	Investigating the signalling functions of newly discovered PTMs	132
5.2.2.	Promising experimental methods	134
5.2.3.	Limitations to be addressed	135
5.2.4.	Engineering signalling systems in synthetic biology	137
6.	Publications	138
6.1.	Relevant publications	138
6.2.	Manuscripts in preparation	138
6.3.	Additional publications	138
6.4.	Publication reprints	140
7.	Bibliography	154

Acronyms and terminology

ASA	A ccessible s urface a rea (for solvent, in square ångströms)
ES cell	E mbyronic s tem cell
iPS cell	induced p luripotent s tem cell
JSD	J ensen– S hannon D ivergence (conservation score)
K-ac	Lysine (K) a cetylation
K-me	Lysine (K) m ethylation
LC-MS/MS	L iquid c hromatography followed by tandem m ass s pectrometry
MW	M olecular w eight
N-gly	N -linked g lycosylation (on asparagines)
PDB	P rotein D ata B ank
PPI	P rotein- p rotein interaction
PTM	P ost- t ranslational m odification
PTM site	One specific residue undergoing post-translational modification
rASA	relative solvent- a ccessible s urface a rea (fraction)
rvET	real- v alued E volutionary T race (conservation score)
S-p	S erine p hosphorylation
T-p	T hreonine p hosphorylation
Y-p	T yrosine p hosphorylation

1. Introduction

Cellular signalling networks function to process information and generate biological outcomes. They operate at multiple levels and in many different ways: for example, by integrating mammalian endocrine information from outside the cell through nuclear hormone receptors, leading to transcriptional responses; by integrating information from the cell surface through receptor tyrosine kinases, leading to autophosphorylation and downstream signalling; and by responding to any number of events within the cell, for instance by continuing a phosphorylation cascade. The net function of their interplay is to enable cellular and organismal homeostasis.

In this dissertation, I will explore the evolutionary, structural and functional features of cellular signalling networks, with a particular focus on the human organism and signalling changes which may lead to disease. Beginning with the evolution of post-translational signalling, the study then leads into stem cell biology and pluripotency.

In this introductory chapter, I will provide an overview of the types and known biological functions of post-translational protein modifications, as well as the methods that are used to detect them. I will then introduce intrinsic disorder as a structural feature of proteins which affects the evolutionary rates of amino acid residues. To illustrate the complex signalling systems enabled by post-translational modifications, I will first focus on the role of the well-studied histone modifications in eukaryotic gene regulation, and then present further examples of complex post-translational signalling mechanisms.

1.1. Post-translational modifications (PTMs)

Post-translational modifications (PTMs) greatly increase the diversity of amino acids available to the cell and enable rapid regulation of protein function. Nearly all proteins are post-translationally modified at some stage during their life cycle (Blom et al., 2004; Komander, 2009). More than 300 types of modification occur physiologically, ranging from the addition of small chemical moieties to entire peptide chains such as ubiquitin (Witze et al., 2007). These modifications greatly expand the diversity of the protein repertoire of an organism (Walsh et al., 2005). Although they also occur in prokaryotes, they are most extensively encountered in eukaryotes, with around 5% of the protein repertoire of higher eukaryotes being

dedicated to PTMs (Walsh et al., 2005). While changes in transcriptional regulation and alternative splicing require intervening synthesis steps before they can take effect, post-translational modifications can rapidly relay messages in the cell by acting on existing proteins in a matter of minutes or seconds (Lissanu Deribe et al., 2010).

To differentiate PTM types, at least three broad distinctions can be made: covalent vs. proteolytic, enzyme-catalysed vs. autocatalytic, and reversible vs. permanent modifications (Walsh et al., 2005; Bischoff and Schlüter, 2012). Limited proteolysis is a common mechanism for the functional maturation of many eukaryotic proteins. One example is the enzymatic removal of two C-terminal amino acids from α -tubulin in the assembly of microtubules, with relevance in processes such as axon regeneration (Ghosh-Roy et al., 2012). Other examples of limited proteolysis include the subcellular targeting of newly translated proteins to the endoplasmic reticulum and mitochondria. This cleavage may be carried out by specialised proteases, such as signal peptidases, or through autoproteolysis, as in the β subunits of proteasomes, and it is generally irreversible (Walsh et al., 2005). One example of a spontaneously occurring non-proteolytic PTM is asparagine deamidation (Robinson and Robinson, 2001). In contrast, most PTMs which involve the covalent addition of a chemical moiety to an amino acid side chain are reversible and enzyme-catalysed (Bischoff and Schlüter, 2012). These features together enable them to function as dynamic and informative signals. For covalent modifications, two additional classifications can be made according to the substrate amino acid and the reaction type (Table 1.1). Notably, 14 of the 20 most commonly occurring, genetically encoded amino acids are substrates for one or more enzyme-catalysed PTM types (Mittal et al., 2006; Mukherjee et al., 2007; Bischoff and Schlüter, 2012). The common mechanistic scheme in these enzyme-catalysed modification reactions is the covalent addition of an electrophilic chemical group to a nucleophilic amino acid side chain containing either a carboxyl, hydroxyl, amino or thiol group (Bischoff and Schlüter, 2012).

Table 1.1: Overview of known enzyme-catalysed PTMs by amino acid and reaction type.

	Cys	Asp	Glu	His	Lys	Met	Asn	Pro	Gln	Arg	Ser	Thr	Trp	Tyr
	C	D	E	H	K	M	N	P	Q	R	S	T	W	Y
Acylation	x				x						x	x		
ADP-ribosylation	x	x	x				x			x	x			
Carboxylation			x											
Disulfide formation	x													
Glycosylation							x				x			
Hydroxylation					x			x						
Isomerisation		x												
Mannosylation													x	
Methylation			x	x	x					x				
Nitration													x	x
Nitrosylation	x													
Oxidation	x					x								
Prenylation	x													
Phosphorylation	x	x		x							x	x		x
Sulfation														x
Transglutamination									x					

In addition to their diversity, PTMs have been shown to affect a large number of substrates, and increasingly precise data is becoming available on their individual substrate residues. Taking into account the entire dataset of PTM sites which I assembled in Chapter 2, one or more sites have been experimentally localised to specific residues for nearly three quarters of the human proteome (Gnad et al., 2007; Pang et al., 2010; Wang et al., 2010; Zhao et al., 2010; Gnad et al., 2010b; Dinkel et al., 2011; Hornbeck et al., 2012; UniProt Consortium, 2012). The functional characterisation of these PTM sites is an ongoing long-term challenge (Lissanu Deribe et al., 2010), and it is expected that the majority of PTM sites still remains to be discovered (Minguez et al., 2012). In many *in vitro* and structural studies, the use of a bacterial expression system may prevent the physiological post-translational modification of the protein being investigated (Sellick et al., 2011), and the functional significance of PTMs may therefore have remained unappreciated in certain cases.

In this dissertation, six of the major PTM types were chosen for study based on the availability of substantial residue-level data (Fig. 1.1). In descending order of the number of sites currently known in humans, these modifications are serine

phosphorylation, threonine phosphorylation, lysine ubiquitination, tyrosine phosphorylation, lysine acetylation and N-linked glycosylation on asparagines.

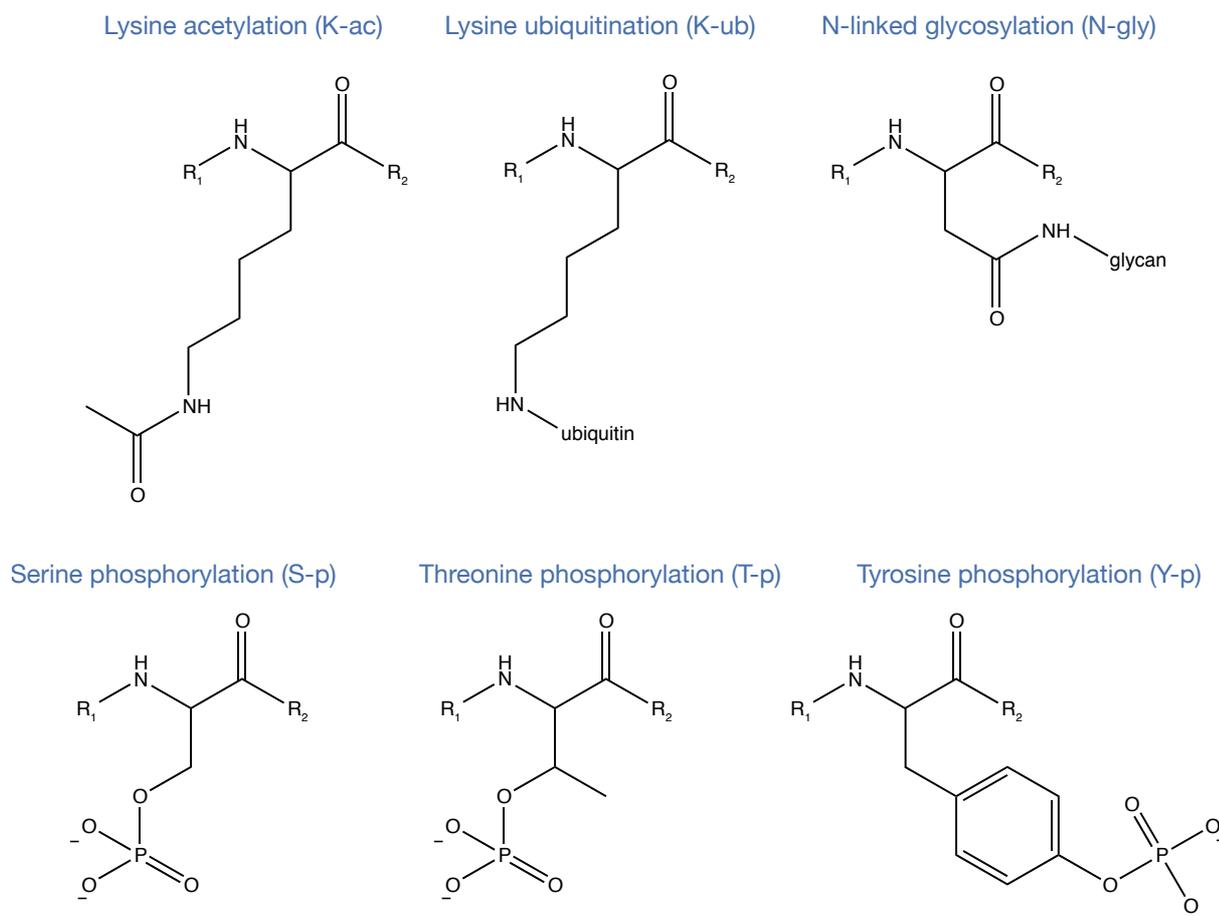


Figure 1.1: Structural overview of the six major PTM types studied in this dissertation. R_1 and R_2 respectively denote the N-terminal and C-terminal continuation of the peptide chain.

An overview of the precise numbers of known human sites for these modifications is given in Table 2.1.

1.1.1. History and methods of PTM identification

Biologically occurring derivatives of the standard amino acids, such as methylated histidine, were first described over 80 years ago (Ackermann et al., 1929). Although many of these are metabolic intermediates, some modified amino acids were later recognised as breakdown products from the catabolism of modified proteins (Baldwin and Carnegie, 1971; Cantoni, 1975).

An overview of methods used to identify PTM sites is given in Figure 1.2. Historically, instances of enzyme-catalysed protein modification have been identified using a range of biochemical methods such as chromatography, electrophoresis and

spectrophotometry (Baldwin and Carnegie, 1971; L'Italien and Laursen, 1979), as well as through assays for the donor cofactors participating in the reaction (Cantoni, 1975). These assays were later supplemented by antibody-based methods for the detection of a specific modification on a protein of interest, e.g. through a Western blot using a pan-methyllysine antibody. Though these methods are effective and provide semi-quantitative information on the extent of modification at a given site (modification penetrance), they are also time-intensive and therefore limited to focused, low-throughput studies.

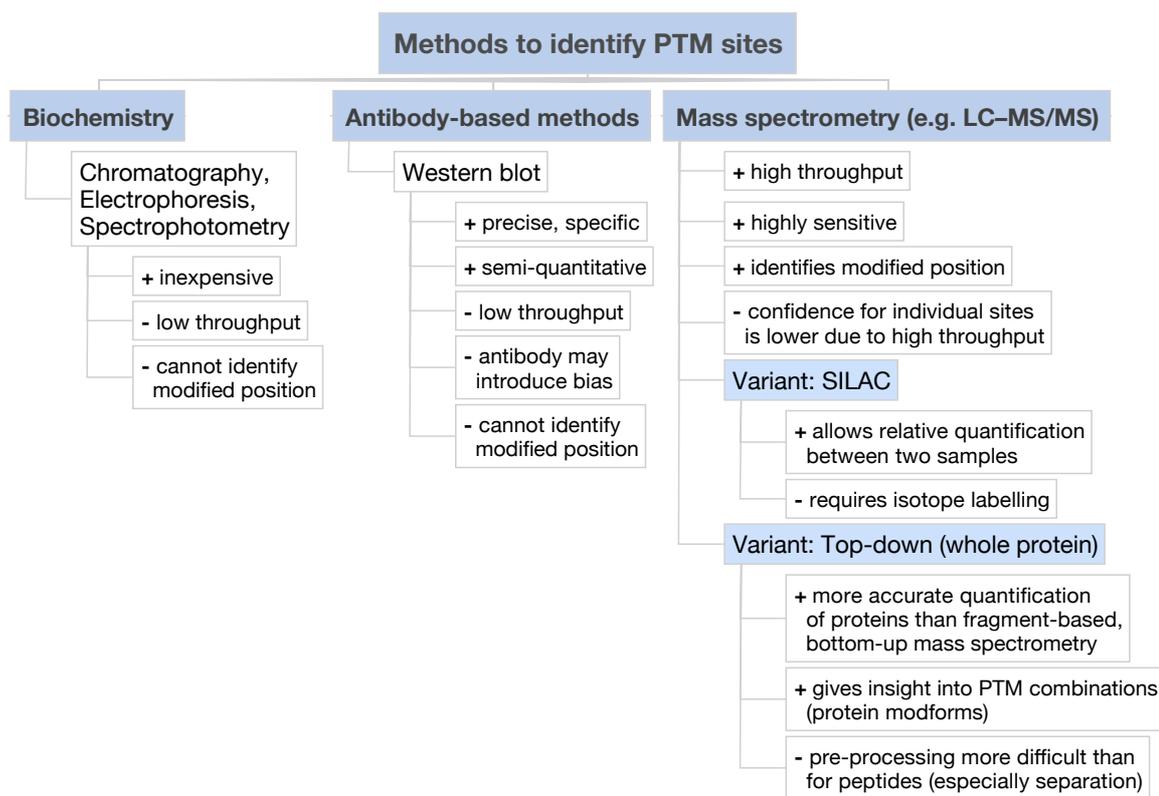


Figure 1.2: Overview of methods used to identify PTM sites, with an indication of their strengths and weaknesses (marked “+” and “-” respectively).

More recently, liquid chromatography combined with tandem mass spectrometry (LC-MS/MS) has emerged as the standard technique for high-throughput, localised PTM site detection (Bischoff and Schlüter, 2012), though numerous variants in instrumentation and methodology exist (Bogdanov and Smith, 2005; Bantscheff et al., 2007; Yates et al., 2009; Bantscheff et al., 2012). Put simply, mass spectrometry allows the identification of peptides by determining their mass-to-charge ratio (m/z), and comparing this figure to a genomically-derived search database of known or

predicted proteins. In tandem mass spectrometry, controlled fragmentation within the instrument yields further information about the identity of the peptide.

One notable application of mass spectrometry is the relative quantification of peptides, including modified peptides, in two samples using the SILAC method (stable isotope labelling by amino acids in cell culture). In this method, an unlabelled control sample is compared to a sample labelled with amino acids containing stable heavy isotopes, such as $^{13}\text{C}_6$ -Lys and $^{13}\text{C}_6^{15}\text{N}_4$ -Arg. These samples can be derived from cell cultures, or from in vivo labelling in model organisms such as *Drosophila melanogaster*, *C. elegans* and *R. norvegicus* (Ong et al., 2002; Yates et al., 2009).

Advances in sample preparation, mass spectrometric technology, experimental techniques and data processing have enabled the identification of several thousand PTM sites within a single experiment (Choudhary et al., 2009). These experiments often include an antibody-based purification step to enrich the sample for modified peptides prior to analysis. Chemical derivatisation of certain modified amino acids can further facilitate their detection through mass shift (Bischoff and Schlüter, 2012). In a similar vein, ubiquitination is generally detected as di-glycine remnants on lysines after trypsin digestion (Xu et al., 2010).

One disadvantage of high-throughput PTM site detection by mass spectrometry is that the confidence of detection at each site is lower than in a classical assay. It may occasionally produce false identifications of post-translational modifications due to factors such as spontaneous deamidation and carbamoylation occurring within a peptide (Stevens et al., 2008; Lin et al., 2010; Palmisano et al., 2012). In at least one study, results of lysine tri-methylation were discarded due to it being near-isobaric to lysine acetylation (Pang et al., 2010). Enrichment using PTM-specific antibodies helps to increase the overall number of PTMs identified in a given experiment, but antibodies may also introduce biases, due to their particular specificities for structural and sequence contexts (Levy et al., 2011). In addition, introducing the mass shifts caused by PTMs into the analysis pipeline complicates the peptide search by greatly increasing the size of the search database. This either limits the overall experiment to the consideration of a small number of PTM types at a time, or to a small number of distinct proteins in a highly pure sample (Baliban et al., 2010; Kersten et al., 2011; Muers, 2011; Tan et al., 2011). As a result, certain highly prevalent PTMs such as phosphorylation are routinely investigated, while others may

go undetected in standard data analysis workflows. This is further exacerbated by the sheer wealth of mass spectrometry data being generated, which may prevent researchers from performing more detailed, manual follow-up analyses of unexplained peaks, which may be present in as many as 50% of the good-quality spectra in an average experiment (Bischoff and Schlüter, 2012). Additionally, standard mass spectrometry does not address the possibility of fractional site occupancy, i.e. that only a part of the proteins in a sample may be modified at a certain residue, which has been observed for phosphorylation (Olsen et al., 2010).

Applying mass spectrometry to intact proteins followed by controlled fragmentation within the instrument (“top-down” mass spectrometry) may offer an alternative to the more conventional “bottom-up” fragmentation–identification approaches, which are most widely used. Bottom-up approaches are generally more successful at identifying proteins in complex mixtures, since peptide mixtures are easier to separate than intact proteins (Bogdanov and Smith, 2005; Yates et al., 2009). However, using intact proteins may allow more meaningful estimation of the absolute quantity of a given protein in a sample by alleviating biases introduced at the peptide level. In addition, the top-down approach may be an especially useful technique for the investigation of different protein “modforms”, i.e. the combinatorial states of multiple PTMs within the same protein (Siuti and Kelleher, 2007; Thomson and Gunawardena, 2009). This type of analysis has already been applied to individual histones (Tweedie-Cullen et al., 2012), and appears very promising for the future combinatorial study of protein modifications.

1.1.2. Architecture of PTM signalling systems

Post-translational modifications can be introduced and removed by specific enzymes, and recognised by specialised protein domains. PTM signalling systems are therefore commonly described as consisting of “writer”, “reader” and “eraser” proteins (Jin and Pawson, 2012). For instance, tyrosine phosphorylation signalling involves tyrosine kinases, SH2-domain proteins and tyrosine phosphatases as writer, reader and eraser proteins, respectively (Seet et al., 2006; Lim and Pawson, 2010). In a chromatin context, reader proteins are often termed “effector” proteins, since they may initiate chromatin remodelling based on the post-translational signals they encounter. These dynamic, reversible systems allow for elaborate combinatorial

logic and signal processing (Ruthenburg et al., 2007; Scott and Pawson, 2009; Lissanu Deribe et al., 2010; Cerone et al., 2011; Grabbe et al., 2011). Multiple modifying and erasing enzymes may compete and maintain a dynamic equilibrium of modified forms (Komander, 2009; Thomson and Gunawardena, 2009), and it has been shown conceptually that the steady-state concentration of protein “modforms” can be derived from the concentrations of the enzymes involved (Thomson and Gunawardena, 2009).

Reader domains such as SH2 domains (for tyrosine phosphorylation), bromodomains (for lysine acetylation) and chromo-like domains of the Royal family (for lysine and arginine methylation) recognise and bind one or more modified residues (Kasten et al., 2004; Seet et al., 2006; Shaw et al., 2007; Taverna et al., 2007; Morinière et al., 2009; Yap and Zhou, 2010; Chen et al., 2011). Other reader domains include small zinc fingers, for instance the PHD zinc finger domain, which can recognise methylated lysines (Mellor, 2006). Histone lysine demethylases often contain reader domains that allow them to read out the methylation state of a lysine residue different from their demethylation substrate (Upadhyay et al., 2011). At least 20 different reader domains for ubiquitination exist, covering the various ubiquitin chain types (Komander, 2009). Taken together, these modular recognition domains are important building blocks of eukaryotic cellular signalling.

For eraser proteins, relatively few serine/threonine phosphatases catalyse the dephosphorylation of thousands of substrates. Specificity is achieved here through regulatory subunits, which may direct the phosphatases to their targets (Shi, 2009). Deubiquitinases also occur in complexes that are thought to mediate their specificity (Sowa et al., 2009). For E2 ubiquitin ligases, specificity can be achieved by E3 ligases that function as adaptor proteins which mediate interactions with substrates (Zeng et al., 2008; Komander, 2009).

It appears to be a general pattern in many PTM signalling systems that the number of modifying enzymes, or of modifier-adaptor complexes as in the substrate targeting of ubiquitination, is much greater than the number of erasing enzymes (Komander, 2009; Bischoff and Schlüter, 2012; Liu and Nash, 2012). For phosphorylation, over 500 human protein kinases work alongside ~150 phosphatases that counteract them (Komander, 2009; Bischoff and Schlüter, 2012). In ubiquitination, one of two ubiquitin-activating enzymes (E1) charges one of 37

ubiquitin-conjugating enzymes (E2) by attaching the ubiquitin C-terminus to a catalytic cysteine residue. One of over 600 E3 ligases then acts as an adaptor protein in targeting the E2-Ub complex to ubiquitination substrates, while enhancing transfer reactivity (Catic et al., 2004; Komander, 2009; Pruneda et al., 2012). These numbers compare to ~85 human deubiquitinases.

1.1.3. Biological functions of individual PTM types

Post-translational modifications have been implicated in nearly all biological processes (Komander, 2009; Lissanu Deribe et al., 2010; Wagner et al., 2011). They can induce changes in the conformation, enzymatic activity, interactions, stability and subcellular localisation of a protein. This includes import and export from the nucleus and other organelles, as well as export from the cell through the secretory system (Nishi et al., 2011; Nussinov et al., 2012; Zhang et al., 2012). PTMs are well-suited for modulating interactions by creating or disrupting hydrophilicity matching between interaction interfaces (Hlevnjak et al., 2010). An overview of the biological functionality of selected PTMs is given below, as well as in Table 1.2.

Table 1.2: Overview of the biological functions of selected, common modification types.

PTM type	Residue(s)	Abbreviations used in this study	Biological functions
Phosphorylation	S, T, Y	S-p, T-p, Y-p	regulation and signal transduction in nearly all cellular processes
Ubiquitination	K	K-ub	targeting proteins for degradation, as well as the regulation of many other processes, with different polyubiquitin chain types having different functions which remain to be explored
Acetylation	K, rarely: S, T	K-ac	regulation of metabolic enzymes via activation, inactivation and destabilisation, as well as many other processes
N-linked glycosylation	N	N-gly	stabilisation of secreted proteins, often required for proper folding, introduced as part of the secretory pathway in the ER and Golgi
O-linked glycosylation	S, T		regulation of some cellular processes, particularly development
Methylation	K, R, rarely: E, H		regulation of several biological processes, including circadian rhythm, splicing and translation
SUMOylation	K		broad range of functions, including regulation of the cell cycle, heat shock response, and the folding and degradation of proteins
Prenylation (lipidation)	C		membrane localisation

Phosphorylation

Phosphorylation is the most prevalent post-translational modification in higher eukaryotes and is found in at least 7,400 substrate proteins in humans (UniProt Consortium, 2012). Estimates have been made that at least 70% of cellular proteins are phosphorylated at some stage, though the degree of fractional occupancy (the

fraction of protein copies that are modified) varies greatly (Olsen et al., 2010). In phosphorylation, a triply negatively charged phosphate group is attached to a serine or, more rarely, a threonine or tyrosine residue. In descending order of approximate frequency, protein phosphorylation can occur on serine (S), threonine (T), tyrosine (Y), cysteine (C), aspartate (D) and histidine (H) residues (Table 1.1).

Phosphorylation is introduced by kinases and removed by phosphatases. Many kinases phosphorylate both serines and threonines, and are termed Ser/Thr kinases. Most Ser/Thr kinases prefer to phosphorylate serines, while most phosphatases prefer to dephosphorylate threonines. This may account for the smaller number of observed threonine phosphorylation sites compared to serine. Most of these serine/threonine kinases choose their substrates based on a ± 4 amino acid motif around the substrate serine or threonine. Although the majority of kinases cannot accommodate a proline in the +1 position, around a quarter of serine/threonine phosphorylation sites are flanked by a +1 proline. This suggests that a small number of kinases may be generating a large number of phosphorylated residues, while others may have a smaller, more specialised set of substrates (Ubersax and Ferrell, 2007). Tyrosine phosphorylation is the rarest type of phosphorylation, and eukaryotic tyrosine phosphorylation appears to be an evolutionary innovation which underwent expansion in the common ancestor of choanoflagellates and metazoans. It has distinct kinases, phosphatases and recognition domains, and is considered promising for creating ectopic signalling circuits in species which do not have these enzymes and proteins endogenously (Lim and Pawson, 2010). The abundance of tyrosine phosphorylation appears to increase with organismal complexity, along with a decrease in proteomic tyrosine content, which may have evolved to avoid excessive nonspecific phosphorylation of tyrosines (Tan et al., 2009b; Lim and Pawson, 2010). Prokaryotes appear to have independently evolved tyrosine kinases, though prokaryotic tyrosine phosphorylation occurs very rarely (Grangeasse et al., 2012).

Relatively low conservation of phosphorylation sites has been reported between bacterial species (Soufi et al., 2008). This may either point to a high degree of plasticity, or to a low degree of experimental coverage of these bacterial phosphoproteomes. More generally, changes in phosphorylation are thought to be a large contributor to phenotypic divergence, which indicates that information on kinase-substrate associations cannot be transferred confidently between species

(Beltrao et al., 2009; Moses and Landry, 2010). This is likely to be the case for certain other PTMs as well.

Ubiquitination

Ubiquitination is involved in almost every cellular process (Komander and Rape, 2012). Most typically, it targets proteins for degradation by the proteasome, through the addition of a chain of at least four K48-linked copies of ubiquitin, which is a highly conserved small protein of ~8 kDa. Ubiquitination is catalysed by a cascade of E1 (activating), E2 (conjugating) and E3 (ligating) ubiquitin ligases (Lu et al., 2008). HECT domain E3 ligases participate directly in the ubiquitination reaction, while RING/U-box domain E3 ligases serve purely as specificity-mediating adaptor proteins (Lu et al., 2008). The functional investigation of ubiquitination is complicated by several factors. In proteasomal degradation, polyubiquitination and the subsequent degradation of the substrate have been reported to follow rapidly (Pierce et al., 2009). Furthermore, the understanding of different ubiquitin chain types (which differ in their functions) is still not fully worked out (Komander, 2009; Ikeda et al., 2010). Ubiquitin can also form mixed chains with other ubiquitin-like modifiers such as SUMO (Danielsen et al., 2010). In addition to targeting substrates to the proteasome for degradation, ubiquitin can also directly mediate destabilisation and unfolding of proteins (Hagai and Levy, 2010).

Acetylation

Acetylation is part of a larger group of post-translational modifications referred to as acylation reactions. In these reactions, an acylating enzyme uses an acyl-CoA donor to transfer the acyl group to a substrate residue. Donors include acetyl-CoA, succinyl-CoA, malonyl-CoA, palmitoyl-CoA and a large number of others (Lin et al., 2012). Acylation can occur on lysines (K) and cysteines (C), and acetylation has also been described on serines (S) and threonines (T) (Table 1.1). Many of the donor molecules used, such as acetyl-CoA and succinyl-CoA, are highly common metabolic intermediates, and it has been found that many metabolic enzymes seem to be regulated by acetylation (Lin et al., 2012). Succinylation and malonylation also frequently occur on metabolic enzymes (Lin et al., 2012). The effects of lysine acetylation have been described as leading to either activation, inactivation or destabilisation of metabolic enzymes, which illustrates a similarly broad functional

range for acetylation as for phosphorylation (Yu et al., 2008; Zhao et al., 2010; Jiang et al., 2011; Lin et al., 2012). Acetylation occurs in both eukaryotes and prokaryotes. For instance, the half-life of the RNase R exoribonuclease in *E. coli* is increased by enzymatic acetylation at one lysine residue, thereby increasing its concentration 3- to 10-fold under stress conditions (Liang et al., 2011). A more in-depth description of lysine acetylation is given in the introduction to Chapter 3, which investigates the human lysine acetylation signalling system.

Glycosylation

Glycosylation is introduced in the endoplasmic reticulum or Golgi apparatus as part of the secretory pathway, and it can occur as N-linked, O-linked or C-linked glycosylation. N-glycosylation occurs on asparagines, O-glycosylation occurs on serines and threonines, and C-glycosylation occurs very rarely on tryptophan. The most common type is N-glycosylation of the highly predictive N-X-S motif, where X is any amino acid, and it is estimated that between one third and one half of eukaryotic secreted proteins are N-glycosylated in the ER (Walsh et al., 2005; Hart and Copeland, 2010). A more recent proteomic study described the primary motif as “N-!P-[S]T-!P” where !P is not proline, occurring in nearly all cases (Zielinska et al., 2010). The same study found an extremely high reproducibility in biological replicates, indicating near-total occupancy of N-glycosylation sites across all copies of a protein. However, the architecture of the branched polysaccharide tree on an individual protein can vary greatly depending on apparently stochastic encounters with glycosylation enzymes (Hart and Copeland, 2010). Its general functions include stabilisation and ensuring the proper folding of many proteins (Walsh et al., 2005; Shental-Bechor and Levy, 2009). O-glycosylation by N-acetylglucosamine (O-GlcNAc) is another common type of glycosylation which has been suggested to have dynamic signalling roles similar to phosphorylation (Hart et al., 2007), and has been shown to be important in zebrafish development (Webster et al., 2009).

Additional PTM types

Arginine methylation of spliceosomal proteins by PRMT5 has been implicated as a potential link between circadian rhythm and alternative splicing (Sanchez et al., 2010), and demethylation of lysines on MYPT1 by LSD1 leads to cell cycle progression (Hamamoto et al., 2010). In the green alga *Chlamydomonas reinhardtii*,

arginine methylation by PRMT1 in response to phototrophic growth conditions activates the cytosolic translation repressor protein NAB1, thereby repressing translation of components of the light-harvesting complex (Blifernez et al., 2011). In plants, rubisco, the essential photosynthetic enzyme, is methylated by RLSMT, a SET-domain lysine methyltransferase (Raunser et al., 2009). A number of other plant proteins are also methylated (Feng et al., 2009). Although the currently known number of human lysine and arginine methylation sites is in the low hundreds, the methyl group donor S-adenosyl-L-methionine (SAM) is the most widely used enzyme substrate after ATP, potentially suggesting that many methylation sites may remain unidentified (Cantoni, 1975; Paik et al., 2007; Sprung et al., 2008). For lysine acetylation, the generation of antibodies of high specificity appears to remain a technical challenge, and the identification of acetylated proteins is complicated by the high concentrations of histones (Guan et al., 2010). Similar problems may exist for lysine and arginine methylation, both of which also occur on histones (Kouzarides, 2007b).

The SUMO proteins, a subfamily of ubiquitin-like modifiers, have a similarly broad range of functions as ubiquitin, including cell cycle regulation, heat shock response, the folding and degradation of proteins (Golebiowski et al., 2009; Creton and Jentsch, 2010).

Lipidation is a group of modifications which are often involved in localising proteins to membranes. S-palmitoylation, a thioester-linked fatty acid modification, can effect this reversibly (Linder and Deschenes, 2007; Maurer-Stroh et al., 2007; Yount et al., 2010; Hang et al., 2011).

Certain reactions can remove a PTM by irreversibly converting a modified amino acid side chain into an atypical derivative. For instance, arginine methylation can be disrupted by arginine deimination, an irreversible modification which produces the uncharged amino acid citrulline (Guo et al., 2011). Similarly, lysine can be irreversibly hydroxylated by Jmjd6 (Webby et al., 2009). Hydroxylation is an oxygen-dependent reaction, and proline and lysine hydroxylation have been found to regulate the activity and half-life of hypoxia-inducible factor alpha (HIF α) (Schofield and Ratcliffe, 2004). In addition, phosphothreonine lyases can convert phosphorylated threonine into dehydrobutyrine (Ribet and Cossart, 2010). Taken together, these types of

reactions provide alternative pathways for permanently altering and inactivating a PTM site by introducing an uncommon amino acid.

1.1.4. Disruption of PTM signalling in infection and disease

Human diseases frequently involve defects in post-translational modifications, and therapeutics targeted at PTM signalling proteins, such as histone deacetylase inhibitors, are being developed as interventions for human diseases such as cancer (Bradner et al., 2010). The leukemogenicity of the oncogenic fusion protein AML1-ETO is dependent on its site-specific lysine acetylation by p300, suggesting that inhibition of this acetyltransferase could be a potential therapeutic approach (Wang et al., 2011a). Phosphorylation and ubiquitination are also implicated in α -synuclein aggregation in Parkinson's disease (Hejjaoui et al., 2010; Lu et al., 2011).

Pathogens can make use of PTMs to influence or disrupt host cell signalling (Broberg and Orth, 2010; Ribet and Cossart, 2010). A notable example is the bacterium *Yersinia pestis*, which caused several historic plague pandemics, including the Plague of Justinian, the devastating "Black Death" in medieval Europe and a worldwide third outbreak originating in Yunnan in the mid-19th century (Haensch et al., 2010). One of its toxins, YopJ, acetylates serine and threonine phosphosites on various host proteins which are part of the MAPKK and NF- κ B pathways, thereby inhibiting the host inflammatory response and inducing apoptosis of immune cells (Mittal et al., 2006; Mukherjee et al., 2006). An acetyltransferase related to *Yersinia* YopJ, termed HopZ1a, has been found in the plant pathogen *Pseudomonas syringae*, where it interferes with the host microtubule network, blocks secretion and suppresses cell wall-mediated defence (Lee et al., 2012). HopZ1a autoacetylates a lysine residue, and is hypothesised to acetylate α -tubulin K252, though its mechanism of action is still unclear. Further examples include the manipulation of ubiquitin and ubiquitin-like modifier signalling (Ribet and Cossart, 2010; Perrett et al., 2011), or direct dephosphorylation of host serines, threonines and tyrosines (la Puerta et al., 2009; Erazo et al., 2011). The apicomplexan parasites *Toxoplasma gondii* and *Plasmodium falciparum* not only have phosphoproteomes similar in size to multicellular eukaryotes, but also a small number of endogenous tyrosine phosphorylation sites (Treeck et al., 2011). At least 20 kinases of

Plasmodium falciparum are thought to be exported into the host cell, where they may interfere with host tyrosine phosphorylation signalling (Nunes et al., 2007).

During infection by *Legionella pneumophila*, intracellular vesicle trafficking is altered due to the phosphocholination and adenylation of Rab proteins by virulence factors, which affects their cycling between membrane-bound and cytosolic states. It has been suggested that similar mechanisms may be employed for cycling even in the absence of infection (Oesterlin et al., 2012). In HIV infection, two important coreceptors bound by HIV during cell entry, CCR5 and CXCR4, are endogenously sulfated on tyrosines, and this modification is required for infection (Farzan et al., 1999; Schnur et al., 2011). Pharmacological interventions which modulate post-translational modifications may therefore represent a promising therapeutic approach in a number of diseases.

1.1.5. Chromatin regulation by PTMs

The nucleosome system, which compacts and regulates chromatin structure by dynamically wrapping DNA around histone octamers, is one of the best examples of the complexity and flexibility of post-translational signalling in higher eukaryotes (Kouzarides, 2007a; Bannister and Kouzarides, 2011; Bartke and Kouzarides, 2011). The recognition of its importance in gene regulation is one of the catalysts which has spurred recent PTM research. Classically, chromatin has been considered divided into transcriptionally active euchromatin and condensed, inactive heterochromatin. It has more recently been suggested that the individual modification marks on histones may combinatorially form a “histone code”, and it has been demonstrated that *Drosophila* chromatin is structured into no less than five distinct chromatin types, which are characterised by a particular PTM state as well as by proteins which recognise specific PTMs (Filion et al., 2010). Histone PTMs regulate chromatin conformation and gene expression, and have extensive developmental roles (Campos and Reinberg, 2009; Nottke et al., 2009).

However, new histone modifications are still being discovered, and their functional elucidation is far from complete (Muers, 2011). Recent in-depth proteomics studies of nucleosomes have uncovered that modifications on the individual histone proteins not only include classical modifications such as serine and threonine phosphorylation, lysine acetylation, and lysine and arginine methylation, but also

include several atypical modifications such as serine and threonine acetylation, tyrosine hydroxylation, as well as lysine butyrylation, crotonylation, and propionylation (Tan et al., 2011; Tweedie-Cullen et al., 2012). The fact that these modifications are still being discovered in a regulatory system as intensively studied as chromatin, along with indications of their dynamic regulation and functionality, could indicate that atypical PTMs also remain undiscovered in other important signalling systems.

At least one type of PTM, the lysine and arginine methylation system centred on histones, is thought to be of prokaryotic evolutionary origin. It most likely first functioned in bacterial peptide methylation, and in hydroxylation systems used in the production of secondary metabolites (Aravind et al., 2011). The last common eukaryotic ancestor most likely already had several lysine and arginine methyltransferases as well as chromodomain-like reader proteins in place, while demethylases are likely to have evolved later, along with cytosine methylation in DNA (Aravind et al., 2011).

As mentioned above, the combination of different histone PTMs may form a combinatorial “histone code”. To illustrate this, lysine methylation can either be transcriptionally activating or repressive, based on the methylation state (mono-, di- or tri-methylation) and the residue modified (Bannister and Kouzarides, 2011). Histone lysine demethylases often contain reader domains that allow them to read out the methylation state of a lysine residue different from their demethylation substrate, potentially allowing them to remove opposing signals, e.g. repressive methylation marks in the presence of activating methylation (Upadhyay et al., 2011).

In histones, there is extensive combinatorial “cross-talk” between PTMs at different residues and of different types, including acetylation, phosphorylation, methylation, sumoylation, ubiquitination and proline isomerization (Latham and Dent, 2007; Ruthenburg et al., 2007; Fischle, 2008; Suganuma and Workman, 2008; Shukla et al., 2009; Filippakopoulos et al., 2012). In addition to PTMs, certain chromatin complexes can integrate the methylation state of DNA, whose methylation at cytosines is generally transcriptionally repressive, but has been found insufficient in predicting gene activity on its own (Filion et al., 2006; Vu et al., 2006; Bartke et al., 2010). In non-histone proteins, cross-talk exists e.g. between ubiquitination and

methylation (Wu et al., 2011), and between ubiquitination and phosphorylation (Huang et al., 2012).

In the chromatin transcriptional regulatory system, two additional layers of complexity are added by the constitution of chromatin remodelling complexes, and by non-coding RNAs (Guttman et al., 2011). The activity of chromatin remodellers can be modulated through the formation of variant complexes, differing in their components (Ho and Crabtree, 2010). This is reminiscent of the modulation of ubiquitin ligases by adaptor proteins, which determine substrate targeting (Komander, 2009). Large intergenic non-coding RNAs (lincRNAs) have additionally been shown to affect the chromatin state by regulating chromatin-modifying complexes, and thereby to affect gene expression (Khalil et al., 2009).

1.1.6. Additional examples of complex combinatorial PTM signalling

In addition to the fascinatingly complex regulatory system centred on eukaryotic chromatin, post-translational signals also mutually influence each other extensively on other proteins (Minguez et al., 2012). This can include cases where PTMs compete for the same sites, as well as more indirect associations. A recent study in the genome-reduced obligate parasitic bacterium *Mycoplasma pneumoniae* found extensive changes in lysine acetylation when phosphorylation was perturbed, and *vice versa*, thereby identifying cross-talk between the two PTMs in a prokaryote (van Noort et al., 2012). Functional interplay of lysine methylation and acetylation, where acetylation of one lysine obstructs the methylation of two closely adjacent lysines, has also been reported in the regulation of mammalian NF- κ B (Yang et al., 2010).

Acetylation significantly overlaps with ubiquitination sites, with competition occurring at around 30% of lysine acetylation sites and around 9% of ubiquitination sites (Wagner et al., 2011). This represents a possible mechanism for the regulation of half-life by preventing addition of a K48-linked polyubiquitin chain (Grönroos et al., 2002; Jin et al., 2004; Caron et al., 2005; Danielsen et al., 2010; Wagner et al., 2011). As an example of interplay between lysine methylation and phosphorylation, methylation of lysine residues on the retinoblastoma protein pRb prevents phosphorylation of adjacent residues and thereby blocks cell cycle progression (Carr et al., 2010). In G9a, a histone methyltransferase, methylation of one lysine residue creates a binding site for HP1 and inhibits phosphorylation (Sampath et al., 2007).

Lysine acetylation also regulates gluconeogenesis by promoting PEPCK1 degradation via recruiting the UBR5 ubiquitin ligase (Jiang et al., 2011).

1.1.7. Structural contexts of PTM sites

Where do modified residues occur in proteins? Regulation by PTMs has been described as occurring either orthosterically or allosterically. Orthosteric effects of PTMs are achieved directly at the site modification, e.g. through direct recognition by a reader domain or through disruption of a binding interface, while allosteric effects are achieved through conformational changes of the protein (Nussinov et al., 2012). PTM types may differ in their effects, acting either mostly electrostatically or sterically (Nussinov et al., 2012).

An important factor for the specificity of kinases and phosphatases is the sequence context they recognise (Ubersax and Ferrell, 2007). Investigations of the context of phosphorylated sites revealed that they most frequently occur in intrinsically disordered regions, which are at least partially unfolded parts of proteins (Iakoucheva et al., 2004).

1.2. Intrinsic disorder in proteins

Protein folding has been described as a multi-dimensional energy landscape, with minima signifying folded states (Hartl and Hayer-Hartl, 2009). However, many proteins may never reach a fully folded state, and may instead carry out their functions in an at least partially disordered native state (Dunker et al., 2008; Gsponer and Babu, 2009; Babu et al., 2012; Lobanov and Galzitskaya, 2012). These intrinsically disordered proteins (IDPs) are especially common in eukaryotes (Gsponer and Babu, 2009). They tend to display relatively short half-lives and their expression is tightly regulated, both of which are advantageous characteristics for their frequent involvement in signalling roles (Gsponer et al., 2008). One canonical type of intrinsically disordered region are disordered linkers, which can connect fully folded domains while allowing conformational flexibility of the protein. This type of arrangement is found in many multi-domain transcription factors, which need to maintain specific DNA contact while recruiting additional proteins, and particularly in the stem cell transcription factors investigated in Chapter 4 (Frankel and Kim, 1991; Xue et al., 2012). Intrinsic disorder is also frequently found at protein interaction

interfaces, where the disordered regions may undergo a disorder-to-order transition upon binding (Fong et al., 2009).

Based on their conservation, intrinsically disordered regions can be classified into three groups: i) fast-evolving disordered regions whose presence is conserved (e.g. linker regions), ii) sequence-conserved disorder (e.g. interaction motifs that fold upon binding), and iii) non-conserved disorder of unclear importance (Bellay et al., 2011). Intrinsic disorder facilitates evolutionary processes such as alternative splicing, domain shuffling and protein modularity (Mittag et al., 2010; Buljan et al., 2012). In these processes, intrinsic disorder may permit increased functional and regulatory diversity while avoiding the constraints posed by the folding of structured domains (Romero et al., 2006). The tissue-specific splicing of disordered protein-protein interaction motifs is one common mechanism which benefits from this flexibility (Buljan et al., 2012; Ellis et al., 2012). Protein-protein interactions via disordered interfaces can also evolve more rapidly than structured interactions (Mosca et al., 2012).

It has been suggested that disordered regions can constitute functional and evolutionary units of similar importance to folded domains without being structured (Tompa et al., 2009; Babu et al., 2012). For instance, these regions may contain important interaction motifs, including linear motifs (LMs), short linear motifs (SLiMs) and molecular recognition features (MoRFs) (Fuxreiter et al., 2007; Stein and Aloy, 2008; Edwards et al., 2011). Importantly, SLiMs frequently contain PTM sites (Davey et al., 2011). In fact, most phosphorylation sites occur in intrinsically disordered regions of proteins, and phosphorylation is enriched outside of classical functional domains (Iakoucheva et al., 2004; Gnad et al., 2009; Holt et al., 2009; Landry et al., 2009; Ba and Moses, 2010). Highlighting the functional importance of disordered regions, it has been suggested that the pathogenicity of disease-associated mutations in disordered regions may sometimes be caused by transition to structure (Vacic and Iakoucheva, 2012).

1.2.1. Biological functions of intrinsically disordered regions

Approximately one third of eukaryotic proteins contain extended intrinsically disordered regions (Ward et al., 2004), and intrinsically disordered regions are a feature of many eukaryotic proteins involved in protein-protein interactions and

regulatory and signalling processes (Xie et al., 2007b; Dunker et al., 2008; Stein and Aloy, 2008; Gsponer and Babu, 2009). Due to their increased conformational flexibility compared to structured proteins, disordered proteins are able to interact with a broader range of partners. They frequently constitute hubs in interaction networks, and may more easily be reused in multiple biological processes (Gsponer and Babu, 2009). These processes include cell differentiation, transcriptional regulation, the cell cycle, mitosis, apoptosis, mRNA processing, splicing and protein transport. Many developmental regulators, ribonucleoproteins, ribosomal proteins, hormones, growth factors, cytokines and neuropeptides are intrinsically disordered (Xie et al., 2007b). Disordered regions have been reported to be important in displaying post-translationally modified sites, such as phosphorylation sites (Gsponer and Babu, 2009; Mittag et al., 2010; Brown et al., 2011). PTM-site containing regions such as these can be acquired as regulatory modules in evolution (Fantini et al., 2010).

Small interaction domains such as zinc fingers frequently occur within extensively disordered regions (Vucetic et al., 2007). The functional importance of disordered regions in protein-protein interactions is further illustrated by hypoxia-inducible factor 1 alpha (HIF-1 α), which binds p53 through a disordered interaction domain when inducing hypoxic p53-mediated apoptosis (Sánchez-Puig et al., 2005). In addition, p53 itself has disordered N- and C-termini which serve as regulatory domains and interaction interfaces (Römer et al., 2006; Joerger and Fersht, 2007; Wells et al., 2008; van Dieck et al., 2009). Phosphorylation-induced interaction of 14-3-3 reader proteins with its C-terminus activates p53's DNA binding activity (Rajagopalan et al., 2008). Lysine acetylation and phosphorylation of the C-terminal domain are further thought to modulate p53's DNA affinity by reduction of its overall positive charge, resulting in more specific DNA interactions (Melero et al., 2011). Disordered regions containing positive charges appear to be of general importance in protein-DNA interactions, where they increase search efficiency by maintaining nonspecific interactions with the phosphate backbone of DNA while allowing various search mechanisms such as sliding, hopping, fly-casting and brachiation (Vuzman and Levy, 2010; Vuzman et al., 2012).

1.2.2. Intrinsic disorder and PTMs

Intrinsic disorder of proteins is a measure of flexibility, and has been used as a predictive feature in the prediction of modification sites, principally based on the assumption that flexible regions of proteins may be more accessible to a modifying enzyme (Iakoucheva et al., 2004; Daily et al., 2005; Tompa, 2005; Xie et al., 2007a; Dunker et al., 2008). As mentioned above, phosphorylation is statistically enriched in disordered regions (Iakoucheva et al., 2004; Gnad et al., 2009; Holt et al., 2009; Landry et al., 2009; Ba and Moses, 2010), as are most other PTMs (Xie et al., 2007a). Disordered short linear motifs also frequently contain PTM sites (Davey et al., 2011). It has further been suggested that the nonspecific DNA interactions of disordered regions in transcription factors described above can be modulated by modifications such as phosphorylation and acetylation (Vuzman et al., 2012).

However, intrinsic disorder at the substrate residue does not appear to be a necessary requirement for PTMs to occur, although it has been described as quite important (Iakoucheva et al., 2004; Daily et al., 2005; Tompa, 2005; Xie et al., 2007a; Dunker et al., 2008). Conceptually, adaptor proteins such as E3 ubiquitin ligases could mediate contact between a charged E2 ligase and its substrate site without particular site specificity on the part of the E2 ligase. Moreover, N-glycosylation and lysine acetylation have been reported to show enrichment in structured regions (Choudhary et al., 2009; Zielinska et al., 2010), as have formylation, protein splicing, oxidation, and covalent attachment of quinones and organic radicals (Xie et al., 2007a). It has further been reported that certain PTMs, including lysine methylation, are not primarily found on the surface of proteins, but can fall within its core (Pang et al., 2007). This is in accord with reports that modification events can result in large structural changes, including domain unfolding and allosteric transitions (Díaz-Moreno et al., 2009; Nussinov et al., 2012; Shental-Bechor et al., 2012; Yuchi et al., 2012). Similarly, it has been shown that mutation of a single amino acid on the periphery of a designed protein's core can lead to fold switching (He et al., 2012b).

1.2.3. Evolution of disordered regions

Intrinsically disordered regions tend to display faster evolutionary divergence than the rest of the protein, most likely since they are under less structural constraint from intramolecular contacts (Brown et al., 2002; 2010; 2011). In addition, solvent-

exposed protein regions can accept higher rates of mutation than buried regions, and disordered regions tend to be more hydrophilic (Sasidharan and Chothia, 2007; Awile et al., 2010). It has been suggested that at least some disordered regions are mostly conserved for their overall charge and structural flexibility (Tompa and Fuxreiter, 2008; Mittag et al., 2010). Providing a connection between intrinsic disorder, PTM sites and evolution, it has been suggested that mutations at PTM sites may be significant drivers of phenotypic evolution between species (Moses and Landry, 2010).

1.2.4. Protein-protein interaction networks

As mentioned above, post-translational modifications can affect the physical interactions a protein is able to make in the cell, either allosterically or by occurring directly within an interaction interface (Nishi et al., 2011). Interfaces which enable the interaction between proteins are generally hydrophobic with complementarity mediating specificity (Chothia and Janin, 1975; Jones and Thornton, 1996), and as mentioned above, certain PTMs such as phosphorylation and acetylation alter the charge of residues, thereby enabling the regulation of protein-protein interactions (PPIs) and potentially fine-tuning their specificity. Protein-protein interaction interfaces are especially interesting since they can be included or excluded by alternative splicing and alternative promoters (Weatheritt et al., 2012). Alternative exons are sometimes expressed in a tissue-specific manner, and have been found enriched in modifiable residues as well as interaction motifs (Buljan et al., 2012; Weatheritt et al., 2012).

Both of these mechanisms, post-translational modification and the generation of alternative mRNAs, can therefore contribute to the rewiring of protein-protein interactions. These interactions are frequently represented as large networks covering much of an organism's proteome, with nodes representing proteins and edges their physical interactions (Levy and Pereira-Leal, 2008; Mosca et al., 2012). A more indirect approach is the use of genetic interaction data, which have also been used to arrive at similar networks and appear to be predictive of physical interactions (Roguev et al., 2013).

It is important to note the difference between obligate and transient protein-protein interactions. Strong interactions are crucial for the functioning of many obligatory

complexes in the cell, especially in housekeeping processes such as transcription and replication, while weaker transient interactions are less well conserved between species (Jones and Thornton, 1996; Mosca et al., 2012). Intrinsically disordered regions are a common characteristic of proteins involved in a large number of interactions, often termed “hub” proteins, and the disordered regions are thought to play a key role in the proteins’ ability to interact with a large variety of binding partners (Mosca et al., 2012). One example of these is the tumour suppressor protein p53, whose disordered N- and C-termini allow it to participate in a large number of protein-protein interactions in humans, as described above (Römer et al., 2006; Joerger and Fersht, 2007; Wells et al., 2008; van Dieck et al., 2009).

Similar to many biological and other networks, the number of interactions per protein tends to approximate a power law distribution, resulting in a small number of highly connected “hub” proteins, and a larger number of proteins with relatively few interactions (Barabási and Oltvai, 2004; Giot et al., 2003; Li et al., 2006). This type of architecture offers resilience to random mutation since the chance of randomly affecting a highly connected protein is relatively low (Albert et al., 2000; Li et al., 2006). This is especially true if functional redundancy is also present, as has been described for p38 MAP kinases in *Drosophila* (Belozarov et al., 2012).

The methods used to generate large-scale data on PPIs include affinity purification coupled with mass spectrometry (Vermeulen et al., 2008), as employed by my collaborators to identify interactors of Oct4 in Chapter 4, and methods with a phenotypic readout such as yeast two-hybrid screens (Cusick et al., 2005), or the use of RNA interference against multiple targets to identify genetic interactions (Roguev et al., 2013). Although the question of interaction strength and functionality is difficult to address with these high-throughput methods, projecting additional layers of information about e.g. gene expression, subcellular localisation, protein structure, sequence features and evolution onto protein-protein interaction networks has been found to be extremely helpful in their biological interpretation (Aloy and Russell, 2006; Dittrich et al., 2008; Fraser et al., 2013; Levy and Pereira-Leal, 2008; Szklarczyk et al., 2011). An approach of this kind was applied to the human lysine acetylation system in Chapter 3.

1.3. Motivation and overview of this dissertation

As illustrated above, post-translational modifications are crucial components of eukaryotic signalling systems, mediating processes in development, homeostasis and disease. Our functional understanding of these modifications is still limited. By applying evolutionary approaches and a data integration framework, we aim to understand the evolutionary histories as well as the overall topologies of these signalling networks, and to highlight their most essential and potentially unappreciated components as candidates for future study.

The dissertation is structured into three primary chapters, preceded by the introduction and followed by a summary section for conclusions and implications.

Chapter 1: Introduction

This chapter presents a literature overview of the diversity of post-translational modifications, the experimental approaches used to identify them, their known biological functions and describes how they may combinatorially influence protein function. Since many modifications occur in fast-evolving disordered regions of proteins, intrinsic disorder was introduced as a structural feature of proteins. Histone modifications were discussed as an example of a highly complex cellular regulatory system mediated by PTMs.

Chapter 2: Evolution of post-translationally modified residues

This chapter was prompted by the open question of how many PTM sites are biologically functional. We investigated this question by quantifying selection pressure on post-translationally modified residues at the species, population and somatic levels, and by analysing conservation and substitution patterns at these sites. The chapter also addresses the increased occurrence of mutations which may either resemble the modified or unmodified state at PTM sites.

Chapter 3: The human lysine acetylation system

Due to the central importance of post-translational signalling systems in cellular homeostasis, we here present a generalised framework for the study of PTM signalling systems through data integration. This method was applied here to

present a concise overview of the human lysine acetylation signalling system, using multiple types of data. Proteins with functions in lysine acetylation were identified, and were then analysed based on their conservation, expression profile, protein-protein interactions and subcellular localisation. By integrating these types of information, the most promising candidate proteins participating in lysine acetylation signalling were systematically highlighted and ranked by priority for targeted functional study.

Chapter 4: The interaction network of Oct4

This chapter aims to understand the downstream regulatory effects of Oct4 activity, and it is the result of a collaboration with Dr. Jyoti Choudhary's group at the Wellcome Trust Sanger Institute. Oct4 is an essential transcription factor which is central to stem cell identity. My collaborators identified physical interactors of Oct4, and this group of proteins represented an important system for analysis using similar methods as in Chapter 3. This chapter presents an in-depth analysis of the Oct4 protein-protein interaction network based on conservation, transcriptional regulation, expression data and phenotypic data, including implication in human disease.

Chapter 5: Conclusions and implications

In this concluding chapter, I highlight key findings and describe their potential implications, including the role of disruptions of post-translational signalling in human disease, and implications for the engineering of signalling systems in synthetic biology.

2. Evolution of post-translationally modified residues

2.1. Introduction

The post-translational modification of proteins is a fundamental means of biological information processing that is important in development, homeostasis and disease. A classical example is phosphorylation, where the interplay of kinases and phosphatases generates complex signalling cascades. A number of dynamic, reversible types of modifications have been studied on a genomic scale in an effort to understand the regulatory code of eukaryotic chromatin. Notably, large-scale mass spectrometry studies have recently identified thousands of residues which undergo modifications such as lysine acetylation. Our functional understanding of these modifications is still limited. By combining several complementary evolutionary approaches, we here aim to investigate the evolutionary histories as well as the overall functional significance of these signalling networks.

This chapter also seeks to address a fundamental problem posed by the high sensitivity and stochastic nature of mass spectrometry (Olsen et al., 2010): do the thousands of post-translationally modified residues which have been identified thus far truly represent biologically significant regulatory events in the cell? A useful indirect means of addressing this question is to determine whether the modified residues are conserved in evolution (Kumar et al., 2009), which is done here at three levels: between species, within populations, and by studying the effects of somatic mutations at PTM sites in individuals. I will then describe intriguing evolutionary patterns that were found in the course of these analyses. Figure 2.1 shows an overview of the methods used in this chapter.

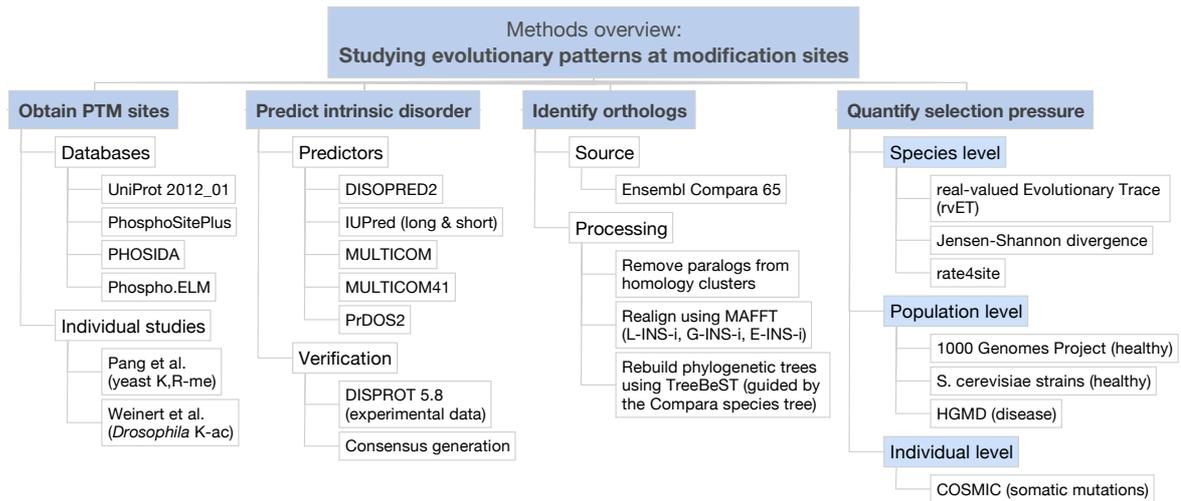


Figure 2.1: A schematic representation of the workflow undertaken here to quantify selection pressure at post-translationally modified residues.

2.1.1. Initial motivation

Protein phosphorylation is one of the most common types of post-translational modification (PTM) studied to date. Theoretical considerations suggest that modification sites involved in the direct regulation of protein activity might be more likely to occur within conserved regions of a protein family (Nühse et al., 2004). On the other hand, modification of unique sites could allow distinct regulation of certain paralogs, isoforms and proteins in a species-specific manner. In the case of paralogs, these sites may still be conserved among true orthologs in related species (Nühse et al., 2004). In a relatively small-scale study in *Arabidopsis thaliana*, observations indicated that serine and threonine phosphorylation sites do appear to be evolutionarily conserved in other plant species (Nühse et al., 2004).

Surprisingly, an initial report on a large-scale phosphorylation dataset stated that phosphorylation sites displayed lower average conservation than the rest of the protein (Gnad et al., 2007). In principle, it can be considered that the cell could easily afford extensive non-functional phosphorylation sites, since even the phosphorylation of a large proportion of proteins would only use up a small fraction of cellular ATP (Lienhard, 2008). It has been suggested that the low conservation of phosphorylation might be due to its effect being frequently exerted through “bulk charge”, i.e. the addition of multiple phosphate groups within a certain region of a protein (Serber and Ferrell, 2007). These would exert their signalling effects through relatively unspecific enhancement or disruption of protein interactions (Holt et al.,

2009). This has been termed the “polyelectrostatic model” of phosphorylation (Borg et al., 2007). In support of this, it has been reported that the overall number of phosphorylation sites for a protein is more conserved than their precise positions (Beltrao et al., 2009; Tan et al., 2009a; Brown et al., 2011). Sites may shift position in rapidly evolving disordered regions (Holt et al., 2009). The conservation of the phosphorylated state could then be due to shared subcellular localisation or conserved protein-protein interactions that frequently allow these phosphorylation substrates to come into contact with kinases (Beltrao et al., 2009).

A case has also been made for widespread non-functional, “noisy” phosphorylation (Lienhard, 2008; Landry et al., 2009; Moses and Landry, 2010; Levy et al., 2012), due to the low substrate specificity of at least some kinases (Joughin et al., 2012). In this case, only 10–40% of disordered phosphorylation sites may be more conserved than disordered control residues (Pearlman et al., 2011), while others have contradicted this for both structured and disordered phosphorylation sites (Chen et al., 2010). More recently, it has been reported in a *Saccharomyces* study that up to 90% of phosphosites may be conserved, though this study was limited to a few hundred high-confidence phosphorylation sites (Ba and Moses, 2010). Another study covering a set of vertebrate species estimated that around 70% of phosphorylation and N-glycosylation sites display better conservation than controls (Gray and Kumar, 2011). A second study of N-glycosylation also found significant conservation at solvent-accessible sites (Park and Zhang, 2011). An additional larger-scale study found evidence of conservation of serine, threonine and tyrosine phosphorylation sites located in loop regions, according to secondary structure prediction (Gnad et al., 2010a). Contrary to expectation, one study covering four vertebrates found that phosphorylation sites in proteins of vertebrate-specific functional modules, such as signalling processes, were more conserved than those in basic functional modules, such as metabolic and genetic processes (Wang et al., 2011b). A study of lysine acetylation found significant conservation from *Drosophila* to humans (Weinert et al., 2011), while studies of ubiquitination reported weak evolutionary conservation of sites (Danielsen et al., 2010; Hagai et al., 2012), and in the most recent study to date, it was estimated that only around 16–26% of PTM sites are more conserved than control residues, and that only these are likely to have significant biological functions (Minguez et al., 2012).

Clearly, the conservation and by proxy the biological functionality of PTM sites is still a matter of intense debate, especially due to the differences in scope in the above studies and due to variations in methodology. The study presented here attempts to resolve this by integrating comprehensive experimental information on the PTM types studied to date with evolutionary data on a large, representative cross-section of species, and by employing multiple up-to-date, independently evaluated methods at each step of the analysis.

2.1.2. Methodological considerations

There are several methodological aspects which we had to consider and correct for.

Intrinsic disorder

First, the study of residue conservation is confounded by the tendency of some modification types to occur predominantly in fast-evolving intrinsically disordered regions. We addressed these differences in structural preferences between specific post-translational modifications by performing separate analyses for structured and disordered residues.

Internal residues

Second, internal residues in the core of a protein evolve more slowly than surface residues, since they are likely to be under much larger structural and therefore evolutionary constraints (Chothia and Gough, 2009; Sasidharan and Chothia, 2007). To generate a control group more closely resembling to modifiable residues, which must be accessible to a modifying enzyme in at least one conformation of the protein, residues predicted to be in the interior of a protein were removed from the structured control. We based this on structural information from the Protein Data Bank (PDB) in those cases where sequence similarity existed to an entry, resulting in the removal of under ten percent of the structured control residues (Velankar et al., 2012).

We also subdivided certain analyses further for core and surface residues, resulting in a total of four distinct structural categories (with predicted disorder being very rare

in the interior). However, since PDB information was only available in around a fifth of cases, we did not make general use of the core/surface distinction in this study.

Functional divergence of paralogs

Additionally, it has been reported in recent studies on *Saccharomyces cerevisiae* that the presence of post-translational modification sites increases the likelihood of retention of both copies of duplicated genes, although the modification sites appear to diverge quickly (Amoutzias et al., 2010; Freschi et al., 2011). This supports a hypothesis of rapid neo- and subfunctionalisation of paralogs, and highlights the importance of post-translational modifications in these processes.

Ortholog identification

Another aspect was the confident identification of orthologs. A variety of advanced homology detection methods exist, and a number of databases have been created to disseminate their results (Altenhoff and Dessimoz, 2009). Ensembl Compara performed very well in an independent evaluation of several homology detection pipelines (Linard et al., 2011). Previous studies of PTM site conservation have only rarely made full use of advanced, duplication-aware pipelines such as Ensembl Compara (Landry et al., 2009; Gnad et al., 2010a; Gray and Kumar, 2011; Weinert et al., 2011). Of these, only the study by Gnad *et al.* used coding sequence data, which can be useful for judging evolutionary distance over shorter time scales. Only two studies used evolutionary rates to quantify conservation (Landry et al., 2009; Gray and Kumar, 2011), though the study by Gray *et al.* did not correct for potential differences in intrinsic disorder occurrence at PTM sites relative to non-modified control sites. Landry *et al.* used a relatively small set of nine vertebrate species, in addition to an analysis in fungi (Landry et al., 2009).

Coding sequences in phylogeny

In this study, the phylogeny-based methods (rvET and Rate4site), and especially the hybrid method rvET, appeared more sensitive in that they detected a larger number of significantly conserved PTM categories, and less cases of significantly faster evolution at PTM sites compared to controls. The difference in performance of the phylogeny-based methods might be explained by their indirect access to cDNA

sequence information through Ensembl Compara, which builds its phylogenetic trees using cDNA information, and also reconciles them with an established species tree (Flicek et al., 2012). This information may be especially valuable since the species studied are mostly mammals and vertebrates (Baldauf, 2003). Their evolutionary distances are relatively small, and classical taxonomic methods including the fossil record can contribute to the construction of their species tree.

Mutations which mimic PTMs

Mutations of phosphorylation sites to negatively charged amino acids can sometimes successfully mimic the effects of phosphorylation, and effectively make the phosphorylated state permanent (Thorsness and Koshland, 1987; Tarrant and Cole, 2009). *Vice versa*, phosphorylation sites may originate in place of negatively charged amino acids (Kurmangaliyev et al., 2011; Pearlman et al., 2011). These cases illustrate that simply distinguishing between the presence and absence of the modifiable amino acid (e.g. serine at a phosphorylation site), as adopted in many previous studies, is likely to be an imperfect method for judging the functional significance of the phosphorylation site in question, since the mutations do not diminish it. This illustrates that sequence entropy- and phylogeny-based conservation scoring methods should be considerably more sensitive in detecting selection pressure at PTM sites.

Heterotachy

In a phenomenon sometimes referred to as heterotachy, evolutionary rates and selection pressure do not necessarily remain constant over time within a lineage (Lopez et al., 2002; Philippe et al., 2003; Kosiol et al., 2008; Roure and Philippe, 2011). This has led to some skepticism about the accuracy of the “molecular clock”, i.e. the estimation of divergence times from sequence data (Ayala, 1999), although other studies have reported this method to be accurate (Subramanian and Lambert, 2011). Another complicating factor may be that mutations in one residue might lead to relaxation of evolutionary pressure at other positions through change in structural constraints (Bridgham et al., 2009).

Heterotachy should not affect the observations made in this study, since PTM sites and control residues were obtained from the same set of proteins. Heterotachy may

imply that the hybrid rvET method could be more robust than the phylogeny-based Rate4site method, which also considers branch lengths. Population bottlenecks and adaptive radiation events may violate this assumption.

Two related factors may result in the loss of branch length additivity, and therefore complicate the reconstruction of sequence evolution: i) extremely high evolutionary rates, and ii) horizontal transfer events (Tria et al., 2010). Neither are likely to be significant in the species used here, which are relatively closely staggered in evolutionary time.

Sequence windows

The authors of the Jensen-Shannon Divergence conservation scoring method (JSD) reported that considering a sequence window, rather than a single residue, may give improved results on various benchmark datasets, including the Catalytic Site Atlas (Porter et al., 2004), ligand binding sites and protein-protein interaction interfaces (Capra and Singh, 2007). The authors of the real-valued Evolutionary Trace (rvET) method also recently reported that taking into account the scores of neighbouring residues, either adjacent structurally or in the primary sequence (± 1 aa), results in an improvement in functional site discovery (Wilkins et al., 2010). Additionally, most serine/threonine kinases appear to target motifs of ± 4 residues in size (Ubersax and Ferrell, 2007). A recent study of interactions by linear motifs found that three-dimensional context contributes $\sim 20\%$ to the binding energy, and is important for binding specificity (Stein and Aloy, 2008). In addition, it has been noted that recognition modules for short linear motifs may display multiple binding specificities, which further increases the number of their potential binding partners (Gfeller et al., 2011). This further suggests that neighbouring protein features may play a part in mediating interaction specificity. However, primary sequence adjacency as implemented by a sequence window can only capture part of these contributions, and three-dimensional structure data is unavailable especially for the disordered PTM sites used here. In this study, we found that the use of a sequence window did not strengthen the PTM site conservation signals.

Amino acid similarity

With regard to amino acid similarity matrices, the authors of the JSD implementation found that incorporating similarities is not always helpful when assessing residue conservation (Capra and Singh, 2007). They caution that the datasets of known functional sites used in evaluations may generally be biased towards absolutely conserved residues, where similarity relationships between amino acids may not be essential for good performance. With the potential exception of multi-specificity serine/threonine kinases, PTM sites would similarly require a single specific residue to be conserved, rendering amino acid similarities unhelpful. This illustrates that conservation scoring methods should be carefully chosen, and that the most elaborate models may not always be the most useful for a given application.

2.2. Methods

Given the diversity of the types of post-translational modification and their potential disorder/structure preferences, the primary challenge in this study was to quantify PTM site conservation (Fig. 2.1). To achieve this, we needed to source experimentally-determined PTM sites, predict their structural properties, obtain clusters of homologous proteins, remove paralogs due to their rapid divergence, align the orthologous sequences, reconstruct their phylogeny, and then quantify conservation using several up-to-date evolutionary scoring methods. The details of these steps are described below.

2.2.1. Obtaining PTM sites

Amino acid sequences as well as experimentally determined post-translational modification sites were extracted from the curated Swiss-Prot subsection of the UniProt database (release 2012_01) (UniProt Consortium, 2012). We also included PTM sites from PhosphoSitePlus version 110311 (Hornbeck et al., 2012), PHOSIDA (obtained 2012-02-03) (Gnad et al., 2010b) and Phospho.ELM version 2011-11 (Dinkel et al., 2011). These databases also include many PTMs other than phosphorylation. Additional experimental lysine acetylation and N-glycosylation sites were obtained from individual large-scale studies (Gnad et al., 2007; Pang et al., 2010; Wang et al., 2010; Zhao et al., 2010). The final compendium, incorporating all PTM site resources that previous major studies have used, consisted of 172,931

sites, of which 89,236 were on human proteins. This dataset comprises the following major PTMs: lysine acetylation (abbreviated “K-ac”), lysine ubiquitination (“K-ub”), N-glycosylation (“N-gly”), and serine (“S-p”), threonine (“T-p”) and tyrosine phosphorylation (“Y-p”). Their structures are shown in Figure 1.1.

2.2.2. Predicting structural properties

Distinction between intrinsically disordered and structured residues

In order to address the different evolutionary rates of intrinsically disordered and structured regions, disordered protein segments were predicted using six methods: DISOPRED2 (Ward et al., 2004), IUPred in “long” and “short” operating modes (Dosztányi et al., 2005), MULTICOM-REFINE, MULTICOM-REFINE with sspro4.1 (Deng et al., 2009), and PrDOS2 (Ishida and Kinoshita, 2007). Although some methods provide various scores in addition to a binary disorder/structure call, only the binary calls per amino acid were retained. A consensus of all six methods was used for verification, and the experimental DISPROT database (release 5.8) (Sickmeier et al., 2007) served as a gold standard control.

PrDOS2 and MULTICOM-REFINE placed first and third respectively in the CASP9 disorder prediction assessment, according to the area under their receiver operating characteristic (ROC) curves (Monastyrskyy et al., 2011). This measure strikes a compromise between the sensitivity and specificity of disorder prediction, i.e. the rates of false negatives and false positives. This metric is the most useful for the present study since both under- and overestimation of intrinsic disorder would be equally detrimental. The second best method, DISOPRED3C, was not yet available for use. The CASP9 assessment also used a highly varied set of prediction targets from diverse species, which should ensure the robust prediction of intrinsic disorder.

It is notable that PrDOS2 and MULTICOM-REFINE disagree quite substantially in their predictions (on 14% of the residues in the UniProtKB/Swiss-Prot proteome studied here). While MULTICOM-REFINE is a purely *ab initio* prediction method, PrDOS2 is a hybrid method which includes *ab initio* prediction as well as a template-based PSI-BLAST approach (Deng et al., 2012). All methods were carried forward in the analysis and produced highly similar results, although MULTICOM-REFINE was later chosen as the primary predictor to be presented in the interest of space. The

proteome-wide agreement among all six disorder predictors in terming residues structured or disordered was 63%, and both MULTICOM-REFINE and PrDOS2 showed agreement with DISPROT on 94% of residues. Overall, 41.7% of all residues were predicted to be intrinsically disordered by MULTICOM-REFINE, the primary predictor chosen in this chapter.

Distinction between core and surface residues

Absolute all-atom solvent-accessible surface areas (ASAs) for all 86,008 proteins contained in the PDBe database (the Protein Data Bank in Europe) as of 2012-11-07 (Velankar et al., 2012) were calculated using the program “PISA”, as contained in version 6.3.0 of the CCP4 structural analysis software package (Krissinel and Henrick, 2005; 2007; Winn et al., 2011). Default parameters were used, but the multimer assembly analysis step was deactivated by recompiling due to failure on a large fraction of PDB structures. Monomeric ASA values were unaffected by this modification. Without assembly analysis, ASA values could be obtained for 76,680 PDB structures (89.2%), the missing structures being predominantly of DNA or RNA or NMR-based. Relative solvent-accessible surface areas (rASAs) were then calculated using a reference table of all-atom ASAs for the central amino acid in Gly-X-Gly tripeptides (Miller et al., 1987). Residues with a relative accessible surface area of >25% were considered to lie on the protein surface. This threshold has been reported to provide the best separation in terms of sequence composition between buried (core) and surface residues (Levy, 2010).

The per-residue core/surface calls collected from PDB were then used to determine likely ASA values for the human protein sequences from UniProt, which form the basis of this chapter. This was done via BLASTP searches with an E-value cutoff of 1, chosen to be slightly more stringent than the default value of 10, while still remaining inclusive of more distant matches. This adjustment excluded only a very small fraction of very low-scoring BLASTP matches. Each UniProt residue’s core/surface status was then obtained from the best-scoring PDB sequence match reported by BLASTP, provided that the aligned residues were identical. In an alternative approach, the average ASA value from identical residues across all BLASTP matches below the E-value cutoff which still contributed more UniProt sequence coverage was used. Highly similar results were observed (data not shown). In this alternative approach, the median absolute disagreement for residues

where multiple overlapping and coverage-contributing BLAST matches were available was 5.5 Å, which was considered an indication that the structures generally agree well in terms of accessible surface area per residue.

Overall, despite the relatively inclusive BLASTP-based approach described above, using structural information from PDB to distinguish between core and surface residues drastically limited the number of investigable PTM sites to around 20% of sites (e.g. from 56,537 to 11,211 residues for the disorder predictor MULTICOM-REFINE and the variation scoring method Rate4site). It also reduced the number of disordered PTM sites to be studied, with MULTICOM-REFINE predicting 41.7% of all residues to be intrinsically disordered, whereas the proportion of disordered residues with identical BLASTP matches in PDB was only 16.3%. As expected, core residues were largely predicted to be structured, with 90.3% of core residues falling into structured regions, indicating good agreement between the disorder predictors and the PDB-based core/surface distinction. Where PDB information was available, approximately half (53.9%) of all structured residues were predicted to fall into the protein core, with the rest lying on the surface, highlighting the general usefulness of the core/surface distinction in identifying residues under different evolutionary constraints.

However, in order to maintain a statistically useful number of sites and to better address disordered residues, we made use of the PDB-based core/surface distinction only in some analyses (e.g. Table 2.2). The primary purpose of this distinction was to remove structured core residues from the control samples, since these residues were expected to be extremely highly conserved and would therefore not represent a useful, structurally equivalent type of control for PTM sites. This affected around 8% of the structured control residues.

2.2.3. Obtaining homology clusters

Protein- and cDNA-level multiple sequence alignments and a species tree covering 56 primarily mammalian and vertebrate species were obtained from Ensembl Compara (Release 65, December 2011) (Flicek et al., 2012). Using homologs from Ensembl offers an advantage over simpler methods that have been widely used in previous studies of PTM site evolution. A bidirectional best hit BLAST search will only return a limited set of ortholog pairs (that are each other's best hits). It might

therefore be too stringent and produce false negatives. A simple one-directional BLAST, on the other hand, makes it difficult to identify a true ortholog among the hits, especially if many closely related paralogs score similarly or if a true ortholog is not present, and ultimately an arbitrary cutoff would have to be chosen.

Compara addresses these problems through a more sophisticated pipeline for building homology clusters, involving a species-based guide tree based on well-established taxonomies, especially for the vertebrates. The pipeline allows deviations from the species tree if there is very strong evidence of lateral transfer, but this is unexpected in the species considered here. Phylogenetic reconstruction becomes an important component of the analysis in evolutionary rate calculation, where either the dendrogram structure alone (rvET, see below) or both structure and branch lengths (Rate4site) are used. A simple alignment of BLAST hits, as used in some studies, essentially discards all knowledge on species relationships. It assumes that the evolutionary distances between sequences are directly measurable from their sequence similarity, which will not be the case if selection pressure has changed in different lineages, or over long time scales. The accuracy of the phylogenetic tree constructed will also be lower in the absence of taxonomic information. Therefore, the use of Ensembl Compara as an ortholog source constitutes a methodological improvement.

2.2.4. Removal of paralogs

In the analysis of protein interfaces, inclusion of diverse homologs including paralogs has been reported to give slightly improved results (Caffrey et al., 2004). For PTM sites, however, the opposite has been observed. Duplicated genes rapidly lose phosphorylation sites as their sequences diverge, suggesting that substantial neo- and subfunctionalisation of regulation via phosphorylation is occurring (Amoutzias et al., 2010; Freschi et al., 2011). In order to remove all potentially functionally divergent proteins, we decided to exclude paralogs when calculating conservation scores.

One-to-one orthologs of human proteins were obtained from Ensembl Compara (Vilella et al., 2009; Flicek et al., 2012), and the homology clusters were filtered to contain only these. This approach discards inparalogs, but retains outparalogs and divides them into separate orthology clusters (Sonnhammer and Koonin, 2002).

Orthology clusters consisting of less than ten sequences were also discarded. This means that at a minimum, each processed orthology cluster spanned all primate species included in Compara, which should ensure sufficient evolutionary distance to assess evolutionary trends.

2.2.5. Alignment of sequences

Alignment among one-to-one orthologs, without functionally divergent paralogs, should result in better alignment quality, and remove gaps originally introduced by distant homologs within a sequence family. The orthologous clusters were therefore aligned using MAFFT, a fast multiple sequence alignment method (Kato and Toh, 2008), which performs very well in independent evaluations (Ahola et al., 2006; Nuin et al., 2006; Golubchik et al., 2007; Liu et al., 2010a; 2010c). MAFFT is also used in major public homology pipelines such as Ensembl Compara (Flicek et al., 2012). Six variants of this method were used. MAFFT can run in either local (L-INS-i), global (G-INS-i), or BLAST-like (E-INS-i) alignment modes. In addition, MAFFT can either construct a guide tree based on sequence similarity, or use the Compara species tree as a starting point.

L-INS-i is the default alignment mode of MAFFT, and is optimised for aligning one central conserved region while allowing terminal variation. This is appropriate given observations of preferential terminal domain acquisition over long evolutionary timescales (Moore et al., 2008). The Ensembl Compara pipeline first attempts the global-like alignment mode G-INS-i, presumably due to the relatively short distances between many species in Compara, and then reverts to the local L-INS-i mode if a consensus with other alignment programs cannot be reached (Flicek et al., 2012). Guide trees for MAFFT were generated from the Compara species tree, and pruned of any species that were no longer present in the paralog-free orthology clusters using Bio::Phylo (Vos et al., 2011). The alignment results from all six variations of MAFFT were carried forward in the analysis. Since the results produced were quite similar, the species tree-based L-INS-i mode was chosen for display in this chapter for simplicity.

2.2.6. Phylogenetic reconstruction

Though horizontal gene transfer is not unheard of in higher eukaryotes, it is exceedingly rare (Dunning Hotopp, 2011), and we therefore assumed that the *Compara* species tree, as established largely through classical taxonomy, is a reliable blueprint for the gene trees. The known instances of lateral transfer to eukaryotes are thus far limited to insects and nematodes, with endosymbiotic bacteria or fungi as the clade of origin (Danchin et al., 2010; Moran and Jarvik, 2010; Acuna et al., 2012).

The phylogenetic tree and branch lengths of the alignments generated by MAFFT were calculated using TreeBeST (Guindon and Gascuel, 2003; Ruan et al., 2008). TreeBeST reconciles gene-specific phylogenetic trees with a genome-wide species tree, and it is a widely used method for phylogenetic inference in public homology detection projects (Ruan et al., 2008; Hulsen et al., 2009; Vilella et al., 2009; Flicek et al., 2012). PAL2NAL was used for generating codon alignments from amino acid alignments, which is useful for evaluating evolutionary distance over shorter time scales (Suyama et al., 2006). All of these steps mirror *Compara*'s approach. Alignments with branch lengths above 100 were discarded, since these only occur when artificially introduced by TreeBeST in order to point to violations of the *Compara* species tree, and they lead to technical problems in computing evolutionary rates. This affects less than ten alignments. The NCBI taxonomy database was used to translate between taxon identifiers and species names (Sayers et al., 2012).

TreeBeST also incorporates both amino acid and back-translated coding sequence (CDS) alignments, in order to cover both long and short evolutionary distances, respectively. The CDS data is especially important in *Compara* since the evolutionary distances between many of the species are relatively small. At the codon level, synonymous mutations can accumulate without significant selection pressure, presenting a more finely resolved picture of evolutionary distance (Baldauf, 2003).

2.2.7. Assessing site conservation using variation scores

Methods for quantifying conservation can be broadly grouped into three categories: i) entropy-based methods, ii) phylogeny-based methods, and iii) hybrid methods (Johansson and Toh, 2010). Johansson *et al.* performed a comparative analysis of 25 conservation scoring methods, graded on their performance at identifying enzyme active site residues from the Catalytic Site Atlas (Porter *et al.*, 2004). Although the various scores display different characteristics, they largely cluster together according to their fundamental methodologies. With this in mind, one method from each category was chosen in this study, according to its performance in the evaluation by Johansson *et al.* and the availability of a software implementation. The three methods used are i) the Jensen-Shannon divergence (JSD), ii) Rate4site, and iii) real-valued Evolutionary Trace (rvET).

Symbol frequency-based approach: Jensen-Shannon divergence (JSD)

The Jensen-Shannon divergence (JSD) is a measure purely based on symbol frequency (Capra and Singh, 2007). It is the best-scoring method in the Johansson *et al.* evaluation (Johansson and Toh, 2010). JSD uses the BLOSUM62 matrix to derive background amino acid frequencies for sites subject to no evolutionary pressure. Then, positions in an alignment that are found to have amino acid distributions very different from this background distribution are proposed to be functionally important or under evolutionary constraint. It also implements a sequence weighting method that rewards sequences that are “surprising” (Capra and Singh, 2007). The authors find that JSD outperforms Rate4site, though only on some of the datasets tested (Capra and Singh, 2007), and another report has also suggested that an entropy-based approach can perform similarly well as the time-intensive Rate4site method (Mihalek *et al.*, 2007). Minor technical problems were encountered for a negligible set of nine proteins whose scores could not be obtained, and which were therefore removed from all analyses. By default, JSD uses a window of ± 3 amino acids. This option was not used in the present study as it did not result in an increase in the observed degree of conservation when studying PTM sites. Alignment columns that consist of $>30\%$ gaps were discarded by JSD by default. The JSD score ranges between 0 and 1. For comparability with the other

two methods, this score was inverted by subtracting its value from 1, so that lower values now signify stronger conservation in accord with the other two methods.

Phylogeny-based approach: Rate4site

Rate4site is a phylogeny-based Empirical Bayesian method, which calculates replacement probabilities for each residue, using the JTT model of evolution (Mayrose et al., 2004). It is the only method incorporating branch lengths in its calculations. In alignments of less than 50 sequences, Rate4site appears to perform slightly worse than other methods on a test set for catalytic site prediction (Johansson and Toh, 2010). Minor technical problems were encountered for a negligible set of three proteins whose scores could not be obtained, and which were therefore removed from all analyses. Phylogenetic trees were passed to Rate4site, and the original tree branch lengths were retained. This ensures the information from the CDS alignment as used by TreeBeST is kept in the branch lengths.

Hybrid approach: Real-valued Evolutionary Trace (rvET)

Real-valued Evolutionary Trace (rvET) is a mixed score incorporating both phylogeny-based and sequence entropy information (Mihalek et al., 2004). It is defined as the sum of the information entropy within an alignment column and a phylogeny-based score, i.e. the number of tree nodes that need to be traversed from the root of the tree to reach a branch in which the alignment column is uniform. The rvET score ranges from a minimum of 1 to higher values, with 1 signifying complete conservation of a residue within an alignment. This method uses the simple Shannon entropy (Shannon, 1951) in combination with phylogenetic information. The more complex von Neumann entropy definition, which additionally takes amino acid similarities into account, was not found to be an improvement (Johansson and Toh, 2010).

Normalisation method

In addition to the raw conservation scores, we repeated our analyses using scores normalised per protein. This normalisation was performed by adjusting all scores in a protein to a mean of zero and a standard deviation of one, i.e. a residue evolving at the average rate of the protein received a normalised score of zero (Fig. 2.4).

2.2.8. Assessing evolutionary patterns at PTM sites

Comparing PTM sites to control sites

This analysis will address the central question whether post-translationally modified residues fall into more evolutionarily constrained regions than control residues. The three scoring methods described above were used to quantify conservation at PTM sites as compared to non-modified amino acids of the same type, within proteins bearing at least one PTM site of the same type, as control sites. Johansson *et al.* note that evolutionary rates should not be compared between alignments, since most scoring methods are dependent on the number of sequences in the alignment. These dependencies cannot be resolved through simple measures, such as dividing by alignment size (Johansson and Toh, 2010). Though all three conservation scoring methods selected here display this dependency, I circumvented this problem by selecting all control sites from the same set of proteins as the modified sites. This type of control also ensures that the proteins can, in principle, be modified, since at least one modification site of the type of interest exists on them. The correlation between the number of PTM sites and control sites per protein was positive ($p < 2.2 \cdot 10^{-16}$, $r = 0.66$, Pearson's product-moment correlation), and stratification by protein (Table 2.2) as well as normalisation of scores (Fig. 2.4) were used to address potential differences in the distribution of PTM sites and control sites across proteins. When choosing the control sites, non-experimental PTM sites annotated as “probable”, “potential” or “by similarity” in UniProt were also avoided, in order to ensure as confidently as possible that control sites were not modified. PTM sites without a suitable number of control residues within the same protein, i.e. less control residues than PTM sites, were not considered in the analysis. This affected less than twenty proteins.

In one previous study of serine phosphorylation conservation, the nearest unmodified serine was used as a control (Kurmangaliyev *et al.*, 2011). However, this might result in a residue with different structural characteristics being chosen, and it does not make full use of the available control sites. We decided to take the following approach: modified residues of a certain structural category were compared to all non-modified control residues of the same structural category within the same set of proteins. Structured residues predicted to fall into the protein core

(approximately 8%) were removed from the analysis. Since the control residues were of the same disorder/structure category, they should be under comparable structural constraints in evolution as the modified residues, providing an adequate background to detect evolutionary rate differences against.

The resulting distributions of variation scores are shown as density estimates, using a Gaussian kernel function, and the R package ggplot2 (Wickham, 2010). Two-tailed Mann-Whitney U tests as well as permutation tests were used to determine whether the estimated distribution function of the PTM site variation scores differs significantly from that of the control sites. For stratified tests, this was done using implementations from the conditional inference R package “coin” (Hothorn et al., 2008). In order to determine the directionality of the differences (i.e. increased or decreased conservation at PTM sites), the medians of the samples were compared, as appropriate for the Mann-Whitney U test.

To judge the extent of conservation of PTM sites with reference to known functional residues, a list of residues participating in confirmed protein-protein interaction interfaces was obtained from a recent study by Levy *et al.* (Levy et al., 2012), and known catalytic residues were obtained from the Catalytic Site Atlas (Porter et al., 2004). To ensure comparability of these variation scores, only residues from the set of proteins which also contained PTM sites were included.

The nature of PTM site substitutions in evolutionary history

It has been noted that phosphorylated residues may act as a reversible mimic of functionally important acidic residues (Nühse et al., 2004). In order to determine whether such mimicry may exist for PTM types other than phosphorylation, the frequencies of substitution by various amino acids at PTM sites were compared to those at control sites using Two Sample Logo (Vacic et al., 2006). The default option for Student’s t-test was used instead of a binomial test due to performance reasons. A significance threshold of $\alpha = 0.001$ was used to correct for multiple testing.

We considered better-conserved residues to be more likely to retain similar functions, and therefore hypothesised that amino acid substitutions at better-conserved sites should be more functionally meaningful than at fast-evolving sites. Therefore, the PTM sites were split into tertiles for each PTM type and disorder category, and the best-conserved third was investigated in detail. Using the top third

strikes a balance between good conservation and a sufficiently large sample size for statistical analysis. The boundaries of the PTM site tertiles were then used to select control sites within the same score range, i.e. residues under similar evolutionary pressure. The resulting control site samples displayed highly similar averages and standard deviations to the PTM site samples.

Tracing PTM sites across multiple species

Since modifications such as lysine acetylation exist even in bacteria (Wang et al., 2010), and since the evolutionary origins of PTM sites have not yet been explored in detail, I generated plots of the conservation profiles of all human PTM sites in my dataset. This analysis is represented as a heatmap figure, produced using NeatMap (Rajaram and Oono, 2010), which shows the breakdown of the conservation patterns for these PTM sites (i.e. primate-specific, mammalian, jawed vertebrates, or even in fungi). More sporadic conservation patterns may be evidence for imprecise polyelectrostatic conservation of certain PTMs, or for only recent evolution of the site. Unmodified control residues are shown for comparison, excluding structured core residues.

2.2.9. Investigating the population and individual levels

Variation at post-translationally modified sites was assessed using data on non-synonymous single-nucleotide polymorphisms (nsSNPs) in populations, and on somatic mutations in individuals. Only missense variants, resulting in a single amino acid change, were used. Neutral nsSNPs from whole-genome shotgun sequencing (“LOWCOV”) were obtained from the coding annotation (released July 2, 2012) of the integrated phase 1 release (version 3) of the 1000 Genomes Project (Consortium et al., 2010). SNPs exclusively identified by exome sequencing (“EXOME”) were discarded from the main analysis due to their very low allele frequencies. While the whole-genome sequencing SNPs occurred at an average allele frequency (corrected for linkage disequilibrium) of ~4%, the exome-sequencing SNPs occurred at a 20-fold lower average allele frequency of ~0.2%. This suggests that these are rare alleles which have not spread in the population, and in some cases they might possibly derive from sequencing errors.

Confirmed disease-causing mutations (“DM”) were obtained from the 2012.2 release of the Human Gene Mutation Database (HGMD) (Stenson et al., 2009). Human somatic mutations in individuals that are causative of cancer were obtained from release 57 of the Catalogue Of Somatic Mutations In Cancer (COSMIC) (Dinkel et al., 2011). To verify if the observations in humans hold true in other organisms, an additional analysis was performed using nsSNP data on 38 different strains of *S. cerevisiae* (Liti et al., 2009). In this dataset, any SNPs with a quality score below 40 were discarded, unless they were present in at least two strains.

For each of these variation datasets, the dependence between the presence of variants and of PTMs was assessed separately for each PTM and disorder type, using Fisher’s exact test. Structured core residues were excluded from the control group.

2.3. Results

Recent large-scale studies have identified a large number of post-translationally modified (PTM) sites in a variety of organisms, with humans, mice and yeast being by far the most intensively studied species. Since the mass spectrometry methods used for their identification are highly sensitive, questions have been raised as to what fraction of the identified sites are modified with biologically meaningful frequency in vivo. We approached this problem by studying conservation scores at PTM sites as compared to unmodified residues, while controlling for structural characteristics. An overview of the human PTM site dataset we obtained is shown in Table 2.1. Orthologs could be found for approximately two thirds of proteins in each category using Ensembl Compara. This was done by searching Ensembl Compara for Ensembl protein identifiers, which were mapped from the UniProtKB/Swiss-Prot identifiers used for PTM site annotation directly via mapping information from UniProtKB (release 2012_01) (UniProt Consortium, 2012).

Table 2.1: Overview of human PTM types. The six PTMs highlighted in bold were chosen for study in this chapter due to the high availability of known sites. The number of modified proteins and of PTM sites for which orthologous proteins could be found in Ensembl Compara are given under “proteins with orthologs” and “sites with orthologs”, respectively. These columns show the final numbers of proteins and sites used in the conservation score-based analyses and conservation profiles. “n.d.” indicates that the number of orthologs was not determined since the modifications were not investigated in this study.

Modification	Amino acid	Proteins	Proteins with orthologs	Sites	Sites with orthologs
Phosphorylation	Serine	7,688	5,410	39,531	25,159
Phosphorylation	Tyrosine	5,470	3,538	12,047	6,787
Phosphorylation	Threonine	5,171	3,631	12,749	8,284
Ubiquitination	Lysine	4,821	3,374	17,434	10,870
Acetylation	Lysine	2,903	2,033	6,505	4,183
other modifications	various	2,824	<i>n.d.</i>	6,040	<i>n.d.</i>
N-linked glycosylation	Asparagine	873	600	2,047	1,406
Methylation	Arginine	86	<i>n.d.</i>	239	<i>n.d.</i>
Methylation	Lysine	82	<i>n.d.</i>	189	<i>n.d.</i>

The resulting alignments of orthologous proteins were then used to calculate the conservation scores. The median number of orthologs (i.e. species) per alignment was 42 (Fig. 2.2), and a minimum threshold of ten species was used to minimise artefacts arising from small alignments. At minimum, an alignment of ten species should span all primates included in Ensembl Compara, ensuring sufficient phylogenetic distance to assess evolutionary trends.

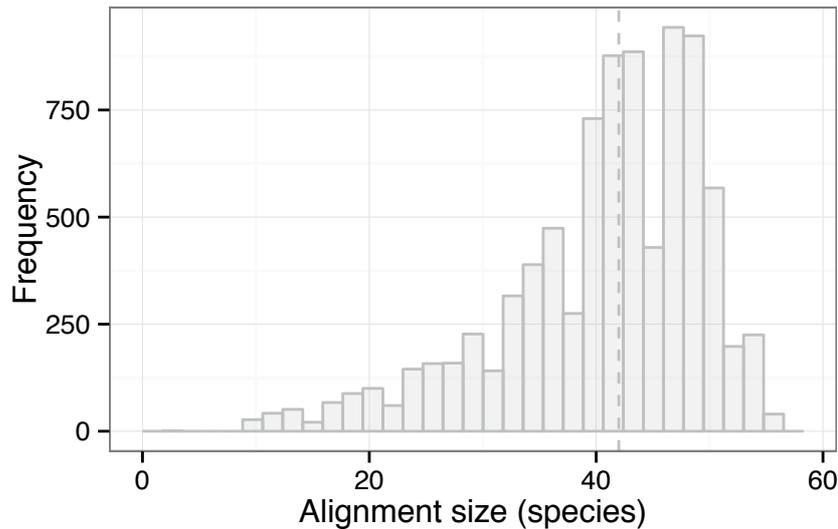
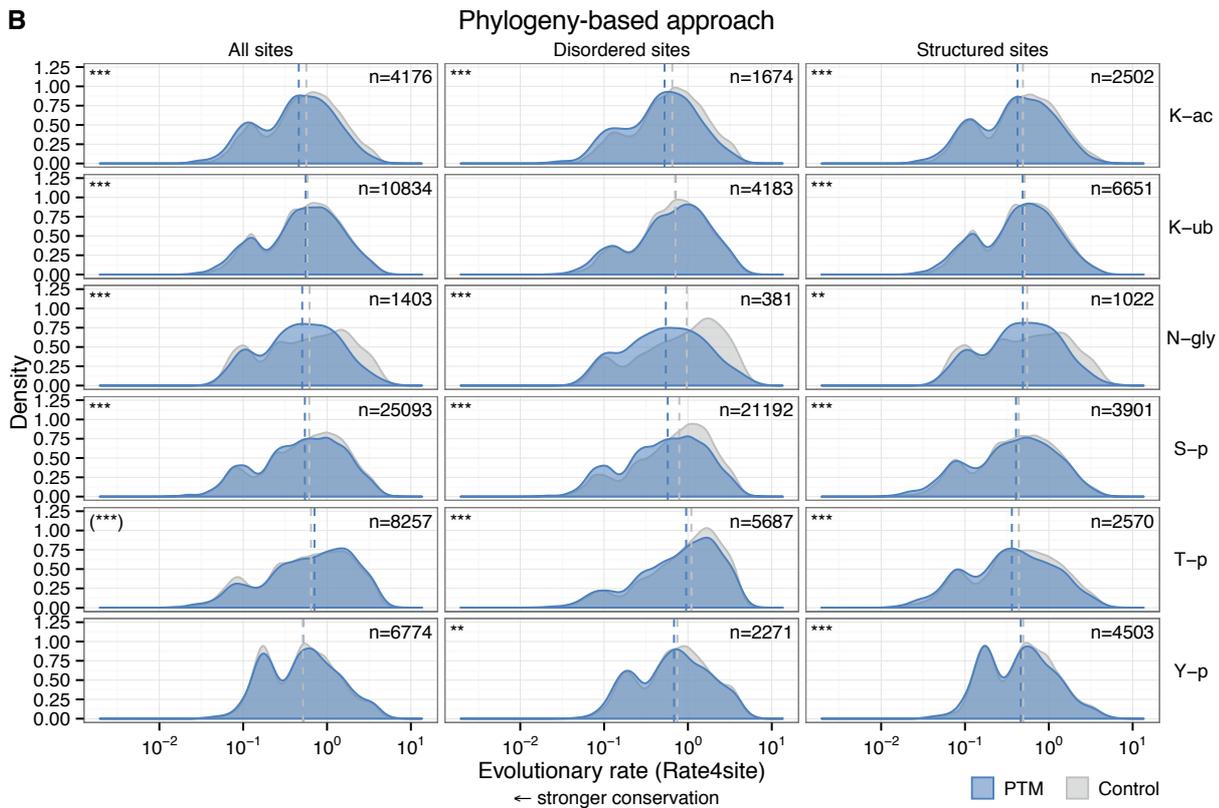
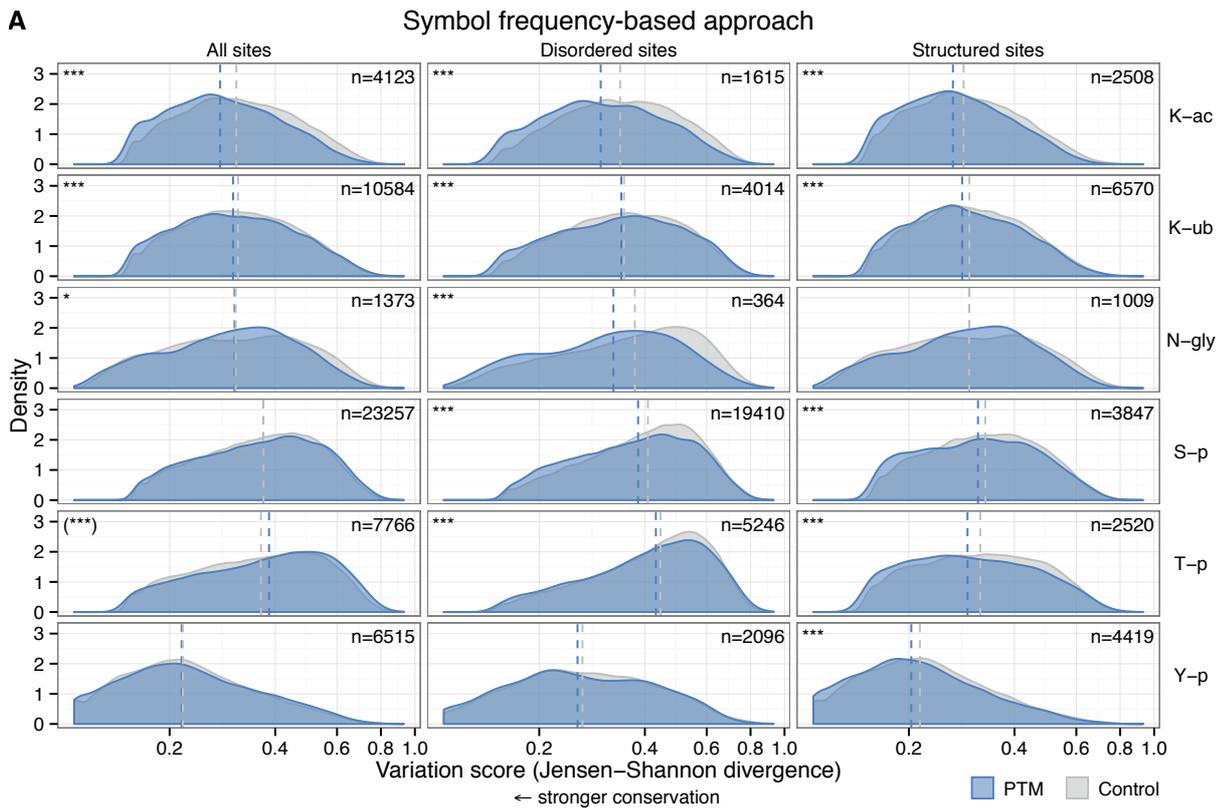


Figure 2.2: Size distribution of the alignments of orthologous proteins used to calculate conservation scores in this chapter. A minimum threshold of 10 species was used. The median number of orthologs (i.e. species) was 42 (see dashed line).

2.3.1. PTM sites are more conserved than structurally similar control residues between species

When analysing the conservation of PTM sites, we found that modification sites were generally significantly better conserved than a structurally equivalent background, as determined via Mann-Whitney U tests. Figure 2.3 shows the conservation of PTM sites compared to control residues according to three methods: i) the Jensen-Shannon divergence (JSD), a purely symbol frequency-based approach (Fig. 2.3A) (Capra and Singh, 2007), ii) Rate4site, a phylogeny-based method (Fig. 2.3B) (Mayrose et al., 2004), and iii) the real-valued Evolutionary Trace score (rvET), a hybrid approach with an entropy-based component as well as a phylogeny-based component (Fig. 2.3C) (Mihalek et al., 2004). The control samples were around eleven times larger on average than the samples of PTM sites.



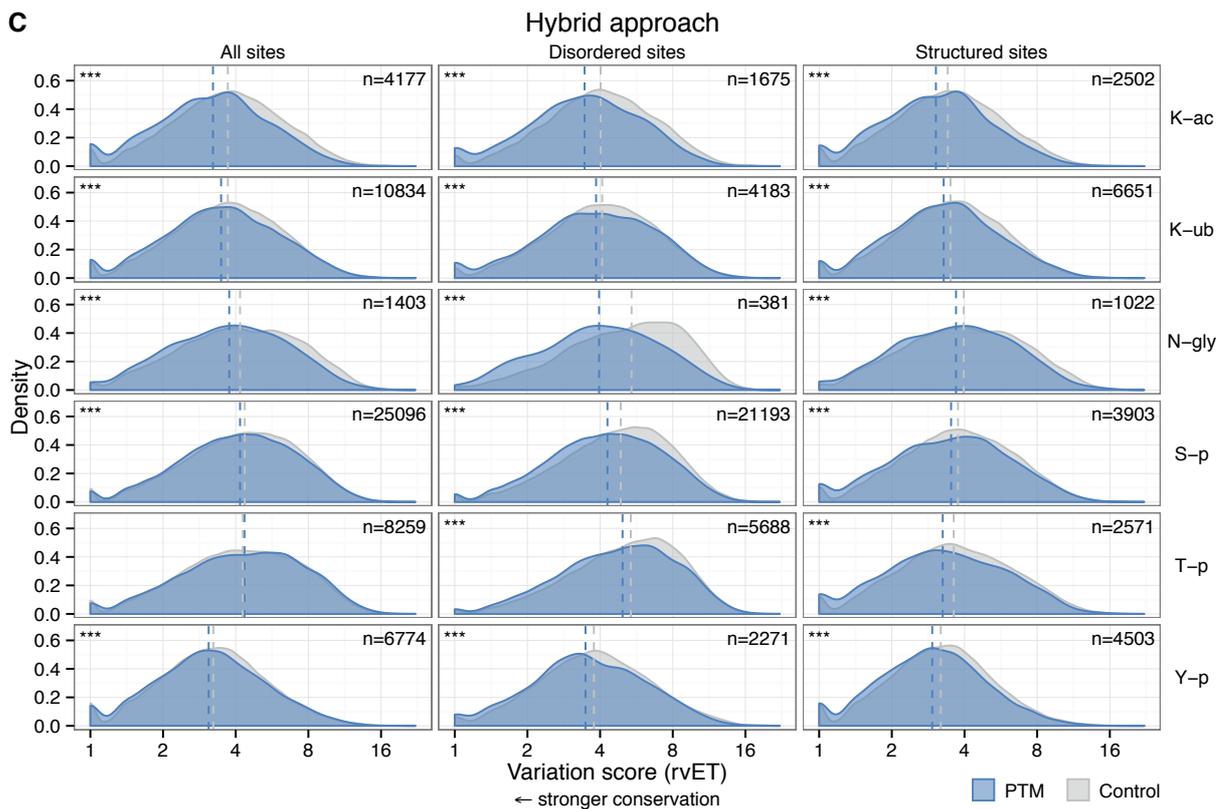


Figure 2.3: Conservation density plots showing conservation of PTM sites compared to control sites, assessed using (A) Jensen-Shannon divergence (JSD), a symbol frequency-based method, (B) Rate4site, a phylogeny-based method, and (C) the real-valued Evolutionary Trace score (rvET), a hybrid approach with an entropy-based component as well as a phylogeny-based component. Low variation scores indicate strong sequence conservation. Horizontally, the leftmost column contains all residues, while the other two show a breakdown into disordered and structured residues. Vertically, the different PTM types are shown. Asterisks indicate the level of statistical significance (*: $p < 0.001$, **: $p < 0.01$, *: $p < 0.05$), as determined by a two-sided Mann-Whitney U test. The dashed lines indicate the median of each sample, in blue for PTM sites and in grey for control residues. Where median PTM site variation is higher than that of controls, these asterisks are shown in round brackets. The size of the PTM site sample in each category is given in the top right corner. The size of the corresponding control sample was around eleven times larger on average.**

The leftmost column in Figure 2.3A–C contains all residues, whereas the other two constitute a breakdown into disordered and structured residues, with internal residues having been removed from the structured control. Notably, when all sites were considered together without distinguishing between disordered and structured residues, the symbol frequency-based JSD and the phylogeny-based Rate4site scoring methods reported that sites of threonine phosphorylation (T-p) were not only non-conserved, but were significantly less conserved than background (Fig. 2.3A). However, once the distinction was made, both disordered and structured

phosphorylated threonines (T-p, Fig. 2.3A–B) were found to be significantly more conserved than controls. Similarly, when considering all sites together, tyrosine phosphorylation sites appeared to evolve at the same rate than unmodified residues for these two methods (Y-p, Fig. 2.3A–B). Once we distinguished between intrinsically disordered and structured residues, structured tyrosine phosphorylation sites were consistently found conserved, though disordered sites were not conserved according to JSD.

Overall, the hybrid scoring approach rvET reported the most PTM types and categories as significantly more conserved than controls, including both disordered and structured PTM sites of all types studied here (Fig. 2.3C). It also did not report threonine phosphorylation sites to evolve significantly faster than controls when not distinguishing between disordered and structured, though this remained a case where no significant difference could be found (T-p, Fig. 2.3C). The phylogeny-based Rate4site performed similarly with the disorder/structure distinction, except for disordered ubiquitination sites (K-ub, Fig. 2.3B), while the symbol frequency-based JSD method reported less cases of significant conservation.

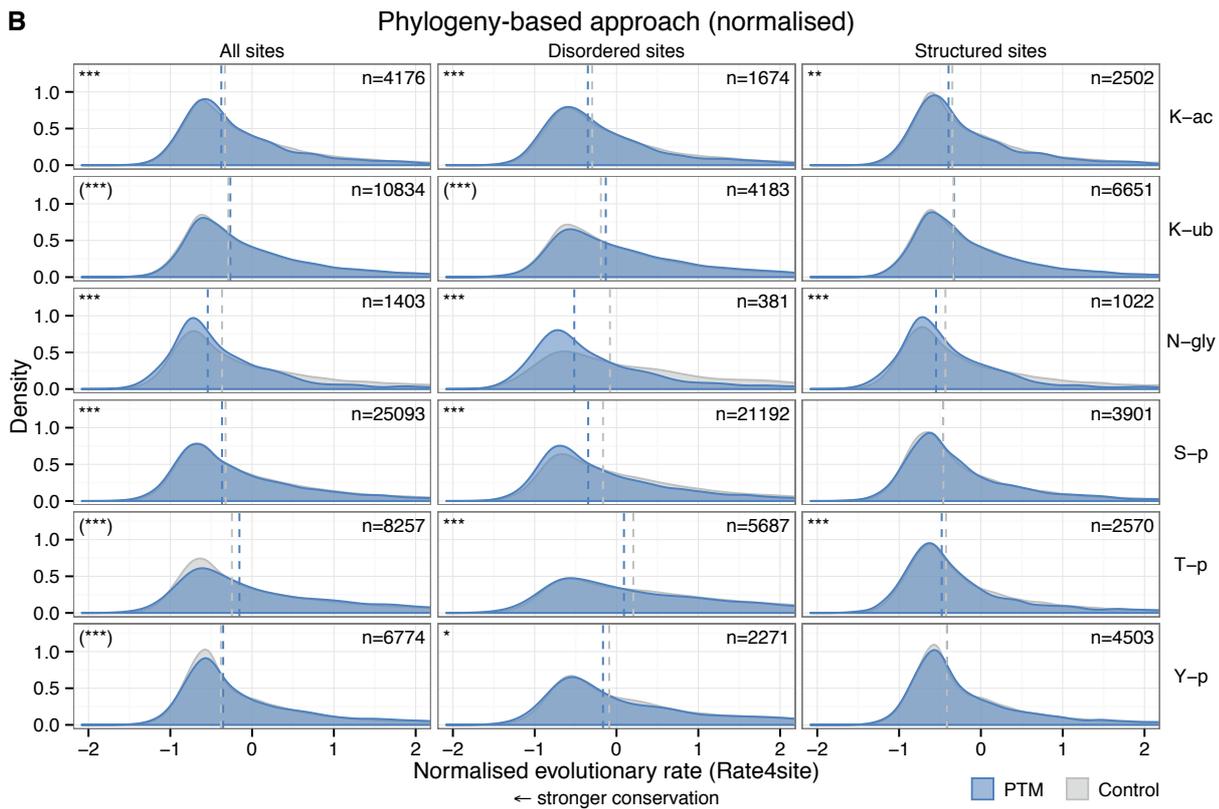
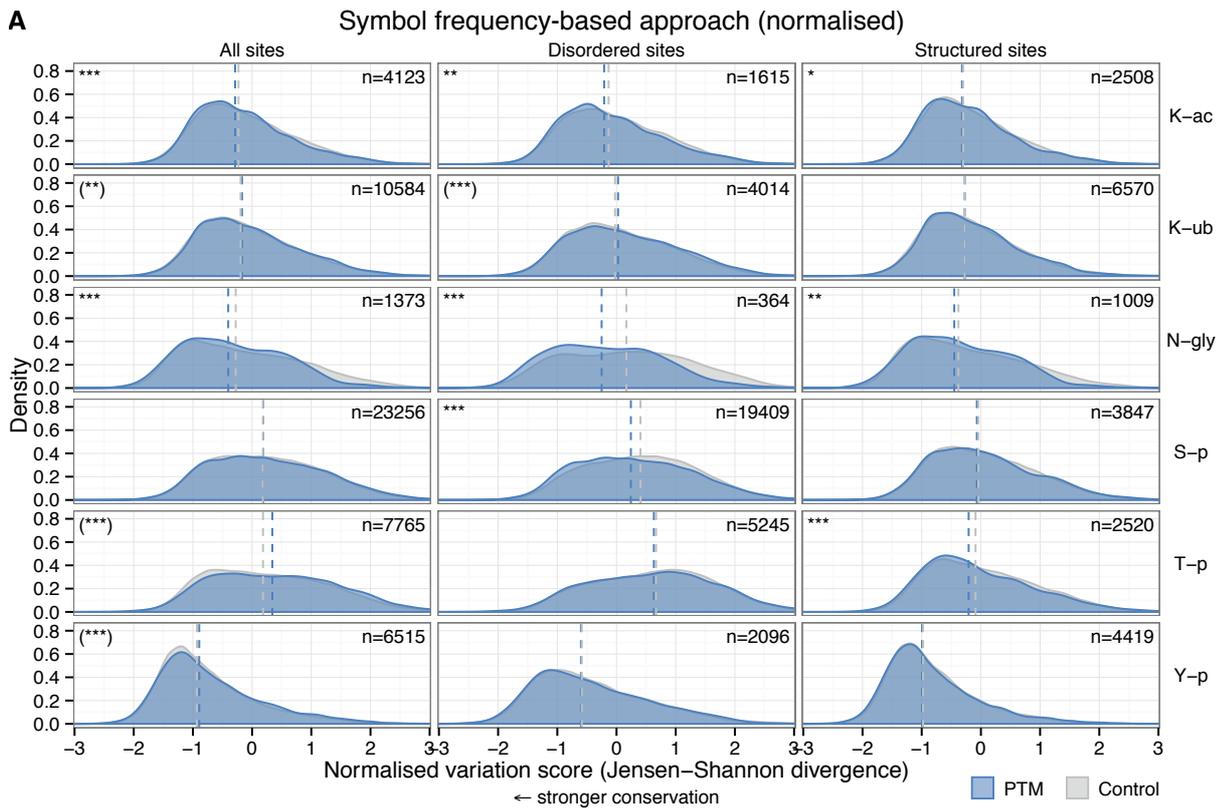
Disordered N-glycosylation sites consistently showed the strongest conservation signal compared to control sites using all three methods (Fig. 2.3A–C). One factor in explaining strong conservation of N-glycosylation sites may be its high site occupancy, i.e. nearly all copies of an N-glycosylation substrate are modified, and the fact that glycosylation is highly important for successful transit of a substrate protein through the secretory pathway due to quality control mechanisms (Ellgaard et al., 1999; Zielinska et al., 2010). The majority of serine and threonine phosphorylation sites occurred in disordered regions, while tyrosine phosphorylation, lysine acetylation, ubiquitination and N-linked glycosylation sites were found to fall more frequently into structured regions. This is in line with previous reports for acetylation and ubiquitination (Norris et al., 2009; Xu et al., 2010).

In order to determine whether PTM sites differed from control sites in their structural characteristics, the fraction of disordered residues for each PTM type was also compared to the random samples of non-modified control sites by using Fisher's exact test. No significant differences were found, indicating that PTM sites did not show a strong preference for either disordered or structured regions, although a

slight general trend towards more frequent disorder at phosphorylation and glycosylation sites was observed (data not shown). For acetylation and ubiquitination sites, the trend was towards slightly decreased disorder, though also not statistically significant.

Notably, the evolutionary rates reported by Rate4site in many cases displayed a bimodal distribution, divided into well-conserved sites on the left and rapidly evolving sites on the right (Fig. 2.3B). This is in contrast to the distributions reported by JSD and rvET, which appear largely unimodal except for a relatively small peak of highly conserved residues in some cases, especially for rvET (Fig. 2.3A and C). Normalisation of Rate4site scores (Fig. 2.4B) removes the bimodal distributions observed for unnormalised Rate4site, resulting in plots more similar to the unnormalised symbol frequency-based and hybrid approaches. Normalisation of Rate4site scores also causes minor differences in statistical significance and introduces more significantly faster-evolving cases, such as for tyrosine phosphorylation and ubiquitination sites. Similarly, normalisation appeared to decrease the sensitivity of the symbol frequency-based and hybrid approaches and leads to a general decrease in the number of categories which were found significantly conserved compared to controls (Fig. 2.4A, C). The hybrid rvET approach still reported all disordered PTM sites except ubiquitination to be significantly more conserved than controls.

Nearly all peaks of the normalised density plots are left-shifted (below the average), indicating a long tail of sites evolving much faster than the average to the right (Fig. 2.4A–C). When normalising the symbol frequency-based and hybrid scores, the highly conserved density peaks seen in some categories in Figure 2.3B and 2.3C also disappear, resulting in more homogenous plots for the three scoring methods. Whether the bimodality and peaks represent artefacts or meaningful differences in conservation between proteins is unclear. These cases may have been caused by relatively small alignments near the minimum alignment size of ten sequences (which still span at least all primates in Ensembl Compara). Since control sites were drawn from the same set of proteins, these smaller alignments should not skew the statistical tests for conservation, although they may introduce slight artefacts in the density plots which normalisation may help to alleviate.



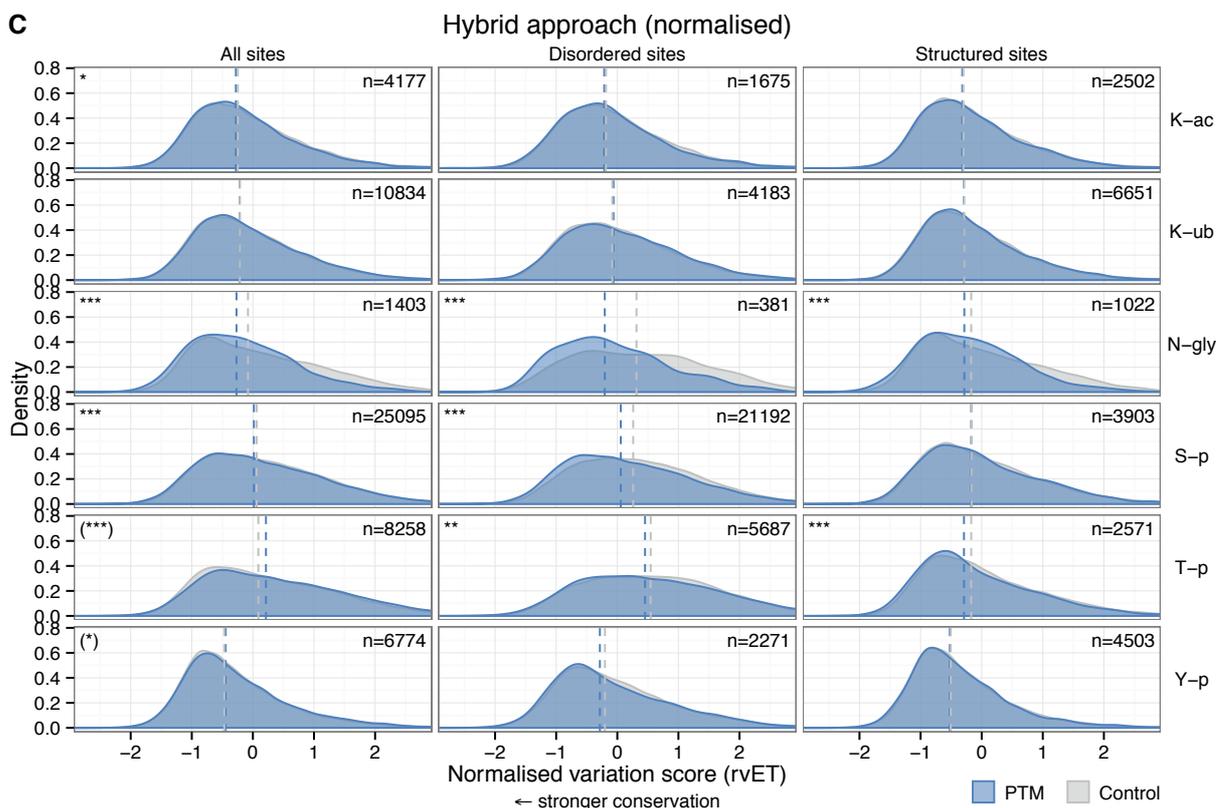


Figure 2.4: Normalised conservation density plots showing conservation of PTM sites compared to control sites for (A) Jensen-Shannon divergence (JSD), a symbol frequency-based method, (B) Rate4site, a phylogeny-based method, and (C) the real-valued Evolutionary Trace score (rvET), a hybrid approach with an entropy-based component as well as a phylogeny-based component. Normalising the Rate4site scores per protein here removes the bimodal distributions observed in Fig. 2.3C. Please see Figure 2.3 for a complete description of the plot content.

In order to systematically verify the usefulness of the structural and other classifications used in our comparisons of PTM sites to control residues above, Mann-Whitney U and permutation tests were carried out on all possible stratifications (Table 2.2). The results suggested that distinguishing simultaneously between different PTM types, disordered and structured residues, as well as core and surface residues (where PDB data was available) gives the clearest difference in conservation scores between modified and control residues, as assessed using p-values (Table 2.2B). Stratification by protein did not appear useful, resulting in insignificant p-values in some cases. Both types of tests reported very similar p-values. The directionality was always towards stronger conservation of the modified residues, as assessed using one-sided Mann-Whitney U tests (data not shown).

Table 2.2: Assessment of different stratification methods in studying the overall conservation of PTM sites. Mann-Whitney *U* tests and permutation tests were performed using all available combinations of classification criteria. (A) Results with the entire dataset, prior to the PDB-based core/surface distinction. (B) Results with the more limited dataset including the core/surface distinction. Here, insignificant p-values are marked in underlined italic text, while the optimal stratification method for each conservation scoring method is highlighted in bold. Overall, distinguishing simultaneously between PTM types, disordered and structured residues, as well as core and surface residues resulted in the most significant observed differences in evolutionary rates between modified and control residues.

Core/surface distinction made?	Conservation score	Stratification method	Stratification				Mann-Whitney <i>U</i> test	Permutation test
			by PTM	by disordered/structured	by core/surface	by protein		
No	JSD (symbol)	none	X	X	N/A	X	<2.2E-16	<2.2E-16
No	JSD (symbol)	ptm	✓	X	N/A	X	<2.2E-16	<2.2E-16
No	JSD (symbol)	disstr	X	✓	N/A	X	<2.2E-16	<2.2E-16
No	JSD (symbol)	prot	X	X	N/A	✓	7.03E-06	0.0059
No	JSD (symbol)	ptm-disstr	✓	✓	N/A	X	<2.2E-16	<2.2E-16
No	JSD (symbol)	ptm-prot	✓	X	N/A	✓	1.80E-09	4.70E-06
No	JSD (symbol)	disstr-prot	X	✓	N/A	✓	<2.2E-16	<2.2E-16
No	JSD (symbol)	ptm-disstr-prot	✓	✓	N/A	✓	<2.2E-16	<2.2E-16
No	Rate4site (phylogeny)	none	X	X	N/A	X	<2.2E-16	<2.2E-16
No	Rate4site (phylogeny)	ptm	✓	X	N/A	X	<2.2E-16	<2.2E-16
No	Rate4site (phylogeny)	disstr	X	✓	N/A	X	<2.2E-16	<2.2E-16
No	Rate4site (phylogeny)	prot	X	X	N/A	✓	<2.2E-16	<2.2E-16
No	Rate4site (phylogeny)	ptm-disstr	✓	✓	N/A	X	<2.2E-16	<2.2E-16
No	Rate4site (phylogeny)	ptm-prot	✓	X	N/A	✓	<2.2E-16	<2.2E-16
No	Rate4site (phylogeny)	disstr-prot	X	✓	N/A	✓	<2.2E-16	<2.2E-16
No	Rate4site (phylogeny)	ptm-disstr-prot	✓	✓	N/A	✓	<2.2E-16	<2.2E-16
No	rvET (hybrid)	none	X	X	N/A	X	<2.2E-16	<2.2E-16
No	rvET (hybrid)	ptm	✓	X	N/A	X	<2.2E-16	<2.2E-16
No	rvET (hybrid)	disstr	X	✓	N/A	X	<2.2E-16	<2.2E-16
No	rvET (hybrid)	prot	X	X	N/A	✓	<2.2E-16	1.11E-15
No	rvET (hybrid)	ptm-disstr	✓	✓	N/A	X	<2.2E-16	<2.2E-16
No	rvET (hybrid)	ptm-prot	✓	X	N/A	✓	<2.2E-16	<2.2E-16
No	rvET (hybrid)	disstr-prot	X	✓	N/A	✓	<2.2E-16	<2.2E-16
No	rvET (hybrid)	ptm-disstr-prot	✓	✓	N/A	✓	<2.2E-16	<2.2E-16

B

Core/surface distinction made?	Conservation score	Stratification method	Stratification				Mann-Whitney U test	Permutation test
			by PTM	by disordered/structured	by core/surface	by protein		
Yes (i.e. PDB only)	JSD (symbol)	none	×	×	×	×	1.32E-08	1.82E-05
Yes (i.e. PDB only)	JSD (symbol)	ptm	✓	×	×	×	1.43E-07	7.44E-05
Yes (i.e. PDB only)	JSD (symbol)	disstr	×	✓	×	×	1.78E-15	1.31E-11
Yes (i.e. PDB only)	JSD (symbol)	coresurf	×	×	✓	×	1.13E-12	2.24E-08
Yes (i.e. PDB only)	JSD (symbol)	prot	×	×	×	✓	<u>0.0582</u>	<u>0.2475</u>
Yes (i.e. PDB only)	JSD (symbol)	ptm-disstr	✓	✓	×	×	1.99E-11	7.24E-09
Yes (i.e. PDB only)	JSD (symbol)	ptm-coresurf	✓	×	✓	×	4.16E-09	3.16E-06
Yes (i.e. PDB only)	JSD (symbol)	ptm-prot	✓	×	×	✓	<u>0.2856</u>	<u>0.4967</u>
Yes (i.e. PDB only)	JSD (symbol)	disstr-coresurf	×	✓	✓	×	<2.2E-16	1.66E-13
Yes (i.e. PDB only)	JSD (symbol)	disstr-prot	×	✓	×	✓	0.0004	0.0029
Yes (i.e. PDB only)	JSD (symbol)	coresurf-prot	×	×	✓	✓	0.0079	0.0431
Yes (i.e. PDB only)	JSD (symbol)	ptm-disstr-coresurf	✓	✓	✓	×	7.53E-12	3.05E-09
Yes (i.e. PDB only)	JSD (symbol)	ptm-disstr-prot	✓	✓	×	✓	<u>0.0685</u>	<u>0.1183</u>
Yes (i.e. PDB only)	JSD (symbol)	ptm-coresurf-prot	✓	×	✓	✓	<u>0.1314</u>	<u>0.2590</u>
Yes (i.e. PDB only)	JSD (symbol)	disstr-coresurf-prot	×	✓	✓	✓	0.0003	0.0020
Yes (i.e. PDB only)	JSD (symbol)	ptm-disstr-coresurf-prot	✓	✓	✓	✓	<u>0.0624</u>	<u>0.1051</u>
Yes (i.e. PDB only)	Rate4site (phylogeny)	none	×	×	×	×	0.0105	0.0074
Yes (i.e. PDB only)	Rate4site (phylogeny)	ptm	✓	×	×	×	0.0012	0.0009
Yes (i.e. PDB only)	Rate4site (phylogeny)	disstr	×	✓	×	×	0.0003	5.63E-05
Yes (i.e. PDB only)	Rate4site (phylogeny)	coresurf	×	×	✓	×	0.0005	0.0004
Yes (i.e. PDB only)	Rate4site (phylogeny)	prot	×	×	×	✓	0.0374	<u>0.0542</u>
Yes (i.e. PDB only)	Rate4site (phylogeny)	ptm-disstr	✓	✓	×	×	2.70E-06	1.09E-06
Yes (i.e. PDB only)	Rate4site (phylogeny)	ptm-coresurf	✓	×	✓	×	6.64E-05	2.56E-05
Yes (i.e. PDB only)	Rate4site (phylogeny)	ptm-prot	✓	×	×	✓	0.0010	0.0035
Yes (i.e. PDB only)	Rate4site (phylogeny)	disstr-coresurf	×	✓	✓	×	6.72E-05	1.71E-05
Yes (i.e. PDB only)	Rate4site (phylogeny)	disstr-prot	×	✓	×	✓	0.0020	0.0020
Yes (i.e. PDB only)	Rate4site (phylogeny)	coresurf-prot	×	×	✓	✓	0.0043	0.0049
Yes (i.e. PDB only)	Rate4site (phylogeny)	ptm-disstr-coresurf	✓	✓	✓	×	2.10E-06	3.06E-07
Yes (i.e. PDB only)	Rate4site (phylogeny)	ptm-disstr-prot	✓	✓	×	✓	6.58E-05	0.0003
Yes (i.e. PDB only)	Rate4site (phylogeny)	ptm-coresurf-prot	✓	×	✓	✓	0.0002	0.0006
Yes (i.e. PDB only)	Rate4site (phylogeny)	disstr-coresurf-prot	×	✓	✓	✓	0.0005	0.0004
Yes (i.e. PDB only)	Rate4site (phylogeny)	ptm-disstr-coresurf-prot	✓	✓	✓	✓	7.17E-05	0.0002
Yes (i.e. PDB only)	rvET (hybrid)	none	×	×	×	×	4.84E-14	7.64E-10
Yes (i.e. PDB only)	rvET (hybrid)	ptm	✓	×	×	×	8.88E-16	2.60E-11
Yes (i.e. PDB only)	rvET (hybrid)	disstr	×	✓	×	×	<2.2E-16	1.71E-14
Yes (i.e. PDB only)	rvET (hybrid)	coresurf	×	×	✓	×	4.44E-16	7.59E-12
Yes (i.e. PDB only)	rvET (hybrid)	prot	×	×	×	✓	<u>0.2210</u>	<u>0.1025</u>
Yes (i.e. PDB only)	rvET (hybrid)	ptm-disstr	✓	✓	×	×	<2.2E-16	1.11E-15
Yes (i.e. PDB only)	rvET (hybrid)	ptm-coresurf	✓	×	✓	×	<2.2E-16	4.17E-13
Yes (i.e. PDB only)	rvET (hybrid)	ptm-prot	✓	×	×	✓	<u>0.0679</u>	0.0239
Yes (i.e. PDB only)	rvET (hybrid)	disstr-coresurf	×	✓	✓	×	<2.2E-16	1.33E-15
Yes (i.e. PDB only)	rvET (hybrid)	disstr-prot	×	✓	×	✓	0.0077	0.0015
Yes (i.e. PDB only)	rvET (hybrid)	coresurf-prot	×	×	✓	✓	<u>0.0525</u>	0.0124
Yes (i.e. PDB only)	rvET (hybrid)	ptm-disstr-coresurf	✓	✓	✓	×	<2.2E-16	2.22E-16
Yes (i.e. PDB only)	rvET (hybrid)	ptm-disstr-prot	✓	✓	×	✓	0.0081	0.0019
Yes (i.e. PDB only)	rvET (hybrid)	ptm-coresurf-prot	✓	×	✓	✓	0.0311	0.0058
Yes (i.e. PDB only)	rvET (hybrid)	disstr-coresurf-prot	×	✓	✓	✓	0.0032	0.0004
Yes (i.e. PDB only)	rvET (hybrid)	ptm-disstr-coresurf-prot	✓	✓	✓	✓	0.0085	0.0012

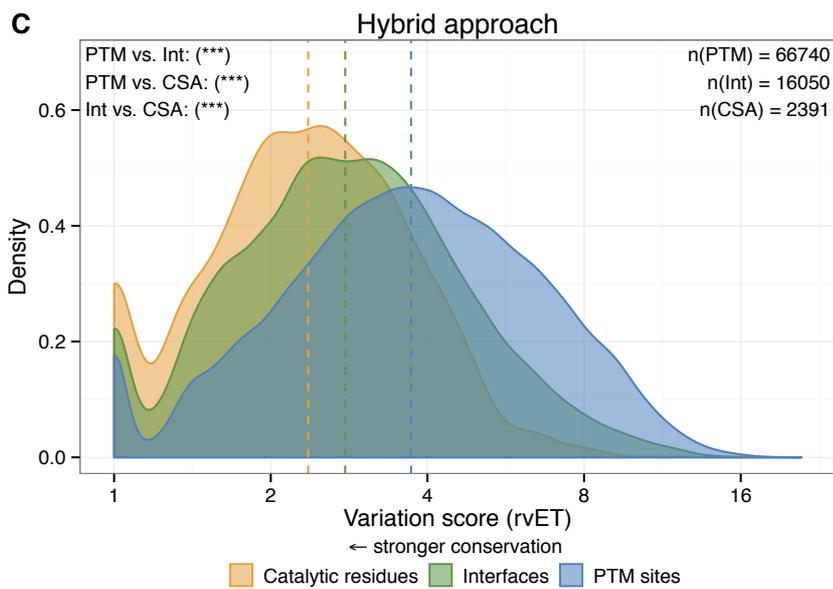
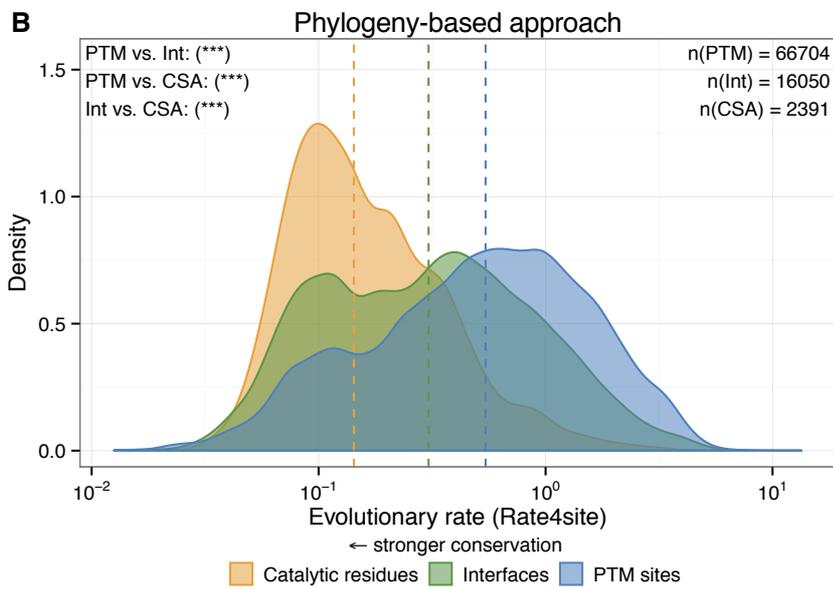
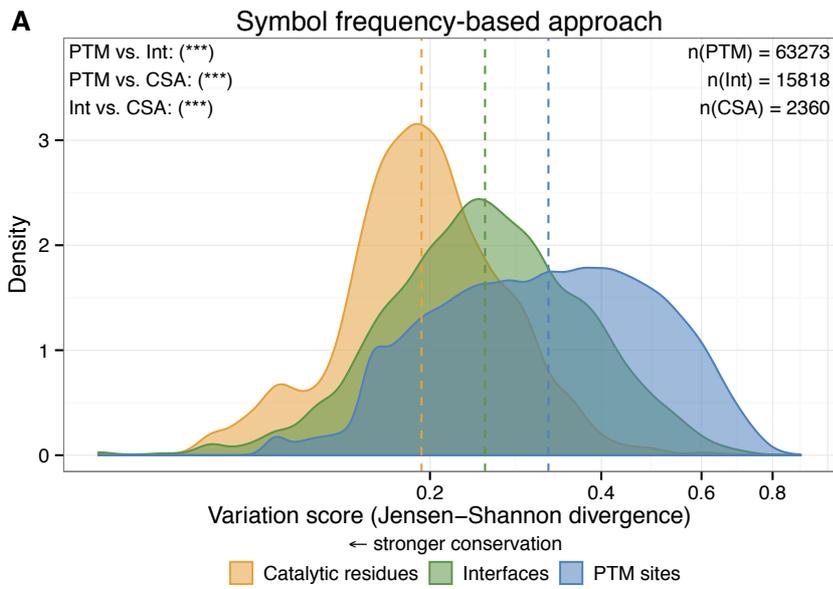


Figure 2.5: Comparison of evolutionary rates for catalytic residues from the Catalytic Site Atlas (“CSA”), protein-protein interaction interfaces (“Int”) and modified residues (“PTM”). Conservation density plots showing conservation assessed using (A) Jensen-Shannon divergence (JSD), a symbol frequency-based method, (B) Rate4site, a phylogeny-based method, and (C) the real-valued Evolutionary Trace score (rvET), a hybrid approach with an entropy-based component as well as a phylogeny-based component. Low variation scores indicate strong sequence conservation. Asterisks indicate the level of statistical significance between samples (*: $p < 0.001$, **: $p < 0.01$, *: $p < 0.05$), as determined by two-sided Mann-Whitney U tests. The dashed lines indicate the median of each sample. Where the median of the first sample is higher than that of the second, these asterisks are shown in round brackets as in previous figures. The size of each sample is given in the top right corner.**

Additionally, to obtain a frame of reference for the observed variation scores, we performed a comparison between PTM sites, residues participating in protein-protein interaction interfaces and catalytic residues (Fig. 2.5). All three categories differed significantly from each other, with catalytic residues being the most highly conserved group and interaction interfaces generally being more conserved than PTM sites. Although catalytic residues were the most numerous among the strongly conserved sites, all three categories overlapped in this area. The bimodal distribution observed for phylogeny-based Rate4site score (Fig. 2.5B) showed this most clearly, displaying an equivalent peak of highly conserved residues in all three categories, and indicating that certain interface residues and PTM sites may be conserved to an equivalent extent as catalytic residues. Similarly, the hybrid rvET score (Fig. 2.5C) displayed a peak of absolutely conserved residues with a score of 1 for all three categories, which was largest for the catalytic residues.

2.3.2. Many PTM sites have ancient evolutionary origins

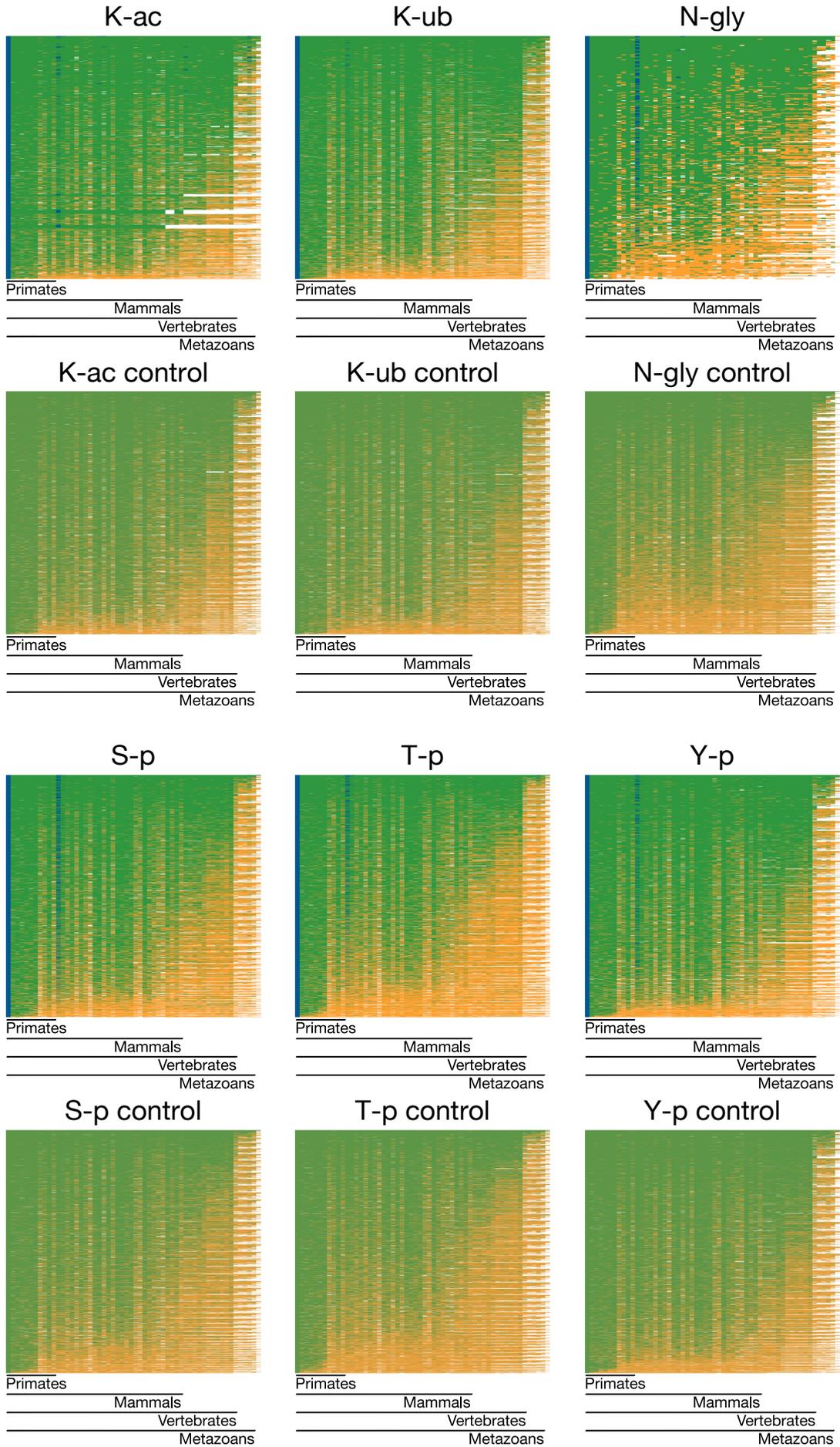
In addition to quantifying conservation at PTM sites using aggregate conservation scores, we also traced the conservation of individual modified residues across species. Figure 2.6 shows conservation profiles for disordered and structured human PTM sites in a large cross-section of other species from Ensembl Compara. These species are ordered according to their phylogenetic relationships in the Compara species tree (Flicek et al., 2012), from human on the left to yeast on the very right. Though there appears to be evidence of mutations being retained within lineages, there is also a considerable amount of variability which does not seem confined to a lineage. This seems to be especially the case for human N-

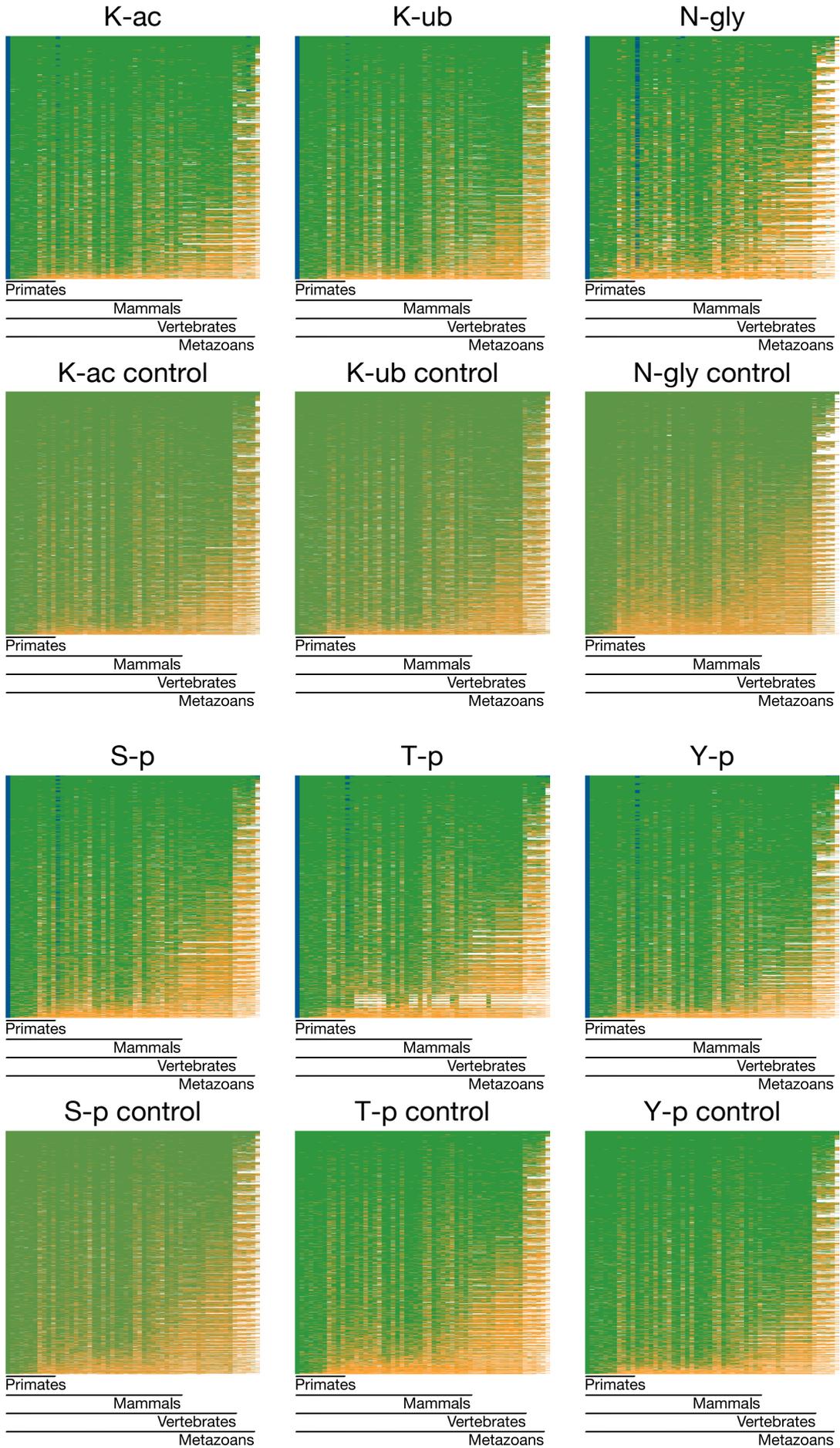
glycosylated residues. In at least some cases, the absence of orthologous residues (white tiles in Fig. 2.6) could also point to insufficient sequencing coverage.

Overall, there is a large amount of conservation of human post-translationally modified residues extending beyond the mammalian lineage. In many cases, there appears to be a boundary upon leaving the gnathostomata (from *Danio rerio* to *Petromyzon marinus*, which is the rightmost vertebrate), or the tetrapoda (from *Xenopus tropicalis* to *Tetraodon nigroviridis*, centrally between the mammalian and vertebrate boundaries). Lysine acetylation, ubiquitination and tyrosine phosphorylation sites are relatively more conserved than the other PTMs studied, though this seems to stem more from the type of amino acid itself, as evidenced by the control residues. Tyrosines appear more conserved than serines and threonines. This may be due to positive selection for tyrosine loss caused by the advent of tyrosine phosphorylation early in the metazoan lineage, potentially resulting in stronger conservation of the tyrosines which were retained (Tan et al., 2009b).

For all PTMs, this analysis illustrates the low experimental coverage of post-translational modifications in most species other than mice and humans. Due to the sparseness of experimental PTM site information for other species, relatively little experimental overlap could be detected, even in the mouse. An exception is N-glycosylation, where large-scale analysis has been undertaken in the mouse, and a good overlap with human sites is clearly evident. Although *S. cerevisiae*, *Drosophila melanogaster* and *C. elegans* are relatively well-studied, surprisingly few human PTM sites appeared to have precisely conserved, modified orthologous residues in these species, in line with a previous report on yeast-human conservation (Minguez et al., 2012).

Interestingly, many acetylated disordered lysines which are also modified in the mouse appear to occur in proteins with no homologs in most mammals. A similar “gap” of PTM sites conserved identically in a certain lineage, but with few homologous proteins outside it is visible for structured threonine phosphorylation sites (T-p, Fig. 2.6B).

A**Disordered PTM sites**

B**Structured PTM sites**

- C**
- residue is conserved and modified
 - residue is conserved
 - residue not conserved
 - no orthologous protein

Figure 2.6: Conservation profiles of different human PTM types and unmodified control residues for (A) disordered and (B) structured sites. The figure legend is shown in panel C. Species are shown on the horizontal axis of each panel, and individual PTM sites or control residues are shown on the vertical axis. Species are arranged according to their phylogenetic relationships, from human on the left to yeast on the right, as obtained from the Ensembl Compara species tree. Green tiles indicate the presence of an identical homologous residue. Blue tiles further indicate that an identical, homologous residue is experimentally known to be modified by the same PTM type. Annotations by similarity were excluded here. In orange tiles, a homologous protein exists, but does not contain an identical residue, while white indicates that no homologous position exists in a species's proteome. Vertically, the residues are arranged in descending order by the number of species with identical homologous residues (marked in green), and in case of ties secondarily by the number of species with a homologous protein (marked in orange), and additionally by the number of known modified identical residues (marked in blue).

2.3.3. Mutations at PTM sites may mimic or avoid the modified state

Since mimetic mutations can mirror the effect of a PTM (Thorsness and Koshland, 1987), we decided to investigate whether PTM sites differ from unmodified residues in the types of residues that are present at orthologous positions. An enrichment in mimetic residues would be an additional indication that PTM sites tend to be functionally important. Although the directionality of substitution in other species cannot be determined since we have no direct knowledge of the ancestral sequences, we will refer to the presence of amino acids which differ from the modifiable human residue as “mutations” or “substitutions” here as a simplification. We decided to use the best-conserved tertile of PTM sites to determine score cutoffs, since well-conserved sites should ensure limited functional divergence. Using the raw, unnormalised scores from the three different conservation scoring methods, significant substitution differences generally did not appear consistently with all three scoring methods. Since normalisation resulted in more uniform score distributions in Figure 2.4, we decided to repeat this analysis with normalised scores.

When the scores were normalised per protein, certain enrichments appeared consistently (Fig. 2.7A). The unnormalised Rate4site result resembled these consistent results from normalised scores slightly more closely than the other methods (Fig. 2.7B), and was therefore included here as the only raw score in the interest of space. This resemblance only existed for structured tyrosines, disordered serines and perhaps lysines (Fig. 2.7B). For disordered tyrosines, the substitution pattern did not differ significantly from control sites, while disordered threonine phosphorylation sites had also not shown a bimodal distribution in their raw Rate4site conservation scores (Fig. 2.3B), which may offer an explanation for the absence of an enrichment in their column here.

We then decided to focus on the normalised conservation scores (Fig. 2.7A) for the substitution analysis, since they were best able to recapture the significant conservation of modified residues which we detected previously in Figures 2.3–4. Despite reporting a relatively small number of PTM categories as significantly more conserved than controls in Fig. 2.4A, the normalised symbol frequency-based JSD method was well able to recapture overrepresentation of these original residues here. This was the case for each disordered PTM type except for lysine acetylation, which JSD reported to be significantly more frequently replaced by negatively-charged glutamate (E) compared to unmodified lysines. The good performance of the JSD method may be explained here by the similarity of its symbol frequency-based approach to the symbol overrepresentation test used here, since JSD may be most likely to retrieve the tertile with the lowest number of substitutions out of the three conservation scoring methods, since it does not take phylogeny into account.

In addition to the original amino acids for each PTM type, we expected the substitution analysis to show an overrepresentation of negatively charged phosphomimics at disordered serine phosphorylation sites (Kurmangaliyev et al., 2011; Pearlman et al., 2011). Though small compared to the enrichment for serine, this was indeed retrieved by all three normalised scoring methods, as well as all raw scores, with Rate4site showing the strongest enrichment of both aspartate (D) and glutamate (E).

We additionally found that aspartate substitutions (D) might be enriched at phosphorylated structured tyrosines, which may also constitute a previously unreported type of phosphomimicry. Intriguingly, acetylated disordered lysines might

preferentially become replaced by negatively charged glutamate (E) according to both the normalised symbol frequency-based JSD score and the raw phylogeny-based Rate4site score. Structured acetylated lysines were enriched in substitutions by either glutamine (Q) and asparagine (N), two structurally similar amino acids with carboxamide side chains, which only differ in their carbon chain length. Disordered N-glycosylated asparagines (N) were found to be significantly enriched in glutamine (Q) substitutions. Since asparagine and glutamine are structurally very similar, mutation to glutamine (Q) could potentially abolish N-glycosylation while maintaining a very similar structure.

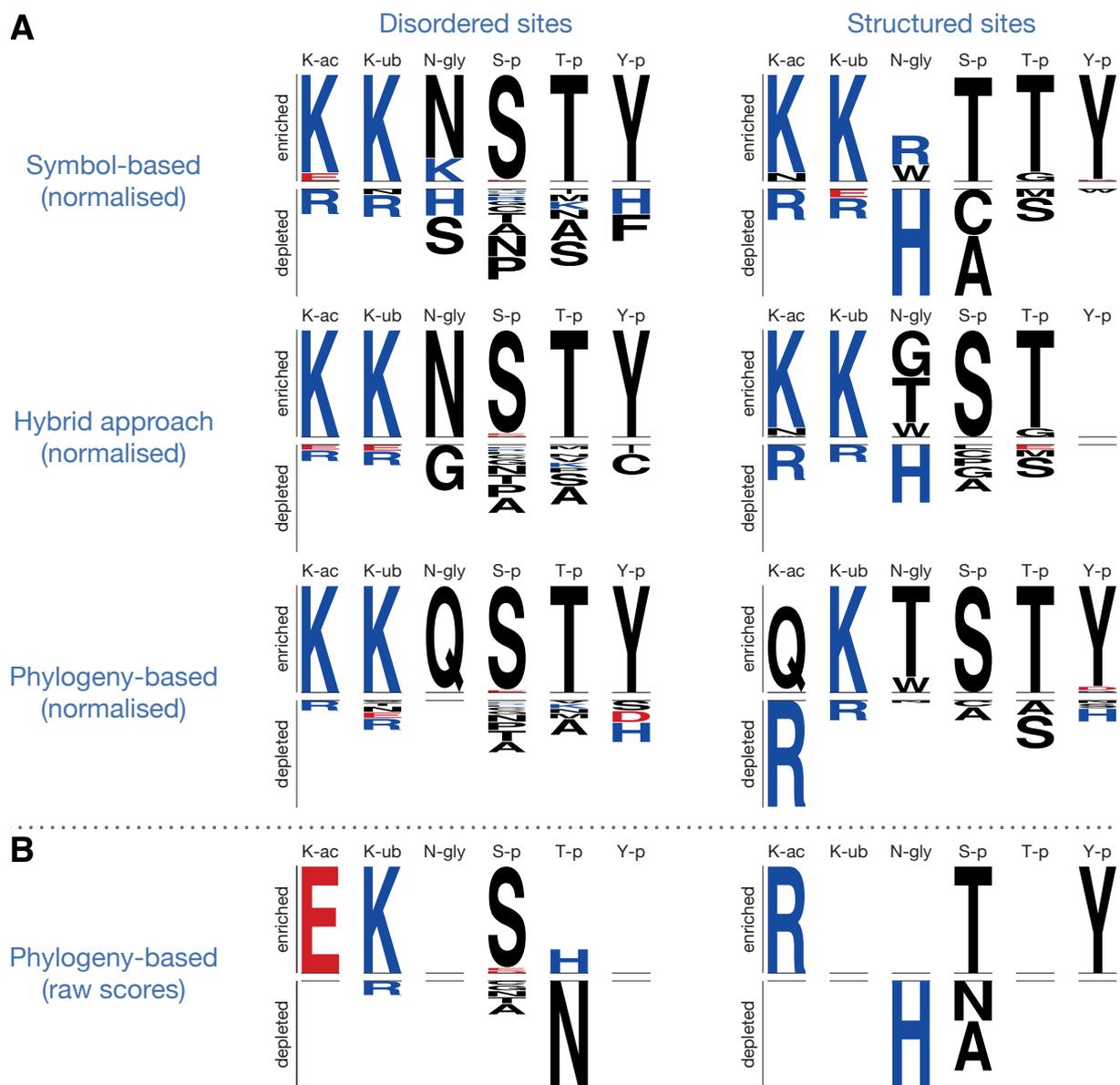


Figure 2.7: Amino acid over- and underrepresentation at well-conserved PTM sites, using the best-conserved third of PTM sites according to (A) the three normalised conservation scores and (B) the phylogeny-based raw Rate4site scores. Amino acids which were significantly more or less frequently encountered at PTM sites compared to similarly conserved control residues are shown in colours according to their charge (blue: positive, black: neutral, red: negative). The analysis is broken down horizontally into disordered and structured residues, as well as vertically by the conservation scoring method used to determine the best-conserved PTM site tertile range. Within each panel, the six PTM types studied are shown left-to-right. Overrepresentation is shown in the upper half of each panel, while underrepresentation is shown in the lower half. The scaling of the one-letter amino acid codes is proportional to their relative degree of over- or underrepresentation within an individual panel column.

2.3.4. PTM sites are more constrained than control residues at the population and somatic levels

Thus far, we have attempted to address the likely functionality of PTM sites by investigating conservation at modified residues between species. However, this approach does not confirm functional relevance of a specific site within the human species. Functional constraints on post-translationally modified sites in humans was therefore assessed by using data on non-synonymous single-nucleotide polymorphisms (nsSNPs) within human populations (Fig. 2.8A), and on disease-causing variants and cancer-associated somatic mutations in individuals (Fig. 2.9). This analysis was further substantiated using variation data from 38 strains of *S. cerevisiae* (Fig. 2.8B) (Liti et al., 2009).

Consistently, we found PTM sites to be significantly less frequently affected by natural variants than unmodified control residues (Fig. 2.8), and significantly more frequently affected by disease mutations (Fig. 2.9). In most non-significant cases in the largest datasets (1000 Genomes and HGMD), this is still maintained as a trend. Notably, our analysis of this intra-species data strongly highlights the conservation of structured PTM sites in addition to disordered sites. While N-glycosylation and threonine phosphorylation were the only PTM types which were consistently reported as more conserved than controls at structured sites using normalised scores in Figure 2.4, all structured PTM types were here found to be either significantly depleted in natural variants (Fig. 2.8A) or enriched in disease mutations (Fig. 2.9A). These findings are supported by an earlier study using a smaller PTM site dataset, which reported that disease mutations are enriched at PTM sites (Li et al., 2010).

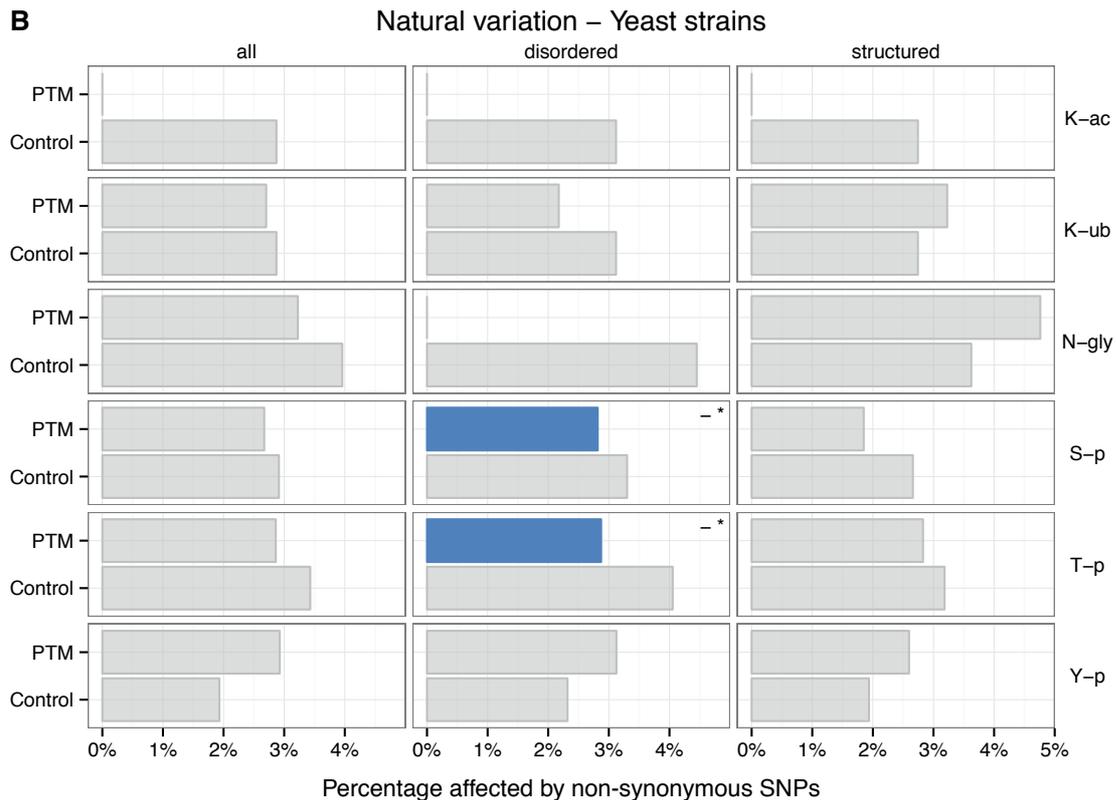
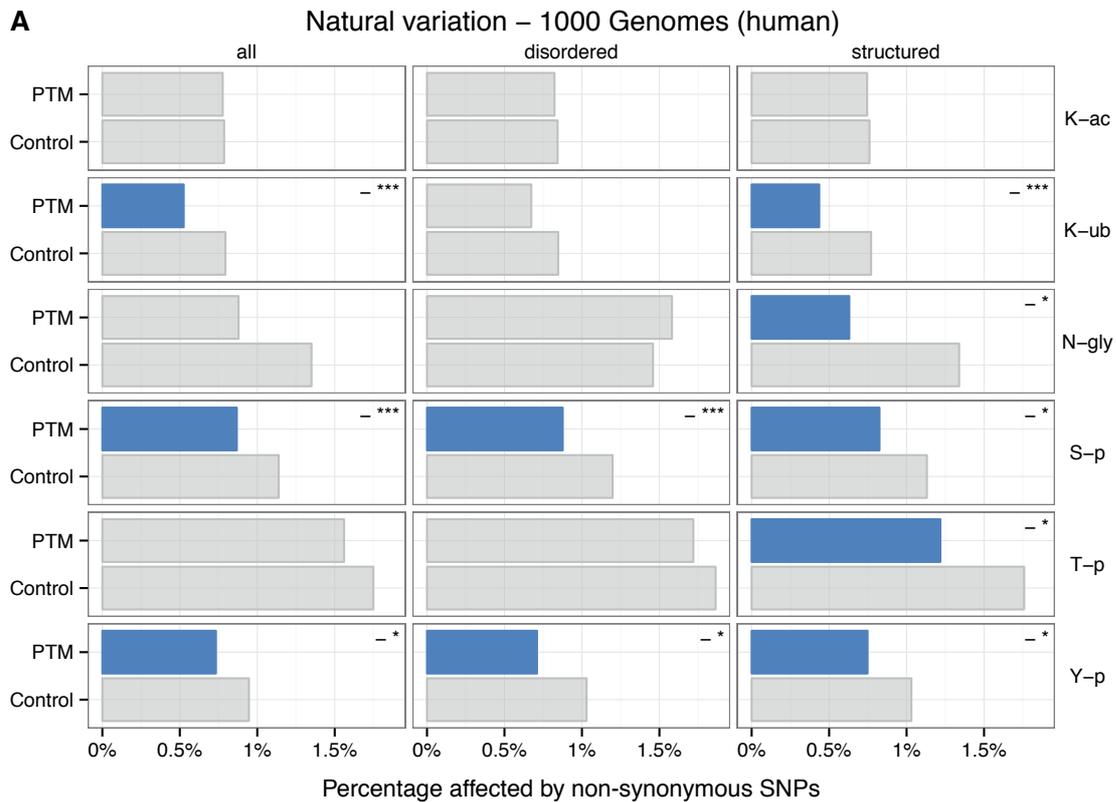


Figure 2.8: Decreased nsSNP incidence at PTM sites in natural variation, assessed using data from (A) the human 1000 Genomes Project and (B) yeast strain variation data. Cases where PTM sites are significantly less often affected by non-synonymous SNPs are highlighted in blue and with a minus symbol. Asterisks indicate the level of statistical significance of the difference (*: $p < 0.001$, **: $p < 0.01$, *: $p < 0.05$), as determined by Fisher's exact test.**

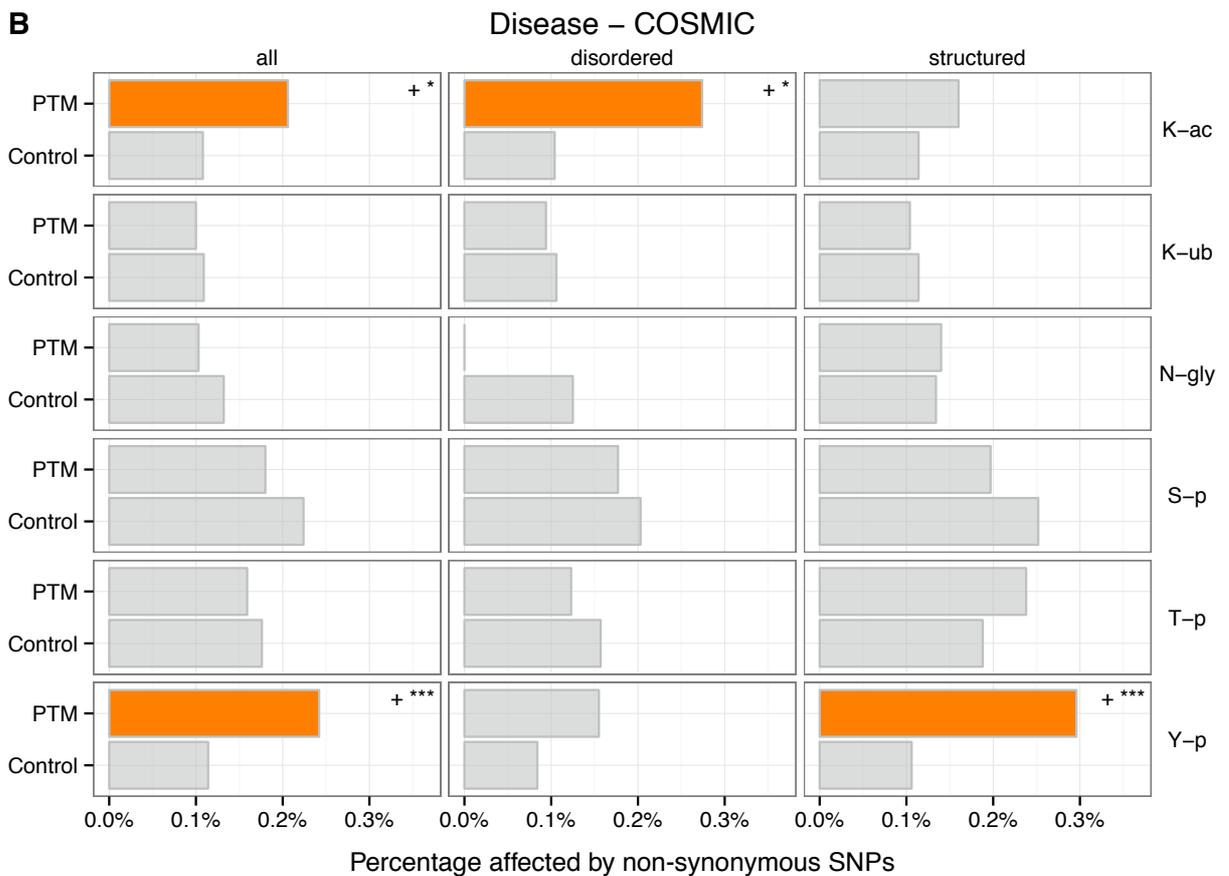
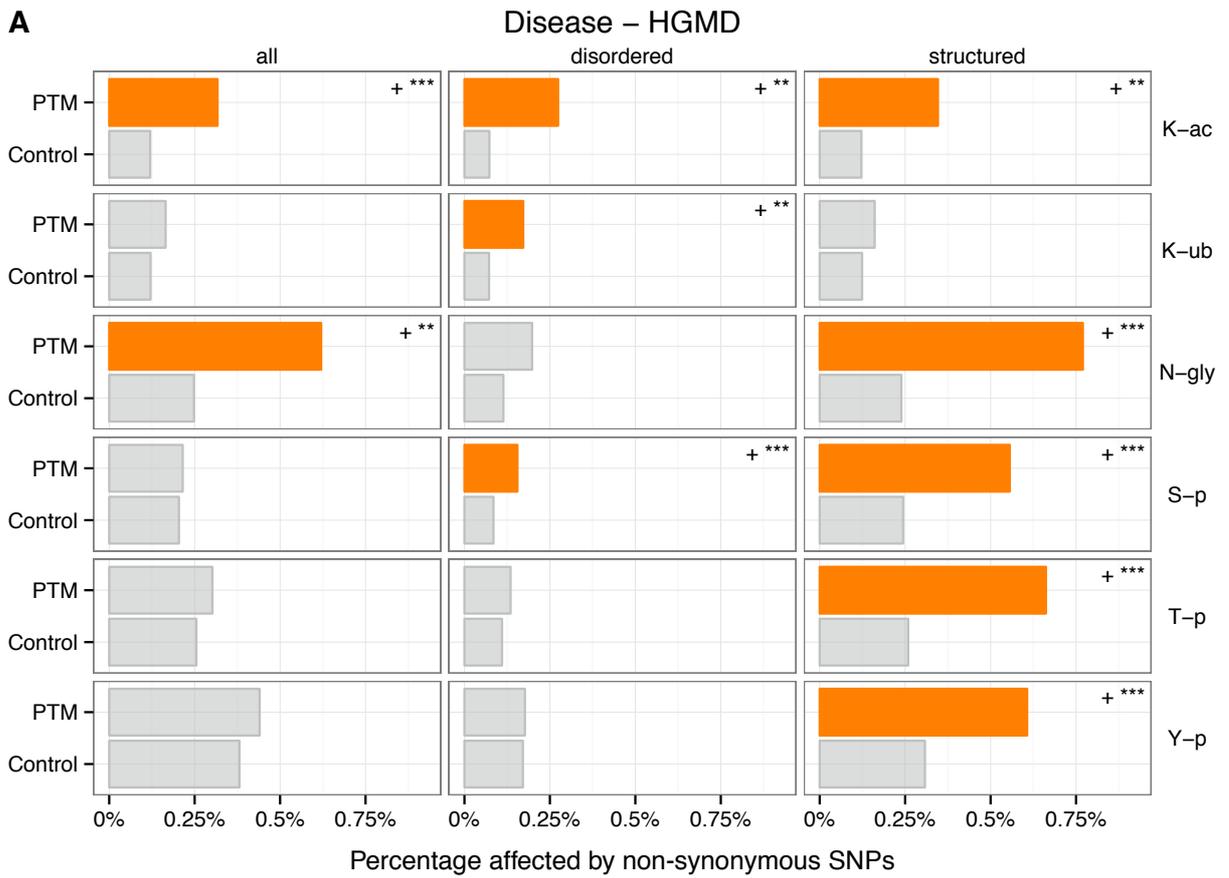


Figure 2.9: Increased nsSNP incidence at PTM sites in disease, assessed using (A) disease mutation data from the Human Gene Mutation Database (HGMD) and (B) cancer-associated somatic mutations from the COSMIC database. Cases where PTM sites are significantly more often affected by non-synonymous SNPs are highlighted in orange and with a plus symbol. Asterisks indicate the level of statistical significance of the difference (*: $p < 0.001$, **: $p < 0.01$, *: $p < 0.05$), as determined by Fisher's exact test.**

2.4. Discussion

The analyses in this chapter were prompted by the open question of whether a large fraction of PTM sites may be biologically nonfunctional. We here investigated this question by quantifying selection pressure acting on post-translationally modified residues at the species, population and somatic levels, and by analysing evolutionary profiles of these sites. We also addressed the tendency of certain modifications to occur in fast-evolving intrinsically disordered regions, and the intriguing subject of mimetic and avoiding mutations at modification sites. We found mild, but significant evidence of increased conservation of at least a fraction of PTM sites at all points.

2.4.1. Quantifying PTM site conservation using conservation scores

Using three different conservation scoring methods, we investigated the conservation of human PTM sites while correcting for structural characteristics affecting evolutionary rates. Though wide ranges of variation in scores were observed, as reported in previous studies of evolutionary rates (Subramanian and Lambert, 2011), we found that disordered and structured modified residues were generally significantly more conserved overall than unmodified controls (Fig. 2.3–4). Using conservation scores normalised per protein however, structured sites for PTM types other than N-glycosylation and threonine phosphorylation appeared to be similarly conserved as control residues (Fig. 2.4).

Well-conserved PTM sites appear similarly conserved as catalytic residues

A comparison was performed between PTM sites, residues participating in protein-protein interaction interfaces and catalytic residues (Fig. 2.5), allowing us to obtain a frame of reference for the observed variation scores. Though catalytic residues were significantly more highly conserved as a group than interaction interfaces, and these

respectively more than PTM sites as a group, all three distributions overlapped in the highly-conserved peak defined by the catalytic residues. Especially the bimodal distribution observed for phylogeny-based Rate4site score (Fig. 2.5B), but also the peak of absolutely conserved residues with a score of 1 for the hybrid rvET score (Fig. 2.5B) showed this most clearly, displaying an equivalent peak of highly conserved residues in all three categories, and indicating that certain interface residues and PTM sites are conserved to an equivalent extent as catalytic residues.

The observed non-conservation of structured PTM sites using normalised scores may be due to high background conservation at structured control sites

Though structured residues likely to be in the protein core were removed from the control samples, the discrepancy we observed between the generally significantly conserved disordered and the less frequently conserved structured sites may make sense in light of the lower background conservation of disordered regions (Brown et al., 2002; 2010; 2011). Against a disordered background, the conservation of functional residues might be much more apparent than against a background under stronger structural constraint. An alternative explanation could be provided for phosphorylation by the polyelectrostatic model, which suggests that phosphorylated residues in disordered regions can present themselves as diffuse charges, involving rapidly changing conformation of the disordered region (Borg et al., 2007). Recent advances in the structural analysis of disordered proteins support this view of disordered proteins as “structural ensembles” that may rapidly transition between conformations (Lange et al., 2008; Marsh and Forman-Kay, 2011). If polyelectrostatic mechanisms are generally important at phosphorylation sites, this may help to explain the increased conservation of disordered phosphorylation sites which we observed.

Overall, our results illustrate the importance of structural distinctions, since prior to the distinction between disordered and structured regions we were frequently unable to detect significant conservation (Fig. 2.3–4). This is further shown in detail in our experimentation with all possible stratification combinations in Table 2.2.

The choice of conservation scoring method is important

The real-valued Evolutionary Trace score, a hybrid method combining an entropy-based and a phylogeny-based component, performed better than alternatives and detected significant conservation of disordered and structured sites for all PTM types studied here when using the raw score. The fact that the three methods performed very similarly in an independent evaluation (Johansson and Toh, 2010), but showed notable differences in results here suggests that the choice of conservation scoring method is of great importance when analysing PTM site conservation. It has been suggested that certain datasets of known functional sites which are used in the evaluation of scoring methods may generally be biased towards absolutely conserved residues (Capra and Singh, 2007). This seems relatively appropriate for PTM sites, unless mimetic mutations are taken into account.

The bimodality of scores observed using Rate4site may or may not be artifactual

Conservation density plots using unnormalised Rate4site scores generally displayed bimodal distributions of evolutionary rates (Fig. 2.3B). In support of multimodality of residue conservation, a previous report has described evolutionary rates as forming several peaks for the majority of proteins: slowly evolving “core” residues, fast-evolving surface residues, and extremely evolvable residues in disordered regions (Tóth-Petróczy and Tawfik, 2011). Although the bimodal distributions at PTM sites reported by Rate4site could be artifactual, potentially due to score dependency on alignment size, they might then also signify that there are two evolutionary classes of PTM sites: i) well-conserved residues that are functionally conserved far back in evolution, and ii) less conserved residues that have either only recently acquired importance in post-translational signalling, or are non-functional PTM sites and potentially only present at low occupancy, i.e. in a small fraction of protein copies (Beltrao et al., 2012). This view can be supported by our results in Figure 2.5C, where the highly-conserved component of the bimodal distribution of Rate4site scores appears to overlap well with the peak for catalytic residues.

Although the other scoring methods did not display clear bimodal distributions (only small peaks of well-conserved residues), the fact that the unnormalised Rate4site

score best retrieved aspartate (D) and glutamate (E) enrichments at disordered serine phosphorylation sites (S-p, Fig. 2.7B) might lend a small degree of support to the idea that the uniquely bimodal distribution of the unnormalised Rate4site scores could represent a meaningful difference between two populations of residues. Rate4site was the only method which explicitly used branch lengths for scoring conservation, which may potentially lend it increased sensitivity. The fact that the bimodality disappeared upon normalisation at the protein level (Fig. 2.4B) may be due to differential conservation of entire proteins, which could be influenced by the presence or absence of functional PTM sites.

The low conservation observed for lysine ubiquitination sites using normalised scores could be due to degradation-associated sites

The relatively low conservation of ubiquitination we observed using normalised scores could potentially be due to relatively unspecific degradation-related ubiquitination. It has been noted that ubiquitination mostly occurs in structured regions, although less frequently if the sites are associated with degradation (Hagai and Levy, 2010; Hagai et al., 2011). This finding, in addition to the higher background conservation at unmodified control residues in structured regions, may explain the absence of observable conservation of ubiquitination sites using normalised scores (Fig. 2.4A–C). Recent studies that distinguish between ubiquitin chain types could be used to address this, but these data are not yet available at a large enough scale (Sims et al., 2012).

PTM sites do not appear to show disorder or structure preference when compared to control residues sampled from the same proteins

Previous studies have generally reported PTM sites to fall preferentially into disordered regions (Iakoucheva et al., 2004; Xie et al., 2007a; Gnad et al., 2009; Holt et al., 2009; Landry et al., 2009; Ba and Moses, 2010). In our analyses, though some PTM types were indeed observed to fall more frequently into either disordered or structured regions, these structural preferences were not significantly different from unmodified control residues of the same amino acid type within the same set of proteins, indicating that PTM sites did not show a significant preference for either disordered or structured regions, although a general trend towards more frequent

disorder at phosphorylation and glycosylation sites was indeed observed. Lysine acetylation and ubiquitination sites were instead found to be slightly, but insignificantly, more frequently structured than their unmodified counterparts. Potentially, then, intrinsic disorder might not be a generally required feature of an accessible, modifiable residue, and this would indicate that intrinsic disorder may not always be a useful characteristic for predicting PTM sites.

2.4.2. Analysis of PTM site evolution using conservation profiles

Evidence of lineage-specific and sporadic conservation of PTM sites

In the conservation profiles of human post-translationally modified residues (Fig. 2.6), there appeared to be a considerable amount of variability which did not seem confined to a lineage. Aside from non-functionality of some PTM sites at least in species other than human, possible explanations for these noisy patterns could include post-translational regulatory divergence in individual species, potentially including mutations which mimic or avoid the modified state (Fig. 2.7), and which may hypothetically underlie phenotypic differences between species. However, many control residues showed similar patterns. Human N-glycosylation sites had especially few orthologous residues in non-mammalian and in non-vertebrate species. This may point to extensive differences in the types of glycoproteins these species generate, which have been reported to be a major factor in speciation (Varki and Altheide, 2005; Varki and Nelson, 2007; Varki et al., 2008; Varki, 2010; Varki et al., 2011). Conversely, human structured acetylated lysines appeared to be conserved in a relatively large number of species, though not drastically more often than control lysines.

Invertebrates, vertebrates and primates appear divergent in PTM signalling

The observation that *Drosophila melanogaster*, *C. elegans* and *S. cerevisiae*, despite being relatively well-studied in terms of PTM sites, have surprisingly few modified residues orthologous to human PTM sites could either be due to non-position specific conservation as in the polyelectrostatic model (Borg et al., 2007), or due to divergence of these post-translational signalling systems at many sites. It has been suggested that the majority of phosphorylation events on both sides may have

evolved after the divergence of higher eukaryotes, *C. elegans* and yeast (Zielinska et al., 2009; Gnad et al., 2010a). This would indicate that more closely related model organisms should be used in the study of human PTM signalling. Corroborating this, a relatively large fraction of disordered serine and threonine phosphorylation sites appears to be specific to primates (Fig. 2.6). This might indicate extensive changes in serine and threonine phosphorylation signalling in the early divergence of the primate lineage.

2.4.3. Analysis of substitution patterns at PTM sites

Mutations at PTM sites tend to resemble either the modified or unmodified state

Substitutions by negatively charged amino acids can successfully mimic the phenotype of phosphorylated residues (Thorsness and Koshland, 1987; Tarrant and Cole, 2009), and *vice versa* some phosphorylation sites appear to have originated from negatively charged amino acids. These amino acids may have been part of an ancestral salt bridge, which phosphorylation can conditionally restore (Nühse et al., 2004; Kurmangaliyev et al., 2011; Pearlman et al., 2011). Conversely, mutation to a negatively charged residue would effectively make the modified state of a phosphorylation site permanent. It should be briefly noted at this point that phosphomimics such as glutamate and aspartate are not always functionally equivalent to phosphorylation, as illustrated by a study of the different aggregation kinetics of α -synuclein mutants, which aggregate as Lewy bodies in Parkinson's disease (Paleologou et al., 2008). Nonetheless, successful PTM mimicry is not limited to phosphorylation. Mutations to glutamine can reconstitute the effects of lysine acetylation (Yang et al., 2010; Liu et al., 2011; Nagaraj et al., 2012; Yu et al., 2012), and leucine and methionine can be used to successfully mimic lysine methylation (Hyland et al., 2011). The importance of PTM mimicry is further illustrated by the fact that PTMs often appear to modulate protein-protein interactions by improving or disrupting hydrophilicity matching, which highlights the importance of charges and electrostatic interactions (Hlevnjak et al., 2010).

We observed a number of interesting overrepresented substitutions at PTM sites (Fig. 2.7). Depending on the scoring method used, substitutions by glutamine (Q)

and asparagine (N) were overrepresented at structured lysine acetylation sites. We hypothesise that either their carbonyl oxygens or their primary amines may at least partially be able to mimic the steric and electrostatic interactions made by acetyllysine at structured sites (Figure 1.1). We also observed overrepresentation of glutamate (E) and aspartate (D) at disordered serine phosphorylation sites (in line with earlier reports), of aspartate (D) at tyrosine phosphorylation sites, and of glutamine (Q) at disordered N-glycosylated asparagines. Substitution of glycosylatable asparagine by glutamine, which only differs in carbon chain length, may be an efficient way of abolishing N-glycosylation while minimally affecting structure otherwise, thereby representing an “avoiding” type of PTM site mutation. At disordered lysine acetylation sites, an overrepresentation of substitutions by glutamate (E) was observed using two scoring methods, and this is discussed in more detail below.

Negatively charged succinylation or malonylation modifications might potentially occur as alternatives at lysine acetylation sites

The enrichment of negative charges at lysine acetylation sites, though inconsistently reported, was a surprising finding to us since lysine itself is positively charged, and acetylated lysine is neutral. In at least one instance, however, a mutation of a lysine acetylation site to glutamate (E) has been reported to mimic the effect of acetylation (Erkman and Kaufman, 2009). In this case, it has been suggested that the key factor is the removal of the positive charge of lysine by any means.

An alternative hypothetical explanation for the enrichment could be the presence of competing lysine succinylation or malonylation at lysine acetylation sites, at a frequency extensive enough to significantly affect substitution patterns. Both succinyl- and malonyllysine are negatively charged, and like acetylation, these modifications use a CoA-containing metabolic intermediate as the donor molecule (Tan et al., 2011; Lin et al., 2012). Succinyl-CoA is an intermediate of the citric acid cycle, while malonyl-CoA is formed from acetyl-CoA by acetyl-CoA carboxylase and is central in the synthesis of long-chain fatty acids (Ellis and Wolfgang, 2012)(Lin et al., 2012). Lysine succinylation and malonylation can be specifically reversed by erasing enzymes (Zhang et al., 2010c; Du et al., 2011). While sirtuins tend to act as protein deacetylases, Sirt5 was found to be a lysine desuccinylase and demalonylase, and the reactions involved are mechanistically equivalent to those

involved in lysine deacetylation by class I sirtuins (Du et al., 2011; Peng et al., 2011). In this light, it is interesting that Sirt5 can also still act as a very weak deacetylase, further highlighting the similarity of the modification reactions by showing cross-reactivity (Olsen, 2012). Sirt5 is found in mitochondria as well as the cytoplasm (Olsen, 2012; He et al., 2012a), although nuclear sites of lysine succinylation and malonylation have also been found on human, mouse, *Drosophila* and yeast histones (Xie et al., 2012). It has been suggested that lysine succinylation may be very abundant in human cells (Du et al., 2011). A useful way to follow up on these findings would be proteomic studies using specific antibody-based enrichment for succinylated or malonylated lysines.

Serine and threonine phosphomimicry may require multiple mutational steps

One prerequisite for mimetic mutations at a functionally important PTM site is that the mimetic residue should be reached in one mutational step. This is not the case for serine and threonine phosphorylation, which indicates that selection pressure may need to be relaxed through gene duplication (Diss et al., 2012). When not under constraint due to phosphorylation, serine has been described as a highly mutationally active residue (Creixell et al., 2012). Interestingly, multinucleotide substitutions have been reported to occur more frequently than would be expected from the random coincidence of individual neutral nucleotide substitutions (Averof et al., 2000; Hodgkinson and Eyre-Walker, 2010; Schrider et al., 2011). Though certain types of multinucleotide substitutions are more likely to occur than others, for instance due to environmental influences such as ultraviolet light leading to dipyrimidine lesions, it can be concluded in principle that any amino acid could be substituted by any other without an intervening fitness gap. Possible mechanisms include homologous recombination and gene conversion, especially over long evolutionary distances (Miyazawa, 2011b), as well as others (Hodgkinson and Eyre-Walker, 2010). Within the primate lineage, multinucleotide substitutions have been reported to be rare (Smith et al., 2003), though in general, allowing them in substitution models leads to a maximum likelihood improvement (Whelan and Goldman, 2004; Miyazawa, 2011a). A more recent study has reported a general enrichment in eukaryotes, including humans (Schrider et al., 2011). Nonetheless, the difficulty of reaching negatively charged phosphomimics through point mutations of serine and threonine phosphosites indicates that these mutations may be limited to

rare duplication events at the root of a lineage. The large number of serine phosphorylation sites (Table 2.1) may have aided in the identification of significant overrepresentation of negative charges at disordered positions in ours and other studies, despite the complex mutational steps required (Fig. 2.7).

The other types of PTM mimicry we observed can proceed through single point mutations

In this study, findings were also made with regard to the feasibility of mimicry for other PTM types. The enrichment of potential aspartate (D) phosphomimics which we observed at structured tyrosine (Y) phosphorylation sites (Fig. 2.7A) can be reached by a single point mutation (UAU to GAU or UAC to GAC), as opposed to glutamate (E), which would require two base changes and was not significantly overrepresented (Watanabe and Yokobori, 2011). It is intriguing to speculate that part of the depletion of tyrosines in higher eukaryotes (Tan et al., 2009b) may be due to phosphomimetic mutations. Additionally, the mutation of disordered lysine acetylation sites (K) to glutamate (E) and of structured lysines to glutamine (Q) both require only one point mutation. This indicates that these hypothetically mimetic mutations may occur first in evolution, and could then be followed by an additional single point mutation leading to transition to either aspartate (D) or asparagine (N), respectively. Similarly, the potential PTM-avoiding mutation of N-glycosylated asparagine (N) to glutamine (Q) can occur through a single point mutation.

2.4.4. Constraints on PTM sites at the population and somatic levels

Structured sites appear under clear selection pressure within populations and under functional constraint in individuals

Our results on PTM site conservation using data on natural variation and on disease mutations indicated that structured PTM sites are in fact also under strong constraint (Fig. 2.8–9), despite the observed absence of significant conservation when using a conservation scoring approach across species (Fig. 2.3–4). However, it should be noted that both the natural variant nsSNPs and the mutations associated with genetic disease from HGMD used here are limited to mutations which are at least non-lethal in early development. This is opposed to the situation across

species, where selection pressure even on essential positions may be relieved by strong functional divergence potentially aided by prior gene duplication. Our interpretation therefore is that these nsSNP datasets will be limited to mutations at structured sites which do not majorly interfere with protein folding, as potentially opposed to the conservation scores. Compared to these control sites, PTM sites then appeared significantly conserved. This indicates that it may be crucial in the analysis of structured PTM site conservation to distinguish between peripheral and core residues, which should be feasible for the majority of the human proteins investigated here through the use of homology modelling (Chothia, 2003; Eswar et al., 2007; Krieger et al., 2009; Braberg et al., 2012).

Ubiquitination and protein half-life may be affected by rare polymorphisms

Additionally, although the low-frequency natural variation nsSNPs identified by the 1000 Genomes Project exclusively through exome sequencing were discarded from this analysis due to their rarity, an investigation of these variants uncovered one significant enrichment of nsSNPs at lysine ubiquitination sites. Aside from this, the rare variants showed no significant preferences with regard to PTMs. It is tempting to speculate that these polymorphisms may modulate protein half-life by preventing polyubiquitination at certain sites, or modulate other types of ubiquitin signalling.

2.4.5. Summary of key points and perspective

In our analyses of the conservation of PTM sites compared to structurally similar background residues, we consistently found mild, but significant evidence of increased conservation of at least a fraction of PTM sites. Some sites were highly conserved, on par with catalytic residues (Fig. 2.5). The distributions of conservation scores at PTM sites, however, also included many residues which evolved rapidly. In addition to non-functionality, rapid divergence and lineage specificity, another explanation for the observed low conservation of a notable fraction of PTM sites may be the occurrence of specific types of mutations, which leave open the possibility of the site being functionally important in a given species.

Mimetic mutations, which resemble the modified amino acid and essentially fix the modified state, and PTM-avoiding mutations, which resemble its unmodified state and thereby disable the modification, may both decrease the apparent conservation

of a functionally important PTM site. Alternatively, a PTM site may arise at a residue which was previously already functional, and allow its function to be regulated (Pearlman et al., 2011). This is partially, but not completely alleviated by using a conservation scoring method which goes beyond determining the simple presence or absence of a modifiable residue, as has been done in this study. Ideally, a conservation scoring method that strongly rewards structural and biochemical similarity between residues should be developed for PTM sites.

Taken together, our results indicate that independently of structural characteristics, human PTM sites are mildly, but significantly more conserved overall than control residues at multiple evolutionary timescales for the six major PTM types studied.

3. The human lysine acetylation system

3.1. Introduction

Lysine acetylation is a common post-translational modification conserved from prokaryotes to humans (van Noort et al., 2012), and affects several thousand proteins in humans (Table 2.1). While the elucidation of its functions has not yet progressed as far as that of phosphorylation signalling, it has generated a large amount of interest due to its implication in transcriptional regulation and epigenetics, as well as in many other cellular processes, including metabolic regulation (Choudhary et al., 2009; Wang et al., 2010; Zhao et al., 2010). In this chapter, we applied a systematic data integration approach in order to identify lysine acetylation signalling proteins with interesting functional properties. In addition to presenting a concise overview of the human lysine acetylation signalling system using multiple types of data, we envision that this integrative approach may serve to highlight potentially under-appreciated candidates for detailed functional investigation.

3.1.1. Mechanisms and functional domains

Post-translational signalling systems are commonly described as being composed of “writer”, “reader” and “eraser” proteins (Seet et al., 2006; Lim and Pawson, 2010). Lysine acetylation is introduced by acetyltransferases, recognised by bromodomain proteins, and removed by deacetylases, which are described here in detail.

Acetyltransferases (Writers)

Acetylation of lysines is catalysed by lysine acetyltransferases (KATs), which were previously also termed histone acetyltransferases (HATs). These enzymes use acetyl-CoA as an acetyl group donor, which is a metabolic intermediate connecting glycolysis to the citric acid cycle. Several domains with acetyltransferase activity have been described: in terms of Pfam domain families, these are MOZ/SAS (PF01853), KAT11 (PF08214) and GNAT (PF00583) (Punta et al., 2012). Mechanistically, all of these domains can catalyse the transfer of the acetyl group of acetyl-CoA to the ϵ -amino group of lysine by a sequential catalytic mechanism of ϵ -amino group deprotonation followed by nucleophilic attack on the cofactor (Mischerikow and Heck, 2011). In addition to the internal acetylation of proteins on

lysine side chains, several enzymes in these families alternatively acetylate a protein's N-terminal α -amino group, which can affect protein half-life (Hwang et al., 2010; Soppa, 2010; Wildes and Wells, 2010).

Bromodomain proteins (Readers)

The canonical reader domain for lysine acetylation is the bromodomain (Pfam: PF00439) (Sanchez and Zhou, 2009; Zhang et al., 2010b; Punta et al., 2012), which appears to be limited to nuclear proteins (Zeng and Zhou, 2002; Mujtaba et al., 2007; Norris et al., 2009), though not all bromodomain proteins display binding to the core histones (Zhang et al., 2010b). Individual bromodomain proteins may display different sequence specificities based on the residues neighbouring the acetylated lysine (Hudson et al., 2000). This has been more recently investigated in detail for 14 bromodomain-containing proteins from *S. cerevisiae* (Zhang et al., 2010b). In certain proteins it occurs as two tandem bromodomains, resulting in the potential for combinatorial recognition of multiple acetylated lysines (Nakamura et al., 2007). Whether equivalently specialised reader domains exist in other compartments, as in the cytoplasm or mitochondria, remains to be determined (Norris et al., 2009). Alternatively, lysine acetylation might function primarily through conformational and electrostatic changes in its substrates outside the nucleus. Another type of domain, the PHD zinc fingers, has been found to very rarely bind acetylated lysine, for instance at H3K14 (Suganuma and Workman, 2011). However, as this domain is much more commonly involved in binding methylated lysine (Mellor, 2006), it was not included as a reader domain here.

Deacetylases (Erasers)

Reversal of lysine acetylation is catalysed by lysine deacetylases (KDACs), which are frequently also termed histone deacetylases (HDACs). At least two subgroups of these enzymes exist, though additional subdivisions have been described (Mischerikow and Heck, 2011). According to Pfam, these two sub-groups are the histone deacetylases (Pfam: PF00850) and sirtuins (Pfam: PF02146) (Punta et al., 2012). Histone deacetylases are Zn^{2+} -dependent enzymes which hydrolyse the acetyllysine amide bond, yielding acetate and the deacetylated lysine-containing substrate protein. Sirtuins, named for their homology to the yeast protein Sir2, are NAD^+ -dependent enzymes. Sirtuins couple protein deacetylation to NAD^+ hydrolysis,

yielding acetyl-ADP-ribose, nicotinamide and the deacetylated lysine-containing substrate. Alternatively, sirtuins have been reported to transfer the ADP-ribosyl moiety of NAD⁺ to the substrate, yielding a mono-ADP-ribosylated substrate protein and nicotinamide (Mischerikow and Heck, 2011).

3.1.2. Biological functions

Chromatin regulation

Lysine acetylation was first identified as a transcriptionally activating post-translational modification of histones (Ogryzko et al., 1996; Spange et al., 2009). A simplified yet relevant model of this activity is that it can neutralise positive charges on histones, leading to a decrease in affinity for the negatively charged phosphate backbone of DNA, thereby mediating an open chromatin conformation. For instance, H4K16 acetylation acts according to the above model by decreasing the affinity between the disordered histone H4 N-terminus and the linker DNA between two nucleosomes, thereby decompacting chromatin fibres (Shahbazian and Grunstein, 2007). Similarly, H3K9 acetylation is observed in transcriptionally active regions. However, the function of lysine acetylation on histones also depend on the specific residue which is acetylated. H3K4 acetylation acts as a repressive signal by decreasing the affinity of certain activating chromodomain-containing proteins for methylated H3K9, thereby leading to heterochromatin formation and transcriptional repression (Xhemalce and Kouzarides, 2010). Other lysines such as acetylated H3K14 may act as a signal bound by reader proteins, while H3K56 acetylation appears to act as an identification mark for newly synthesised histones during S-phase (Suganuma and Workman, 2011).

Connection with metabolism

In addition to many other cellular processes, especially those involving large protein complexes (Choudhary et al., 2009), lysine acetylation was shown to regulate metabolic enzymes in both humans and prokaryotes (Wang et al., 2010; Zhao et al., 2010). This is interesting since the cofactor used in acetylation, acetyl-CoA, is a common metabolic intermediate. It has been found in both humans and in *Salmonella* that many metabolic enzymes seem to be regulated by acetylation

(Wang et al., 2010; Zhao et al., 2010; Lin et al., 2012). The effects of lysine acetylation on the functions of these enzymes are diverse, and have been described as leading to either their activation, inactivation or destabilisation (Yu et al., 2008; Lammers et al., 2010; Zhao et al., 2010; Jiang et al., 2011; Lin et al., 2012). This illustrates a broad range of effects for acetylation, similar to that of phosphorylation. As an additional example, the half-life of the RNase R exoribonuclease in *E. coli* is increased by enzymatic acetylation at one lysine residue, thereby increasing its concentration 3- to 10-fold under stress conditions (Liang et al., 2011).

In *Salmonella*, a positive correlation has been observed between the activity of glucose metabolism and the acetylation of various enzymes. Here, acetylation of certain enzymes has been described as favouring the metabolism of glucose by increasing enzyme activity in glycolysis and the citric acid cycle. In contrast, a decrease in acetylation of certain enzymes was described to favour the glyoxylate pathway and gluconeogenesis (Lin et al., 2012). A correlation between deacetylation state of certain metabolic enzymes and the concentration of acetyl-CoA has also been reported (Wang et al., 2010; Cai et al., 2011). Caloric restriction is associated with deacetylation of mitochondrial proteins (Schwer et al., 2009; Bao and Sack, 2010), which appears to be catalysed by mitochondrial sirtuins, and redox stress also affects lysine acetylation (Bao and Sack, 2010). These connections suggest that lysine acetylation signalling may be relevant to the study of human diseases including diabetes, obesity and cancer. Concordantly, histone deacetylase inhibitors have been developed for use as anti-cancer and anti-inflammatory treatments (Bantscheff et al., 2011).

Additional functions

Since lysine acetylation removes a positive charge, its effects could potentially be similar to those of phosphorylation, where negative charges are introduced (Kouzarides, 2000; Norris et al., 2009). Similarly to phosphorylation, lysine acetylation has been found to exert a wide range of effects. Effects on enzymatic activity, protein-protein interactions, DNA binding, protein stability and subcellular localisation have been documented (Glozak et al., 2005; Scroggins et al., 2007; Sun et al., 2007; Sadoul et al., 2008; Fantini et al., 2010; Liarzi et al., 2010; Xiong et al., 2010). As in the charge-based mechanism described above for chromatin

decompaction, lysine acetylation can also decrease the DNA affinity of transcription factors (Aran-Guiu et al., 2010).

Proteins can be acetylated at the α -amino group of the N-terminal amino acid, or at the ϵ -amino group of internal lysines. N-terminal acetylation seems to occur co-translationally, and is very common in eukaryotes (Wildes and Wells, 2010) and certain archaea, but extremely rare in bacteria (Soppa, 2010). N-terminal acetylation is thought to be a factor in determining protein half-life (Hwang et al., 2010). In biotechnology, acetylation has been used to create highly negatively charged enzymes that are less prone to aggregation and display increased stability (Shaw et al., 2008). Overall protein charge has also been implicated as a determinant of protein aggregation in neurodegenerative diseases (Shaw et al., 2010).

Lysine acetylation has also been implicated in neurological roles, including neurodegenerative disease. Inhibiting the deacetylase Sirt2 has been reported to counteract neurodegeneration in a model of Parkinson's disease, and acetylation has similarly been reported to target mutant huntingtin for degradation in Huntington's disease (Outeiro et al., 2007; Jeong et al., 2009). It has also been reported that neurotransmission affects lysine acetylation of a number of proteins, as well as of histones in distinct brain regions (Sen and Snyder, 2010).

Lysine acetylation is additionally important in the endogenous regulation of microtubule assembly, which an acetyltransferase of the plant pathogen *Pseudomonas syringae* has been shown to disrupt (Lee et al., 2012). Other pathogens also use acetyltransferases as virulence factors (Mittal et al., 2006; Mukherjee et al., 2006; Tasset et al., 2010).

3.1.3. Motivation and analysis outline

Due to the central importance of post-translational signalling systems such as lysine acetylation in cellular homeostasis, we were prompted to develop and refine a general framework for the study of PTM signalling systems by data integration. This method is applied here to lysine acetylation, a highly common PTM which displays a wide range of functions in higher eukaryotes as well as prokaryotes. Proteins with functions in lysine acetylation were identified, and were then systematically analysed based on their conservation, expression profile, protein-protein interactions and subcellular localisation. By integrating these types of information, the most

promising candidate proteins participating in lysine acetylation signalling were systematically highlighted and ranked by priority for targeted followup studies.

3.2. Methods

3.2.1. Identification of proteins with functions in lysine acetylation

Human proteins containing one or more lysine acetylation writer, reader and eraser Pfam domains were retrieved from UniProtKB/Swiss-Prot (release 2012_01), which provides Pfam domain annotation for its proteins (Punta et al., 2012; UniProt Consortium, 2012). For writers, domains used were: “MOZ/SAS” (PF01853), “KAT11 family” (PF08214) and “Acetyltransferase (GNAT) family” (PF00583). For readers, the domain was the “Bromodomain” (PF00439). For erasers, these were “Hist_deacetyl” (PF00850) and “Sir2 family” (PF02146). Since it has been noted that some acetyltransferases are quite divergent and cannot be easily grouped by sequence homology (Mischerikow and Heck, 2011), we additionally performed a manual search of UniProtKB/Swiss-Prot for proteins with annotation relating to acetylation. In this search, we retrieved additional confirmed acetyltransferases, as well as proteins which were annotated to contain bromodomains, but not according to Pfam. This may potentially be due to a relatively strict, limited definition of the Pfam bromodomain Hidden Markov Model (HMM). We therefore additionally included all proteins with a match to the bromodomain superfamily definition (SCOP ID 47370) in the SUPERFAMILY database (Gough et al., 2001; Wilson et al., 2007; 2009). This resulted in the addition of three additional reader proteins with significant matches to SUPERFAMILY’s definition of the bromodomain family.

3.2.2. Analysis of evolutionary conservation

Two complementary analyses of evolutionary conservation of the lysine acetylation-related candidate proteins were performed, at different evolutionary distances. For an analysis largely focused on vertebrates and shorter evolutionary distances, aligned protein homology clusters were obtained from Ensembl Compara (Release 65, December 2011) (Flicek et al., 2012). In order to retrieve the best-matching ortholog for each human protein, the global pairwise sequence identity percentage of the best-matching pair of sequences corresponding to the human Ensembl gene

identifiers (as mapped to by the original UniProtKB/Swiss-Prot identifiers) was used to generate the protein conservation profile figure. A complementary analysis of more distantly related species was performed analogously using ortholog pairs from InParanoid 7 (Ostlund et al., 2010). Since InParanoid does not provide alignments, sequence identities of pairs were calculated using global pairwise Needleman-Wunsch alignments. For comparison of the average conservation of the lysine acetylation-related proteins to other groups of interest, a list of human transcription factors was obtained (Vaquerizas et al., 2009), and human essential genes were obtained from version 6.8 of the DEG database (Zhang and Lin, 2009). Missing orthologs were considered to have a sequence identity of zero when calculating average conservation. The matrices were visualised using Genesis 1.7.6 (Sturn et al., 2002), and complete-linkage hierarchical clustering was used to cluster the genes in the conservation profile figures.

3.2.3. Subcellular localisation

Subcellular localisation information was obtained from UniProtKB/Swiss-Prot 2012_01, and terms were manually simplified to “Nuclear” or “Cytoplasmic”. Proteins which were annotated as both were termed “Shuttling” proteins, since they are presumably able to change their localisation between the nucleus and cytoplasm. Proteins without subcellular localisation annotation in UniProt were labelled “Unknown”.

3.2.4. Protein-protein interaction network

Known protein-protein interactions were obtained from STRING 9.0.5 (Szklarczyk et al., 2011). Only experimental and “knowledge-based” interactions from external databases were used, with the default “medium confidence” of 0.4 as the threshold. Cytoscape 2.8.3 was used to generate the network figure using an edge-weighted force-directed layout (“BioLayout”), weighted using STRING’s combined confidence score (Smoot et al., 2011). Statistical assessment of the number of proteins in a given category which interact with at least one lysine acetylation substrate, compared to the proteome-wide figure, was done using Fisher’s exact test.

3.2.5. Expression profiles and co-expression network

Human expression data was obtained from the BioGPS “U133A/GNF1H Gene Atlas” (Su et al., 2004) and log₂-transformed to achieve a more Gaussian-like value distribution, which is required for assessing correlations using the Pearson product-moment correlation coefficient (Ballman, 2008). Up-to-date mappings of probe sets to transcripts were obtained for the AffyMetrix U133A chip from Ensembl (release 65, December 2011) (Flicek et al., 2012). Since recent probe set mappings were not available for the custom GNF1H chip, its expression values were not used. We chose this expression dataset due to its high reported reproducibility and linear relationship between abundance at the transcript and protein levels (Lage et al., 2008). Six of the 84 samples were from human cancer cell lines, rather than healthy human tissues. While these are shown in the expression matrix figure, only healthy tissues were considered when quantifying co-expression for the co-expression network. The Pearson product-moment correlation coefficient was used for detecting significant co-expression via Student’s t-distribution, with Bonferroni correction for multiple testing. Cytoscape 2.8.3 was used to generate the network figure using an edge-weighted force-directed layout (“BioLayout”), weighted using the correlation coefficients, with minor manual adjustments to reduce overlap (Smoot et al., 2011). We also quantified the expression breadth of genes as 1- τ , where τ is the tissue specificity index

$$\tau = \frac{\sum_{i=1}^N (1 - x_i)}{N - 1}$$

and where N is the number of tissues and x_i is the expression profile component normalised by the maximal component value (Yanai et al., 2005).

3.2.6. Candidate ranking

The candidate score is a rank-based measure which takes into account a protein’s average conservation according to the Compara-based approach, its number of protein-protein interactions, number of co-expressed genes and its expression breadth across tissues, all with equal weight. It was calculated as follows: for each gene in each column, the fraction of lower- or equal-ranking genes in a given column was calculated, and the final score is composed of the average of these column

scores. Genes which had no information available in a column were not considered for its scoring. The final score thus ranges between 0 (for a protein ranking lowest or unranked in all categories) and 1 (for a protein ranking highest in all categories). When generating the ranked summary table, information on genetic disease mutations from HGMD (Stenson et al., 2009) and OMIM (Amberger et al., 2009) as well as cancer-causing mutations from the Cancer Gene Census (Futreal et al., 2004) were also integrated as an additional metric of the functional importance of genes. Additionally, significant up- or down-regulation of the mouse orthologs (identified using Ensembl Compara) in the differentiation of embryonic stem cells as determined using expression data from StemBase (Sandie et al., 2009) is shown. This methodology is described in detail in Chapter 4, section 4.2.3.

3.3. Results

3.3.1. Identification of 100 lysine acetylation-related candidate proteins

By searching for functional Pfam domains as well as manually for proteins either confirmed or likely to be functionally involved in lysine acetylation (as either writer, reader or eraser proteins), we were able to identify a set of 100 human candidate proteins (Fig. 3.1). Notably, four proteins fell into both the writer and reader categories, indicating that they might be able to integrate the modification state of an adjacent lysine in the substrate targeting of their acetyltransferase activity. This writer-reader overlap and the size of the functional categories is reiterated as a Venn diagram in Figure 3.2. Acetyltransferases and bromodomain reader proteins made up the largest groups, consisting of 43 proteins each, with an overlap of 4 proteins containing both types of domain. The smallest group were the deacetylases, consisting of 18 proteins and showing no overlaps.

Writers	Readers	Erasers
acetyltransferases	bromodomain proteins	deacetylases

Protein domains

Acetyltransf_1 / GNAT	Bromodomain	Hist_deacetyl
KAT11 family		Sir2 family (sirtuins)
MOZ/SAS family		

Human genes

AANAT, CDY1, CDY2A, CDYL, CLOCK, CSRP2BP, ELP3, ESCO1, ESCO2, GNP NAT1, GTF3C4, HAT1, KAT5, KAT6A, KAT6B, KAT7, KAT8, MGEA5, NAA10, NAA11, NAA20, NAA30, NAA40, NAA50, NAA60, NAT6, NAT8, NAT8B, NAT8L, NAT9, NAT10, NAT14, NAT16, NCOA1, NCOA2, NCOA3, SAT1, SAT2, SATL1 CREBBP, EP300, KAT2A, KAT2B	ASH1L, ATAD2, ATAD2B, BAZ1A, BAZ1B, BAZ2A, BAZ2B, BPTF, BRD1, BRD2, BRD3, BRD4, BRD7, BRD8, BRD9, BRDT, BRPF1, BRPF3, BRWD1, BRWD3, CECR2, KIAA2026, MLL, PBRM1, PHIP, SMARCA2, SMARCA4, SP100, SP110, SP140, SP140L, TAF1, TAF1L, TRIM24, TRIM28, TRIM33, TRIM66, ZMYND8, ZMYND11 CREBBP, EP300, KAT2A, KAT2B	HDAC1, HDAC2, HDAC3, HDAC4, HDAC5, HDAC6, HDAC7, HDAC8, HDAC9, HDAC10, HDAC11, SIRT1, SIRT2, SIRT3, SIRT4, SIRT5, SIRT6, SIRT7
--	---	--

Figure 3.1: Functional classes, Pfam domains and gene symbols for putative components of the human lysine acetylation system. Four genes fall into both the writer and reader categories.

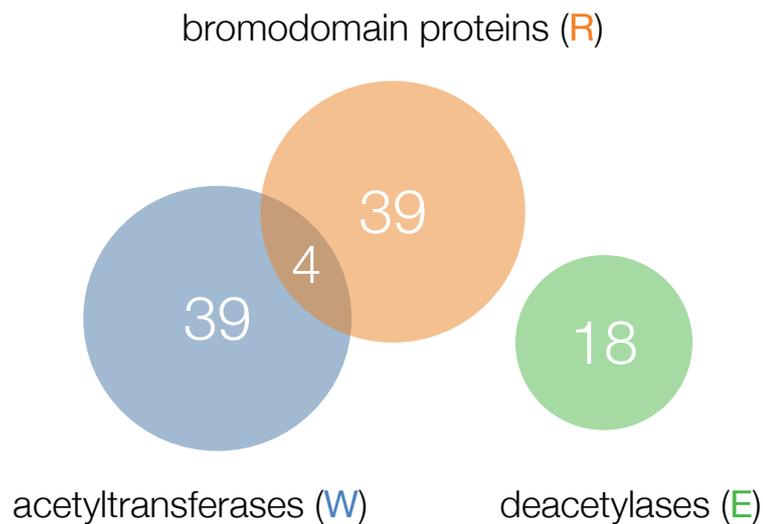


Figure 3.2: Venn diagram showing the number of human proteins for lysine acetylation-related functional classes.

3.3.2. Nearly all candidate proteins are highly conserved in vertebrates

Figure 3.3 shows the conservation of the human lysine acetylation proteins, as determined using homology information from Ensembl Compara. A strikingly high degree of conservation is apparent, with the large majority of proteins having orthologs in all vertebrates. Interestingly, the potentially combinatorial writer-reader proteins (WR) were the best-conserved category overall, despite their potentially more complex architecture of at least two distinct domains. Erasers and readers showed similarly strong conservation to each other, with the notable exception of the bromodomain reader proteins KIAA2026, SP100, SP110, SP140 and SP140L. SP110 and SP140 are components of nuclear bodies, which are also referred to as nuclear dots or PML bodies (UniProt Consortium, 2012). Although orthologs of these proteins apparently exist in all mammals, they are highly divergent outside the primate lineage (Fig. 3.3). This suggests that lysine acetylation signalling mechanisms linked to nuclear bodies, which might use these five proteins as readers, may have evolved at the base of the primate lineage. Writers as a group displayed the most cases where orthologs were absent in species, often in patterns not constricted to a lineage, though many writers were also highly conserved even in non-vertebrates.

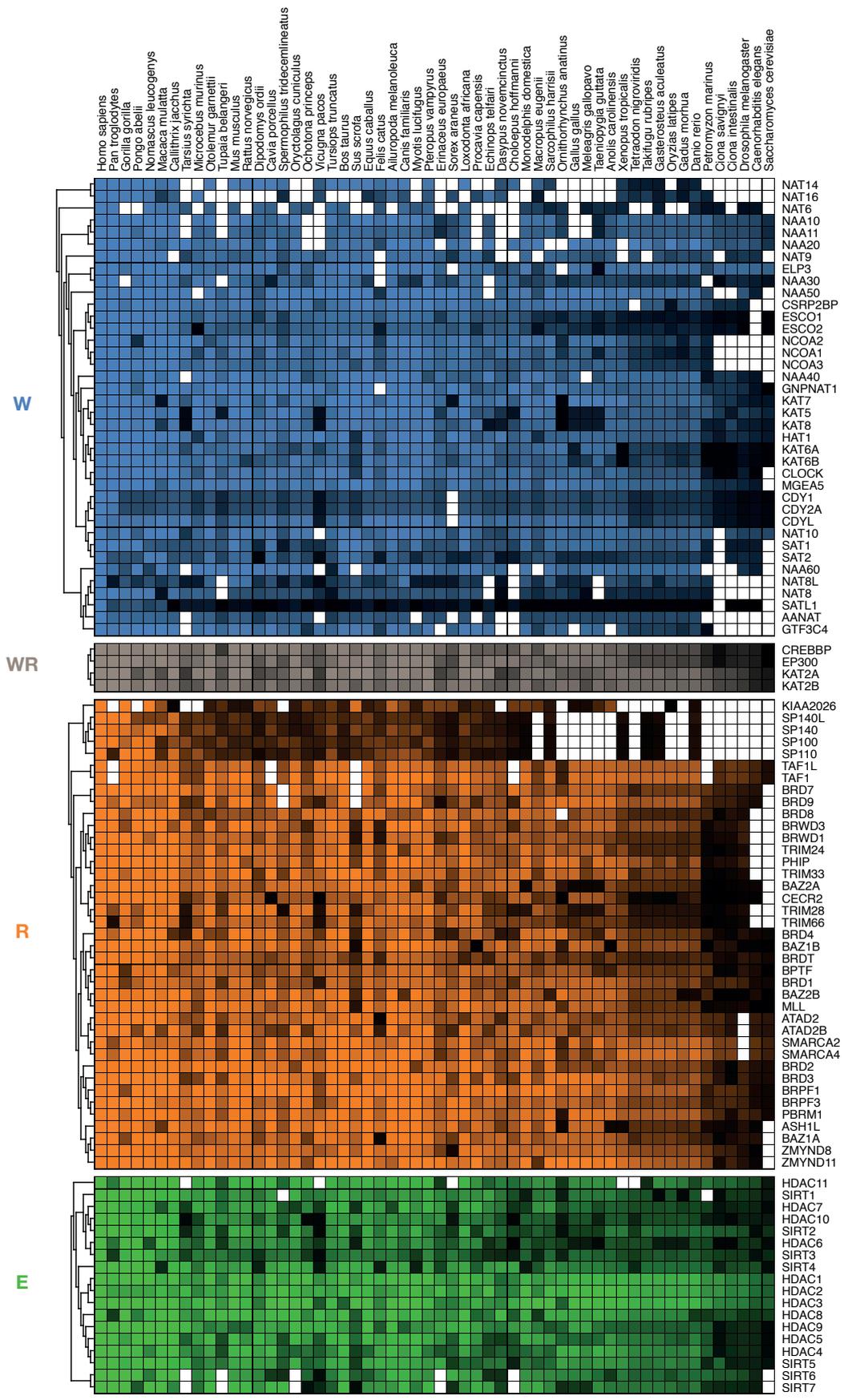


Figure 3.3: Conservation of human lysine acetylation proteins, based on Ensembl Compara. Columns represent species, while rows represent genes. Rows are coloured according to the gene's functional category, and shaded according to the sequence identity between the human protein and its best-matching ortholog in a species. Missing orthologs are shown in white.

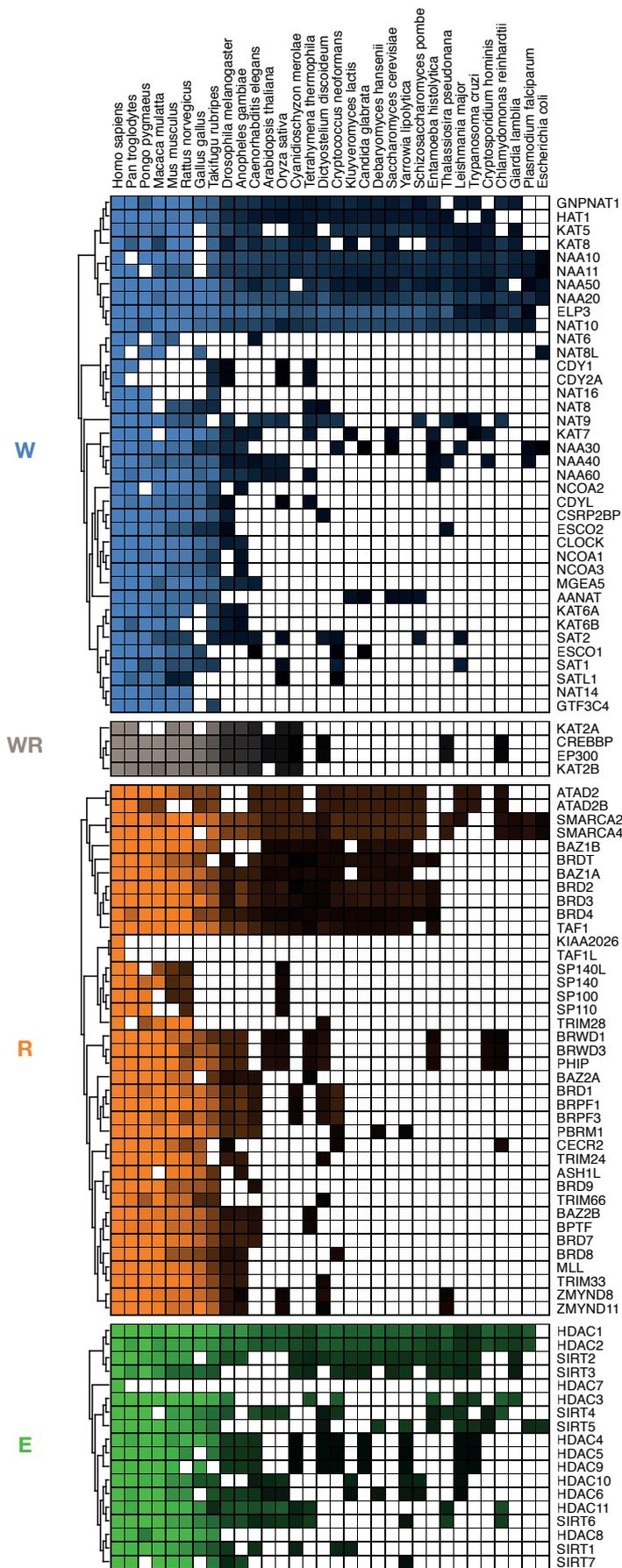


Figure 3.4: Conservation of human lysine acetylation enzymes, based on InParanoid. Fewer species were covered, but with longer evolutionary distances. For a complete description of the plot content please see Figure 3.3.

Across the species in Ensembl Compara, the lysine acetylation-related proteins were significantly better conserved than the human proteome-wide average ($p = 8.7^{-137}$, Mann-Whitney U test), at an average sequence identity of 63.4% compared to the human average of 51.2%. Their conservation is on par with that of essential genes at 64.0% (though lower, $p = 6.2^{-5}$, Mann-Whitney U test). This also makes them significantly more conserved than the average for human transcription factors of 53.3% ($p = 6.7^{-86}$, Mann-Whitney U test), which have been shown to display a relatively high degree of evolutionary plasticity in the human lineage (De et al., 2008; López-Bigas et al., 2008). In both transcription factors and in the lysine acetylation signalling proteins studied here, there appears to be substantial variation between highly conserved factors likely to be essential on the one hand, and less conserved, dynamically evolving proteins on the other, as evidenced by their respective standard deviations in sequence identity of 34.6% and 30.3%, as well as the gaps in their conservation profiles (Fig. 3.3, Fig. 3.4). As a category, the bromodomain proteins (readers) displayed the lowest conservation at an average of 61.2%, and high variation at a standard deviation of 30.5%, followed by the acetyltransferases (writers) with 63% average sequence identity.

A complementary approach based on a resource containing a much larger number of non-vertebrate species, InParanoid, is shown in Figure 3.4. Both here and in Figure 3.3, the conservation of the best putative ortholog is indicated based on global sequence identity, which means that large changes such as domain acquisitions, a common mechanism in the evolution of phosphorylation signalling systems and proteins in general (Chothia and Gough, 2009; Jin and Pawson, 2012), should be clearly visible as decreased sequence identity.

In the alternative InParanoid approach, the human lysine acetylation candidate proteins are also well-conserved within the vertebrates. In addition, since InParanoid covers a much wider phylogenetic range, divisions become apparent between those candidates which are conserved across very long evolutionary distances, and others which appear to have arisen more recently, most likely by expansion of protein families through gene duplication. The most evolutionary ancient proteins include the acetyltransferase and elongator complex component ELP3 as well as the histone deacetylases HDAC1 and HDAC2. These appeared to have orthologs in all eukaryotes, and they were only non-conserved in *E. coli* (the rightmost column).

In three cases (KIAA2026, HDAC7 and TAF1L), the InParanoid approach reported that no orthologs were present outside of humans, though Compara did find a number of orthologs comparable to other candidates. These were artefacts caused by identifier mapping problems to the relatively old InParanoid release, and they did not affect any of the other genes.

3.3.3. Many substrates are exclusively located in the cytoplasm

More than half of the writer and eraser candidates are observed in the cytoplasm at least some of the time (cytoplasmic or “shuttling”). Their substrates are most frequently observed exclusively in the cytoplasm (Figure 3.5). Conversely, bromodomain reader proteins appear absent from the cytoplasmic environment (unless shuttling, 15% of proteins), perhaps hinting at the existence of other specialised cytoplasmic reader domains yet to be identified in this part of the cell.

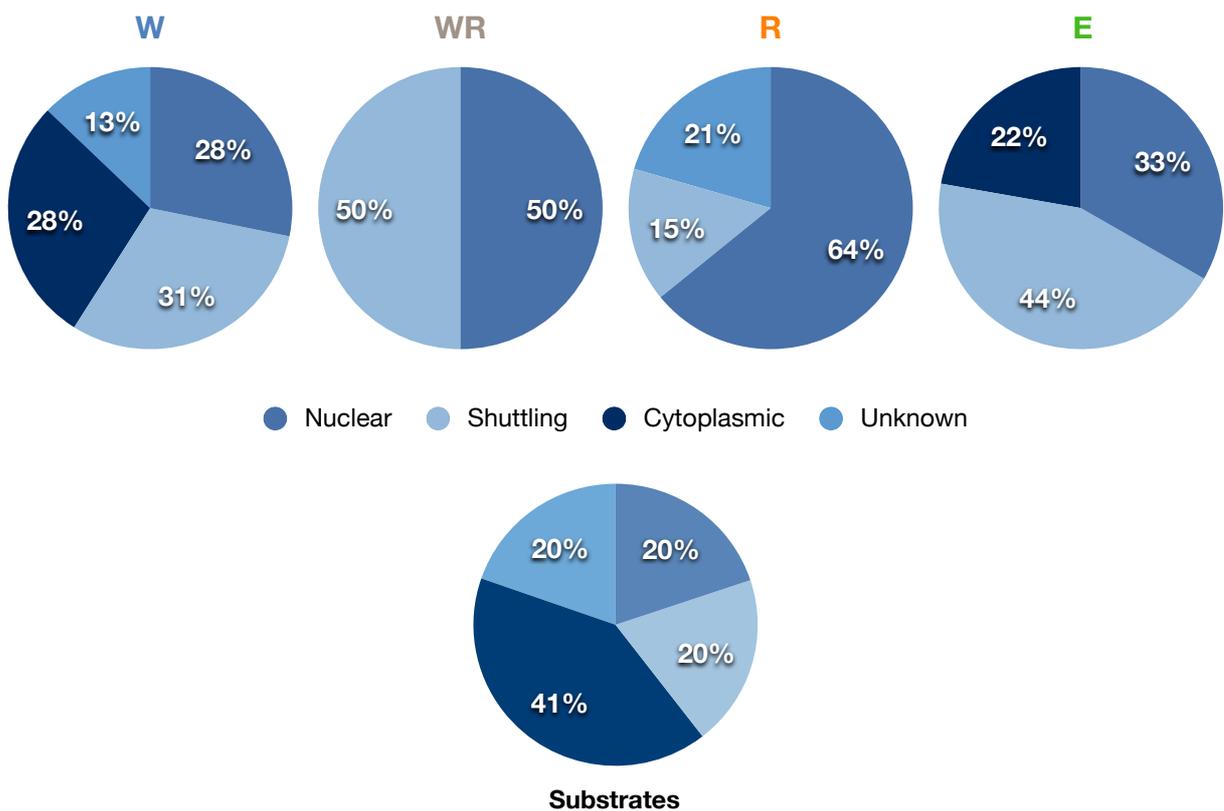


Figure 3.5: Subcellular localisation of lysine acetylation-related proteins and of lysine acetylation substrates.

3.3.4. Functional classes tend to share physical interactions

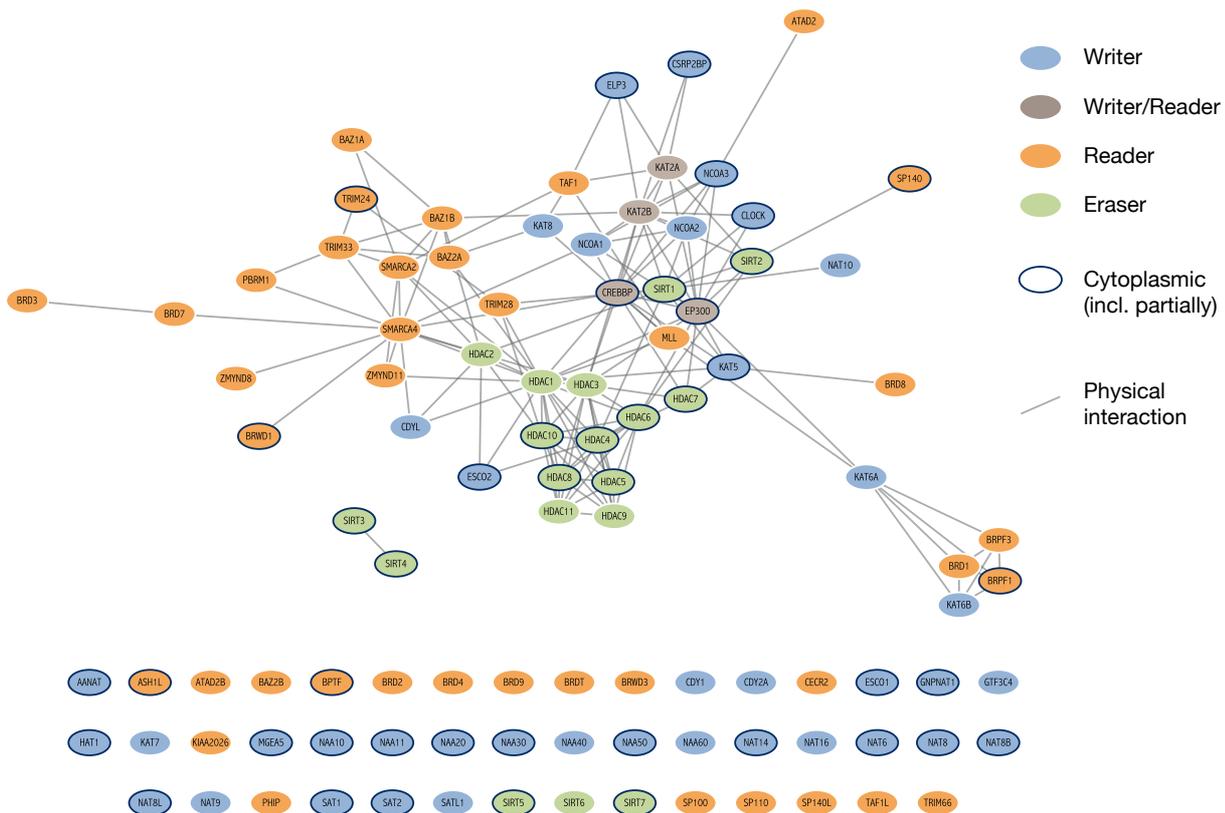


Figure 3.6: Protein-protein interaction network, showing known physical interactions among human lysine acetylation-related proteins. Node colours indicate functional category (blue: writers, brown: writer-readers, orange: readers, green: erasers). Dark blue node borders indicate either cytoplasmic or “shuttling” subcellular localisation.

Analysing the protein-protein interaction network of the lysine acetylation-related candidates revealed that functional classes (writers, readers and erasers) tend to cluster together in a force-directed layout (Fig. 3.6). This could either be due to a similar set of interaction partners, or due to mutual interactions. For histone deacetylase-family erasers (HDAC1 to HDAC11), both seem to be the case. Acetyltransferases seem to cluster together due to shared interactors, which may either be specificity-mediating adaptor proteins reminiscent of E3 ubiquitin ligases, or they might also be acetylation substrates. Readers tend to interact with one another, suggesting that they might form part of larger complexes which might be performing combinatorial readout of multiple acetylation sites. SMARCA4 interacted with the largest number of other reader proteins, and is one of the most highly conserved reader proteins (Fig. 3.3). Additionally, we found that around half of the candidate proteins were not known to physically interact with any other lysine acetylation-related proteins.

Table 3.1: Percentage of proteins in different functional categories which physically interact with lysine acetylation substrates. Fisher’s exact test was used to compare the number of proteins which interact with at least one acetylated substrate to that across the entire human proteome.

Category	Interacting with K-ac substrates	Not interacting with K-ac substrates	Percentage interacting	p-value
All K-ac-related proteins	66	34	66%	2E-06
W	23	16	59%	0.04
WR	4	0	100%	0.03
R	23	16	59%	0.04
E	16	2	88.9%	5.4E-05
K-ac substrates	2099	804	72.3%	9.8E-205
Other human proteins	7270	10019	42%	
<i>All Nuclear</i>	<i>1628</i>	<i>1222</i>	<i>57.1%</i>	
<i>All Shuttling</i>	<i>1381</i>	<i>496</i>	<i>73.6%</i>	
<i>All Cytoplasmic</i>	<i>4951</i>	<i>5470</i>	<i>47.5%</i>	
<i>All Unknown</i>	<i>1437</i>	<i>3657</i>	<i>28.2%</i>	

In an analysis of the known physical interactions of the different functional categories of lysine acetylation-related proteins with acetylation substrates, it was found that especially eraser proteins were highly likely to interact with acetylated proteins (Table 3.1). Nearly 89% of eraser proteins, as well as all of the combinatorial writer-reader proteins displayed at least one such substrate interaction. In total, 66% of all lysine acetylation-related candidate proteins studied here were known to interact with acetylation substrates, significantly higher than the proteome-wide percentage of 42% ($p = 2^{-6}$, Fisher’s exact test). Acetylation substrates also frequently made interactions with other acetylated proteins, at a percentage of 72.3% ($p = 9.8^{-205}$, Fisher’s exact test). These numbers may indicate strong functional links between different acetylation-related proteins and substrates, as well as among the acetylated substrates themselves.

In terms of subcellular localisation, “shuttling” proteins, which can be found both in the nucleus and the cytoplasm, were among the most frequent interactors with acetylation substrates, at 73.6%. Conversely, only 28.2% of proteins with unknown localisation interacted with acetylated proteins.

3.3.5. Many candidates show significantly similar expression profiles

Co-expression analysis can be used to identify functionally related genes (Mason et al., 2009; Nayak et al., 2009). This type of analysis further allows us to highlight functionally related proteins within the lysine acetylation system which are not necessarily physical interactors, but which may be active in concert in similar sets of tissues (Fig. 3.7). This figure also shows that the candidate proteins varied widely both in their average expression and in their expression between individual tissues. Some proteins were very highly expressed in some tissues, but were otherwise generally inactive, indicating specialised rather than universal roles. We quantified this aspect by calculating the expression breadth for each gene, which allowed us to highlight candidates which were expressed across a large number of tissues.

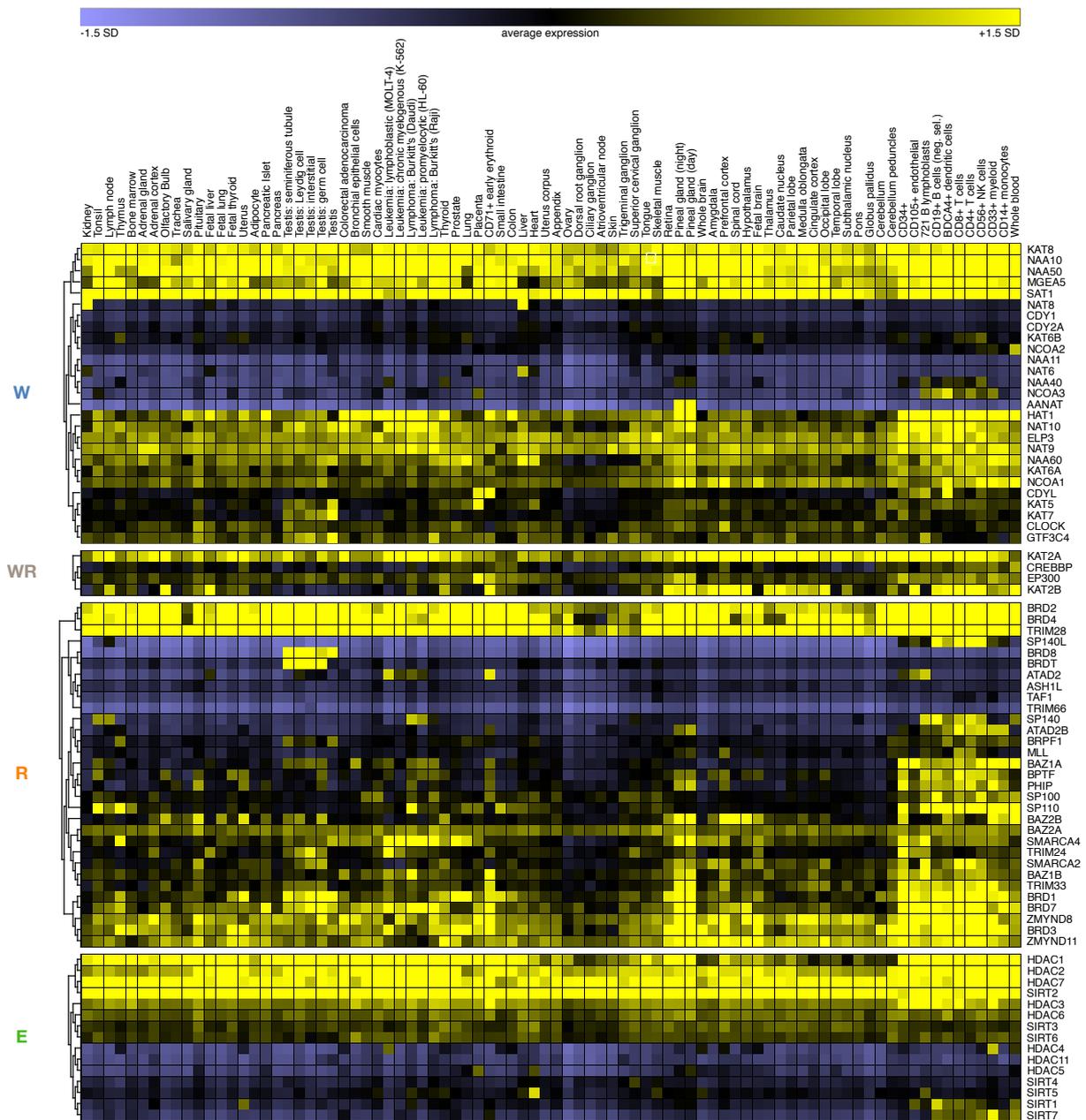


Figure 3.7: Expression profiles of lysine acetylation-related proteins in human tissues. Columns correspond to different tissues, while rows show individual genes. The colour gradient proceeds from blue for low expression (1.5 standard deviations below the chip-wide average) through black (chip-wide average expression) to yellow for high expression (up to 1.5 standard deviations above the chip-wide average). Complete-linkage hierarchical clustering was performed to arrive at the final layout.

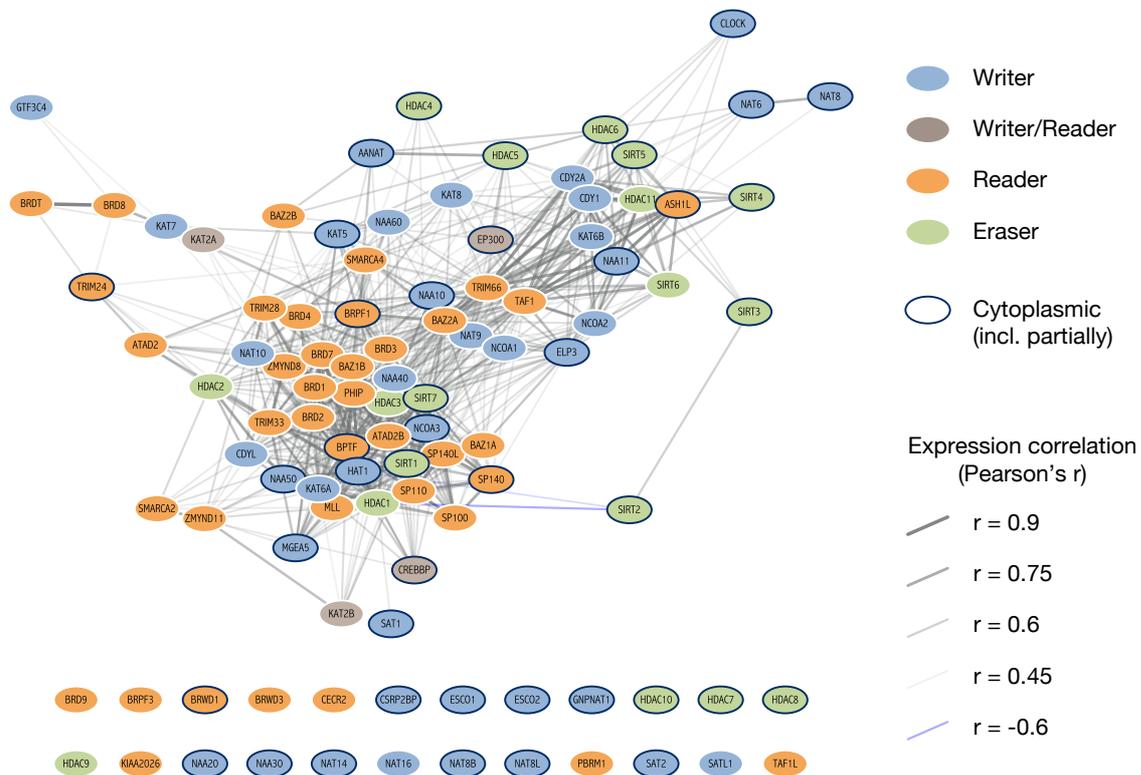


Figure 3.8: Co-expression network of lysine acetylation-related proteins in healthy human tissues. Genes are represented as nodes, while edges represent significant co-expression between two genes. The line thickness and shading of the edges represent the strength of co-expression, with blue representing inverse correlation.

Figure 3.8 shows significant expression correlations between the candidate proteins, including between candidates belonging to separate functional classes. Notably, this analysis revealed a larger number of links between proteins than their physical interaction network (Fig. 3.6), with fewer candidates appearing isolated. Here, candidates were frequently co-expressed across functional categories, indicating significant functional associations between e.g. acetyltransferases and readers. This makes mechanistic sense, since certain reader proteins may only need to be expressed if a corresponding writer enzyme is present. In a similar potential mechanistic connection, expression of the histone deacetylase SIRT2 was found to be significantly anti-correlated with the acetyltransferase HAT1 (as well as HDAC1). Both SIRT2 and HAT1 act in histone H4 acetylation, respectively deacetylating H4K16 and acetylating H4K5/H4K12 (Kouzarides, 2007), which lends support to the functional significance of their anti-correlated expression patterns.

3.3.6. Candidate ranking

In addition to presenting a concise overview of the human lysine acetylation signalling system using multiple types of data, one primary motivation of this chapter was to integrate this information in order to highlight the most promising candidate proteins participating in lysine acetylation signalling, and to systematically prioritise them for targeted followup studies. This is done here by presenting an informative summary table of information on all candidates, as well as by ranking them using a scoring function (Table 3.2). The scoring function takes into account a candidate protein's average conservation according to the Compara analysis, its number of protein-protein interactions, the number of co-expressed genes, and the expression breadth (fraction of tissues where a gene is expressed) of the candidate.

Table 3.2: Candidate list ranked by candidate score. This table lists human lysine acetylation-related proteins according to their apparent functional importance. The candidate score is a rank-based measure, which takes into account the four central columns (conservation according to the Compara analysis, protein-protein interactions, co-expression and expression breadth) with equal weight.

Gene	Class	Subcellular localisation	Disease implications	Conservation (Average sequence identity)	Known protein-protein interactions	Degree of co-expression	Expression breadth	Expression trend in ESC differentiation	Candidate score	Candidate rank
HDAC3	E	Nucleus		87%	18	41	0.66		0.87	1
HDAC1	E	Nucleus		87%	24	34	0.66		0.85	2
HDAC2	E	Nucleus		86%	19	15	0.73		0.76	3
BAZ2A	R	Nucleus		61%	6	46	0.87		0.73	4
BRPF1	R	Both		81%	4	35	0.69		0.73	5
KAT6A	W	Nucleus		68%	7	30	0.75		0.71	6
TAF1	R	Nucleus	HGMD/OMIM	69%	3	38	0.72		0.68	7
EP300	WR	Both	CGC/HGMD/OMIM	74%	19	12	0.70	Down	0.67	8
BAZ1B	R	Nucleus		67%	6	38	0.67		0.67	9
SMARCA4	R	Nucleus	CGC/HGMD/OMIM	74%	17	20	0.64		0.67	10
ELP3	W	Both		81%	1	25	0.78		0.66	11
HDAC11	E	Nucleus		68%	9	19	0.79		0.65	12
CREBBP	WR	Both	CGC/HGMD/OMIM	74%	18	11	0.66		0.64	13
TRIM33	R	Nucleus	CGC/OMIM	68%	7	29	0.65		0.63	14
NCOA1	W	Nucleus	CGC/HGMD	68%	7	24	0.68		0.63	15
BRD3	R	Nucleus	CGC	74%	2	34	0.67		0.63	16
KAT8	W	Nucleus		69%	5	15	0.82	Down	0.62	17
SMARCA2	R	Nucleus	HGMD/OMIM	76%	11	8	0.64	Up	0.60	18
BRD7	R	Nucleus		66%	4	37	0.66		0.60	19
HDAC6	E	Both		52%	11	16	0.86	Down	0.59	20
ZMYND11	R	Nucleus	HGMD	80%	3	11	0.74	Up	0.59	21
BRD4	R	Nucleus	CGC	60%	8	21	0.74		0.59	22
BRD2	R	Nucleus		73%	1	29	0.70		0.58	23
SIRT1	E	Both	HGMD	62%	14	33	0.53	Down	0.57	24
TRIM28	R	Nucleus		53%	7	22	0.73	Down	0.57	25
CLOCK	W	Both	HGMD	76%	6	5	0.71		0.56	26
NAT10	W	Nucleus		77%	1	29	0.61		0.56	27
NAA10	W	Both	HGMD/OMIM	68%		27	0.82		0.55	28
NAA50	W	Cytoplasm		82%		21	0.71		0.55	29
MLL	R	Nucleus	CGC	69%	5	26	0.61		0.55	30
NCOA3	W	Both	HGMD	68%	7	29	0.43	Down	0.55	31
NCOA2	W	Nucleus	CGC	73%	8	19	0.54		0.54	32
KAT5	W	Both		72%	6	19	0.63	Down	0.54	33
BRD1	R	Nucleus	HGMD	64%	4	34	0.61		0.54	34
NAA40	W	Unknown		77%		40	0.55		0.53	35
KAT6B	W	Nucleus	HGMD/OMIM	66%	4	18	0.73	Down	0.53	36
KAT2B	WR	Nucleus		75%	15	8	0.47		0.53	37
HDAC5	E	Both		69%	10	16	0.54		0.52	38
KAT2A	WR	Nucleus		76%	6	5	0.64	Down	0.51	39
ASH1L	R	Both		67%		22	0.79	Up	0.50	40
TRIM66	R	Nucleus		52%		37	0.79		0.50	41
HDAC7	E	Both		61%	9		0.76		0.49	42
NAT9	W	Unknown		62%		25	0.80		0.48	43
SIRT2	E	Cytoplasm		60%	7	3	0.79	Up	0.48	44
HAT1	W	Both		77%		27	0.56	Down	0.48	45
ZMYND8	R	Unknown		72%	1	21	0.62		0.47	46
CDYL	W	Nucleus		73%	3	20	0.52	Down	0.46	47
PHIP	R	Nucleus	HGMD	71%		31	0.55		0.44	48
MGEA5	W	Both	HGMD	80%		17	0.63		0.44	49
NAA60	W	Unknown		70%		15	0.71		0.43	50

Gene	Class	Subcellular localisation	Disease implications	Conservation (Average sequence identity)	Known protein-protein interactions	Degree of co-expression	Expression breadth	Expression trend in ESC differentiation	Candidate score	Candidate rank
HDAC4	E	Both	HGMD/OMIM	68%	12	5	0.43		0.42	51
CDY1	W	Nucleus		49%		22	0.81		0.42	52
CDY2A	W	Nucleus		50%		22	0.80		0.42	53
HDAC8	E	Both		76%	9				0.40	54
SIRT4	E	Cytoplasm		66%	1	8	0.71		0.40	55
BPTF	R	Both		61%		36	0.57		0.38	56
ATAD2B	R	Unknown		63%		38	0.48		0.38	57
TRIM24	R	Both	OMIM	65%	5	8	0.59		0.38	58
SIRT3	E	Cytoplasm	HGMD	51%	2	4	0.83		0.38	59
HDAC9	E	Nucleus	HGMD	70%	11				0.38	60
SIRT6	E	Nucleus		59%		14	0.81	Down	0.38	61
SIRT5	E	Cytoplasm	HGMD	68%	2	13	0.55		0.37	62
BAZ1A	R	Nucleus		66%	2	20	0.51		0.37	63
KAT7	W	Nucleus		80%		6	0.59		0.36	64
SAT1	W	Cytoplasm	HGMD/OMIM	65%	1	1	0.73		0.36	65
SIRT7	E	Both		58%		38	0.44	Up	0.34	66
NAA11	W	Both		61%		20	0.65		0.33	67
CSRP2BP	W	Both		78%	2				0.30	68
BRPF3	R	Unknown		73%	4			Down	0.30	69
PBRM1	R	Nucleus	CGC	75%	2			Up	0.28	70
HDAC10	E	Both		54%	13				0.27	71
GTF3C4	W	Nucleus		63%		2	0.67		0.25	72
ATAD2	R	Nucleus		65%	1	9	0.43		0.25	73
BAZ2B	R	Nucleus		67%		8	0.54		0.25	74
GNPNAT1	W	Cytoplasm		81%				Down	0.24	75
SP100	R	Nucleus		27%		20	0.59	Up	0.23	76
SP140	R	Both		28%	1	21	0.39		0.23	77
SP110	R	Nucleus	HGMD/OMIM	27%		26	0.45		0.22	78
SP140L	R	Unknown		26%		30	0.27		0.21	79
NAA20	W	Both		75%					0.20	80
BRD8	R	Nucleus		61%	1	5	0.24		0.17	81
ESCO2	W	Both	HGMD/OMIM	56%	3			Down	0.15	82
ESCO1	W	Both		59%	2				0.15	83
BRD9	R	Unknown		62%	1				0.15	84
BRWD1	R	Both	HGMD	60%	1				0.12	85
BRWD3	R	Unknown	HGMD/OMIM	65%					0.11	86
TAF1L	R	Nucleus	HGMD	64%					0.10	87
NAT6	W	Cytoplasm		35%		8	0.43	Up	0.10	88
AANAT	W	Cytoplasm	HGMD/OMIM	48%		8	0.22		0.09	89
SAT2	W	Cytoplasm		51%	1				0.09	90
BRDT	R	Nucleus		54%		3	0.33		0.09	91
NAA30	W	Cytoplasm		58%					0.06	92
NAT8	W	Cytoplasm		46%		2	0.28		0.05	93
CECR2	R	Unknown		52%				Down	0.05	94
NAT14	W	Cytoplasm		46%					0.03	95
NAT8L	W	Cytoplasm	HGMD/OMIM	40%					0.02	96
NAT16	W	Unknown		30%					0.02	97
KIAA2026	R	Unknown		27%					0.01	98
SATL1	W	Unknown	HGMD	17%					0.00	99
NAT8B	W	Cytoplasm							0.00	100

The three very highly conserved deacetylases HDAC1–3, the widely expressed reader proteins BAZ2A, BRPF1 and TAF1, and the acetyltransferases ELP3 and KAT6A appear to be the most universally important components of the human lysine acetylation signalling system across tissues and species, according to the criteria used here. Strangely, both TAF1 and ELP3 are reported by Compara as absent in some species, such as *Ornithorhynchus anatinus* (platypus), *Felis catus* (cat) and *Sus scrofa* (pig). InParanoid reports full conservation of these genes within its set of

species, which does not contain these three species. These might therefore be false negatives due to insufficient sequencing coverage.

The genetic disease associations from HGMD and OMIM are spread relatively evenly across the table, while the nearly exclusively somatic cancer-causative mutations from CGC appear more frequently at the head of the list. This could be explained by the consideration that essential genes, where any mutations are likely to be embryonically lethal, should not tend to appear in the genetic disease databases (HGMD and OMIM), although these essential genes are of the highest functional importance.

Several of the lysine acetylation-related genes were either significantly up- or down-regulated in differentiated cell types, as compared to embryonic stem cells (ESCs). This could be explained in light of the extensive chromatin remodelling that takes place during differentiation, and in which changes in histone acetylation play an important role (Rasmussen, 2003; Azuara et al., 2006; Meshorer and Misteli, 2006).

3.4. Discussion

Several potentially interesting aspects of the human lysine acetylation system were found during the above analyses. The proportions of the writer and eraser categories (Fig. 3.1, Fig. 3.2) were similar to the constitution of the phosphorylation and ubiquitination signalling systems, which contain a much smaller number of erasing enzymes compared to the number of modifying enzymes (for phosphorylation), or compared to the number of distinct modifying complexes (for ubiquitination) (Komander, 2009; Bischoff and Schlüter, 2012). Potentially, this might indicate that erasers are primarily responsible for maintaining a dynamic equilibrium of acetylation and deacetylation at a large number of substrate sites, while the writers may have primary signalling roles at a more specific set of sites. Alternatively, it may indicate that a portion of lysine acetylation sites are not reversible.

In addition, some writer enzymes displayed higher divergence than readers or erasers, indicating that they may potentially be the most evolvable part of the system among the vertebrates (Fig. 3.3). The conservation profiles also highlighted several strikingly conserved genes, including the acetyltransferases ELP3, NAT10 and NAA20, the writer-reader proteins CREBBP, EP300, KAT2A, and KAT2B, the reader proteins SMARCA2 and SMARCA4, and the deacetylases HDAC1, HDAC2

and HDAC3. All of these except NAA20 also scored in the top third of the ranked table, potentially indicating that conservation is well-correlated with the other features included in the ranking method.

The conservation profiles generated using InParanoid, which covers many non-vertebrate species, appeared to show interesting evidence of family expansions by gene duplication. Similar lineage-specific expansions have been described in the evolution of the lysine methylation system (Aravind et al., 2011).

With only two exceptions, the more highly-ranked half of the candidates are located at least partially in the nucleus, or their localisation was still unknown. Lower-ranked candidates more frequently occurred in the cytoplasm. This suggests that the most essential functions of lysine acetylation tend to be carried out within the nucleus. Since the higher-ranked half also display higher conservation, a remote speculation might be that lysine acetylation first fulfilled essential roles in the nucleus before diversifying into cytoplasmic roles over evolutionary time, perhaps accompanying the family expansions observed in Figure 3.4.

The results on the interaction network are in accordance with the consideration that the functional interactions between writers, readers and erasers can be separated in time and space, since they may take place on a substrate (Fig. 3.8). Aside from such indirect functional interactions, it is conceivable that some of the relative sparsity of the experimentally-determined protein-protein interaction network (Fig. 3.7) compared to the co-expression network (Fig. 3.8) may be due to physical interactions which remain unidentified. Though expression data is not an ideal predictor of protein concentration, studies in human cell lines and on the expression dataset used here have reported good correlations between mRNA and protein concentrations (Ballman, 2008; Lundberg et al., 2010; Vogel et al., 2010; Nagaraj et al., 2011). This indicates that the co-expressed genes might form stoichiometric complexes, and that co-expression could potentially point to direct physical interactions.

The writer, reader and eraser proteins as well as acetylation substrates frequently made physical interactions with other acetylated proteins (Table 3.2). This may indicate strong functional links between different acetylation-related proteins and substrates, as well as among the acetylated substrates themselves. In addition, “shuttling” proteins, which can be found both in the nucleus and the cytoplasm,

were among the most frequent interactors with acetylation substrates (73.6%). Certain shuttling proteins might therefore take part in acetylation-dependent signal transduction mechanisms across the nuclear membrane.

In summary, this chapter has given an overview of the writer, reader and eraser proteins involved in the human lysine acetylation signalling system by using multiple types of data, including the conservation, expression patterns, protein-protein interactions and subcellular localisations of its components. This information was then integrated into a ranked summary table order to systematically highlight promising candidate proteins for detailed further investigation (Table 3.2). The approach we used here can be thought of as a generalised framework which can be applied to the study of PTM signalling systems, and which makes use of pre-existing large-scale datasets.

4. The interaction network of Oct4

This chapter is the result of a collaboration with Dr. Jyoti Choudhary's group at the Wellcome Trust Sanger Institute, published in Pardo *et al.* (Pardo *et al.*, 2010), an article of which I am a joint first author. My collaborators identified physical interactors of Oct4, a transcription factor with a central role in the establishment and maintenance of stem cell identity. Using similar methods to those presented in Chapter 3, I performed an analysis of its potential signalling functions. Here, I will present an in-depth analysis of the Oct4 interactome based on transcriptional regulation, expression data, sequence conservation and phenotypic data, including human disease phenotypes and cancers.

4.1. Introduction

Embryonic stem cells hold great promise in regenerative medicine due to their ability to differentiate into all three germ layers (endo-, meso- and ectoderm), and therefore to form all cell types of an adult organism. This property is termed pluripotency (Evans and Kaufman, 1981; Martin, 1981; Barberi *et al.*, 2005). Their envisioned uses include the regeneration of tissues damaged by injury or degenerative diseases, growing autologous patient tissues for drug tolerance and efficacy tests, and even the regeneration of complex organs through tissue engineering (Park *et al.*, 2008; Yamzon *et al.*, 2008; Zhou *et al.*, 2008; Macarthur *et al.*, 2009). They may also find applications in the generation of well-defined cell populations as models, which could be derived from them using defined combinations of factors (Enver *et al.*, 2009; Hyman and Simons, 2011; Spence *et al.*, 2011). A continuing challenge, however, is the discovery of a safe and ethically acceptable means of obtaining cells with pluripotent capabilities. Induced pluripotent stem cells (iPS cells), which can be generated by reprogramming adult cells using specific combinations of transcription factors, are one viable option, though embryonic stem cells are still considered superior in their homogeneity and therapeutic safety (Belmonte *et al.*, 2009). Differences in gene expression can exist between iPS cells and embryonic stem cells (Chin *et al.*, 2009), and some iPS cells may carry chromosomal aberrations (Mayshar *et al.*, 2010), though gene expression noise may also be a normal feature of ES cells which has been associated with the maintenance of pluripotency (Trott *et al.*, 2012).

Induced pluripotent cells capable of differentiating into all three germ layers were first generated from mouse fibroblasts by ectopic expression of four transcription factors: Oct4, Sox2, Klf4 and c-Myc (Takahashi and Yamanaka, 2006). This feat was then repeated with human fibroblasts (Takahashi et al., 2007). One of these four transcription factors, c-Myc, is an oncogene which was later shown to be expendable in iPS cell generation, though at the cost of reduced reprogramming efficiency (Dang, 1999; Nakagawa et al., 2008). While most approaches used viral delivery for the ectopic expression of the reprogramming factors, stable reprogramming by direct delivery of proteins fused with a cell-penetrating peptide has also been described, and alleviates the risk of mutagenesis (Kim et al., 2009a).

Another of the four reprogramming transcription factors, Oct4, has been shown to be key in maintaining embryonic stem cell identity, and is able to ectopically induce pluripotency in adult neural stem cells on its own (Kim et al., 2009b). Since combinatorial interactions of transcription factors are a common feature in the establishment and maintenance of cell type (Ravasi et al., 2010), insight into Oct4's physical interaction partners should illuminate how the pluripotent state is established and maintained. In this study, my collaborators identified a considerably expanded set of Oct4-binding proteins in mouse embryonic stem cells. By integrating functional annotation with data on gene expression, transcriptional regulation, known protein-protein interactions, mutant phenotypes, disease and cancer, I here present a comprehensive analysis of the Oct4 interactome as determined through affinity purification and mass spectrometry.

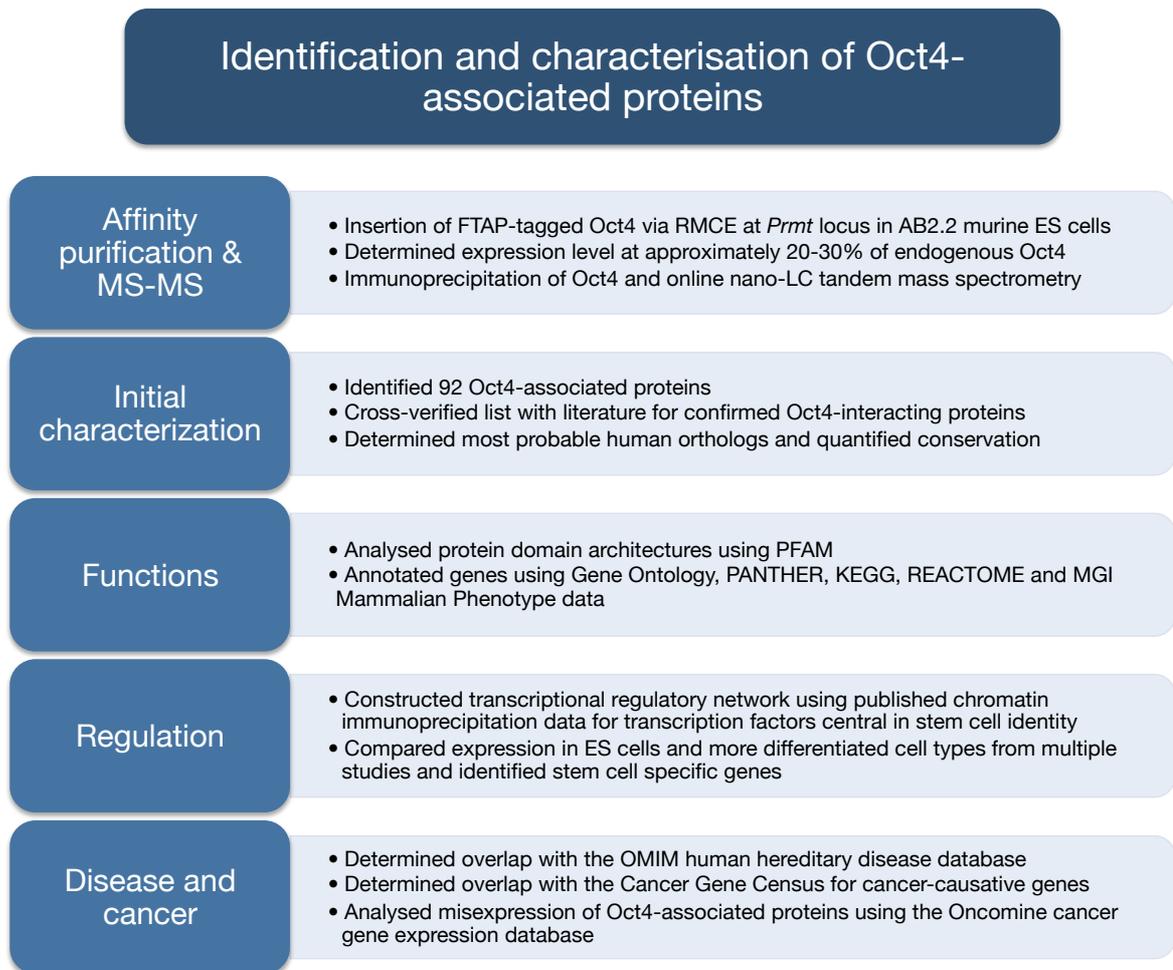


Figure 4.1: An overview of the workflow undertaken to characterise the Oct4 interactome.

4.2. Methods

4.2.1. Identification of Oct4 interactors

My collaborators identified Oct4-interacting proteins in mouse embryonic stem cells through an affinity purification and mass spectrometry approach. Using recombination-mediated cassette exchange (RMCE), my collaborators introduced an FTAP-tagged copy of Oct4 into murine AB2.2 embryonic stem cells at the *Prmt* locus (Pardo et al., 2010). The tagged copy was expressed at a low level in parallel with endogenous Oct4 (approximately 20–30% of the endogenous protein's concentration). By using an FTAP-tag specific antibody, the tagged Oct4 was then immunoprecipitated in conjunction with interacting proteins. Three separate single-step purifications were performed. Tandem mass spectrometry (MS/MS) in conjunction with nano-liquid chromatography (nano-LC) was then used to identify the extracted proteins, and proteins appearing in all three Oct4-FTAP purifications

but not in controls were retained. The data are deposited in the Proteomics Identifications Database (PRIDE) under the accession numbers 12005 through 12010 (Vizcaíno et al., 2009). Using a data integration approach, I then analysed this high-confidence set of 92 Oct4 interactors.

4.2.2. Integrating known protein-protein interactions

To study the interactions among the Oct4 interactors and identify well-described protein complexes, experimental protein-protein interactions were obtained from the STRING database (Jensen et al., 2009). “Knowledge-based” interactions imported from other major databases were also included. The resulting network was visualised using Cytoscape 2.6.3 (Cline et al., 2007).

4.2.3. Expression at various differentiation stages

For the analysis of expression at different stages of differentiation, data were obtained for 43 mouse samples in StemBase (Sandie et al., 2009), originating from 16 studies with Affymetrix MOE430A microarray chips, as used in an Oct4 expression profiling study covering murine ESCs, embryonal carcinoma cell lines, and several early differentiated lineages (Campbell et al., 2007). Expression data were available for 70 of the 92 Oct4-associated proteins. Where multiple probes were available, the arithmetic mean expression was used. Student’s t-test was used for identifying genes differentially expressed in ESCs as compared to more differentiated cell types, and the Bonferroni correction was used for multiple testing. Expression values were log₂-transformed and colour-coded as a gradient from blue (more than 1.5 times the standard deviation below the global microarray mean) via black (microarray mean) to yellow (more than 1.5 times the standard deviation above the mean). Average-linkage hierarchical clustering using using Genesis 1.7.6 (Sturn et al., 2002) was performed to arrive at the final layout.

4.2.4. Transcriptional regulation

ChIP-on-chip data were obtained for Oct4 and eight other transcription factors (Kim et al., 2008). The significance of the number of Oct4-associated proteins regulated by these factors was assessed against 1000 random sets of 92 genes. The protein

interaction network was generated with Cytoscape 2.6.3 (Cline et al., 2007), with a spring-embedded layout.

4.2.5. Sequence identity between mouse and human

Orthologous human proteins were identified with the g:Profiler orthology search tool (Reimand et al., 2007). Where g:Profiler was unable to provide an ortholog, NCBI BLASTP was used to find the best-matching human protein (Altschul et al., 1990). The one-to-one orthologous pairs were globally aligned with the Needleman-Wunsch algorithm (Needleman and Wunsch, 1970). For assessment of the degree of conservation between the Oct4-associated proteins and their orthologs, sequence identities of all mouse-to-human ortholog pairs sequence lengths comparable to the Oct4-associated proteins in ENSEMBL release 57 were compared via a Mann-Whitney U test.

4.2.6. Developmental, disease and cancer phenotypes

Mammalian Phenotype Ontology annotations were obtained from the Mouse Genome Informatics project (Blake et al., 2009), human disease associations were obtained from OMIM (Hamosh et al., 2005), and known cancer-causing mutations in genes were obtained from the Cancer Gene Census (Futreal et al., 2004). Student's t-test was used to assess whether the observed numbers of Oct4 interactors with lethal phenotypes or disease, and cancer associations differ significantly from those found in 1000 random sets of 92 genes.

4.2.7. Misexpression in cancer types

Data on significantly misexpressed human orthologs of the Oct4 interactors were obtained from the Oncomine database (Rhodes et al., 2007). To correct for multiple testing and expression noise, genes were considered misexpressed only if Oncomine reported a p-value below 10^{-10} . This resulted in a manageable number of significant misexpression events. The network figure was generated using Cytoscape 2.8.2 (Cline et al., 2007), using a force-directed layout.

4.2.8. Functional enrichment according to PANTHER annotation

The PANTHER functional annotation resource was used to identify annotation terms which were significantly overrepresented among the Oct4 interactors compared to a genome-wide background in the mouse (Thomas et al., 2003).

4.2.9. Overrepresented domains

Protein domains were identified with Pfam 24.0, and genome-wide and nuclear frequencies were calculated from domain annotations and experimentally determined subcellular localisation information in UniProtKB/Swiss-Prot release 15.15 (UniProt Consortium, 2010). Fisher's exact test was used to statistically assess the differences in domain occurrence.

4.3. Results

By integrating our list of Oct4-associated proteins obtained through affinity purification and mass spectrometry with additional information, we hope to gain knowledge on the relative importance of the candidate proteins in stem cell identity and function. This information is then envisioned to be used to prioritise the most promising Oct4 interactors for further studies.

4.3.1. Identification of Oct4 interactors

Using a tandem affinity purification approach followed by LC-MS/MS (liquid chromatography coupled to tandem mass spectrometry) in triplicate, my collaborators at the Sanger Institute were able to identify 92 Oct4-interacting proteins which were part of a diverse set of functional categories and complexes (Table 4.1). Sox2 and Nanog, two of the best-characterised Oct4 interactors (Reményi et al., 2003), were identified in only some of the three replicate experiments, and are therefore not included in this list. Overall, Oct4 associated with a varied set of proteins, including components of chromatin remodelling complexes, transcription factors and enzymes involved in histone and DNA methylation.

Table 4.1: Oct4-associated proteins classified into protein complexes and functional categories. Gene names and sequence accessions for the IPI database are given, and the number of unique peptides in each of the three replicate experiments is shown. MW: molecular weight (Da).

Gene name	Accession	Description	MW	Exp I	Exp II	Exp III
Bait						
Oct4 (Pou5f1)	IPI00117218	POU domain, class 5, transcription factor 1	38705.35	20	19	21
NuRD complex						
Chd4	IPI00396802	Chromodomain-helicase-DNA-binding protein 4	219096.34	21	40	33
Gatad2a	IPI00625995	p66 alpha isoform a	67762.39	13	11	16
Gatad2b	IPI00128615	Isoform 1 of Transcriptional repressor p66-beta	65712.08	13	7	16
Mbd3	IPI00131067	Isoform 1 of Methyl-CpG-binding domain protein 3	32168	4	6	4
Mta1	IPI00624969	Mta1 protein	80019.75	13	14	16
Mta2	IPI00128230	Metastasis-associated protein MTA2	75723.93	25	17	21
Mta3	IPI00125745	Isoform 1 of Metastasis-associated protein MTA3	67719.08	10	9	7
Hdac1	IPI00114232	Histone deacetylase 1	55609.93	11	12	11
Spalt-like transcriptional repressors						
Sall1	IPI00342267	Sal-like 1	141745.1	11	17	18
Sall3	IPI00123404	Isoform 1 of Sal-like protein 3	140610.62	5	5	6
Sall4	IPI00475164	Isoform 1 of Sal-like protein 4	114711.29	35	29	28
BAF complex						
Smarca4	IPI00875789	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily a, member 4, isoform CRA_b	181913.68	3	5	1
Smarcc1	IPI00125662	Isoform 1 of SWI/SNF complex subunit SMARCC1	123326.28	7	3	4
Actl6a	IPI00323660	Actin-like protein 6A	47930.54	4	1	1
FACT complex						
Ssrp1	IPI00407571	Isoform 2 of FACT complex subunit SSRP1	81766.73	36	12	32
Supt16h	IPI00120344	FACT complex subunit SPT16	120319.5	64	34	56
LSD1 complex						
Aof2	IPI00648295	Amine oxidase (Flavin containing) domain 2	95113.5	9	6	7
Rcor2	IPI00226581	REST corepressor 2	58042.89	5	6	2
ISWI chromatin remodelling complex						
Smarca5	IPI00396739	SWI/SNF-related matrix-associated actin-dependent regulator of chromatin subfamily A member 5	122291.43	5	7	2
INO80 chromatin-remodeling complex						
Ino80	IPI00378561	Isoform 1 of Putative DNA helicase INO80 complex homolog 1	177265.25	5	6	1
Nfrkb	IPI00274469	Nuclear factor related to kappa-B-binding protein	139134.5	22	11	4
Actl6a	IPI00323660	Actin-like protein 6A	47930.54	4	1	1
Histone chaperone complex						
Asf1a	IPI00132452	Histone chaperone ASF1A	23099.19	3	1	3
Cabin1	IPI00380107	Calcineurin binding protein 1	245584.2	8	18	7
Hira	IPI00123694	Isoform Long of Protein HIRA	113235.4	10	7	10
Ubn2	IPI00854896	Isoform 4 of Uncharacterized protein KIAA2030	142523.51	17	19	11
Transcription factors						
Arid3b	IPI00277032	Isoform 1 of AT-rich interactive domain-containing protein 3B	61091.99	9	8	3
Atf2	IPI00110172	Isoform 1 of Cyclic AMP-dependent transcription factor ATF-2	52550.73	4	5	5
Creb1	IPI00119924	Isoform 1 of cAMP response element-binding protein	36879.65	3	3	3
Ctbp1	IPI00128155	Isoform 1 of C-terminal-binding protein 1	48170.7	9	4	3
Ctbp2	IPI00856974	Isoform 2 of C-terminal-binding protein 2	107801.19	10	3	4
Klf4	IPI00120384	Kruppel-like factor 4	52531.81	1	1	2
Mitf	IPI00125758	Isoform A of Microphthalmia-associated transcription factor	59160.39	2	3	1
Nfyc	IPI00108204	Nuclear transcription factor Y subunit gamma	37230.86	7	3	1
Sp1	IPI00323887	Isoform 1 of Transcription factor Sp1	81309.84	5	4	6
Tcfe3	IPI00380308	Isoform 1 of Transcription factor E3	61555.7	9	5	7
Tcf7	IPI00314502	Transcription factor EB	52638.18	2	3	2
Zbtb10	IPI00223276	Zinc finger and BTB domain containing 10 isoform 1	118071.48	9	5	12
Zbtb2	IPI00652356	Putative uncharacterized protein	58189.61	1	1	1
Zbtb43	IPI00230530	Zinc finger protein 297B isoform a	57551.37	2	1	3

Gene name	Accession	Description	MW	Exp I	Exp II	Exp III
Transcription factors						
Zfhx3	IPI00475055	AT motif binding factor 1	410697.11	3	36	1
Zfp217	IPI00758403	Zinc finger protein 217	115181.65	11	10	4
Zfp219	IPI00469594	Zinc finger protein 219, isoform CRA_a	78831.33	4	6	7
Zfp513	IPI00830836	Isoform 1 of Zinc finger protein 513	59968.12	1	2	1
Zic2	IPI00127145	Zinc finger protein ZIC 2	55546.36	1	1	1
Zscan4b	IPI00755380	Similar to Gene model 397	58667.99	4	6	7
Regulation of transcription						
Acin1	IPI00121136	Isoform 1 of Apoptotic chromatin condensation inducer in the nucleus	151000.03	12	2	13
Brwd1	IPI00121655	Isoform A of Bromodomain and WD repeat-containing protein 1	262057.12	1	2	2
Hcfc1	IPI00828490	Host cell factor C1	216798.82	19	8	12
Ifi202b	IPI00126725	Interferon-activable protein 202	50727.44	4	6	3
Phf17	IPI00453799	Isoform 1 of Protein Jade-1	95434.25	4	2	4
Rfx2	IPI00406298	DNA-binding protein RFX2	76998.5	8	1	1
General transcription						
Ttf2	IPI00112371	Transcription termination factor 2	126706.43	5	2	2
Recombination/repair						
Lig3	IPI00124272	Isoform Alpha of DNA ligase 3	114656.59	5	6	1
Msh6	IPI00310173	MutS homolog 6	152813.37	10	4	5
Parp1	IPI00139168	Putative uncharacterized protein	113491.6	24	17	33
Top2a	IPI00122223	DNA topoisomerase 2-alpha	173567.4	30	44	12
Xrcc1	IPI00118139	DNA repair protein XRCC1	69270.68	6	3	1
Xrcc5	IPI00321154	ATP-dependent DNA helicase 2 subunit 2	83802.29	6	5	13
Xrcc6	IPI00132424	ATP-dependent DNA helicase 2 subunit 1	69726.04	12	5	12
Replication						
Rpa1	IPI00124520	Replication protein A 70 kDa DNA-binding subunit	69620.87	2	3	2
Rpa3	IPI00132128	Replication protein A 14 kDa subunit	13688.99	1	1	1
Helicases						
Chd1	IPI00107999	Chromodomain-helicase-DNA-binding protein 1	197601.13	8	9	7
Chd3	IPI00675483	Chromodomain helicase DNA binding protein 3	234196.54	5	6	6
Chd5	IPI00875673	chromodomain helicase DNA binding protein 5 isoform1	224027.21	10	9	11
Dhx9	IPI00339468	Isoform 2 of ATP-dependent RNA helicase A	150907.1	15	15	31
Hells	IPI00121431	Isoform 1 of Lymphocyte-specific helicase	95806.47	3	1	1
Histones						
Hist1h3e	IPI00282848	Histone cluster 2, H3c1 isoform 2	20348.12	8	9	4
Hist1h4b	IPI00407339	Histone H4	11360.38	11	13	6
Hist3h2bb	IPI00229539	Histone cluster 3, H2bb	17248.15	9	7	4
Heterogeneous nuclear ribonucleoproteins						
Hnrnpab	IPI00277066	Heterogeneous nuclear ribonucleoprotein A/B isoform 1	36302.44	3	2	2
Hnrnpl	IPI00620362	Heterogeneous nuclear ribonucleoprotein L	64550.49	10	4	8
Hnrnpu	IPI00458583	Putative uncharacterized protein	88661.02	8	2	12
Histone ubiquitination (E3 ubiquitin ligase complex)						
Cul4b	IPI00224689	Cullin 4B	111314	7	3	9
Ddb1	IPI00316740	DNA damage-binding protein 1	128026.73	7	5	16
Enzymes						
Cad	IPI00380280	Carbamoyl-phosphate synthetase 2, aspartate transcarbamylase, and dihydroorotase	245649.59	4	18	11
Dnmt3a	IPI00131694	Isoform 1 of DNA (cytosine-5)-methyltransferase 3A	103203.53	4	4	7
Dnmt3l	IPI00109459	DNA (cytosine-5)-methyltransferase 3-like	49159.08	7	3	7
Myst2	IPI00228457	Isoform 2 of Histone acetyltransferase MYST2	67588.54	4	2	3
Ogt	IPI00420870	Isoform 1 of UDP-N-acetylglucosamine--peptide N-acetylglucosaminyltransferase 110 kDa subunit	118131.41	14	4	10
P4ha1	IPI00399959	Isoform 2 of Prolyl 4-hydroxylase subunit alpha-1	61132.82	14	5	5
Ppp2r1a	IPI00310091	Serine/threonine-protein phosphatase 2A 65 kDa regulatory subunit A alpha isoform	66079.23	3	4	1
Trim24	IPI00227778	Isoform Short of Transcription intermediary factor 1-alpha	114824.79	3	4	6
Trim33	IPI00409904	Isoform Alpha of E3 ubiquitin-protein ligase TRIM33	125931.28	2	2	2
Karyopherins						
Kpna2	IPI00124973	Importin subunit alpha-2	58234.28	7	7	2
Kpna3	IPI00230429	Importin subunit alpha-3	58193	2	1	2

Gene name	Accession	Description	MW	Exp I	Exp II	Exp III
Chaperones						
Dnaja1	IPI00132208	DnaJ homolog subfamily A member 1	45580.73	6	1	5
Proteasome						
Psmb6	IPI00119239	Proteasome subunit beta type-6	25590.57	2	2	1
Nuclear assembly/organisation						
Emd	IPI00114401	Emerin	29417.38	2	1	1
Matr3	IPI00453826	Matrin-3	95085.04	14	6	11
Miscellaneous						
Amotl2	IPI00263333	Isoform 1 of Angiomotin-like protein 2	85454.32	1	4	2
Cubn	IPI00889898	Cubilin	407679.63	3	7	2
Nudc	IPI00132942	Nuclear migration protein nudC	38334.29	2	2	2

4.3.2. Integrating known protein-protein interactions

Many of the Oct4-associated proteins provide links to interesting cellular systems. These include DNA repair proteins, components of the NuRD and BAF chromatin remodelling complexes, components of mRNP granules (compact cytoplasmic assemblies of translationally inactive mRNAs of uncertain function) (Erickson and Lykke-Andersen, 2011), enzymes involved in *de novo* DNA methylation and a particularly large set of transcription factors and chromatin remodellers (Fig. 4.1). These transcriptional regulators may be downstream effectors in Oct4 signalling, mediating the large-scale chromatin changes observed in embryonic stem cells (Rasmussen, 2003; Azuara et al., 2006; Meshorer and Misteli, 2006).

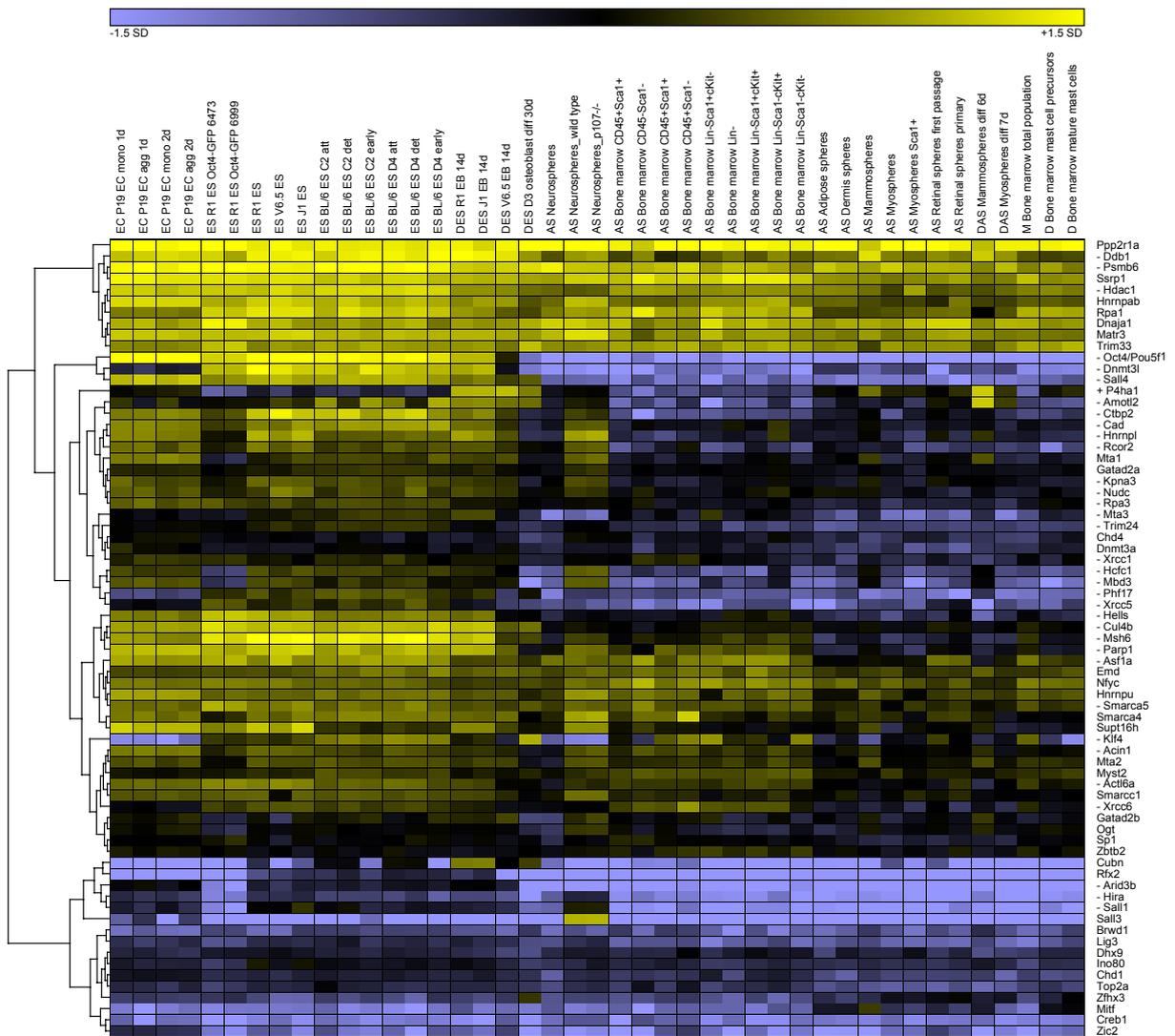


Figure 4.2: One third of the genes are significantly downregulated in differentiation. Expression of Oct4-associated genes in embryonic stem cells and differentiated cell types, based on microarray data. Expression values were log₂-transformed and colour-coded as a gradient from blue (more than 1.5 times the standard deviation below the global microarray mean) via black (microarray mean) to yellow (more than 1.5 times the standard deviation above the mean). Columns correspond to experimental samples, arranged as follows: embryonal carcinoma P19 (EC), ES cells (ES), differentiating embryonic stem cells (DES), adult stem cells (AS), differentiated adult stem cells (DAS), mixed cells (M), and differentiated cells (D). Genes whose expression is significantly up- or downregulated in differentiation are marked “+” or “-” respectively. Differential regulation was assessed using Student’s t-test with the Bonferroni correction for multiple testing. Average-linkage hierarchical clustering was performed to arrive at the final layout.

4.3.4. Regulation by stem cell transcription factors

Approximately half of the Oct4-associated proteins are transcriptionally regulated by key transcription factors participating in stem cell regulatory networks (Fig. 4.3, Fig.

4.4). For this analysis, published promoter binding data from chromatin immunoprecipitation experiments was obtained for Dax1 (Nr0b1), Klf4, Myc, Nac1 (Btbd14b, Nacc1), Nanog, Rex1 (Zfp42), Zfp281 and Sox2. All nine are thought to be central in the maintenance of embryonic stem cell identity (Kim et al., 2008). 48 of the 92 interactor promoters (as well as Oct4's own) are bound by at least one of these factors. Oct4 as a transcription factor has 10 regulatory targets among its interactors, including its own promoter.

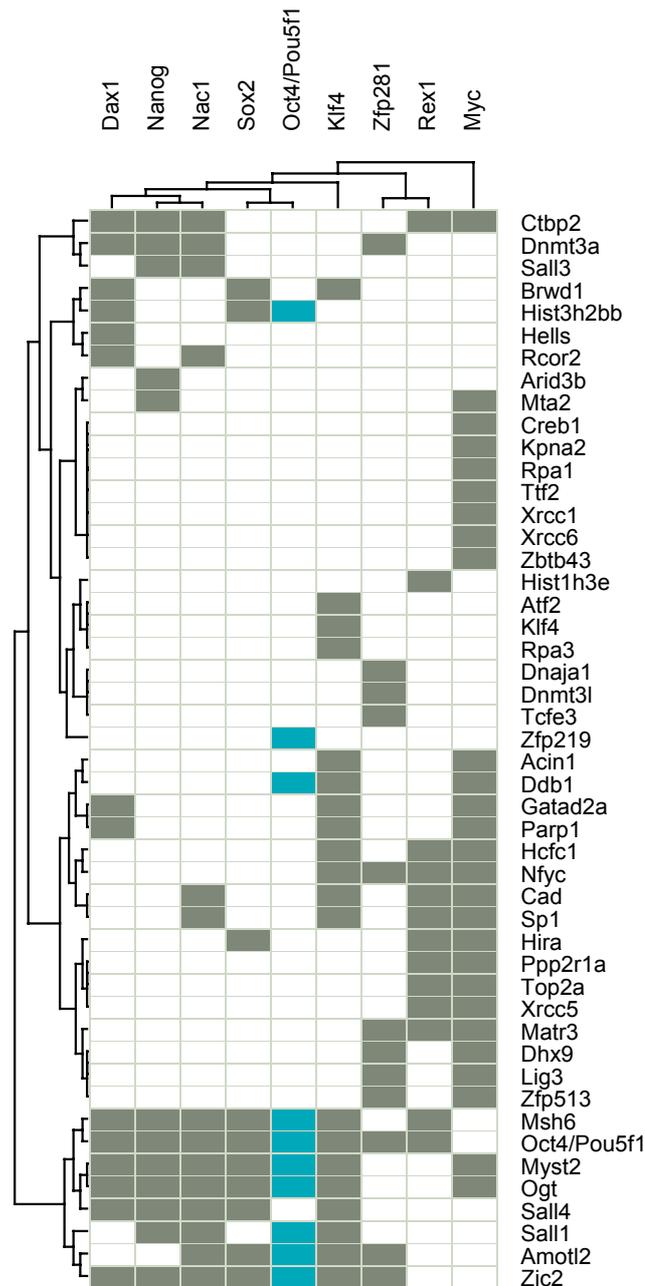


Figure 4.3: Regulation of Oct4 interactors by key stem cell transcription factors. This matrix shows promoter binding based on co-immunoprecipitation of Oct4 and 8 other transcription factors thought to be important in stem cell identity. Stem cell transcription factors correspond to columns while their target genes among the Oct4 interactors correspond to rows.

In a network representation of the same data (Figure 4.4), it becomes especially apparent that while Nanog, Nac1, Dax1 and Sox2 share many targets with Oct4 itself, c-Myc (Myc) has several exclusive targets, including the DNA repair proteins Xrcc1 and Xrcc6.

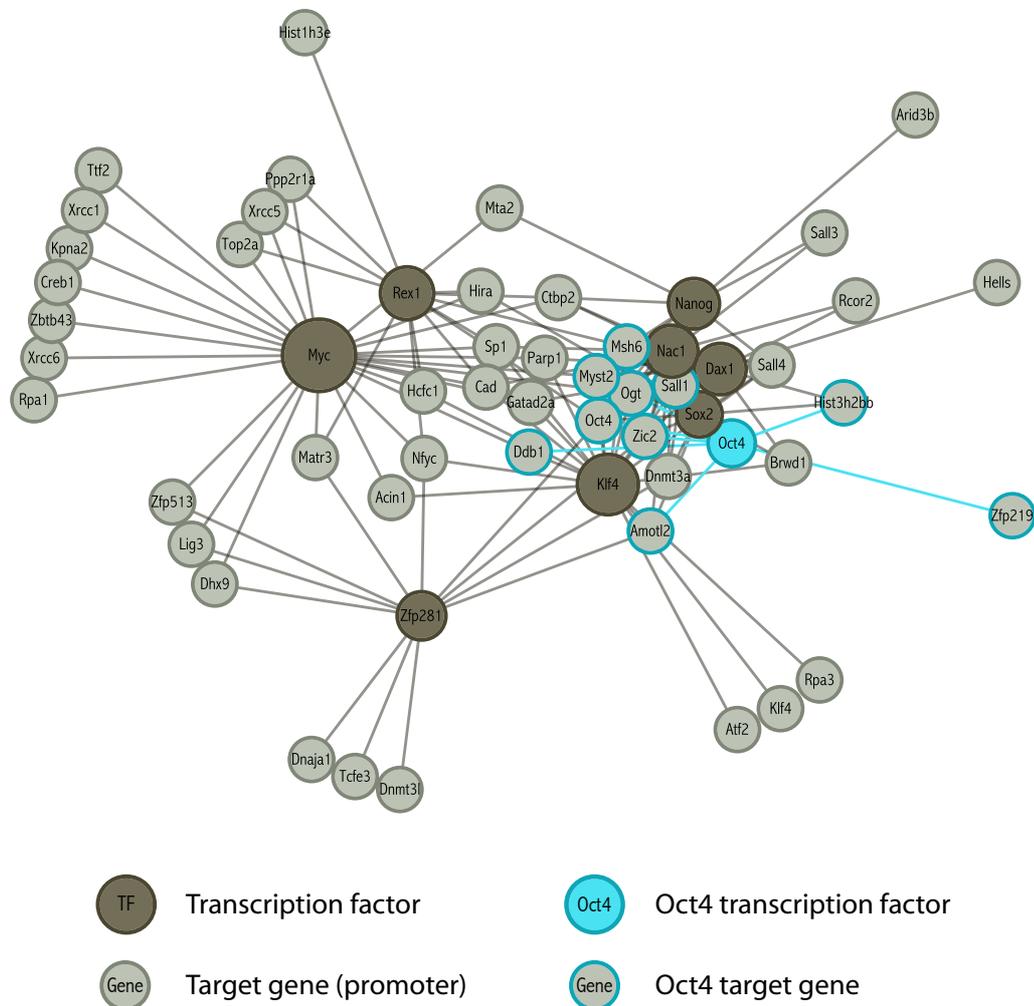


Figure 4.4: Regulation of Oct4 interactors by key stem cell transcription factors, presented as a network. The stem cell transcription factors are shown in dark grey, while their target genes among the Oct4-associated proteins are shown in light grey. Oct4 and its regulatory targets are highlighted in blue.

4.3.5. Sequence identity between mouse and human

The Oct4-associated proteins display significantly sequence conservation between mouse and human one-to-one orthologs (Fig. 4.5). The difference is significant at $p = 1.3 \cdot 10^{-20}$ (Mann-Whitney U test). Aside from a small number of outliers, this suggests that the identified Oct4-associated proteins are part of highly conserved systems

which are likely to be similarly important and similar in function in humans and other mammals.

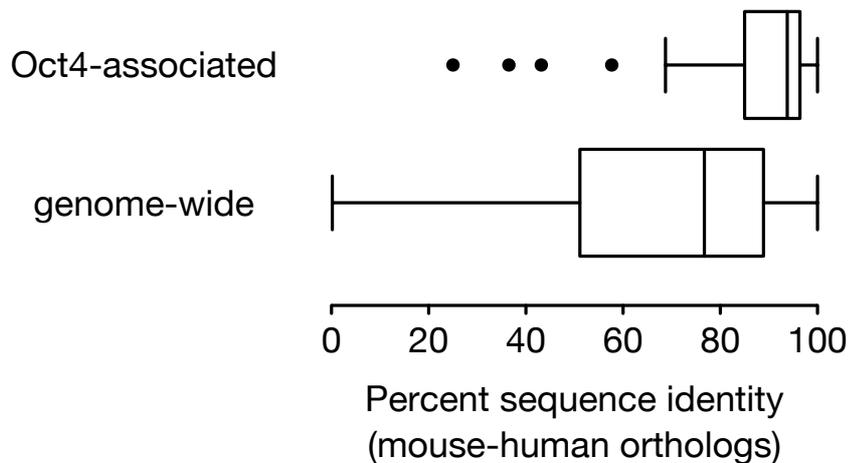


Figure 4.5: Oct4-associated proteins are significantly more conserved between human and mouse. Box plots with Tukey whiskers, showing sequence identity percentages as determined by global sequence alignment between the identified Oct4-associated proteins and their human one-to-one orthologs. Genome-wide sequence identities for all mouse-to-human one-to-one ortholog pairs contained in Ensembl Compara are shown for comparison.

4.3.6. Developmental, disease and cancer phenotypes

Knockout phenotype information in mice was available for 50 of the 92 Oct4-associated proteins, including Oct4 itself (Fig. 4.6). All 50 genes were annotated with non-normal phenotypes in at least one study. 42 out of the 50 Oct4 interactors which have been studied to date are essential for early mouse development, and mutations in these result in either pre-, peri- or postnatal lethality. In humans, a number of the one-to-one orthologs are associated with cancer and inherited developmental disorders (Table 4.2).

Mouse phenotype categories of Oct4 interactor mutants

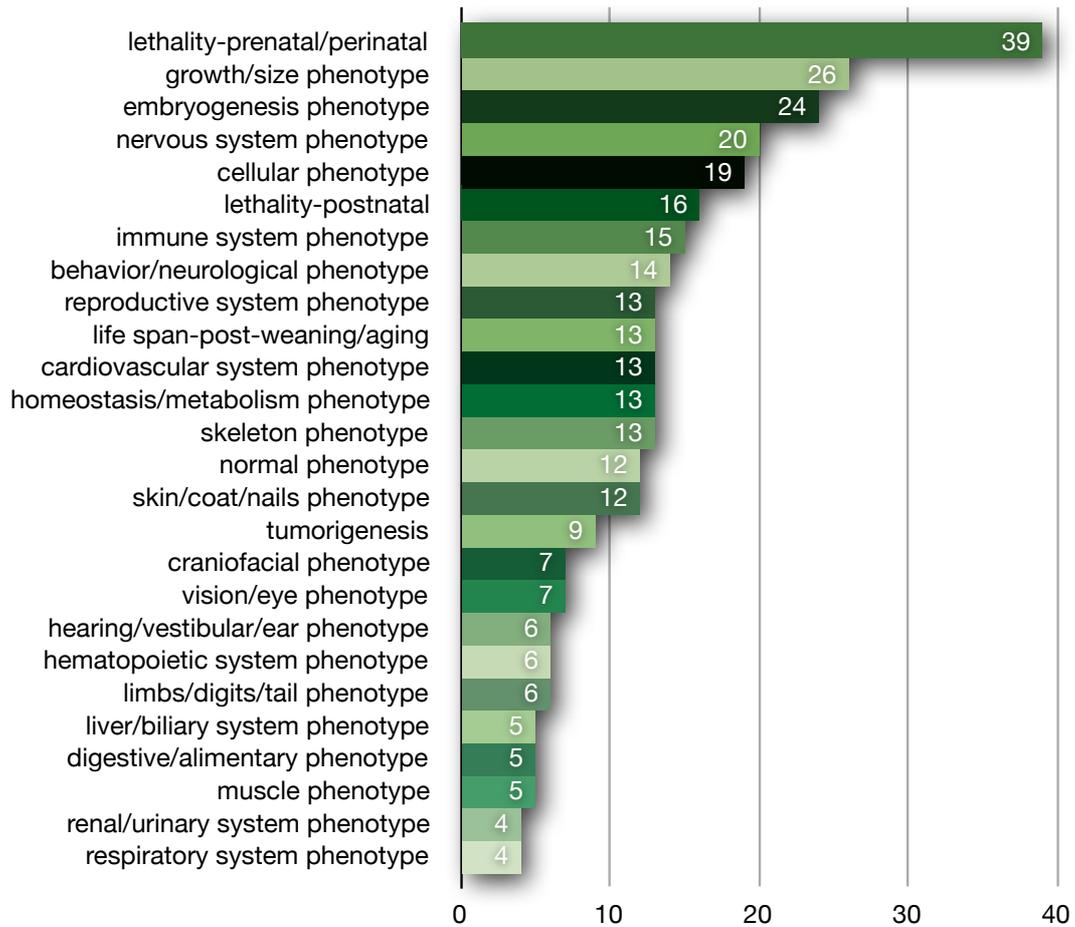


Figure 4.6: Phenotypic analysis. List of phenotypes caused by spontaneous, induced, and/or genetically-engineered mutations in the genes encoding the Oct4-associated proteins as retrieved from the MGI resource. The number of genes with a particular phenotype is given.

Table 4.2: Human hereditary disease associations for the human orthologs of Oct4-interacting proteins, obtained from the OMIM database.

Gene	Disorder type	Disorder
CREB1	Cancer	Histiocytoma, angiomatoid fibrous, somatic
CUBN	Hematological	Megaloblastic anemia-1, Finnish type
CUL4B	Multiple	Mental retardation syndrome, X-linked, Cabezas type Mental retardation-hypotonic facies syndrome, X-linked, 2
EMD	Muscular	Emery-Dreifuss muscular dystrophy
MATR3	Muscular	Myopathy, distal 2
MITF	Multiple	Tietz syndrome Waardenburg syndrome, type IIA Waardenburg syndrome/ocular albinism, digenic
MSH6	Cancer	Colorectal cancer, hereditary nonpolyposis, type 5 Endometrial cancer, familial Mismatch repair cancer syndrome
SALL1	Multiple	Townes-Brocks branchiootorenal-like syndrome Townes-Brocks syndrome
SALL4	Multiple	Duane-radial ray syndrome IVIC syndrome
TFE3	Cancer	Renal cell carcinoma, papillary, 1
TRIM24	Cancer	Thyroid carcinoma, papillary
TRIM33	Cancer	Thyroid carcinoma, papillary
ZFH3	Cancer	Prostate cancer, susceptibility to
ZIC2	Developmental	Holoprosencephaly-5

Table 4.3: Cancer-causative genes among the human orthologs of the Oct4-interacting proteins, according to the Cancer Gene Census project.

Gene	Mutation	Tissue	Cancer type
CREB1	translocation	mesenchymal	clear cell sarcoma, angiomatoid fibrous histiocytoma
MITF	amplification	epithelial	melanoma
MSH6	missense, nonsense, frameshift, splice site	epithelial	colorectal (somatic) colorectal, endometrial, ovarian (germline) non-polyposis colorectal cancer (hereditary)
POU5F1	translocation	mesenchymal	sarcoma
SMARCA4	frameshift, nonsense, missense	epithelial	NSCLC (non-small cell lung carcinoma)
TFE3	translocation	epithelial	papillary renal, alveolar soft part sarcoma, renal
TFEB	translocation	epithelial mesenchymal	renal, childhood epithelioid
TRIM24	translocation	blood	APL (acute promyelocytic leukemia)
TRIM33	translocation	epithelial	papillary thyroid

4.3.7. Misexpression in cancer types

We determined significant over- and underexpression of the Oct4 interactors in various types of cancer using the OncoPrint resource (Rhodes et al., 2007). Msh6, a mismatch repair gene implicated in hereditary predisposition to endometrial cancer (Table 4.3) (Ostergaard et al., 2005; Amberger et al., 2009), was here found significantly underexpressed in ovarian adenocarcinoma (Fig. 4.7). Two additional genes, Trim24 and Trim33, have been implicated in childhood papillary thyroid carcinoma (Klugbauer and Rabes, 1999). Here, we found Trim24 to be overexpressed in glioblastoma in addition to bladder and liver carcinomata, while Trim33 was underexpressed in glioblastoma and diffuse large B-cell lymphoma. Tfe3 (Tcf3) has been implicated in papillary renal cell carcinoma (Sidhar et al., 1996; Weterman et al., 1996), and was here found to be underexpressed in acute B-cell lymphoblastic childhood leukemia. Oct4-associated genes were always found consistently up- or downregulated where multiple studies were available in a specific cancer category (~11% of genes). This may lend additional confidence to the cancer misregulation events we observed.

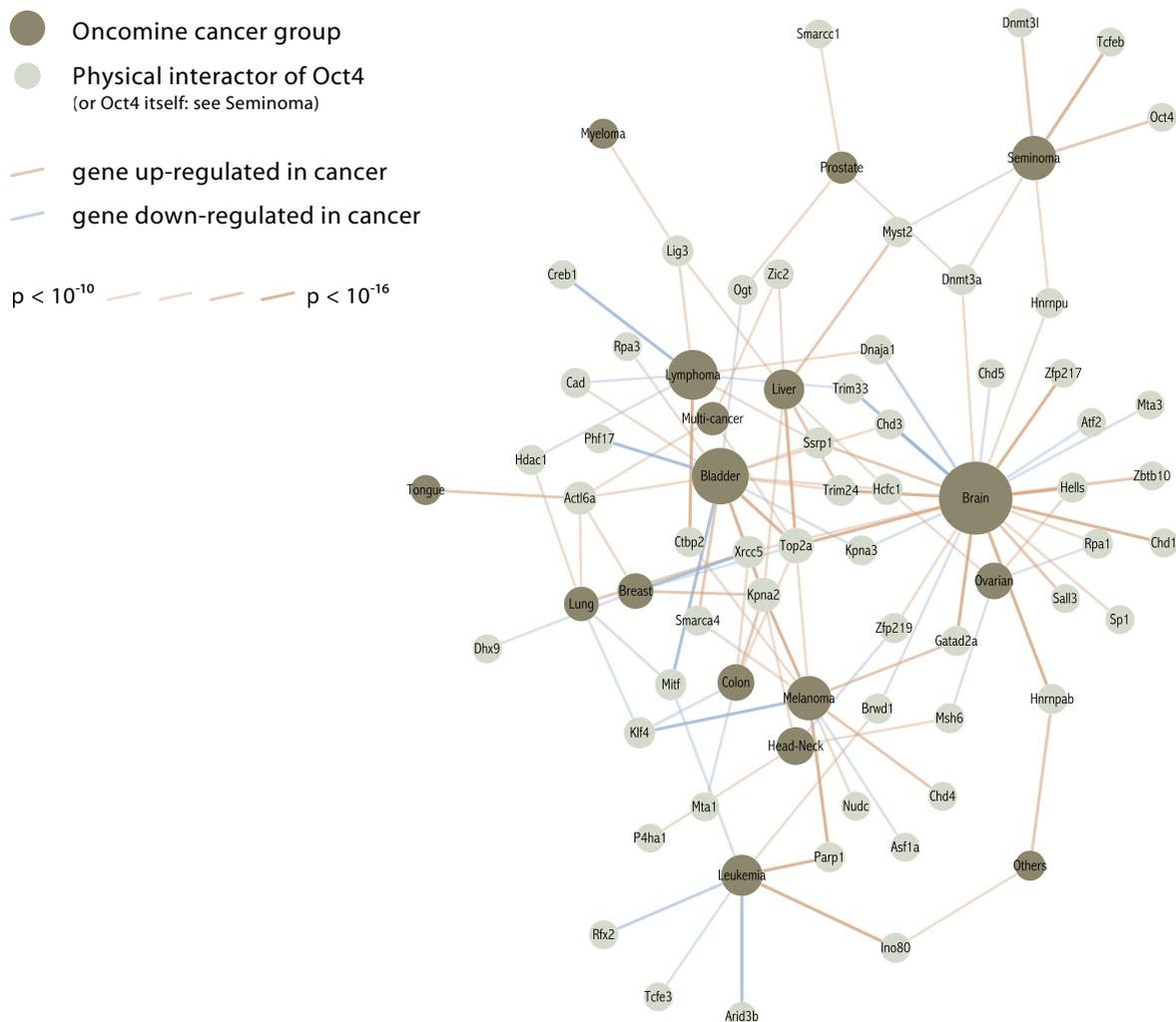


Figure 4.7: Gene misexpression network for the identified Oct4 interactors in a comprehensive set of cancer types. Nodes represent a gene or cancer type, while edges (connections) represent significant misexpression. Edge thickness indicates the misexpression p-value (ranging from $p < 10^{-10}$ to $p < 10^{-16}$).

4.3.8. Functional enrichment according to PANTHER annotation

The Oct4 interactors were enriched in components of the Wnt signalling pathway (Table 4.4), which is known to be important in the achievement and maintenance of stem cell identity (Wend et al., 2010). Additionally, 65% of genes were annotated as nucleic acid-binding, and 33% as transcription factors. This supports the view that the Oct4 interactors could often be downstream transcriptional regulators.

Table 4.4: Functional category analysis for enriched PANTHER terms among Oct4-associated proteins. An analysis for Pathway (PW), Biological Process (BP) and Molecular Function (MF) enrichment was carried out using the mouse genome as the background dataset.

Category	Term	% of Genes	Count	Expected	p-Value
PW	Wnt signaling pathway	8.60%	8	1.3	7.2E-03
BP	Nucleoside, nucleotide and nucleic acid metabolism	73.12%	68	12.0	9.0E-39
BP	mRNA transcription	48.39%	45	7.1	1.2E-23
BP	mRNA transcription regulation	36.56%	34	5.2	7.0E-17
BP	DNA metabolism	13.98%	13	1.0	5.3E-09
BP	Chromatin packaging and remodeling	11.83%	11	0.8	1.2E-07
BP	DNA repair	8.60%	8	0.5	6.1E-06
BP	DNA recombination	4.30%	4	0.1	2.4E-03
BP	Other mRNA transcription	3.23%	3	0.1	6.3E-03
MF	Nucleic acid binding	64.52%	60	11.3	9.0E-31
MF	Transcription factor	33.33%	31	7.3	5.3E-11
MF	Zinc finger transcription factor	17.20%	16	3.3	2.8E-05
MF	Helicase	12.90%	12	0.5	6.5E-11
MF	DNA helicase	11.83%	11	0.2	3.3E-13
MF	Chromatin/chromatin-binding protein	11.83%	11	0.5	8.7E-10
MF	Histone	5.38%	5	0.4	5.5E-03
MF	DNA methyltransferase	3.23%	3	0.1	3.7E-03

4.3.9. Overrepresented domains

Protein domains can give information on functionality, and certain domains are associated with specific contexts, such as chromatin remodelling and gene regulation. Table 4.5 shows an analysis of domain overrepresentation among the Oct4 interactors, relative to the proteome-wide and the nuclear background frequencies.

Table 4.5: Pfam domain occurrence, function and overrepresentation. The number of occurrences of a certain domain within the Oct4 interacting set is given, as well as the number of distinct proteins bearing the domain. Domain functions were obtained from Pfam annotations. Fold enrichment was calculated comparing the domain composition of Oct4 partners to both the nuclear subset and complete set of UniProtKB/Swiss-Prot proteins with subcellular localisation information. This estimates domain overrepresentation in the Oct4-associated group. Only domains significantly enriched are shown.

Pfam domain	Domain description	Oct4 interactors with this domain	cf. nuclear proteins (UniProtKB/Swiss-Prot)		cf. all proteins (UniProtKB/Swiss-Prot)	
			expected	p-Value	expected	p-Value
Helicase_C	Helicase conserved C-terminal domain	10	1.7	2.2E-05	0.5	3.3E-10
SNF2_N	SNF2 family N-terminal domain	9	0.5	3.4E-08	0.1	1.1E-13
zf-C2H2	Zinc finger, C2H2 type	8	8.3	1.000	1.4	9.8E-05
PHD	PHD-finger	6	1.1	0.002	0.4	2.9E-06
ELM2	ELM2 domain	4	0.1	2.5E-05	0.1	1.0E-06
GATA	GATA zinc finger	4	0.4	0.002	0.1	2.1E-06
Chromo	chromo' (CHRromatin Organisation MODifier) domain	4	0.6	0.005	0.1	6.2E-06
Bromodomain	Bromodomain	4	0.5	0.004	0.2	2.9E-05
CHDCT2	CHDCT2 (NUC038) domain	3	0.0	1.1E-04	0.0	3.3E-07
DUF1086	Domain of Unknown Function (DUF1086)	3	0.0	1.1E-04	0.0	3.3E-07
DUF1087	Domain of Unknown Function (DUF1087)	3	0.0	1.1E-04	0.0	3.3E-07
CHDNT	CHDNT (NUC034) domain	3	0.0	1.1E-04	0.0	3.3E-07
DUF3371	Domain of unknown function (DUF3371)	3	0.1	0.000	0.0	6.4E-06
BAH	BAH domain	3	0.1	0.001	0.1	5.1E-05
SAP	SAP domain	3	0.4	0.014	0.1	1.7E-04
Ku	Ku70/Ku80 beta-barrel domain	2	0.0	0.002	0.0	4.8E-05
Ku_N	Ku70/Ku80 N-terminal alpha/beta domain	2	0.0	0.002	0.0	4.8E-05
Ku_C	Ku70/Ku80 C-terminal arm	2	0.0	0.002	0.0	4.8E-05
zf-PARP	Poly(ADP-ribose) polymerase and DNA-Ligase Zn-finger region	2	0.0	0.002	0.0	4.8E-05
2-Hacid_dh_C	D-isomer specific 2-hydroxyacid dehydrogenase, NAD binding domain	2	0.0	0.002	0.0	2.8E-04
2-Hacid_dh	D-isomer specific 2-hydroxyacid dehydrogenase, catalytic domain	2	0.0	0.002	0.0	2.8E-04

4.4. Discussion

The expression of specific sets of regulatory transcription factors is used by the cell to control differentiation (Macarthur et al., 2009). Oct4 is an essential transcription factor which is central to stem cell identity and has been used in the generation of induced pluripotent stem cells from fibroblasts (Takahashi and Yamanaka, 2006). In this study, we aimed to understand the regulatory effects of Oct4 in collaboration with Dr. Jyoti Choudhary's group at the Wellcome Trust Sanger Institute. My

collaborators identified physical interactors of Oct4, and this group of proteins represented an important system for analysis using similar methods as in Chapter 3. We performed an in-depth analysis of the Oct4 protein-protein interaction network based on conservation, transcriptional regulation, expression data and phenotypic data, including implication in human disease.

By integrating these types of information, we found that Oct4 associated with a varied set of 92 proteins, including regulators of gene expression and modulators of Oct4 function. Half of its physical interaction partners are transcriptionally regulated by Oct4 itself or by other stem cell transcription factors (Figure 4.3), and one third display a significant down-regulation in expression upon cell differentiation (Figure 4.2), supporting their functionality in the maintenance of pluripotency. The majority of the Oct4-associated proteins studied to date show early lethal phenotypes when mutated in mice (Figure 4.6), highlighting their importance in development or in basal functions such as DNA repair. A significant fraction of the human orthologs is also associated with inherited disorders (Table 4.2) or causatively mutated in certain cancers (Table 4.3).

Oct4 interacts with several basal transcriptional regulators, such as Ttf2, as well as with proteins involved in DNA replication, recombination and repair, which displayed high expression throughout the ES cells and more differentiated cell types (Fig. 4.2). The alternative pattern shown by one third of the interactors and by Oct4 itself is of a strong switch-like decrease in expression in more differentiated cell types. This specific expression in embryonic stem cells supporting their functionality in the maintenance of pluripotency, supporting their functionality in the maintenance of pluripotency. Among the physical interactors of Oct4, P4ha1 was the only gene which showed increased expression in differentiating cell types, mammospheres and myospheres compared to ES cells. This expression pattern be explained by its function in the post-translational modification of collagen by proline hydroxylation. Its physical interaction with Oct4 in ES cells may perhaps point to additional hydroxylation substrates, whose modification might functionally relate to pluripotency.

Interestingly, in addition to P4ha1, several of the other physical interactors of Oct4 we identified were also writers, readers, erasers or adaptor proteins involved in PTM signalling through lysine acetylation, methylation, phosphorylation, ubiquitination

and O-glycosylation (UniProt Consortium, 2012). Specifically, Brwd1 is a bromodomain lysine acetylation reader protein which interacts with SMARCA4, one of the most highly ranked lysine acetylation-related proteins we identified (Table 3.2). Kat7 is a histone acetyltransferase, while Hdac1 is a histone deacetylase. Kdm1a is a histone demethylase acting on H3K4 and H3K9. Ppp2r1a is a phosphatase adaptor protein, mediating specificity of the catalytic subunit of protein phosphatase 2A. Chd1 and Chd4 are histone lysine methylation reader proteins for H3K4 di- and trimethylation. Trim33 is an E3 ubiquitin ligase, while Ogt is an O-glycosylation writer enzyme, transferring N-acetylglucosamine (O-GlcNAc) to serines and threonines and histones and other proteins. Intriguingly, Oct4 itself is modified by O-GlcNAc in addition to phosphorylation and sumoylation (Wagner and Cooney, 2009; Webster et al., 2009), which suggests that Ogt might be involved in modulating its activity. Ogt is in turn transcriptionally regulated by Oct4 and several other stem cell transcription factors (Fig. 4.3). Oct4 is also modified by phosphorylation and sumoylation (Wagner and Cooney, 2009). Together, these physical interactors of Oct4 may be involved in PTM-based signalling requiring the expression of Oct4, in line with the fact that several layers of regulation such as DNA methylation, chromatin remodelling and various transcription factors are involved in the establishment and maintenance of pluripotency (Hanna et al., 2010).

The presence of Sox2 and Nanog, two well-characterised Oct4 interactors, in only some of the experimental replicates could indicate a degree of transience and potentially point to additional regulation of these interactions. Our study of the Oct4 interactome (Pardo et al., 2010) agreed with a separate independent study on a relatively small number of interactors (19 of the 54 identified in the independent study) (van den Berg et al., 2010). The reason for the observed disparity in interacting proteins may be due to differences in cell lines, conditions, sample preparation, purification techniques or mass spectrometry equipment and data processing. Some interactions between proteins may also be non-specific, which further increases the level of experimental noise (Vermeulen et al., 2008). It therefore becomes important to integrate additional information to independently corroborate an initial list of co-purified proteins, as we have done in this study.

To summarise, the annotated Oct4 interactome presented here provides a resource which can guide further investigation of Oct4's functional context with the aim to illuminate the regulation of pluripotency and differentiation. It may also aid in the

identification of additional factors which could help improve the efficiency and safety of the generation of induced pluripotent stem cells for use in regenerative medicine.

5. Conclusions and implications

In this concluding chapter, I will highlight my most important findings and describe their potential implications, including the role of disruptions of post-translational signalling in human disease, and implications for the engineering of signalling systems in synthetic biology.

5.1. Key findings

5.1.1. Evolution of post-translationally modified residues

Significant selection pressure exists for all PTM types studied

Taken together, our results regarding the constraints on post-translationally modified residues at the species, population and individual levels constitute the finding that there is significant conservation for all six PTM types studied, in both structured and disordered regions of proteins. Between species, conservation at PTM sites was investigated using three alternative scoring methods, and was found to be most apparent in disordered regions even when using normalised scores. Using the original raw scores, the conservation scoring method rvET (a hybrid approach based on phylogeny and symbol entropy) reported significant conservation of all PTM types in both disordered and structured regions. Disordered sites of serine phosphorylation and N-glycosylation as well as structured sites of threonine phosphorylation were consistently reported as conserved by all methods, including normalised scores. Normalisation of the scores per protein generally appeared to decrease detection of PTM site conservation. The less frequent observation of PTM type conservation in structured regions at least when using normalised scores may be due to their higher intrinsic structural constraints, whereas in disordered regions the conservation of specific residues could be more apparent against their fast-evolving background. At the population level, we found PTM sites to be significantly less variable than control residues, and significantly more prone to be affected by disease mutations in individuals. Taken together, our results indicated significant constraints on both structured and disordered sites for all PTM types studied.

Mimetic mutations hint at new signalling systems and may decrease the apparent conservation of PTM sites

Having quantitatively established that significant selection pressure exists on all six PTM types, independent of structural characteristics, we investigated the specifics of residue evolution. By investigating the conservation profiles of PTM sites, we found that while many modified residues appear to have originated before the divergence of the mammalian lineage, a notable fraction of serine and threonine phosphorylation sites appeared to be primate-specific. An analysis of substitutions at PTM sites uncovered evidence of significant mimetic and avoiding mutations at PTM sites, which respectively resemble either the modified or the unmodified state of the residue. Based on the significant overrepresentation of negatively charged amino acid substitutions at lysine acetylation sites observed using some scoring methods, we hypothesised that acylation reactions using negatively charged groups, such as succinylation and malonylation, could alternatively be taking place at these positions, at a frequency sufficient to result in an evolutionary signal. We were also able to confirm a previously reported enrichment in phosphomimetic glutamate substitutions at disordered serine phosphorylation sites (Kurmangaliyev et al., 2011), as well as a similar enrichment of aspartate substitutions for both disordered serine structured tyrosine phosphorylation sites. An example of avoiding mutations was found in the substitution of glutamine for asparagine at disordered N-glycosylation sites using some scoring methods, which is structurally highly similar but does not undergo N-glycosylation. Taken together, we consider these mutations to contribute to a decrease in the degree of conservation observed at PTM sites, despite their functionality.

5.1.2. The human lysine acetylation system

Evolutionary plasticity of a PTM-based signalling system

Our analysis of the enzymes and reader proteins which make up the human lysine acetylation signalling system highlighted its strong conservation across a set of mostly vertebrate species, on par with known essential genes. As a group, the reader (bromodomain) and writer (acetyltransferase) proteins appeared the least well conserved, indicating that they may be the more evolvable components of the

system compared to the deacetylases. Several other interesting aspects were found in the course of the analyses. As in the phosphorylation and ubiquitination systems, the proportion of erasing enzymes was small compared to the reader and writer categories. This may either indicate that erasers tend to primarily maintain an equilibrium of acetylation and deacetylation at a large number of substrate sites, while the writers might carry out primary signalling at a more limited set of sites, or it may indicate that a fraction of acetylated lysines remain permanently modified. The conservation profiles also highlighted a limited number of genes which were conserved in nearly all species studied, as well as evidence of lineage-specific family expansions presumably by gene duplication. Similar expansions have been described in the lysine methylation system (Aravind et al., 2011). We also developed a scoring method to prioritise proteins likely to be of particular importance in lysine acetylation signalling. This method rewarded candidate proteins with high sequence conservation among vertebrates, a large number of known protein-protein interactions, co-expression with many other lysine acetylation-related proteins and expression in a large number of cell types. The resulting ranking showed that based on their localisation, most proteins which fulfil these criteria appear to be associated at least partially with nuclear functions, which may support an ancestral role for lysine acetylation in gene regulation via histone modification which was followed by functional diversification. This is further supported by the set of genes which we identified as conserved in nearly all species studied, which includes histone deacetylases HDAC1–3 and the elongator complex protein ELP3, a histone acetyltransferase, among others.

[A generalised framework for the study of PTM signalling systems](#)

Beginning with an initial list of proteins of interest in three functional classes, the analyses performed here build on publicly available information sources which cover most of the human proteome. Our study therefore demonstrates the application of a generalised, transferrable framework for the analysis of human signalling systems. We envision that this analysis framework could prove useful in identifying the central components, and therefore the physiological functionality, of newly described signalling systems such as lysine succinylation, once the biochemical characterisation of relevant protein domains has progressed slightly further (Lin et al., 2012).

5.1.3. The interaction network of Oct4

Identification of novel Oct4 interactors

We experimentally identified a set of 92 proteins which were co-purified with Oct4, thereby expanding the number of known Oct4 interactors at least three-fold (Pardo et al., 2010). Half of these interaction partners were transcriptionally regulated by Oct4 or by other stem cell transcription factors, and one third displayed significant down-regulation in expression upon ES cell differentiation, supporting their functionality in the maintenance of pluripotency. Many also showed lethal phenotypes when mutated in mice, or association with genetic disease and cancer in humans, highlighting their essentiality in development, transcriptional regulation and in basal functions such as DNA repair.

Oct4 is regulated via O-glycosylation by Ogt

Several of the physical interactors of Oct4 we identified are writers, readers, erasers and adaptor proteins involved in PTM signalling, which may have specific signalling roles when Oct4 is expressed. Both our own study of the Oct4 interactome and an independent study identified an interaction between Oct4 and Ogt, an N-acetylglucosamine (O-GlcNAc) transferase (Pardo et al., 2010; van den Berg et al., 2010). It had previously been found that Oct4 is modified by O-GlcNAc, and that its modification is important in zebrafish development (Webster et al., 2009). More recently, it has been found that O-glycosylation by Ogt causes Oct4 to upregulate several target genes with importance in embryonic stem cell identity, and that this modification is dynamically counteracted by an O-glycosylase, Oga (Jang et al., 2012). In differentiation, the interaction between Ogt and Oct4 is rapidly lost along with expression of the target genes (Jang et al., 2012). According to our findings, Ogt is not significantly down-regulated upon differentiation, suggesting that its decrease in activity on Oct4 might be mediated by another signalling mechanism (Fig. 4.2).

5.2. Future directions

Our studies of the lysine acetylation signalling system and of the Oct4 interactome illustrate the value of data integration approaches for identifying important

components in biological systems, beginning from a list of candidates. This information can now be used to prioritise promising functional aspects and individual proteins for further study, as in the example of Ogt. Ultimately, I hope the results presented will prove useful in identifying new pharmacological targets in diseases such as cancer.

5.2.1. Investigating the signalling functions of newly discovered PTMs

Extracting biological meaning from large-scale datasets is a non-trivial ongoing challenge. The key appears to be the emergent knowledge that results from integrating multiple sets of related, complementary and mutually enhancing information. Today, more than 300 types of post-translational modification are known to occur physiologically in the cell (Witze et al., 2007). Meanwhile, unexplained spectra are still routinely being observed in mass spectrometry studies, potentially pointing to as yet unidentified modifications (Bischoff and Schlüter, 2012). Serine and threonine acetylation have only recently been discovered (Mittal et al., 2006; Mukherjee et al., 2006), as have aspartate and glutamate methylation (Sprung et al., 2008), followed by lysine succinylation, malonylation, crotonylation and various other types of acylation (Zhang et al., 2010c; Du et al., 2011; Tan et al., 2011; Lin et al., 2012; Xie et al., 2012). Several of these have been shown to be reversible and to differ in their occurrence between cell lines, lending them promise in dynamic signalling roles. An in-depth analysis of the already intensely studied histone proteins recently found 67 previously undescribed modification sites, including lysine formylation, tyrosine hydroxylation and lysine crotonylation sites (Tan et al., 2011). It was further shown that lysine crotonylation appears to be a transcriptionally activating signal, and that it is evolutionarily conserved in several species including yeast, mouse and human.

A large number of additional and atypical histone modifications have recently been discovered by in-depth mass spectrometry studies, and their functional elucidation is far from complete (Muers, 2011; Tan et al., 2011; Tweedie-Cullen et al., 2012). The fact that these modifications are still being discovered in a regulatory system as intensively studied as chromatin, along with indications of their dynamic regulation and functionality, suggests that atypical PTMs also remain undiscovered in other important signalling systems. For instance, serine and threonine acetylation of a

MAPK kinase by a pathogenic bacterium, *Yersinia pestis*, have been described, and these interfered with phosphorylation of the same residues (Mittal et al., 2006; Mukherjee et al., 2006). Additional reports of serine and threonine acetylation have been made on kinases in yeast (Zhang et al., 2010a). It remains to be investigated whether acetylation of serine and threonine phosphorylation sites could be a more widespread endogenous mechanism of regulating phosphorylation sites through cross-talk with acetylation. Extensive competition for substrate residues has also been documented between ubiquitination and lysine acetylation, where acetylation could extend protein half-life by preventing polyubiquitination, among other effects (Grönroos et al., 2002; Jin et al., 2004; Caron et al., 2005; Danielsen et al., 2010; Wagner et al., 2011). The direct competition of multiple PTM types for individual residues further adds to the number of combinatorial states of PTM sites in a protein, and is likely to be an important area for future research on cellular information processing and decision making.

The presence of moonlighting enzymes that are able to modify proteins in addition to other canonical role further hints that our current picture of PTM signalling is still incomplete (Bansal et al., 2008; Baghel et al., 2011). At least in prokaryotes, not all acetylation events appear to require acetyl-CoA, since alternative enzymes which use acetoxycoumarins as acetyl group donors have been shown to exist (Bansal et al., 2008).

In plants, insects and mammals, there have been reports that proteins, calcium signals, vesicles and entire organelles can be exchanged between cells, either through plasmodesmata or tunnelling nanotubes (Rustom et al., 2004; Gallagher and Benfey, 2005; Chauveau et al., 2010). It is tempting to speculate that post-translational signalling systems might sometimes extend directly between adjacent cells, potentially through the exchange of donor cofactors such as acetyl-CoA, or through modified proteins or peptides, which might be passed along these channels as intercellular signals. Similar systems exist for the propagation of RNAi-based antiviral defence activation in plants (Garcia-Ruiz et al., 2010; Uddin and Kim, 2011), and in the form of neuropeptides and small peptide hormones in animals (Husson et al., 2005; Reiher et al., 2011).

These discoveries, combined with the results on PTM site conservation and mimetic mutations presented here, point to a large component of cellular signalling that still

remains to be explored, and which may have implications in human disease. It therefore appears clear that there is scope for new discoveries in the analysis of post-translational signalling systems. As illustrated by a very recent study of the genome-reduced parasitic bacterium *Mycoplasma pneumoniae*, extensive bi-directional cross-talk exists between lysine acetylation and phosphorylation even in simple organisms (van Noort et al., 2012). This also points to the fact that PTM systems should not only be studied in isolation if their functions are to be elucidated. In addition to post-translational modifications, modifications at the RNA level might also serve crucial functions (Kim et al., 2010). Like proteins, DNA and RNA have also been found to be enzymatically modified, with regulatory effects on transcription, RNA editing and the subcellular localisation of specific tRNAs (Kaneko et al., 2003; Iyer et al., 2009). One common example is DNA methylation in CpG islands, which has a repressive effect on transcription (Iyer et al., 2009). Another example is the hydroxymethylation of cytosine by the TET enzyme, which may have an epigenetic role in ES cells (Tahiliani et al., 2009). The complexities of mammalian translational processes, including alternative translation initiation sites, are still far from understood (Ingolia et al., 2011). The same is true for alternative splicing and alternative transcription, which contribute to diversity at the molecular level, between cell types, and between organismal phenotypes (Gan et al., 2011; McManus and Graveley, 2011; Pal et al., 2011; Barrie et al., 2012). By integrating these layers of signalling into models, we may increasingly be able to explain aspects which may currently be considered biological noise, but may in fact arise from complexity. However, certain types of biological noise may also be adaptive at the population level, as it can function to generate phenotypic diversity (Veening et al., 2008; Fraser and Kaern, 2009; Chalancon et al., 2012). At the single cell level, expression noise of stem cell transcription factors has been described to contribute to the maintenance of pluripotency in ES cells (Trott et al., 2012).

5.2.2. Promising experimental methods

The ability to synthesise proteins with defined modifications through the incorporation of stop codon-encoded unnatural amino acids should prove invaluable to the study of the effects of individual PTMs (Liu et al., 2010b; Chin, 2011). By expanding the genetic code, site-specific encoding of PTMs such as ubiquitination, phosphorylation and others has been achieved in *E. coli*, *S. cerevisiae*, *C. elegans*

and *Drosophila* (Hancock et al., 2010; Neumann et al., 2010; Greiss and Chin, 2011; Virdee et al., 2011; Bianco et al., 2012; Chin, 2012; Davis and Chin, 2012). Additionally, the ability to generate phosphorylated peptides based on the presence of a leader sequence by using engineered class II lantipeptide synthetases in *E. coli* is promising (Thibodeaux and van der Donk, 2012). Another method which should prove extremely useful for the precisely targeted perturbation of PTM signalling is the use of genetically encoded photocaged residues. The caging groups on these residues can block their function until removed by light. Photocontrol of light-switchable serine and tyrosine kinases (Gautier et al., 2011; Arbely et al., 2012), of protein-protein interactions (Levskaia et al., 2009) and of protein localisation to the nucleus (Gautier et al., 2010) have been demonstrated.

Applying mass spectrometry to intact proteins followed by controlled fragmentation within the instrument (“top-down” mass spectrometry) offers an interesting alternative to the more conventional “bottom-up” fragmentation–identification approaches which are most widely used and generally offer higher rates of protein identification (Bogdanov and Smith, 2005; Yates et al., 2009). However, using intact proteins opens up the investigation of different protein “modforms”, i.e. the combinatorial states of multiple PTMs within the same protein (Siuti and Kelleher, 2007; Tweedie-Cullen et al., 2012). This type of analysis has already been demonstrated for individual histones (Tweedie-Cullen et al., 2012), and appears promising for the future combinatorial study of protein modifications, which have been described as akin to a “molecular barcode” in terms of information content (Benayoun and Veitia, 2009).

5.2.3. Limitations to be addressed

A subject that remains to be studied experimentally in detail is the plasticity of PTM sites between individuals and between related species. To a degree, this has been attempted in this dissertation, but our analyses also illustrated that experimental coverage of PTM sites remains very low outside of the standard model organisms (Fig. 2.4). A significant potential aspect of divergence between species therefore remains unexplored at present. To further highlight this, it has been shown that glycosylation involving sialic acids has uniquely diverged in humans, with medical implications, amid other reports illustrating that significant signalling differences

exist even between species as closely related as humans and chimpanzees (Varki and Altheide, 2005; Varki and Nelson, 2007; Varki et al., 2008; Varki, 2010; Varki et al., 2011). Glycan divergence on cell surfaces, possibly driven by pathogens, has been described as a possible important contributor to the speciation process itself (Varki, 2006). Integrating species-, cell type- and condition-specific knowledge on PTM signalling networks with protein-protein interaction networks is therefore a promising approach to achieve greater clarity in the understanding of signalling by PTMs (Minguez et al., 2012).

The precise concentration of the transcription factor Oct4 is key in regulating an embryonic stem cell's decision between differentiation and pluripotency (Okita and Yamanaka, 2010). In differentiation, the directionality of Oct4 expression change affects the type of differentiation: an increased expression level leads to ES cell differentiation into trophoblast, while a decreased level leads to differentiation into mesoderm or endoderm. Similarly, it is likely that the fraction of proteins which are modified (PTM site occupancy), as well as the concentrations of writer, reader and eraser proteins, could be important in post-translational signalling. In yeast, it has been described that sites of low occupancy are less well conserved (Levy et al., 2012). Large-scale studies on human cells would be useful to address this, and SILAC has been demonstrated as a quantification method for fractional occupancy of phosphorylation (Olsen et al., 2010). SILAC has also been used for measurement of the proteome-wide modification state changes induced upon perturbation of either a class of PTM enzymes (Pan et al., 2009), and of a single PTM enzyme (Hilger et al., 2009). In future, quantitative methods for the analysis of dynamic systems may then prove useful in analysing large-scale PTM signalling networks, especially those where a correlation exists with a donor molecule such as acetyl-CoA (Kim and Motter, 2009; Motter, 2010; Nishikawa and Motter, 2010).

Our study of PTM site evolution had two limitations which are non-trivial to address at present. The first relates to alternative splicing, which is very common in mammals. 92–94% of human genes undergo alternative splicing, and 86% of these display a minor isoform frequency of 15% or more (Wang et al., 2008). In this study, we avoided this additional level of complexity by using the canonical, usually longest transcript as recorded in the UniProtKB/Swiss-Prot database. This was done since the transfer of modification site annotation to alternative isoforms could not be done with confidence. In future, it would be interesting to investigate whether alternatively

spliced exons display an increase in the number of PTM sites they encode. Tissue-specific exons have been found to contain an enrichment of linear motifs and to have important roles in rewiring interaction networks (Buljan et al., 2012; Davis et al., 2012; Ellis et al., 2012; Weatheritt and Gibson, 2012; Weatheritt et al., 2012). The inclusion or exclusion of a PTM site through alternative splicing could therefore represent an effective mechanism for the modulation of signalling.

Secondly, efforts have been undertaken to dissect the dynamic nature of PTM signalling networks (Price et al., 2010), and to establish quantitative links between the concentrations of modifying and erasing enzymes, and the concentrations of individual modified forms (“modforms”) of proteins (Thomson and Gunawardena, 2009). However, due to the limitations that accompanied the collection of known PTM sites from a large number of separate studies, experiments and conditions, we have had to consider PTM sites as mostly static and mutually independent for the purposes of this dissertation. Hopefully, large-scale information on the conditions, tissues, and combinations in which specific PTMs occur together will become available in future, which would allow a global elucidation of combinatorial PTM signalling.

5.2.4. Engineering signalling systems in synthetic biology

In future, the design of complex post-translational signalling circuits for information processing is envisioned in synthetic biology, which is becoming increasingly standardised and efficient (Anderson et al., 2010). The introduction of new signalling circuits might be done especially easily by transgenically introducing systems which may be endogenously absent (Lim, 2010). Eventually, these methods might allow a ground-up reconstruction of biological systems from parts whose characteristics we understand (Hockenberry and Jewett, 2012), which may finally put biology on an even footing with the approaches used in chemistry and physics: reductionism followed by engineering (Bashor et al., 2010). Ultimately, this may be the only approach which can fully allow us to understand how the beautiful complexity of biological systems arises from their smallest parts.

6. Publications

6.1. Relevant publications

An expanded Oct4 interaction network: Implications for stem cell biology, development, and disease. *Pardo, M.**, **Lang, B.***, Yu, L., Prosser, H., Bradley, A., Babu, M.M.**, Choudhary, J.** *Cell Stem Cell* (2010) vol. 6 (4) pp. 382-95.

*, **: Contributed equally.

6.2. Manuscripts in preparation

Selection pressure on primate-specific post-translational modification sites and their relevance in human disease. **Lang, B.**, Chavali, S., Babu, M.M. (2012, in preparation)

Prioritising components of the human lysine acetylation system for further study through network analysis. **Lang, B.**, Babu, M.M. (2012, in preparation)

Evolution and network topologies of post-translational signalling systems. **Lang, B.**, Chavali, S., Babu, M.M. (2012, in preparation)

The role of intrinsic disorder in protein turnover and the evolution of protein half-life. *van der Lee, R.*, **Lang, B.**, Kruse, K., Gsponer, J., Sánchez de Groot, N., Fuxreiter, M., Babu, M.M. (2012, in preparation)

6.3. Additional publications

Evolutionary selection for protein aggregation. *Sanchez de Groot, N.*, *Torrent, M.*, *Villar-Piqué, A.*, **Lang, B.**, *Ventura, S.*, *Gsponer, J.*, *Babu, M.M.* *Biochemical Society Transactions* (2012) vol. 40 (5) pp. 1032-7.

Proteomic analysis of mitotic RNA polymerase II reveals novel interactors and association with proteins dysfunctional in disease. *Moeller, A.*, *Xie, S.Q.*, *Hosp, F.*, **Lang, B.**, *Phatnani, H.P.*, *James, S.*, *Ramirez, F.*, *Collin, G.B.*, *Naggert, J.K.*, *Babu, M.M.*, *Greenleaf, A.L.*, *Selbach, M.* and *Pombo A.* *Molecular & Cellular Proteomics* (2012) vol. 11 (6): M111.011767.

Methods to Reconstruct and Compare Transcriptional Regulatory Networks. *Babu, M.M., Lang, B. and Aravind, L. Methods in Molecular Biology (2009) vol. 541 pp. 163-80.*

High affinity DNA binding sites for H-NS provide a molecular basis for selective silencing within proteobacterial genomes. *Lang, B., Blot, N., Bouffartigues, E., Buckle, M., Geertz, M., Gualerzi, C., Mavathur, R., Muskhelshvili, G., Pon, C., Rimsky, S., Stella, S., Babu, M.M. and Travers, A. Nucleic Acids Res (2007) vol. 35 (18) pp. 6330-7.*

An Expanded Oct4 Interaction Network: Implications for Stem Cell Biology, Development, and Disease

Mercedes Pardo,^{1,4,*} Benjamin Lang,^{3,4} Lu Yu,¹ Haydn Prosser,² Allan Bradley,² M. Madan Babu,^{3,5} and Jyoti Choudhary^{1,5,*}

¹Proteomic Mass Spectrometry, Wellcome Trust Sanger Institute, Hinxton, Cambridgeshire CB10 1SA, UK

²Mouse Genomics, Wellcome Trust Sanger Institute, Hinxton, Cambridgeshire CB10 1SA, UK

³MRC Laboratory of Molecular Biology, Cambridge CB2 0QH, UK

⁴These authors contributed equally to this work

⁵These authors contributed equally to this work

*Correspondence: mp3@sanger.ac.uk (M.P.), jc4@sanger.ac.uk (J.C.)

DOI 10.1016/j.stem.2010.03.004

SUMMARY

The transcription factor Oct4 is key in embryonic stem cell identity and reprogramming. Insight into its partners should illuminate how the pluripotent state is established and regulated. Here, we identify a considerably expanded set of Oct4-binding proteins in mouse embryonic stem cells. We find that Oct4 associates with a varied set of proteins including regulators of gene expression and modulators of Oct4 function. Half of its partners are transcriptionally regulated by Oct4 itself or other stem cell transcription factors, whereas one-third display a significant change in expression upon cell differentiation. The majority of Oct4-associated proteins studied to date show an early lethal phenotype when mutated. A fraction of the human orthologs is associated with inherited developmental disorders or causative of cancer. The Oct4 interactome provides a resource for dissecting mechanisms of Oct4 function, enlightening the basis of pluripotency and development, and identifying potential additional reprogramming factors.

INTRODUCTION

Two characteristics define embryonic stem cells (ESCs), self-renewal ability and pluripotency. Recently, ectopic expression of combinations of transcription factors (Oct4, Nanog, Sox2, c-Myc, Esrrb, and Klf4) has been shown to reprogram mouse and human fibroblasts into a pluripotent state (Kaji et al., 2009; Okita et al., 2007; Takahashi et al., 2007; Takahashi and Yamanaka, 2006; Woltjen et al., 2009; Yu et al., 2007). The induced pluripotent stem cells (iPSCs) are very similar to ESCs and retain the ability to self-renew and differentiate into the three germ layers and thus promise great therapeutic potential in regenerative medicine (Amabile and Meissner, 2009; Maherli et al., 2007; Wernig et al., 2007). Despite the recent flurry of studies, our understanding of the molecular mechanisms and players that drive ESC self-renewal and differentiation is still limited.

The POU transcription factor Oct4, also termed Pou5f1, is a central player in ESC self-renewal and differentiation into specific lineages. Levels of Oct4 must be tightly regulated to

maintain the ESC status. A decrease in Oct4 levels by 50% induces differentiation toward the trophectoderm lineage, whereas a 50% increase causes differentiation into mesoderm and endoderm (Niwa et al., 2000; Shimosaki et al., 2003). Oct4 plays an essential role in early development given that loss of Oct4 in the mouse embryo causes the failure of the inner cell mass to develop (Nichols et al., 1998).

Oct4 regulates transcriptional programs to maintain ESC pluripotency primarily in collaboration with transcription factors Sox2 and Nanog (Boyer et al., 2005; Chew et al., 2005; Pan et al., 2006). Several genome-wide analyses of regulatory targets of key pluripotency factors has led to the identification of sets of jointly regulated or bound targets, highlighting a complex transcriptional circuitry responsible for ESC maintenance (Babaie et al., 2007; Boyer et al., 2005; Ivanova et al., 2006; Kim et al., 2008; Loh et al., 2006; Matoba et al., 2006). Also recently, various other factors have been functionally linked to Oct4 and Nanog, after identification of their binding partners by affinity purification and mass spectrometry (Wang et al., 2006; Liang et al., 2008). These studies have revealed a compact regulatory module responsible for ESC pluripotency (Orkin et al., 2008).

To further elucidate the ESC transcriptional network, we have carried out an unbiased and extensive study of Oct4-associated proteins, using an affinity purification and mass spectrometry approach. In contrast with a previous similar study (Wang et al., 2006), epitope-tagged Oct4 was expressed under the control of Oct4's endogenous promoter to keep the natural transcriptional regulation. The epitope tagging strategy circumvents the need for specific antibodies and facilitates a generic purification procedure that results in cleaner and higher yield samples than traditional immunoprecipitation experiments. Our data significantly expands the current repertoire of Oct4-associated proteins, thereby shedding light on the complex regulatory circuitries of ESCs. The Oct4 interactome provides a useful resource to investigate the mechanisms of Oct4 function and regulation and to explore the basic principles underlying stem cell biology.

RESULTS

Efficient FTAP Tagging of Oct4 by Recombineering and Single-Copy BAC Transgenesis

To investigate the molecular network around Oct4/Pou5f1, we used an epitope-tagging affinity purification strategy.

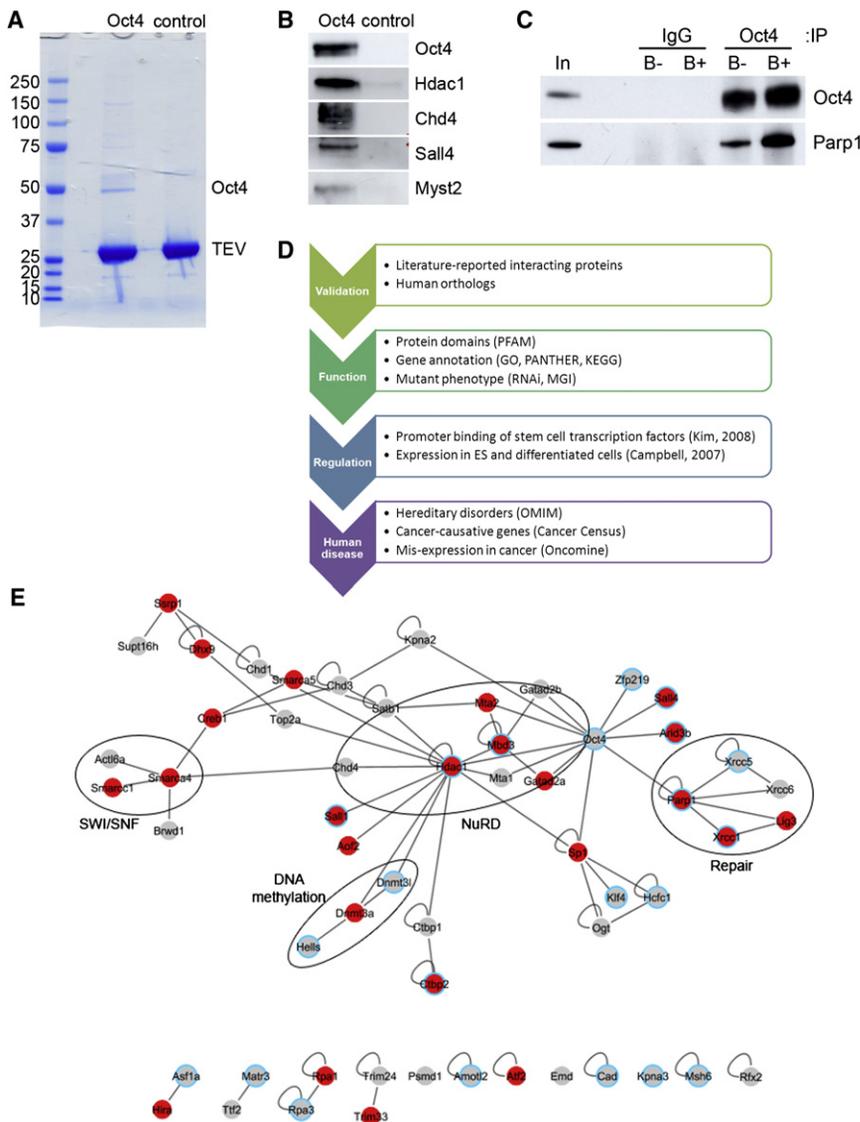


Figure 1. Analysis of Oct4-Interacting Proteins

(A) Typical Oct4-FTAP and control purifications. Molecular weight markers (kDa) are shown. (B) Western blots confirming some of the interacting proteins identified by mass spectrometry. (C) Western blot showing co-immunoprecipitation of endogenous Parp1 with Oct4 in the presence (B+) or absence (B-) of benzamide. In denotes whole cell extract. (D) Workflow tracing the systems analyses of the Oct4 interactome. (E) Network of protein-protein interactions within the Oct4 dataset. Blue circles are proteins downregulated upon ES cell differentiation. Red fill indicates proteins whose absence results in embryonic lethality in the mouse. See also Figure S1.

ifications on whole-cell extracts from both Oct4-FTAP-expressing and control unmodified cells (Figure 1A). Eluates were separated by gel electrophoresis, and whole lanes were excised into several regions, digested, and analyzed by nano-liquid chromatography/tandem mass spectrometry (LC-MS/MS). MS results files from each lane were merged and searched against IPI with Mascot. The data is available in the PRIDE database (Martens et al., 2005) (www.ebi.ac.uk/pride). The data was converted with the PRIDE Converter (Barsnes et al., 2009) (<http://code.google.com/p/pride-converter>). The criteria for peptide and protein identification are detailed in Experimental Procedures. Mass spectrometry analysis resulted in the identification of 92 proteins (excluding Oct4 itself) that were present in all Oct4-FTAP purifications, but not in controls (Table 1).

We modified the SPA tag (Zeghouf et al., 2004) containing the 3x FLAG epitope and a calmodulin binding peptide (CBP) separated by a TEV cleavage site, by adding an extra TEV site to improve cleavage efficiency (Figure S1A available online). The FTAP was fused at the C terminus of the Oct4 coding region by recombining into a BAC clone containing full-length Oct4. This was then integrated into the *Hprt* locus of ESCs by recombinase-mediated cassette exchange (RMCE) (Prosser et al., 2008). The whole procedure is depicted in Figure S1B. Expression levels of the Oct4-FTAP fusion protein were ~30% that of endogenous Oct4 expressed from two alleles (Figure S1C), close to what should be expected given that it is expressed from an extra copy of the gene and avoiding interference with the ESC phenotype, as shown by the expression of ESC markers by the transgenic clone (Figure S1C).

Identification of Oct4-Associated Proteins

The tandem affinity tag allows single- and double-affinity purifications. We first performed three independent one-step purifi-

ing proteins was confirmed by Western blotting (Figure 1B). These data considerably expand the list of published Oct4 binding partners and represent a major extension of the sets reported in two similar studies (Liang et al., 2008; Wang et al., 2006). We detected 13 previously identified Oct4 interacting proteins in our study (Table S6). These included Sall4, Arid3b, Zfp219, and Sp1 (Wang et al., 2006), Kpna2 (Li et al., 2008), Parp1 (Gao et al., 2009), and NuRD complex members Hdac1, Mta1/2, and Gatad2a/b (Liang et al., 2008; Wang et al., 2006). Furthermore, we also identified Sox2 and Nanog, two of the best characterized Oct4 binding partners (Ambrosetti et al., 1997; Chew et al., 2005; Liang et al., 2008; Wang et al., 2006), and Zfp281, Requiem/Dpf2, Yy1, RYBP, Dax1, Esrrb, and Arid3a, recently shown to physically interact with Oct4 (Donohoe et al., 2009; Sun et al., 2009; van den Berg et al., 2008; Wang et al., 2006; Wang et al., 2008) in one or two (Arid3a and Esrrb) purifications, but because of our strict criteria of result reproducibility, we did not include them in the final data set. We also identified proteins reported to be linked to Oct4 through association

Table 1. Oct4-Associated Proteins Classified into Protein Complexes and/or Functional Categories

Complex/ Protein Class	Gene Name	Accession	Description	MW	Exp I	Exp II	Exp III
Bait	Pou5f1	IPI00117218	POU domain, class 5, transcription factor 1	38705.35	20	19	21
NuRD Complex							
	Chd4	IPI00396802	chromodomain-helicase-DNA-binding protein 4	219096.34	21	40	33
	Gatad2a	IPI00625995	p66 alpha isoform a	67762.39	13	11	16
	Gatad2b	IPI00128615	isoform 1 of transcriptional repressor p66-beta	65712.08	13	7	16
	Mbd3	IPI00131067	isoform 1 of methyl-CpG-binding domain protein 3	32168	4	6	4
	Mta1	IPI00624969	Mta1 protein	80019.75	13	14	16
	Mta2	IPI00128230	metastasis-associated protein MTA2	75723.93	25	17	21
	Mta3	IPI00125745	isoform 1 of metastasis-associated protein MTA3	67719.08	10	9	7
	Hdac1	IPI00114232	histone deacetylase 1	55609.93	11	12	11
Spalt-like Transcriptional Repressors							
	Sall1	IPI00342267	Sal-like 1	141745.1	11	17	18
	Sall3	IPI00123404	isoform 1 of Sal-like protein 3	140610.62	5	5	6
	Sall4	IPI00475164	isoform 1 of Sal-like protein 4	114711.29	35	29	28
BAF Complex							
	Smarca4	IPI00875789	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily a, member 4, isoform CRA_b	181913.68	3	5	1
	Smarcc1	IPI00125662	isoform 1 of SWI/SNF complex subunit SMARCC1	123326.28	7	3	4
	Actl6a	IPI00323660	actin-like protein 6A	47930.54	4	1	1
FACT Complex							
	Ssrp1	IPI00407571	isoform 2 of FACT complex subunit SSRP1	81766.73	36	12	32
	Supt16h	IPI00120344	FACT complex subunit SPT16	120319.5	64	34	56
LSD1 Complex							
	Aof2	IPI00648295	amine oxidase (Flavin containing) domain 2	95113.5	9	6	7
	Rcor2	IPI00226581	REST corepressor 2	58042.89	5	6	2
ISWI Chromatin Remodeling Complex	Smarca5	IPI00396739	SWI/SNF-related matrix-associated actin-dependent regulator of chromatin subfamily A member 5	122291.43	5	7	2
INO80 Chromatin-Remodeling Complex							
	Ino80	IPI00378561	isoform 1 of putative DNA helicase INO80 complex homolog 1	177265.25	5	6	1
	Nfrkb	IPI00274469	nuclear factor related to kappa-B-binding protein	139134.5	22	11	4
	Actl6a	IPI00323660	actin-like protein 6A	47930.54	4	1	1
Histone Chaperone Complex							
	Asf1a	IPI00132452	histone chaperone ASF1A	23099.19	3	1	3
	Cabin1	IPI00380107	calcineurin binding protein 1	245584.2	8	18	7
	Hira	IPI00123694	isoform long of protein HIRA	113235.4	10	7	10
	Ubn2	IPI00854896	isoform 4 of uncharacterized protein KIAA2030	142523.51	17	19	11
Transcription Factors							
	Arid3b	IPI00277032	isoform 1 of AT-rich interactive domain-containing protein 3B	61091.99	9	8	3
	Atf2	IPI00110172	isoform 1 of cyclic AMP-dependent transcription factor ATF-2	52550.73	4	5	5
	Creb1	IPI00119924	isoform 1 of cAMP response element-binding protein	36879.65	3	3	3
	Ctbp1	IPI00128155	isoform 1 of C-terminal-binding protein 1	48170.7	9	4	3
	Ctbp2	IPI00856974	isoform 2 of C-terminal-binding protein 2	107801.19	10	3	4
	Klf4	IPI00120384	Kruppel-like factor 4	52531.81	1	1	2

Table 1. Continued

Complex/ Protein Class	Gene Name	Accession	Description	MW	Exp I	Exp II	Exp III
	Mitf	IPI00125758	isoform A of microphthalmia-associated transcription factor	59160.39	2	3	1
	Nfyc	IPI00108204	nuclear transcription factor Y subunit gamma	37230.86	7	3	1
	Sp1	IPI00323887	isoform 1 of transcription factor Sp1	81309.84	5	4	6
	Tcfe3	IPI00380308	isoform 1 of transcription factor E3	61555.7	9	5	7
	Tcfef	IPI00314502	transcription factor EB	52638.18	2	3	2
	Zbtb10	IPI00223276	zinc finger and BTB domain containing 10 isoform 1	118071.48	9	5	12
	Zbtb2	IPI00652356	putative uncharacterized protein	58189.61	1	1	1
	Zbtb43	IPI00230530	zinc finger protein 297B isoform a	57551.37	2	1	3
	Zfhx3	IPI00475055	AT motif binding factor 1	410697.11	3	36	1
	Zfp217	IPI00758403	zinc finger protein 217	115181.65	11	10	4
	Zfp219	IPI00469594	zinc finger protein 219, isoform CRA_a	78831.33	4	6	7
	Zfp513	IPI00830836	isoform 1 of Zinc finger protein 513	59968.12	1	2	1
	Zic2	IPI00127145	zinc finger protein ZIC 2	55546.36	1	1	1
	Zscan4b	IPI00755380	similar to Gene model 397	58667.99	4	6	7
Regulation of Transcription							
	Acin1	IPI00121136	isoform 1 of apoptotic chromatin condensation inducer in the nucleus	151000.03	12	2	13
	Brwd1	IPI00121655	isoform A of bromodomain and WD repeat-containing protein 1	262057.12	1	2	2
	Hcfc1	IPI00828490	host cell factor C1	216798.82	19	8	12
	Ifi202b	IPI00126725	interferon-activable protein 202	50727.44	4	6	3
	Phf17	IPI00453799	isoform 1 of Protein Jade-1	95434.25	4	2	4
	Rfx2	IPI00406298	DNA-binding protein RFX2	76998.5	8	1	1
General Transcription	Ttf2	IPI00112371	transcription termination factor 2	126706.43	5	2	2
Recombination/Repair							
	Lig3	IPI00124272	isoform Alpha of DNA ligase 3	114656.59	5	6	1
	Msh6	IPI00310173	MutS homolog 6	152813.37	10	4	5
	Parp1	IPI00139168	putative uncharacterized protein	113491.6	24	17	33
	Top2a	IPI00122223	DNA topoisomerase 2-alpha	173567.4	30	44	12
	Xrcc1	IPI00118139	DNA repair protein XRCC1	69270.68	6	3	1
	Xrcc5	IPI00321154	ATP-dependent DNA helicase 2 subunit 2	83802.29	6	5	13
	Xrcc6	IPI00132424	ATP-dependent DNA helicase 2 subunit 1	69726.04	12	5	12
Replication							
	Rpa1	IPI00124520	replication protein A 70 kDa DNA-binding subunit	69620.87	2	3	2
	Rpa3	IPI00132128	replication protein A 14 kDa subunit	13688.99	1	1	1
Helicases							
	Chd1	IPI00107999	chromodomain-helicase-DNA-binding protein 1	197601.13	8	9	7
	Chd3	IPI00675483	chromodomain helicase DNA binding protein 3	234196.54	5	6	6
	Chd5	IPI00875673	chromodomain helicase DNA binding protein 5 isoform1	224027.21	10	9	11
	Dhx9	IPI00339468	isoform 2 of ATP-dependent RNA helicase A	150907.1	15	15	31
	Hells	IPI00121431	isoform 1 of lymphocyte-specific helicase	95806.47	3	1	1
Histones							
	Hist1h3e	IPI00282848	histone cluster 2, H3c1 isoform 2	20348.12	8	9	4
	Hist1h4b	IPI00407339	histone H4	11360.38	11	13	6
	Hist3h2bb	IPI00229539	histone cluster 3, H2bb	17248.15	9	7	4

(Continued on next page)

Table 1. Continued

Complex/ Protein Class	Gene Name	Accession	Description	MW	Exp I	Exp II	Exp III
Heterogeneous Nuclear Ribonucleoproteins							
	Hnrnpab	IPI00277066	heterogeneous nuclear ribonucleoprotein A/B isoform 1	36302.44	3	2	2
	Hnrnpl	IPI00620362	heterogeneous nuclear ribonucleoprotein L	64550.49	10	4	8
	Hnrnpu	IPI00458583	putative uncharacterized protein	88661.02	8	2	12
Histone Ubiquitination (E3 Ubiquitin Ligase Complex)							
	Cul4b	IPI00224689	Cullin 4B	111314	7	3	9
	Ddb1	IPI00316740	DNA damage-binding protein 1	128026.73	7	5	16
Enzymes							
	Cad	IPI00380280	carbamoyl-phosphate synthetase 2, aspartate transcarbamylase, and dihydroorotase	245649.59	4	18	11
	Dnmt3a	IPI00131694	isoform 1 of DNA (cytosine-5)-methyltransferase 3A	103203.53	4	4	7
	Dnmt3l	IPI00109459	DNA (cytosine-5)-methyltransferase 3-like	49159.08	7	3	7
	Myst2	IPI00228457	isoform 2 of histone acetyltransferase MYST2	67588.54	4	2	3
	Ogt	IPI00420870	isoform 1 of UDP-N-acetylglucosamine-peptide N-acetylglucosaminyltransferase 110 kDa subunit	118131.41	14	4	10
	P4ha1	IPI00399959	isoform 2 of prolyl 4-hydroxylase subunit alpha-1	61132.82	14	5	5
	Ppp2r1a	IPI00310091	Serine/threonine-protein phosphatase 2A 65 kDa regulatory subunit A alpha isoform	66079.23	3	4	1
	Trim24	IPI00227778	isoform short of transcription intermediary factor 1-alpha	114824.79	3	4	6
	Trim33	IPI00409904	isoform alpha of E3 ubiquitin-protein ligase TRIM33	125931.28	2	2	2
Karyopherins							
	Kpna2	IPI00124973	Importin subunit alpha-2	58234.28	7	7	2
	Kpna3	IPI00230429	Importin subunit alpha-3	58193	2	1	2
Chaperones	Dnaja1	IPI00132208	DnaJ homolog subfamily A member 1	45580.73	6	1	5
Proteasome	Psmb6	IPI00119239	proteasome subunit beta type-6	25590.57	2	2	1
Nuclear Assembly/Organization							
	Emd	IPI00114401	Emerin	29417.38	2	1	1
	Matr3	IPI00453826	Matrin-3	95085.04	14	6	11
Miscellaneous							
	Amotl2	IPI00263333	isoform 1 of Angiomotin-like protein 2	85454.32	1	4	2
	Cubn	IPI00889898	Cubilin	407679.63	3	7	2
	Nudc	IPI00132942	nuclear migration protein nudC	38334.29	2	2	2

The number of unique peptides for three independent experiments is shown. MW, molecular weight. See also Table S6.

with some of its interactors, namely Sall1 and Smarcc1 (Wang et al., 2006). Eight previously identified Oct4-interacting proteins were either not detected, namely EWS, NF45, Cdk1 (Wang et al., 2006), and Zfp206 (Yu et al., 2009), or found also in controls, such as beta-catenin (Takao et al., 2007), Hdac2 (Liang et al., 2008), Ctf (Donohoe et al., 2009), and Wwp2 (Xu et al., 2009; Xu et al., 2004).

We next performed tandem affinity purification, although yields were not high because of the low levels of tagged Oct4 and purification efficiency. We identified seven proteins of the 92, mainly members of NuRD, Sall proteins, and transcription factors E3 and EB (Table S6). We believe these constitute the highest-affinity interactors, given that they can endure a more stringent purification. Although yielding small numbers of interactors, this purification is still at the level of previous similar studies (Liang et al., 2008; Wang et al., 2006).

For confirmation, we immunoprecipitated endogenous Oct4 from whole-cell extracts of untagged feeder-free E14 mouse ESCs in duplicate and analyzed immunoprecipitates by mass spectrometry. Forty-six proteins reproducibly overlapped with the FTAP data set (Table S6). We detected all proteins identified in a similar experiment (Liang et al., 2008). Not surprisingly, proteins that were reproducibly copurified with endogenous Oct4 tended to be more abundant in our single-affinity purification data set.

To address whether the interactions detected were due to coassembly of factors on chromatin, we then immunoprecipitated Oct4 in the presence of DNase treatment with benzonase. Western blotting showed that Parp1, a ubiquitous DNA-binding protein, coimmunoprecipitates with Oct4 even in the absence of DNA (Figure 1C). Preliminary purification experiments with a differently tagged Oct4-FTAP cell line suggested that other

DNA-binding proteins such as ligase 3 and topoisomerase 2a also copurify with Oct4 in the absence of DNA (M.P. and S.P. Shen, data not shown). This suggests that the interactions we detect are not DNA mediated.

Summing up, over 50% of the Oct4-associated proteins (47 of 92) varying in abundance across our data set have been confirmed by independent means, suggesting that the data set we provide here is a bona fide set of Oct4 binding partners.

Functional Annotation Analysis of Oct4-Associated Proteins

To uncover general trends in the functions of the Oct4-interacting proteins, we carried out computational systems-level analyses in a workflow depicted in Figure 1D. We first performed a functional annotation analysis using DAVID 2008 (Dennis et al., 2003) and the PANTHER database (Thomas et al., 2003). We found an enrichment of GO terms such as nucleus, chromosome, and chromatin in the cellular component ontology; nucleic acid binding, protein binding, transcription factor activity, in the molecular function ontology; and transcription, regulation of gene expression, and embryonic development in the biological process ontology (Figure S2 and Table S2). This indicates that GO terms associated to Oct4 are highly represented within the list of Oct4-copurifying proteins, adding consistency to the data set. Twenty proteins in the data set (21%) are annotated with the GO term “transcription factor activity.” Oct4 has been shown to associate with several transcription factors, and our results agree with the notion that combinatorial binding among pluripotency factors, which contributes to achieving specificity in gene regulation, may be a frequent pattern in ESCs (Chambers and Tomlinson, 2009).

We also analyzed the enrichment of proteins involved in cellular pathways. DAVID analysis detected an enrichment of proteins involved in the control of gene expression by vitamin D nuclear hormone receptor, mainly members of the FACT and SWI/SNF complexes. The data set also contains several proteins involved in the nuclear part of the Wnt signaling pathway, as revealed by PANTHER analysis. The Wnt pathway is involved in stem cell maintenance (Anton et al., 2007; Sato et al., 2004), possibly by modulating levels of pluripotency factors Oct4, Nanog, and Sox2 (Kalmar et al., 2009).

We then analyzed the domain composition of Oct4-interacting proteins and detected a significant abundance of DNA-binding and chromatin-related domains (Tables S3 and S4). Highly represented domains were DEAD/DEAH box helicase, SNF2-related, PHD and zinc fingers, and chromo, bromo, and homeo-box domains, all of which are either involved in ATP-dependent chromatin remodeling or bind DNA and posttranslationally modified nucleosomes, thereby influencing gene expression.

The data set was manually classified into known protein complexes and functional categories. As shown in Table 1 and supported by the GO and PANTHER analyses, Oct4 associates mainly with transcriptional regulators, but also with a variety of other chromatin binding proteins involved in DNA replication, recombination, and repair, proteins involved in nuclear assembly and/or organization, and diverse enzymes, some of which are responsible for addition of posttranslational modifications.

To gain an overall view of the previously known interactions among Oct4-associated proteins, we retrieved interaction data

from INTACT, HRPD, and MINT for the data set and represented them as a protein interaction network (Figure 1E). The network comprises 80 known interactions for 57 of the proteins including Oct4. Repressor complexes NuRD and SWI/SNF and DNA repair and de novo DNA methylation modules are apparent in the network.

Transcriptional Regulation of Oct4-Associated Proteins

Many known Oct4 binding proteins are ESC-specific factors (Wang et al., 2008). However, Oct4 has also been shown to interact with more general modulators of transcription that are expressed ubiquitously, such as members of the NuRD complex (Liang et al., 2008; Orkin et al., 2008; Wang et al., 2006). We investigated the patterns of expression of the Oct4-associated data set in cells at different stages of differentiation, including embryonic carcinoma, embryonic stem cells, embryoid bodies, and various differentiated cell types, on the basis of transcriptomics data (Campbell et al., 2007). Protein abundances were fairly varied and most interactors maintained near constant expression across the samples analyzed (Figure 2). This suggests that Oct4 interacts mostly with proteins that are ubiquitously expressed in both differentiated and undifferentiated cells. After statistical analysis, 33 Oct4-interacting proteins were found to be significantly less expressed in differentiated cells compared to ESCs, in correlation with Oct4's expression pattern (Figure 2 and Table S6). Among these are the DNA methylation regulatory factor Dnmt3l and the developmentally important transcription factors Klf4, Sall1, and Sall4. We observed that many complexes or interacting pairs in the interaction network contained at least one member significantly downregulated upon ESC differentiation (Figure 1E), possibly conferring an ESC-specific role.

Regulation of gene expression involves complex dynamics employing sequence-specific DNA binding proteins that form the transcriptional regulatory network (Babu et al., 2004; Jothi et al., 2009; Luscombe et al., 2004). Transcription factors often operate in feedback loops, whereby the expression of a transcriptional target modulates the function of the transcription factor itself (Figure 3A). Pluripotency factors in ESCs are no exception and show a high degree of transcriptional auto and interregulation (Orkin et al., 2008). We next investigated whether the promoters of the Oct4-associated gene set contain binding sites for transcription factors that are central in the establishment and maintenance of ESC identity. Promoter binding sites for nine such transcription factors have previously been identified by ChIP-on-chip (Kim et al., 2008). These include Oct4, Dax1, Klf4, c-Myc, Nac1, Nanog, Rex1, Zfp281, and Sox2.

Nine of the 92 Oct4-associated proteins were found to be transcriptionally regulated by Oct4 itself in mouse ESCs, and 51% of genes encoding Oct4 partners are targets of at least one key ESC transcription factor (Figure S3). This concurs with findings by others on a much smaller data set (Orkin et al., 2008; Wang et al., 2006). To assess whether this is statistically significant, we compared the results to 1000 randomly generated sets of 92 proteins. The expected percentage of promoter binding by ESC transcription factors was only 28% ($Z = 4.45$, $p < 10^{-15}$), indicating that it is a significant trait of the data set. Several genes in the interaction set are common targets of multiple transcription factors (20 of 92 are targets of at least three transcription

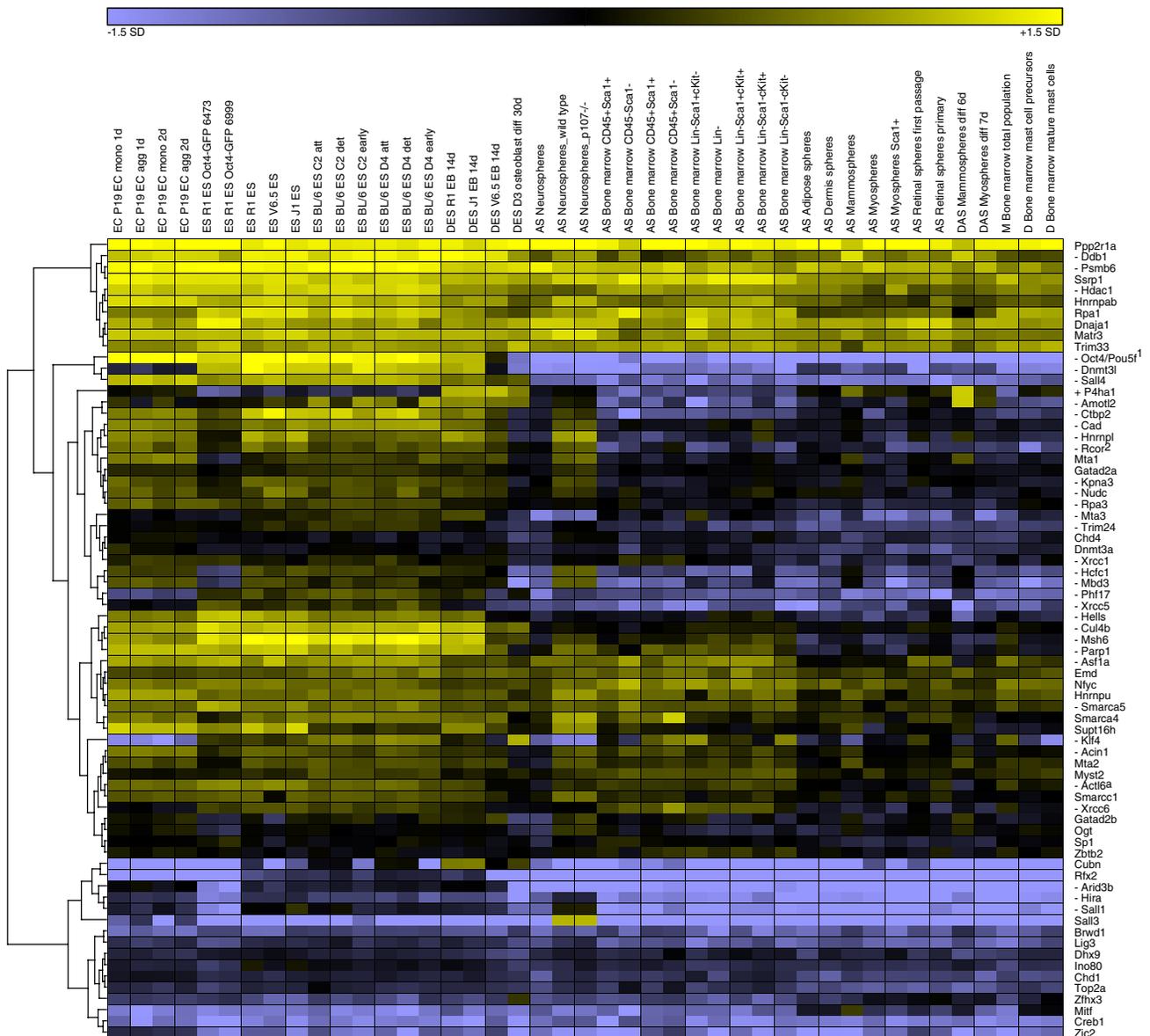


Figure 2. Expression of Oct4-associated proteins in ESCs and Differentiated Cell Types Based on Microarray Data

Columns correspond to experimental samples, arranged as follows: embryonal carcinoma P19 (EC), ES cells (ES), differentiating embryonic stem cells (DES), adult stem cells (AS), differentiated adult stem cells (DAS), mixed cells (M), and differentiated cells (D). Average-linkage hierarchical clustering was performed to arrive at the final layout. Genes whose expression is significantly up or downregulated in differentiation are marked.

factors), making it likely that they have central roles in pluripotency and self-renewal. Ten of these are significantly downregulated in differentiation, and all but two show a downregulation trend (Table S6), in agreement with the hypothesis that genes bound by multiple factors are active in ESCs and become repressed as cells differentiate (Chambers and Tomlinson, 2009; Kim et al., 2008; Orkin et al., 2008). We constructed a regulatory network by integrating promoter target data for the nine stem cell transcription factors with the list of Oct4 binding proteins (Figure 3B). Several of the transcription factors cluster together because of shared targets (e.g., Sox2, Nanog, Nac1, and Dax1), whereas c-Myc and Klf4 exclusively target certain groups of Oct4-interacting factors. This agrees with

the genome-wide trend of the c-Myc target set, which is largely distinct from the rest of the pluripotency factors (Kim et al., 2008).

Role of Oct4 Interactome in Mouse Embryonic Stem Cell Biology and Development

We next explored the consequences of loss of Oct4-interacting proteins in ESCs or mouse development. Five Oct4-interacting proteins have been identified as required for stem cell self-renewal in large scale RNAi screens (Ding et al., 2009; Hu et al., 2009). Literature searches ascribed a role in ESC self-renewal or pluripotency to an additional nine Oct4-interacting proteins (Table S6).

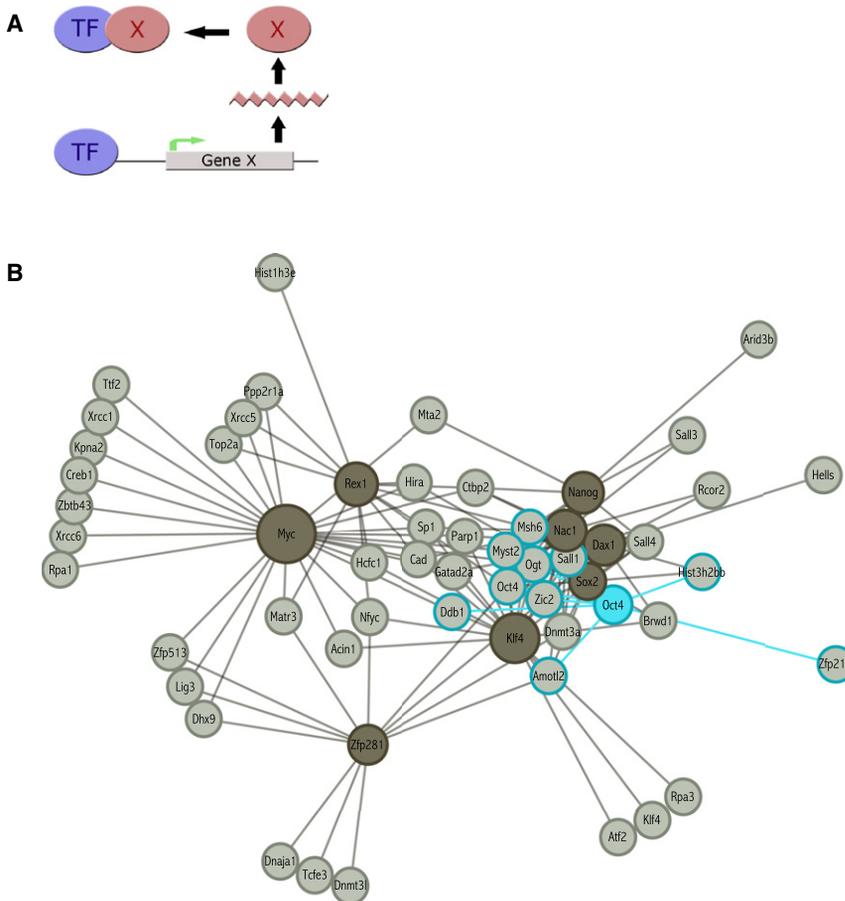


Figure 3. Transcriptional Regulation of Oct4-Associated Proteins

(A) Scheme of transcription factor feedback regulation.

(B) Regulatory network of targets of ESC transcription factors among Oct4 partners. Stem cell transcription factors and their target genes among Oct4-associated proteins are shown in dark and light gray, respectively. Oct4 and its regulatory targets are highlighted in blue. See also Figure S3.

phenotype. These results indicate a high level of requirement for components of the Oct4 network in early mouse development. Although feedback loops are expected to add to the robustness of a transcriptional regulatory network, the high frequency with which mutation of single Oct4 partners causes severe early developmental phenotypes suggests they are essential downstream regulatory hubs.

Involvement of Oct4-Interacting Proteins in Human Disease and Cancer

Given the extent of their part in mouse development and the current excitement about the cancer stem cell hypothesis, we next explored a possible role of Oct4-associated proteins in human disease. Human orthologs were identified for all

Loss-of-function phenotypes in mice were available in the MGI database for 49 Oct4-associated proteins. All 49 show diverse phenotypes when absent or mutated (Figure 4 and Table S5). Significantly, 83% (41 of 49) of the studied knockout alleles of the interaction set showed embryonic and/or perinatal lethality, with over 60% (30 of 49) being embryonic lethal (Figure 1E). Similar analyses on random control data sets allowed us to conclude that the result is significant ($Z = 7.48, p < 10^{-15}$). In addition, 41% (20 of 49) showed an abnormal development

Oct4-associated proteins and sequence identities determined between mouse and human (Figure S4). All Oct4-associated proteins were found to be highly conserved, with a median sequence identity of 94%, compared to 77% genomic median. This strong sequence conservation implies that the findings reported here could be applied to human ESC biology.

We next investigated the involvement of the human orthologs in human disease and development of cancer interrogating the OMIM database and the Cancer Gene Census, which records genes whose mutation has been causally linked to cancer. Genes encoding 14 of 92 Oct4-associated proteins are implicated in one or more hereditary diseases, mostly of developmental nature, with six of them predisposing to certain types of cancer (Table 2). Somatic mutations in eight Oct4-associated proteins and Oct4 itself were found to be responsible for different types of cancer, often through gene translocations, presumably affecting their regulation (Table 3). Statistical analysis on random sets indicated that the observed numbers of Oct4-interacting proteins linked to human disease ($Z = 1.06, p < 10^{-15}$) and cancer ($Z = 4.43, p < 10^{-15}$) are significantly higher than expected.

In light of the central role of Oct4 in pluripotency and the cancer stem cell hypothesis, we investigated which of Oct4's physical interactors are misexpressed in cancer using the Oncomine human cancer expression database. A large fraction (60%) of the Oct4 interactors show misexpression in at least

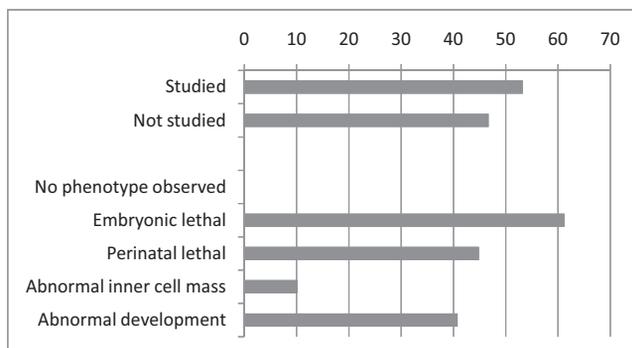


Figure 4. Distribution of Phenotypes Caused by Mutations in the Genes Encoding Oct4-Associated Proteins

Numbers are percentage of genes studied. Full data are shown in Table S5.

Table 2. Genetic Disease Associations of Oct4-Interacting Proteins

Gene	Disorder Type	Disorder
CREB1	cancer	histiocytoma, angiomatoid fibrous, somatic
CUBN	hematological	megaloblastic anemia-1, Finnish type
CUL4B	multiple	mental retardation syndrome, X-linked, Cabezas type mental retardation-hypotonic facies syndrome, X-linked, 2
EMD	muscular	Emery-Dreifuss muscular dystrophy
MATR3	muscular	myopathy, distal 2
MITF	multiple	Tietz syndrome Waardenburg syndrome, type IIA Waardenburg syndrome/ocular albinism, digenic
MSH6	cancer	colorectal cancer, hereditary nonpolyposis, type 5 endometrial cancer, familial mismatch repair cancer syndrome
SALL1	multiple	Townes-Brocks branchiootorenal-like syndrome Townes-Brocks syndrome
SALL4	multiple	Duane-radial ray syndrome IVIC syndrome
TFE3	cancer	renal cell carcinoma, papillary, 1
TRIM24	cancer	thyroid carcinoma, papillary
TRIM33	cancer	thyroid carcinoma, papillary
ZFH3	cancer	prostate cancer, susceptibility to
ZIC2	developmental	holoprosencephaly-5

See also Figure S4.

one cancer type, providing a degree of additional support to the connection between stem cell identity and cancer.

DISCUSSION

The characterization of protein-protein interactions is a very efficient strategy for understanding protein function and regulation. The development of high-affinity tags, including the TAP (Rigaut

et al., 1999) and in vivo biotinylation tag (de Boer et al., 2003), in combination with advances in mass spectrometry that now allow protein identification with high sensitivity and accuracy, has recently produced several protein interaction network reports. However, most studies in the literature rely on cDNA overexpression driven by exogenous promoters or transgenic random integration approaches.

We report here an epitope-tagging strategy for the purification of protein complexes in mouse ESCs. We introduced the tag by recombineering into a full-length Oct4-containing BAC and then integrated this in a precise location in the mouse genome. This approach has the advantage of maintaining the endogenous promoter and therefore natural transcriptional regulation. The technology is amenable to high-throughput delivery, as recently demonstrated by random integration of tagged BAC transgenes (Poser et al., 2008), and should greatly facilitate systematic tagging of genes and analysis of protein complexes with roles in development in different contexts, be it in stem cells, differentiated cell types, or even mouse tissue (Fernández et al., 2009).

The affinity purification method described here is rapid, with the goal of capturing weak or short-lived interactions. Previous proteomic studies of Oct4 protein complexes have relied on lengthy single or tandem purifications from nuclear extracts with streptavidin capture (Wang et al., 2006) or anti-Oct4 antibodies (Liang et al., 2008) and yielded small data sets, very similar to our tandem purification data set. We identified all of the partners reported by the Liang study except Hdac2, and only five Oct4 partners found in the Wang study were not detected in our data set, maybe because of our use of whole extracts. Indeed, our approach has produced by far the most extensive analysis of Oct4-associated proteins to date.

By using whole extracts, thereby not restricting the analysis to the nuclear environment, our data set encompasses diverse aspects of the life of Oct4, both nuclear and nonnuclear. The broad data set puts Oct4 at the center of diverse cellular processes that can have an impact on aspects of stem cell biology (Figure 5), the most interesting of which are discussed below.

Oct4 can both activate and repress transcriptional targets in mouse and human ESCs (Babaie et al., 2007; Loh et al., 2006). To date, Oct4 has been shown to be associated mainly with

Table 3. Cancer-Causative Genes among the Oct4-Interacting Proteins

Gene	Mutation	Tissue	Cancer Type
CREB1	translocation	mesenchymal	clear cell sarcoma, angiomatoid fibrous histiocytoma
MITF	amplification	epithelial	melanoma
MSH6	Missense, nonsense, frameshift, splice site	epithelial	colorectal (somatic) colorectal, endometrial, ovarian (germline) non-polyposis colorectal cancer (hereditary)
POU5F1	translocation	mesenchymal	sarcoma
SMARCA4	frameshift, nonsense, missense	epithelial	NSCLC (non-small cell lung carcinoma)
TFE3	translocation	epithelial	papillary renal, alveolar soft part sarcoma, renal
TFEB	translocation	epithelial mesenchymal	renal, childhood epithelioid
TRIM24	translocation	blood	APL (acute promyelocytic leukemia)
TRIM33	translocation	epithelial	papillary thyroid

See also Figure S4.

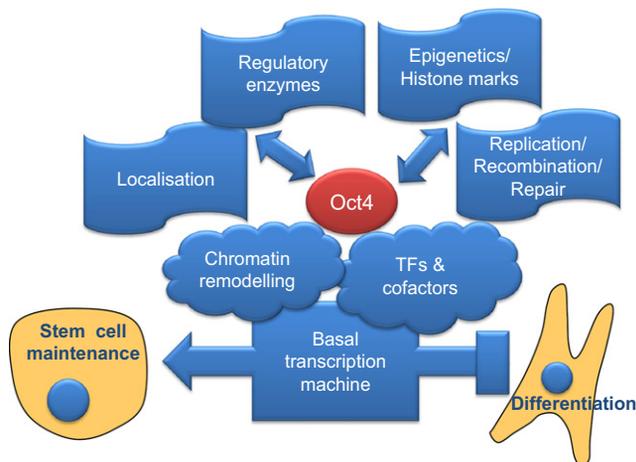


Figure 5. Schematic Model of the Oct4 Interactome

members of repressor complexes NuRD and SWI/SNF (Liang et al., 2008; Wang et al., 2006). We found both among our data set of Oct4-copurifying proteins. NuRD, a histone deacetylase complex, was the most prominent, further confirming this link. Sall4, a well-known Oct4 partner, and other members of the Spalt-like family of transcriptional cofactors have been shown to associate to NuRD (Lauberth and Rauchman, 2006), raising the possibility that they may bridge the interaction between Oct4 and NuRD. This hypothesis is also supported by the similar amounts in which they are detected in our experiments. We also found several subunits of the SWI/SNF nucleosome-remodeler complex, some of which have previously been linked to Nanog (Liang et al., 2008), confirming the link to this chromatin remodeling complex.

Also among Oct4 binding proteins we found various molecules involved in positive regulation of transcription, including several activators and coactivators and chromatin-modifying enzymes such as *Myst2*, a histone H4 acetyltransferase (Doyon et al., 2006; Sterner and Berger, 2000). In addition, we detected *Ttf2*, a component of the general transcription machinery, providing evidence of a physical link between pluripotency factors and basal transcription players. The Oct4 interactome included other basal DNA-process-related factors such as proteins involved in DNA replication, recombination, and repair. This could explain why many of the Oct4-interacting proteins are ubiquitously expressed in both differentiated and undifferentiated cells. Our experiments suggest that the interaction is not DNA mediated, given that copurification of DNA-binding proteins still occurs upon DNA elimination by benzonase.

Importantly, we have uncovered enzymes with a potential role in Oct4 regulation. *Ogt* is responsible for posttranslational addition of O-linked N-acetylglucosamine (O-GlcNAc), a regulatory protein modification similar to phosphorylation possibly working in concert with it (Kamemura and Hart, 2003). Oct4 is modified by O-GlcNAc in human ESCs (Webster et al., 2009), and *Sp1*, one of Oct4 partners, is too (Jackson and Tjian, 1988). A thorough analysis of O-GlcNAc modification in the Oct4 interactome might yield important insight into dynamic modulation of stem cell factors. Posttranslational modification of transcription factors and cofactors is proving to be a critical

component of the regulation of gene transcription in general, and important specifically in stem cell biology (Brill et al., 2009; Van Hoof et al., 2009).

Half of Oct4-associated proteins seem to be directly regulated by transcription factors with key roles in stem cell pluripotency and/or reprogramming. This is also a characteristic of pluripotency networks derived from smaller data sets from different entry points (Orkin et al., 2008; Wang et al., 2006). This indicates that even in the expanded and functionally diverse network, this attribute still holds true, supporting a previously unsuspected role in stem cell biology for some of the proteins we identify here.

Expression of Oct4 decreases in a switch-like fashion as ESCs differentiate into lineage-specific cell types, including progenitor cells. Our analysis has uncovered 33 physical interactors of Oct4 that share this trend. Among these are several transcription factors, such as the DNA methyltransferase 3-like regulatory protein *Dnmt3l*, which stimulates genomic imprinting in germ cells (Bourc'his et al., 2001; Gowher et al., 2005). This is consistent with a recent report demonstrating that treatment with DNA methyltransferase inhibitors can improve the efficiency of the reprogramming process of differentiated cells (Mikkelsen et al., 2008). Therefore, the 33 interactors upregulated in ESCs and the transcription factors that regulate them might be interesting candidates whose expression could be manipulated to facilitate reprogramming.

We find that loss of function of most Oct4-associated genes studied to date results in embryonic or perinatal lethality, suggesting that many serve crucial functions in development. Interestingly, most Oct4-binding proteins linked to a human hereditary disorder (13 of 14), mostly developmental or cancer predisposition, give rise to a related phenotype when absent in the mouse. We find cancer-associated genes, either causal or predisposing, to be transcriptional regulators involved in processes relating to the cell cycle, differentiation, and DNA repair, acting through chromatin remodeling, signaling, or transcription factor activity. These results implicate the orthologs of Oct4-interacting proteins in roles in human development and cancer, and therefore the data presented here should be useful in elucidating their part in human disease.

In summation, the extensive systems-level analyses described here compiling data sets of currently available genome-wide studies provide an integrated vision of the Oct4 interactome. Detailed investigation of this information should facilitate the choice of candidate factors to test for roles in ESC maintenance, differentiation, and reprogramming and provide great insight into the transcriptional regulation of ESC biology.

EXPERIMENTAL PROCEDURES

Generation of FTAP-Tagged Oct4 ESCs

Full details are provided in [Supplemental Experimental Procedures](#). In brief, the FTAP epitope tag (3×FLAG-2×TEV-CBP) sequence was synthesized as two DNA fragments by annealing overlapping complementary oligonucleotide molecules with PCR. The two fragments were cloned into a modified version of recombining vector PL450 (Liu et al., 2003) for pCTR9 creation. The correctness of the FTAP tag within pCTR9 was confirmed by sequencing. Homology arms for recombining were PCR amplified from the *Oct4* containing C57Black/6J derived BAC clone (RPC1 23-213M12) and cloned into

the recombining vector to create pCTS1 (Figure S1). The 5' homology arm creates an in-frame fusion between the Oct4 C-terminal coding sequence and the FTAP tag coding sequence, while deleting the stop codon. A fragment for recombining the FTAP tag sequence into the Oct4-containing BAC (RPC1 23-213M12) was generated by digesting clone pCTS1. Correct recombination into *E. coli* DH10B containing BAC clone RPC1 23-213M12 was confirmed by Southern analysis of BAC DNA with homology arm-specific DNA probes for all six tagged BAC clones tested.

ESC cultures, electroporation, and mini-Southern-blot analysis of ESC clones were as described previously (Ramírez-Solis et al., 1993). Integration of single-copy BAC transgenes at the *Hprt* locus by recombinase-mediated cassette exchange (RMCE) has been described previously (Prosser et al., 2008). For RMCE integration of tagged Oct4 BAC insert into *hprt*^{tm(rmce1)Brd} allele of CCI18#1.6G, cells were cotransfected with pCAGGS-Cre (Araki et al., 1997) and the RPC1 23-213M12 BAC clone carrying an integrated copy of the FTAP tag cassette and neomycin resistance gene. Double-resistant colonies were isolated after successive selection with G418 (200 mg/ml) and 6-TG (10 μ M). Site-specific BAC integration was very efficient, as verified by Southern analysis with *Hprt* flanking probes, with 19 of 23 double-resistant colonies showing correct single-copy integration. For removal of the selection cassette, the verified ESC clones were transfected with pCAGGS-Flpe (Schaff et al., 2001) and then selected with FIAU (200 nM). FIAU-resistant subclones were assessed for selection cassette deletion by Southern blotting. Absence of a hybridizing 5 kb fragment indicated successful deletion of the selection cassette. Transgenic clones were analyzed for expression of tagged *Pou5f1* by Western blotting, demonstrating that 60% of clones expressed the Oct4-FTAP fusion protein.

Affinity Purification

Murine ESCs expressing Oct4-FTAP or wild-type control cells (AB2.2) were separated from feeders by trypsinization and incubation on gelatin-coated plates for 60 min. Whole-cell extracts were incubated with anti-FLAG M2 Dynal beads in buffer containing 150 mM NaCl and 0.1% NP-40 for 90 min at 4°C. Anti-FLAG Dynal beads were prepared by crosslinking M2 FLAG antibody (Sigma) to Protein G-Dynal beads (Invitrogen) in accordance with the manufacturer's instructions. Bound complexes were eluted with AcTEV protease (Invitrogen). For tandem affinity purification, the TEV eluate was incubated with calmodulin resin (Stratagene) for 60 min at 4°C. Elution was carried out by Ca chelation with EGTA. TEV or EGTA eluates were concentrated in Vivaspin 500 PES centrifugal filters (Vivascience), reduced with 1 mM DTT, and alkylated with 2 mM iodoacetamide prior to sample fractionation by polyacrylamide gel electrophoresis with Novex NuPAGE Bis-Tris 4%–12% gels (Invitrogen). Gels were stained with colloidal Coomassie (Sigma) according to Rowley (Rowley et al., 2000). Whole lanes were cut in 24 slices, destained completely, and digested with trypsin (sequencing grade, Roche). Peptides were extracted with 0.5% formic acid 50% acetonitrile and dried in a Speed Vac (Thermo).

Immunoprecipitation and Western Blotting

Oct4 complexes were immunoprecipitated with an Oct4 antibody (Santa Cruz) coupled to Dynal-Protein G beads (Invitrogen). Immunoprecipitates were eluted by boiling in 1 \times LDS loading buffer (Invitrogen) and separated by LDS-PAGE (Invitrogen). Western blotting was carried out with antibodies from Abcam (Parp1, Sall4, and Myst2), Bethyl Laboratories (Chd4), or Santa Cruz (Oct4 and Hdac1).

Mass Spectrometry and Data Analysis

Peptides were redissolved in 0.5% formic acid and analyzed with online nanoLC-MS/MS on a LTQ FT mass spectrometer (Thermo Fisher Scientific) coupled with an Ultimate 3000 Nano/Capillary LC System (Dionex). Samples were first loaded and desalted on a trap (0.3 mm id \times 5 mm) at 25 μ L/min with 0.1% formic acid for 5 min and then separated on an analytical column (75 μ m id \times 15 cm) (both PepMap C18, LC Packings) over a 30 min linear gradient of 4%–32% CH₃CN/0.1% formic acid at 300 nL/min. The LTQ FT was operated in standard data-dependent acquisition. The survey scans (*m/z* 400–2000) were acquired on the FT-ICR at a resolution of 100,000 at *m/z* 400, and one microscan was acquired per spectrum. The three most abundant multiply charged ions with a minimal intensity of 1000 counts were subject to MS/MS in the linear ion trap at an isolation width of 3 Th. Dynamic

exclusion width was set at \pm 10 ppm for 45 s. The automatic gain control target value was regulated at 5E5 for FT and 1E4 for the ion trap, with maximum injection time at 1000 ms for FT and 200 ms for the ion trap, respectively.

The raw files were processed with BioWorks (Thermo). Database searches were performed with Mascot v.2.1 (Matrix Science) against the mouse IPI database (v. January 2009). The search parameters were: Trypsin/P with two missed cleavages, 10 ppm mass tolerance for MS, 0.5 Da tolerance for MS/MS, fixed modification Carbamidomethyl (C), and variable modifications of Acetyl (Protein N-term), Deamidated (NQ), Dioxidation (M), Formyl (N-term), Gln- > pyro-Glu (N-term Q), Methyl (E), and Oxidation (M). Decoy database searches were performed at the same time as the real searches, resulting in false discovery rates under 5%.

Only peptides with scores above 20 were used in the analysis. Protein identification required at least one high-confidence peptide (peptide score above identity threshold, $e \leq 0.05$, length > 8 aas, precursor ion mass accuracy < 5 ppm where $e \geq 0.005$, peptide hit rank 1, and delta peptide score > 10). There is increased risk of false discovery when a protein is identified by only one peptide. Thus, all peptides identifying a protein without additional support met the strict confidence requirements above and were manually verified. Precursor ion mass accuracies of these peptides are shown in Table S1. Mascot results were clustered to 95% protein homology to collapse highly homologous sequences corresponding to the same gene, and all lists for target and control purifications were compared in parallel. External contaminants (keratins, albumin, casein, and TEV protease) were excluded from the list. In the final list of Oct4-associated proteins we report only proteins identified in all three replicates. We have chosen one representative of each protein cluster, the one with the highest number of peptide matches, meaningful gene symbol, and highest molecular weight.

Computational and Systems-Level Analysis

Orthologous human proteins were identified with the g:Profiler orthology search tool (Reimand et al., 2007) or NCBI BLASTP and aligned with the Needleman-Wunsch algorithm. For assessment of the degree of conservation between the Oct4-associated proteins and their orthologs, sequence identities of all mouse-to-human ortholog pairs of comparable sequence length in ENSEMBL release 57 were compared via a Mann-Whitney U test.

Domains were identified with Pfam 24.0, and genome-wide frequencies were calculated from domain annotations in UniProtKB/Swiss-Prot release 15.15 (UniProt Consortium, 2010).

Mammalian Phenotype Ontology annotations were obtained from the Mouse Genome Informatics project (Blake et al., 2009), human disease associations were obtained from OMIM (Hamosh et al., 2005), and known cancer-causing mutations in genes were obtained from the Cancer Gene Census (Futreal et al., 2004). Student's t test was used for assessing the significance of the observed numbers of Oct4-associated genes with lethal phenotypes, disease, and cancer associations against 1000 random sets of 92 genes, in each case.

ChIP-on-chip data were obtained for Oct4 and eight other transcription factors (Kim et al., 2008). The significance of the number of Oct4-associated proteins regulated by these factors was assessed against 1000 random sets of 92 genes. The protein interaction network was generated with Cytoscape 2.6.3 (Cline et al., 2007), with a spring-embedded layout.

For the analysis of expression at different stages of differentiation, data were obtained for 43 mouse samples in StemBase (Sandie et al., 2009), originating from 16 studies with Affymetrix MOE430A microarray chips, as used in an Oct4 expression profiling study (Campbell et al., 2007) covering murine ESCs, embryonal carcinoma cell lines, and several early differentiated lineages. Expression data was available for 70 of the 92 Oct4-associated proteins. Where multiple probes were available, expression was averaged. Student's t test was used for identifying genes differentially expressed in ESCs as compared to more differentiated cell types (Bonferroni-corrected for multiple testing). Expression values were log₂-transformed and color-coded as a gradient from blue (more than twice the standard deviation below the global microarray mean) via black (microarray mean) to yellow (more than twice the standard deviation above the mean). Average-linkage hierarchical clustering was performed to arrive at the final layout.

Data on significantly misexpressed genes was curated from the Oncomine human cancer expression database (Rhodes et al., 2007). Genes were

considered mis-expressed below a p-value threshold of 10^{-10} (Bonferroni-corrected for multiple testing).

SUPPLEMENTAL INFORMATION

Supplemental Information includes four figures, six tables, and Supplemental Experimental Procedures and can be found with this article online at doi:10.1016/j.stem.2010.03.004.

ACKNOWLEDGMENTS

We would like to thank Frances Law and Alastair Beasley for technical assistance and Sajani Swamy and Parthiban Vijayarangakannan for informatics support. The work described here was supported by the Wellcome Trust and the Herchel Smith Research Studentship Fund (B.L.).

Received: December 18, 2009

Revised: March 10, 2010

Accepted: March 16, 2010

Published: April 1, 2010

REFERENCES

- Amabile, G., and Meissner, A. (2009). Induced pluripotent stem cells: Current progress and potential for regenerative medicine. *Trends Mol. Med.* *15*, 59–68.
- Ambrosetti, D.C., Basilico, C., and Dailey, L. (1997). Synergistic activation of the fibroblast growth factor 4 enhancer by Sox2 and Oct-3 depends on protein-protein interactions facilitated by a specific spatial arrangement of factor binding sites. *Mol. Cell. Biol.* *17*, 6321–6329.
- Anton, R., Kestler, H.A., and Kühl, M. (2007). Beta-catenin signaling contributes to stemness and regulates early differentiation in murine embryonic stem cells. *FEBS Lett.* *581*, 5247–5254.
- Araki, K., Imaizumi, T., Okuyama, K., Oike, Y., and Yamamura, K. (1997). Efficiency of recombination by Cre transient expression in embryonic stem cells: Comparison of various promoters. *J. Biochem.* *122*, 977–982.
- Babaie, Y., Herwig, R., Greber, B., Brink, T.C., Wruck, W., Groth, D., Lehrach, H., Burdon, T., and Adjaye, J. (2007). Analysis of Oct4-dependent transcriptional networks regulating self-renewal and pluripotency in human embryonic stem cells. *Stem Cells* *25*, 500–510.
- Babu, M.M., Luscombe, N.M., Aravind, L., Gerstein, M., and Teichmann, S.A. (2004). Structure and evolution of transcriptional regulatory networks. *Curr. Opin. Struct. Biol.* *14*, 283–291.
- Barsnes, H., Vizcaino, J.A., Eidhammer, I., and Martens, L. (2009). PRIDE Converter: Making proteomics data-sharing easy. *Nat. Biotechnol.* *27*, 598–599.
- Blake, J.A., Bult, C.J., Eppig, J.T., Kadin, J.A., Richardson, J.E., and Mouse Genome Database Group. (2009). The Mouse Genome Database genotypes: Phenotypes. *Nucleic Acids Res.* *37* (Database issue), D712–D719.
- Bourc'his, D., Xu, G.L., Lin, C.S., Bollman, B., and Bestor, T.H. (2001). Dnmt3L and the establishment of maternal genomic imprints. *Science* *294*, 2536–2539.
- Boyer, L.A., Lee, T.I., Cole, M.F., Johnstone, S.E., Levine, S.S., Zucker, J.P., Guenther, M.G., Kumar, R.M., Murray, H.L., Jenner, R.G., et al. (2005). Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell* *122*, 947–956.
- Brill, L.M., Xiong, W., Lee, K.B., Ficarro, S.B., Crain, A., Xu, Y., Tersikh, A., Snyder, E.Y., and Ding, S. (2009). Phosphoproteomic analysis of human embryonic stem cells. *Cell Stem Cell* *5*, 204–213.
- Campbell, P.A., Perez-Iratxeta, C., Andrade-Navarro, M.A., and Rudnicki, M.A. (2007). Oct4 targets regulatory nodes to modulate stem cell function. *PLoS ONE* *2*, e553.
- Chambers, I., and Tomlinson, S.R. (2009). The transcriptional foundation of pluripotency. *Development* *136*, 2311–2322.
- Chew, J.L., Loh, Y.H., Zhang, W., Chen, X., Tam, W.L., Yeap, L.S., Li, P., Ang, Y.S., Lim, B., Robson, P., and Ng, H.H. (2005). Reciprocal transcriptional regulation of Pou5f1 and Sox2 via the Oct4/Sox2 complex in embryonic stem cells. *Mol. Cell. Biol.* *25*, 6031–6046.
- Cline, M.S., Smoot, M., Cerami, E., Kuchinsky, A., Landys, N., Workman, C., Christmas, R., Avila-Campilo, I., Creech, M., Gross, B., et al. (2007). Integration of biological networks and gene expression data using Cytoscape. *Nat. Protoc.* *2*, 2366–2382.
- de Boer, E., Rodriguez, P., Bonte, E., Krijgsveld, J., Katsantoni, E., Heck, A., Grosveld, F., and Strouboulis, J. (2003). Efficient biotinylation and single-step purification of tagged transcription factors in mammalian cells and transgenic mice. *Proc. Natl. Acad. Sci. USA* *100*, 7480–7485.
- Dennis, G., Jr., Sherman, B.T., Hosack, D.A., Yang, J., Gao, W., Lane, H.C., and Lempicki, R.A. (2003). DAVID: Database for annotation, visualization, and integrated discovery. *Genome Biol.* *4*, P3.
- Ding, L., Paszkowski-Rogacz, M., Nitzsche, A., Slabicki, M.M., Heninger, A.K., de Vries, I., Kittler, R., Junqueira, M., Shevchenko, A., Schulz, H., et al. (2009). A genome-scale RNAi screen for Oct4 modulators defines a role of the Paf1 complex for embryonic stem cell identity. *Cell Stem Cell* *4*, 403–415.
- Donohoe, M.E., Silva, S.S., Pinter, S.F., Xu, N., and Lee, J.T. (2009). The pluripotency factor Oct4 interacts with Ctfc and also controls X-chromosome pairing and counting. *Nature* *460*, 128–132.
- Doyon, Y., Cayrou, C., Ullah, M., Landry, A.J., Côté, V., Selleck, W., Lane, W.S., Tan, S., Yang, X.J., and Côté, J. (2006). ING tumor suppressor proteins are critical regulators of chromatin acetylation required for genome expression and perpetuation. *Mol. Cell* *21*, 51–64.
- Fernández, E., Collins, M.O., Uren, R.T., Kopanitsa, M.V., Komiyama, N.H., Croning, M.D., Zografos, L., Armstrong, J.D., Choudhary, J.S., and Grant, S.G. (2009). Targeted tandem affinity purification of PSD-95 recovers core postsynaptic complexes and schizophrenia susceptibility proteins. *Mol. Syst. Biol.* *5*, 269.
- Futreal, P.A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., Rahman, N., and Stratton, M.R. (2004). A census of human cancer genes. *Nat. Rev. Cancer* *4*, 177–183.
- Gao, F., Kwon, S.W., Zhao, Y., and Jin, Y. (2009). PARP1 poly(ADP-ribose) at Sox2 to control Sox2 protein levels and FGF4 expression during embryonic stem cell differentiation. *J. Biol. Chem.* *284*, 22263–22273.
- Gowher, H., Liebert, K., Hermann, A., Xu, G., and Jeltsch, A. (2005). Mechanism of stimulation of catalytic activity of Dnmt3A and Dnmt3B DNA-(cytosine-C5)-methyltransferases by Dnmt3L. *J. Biol. Chem.* *280*, 13341–13348.
- Hamosh, A., Scott, A.F., Amberger, J.S., Bocchini, C.A., and McKusick, V.A. (2005). Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* *33* (Database issue), D514–D517.
- Hu, G., Kim, J., Xu, Q., Leng, Y., Orkin, S.H., and Elledge, S.J. (2009). A genome-wide RNAi screen identifies a new transcriptional module required for self-renewal. *Genes Dev.* *23*, 837–848.
- Ivanova, N., Dobrin, R., Lu, R., Kotenko, I., Levorse, J., DeCoste, C., Schafer, X., Lun, Y., and Lemischka, I.R. (2006). Dissecting self-renewal in stem cells with RNA interference. *Nature* *442*, 533–538.
- Jackson, S.P., and Tjian, R. (1988). O-glycosylation of eukaryotic transcription factors: Implications for mechanisms of transcriptional regulation. *Cell* *55*, 125–133.
- Jothi, R., Balaji, S., Wuster, A., Grochow, J.A., Gsponer, J., Przytycka, T.M., Aravind, L., and Babu, M.M. (2009). Genomic analysis reveals a tight link between transcription factor dynamics and regulatory network architecture. *Mol. Syst. Biol.* *5*, 294.
- Kaji, K., Norrby, K., Paca, A., Mileikovskiy, M., Mohseni, P., and Woltjen, K. (2009). Virus-free induction of pluripotency and subsequent excision of reprogramming factors. *Nature* *458*, 771–775.
- Kalmar, T., Lim, C., Hayward, P., Muñoz-Descalzo, S., Nichols, J., Garcia-Ojalvo, J., and Martinez Arias, A. (2009). Regulated fluctuations in nanog expression mediate cell fate decisions in embryonic stem cells. *PLoS Biol.* *7*, e1000149.
- Kamemura, K., and Hart, G.W. (2003). Dynamic interplay between O-glycosylation and O-phosphorylation of nucleocytoplasmic proteins: A new paradigm

- for metabolic control of signal transduction and transcription. *Prog. Nucleic Acid Res. Mol. Biol.* **73**, 107–136.
- Kim, J., Chu, J., Shen, X., Wang, J., and Orkin, S.H. (2008). An extended transcriptional network for pluripotency of embryonic stem cells. *Cell* **132**, 1049–1061.
- Lauberth, S.M., and Rauchman, M. (2006). A conserved 12-amino acid motif in Sall1 recruits the nucleosome remodeling and deacetylase corepressor complex. *J. Biol. Chem.* **281**, 23922–23931.
- Li, X., Sun, L., and Jin, Y. (2008). Identification of karyopherin- α 2 as an Oct4 associated protein. *J. Genet. Genomics* **35**, 723–728.
- Liang, J., Wan, M., Zhang, Y., Gu, P., Xin, H., Jung, S.Y., Qin, J., Wong, J., Cooney, A.J., Liu, D., and Songyang, Z. (2008). Nanog and Oct4 associate with unique transcriptional repression complexes in embryonic stem cells. *Nat. Cell Biol.* **10**, 731–739.
- Liu, P., Jenkins, N.A., and Copeland, N.G. (2003). A highly efficient recombining-based method for generating conditional knockout mutations. *Genome Res.* **13**, 476–484.
- Loh, Y.H., Wu, Q., Chew, J.L., Vega, V.B., Zhang, W., Chen, X., Bourque, G., George, J., Leong, B., Liu, J., et al. (2006). The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. *Nat. Genet.* **38**, 431–440.
- Luscombe, N.M., Babu, M.M., Yu, H., Snyder, M., Teichmann, S.A., and Gerstein, M. (2004). Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature* **431**, 308–312.
- Maherali, N., Sridharan, R., Xie, W., Utikal, J., Eminli, S., Arnold, K., Stadtfeld, M., Yachechko, R., Tchiew, J., Jaenisch, R., et al. (2007). Directly reprogrammed fibroblasts show global epigenetic remodeling and widespread tissue contribution. *Cell Stem Cell* **1**, 55–70.
- Martens, L., Hermjakob, H., Jones, P., Adamski, M., Taylor, C., States, D., Gevaert, K., Vandekerckhove, J., and Apweiler, R. (2005). PRIDE: The proteomics identifications database. *Proteomics* **5**, 3537–3545.
- Matoba, R., Niwa, H., Masui, S., Ohtsuka, S., Carter, M.G., Sharov, A.A., and Ko, M.S. (2006). Dissecting Oct3/4-regulated gene networks in embryonic stem cells by expression profiling. *PLoS ONE* **1**, e26.
- Mikkelsen, T.S., Hanna, J., Zhang, X., Ku, M., Wernig, M., Schorderet, P., Bernstein, B.E., Jaenisch, R., Lander, E.S., and Meissner, A. (2008). Dissecting direct reprogramming through integrative genomic analysis. *Nature* **454**, 49–55.
- Nichols, J., Zevnik, B., Anastasiadis, K., Niwa, H., Klewe-Nebenius, D., Chambers, I., Schöler, H., and Smith, A. (1998). Formation of pluripotent stem cells in the mammalian embryo depends on the POU transcription factor Oct4. *Cell* **95**, 379–391.
- Niwa, H., Miyazaki, J., and Smith, A.G. (2000). Quantitative expression of Oct-3/4 defines differentiation, dedifferentiation or self-renewal of ES cells. *Nat. Genet.* **24**, 372–376.
- Okita, K., Ichisaka, T., and Yamanaka, S. (2007). Generation of germline-competent induced pluripotent stem cells. *Nature* **448**, 313–317.
- Orkin, S.H., Wang, J., Kim, J., Chu, J., Rao, S., Theunissen, T.W., Shen, X., and Levasseur, D.N. (2008). The transcriptional network controlling pluripotency in ES cells. *Cold Spring Harb. Symp. Quant. Biol.* **73**, 195–202.
- Pan, G., Li, J., Zhou, Y., Zheng, H., and Pei, D. (2006). A negative feedback loop of transcription factors that controls stem cell pluripotency and self-renewal. *FASEB J.* **20**, 1730–1732.
- Poser, I., Sarov, M., Hutchins, J.R., Hériché, J.K., Toyoda, Y., Pozniakovskiy, A., Weigl, D., Nitzsche, A., Hegemann, B., Bird, A.W., et al. (2008). BAC TransgeneOmics: A high-throughput method for exploration of protein function in mammals. *Nat. Methods* **5**, 409–415.
- Prosser, H.M., Rzdzinska, A.K., Steel, K.P., and Bradley, A. (2008). Mosaic complementation demonstrates a regulatory role for myosin VIIa in actin dynamics of stereocilia. *Mol. Cell.* **28**, 1702–1712.
- Ramírez-Solis, R., Zheng, H., Whiting, J., Krumlauf, R., and Bradley, A. (1993). Hoxb-4 (Hox-2.6) mutant mice show homeotic transformation of a cervical vertebra and defects in the closure of the sternal rudiments. *Cell* **73**, 279–294.
- Reimand, J., Kull, M., Peterson, H., Hansen, J., and Vilo, J. (2007). g:Profiler—a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic Acids Res.* **35** (*Web Server issue*), W193–200.
- Rhodes, D.R., Kalyana-Sundaram, S., Mahavisno, V., Varambally, R., Yu, J., Briggs, B.B., Barrette, T.R., Anstet, M.J., Kincead-Beal, C., Kulkarni, P., et al. (2007). OncoPrint 3.0: Genes, pathways, and networks in a collection of 18,000 cancer gene expression profiles. *Neoplasia* **9**, 166–180.
- Rigaut, G., Shevchenko, A., Rutz, B., Wilm, M., Mann, M., and Séraphin, B. (1999). A generic protein purification method for protein complex characterization and proteome exploration. *Nat. Biotechnol.* **17**, 1030–1032.
- Rowley, A., Choudhary, J.S., Marzioch, M., Ward, M.A., Weir, M., Solari, R.C., and Blackstock, W.P. (2000). Applications of protein mass spectrometry in cell biology. *Methods* **20**, 383–397.
- Sandie, R., Palidwor, G.A., Huska, M.R., Porter, C.J., Krzyzanowski, P.M., Muro, E.M., Perez-Iratxeta, C., and Andrade-Navarro, M.A. (2009). Recent developments in StemBase: A tool to study gene expression in human and murine stem cells. *BMC Res Notes.* **2**, 39.
- Sato, N., Meijer, L., Skaltsounis, L., Greengard, P., and Brivanlou, A.H. (2004). Maintenance of pluripotency in human and mouse embryonic stem cells through activation of Wnt signaling by a pharmacological GSK-3-specific inhibitor. *Nat. Med.* **10**, 55–63.
- Schaft, J., Ashery-Padan, R., van der Hoeven, F., Gruss, P., and Stewart, A.F. (2001). Efficient FLP recombination in mouse ES cells and oocytes. *Genesis* **31**, 6–10.
- Shimozaki, K., Nakashima, K., Niwa, H., and Taga, T. (2003). Involvement of Oct3/4 in the enhancement of neuronal differentiation of ES cells in neurogenesis-inducing cultures. *Development* **130**, 2505–2512.
- Sterner, D.E., and Berger, S.L. (2000). Acetylation of histones and transcription-related factors. *Microbiol. Mol. Biol. Rev.* **64**, 435–459.
- Sun, C., Nakatake, Y., Akagi, T., Ura, H., Matsuda, T., Nishiyama, A., Koide, H., Ko, M.S., Niwa, H., and Yokota, T. (2009). Dax1 binds to Oct3/4 and inhibits its transcriptional activity in embryonic stem cells. *Mol. Cell. Biol.* **29**, 4574–4583.
- Takahashi, K., and Yamanaka, S. (2006). Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* **126**, 663–676.
- Takahashi, K., Tanabe, K., Ohnuki, M., Narita, M., Ichisaka, T., Tomoda, K., and Yamanaka, S. (2007). Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell* **131**, 861–872.
- Takao, Y., Yokota, T., and Koide, H. (2007). Beta-catenin up-regulates Nanog expression through interaction with Oct-3/4 in embryonic stem cells. *Biochem. Biophys. Res. Commun.* **353**, 699–705.
- Thomas, P.D., Kejariwal, A., Campbell, M.J., Mi, H., Diemer, K., Guo, N., Ladunga, I., Ulitsky-Lazareva, B., Muruganujan, A., Rabkin, S., et al. (2003). PANTHER: A browsable database of gene products organized by biological function, using curated protein family and subfamily classification. *Nucleic Acids Res.* **31**, 334–341.
- UniProt Consortium. (2010). The universal protein resource (UniProt) in 2010. *Nucleic Acids Res.* **38** (*Database issue*), D142–D148.
- van den Berg, D.L., Zhang, W., Yates, A., Engelen, E., Takacs, K., Bezstarosti, K., Demmers, J., Chambers, I., and Poot, R.A. (2008). Estrogen-related receptor beta interacts with Oct4 to positively regulate Nanog gene expression. *Mol. Cell. Biol.* **28**, 5986–5995.
- Van Hoof, D., Muñoz, J., Braam, S.R., Pinkse, M.W., Linding, R., Heck, A.J., Mummery, C.L., and Krijgsvel, J. (2009). Phosphorylation dynamics during early differentiation of human embryonic stem cells. *Cell Stem Cell* **5**, 214–226.
- Wang, J., Rao, S., Chu, J., Shen, X., Levasseur, D.N., Theunissen, T.W., and Orkin, S.H. (2006). A protein interaction network for pluripotency of embryonic stem cells. *Nature* **444**, 364–368.
- Wang, Z.X., Teh, C.H., Chan, C.M., Chu, C., Rossbach, M., Kunarso, G., Allapitchay, T.B., Wong, K.Y., and Stanton, L.W. (2008). The transcription factor Zfp281 controls embryonic stem cell pluripotency by direct activation and repression of target genes. *Stem Cells* **26**, 2791–2799.

Webster, D.M., Teo, C.F., Sun, Y., Wloga, D., Gay, S., Klonowski, K.D., Wells, L., and Dougan, S.T. (2009). O-GlcNAc modifications regulate cell survival and epiboly during zebrafish development. *BMC Dev. Biol.* 9, 28.

Wernig, M., Meissner, A., Foreman, R., Brambrink, T., Ku, M., Hochedlinger, K., Bernstein, B.E., and Jaenisch, R. (2007). In vitro reprogramming of fibroblasts into a pluripotent ES-cell-like state. *Nature* 448, 318–324.

Woltjen, K., Michael, I.P., Mohseni, P., Desai, R., Mileikovsky, M., Hämmäläinen, R., Cowling, R., Wang, W., Liu, P., Gertsenstein, M., et al. (2009). piggyBac transposition reprograms fibroblasts to induced pluripotent stem cells. *Nature* 458, 766–770.

Xu, H.M., Liao, B., Zhang, Q.J., Wang, B.B., Li, H., Zhong, X.M., Sheng, H.Z., Zhao, Y.X., Zhao, Y.M., and Jin, Y. (2004). Wwp2, an E3 ubiquitin ligase that targets transcription factor Oct-4 for ubiquitination. *J. Biol. Chem.* 279, 23495–23503.

Xu, H., Wang, W., Li, C., Yu, H., Yang, A., Wang, B., and Jin, Y. (2009). WWP2 promotes degradation of transcription factor OCT4 in human embryonic stem cells. *Cell Res.* 19, 561–573.

Yu, J., Vodyanik, M.A., Smuga-Otto, K., Antosiewicz-Bourget, J., Frane, J.L., Tian, S., Nie, J., Jonsdottir, G.A., Ruotti, V., Stewart, R., et al. (2007). Induced pluripotent stem cell lines derived from human somatic cells. *Science* 318, 1917–1920.

Yu, H.B., Kunarso, G., Hong, F.H., and Stanton, L.W. (2009). Zfp206, Oct4, and Sox2 are integrated components of a transcriptional regulatory network in embryonic stem cells. *J. Biol. Chem.* 284, 31327–31335.

Zeghouf, M., Li, J., Butland, G., Borkowska, A., Canadien, V., Richards, D., Beattie, B., Emili, A., and Greenblatt, J.F. (2004). Sequential Peptide Affinity (SPA) system for the identification of mammalian and bacterial protein complexes. *J. Proteome Res.* 3, 463–468.

7. Bibliography

Ackermann, D., Timpe, O., and Poller, K. (1929). Über das Anserin, einen neuen Bestandteil der Vogelmuskulatur. *Hoppe-Seylers Zeitschrift für Physiologische Chemie* 183, 1–10.

Acuna, R., Padilla, B.E., Florez-Ramos, C.P., Rubio, J.D., Herrera, J.C., Benavides, P., Lee, S.J., Yeats, T.H., Egan, A.N., Doyle, J.J., et al. (2012). Adaptive horizontal transfer of a bacterial gene to an invasive insect pest of coffee. *Proc Natl Acad Sci USA*.

Ahola, V., Aittokallio, T., Vihinen, M., and Uusipaikka, E. (2006). A statistical score for assessing the quality of multiple sequence alignments. *BMC Bioinformatics* 7, 484.

Albert, R., Jeong, H., and Barabasi, A. (2000). Error and attack tolerance of complex networks. *Nature* 406, 378–382.

Aloy, P., and Russell, R.B. (2006). Structural systems biology: modelling protein interactions. *Nat Rev Mol Cell Biol* 7, 188–197.

Altenhoff, A.M., and Dessimoz, C. (2009). Phylogenetic and functional assessment of orthologs inference projects and methods. *PLoS Comput Biol* 5, e1000262.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. *J Mol Biol* 215, 403–410.

Amberger, J., Bocchini, C.A., Scott, A.F., and Hamosh, A. (2009). McKusick's Online Mendelian Inheritance in Man (OMIM). *Nucleic Acids Res* 37, D793–D796.

Amoutzias, G.D., He, Y., Gordon, J., Mossialos, D., Oliver, S.G., and Van de Peer, Y. (2010). Posttranslational regulation impacts the fate of duplicated genes. *Proc Natl Acad Sci USA* 107, 2967–2971.

Anderson, J.C., Dueber, J.E., Leguia, M., Wu, G.C., Goler, J.A., Arkin, A.P., and Keasling, J.D. (2010). BglBricks: A flexible standard for biological part assembly. *J Biol Eng* 4, 1.

Aran-Guiu, X., Ortiz-Lombardía, M., Oliveira, E., Bonet Costa, C., Odena, M.A., Bellido, D., and Bernués, J. (2010). Acetylation of GAGA Factor Modulates Its Interaction with DNA. *Biochemistry*.

Aravind, L., Abhiman, S., and Iyer, L.M. (2011). Natural history of the eukaryotic chromatin protein methylation system. *Prog Mol Biol Transl Sci* 101, 105–176.

Arbely, E., Torres-Kolbus, J., Deiters, A., and Chin, J.W. (2012). Photocontrol of tyrosine phosphorylation in mammalian cells via genetic encoding of photocaged tyrosine. *J Am Chem Soc* 134, 11912–11915.

- Averof, M., Rokas, A., Wolfe, K.H., and Sharp, P.M. (2000). Evidence for a high frequency of simultaneous double-nucleotide substitutions. *Science* 287, 1283–1286.
- Awile, O., Krisko, A., Sbalzarini, I.F., and Zagrovic, B. (2010). Intrinsically disordered regions may lower the hydration free energy in proteins: a case study of nudix hydrolase in the bacterium *Deinococcus radiodurans*. *PLoS Comput Biol* 6, e1000854.
- Ayala, F.J. (1999). Molecular clock mirages. *BioEssays* 21, 71–75.
- Azuara, V., Perry, P., Sauer, S., Spivakov, M., Jørgensen, H.F., John, R.M., Gouti, M., Casanova, M., Warnes, G., Merckenschlager, M., et al. (2006). Chromatin signatures of pluripotent cell lines. *Nat Cell Biol* 8, 532–538.
- Ba, A.N.N., and Moses, A.M. (2010). Evolution of characterized phosphorylation sites in budding yeast. *Mol Biol Evol* 27, 2027–2037.
- Babu, M.M., Kriwacki, R.W., and Pappu, R.V. (2012). Versatility from Protein Disorder. *Science* 337, 1460–1461.
- Baghel, A.S., Tandon, R., Gupta, G., Kumar, A., Sharma, R.K., Aggarwal, N., Kathuria, A., Saini, N.K., Bose, M., Prasad, A.K., et al. (2011). Characterization of protein acyltransferase function of recombinant purified GlnA1 from *Mycobacterium tuberculosis*: A moon lighting property. *Microbiol. Res.* 166, 662–672.
- Baldauf, S.L. (2003). Phylogeny for the faint of heart: a tutorial. *Trends Genet* 19, 345–351.
- Baldwin, G.S., and Carnegie, P.R. (1971). Isolation and partial characterization of methylated arginines from the encephalitogenic basic protein of myelin. *Biochem J* 123, 69–74.
- Baliban, R.C., DiMaggio, P.A., Plazas-Mayorca, M.D., Young, N.L., Garcia, B.A., and Floudas, C.A. (2010). A novel approach for untargeted post-translational modification identification using integer linear optimization and tandem mass spectrometry. *Mol Cell Proteomics* 9, 764–779.
- Ballman, K.V. (2008). Genetics and genomics: gene expression microarrays. *Circulation* 118, 1593–1597.
- Bannister, A.J., and Kouzarides, T. (2011). Regulation of chromatin by histone modifications. *Cell Res* 21, 381–395.
- Bansal, S., Ponnan, P., Raj, H.G., Weintraub, S.T., Chopra, M., Kumari, R., Saluja, D., Kumar, A., Tyagi, T.K., Singh, P., et al. (2008). Autoacetylation of Purified Calreticulin Transacetylase Utilizing Acetoxycoumarin as the Acetyl Group Donor. *Appl Biochem Biotechnol* 157, 285–298.

Bantscheff, M., Hopf, C., Savitski, M.M., Dittmann, A., Grandi, P., Michon, A.-M., Schlegl, J., Abraham, Y., Becher, I., Bergamini, G., et al. (2011). Chemoproteomics profiling of HDAC inhibitors reveals selective targeting of HDAC complexes. *Nat Biotechnol* 29, 255–265.

Bantscheff, M., Lemeer, S., Savitski, M.M., and Kuster, B. (2012). Quantitative mass spectrometry in proteomics: critical review update from 2007 to the present. *Anal Bioanal Chem* 404, 939–965.

Bantscheff, M., Schirle, M., Sweetman, G., Rick, J., and Kuster, B. (2007). Quantitative mass spectrometry in proteomics: a critical review. *Anal Bioanal Chem* 389, 1017–1031.

Bao, J., and Sack, M.N. (2010). Protein deacetylation by sirtuins: delineating a post-translational regulatory program responsive to nutrient and redox stressors. *Cell Mol Life Sci* 67, 3073–3087.

Barabási, A.-L., and Oltvai, Z.N. (2004). Network biology: understanding the cell's functional organization. *Nat Rev Genet* 5, 101–113.

Barberi, T., Willis, L.M., Socci, N.D., and Studer, L. (2005). Derivation of multipotent mesenchymal precursors from human embryonic stem cells. *PLoS Med* 2, e161.

Barrie, E.S., Smith, R.M., Sanford, J.C., and Sadee, W. (2012). mRNA transcript diversity creates new opportunities for pharmacological intervention. *Mol. Pharmacol.* 81, 620–630.

Bartke, T., and Kouzarides, T. (2011). Decoding the chromatin modification landscape. *Cell Cycle* 10, 182.

Bartke, T., Vermeulen, M., Xhemalce, B., Robson, S.C., Mann, M., and Kouzarides, T. (2010). Nucleosome-interacting proteins regulated by DNA and histone methylation. *Cell* 143, 470–484.

Bashor, C.J., Horwitz, A.A., Peisajovich, S.G., and Lim, W.A. (2010). Rewiring cells: synthetic biology as a tool to interrogate the organizational principles of living systems. *Annu Rev Biophys* 39, 515–537.

Bellay, J., Han, S., Michaut, M., Kim, T., Costanzo, M., Andrews, B.J., Boone, C., Bader, G.D., Myers, C.L., and Kim, P.M. (2011). Bringing order to protein disorder through comparative genomics and genetic interactions. *Genome Biol* 12, R14.

Belmonte, J., Ellis, J., Hochedlinger, K., and Yamanaka, S. (2009). Induced pluripotent stem cells and reprogramming: seeing the science through the hype. *Nat Rev Genet*.

- Belozherov, V.E., Lin, Z.-Y., Gingras, A.-C., McDermott, J.C., and Michael Siu, K.W. (2012). High-resolution protein interaction map of the *Drosophila melanogaster* p38 mitogen-activated protein kinases reveals limited functional redundancy. *Mol Cell Biol* 32, 3695–3706.
- Beltrao, P., Albanèse, V., Kenner, L.R., Swaney, D.L., Burlingame, A., Villén, J., Lim, W.A., Fraser, J.S., Frydman, J., and Krogan, N.J. (2012). Systematic functional prioritization of protein posttranslational modifications. *Cell* 150, 413–425.
- Beltrao, P., Trinidad, J.C., Fiedler, D., Roguev, A., Lim, W.A., Shokat, K.M., Burlingame, A.L., and Krogan, N.J. (2009). Evolution of phosphoregulation: comparison of phosphorylation patterns across yeast species. *PLoS Biol* 7, e1000134.
- Benayoun, B.A., and Veitia, R.A. (2009). A post-translational modification code for transcription factors: sorting through a sea of signals. *Trends Cell Biol* 19, 189–197.
- Bianco, A., Townsley, F.M., Greiss, S., Lang, K., and Chin, J.W. (2012). Expanding the genetic code of *Drosophila melanogaster*. *Nat Chem Biol* 8, 748–750.
- Bischoff, R., and Schlüter, H. (2012). Amino acids: Chemistry, functionality and selected non-enzymatic post-translational modifications. *J Proteomics* 75, 2275–2296.
- Blake, J.A., Bult, C.J., Eppig, J.T., Kadin, J.A., Richardson, J.E., and Mouse Genome Database Group (2009). The Mouse Genome Database genotypes::phenotypes. *Nucleic Acids Res* 37, D712–D719.
- Blifernéz, O., Wobbe, L., Niehaus, K., and Kruse, O. (2011). Protein arginine methylation modulates light-harvesting antenna translation in *Chlamydomonas reinhardtii*. *Plant J* 65, 119–130.
- Blom, N., Sicheritz-Pontén, T., Gupta, R., Gammeltoft, S., and Brunak, S. (2004). Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence. *Proteomics* 4, 1633–1649.
- Bogdanov, B., and Smith, R.D. (2005). Proteomics by FTICR mass spectrometry: top down and bottom up. *Mass Spectrom. Rev.* 24, 168–200.
- Borg, M., Mittag, T., Pawson, T., Tyers, M., Forman-Kay, J.D., and Chan, H.S. (2007). Polyelectrostatic interactions of disordered ligands suggest a physical basis for ultrasensitivity. *Proc Natl Acad Sci USA* 104, 9650–9655.
- Braberg, H., Webb, B.M., Tjioe, E., Pieper, U., Sali, A., and Madhusudhan, M.S. (2012). SALIGN: a web server for alignment of multiple protein sequences and structures. *Bioinformatics* 28, 2072–2073.

- Bradner, J.E., West, N., Grachan, M.L., Greenberg, E.F., Haggarty, S.J., Warnow, T., and Mazitschek, R. (2010). Chemical phylogenetics of histone deacetylases. *Nat Chem Biol* 6, 238–243.
- Bridgham, J.T., Ortlund, E.A., and Thornton, J.W. (2009). An epistatic ratchet constrains the direction of glucocorticoid receptor evolution. *Nature* 461, 515–519.
- Broberg, C.A., and Orth, K. (2010). Tipping the balance by manipulating post-translational modifications. *Curr Opin Microbiol.* 13, 34–40.
- Brown, C.J., Johnson, A.K., and Daughdrill, G.W. (2010). Comparing models of evolution for ordered and disordered proteins. *Mol Biol Evol* 27, 609–621.
- Brown, C.J., Johnson, A.K., Dunker, A.K., and Daughdrill, G.W. (2011). Evolution and disorder. *Curr Opin Struct Biol.*
- Brown, C.J., Takayama, S., Campen, A.M., Vise, P., Marshall, T.W., Oldfield, C.J., Williams, C.J., and Dunker, A.K. (2002). Evolutionary rate heterogeneity in proteins with long disordered regions. *J Mol Evol* 55, 104–110.
- Buljan, M., Chalancon, G., Eustermann, S., Wagner, G.P., Fuxreiter, M., Bateman, A., and Babu, M.M. (2012). Tissue-Specific Splicing of Disordered Segments that Embed Binding Motifs Rewires Protein Interaction Networks. *Mol Cell* 46, 871–883.
- Caffrey, D.R., Somaroo, S., Hughes, J.D., Mintseris, J., and Huang, E.S. (2004). Are protein-protein interfaces more conserved in sequence than the rest of the protein surface? *Protein Sci* 13, 190–202.
- Cai, L., Sutter, B.M., Li, B., and Tu, B.P. (2011). Acetyl-CoA induces cell growth and proliferation by promoting the acetylation of histones at growth genes. *Mol Cell* 42, 426–437.
- Campbell, P.A., Perez-Iratxeta, C., Andrade-Navarro, M.A., and Rudnicki, M.A. (2007). Oct4 targets regulatory nodes to modulate stem cell function. *PLoS ONE* 2, e553.
- Campos, E.I., and Reinberg, D. (2009). Histones: annotating chromatin. *Annu Rev Genet* 43, 559–599.
- Cantoni, G.L. (1975). Biological methylation: selected aspects. *Annu Rev Biochem* 44, 435–451.
- Capra, J.A., and Singh, M. (2007). Predicting functionally important residues from sequence conservation. *Bioinformatics* 23, 1875–1882.

- Caron, C., Boyault, C., and Khochbin, S. (2005). Regulatory cross-talk between lysine acetylation and ubiquitination: role in the control of protein stability. *BioEssays* 27, 408–415.
- Carr, S.M., Munro, S., Kessler, B., Oppermann, U., and La Thangue, N.B. (2010). Interplay between lysine methylation and Cdk phosphorylation in growth control by the retinoblastoma protein. *Embo J*.
- Catic, A., Collins, C., Church, G.M., and Ploegh, H.L. (2004). Preferred in vivo ubiquitination sites. *Bioinformatics* 20, 3302–3307.
- Cerone, L., Muñoz-Garcia, J., and Neufeld, Z. (2011). Integrating multiple signals into cell decisions by networks of protein modification cycles. *Biophys J* 101, 1590–1596.
- Chalancon, G., Ravarani, C.N.J., Balaji, S., Martinez Arias, A., Aravind, L., Jothi, R., and Babu, M.M. (2012). Interplay between gene expression noise and regulatory network architecture. *Trends Genet* 28, 221–232.
- Chauveau, A., Aucher, A., Eissmann, P., Vivier, E., and Davis, D.M. (2010). Membrane nanotubes facilitate long-distance interactions between natural killer cells and target cells. *Proc Natl Acad Sci USA* 107, 5545–5550.
- Chen, C., Nott, T.J., Jin, J., and Pawson, T. (2011). Deciphering arginine methylation: Tudor tells the tale. *Nat Rev Mol Cell Biol* 12, 629–642.
- Chen, S.C.-C., Chen, F.-C., and Li, W.-H. (2010). Phosphorylated and nonphosphorylated serine and threonine residues evolve at different rates in mammals. *Mol Biol Evol* 27, 2548–2554.
- Chin, J.W. (2011). Reprogramming the genetic code. *Embo J* 30, 2312–2324.
- Chin, J.W. (2012). Molecular biology. Reprogramming the genetic code. *Science* 336, 428–429.
- Chin, M.H., Mason, M.J., Xie, W., Volinia, S., Singer, M., Peterson, C., Ambartsumyan, G., Aimiwu, O., Richter, L., Zhang, J., et al. (2009). Induced pluripotent stem cells and embryonic stem cells are distinguished by gene expression signatures. *Cell Stem Cell* 5, 111–123.
- Chothia, C., and Janin, J. (1975). Principles of protein-protein recognition. *Nature* 256, 705–708.
- Chothia, C. (2003). Evolution of the Protein Repertoire. *Science* 300, 1701–1703.
- Chothia, C., and Gough, J. (2009). Genomic and structural aspects of protein evolution. *Biochem J* 419, 15–28.

Choudhary, C., Kumar, C., Gnad, F., Nielsen, M.L., Rehman, M., Walther, T.C., Olsen, J.V., and Mann, M. (2009). Lysine acetylation targets protein complexes and co-regulates major cellular functions. *Science* 325, 834–840.

Cline, M.S., Smoot, M., Cerami, E., Kuchinsky, A., Landys, N., Workman, C., Christmas, R., Avila-Campilo, I., Creech, M., Gross, B., et al. (2007). Integration of biological networks and gene expression data using Cytoscape. *Nat Protoc* 2, 2366–2382.

Consortium, 1.G.P., Durbin, R.M., Abecasis, G.R., Altshuler, D.L., Auton, A., Brooks, L.D., Durbin, R.M., Gibbs, R.A., Hurles, M.E., and McVean, G.A. (2010). A map of human genome variation from population-scale sequencing. *Nature* 467, 1061–1073.

Creixell, P., Schoof, E.M., Tan, C.S.H., and Linding, R. (2012). Mutational properties of amino acid residues: implications for evolvability of phosphorylatable residues. *Philos Trans R Soc Lond, B, Biol Sci* 367, 2584–2593.

Creton, S., and Jentsch, S. (2010). SnapShot: The SUMO System. *Cell* 143, 848–848.e1.

Cusick, M.E., Klitgord, N., Vidal, M., and Hill, D.E. (2005). Interactome: gateway into systems biology. *Hum Mol Genet* 14 Spec No. 2, R171–R181.

Daily, K., Radivojac, P., and Dunker, A. (2005). Intrinsic disorder and protein modifications: building an SVM predictor for methylation. *IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology, CIBCB* 475–481.

Danchin, E.G.J., Rosso, M.-N., Vieira, P., de Almeida-Engler, J., Coutinho, P.M., Henrissat, B., and Abad, P. (2010). Multiple lateral gene transfers and duplications have promoted plant parasitism ability in nematodes. *Proc Natl Acad Sci USA* 107, 17651–17656.

Dang, C. (1999). Function of the c-Myc Oncogenic Transcription Factor. *Experimental Cell Research* 253, 63–77.

Danielsen, J.M.R., Sylvestersen, K.B., Bekker-Jensen, S., Szklarczyk, D., Poulsen, J.W., Horn, H., Jensen, L.J., Mailand, N., and Nielsen, M.L. (2010). Mass spectrometric analysis of lysine ubiquitylation reveals promiscuity at site level. *Mol Cell Proteomics*.

Davey, N.E., Haslam, N.J., Shields, D.C., and Edwards, R.J. (2011). SLiMSearch 2.0: biological context for short linear motifs in proteins. *Nucleic Acids Res*.

Davis, L., and Chin, J.W. (2012). Designer proteins: applications of genetic code expansion in cell biology. *Nat Rev Mol Cell Biol* 13, 168–182.

Davis, M.J., Shin, C.J., Jing, N., and Ragan, M.A. (2012). Rewiring the dynamic interactome. *Mol Biosyst* 8, 2054–66–2013.

- De, S., López-Bigas, N., and Teichmann, S.A. (2008). Patterns of evolutionary constraints on genes in humans. *BMC Evol Biol* 8, 275.
- Deng, X., Eickholt, J., and Cheng, J. (2009). PreDisorder: ab initio sequence-based prediction of protein disordered regions. *BMC Bioinformatics* 10, 436.
- Deng, X., Eickholt, J., and Cheng, J. (2012). A comprehensive overview of computational protein disorder prediction methods. *Mol Biosyst* 8, 114–121.
- Díaz-Moreno, I., Hollingworth, D., Frenkiel, T.A., Kelly, G., Martin, S., Howell, S., García-Mayoral, M., Gherzi, R., Briata, P., and Ramos, A. (2009). Phosphorylation-mediated unfolding of a KH domain regulates KSRP localization via 14-3-3 binding. *Nat Struct Mol Biol* 16, 238–246.
- Dinkel, H., Chica, C., Via, A., Gould, C.M., Jensen, L.J., Gibson, T.J., and Diella, F. (2011). Phospho.ELM: a database of phosphorylation sites--update 2011. *Nucleic Acids Res* 39, D261–D267.
- Diss, G., Freschi, L., and Landry, C.R. (2012). Where do phosphosites come from and where do they go after gene duplication? *Int J Evol Biol* 2012, 843167.
- Dittrich, M.T., Klau, G.W., Rosenwald, A., Dandekar, T., and Müller, T. (2008). Identifying functional modules in protein-protein interaction networks: an integrated exact approach. *Bioinformatics* 24, i223–i231.
- Dosztányi, Z., Csizmók, V., Tompa, P., and Simon, I. (2005). The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *J Mol Biol* 347, 827–839.
- Du, J., Zhou, Y., Su, X., Yu, J.J., Khan, S., Jiang, H., Kim, J., Woo, J., Kim, J.H., Choi, B.H., et al. (2011). Sirt5 is a NAD-dependent protein lysine demalonylase and desuccinylase. *Science* 334, 806–809.
- Dunker, A.K., Silman, I., Uversky, V.N., and Sussman, J.L. (2008). Function and structure of inherently disordered proteins. *Curr Opin Struct Biol* 18, 756–764.
- Dunning Hotopp, J.C. (2011). Horizontal gene transfer between bacteria and animals. *Trends Genet* 27, 157–163.
- Edwards, R.J., Davey, N.E., Brien, K.O., and Shields, D.C. (2011). Interactome-wide prediction of short, disordered protein interaction motifs in humans. *Mol Biosyst* 8, 282.
- Ellgaard, L., Molinari, M., and Helenius, A. (1999). Setting the standards: quality control in the secretory pathway. *Science* 286, 1882–1888.

Ellis, J.M., and Wolfgang, M.J. (2012). A Genetically Encoded Metabolite Sensor for Malonyl-CoA. *Chem Biol* 19, 1333–1339.

Ellis, J.D., Barrios-Rodiles, M., Çolak, R., Irimia, M., Kim, T., Calarco, J.A., Wang, X., Pan, Q., O'Hanlon, D., Kim, P.M., et al. (2012). Tissue-specific alternative splicing remodels protein-protein interaction networks. *Mol Cell* 46, 884–892.

Enver, T., Pera, M., Peterson, C., and Andrews, P.W. (2009). Stem cell states, fates, and the rules of attraction. *Cell Stem Cell* 4, 387–397.

Erazo, A., Yee, M.B., Banfield, B.W., and Kinchington, P.R. (2011). The alphaherpesvirus US3/ORF66 protein kinases direct phosphorylation of the nuclear matrix protein matrin 3. *Journal of Virology* 85, 568–581.

Erickson, S.L., and Lykke-Andersen, J. (2011). Cytoplasmic mRNP granules at a glance. *J Cell Sci* 124, 293–297.

Erkman, J.A., and Kaufman, P.D. (2009). A negatively charged residue in place of histone H3K56 supports chromatin assembly factor association but not genotoxic stress resistance. *DNA Repair (Amst)* 8, 1371–1379.

Eswar, N., Webb, B., Martí-Renom, M.A., Madhusudhan, M.S., Eramian, D., Shen, M.-Y., Pieper, U., and Sali, A. (2007). Comparative protein structure modeling using MODELLER. *Current Protocols in Protein Science / Editorial Board, John E Coligan [Et Al] Chapter 2, Unit2.9.*

Evans, M.J., and Kaufman, M.H. (1981). Establishment in culture of pluripotential cells from mouse embryos. *Nature* 292, 154–156.

Fantini, D., Vascotto, C., Marasco, D., D'Ambrosio, C., Romanello, M., Vitagliano, L., Pedone, C., Poletto, M., Cesaratto, L., Quadrioglio, F., et al. (2010). Critical lysine residues within the overlooked N-terminal domain of human APE1 regulate its biological functions. *Nucleic Acids Res.*

Farzan, M., Mirzabekov, T., Kolchinsky, P., Wyatt, R., Cayabyab, M., Gerard, N.P., Gerard, C., Sodroski, J., and Choe, H. (1999). Tyrosine sulfation of the amino terminus of CCR5 facilitates HIV-1 entry. *Cell* 96, 667–676.

Feng, B., Li, L., Zhou, X., Stanley, B., and Ma, H. (2009). Analysis of the Arabidopsis floral proteome: detection of over 2 000 proteins and evidence for posttranslational modifications. *J Integr Plant Biol* 51, 207–223.

- Filion, G.J., van Bommel, J.G., Braunschweig, U., Talhout, W., Kind, J., Ward, L.D., Brugman, W., de Castro, I.J., Kerkhoven, R.M., Bussemaker, H.J., et al. (2010). Systematic protein location mapping reveals five principal chromatin types in *Drosophila* cells. *Cell* *143*, 212–224.
- Filion, G.J.P., Zhenilo, S., Salozhin, S., Yamada, D., Prokhortchouk, E., and Defossez, P.-A. (2006). A family of human zinc finger proteins that bind methylated DNA and repress transcription. *Mol Cell Biol* *26*, 169–181.
- Filippakopoulos, P., Picaud, S., Mangos, M., Keates, T., Lambert, J.-P., Barsyte-Lovejoy, D., Felletar, I., Volkmer, R., Müller, S., Pawson, T., et al. (2012). Histone recognition and large-scale structural analysis of the human bromodomain family. *Cell* *149*, 214–231.
- Fischle, W. (2008). Talk is cheap--cross-talk in establishment, maintenance, and readout of chromatin modifications. *Genes Dev* *22*, 3375–3382.
- Flicek, P., Amode, M.R., Barrell, D., Beal, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S., et al. (2012). Ensembl 2012. *Nucleic Acids Res* *40*, D84–D90.
- Fong, J.H., Shoemaker, B.A., Garbuzynskiy, S.O., Lobanov, M.Y., Galzitskaya, O.V., and Panchenko, A.R. (2009). Intrinsic disorder in protein interactions: insights from a comprehensive structural analysis. *PLoS Comput Biol* *5*, e1000316.
- Frankel, A.D., and Kim, P.S. (1991). Modular structure of transcription factors: implications for gene regulation. *Cell* *65*, 717–719.
- Fraser, D., and Kaern, M. (2009). A chance at survival: gene expression noise and phenotypic diversification strategies. *Mol Microbiol* *71*, 1333–1340.
- Fraser, J.S., Gross, J.D., and Krogan, N.J. (2013). From systems to structure: bridging networks and mechanism. *Mol Cell* *49*, 222–231.
- Freschi, L., Courcelles, M., Thibault, P., Michnick, S.W., and Landry, C.R. (2011). Phosphorylation network rewiring by gene duplication. *Mol Syst Biol* *7*, 504.
- Futreal, P.A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., Rahman, N., and Stratton, M.R. (2004). A census of human cancer genes. *Nat Rev Cancer* *4*, 177–183.
- Fuxreiter, M., Tompa, P., and Simon, I. (2007). Local structural disorder imparts plasticity on linear motifs. *Bioinformatics* *23*, 950–956.
- Gallagher, K.L., and Benfey, P.N. (2005). Not just another hole in the wall: understanding intercellular protein trafficking. *Genes Dev* *19*, 189–195.

Gan, X., Stegle, O., Behr, J., Steffen, J.G., Drewe, P., Hildebrand, K.L., Lyngsoe, R., Schultheiss, S.J., Osborne, E.J., Sreedharan, V.T., et al. (2011). Multiple reference genomes and transcriptomes for *Arabidopsis thaliana*. *Nature* 477, 419–423.

Garcia-Ruiz, H., Takeda, A., Chapman, E.J., Sullivan, C.M., Fahlgren, N., Brempelis, K.J., and Carrington, J.C. (2010). *Arabidopsis* RNA-dependent RNA polymerases and dicer-like proteins in antiviral defense and small interfering RNA biogenesis during Turnip Mosaic Virus infection. *Plant Cell* 22, 481–496.

Gautier, A., Deiters, A., and Chin, J.W. (2011). Light-activated kinases enable temporal dissection of signaling networks in living cells. *J Am Chem Soc* 133, 2124–2127.

Gautier, A., Nguyen, D.P., Lusic, H., An, W., Deiters, A., and Chin, J.W. (2010). Genetically encoded photocontrol of protein localization in mammalian cells. *J Am Chem Soc* 132, 4086–4088.

Gfeller, D., Butty, F., Wierzbicka, M., Verschueren, E., Vanhee, P., Huang, H., Ernst, A., Dar, N., Stagljar, I., Serrano, L., et al. (2011). The multiple-specificity landscape of modular peptide recognition domains. *Mol Syst Biol* 7, 484.

Ghosh-Roy, A., Goncharov, A., Jin, Y., and Chisholm, A.D. (2012). Kinesin-13 and Tubulin Posttranslational Modifications Regulate Microtubule Growth in Axon Regeneration. *Developmental Cell*.

Giot, L., Bader, J.S., Brouwer, C., Chaudhuri, A., Kuang, B., Li, Y., Hao, Y.L., Ooi, C.E., Godwin, B., Vitols, E., et al. (2003). A protein interaction map of *Drosophila melanogaster*. *Science* 302, 1727–1736.

Glozak, M.A., Sengupta, N., Zhang, X., and Seto, E. (2005). Acetylation and deacetylation of non-histone proteins. *Gene* 363, 15–23.

Gnad, F., de Godoy, L.M.F., Cox, J., Neuhauser, N., Ren, S., Olsen, J.V., and Mann, M. (2009). High-accuracy identification and bioinformatic analysis of in vivo protein phosphorylation sites in yeast. *Proteomics* 9, 4642–4652.

Gnad, F., Forner, F., Zielinska, D.F., Birney, E., Gunawardena, J., and Mann, M. (2010a). Evolutionary constraints of phosphorylation in eukaryotes, prokaryotes, and mitochondria. *Mol Cell Proteomics* 9, 2642–2653.

Gnad, F., Gunawardena, J., and Mann, M. (2010b). PHOSIDA 2011: the posttranslational modification database. *Nucleic Acids Res*.

- Gnad, F., Ren, S., Cox, J., Olsen, J.V., Macek, B., Oroshi, M., and Mann, M. (2007). PHOSIDA (phosphorylation site database): management, structural and evolutionary investigation, and prediction of phosphosites. *Genome Biol* 8, R250.
- Golebiowski, F., Matic, I., Tatham, M.H., Cole, C., Yin, Y., Nakamura, A., Cox, J., Barton, G.J., Mann, M., and Hay, R.T. (2009). System-wide changes to SUMO modifications in response to heat shock. *Sci Signal* 2, ra24.
- Golubchik, T., Wise, M.J., Eastal, S., and Jermin, L.S. (2007). Mind the gaps: evidence of bias in estimates of multiple sequence alignments. *Mol Biol Evol* 24, 2433–2442.
- Gough, J., Karplus, K., Hughey, R., and Chothia, C. (2001). Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. *J Mol Biol* 313, 903–919.
- Grabbe, C., Husnjak, K., and Dikic, I. (2011). The spatial and temporal organization of ubiquitin networks. *Nat Rev Mol Cell Biol* 12, 295–307.
- Grangeasse, C., Nessler, S., and Mijakovic, I. (2012). Bacterial tyrosine kinases: evolution, biological function and structural insights. *Philos Trans R Soc Lond, B, Biol Sci* 367, 2640–2655.
- Gray, V.E., and Kumar, S. (2011). Rampant Purifying Selection Conserves Positions with Post-Translational Modifications in Human Proteins. *Mol Biol Evol*.
- Greiss, S., and Chin, J.W. (2011). Expanding the genetic code of an animal. *J Am Chem Soc* 133, 14196–14199.
- Grönroos, E., Hellman, U., Heldin, C.-H., and Ericsson, J. (2002). Control of Smad7 stability by competition between acetylation and ubiquitination. *Mol Cell* 10, 483–493.
- Gsponer, J., and Babu, M.M. (2009). The rules of disorder or why disorder rules. *Prog Biophys Mol Biol* 99, 94–103.
- Gsponer, J., Futschik, M.E., Teichmann, S.A., and Babu, M.M. (2008). Tight regulation of unstructured proteins: from transcript synthesis to protein degradation. *Science* 322, 1365–1368.
- Guan, K.-L., Yu, W., Lin, Y., Xiong, Y., and Zhao, S. (2010). Generation of acetylysine antibodies and affinity enrichment of acetylated peptides. *Nat Protoc* 5, 1583–1595.
- Guindon, S., and Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* 52, 696–704.

Guo, Q., Bedford, M.T., and Fast, W. (2011). Discovery of peptidylarginine deiminase-4 substrates by protein array: antagonistic citrullination and methylation of human ribosomal protein S2. *Mol Biosyst* 7, 2286–2295.

Guttman, M., Donaghey, J., Carey, B.W., Garber, M., Grenier, J.K., Munson, G., Young, G., Lucas, A.B., Ach, R., Bruhn, L., et al. (2011). lincRNAs act in the circuitry controlling pluripotency and differentiation. *Nature*.

Haensch, S., Bianucci, R., Signoli, M., Rajerison, M., Schultz, M., Kacki, S., Vermunt, M., Weston, D.A., Hurst, D., Achtman, M., et al. (2010). Distinct clones of *Yersinia pestis* caused the black death. *PLoS Pathog* 6, e1001134.

Hagai, T., and Levy, Y. (2010). Ubiquitin not only serves as a tag but also assists degradation by inducing protein unfolding. *Proc Natl Acad Sci USA* 107, 2001–2006.

Hagai, T., Azia, A., Tóth-Petróczy, A., and Levy, Y. (2011). Intrinsic disorder in ubiquitination substrates. *J Mol Biol* 412, 319–324.

Hagai, T., Tóth-Petróczy, A., Azia, A., and Levy, Y. (2012). The origins and evolution of ubiquitination sites. *Mol Biosyst* 8, 1865–1877.

Hamamoto, R., Cho, H.-S., Suzuki, T., Dohmae, N., Hayami, S., Unoki, M., Yoshimatsu, M., Toyokawa, G., Takawa, M., Chen, T., et al. (2010). Demethylation of RB regulator MYPT1 by histone demethylase LSD1 promotes cell cycle progression in cancer cells. *Cancer Research*.

Hamosh, A., Scott, A.F., Amberger, J.S., Bocchini, C.A., and McKusick, V.A. (2005). Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res* 33, D514–D517.

Hancock, S.M., Uprety, R., Deiters, A., and Chin, J.W. (2010). Expanding the genetic code of yeast for incorporation of diverse unnatural amino acids via a pyrrolysyl-tRNA synthetase/tRNA pair. *J Am Chem Soc* 132, 14819–14824.

Hang, H.C., Wilson, J.P., and Charron, G. (2011). Bioorthogonal chemical reporters for analyzing protein lipidation and lipid trafficking. *Acc Chem Res* 44, 699–708.

Hanna, J.H., Saha, K., and Jaenisch, R. (2010). Pluripotency and cellular reprogramming: facts, hypotheses, unresolved issues. *Cell* 143, 508–525.

Hart, G.W., and Copeland, R.J. (2010). Glycomics hits the big time. *Cell* 143, 672–676.

Hart, G.W., Housley, M.P., and Slawson, C. (2007). Cycling of O-linked beta-N-acetylglucosamine on nucleocytoplasmic proteins. *Nature* 446, 1017–1022.

- Hartl, F.U., and Hayer-Hartl, M. (2009). Converging concepts of protein folding in vitro and in vivo. *Nat Struct Mol Biol* 16, 574–581.
- He, W., Newman, J.C., Wang, M.Z., Ho, L., and Verdin, E. (2012a). Mitochondrial sirtuins: regulators of protein acylation and metabolism. *Trends Endocrinol. Metab.* 23, 467–476.
- He, Y., Chen, Y., Alexander, P.A., Bryan, P.N., and Orban, J. (2012b). Mutational tipping points for switching protein folds and functions. *Structure* 20, 283–291.
- Hejjaoui, M., Haj-Yahya, M., Kumar, K.S.A., Brik, A., and Lashuel, H.A. (2010). Towards Elucidation of the Role of Ubiquitination in the Pathogenesis of Parkinson's Disease with Semisynthetic Ubiquitinated α -Synuclein. *Angew. Chem. Int. Ed.* 50, 405–409.
- Hilger, M., Bonaldi, T., Gnad, F., and Mann, M. (2009). Systems-wide analysis of a phosphatase knock-down by quantitative proteomics and phosphoproteomics. *Mol Cell Proteomics* 8, 1908–1920.
- Hlevnjak, M., Zitkovic, G., and Zagrovic, B. (2010). Hydrophilicity matching - a potential prerequisite for the formation of protein-protein complexes in the cell. *PLoS ONE* 5, e11169.
- Ho, L., and Crabtree, G.R. (2010). Chromatin remodelling during development. *Nature* 463, 474–484.
- Hockenberry, A.J., and Jewett, M.C. (2012). Synthetic in vitro circuits. *Curr Opin Chem Biol* 16, 253–259.
- Hodgkinson, A., and Eyre-Walker, A. (2010). Human triallelic sites: evidence for a new mutational mechanism? *Genetics* 184, 233–241.
- Holt, L.J., Tuch, B.B., Villén, J., Johnson, A.D., Gygi, S.P., and Morgan, D.O. (2009). Global analysis of Cdk1 substrate phosphorylation sites provides insights into evolution. *Science* 325, 1682–1686.
- Hornbeck, P.V., Kornhauser, J.M., Tkachev, S., Zhang, B., Skrzypek, E., Murray, B., Latham, V., and Sullivan, M. (2012). PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. *Nucleic Acids Res* 40, D261–D270.
- Hothorn, T., Hornik, K., van de Wiel, M.A., and Zeileis, A. (2008). Implementing a Class of Permutation Tests: The coin Package. *J Stat Softw* 28, 1–23.
- Huang, O.W., Ma, X., Yin, J., Flinders, J., Maurer, T., Kayagaki, N., Phung, Q., Bosanac, I., Arnott, D., Dixit, V.M., et al. (2012). Phosphorylation-dependent activity of the deubiquitinase DUBA. *Nat Struct Mol Biol* 19, 171–175.

Hudson, B.P., Martinez-Yamout, M.A., Dyson, H.J., and Wright, P.E. (2000). Solution structure and acetyl-lysine binding activity of the GCN5 bromodomain. *J Mol Biol* 304, 355–370.

Hulsen, T., Groenen, P.M.A., de Vlieg, J., and Alkema, W. (2009). PhyloPat: an updated version of the phylogenetic pattern database contains gene neighborhood. *Nucleic Acids Res* 37, D731–D737.

Husson, S.J., Clynen, E., Baggerman, G., De Loof, A., and Schoofs, L. (2005). Discovering neuropeptides in *Caenorhabditis elegans* by two dimensional liquid chromatography and mass spectrometry. *Biochem Biophys Res Commun* 335, 76–86.

Hwang, C.-S., Shemorry, A., and Varshavsky, A. (2010). N-terminal acetylation of cellular proteins creates specific degradation signals. *Science* 327, 973–977.

Hyland, E.M., Molina, H., Poorey, K., Jie, C., Xie, Z., Dai, J., Qian, J., Bekiranov, S., Auble, D.T., Pandey, A., et al. (2011). An evolutionarily “young” lysine residue in histone H3 attenuates transcriptional output in *Saccharomyces cerevisiae*. *Genes Dev* 25, 1306–1319.

Hyman, A.H., and Simons, K. (2011). The new cell biology: Beyond HeLa cells. *Nature* 480, 34.

Iakoucheva, L.M., Radivojac, P., Brown, C.J., O'Connor, T.R., Sikes, J.G., Obradovic, Z., and Dunker, A.K. (2004). The importance of intrinsic disorder for protein phosphorylation. *Nucleic Acids Res* 32, 1037–1049.

Ikeda, F., Crosetto, N., and Dikic, I. (2010). What determines the specificity and outcomes of ubiquitin signaling? *Cell* 143, 677–681.

Ingolia, N.T., Lareau, L.F., and Weissman, J.S. (2011). Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell* 147, 789–802.

Ishida, T., and Kinoshita, K. (2007). PrDOS: prediction of disordered protein regions from amino acid sequence. *Nucleic Acids Res* 35, W460–W464.

Iyer, L.M., Tahiliani, M., Rao, A., and Aravind, L. (2009). Prediction of novel families of enzymes involved in oxidative and other complex modifications of bases in nucleic acids. *Cell Cycle* 8, 1698–1710.

Jang, H., Kim, T.W., Yoon, S., Choi, S.-Y., Kang, T.-W., Kim, S.-Y., Kwon, Y.-W., Cho, E.-J., and Youn, H.-D. (2012). O-GlcNAc regulates pluripotency and reprogramming by directly acting on core components of the pluripotency network. *Cell Stem Cell* 11, 62–74.

- Jensen, L.J., Kuhn, M., Stark, M., Chaffron, S., Creevey, C., Muller, J., Doerks, T., Julien, P., Roth, A., Simonovic, M., et al. (2009). STRING 8--a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res* 37, D412–D416.
- Jeong, H., Then, F., Melia, T.J., Mazzulli, J.R., Cui, L., Savas, J.N., Voisine, C., Paganetti, P., Tanese, N., Hart, A.C., et al. (2009). Acetylation targets mutant huntingtin to autophagosomes for degradation. *Cell* 137, 60–72.
- Jiang, W., Wang, S., Xiao, M., Lin, Y., Zhou, L., Lei, Q., Xiong, Y., Guan, K.-L., and Zhao, S. (2011). Acetylation regulates gluconeogenesis by promoting PEPCK1 degradation via recruiting the UBR5 ubiquitin ligase. *Mol Cell* 43, 33–44.
- Jin, J., and Pawson, T. (2012). Modular evolution of phosphorylation-based signalling systems. *Philos Trans R Soc Lond, B, Biol Sci* 367, 2540–2555.
- Jin, Y.-H., Jeon, E.-J., Li, Q.-L., Lee, Y.H., Choi, J.-K., Kim, W.-J., Lee, K.-Y., and Bae, S.-C. (2004). Transforming growth factor-beta stimulates p300-dependent RUNX3 acetylation, which inhibits ubiquitination-mediated degradation. *J Biol Chem* 279, 29409–29417.
- Joerger, A.C., and Fersht, A.R. (2007). Structure-function-rescue: the diverse nature of common p53 cancer mutants. *Oncogene* 26, 2226–2242.
- Johansson, F., and Toh, H. (2010). A comparative study of conservation and variation scores. *BMC Bioinformatics* 11, 388.
- Jones, S., and Thornton, J.M. (1996). Principles of protein-protein interactions. *Proc Natl Acad Sci USA* 93, 13–20.
- Joughin, B.A., Liu, C., Lauffenburger, D.A., Hogue, C.W.V., and Yaffe, M.B. (2012). Protein kinases display minimal interpositional dependence on substrate sequence: potential implications for the evolution of signalling networks. *Philos Trans R Soc Lond, B, Biol Sci* 367, 2574–2583.
- Kaneko, T., Suzuki, T., Kapushoc, S.T., Rubio, M.A., Ghazvini, J., Watanabe, K., Simpson, L., and Suzuki, T. (2003). Wobble modification differences and subcellular localization of tRNAs in *Leishmania tarentolae*: implication for tRNA sorting mechanism. *Embo J* 22, 657–667.
- Kasten, M., Szerlong, H., Erdjument-Bromage, H., Tempst, P., Werner, M., and Cairns, B.R. (2004). Tandem bromodomains in the chromatin remodeler RSC recognize acetylated histone H3 Lys14. *Embo J* 23, 1348–1359.
- Katoh, K., and Toh, H. (2008). Recent developments in the MAFFT multiple sequence alignment program. *Brief Bioinformatics* 9, 286–298.

Kersten, R.D., Yang, Y.-L., Xu, Y., Cimermancic, P., Nam, S.-J., Fenical, W., Fischbach, M.A., Moore, B.S., and Dorrestein, P.C. (2011). A mass spectrometry-guided genome mining approach for natural product peptidogenomics. *Nat Chem Biol* 7, 794–802.

Khalil, A.M., Guttman, M., Huarte, M., Garber, M., Raj, A., Rivea Morales, D., Thomas, K., Presser, A., Bernstein, B.E., van Oudenaarden, A., et al. (2009). Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proc Natl Acad Sci USA* 106, 11667–11672.

Kim, D., Kim, C.-H., Moon, J.-I., Chung, Y.-G., Chang, M.-Y., Han, B.-S., Ko, S., Yang, E., Cha, K.Y., Lanza, R., et al. (2009a). Generation of human induced pluripotent stem cells by direct delivery of reprogramming proteins. *Cell Stem Cell* 4, 472–476.

Kim, D.-H., and Motter, A.E. (2009). Slave nodes and the controllability of metabolic networks. *arXiv q-bio.MN*.

Kim, J., Chu, J., Shen, X., Wang, J., and Orkin, S.H. (2008). An extended transcriptional network for pluripotency of embryonic stem cells. *Cell* 132, 1049–1061.

Kim, J.B., Sebastiano, V., Wu, G., Araúzo-Bravo, M.J., Sasse, P., Gentile, L., Ko, K., Ruau, D., Ehrlich, M., van den Boom, D., et al. (2009b). Oct4-induced pluripotency in adult neural stem cells. *Cell* 136, 411–419.

Kim, Y.-K., Heo, I., and Kim, V.N. (2010). Modifications of Small RNAs and Their Associated Proteins. *Cell* 143, 703–709.

Klugbauer, S., and Rabes, H.M. (1999). The transcription coactivator HTIF1 and a related protein are fused to the RET receptor tyrosine kinase in childhood papillary thyroid carcinomas. *Oncogene* 18, 4388–4393.

Komander, D. (2009). The emerging complexity of protein ubiquitination. *Biochem Soc Trans* 37, 937–953.

Komander, D., and Rape, M. (2012). The ubiquitin code. *Annu Rev Biochem* 81, 203–229.

Kosiol, C., Vinar, T., da Fonseca, R.R., Hubisz, M.J., Bustamante, C.D., Nielsen, R., and Siepel, A. (2008). Patterns of positive selection in six Mammalian genomes. *PLoS Genet* 4, e1000144.

Kouzarides, T. (2000). Acetylation: a regulatory modification to rival phosphorylation? *Embo J* 19, 1176–1179.

Kouzarides, T. (2007a). Chromatin modifications and their function. *Cell* 128, 693–705.

Kouzarides, T. (2007b). SnapShot: Histone-modifying enzymes. *Cell* 131, 822.

- Krieger, E., Joo, K., Lee, J., Lee, J., Raman, S., Thompson, J., Tyka, M., Baker, D., and Karplus, K. (2009). Improving physical realism, stereochemistry, and side-chain accuracy in homology modeling: Four approaches that performed well in CASP8. *Proteins 77 Suppl 9*, 114–122.
- Krissinel, E., and Henrick, K. (2005). Detection of protein assemblies in crystals. *Computational Life Sciences* 163–174.
- Krissinel, E., and Henrick, K. (2007). Inference of macromolecular assemblies from crystalline state. *J Mol Biol* 372, 774–797.
- Kumar, P., Henikoff, S., and Ng, P.C. (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc* 4, 1073–1081.
- Kurmangaliyev, Y.Z., Goland, A., and Gelfand, M.S. (2011). Evolutionary patterns of phosphorylated serines. *Biology Direct* 2010 5:6 6, 8.
- L'Italien, J.J., and Laursen, R.A. (1979). Location of the site of methylation in elongation factor Tu. *FEBS Lett* 107, 359–362.
- la Puerta, de, M.L., Trinidad, A.G., del Carmen Rodríguez, M., Bogetz, J., Sánchez Crespo, M., Mustelin, T., Alonso, A., and Bayón, Y. (2009). Characterization of new substrates targeted by Yersinia tyrosine phosphatase YopH. *PLoS ONE* 4, e4431.
- Lage, K., Hansen, N.T., Karlberg, E.O., Eklund, A.C., Roque, F.S., Donahoe, P.K., Szallasi, Z., Jensen, T.S., and Brunak, S. (2008). A large-scale analysis of tissue-specific pathology and gene expression of human disease genes and complexes. *Proc Natl Acad Sci USA* 105, 20870–20875.
- Lammers, M., Neumann, H., Chin, J.W., and James, L.C. (2010). Acetylation regulates cyclophilin A catalysis, immunosuppression and HIV isomerization. *Nat Chem Biol* 6, 331–337.
- Landry, C.R., Levy, E.D., and Michnick, S.W. (2009). Weak functional constraints on phosphoproteomes. *Trends Genet* 25, 193–197.
- Lange, O.F., Lakomek, N.-A., Farès, C., Schröder, G.F., Walter, K.F.A., Becker, S., Meiler, J., Grubmüller, H., Griesinger, C., and de Groot, B.L. (2008). Recognition dynamics up to microseconds revealed from an RDC-derived ubiquitin ensemble in solution. *Science* 320, 1471–1475.
- Latham, J.A., and Dent, S.Y.R. (2007). Cross-regulation of histone modifications. *Nat Struct Mol Biol* 14, 1017–1024.

Lee, A.H.-Y., Hurley, B., Felsensteiner, C., Yea, C., Ckurshumova, W., Bartetzko, V., Wang, P.W., Van Quach, Lewis, J.D., Liu, Y.C., et al. (2012). A Bacterial Acetyltransferase Destroys Plant Microtubule Networks and Blocks Secretion. *PLoS Pathog* 8, e1002523.

Levskaya, A., Weiner, O.D., Lim, W.A., and Voigt, C.A. (2009). Spatiotemporal control of cell signalling using a light-switchable protein interaction. *Nature* 461, 997–1001.

Levy, E.D., and Pereira-Leal, J.B. (2008). Evolution and dynamics of protein interactions and networks. *Curr Opin Struct Biol* 18, 349–357.

Levy, E.D. (2010). A simple definition of structural regions in proteins and its use in analyzing interface evolution. *J Mol Biol* 403, 660–670.

Levy, D., Liu, C.L., Yang, Z., Newman, A.M., Alizadeh, A.A., Utz, P.J., and Gozani, O. (2011). A proteomic approach for the identification of novel lysine methyltransferase substrates. *Epigenetics & Chromatin* 2010 3:4 4, 19.

Levy, E.D., Michnick, S.W., and Landry, C.R. (2012). Protein abundance is key to distinguish promiscuous from functional phosphorylation based on evolutionary information. *Philos Trans R Soc Lond, B, Biol Sci* 367, 2594–2606.

Levy, E.D., De, S., and Teichmann, S.A. (2012). Cellular crowding imposes global constraints on the chemistry and evolution of proteomes. *Proc Natl Acad Sci USA* 109, 20461–20466.

Li, D., Li, J., Ouyang, S., Wang, J., Wu, S., Wan, P., Zhu, Y., Xu, X., and He, F. (2006). Protein interaction networks of *Saccharomyces cerevisiae*, *Caenorhabditis elegans* and *Drosophila melanogaster*: Large-scale organization and robustness. *Proteomics* 6, 456–461.

Li, S., Iakoucheva, L.M., Mooney, S.D., and Radivojac, P. (2010). Loss of post-translational modification sites in disease. *Pac Symp Biocomput* 337–347.

Liang, W., Malhotra, A., and Deutscher, M.P. (2011). Acetylation regulates the stability of a bacterial protein: growth stage-dependent modification of RNase R. *Mol Cell* 44, 160–166.

Liarzi, O., Barak, R., Bronner, V., Dines, M., Sagi, Y., Shainskaya, A., and Eisenbach, M. (2010). Acetylation represses the binding of CheY to its target proteins. *Mol Microbiol* 76, 932–943.

Lienhard, G.E. (2008). Non-functional phosphorylations? *Trends Biochem Sci* 33, 351–352.

Lim, W.A. (2010). Designing customized cell signalling circuits. *Nat Rev Mol Cell Biol* 11, 393–403.

Lim, W.A., and Pawson, T. (2010). Phosphotyrosine signaling: evolving a new cellular communication system. *Cell* 142, 661–667.

- Lin, H., Su, X., and He, B. (2012). Protein lysine acylation and cysteine succination by intermediates of energy metabolism. *ACS Chem Biol* 120509224428009.
- Lin, J., Xie, Z., Zhu, H., and Qian, J. (2010). Understanding protein phosphorylation on a systems level. *Brief Funct Genomics* 9, 32–42.
- Linard, B., Thompson, J.D., Poch, O., and Lecompte, O. (2011). OrthoInspector: comprehensive orthology analysis and visual exploration. *BMC Bioinformatics* 12, 11.
- Linder, M.E., and Deschenes, R.J. (2007). Palmitoylation: policing protein stability and traffic. *Nat Rev Mol Cell Biol* 8, 74–84.
- Lissanu Deribe, Y., Pawson, T., and Dikic, I. (2010). Post-translational modifications in signal integration. *Nat Struct Mol Biol* 17, 666–672.
- Liti, G., Carter, D.M., Moses, A.M., Warringer, J., Parts, L., James, S.A., Davey, R.P., Roberts, I.N., Burt, A., Koufopanou, V., et al. (2009). Population genomics of domestic and wild yeasts. *Nature* 458, 337–341.
- Liu, B.A., and Nash, P.D. (2012). Evolution of SH2 domains and phosphotyrosine signalling networks. *Philos Trans R Soc Lond, B, Biol Sci* 367, 2556–2573.
- Liu, K., Linder, C.R., and Warnow, T. (2010a). Multiple sequence alignment: a major challenge to large-scale phylogenetics. *PLoS Curr* 2, RRN1198.
- Liu, W.R., Wang, Y.-S., and Wan, W. (2010b). Synthesis of proteins with defined posttranslational modifications using the genetic noncanonical amino acid incorporation approach. *Mol Biosyst*.
- Liu, Y., Lu, C., Yang, Y., Fan, Y., Yang, R., Liu, C.-F., Korolev, N., and Nordenskiöld, L. (2011). Influence of histone tails and H4 tail acetylations on nucleosome-nucleosome interactions. *J Mol Biol* 414, 749–764.
- Liu, Y., Schmidt, B., and Maskell, D.L. (2010c). MSAProbs: multiple sequence alignment based on pair hidden Markov models and partition function posterior probabilities. *Bioinformatics* 26, 1958–1964.
- Lobanov, M.Y., and Galzitskaya, O.V. (2012). Occurrence of disordered patterns and homorepeats in eukaryotic and bacterial proteomes. *Mol Biosyst* 8, 327–337.
- Lopez, P., Casane, D., and Philippe, H. (2002). Heterotachy, an important process of protein evolution. *Mol Biol Evol* 19, 1–7.
- López-Bigas, N., De, S., and Teichmann, S.A. (2008). Functional protein divergence in the evolution of Homo sapiens. *Genome Biol* 9, R33.

- Lu, J.-Y., Lin, Y.-Y., Qian, J., Tao, S.-C., Zhu, J., Pickart, C., and Zhu, H. (2008). Functional dissection of a HECT ubiquitin E3 ligase. *Mol Cell Proteomics* 7, 35–45.
- Lu, Y., Prudent, M., Fauvet, B., Lashuel, H.A., and Girault, H.H. (2011). Phosphorylation of α -Synuclein at Y125 and S129 Alters Its Metal Binding Properties: Implications for Understanding the Role of α -Synuclein in the Pathogenesis of Parkinson's Disease and Related Disorders. *ACS Chem. Neurosci.* 2, 667–675.
- Lundberg, E., Fagerberg, L., Klevebring, D., Matic, I., Geiger, T., Cox, J., Algenäs, C., Lundberg, J., Mann, M., and Uhlen, M. (2010). Defining the transcriptome and proteome in three functionally different human cell lines. *Mol Syst Biol* 6, 450.
- Macarthur, B.D., Ma'ayan, A., and Lemischka, I.R. (2009). Systems biology of stem cell fate and cellular reprogramming. *Nat Rev Mol Cell Biol* 10, 672–681.
- Marsh, J.A., and Forman-Kay, J.D. (2011). Ensemble modeling of protein disordered states: Experimental restraint contributions and validation. *Proteins*.
- Martin, G.R. (1981). Isolation of a pluripotent cell line from early mouse embryos cultured in medium conditioned by teratocarcinoma stem cells. *Proc Natl Acad Sci USA* 78, 7634–7638.
- Mason, M.J., Fan, G., Plath, K., Zhou, Q., and Horvath, S. (2009). Signed weighted gene co-expression network analysis of transcriptional regulation in murine embryonic stem cells. *BMC Genomics* 10, 327.
- Maurer-Stroh, S., Koranda, M., Benetka, W., Schneider, G., Sirota, F.L., and Eisenhaber, F. (2007). Towards complete sets of farnesylated and geranylgeranylated proteins. *PLoS Comput Biol* 3, e66.
- Mayrose, I., Graur, D., Ben-Tal, N., and Pupko, T. (2004). Comparison of site-specific rate-inference methods for protein sequences: empirical Bayesian methods are superior. *Mol Biol Evol* 21, 1781–1791.
- Mayshar, Y., Ben-David, U., Lavon, N., Biancotti, J.-C., Yakir, B., Clark, A.T., Plath, K., Lowry, W.E., and Benvenisty, N. (2010). Identification and classification of chromosomal aberrations in human induced pluripotent stem cells. *Cell Stem Cell* 7, 521–531.
- McManus, C.J., and Graveley, B.R. (2011). RNA structure and the mechanisms of alternative splicing. *Curr Opin Genet Dev* 21, 373–379.

- Melero, R., Rajagopalan, S., Lázaro, M., Joerger, A.C., Brandt, T., Veprintsev, D.B., Lasso, G., Gil, D., Scheres, S.H.W., Carazo, J.M., et al. (2011). Electron microscopy studies on the quaternary structure of p53 reveal different binding modes for p53 tetramers in complex with DNA. *Proc Natl Acad Sci USA* *108*, 557–562.
- Mellor, J. (2006). It takes a PHD to read the histone code. *Cell* *126*, 22–24.
- Meshorer, E., and Misteli, T. (2006). Chromatin in pluripotent embryonic stem cells and differentiation. *Nat Rev Mol Cell Biol* *7*, 540–546.
- Mihalek, I., Res, I., and Lichtarge, O. (2004). A family of evolution-entropy hybrid methods for ranking protein residues by importance. *J Mol Biol* *336*, 1265–1282.
- Mihalek, I., Res, I., and Lichtarge, O. (2007). Background frequencies for residue variability estimates: BLOSUM revisited. *BMC Bioinformatics* *8*, 488.
- Miller, S., Janin, J., Lesk, A.M., and Chothia, C. (1987). Interior and surface of monomeric proteins. *J Mol Biol* *196*, 641–656.
- Minguez, P., Parca, L., Diella, F., Mende, D.R., Kumar, R., Helmer-Citterich, M., Gavin, A.-C., van Noort, V., and Bork, P. (2012). Deciphering a global network of functionally associated post-translational modifications. *Mol Syst Biol* *8*, 599.
- Mischerikow, N., and Heck, A.J.R. (2011). Targeted large-scale analysis of protein acetylation. *Proteomics* *11*, 571–589.
- Mittag, T., Kay, L.E., and Forman-Kay, J.D. (2010). Protein dynamics and conformational disorder in molecular recognition. *J Mol Recognit* *23*, 105–116.
- Mittal, R., Peak-Chew, S.-Y., and McMahon, H.T. (2006). Acetylation of MEK2 and I kappa B kinase (IKK) activation loop residues by YopJ inhibits signaling. *Proc Natl Acad Sci USA* *103*, 18574–18579.
- Miyazawa, S. (2011a). Advantages of a mechanistic codon substitution model for evolutionary analysis of protein-coding sequences. *PLoS ONE* *6*, e28892.
- Miyazawa, S. (2011b). Selective constraints on amino acids estimated by a mechanistic codon substitution model with multiple nucleotide changes. *PLoS ONE* *6*, e17244.
- Monastyrskyy, B., Fidelis, K., Moulton, J., Tramontano, A., and Kryzhanovych, A. (2011). Evaluation of disorder predictions in CASP9. *Proteins* *79*, 107–118.
- Moore, A.D., Björklund, A.K., Ekman, D., Bornberg-Bauer, E., and Elofsson, A. (2008). Arrangements in the modular evolution of proteins. *Trends Biochem Sci* *33*, 444–451.

- Moran, N.A., and Jarvik, T. (2010). Lateral transfer of genes from fungi underlies carotenoid production in aphids. *Science* 328, 624–627.
- Morinière, J., Rousseaux, S., Steuerwald, U., Soler-López, M., Curtet, S., Vitte, A.-L., Govin, J., Gaucher, J., Sadoul, K., Hart, D.J., et al. (2009). Cooperative binding of two acetylation marks on a histone tail by a single bromodomain. *Nature* 461, 664–668.
- Mosca, R., Pache, R.A., and Aloy, P. (2012). The role of structural disorder in the rewiring of protein interactions through evolution. *Mol Cell Proteomics*.
- Moses, A.M., and Landry, C.R. (2010). Moving from transcriptional to phospho-evolution: generalizing regulatory evolution? *Trends Genet* 26, 462–467.
- Motter, A.E. (2010). Improved network performance via antagonism: From synthetic rescues to multi-drug combinations. *BioEssays* 32, 236–245.
- Muers, M. (2011). Chromatin: a haul of new histone modifications. *Nat Rev Genet* 12, 744.
- Mujtaba, S., Zeng, L., and Zhou, M.-M. (2007). Structure and acetyl-lysine recognition of the bromodomain. *Oncogene* 26, 5521–5527.
- Mukherjee, S., Hao, Y.-H., and Orth, K. (2007). A newly discovered post-translational modification--the acetylation of serine and threonine residues. *Trends Biochem Sci* 32, 210–216.
- Mukherjee, S., Keitany, G., Li, Y., Wang, Y., Ball, H.L., Goldsmith, E.J., and Orth, K. (2006). Yersinia YopJ acetylates and inhibits kinase activation by blocking phosphorylation. *Science* 312, 1211–1214.
- Nagaraj, N., Wisniewski, J.R., Geiger, T., Cox, J., Kircher, M., Kelso, J., Pääbo, S., and Mann, M. (2011). Deep proteome and transcriptome mapping of a human cancer cell line. *Mol Syst Biol* 7, 548.
- Nagaraj, R.H., Nahomi, R.B., Shanthakumar, S., Linetsky, M., Padmanabha, S., Pasupuleti, N., Wang, B., Santhoshkumar, P., Panda, A.K., and Biswas, A. (2012). Acetylation of α A-crystallin in the human lens: effects on structure and chaperone function. *Biochim Biophys Acta* 1822, 120–129.
- Nakagawa, M., Koyanagi, M., Tanabe, K., Takahashi, K., Ichisaka, T., Aoi, T., Okita, K., Mochiduki, Y., Takizawa, N., and Yamanaka, S. (2008). Generation of induced pluripotent stem cells without Myc from mouse and human fibroblasts. *Nat Biotechnol* 26, 101–106.

- Nakamura, Y., Umehara, T., Nakano, K., Jang, M.K., Shirouzu, M., Morita, S., Uda-Tochio, H., Hamana, H., Terada, T., Adachi, N., et al. (2007). Crystal structure of the human BRD2 bromodomain: insights into dimerization and recognition of acetylated histone H4. *J Biol Chem* 282, 4193–4201.
- Nayak, R., Kearns, M., Spielman, R., and Cheung, V. (2009). Coexpression network based on natural variation in human gene expression reveals gene interactions and functions. *Genome Res.*
- Needleman, S.B., and Wunsch, C.D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol* 48, 443–453.
- Neumann, H., Slusarczyk, A.L., and Chin, J.W. (2010). De Novo Generation of Mutually Orthogonal Aminoacyl-tRNA Synthetase/tRNA Pairs. *J Am Chem Soc* 132, 2142–2144.
- Nishi, H., Hashimoto, K., and Panchenko, A.R. (2011). Phosphorylation in protein-protein binding: effect on stability and function. *Structure* 19, 1807–1815.
- Nishikawa, T., and Motter, A.E. (2010). Network synchronization landscape reveals compensatory structures, quantization, and the positive effect of negative interactions. *Proc Natl Acad Sci USA* 107, 10342–10347.
- Norris, K., Lee, J., and Yao, T. (2009). Acetylation Goes Global: The Emergence of Acetylation Biology. *Sci Signal* 2, pe76.
- Nottke, A., Colaiácovo, M.P., and Shi, Y. (2009). Developmental roles of the histone lysine demethylases. *Development* 136, 879–889.
- Nuin, P.A.S., Wang, Z., and Tillier, E.R.M. (2006). The accuracy of several multiple sequence alignment programs for proteins. *BMC Bioinformatics* 7, 471.
- Nunes, M.C., Goldring, J.P.D., Doerig, C., and Scherf, A. (2007). A novel protein kinase family in *Plasmodium falciparum* is differentially transcribed and secreted to various cellular compartments of the host cell. *Mol Microbiol* 63, 391–403.
- Nussinov, R., Tsai, C.-J., Xin, F., and Radivojac, P. (2012). Allosteric post-translational modification codes. *Trends Biochem Sci.*
- Nühse, T.S., Stensballe, A., Jensen, O.N., and Peck, S.C. (2004). Phosphoproteomics of the *Arabidopsis* plasma membrane and a new phosphorylation site database. *Plant Cell* 16, 2394–2405.
- Oesterlin, L.K., Goody, R.S., and Itzen, A. (2012). Posttranslational modifications of Rab proteins cause effective displacement of GDP dissociation inhibitor. *Proc Natl Acad Sci USA* 109, 5621–5626.

Ogryzko, V.V., Schiltz, R.L., Russanova, V., Howard, B.H., and Nakatani, Y. (1996). The transcriptional coactivators p300 and CBP are histone acetyltransferases. *Cell* 87, 953–959.

Okita, K., and Yamanaka, S. (2010). Induction of pluripotency by defined factors. *Experimental Cell Research*.

Olsen, C.A. (2012). Expansion of the Lysine Acylation Landscape. *Angew Chem Int Ed Engl*.

Olsen, J.V., Vermeulen, M., Santamaria, A., Kumar, C., Miller, M.L., Jensen, L.J., Gnad, F., Cox, J., Jensen, T.S., Nigg, E.A., et al. (2010). Quantitative phosphoproteomics reveals widespread full phosphorylation site occupancy during mitosis. *Sci Signal* 3, ra3.

Ong, S.-E., Blagoev, B., Kratchmarova, I., Kristensen, D.B., Steen, H., Pandey, A., and Mann, M. (2002). Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol Cell Proteomics* 1, 376–386.

Ostergaard, J.R., Sunde, L., and Okkels, H. (2005). Neurofibromatosis von Recklinghausen type I phenotype and early onset of cancers in siblings compound heterozygous for mutations in MSH6. *Am J Med Genet A* 139A, 96–105; discussion96.

Ostlund, G., Schmitt, T., Forslund, K., Köstler, T., Messina, D.N., Roopra, S., Frings, O., and Sonnhammer, E.L.L. (2010). InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. *Nucleic Acids Res* 38, D196–D203.

Outeiro, T.F., Kontopoulos, E., Altmann, S.M., Kufareva, I., Strathearn, K.E., Amore, A.M., Volk, C.B., Maxwell, M.M., Rochet, J.-C., McLean, P.J., et al. (2007). Sirtuin 2 inhibitors rescue alpha-synuclein-mediated toxicity in models of Parkinson's disease. *Science* 317, 516–519.

Paik, W.K., Paik, D.C., and Kim, S. (2007). Historical review: the field of protein methylation. *Trends Biochem Sci* 32, 146–152.

Pal, S., Gupta, R., Kim, H., Wickramasinghe, P., Baubet, V., Showe, L.C., Dahmane, N., and Davuluri, R.V. (2011). Alternative transcription exceeds alternative splicing in generating the transcriptome diversity of cerebellar development. *Genome Res* 21, 1260–1272.

Paleologou, K.E., Schmid, A.W., Rospigliosi, C.C., Kim, H.Y., Lamberto, G.R., Fredenburg, R.A., Lansbury, P.T., Fernandez, C.O., Eliezer, D., Zweckstetter, M., et al. (2008). Phosphorylation at Ser-129 but Not the Phosphomimics S129E/D Inhibits the Fibrillation of -Synuclein. *J Biol Chem* 283, 16895–16905.

Palmisano, G., Melo-Braga, M.N., Engholm-Keller, K., Parker, B.L., and Larsen, M.R. (2012). Chemical deamidation: a common pitfall in large-scale N-linked glycoproteomic mass spectrometry-based analyses. *J Proteome Res*.

- Pan, C., Olsen, J.V., Daub, H., and Mann, M. (2009). Global effects of kinase inhibitors on signaling networks revealed by quantitative phosphoproteomics. *Mol Cell Proteomics* 8, 2796–2808.
- Pang, C.N.I., Gasteiger, E., and Wilkins, M.R. (2010). Identification of arginine- and lysine-methylation in the proteome of *Saccharomyces cerevisiae* and its functional implications. *BMC Genomics* 11, 92.
- Pang, C.N.I., Hayen, A., and Wilkins, M.R. (2007). Surface accessibility of protein post-translational modifications. *J Proteome Res* 6, 1833–1845.
- Pardo, M., (null), A., Yu, L., Prosser, H., Bradley, A., Babu, M.M., and Choudhary, J. (2010). An expanded Oct4 interaction network: implications for stem cell biology, development, and disease. *Cell Stem Cell* 6, 382–395.
- Park, C., and Zhang, J. (2011). Genome-wide evolutionary conservation of N-glycosylation sites. *Mol Biol Evol* 28, 2351–2357.
- Park, I.-H., Arora, N., Huo, H., Maherali, N., Ahfeldt, T., Shimamura, A., Lensch, M.W., Cowan, C., Hochedlinger, K., and Daley, G.Q. (2008). Disease-Specific Induced Pluripotent Stem Cells. *Cell* 134, 877–886.
- Pearlman, S.M., Serber, Z., and Ferrell, J.E. (2011). A mechanism for the evolution of phosphorylation sites. *Cell* 147, 934–946.
- Peng, C., Lu, Z., Xie, Z., Cheng, Z., Chen, Y., Tan, M., Luo, H., Zhang, Y., He, W., Yang, K., et al. (2011). The first identification of lysine malonylation substrates and its regulatory enzyme. *Mol Cell Proteomics* 10, M111.012658.
- Perrett, C.A., Lin, D.Y.-W., and Zhou, D. (2011). Interactions of bacterial proteins with host eukaryotic ubiquitin pathways. *Front Microbiol* 2, 143.
- Philippe, H., Casane, D., Gribaldo, S., Lopez, P., and Meunier, J. (2003). Heterotachy and functional shift in protein evolution. *IUBMB Life* 55, 257–265.
- Pierce, N.W., Kleiger, G., Shan, S.-O., and Deshaies, R.J. (2009). Detection of sequential polyubiquitylation on a millisecond timescale. *Nature* 462, 615–619.
- Porter, C.T., Bartlett, G.J., and Thornton, J.M. (2004). The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Res* 32, D129–D133.
- Price, J.C., Guan, S., Burlingame, A., Prusiner, S.B., and Ghaemmaghami, S. (2010). Analysis of proteome dynamics in the mouse brain. *Proc Natl Acad Sci USA* 107, 14508–14513.

Pruneda, J.N., Littlefield, P.J., Soss, S.E., Nordquist, K.A., Chazin, W.J., Brzovic, P.S., and Kleivit, R.E. (2012). Structure of an E3:E2~Ub Complex Reveals an Allosteric Mechanism Shared among RING/U-box Ligases. *Mol Cell*.

Punta, M., Coggill, P.C., Eberhardt, R.Y., Mistry, J., Tate, J., Boursnell, C., Pang, N., Forslund, K., Ceric, G., Clements, J., et al. (2012). The Pfam protein families database. *Nucleic Acids Res* *40*, D290–D301.

Rajagopalan, S., Jaulent, A.M., Wells, M., Veprintsev, D.B., and Fersht, A.R. (2008). 14-3-3 activation of DNA binding of p53 by enhancing its association into tetramers. *Nucleic Acids Res* *36*, 5983–5991.

Rajaram, S., and Oono, Y. (2010). NeatMap--non-clustering heat map alternatives in R. *BMC Bioinformatics* *11*, 45.

Rasmussen, T.P. (2003). Embryonic stem cell differentiation: a chromatin perspective. *Reprod. Biol. Endocrinol.* *1*, 100.

Raunser, S., Magnani, R., Huang, Z., Houtz, R.L., Trievel, R.C., Penczek, P.A., and Walz, T. (2009). Rubisco in complex with Rubisco large subunit methyltransferase. *Proc Natl Acad Sci USA* *106*, 3160–3165.

Ravasi, T., Suzuki, H., Cannistraci, C.V., Katayama, S., Bajic, V.B., Tan, K., Akalin, A., Schmeier, S., Kanamori-Katayama, M., Bertin, N., et al. (2010). An atlas of combinatorial transcriptional regulation in mouse and man. *Cell* *140*, 744–752.

Reiher, W., Shirras, C., Kahnt, J., Baumeister, S., Isaac, R.E., and Wegener, C. (2011). Peptidomics and peptide hormone processing in the *Drosophila* midgut. *J Proteome Res* *10*, 1881–1892.

Reimand, J., Kull, M., Peterson, H., Hansen, J., and Vilo, J. (2007). g:Profiler--a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic Acids Res* *35*, W193–W200.

Reményi, A., Lins, K., Nissen, L.J., Reinbold, R., Schöler, H.R., and Wilmanns, M. (2003). Crystal structure of a POU/HMG/DNA ternary complex suggests differential assembly of Oct4 and Sox2 on two enhancers. *Genes Dev* *17*, 2048–2059.

Rhodes, D.R., Kalyana-Sundaram, S., Mahavisno, V., Varambally, R., Yu, J., Briggs, B.B., Barrette, T.R., Anstet, M.J., Kincead-Beal, C., Kulkarni, P., et al. (2007). OncoPrint 3.0: genes, pathways, and networks in a collection of 18,000 cancer gene expression profiles. *Neoplasia* *9*, 166–180.

- Ribet, D., and Cossart, P. (2010). Pathogen-Mediated Posttranslational Modifications: A Re-emerging Field. *Cell* 143, 694–702.
- Robinson, N.E., and Robinson, A.B. (2001). Prediction of protein deamidation rates from primary and three-dimensional structure. *Proc Natl Acad Sci USA* 98, 4367–4372.
- Roguev, A., Talbot, D., Negri, G.L., Shales, M., Cagney, G., Bandyopadhyay, S., Panning, B., and Krogan, N.J. (2013). Quantitative genetic-interaction mapping in mammalian cells. *Nat Methods*.
- Romero, P.R., Zaidi, S., Fang, Y.Y., Uversky, V.N., Radivojac, P., Oldfield, C.J., Cortese, M.S., Sickmeier, M., LeGall, T., Obradovic, Z., et al. (2006). Alternative splicing in concert with protein intrinsic disorder enables increased functional diversity in multicellular organisms. *Proc Natl Acad Sci USA* 103, 8390–8395.
- Roure, B., and Philippe, H. (2011). Site-specific time heterogeneity of the substitution process and its impact on phylogenetic inference. *BMC Evol Biol* 11, 17.
- Römer, L., Klein, C., Dehner, A., Kessler, H., and Buchner, J. (2006). p53--a natural cancer killer: structural insights and therapeutic concepts. *Angew Chem Int Ed Engl* 45, 6440–6460.
- Ruan, J., Li, H., Chen, Z., Coghlan, A., Coin, L.J.M., Guo, Y., Hériché, J.-K., (null), Kristiansen, K., Li, R., et al. (2008). TreeFam: 2008 Update. *Nucleic Acids Res* 36, D735–D740.
- Rustom, A., Saffrich, R., Markovic, I., Walther, P., and Gerdes, H.-H. (2004). Nanotubular highways for intercellular organelle transport. *Science* 303, 1007–1010.
- Ruthenburg, A.J., Li, H., Patel, D.J., and Allis, C.D. (2007). Multivalent engagement of chromatin modifications by linked binding modules. *Nat Rev Mol Cell Biol* 8, 983–994.
- Sadoul, K., Boyault, C., Pabion, M., and Khochbin, S. (2008). Regulation of protein turnover by acetyltransferases and deacetylases. *Biochimie* 90, 306–312.
- Sampath, S.C., Marazzi, I., Yap, K.L., Sampath, S.C., Krutchinsky, A.N., Mecklenbräuker, I., Viale, A., Rudensky, E., Zhou, M.-M., Chait, B.T., et al. (2007). Methylation of a histone mimic within the histone methyltransferase G9a regulates protein complex assembly. *Mol Cell* 27, 596–608.
- Sanchez, R., and Zhou, M.-M. (2009). The role of human bromodomains in chromatin biology and gene transcription. *Curr Opin Drug Discov Devel* 12, 659–665.

Sanchez, S.E., Petrillo, E., Beckwith, E.J., Zhang, X., Rugnone, M.L., Hernando, C.E., Cuevas, J.C., Godoy Herz, M.A., Depetris-Chauvin, A., Simpson, C.G., et al. (2010). A methyl transferase links the circadian clock to the regulation of alternative splicing. *Nature* 468, 112–116.

Sandie, R., Palidwor, G.A., Huska, M.R., Porter, C.J., Krzyzanowski, P.M., Muro, E.M., Perez-Iratxeta, C., and Andrade-Navarro, M.A. (2009). Recent developments in StemBase: a tool to study gene expression in human and murine stem cells. *BMC Res Notes* 2, 39.

Sasidharan, R., and Chothia, C. (2007). The selection of acceptable protein mutations. *Proc Natl Acad Sci USA* 104, 10080–10085.

Sayers, E.W., Barrett, T., Benson, D.A., Bolton, E., Bryant, S.H., Canese, K., Chetvernin, V., Church, D.M., Dicuccio, M., Federhen, S., et al. (2012). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 40, D13–D25.

Sánchez-Puig, N., Veprintsev, D.B., and Fersht, A.R. (2005). Binding of natively unfolded HIF-1 α ODD domain to p53. *Mol Cell* 17, 11–21.

Schnur, E., Noah, E., Ayzenshtat, I., Sargsyan, H., Inui, T., Ding, F.-X., Arshava, B., Sagi, Y., Kessler, N., Levy, R., et al. (2011). The conformation and orientation of a 27-residue CCR5 peptide in a ternary complex with HIV-1 gp120 and a CD4-mimic peptide. *J Mol Biol* 410, 778–797.

Schofield, C.J., and Ratcliffe, P.J. (2004). Oxygen sensing by HIF hydroxylases. *Nat Rev Mol Cell Biol* 5, 343–354.

Schrider, D.R., Hourmozdi, J.N., and Hahn, M.W. (2011). Pervasive Multinucleotide Mutational Events in Eukaryotes. *Current Biology* 21, 1051–1054.

Schwer, B., Eckersdorff, M., Li, Y., Silva, J.C., Fermin, D., Kurtev, M.V., Giallourakis, C., Comb, M.J., Alt, F.W., and Lombard, D.B. (2009). Calorie restriction alters mitochondrial protein acetylation. *Aging Cell* 8, 604–606.

Scott, J.D., and Pawson, T. (2009). Cell signaling in space and time: where proteins come together and when they're apart. *Science* 326, 1220–1224.

Scroggins, B.T., Robzyk, K., Wang, D., Marcu, M.G., Tsutsumi, S., Beebe, K., Cotter, R.J., Felts, S., Toft, D., Karnitz, L., et al. (2007). An acetylation site in the middle domain of Hsp90 regulates chaperone function. *Mol Cell* 25, 151–159.

Seet, B.T., Dikic, I., Zhou, M.-M., and Pawson, T. (2006). Reading protein modifications with interaction domains. *Nat Rev Mol Cell Biol* 7, 473–483.

- Sekhon, J.S. (2011). Multivariate and Propensity Score Matching Software with Automated Balance Optimization: The Matching Package for R. *J Stat Softw* 42, 1–52.
- Sellick, C.A., Hansen, R., Stephens, G.M., Goodacre, R., and Dickson, A.J. (2011). Metabolite extraction from suspension-cultured mammalian cells for global metabolite profiling. *Nat Protoc* 6, 1241–1249.
- Sen, N., and Snyder, S.H. (2010). Protein modifications involved in neurotransmitter and gasotransmitter signaling. *Trends in Neurosciences*.
- Serber, Z., and Ferrell, J.E. (2007). Tuning bulk electrostatics to regulate protein function. *Cell* 128, 441–444.
- Shahbazian, M.D., and Grunstein, M. (2007). Functions of site-specific histone acetylation and deacetylation. *Annu Rev Biochem* 76, 75–100.
- Shannon, C.E. (1951). Prediction and Entropy of Printed English, Pp. 50-64 in the Bell System Technical Journal, Vol. XXX, No. 1 (A T & T).
- Shaw, B.F., Moustakas, D.T., Whitelegge, J.P., and Faull, K.F. (2010). *Advances in Protein Chemistry and Structural Biology* (Elsevier).
- Shaw, B.F., Schneider, G.F., Bilgiçer, B., Kaufman, G.K., Neveu, J.M., Lane, W.S., Whitelegge, J.P., and Whitesides, G. (2008). Lysine acetylation can generate highly charged enzymes with increased resistance toward irreversible inactivation. *Protein Sci* 17, 1446–1455.
- Shaw, N., Zhao, M., Cheng, C., Xu, H., Saarikettu, J., Li, Y., Da, Y., Yao, Z., Silvennoinen, O., Yang, J., et al. (2007). The multifunctional human p100 protein “hooks” methylated ligands. *Nat Struct Mol Biol* 14, 779–784.
- Shental-Bechor, D., and Levy, Y. (2009). Folding of glycoproteins: toward understanding the biophysics of the glycosylation code. *Curr Opin Struct Biol* 19, 524–533.
- Shental-Bechor, D., Smith, M.T.J., MacKenzie, D., Broom, A., Marcovitz, A., Ghashut, F., Go, C., Bralha, F., Meiering, E.M., and Levy, Y. (2012). Nonnative interactions regulate folding and switching of myristoylated protein. *Proc Natl Acad Sci USA*.
- Shi, Y. (2009). Serine/threonine phosphatases: mechanism through structure. *Cell* 139, 468–484.
- Shukla, A., Chaurasia, P., and Bhaumik, S. (2009). Histone methylation and ubiquitination with their cross-talk and roles in gene expression and stability. *Cell Mol Life Sci* 66, 1419–1433.

Sickmeier, M., Hamilton, J.A., LeGall, T., Vacic, V., Cortese, M.S., Tantos, A., Szabo, B., Tompa, P., Chen, J., Uversky, V.N., et al. (2007). DisProt: the Database of Disordered Proteins. *Nucleic Acids Res* 35, D786–D793.

Sidhar, S.K., Clark, J., Gill, S., Hamoudi, R., Crew, A.J., Gwilliam, R., Ross, M., Linehan, W.M., Birdsall, S., Shipley, J., et al. (1996). The t(X;1)(p11.2;q21.2) translocation in papillary renal cell carcinoma fuses a novel gene PRCC to the TFE3 transcription factor gene. *Hum Mol Genet* 5, 1333–1338.

Sims, J.J., Scavone, F., Cooper, E.M., Kane, L.A., Youle, R.J., Boeke, J.D., and Cohen, R.E. (2012). Polyubiquitin-sensor proteins reveal localization and linkage-type dependence of cellular ubiquitin signaling. *Nat Methods* 9, 303–309.

Siuti, N., and Kelleher, N.L. (2007). Decoding protein modifications using top-down mass spectrometry. *Nat Methods* 4, 817–821.

Smith, N.G.C., Webster, M.T., and Ellegren, H. (2003). A low rate of simultaneous double-nucleotide mutations in primates. *Mol Biol Evol* 20, 47–53.

Smoot, M.E., Ono, K., Ruscheinski, J., Wang, P.-L., and Ideker, T. (2011). Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics* 27, 431–432.

Sonnhammer, E.L.L., and Koonin, E.V. (2002). Orthology, paralogy and proposed classification for paralog subtypes. *Trends Genet* 18, 619–620.

Soppa, J. (2010). Protein acetylation in archaea, bacteria, and eukaryotes. *Archaea* 2010.

Soufi, B., Gnad, F., Jensen, P.R., Petranovic, D., Mann, M., Mijakovic, I., and Macek, B. (2008). The Ser/Thr/Tyr phosphoproteome of *Lactococcus lactis* IL1403 reveals multiply phosphorylated proteins. *Proteomics* 8, 3486–3493.

Sowa, M.E., Bennett, E.J., Gygi, S.P., and Harper, J.W. (2009). Defining the human deubiquitinating enzyme interaction landscape. *Cell* 138, 389–403.

Spange, S., Wagner, T., Heinzl, T., and Krämer, O.H. (2009). Acetylation of non-histone proteins modulates cellular signalling at multiple levels. *Int. J. Biochem. Cell Biol.* 41, 185–198.

Spence, J.R., Mayhew, C.N., Rankin, S.A., Kuhar, M.F., Vallance, J.E., Tolle, K., Hoskins, E.E., Kalinichenko, V.V., Wells, S.I., Zorn, A.M., et al. (2011). Directed differentiation of human pluripotent stem cells into intestinal tissue in vitro. *Nature* 470, 105–109.

Sprung, R., Chen, Y., Zhang, K., Cheng, D., Zhang, T., Peng, J., and Zhao, Y. (2008). Identification and validation of eukaryotic aspartate and glutamate methylation in proteins. *J Proteome Res* 7, 1001–1006.

- Stein, A., and Aloy, P. (2008). Contextual specificity in peptide-mediated protein interactions. *PLoS ONE* 3, e2524.
- Stenson, P.D., Mort, M., Ball, E.V., Howells, K., Phillips, A.D., Thomas, N.S., and Cooper, D.N. (2009). The Human Gene Mutation Database: 2008 update. *Genome Medicine* 1, 13.
- Stevens, S.M., Prokai-Tatrai, K., and Prokai, L. (2008). Factors that contribute to the misidentification of tyrosine nitration by shotgun proteomics. *Mol Cell Proteomics* 7, 2442–2451.
- Sturn, A., Quackenbush, J., and Trajanoski, Z. (2002). Genesis: cluster analysis of microarray data. *Bioinformatics* 18, 207–208.
- Su, A.I., Wiltshire, T., Batalov, S., Lapp, H., Ching, K.A., Block, D., Zhang, J., Soden, R., Hayakawa, M., Kreiman, G., et al. (2004). A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci USA* 101, 6062–6067.
- Subramanian, S., and Lambert, D.M. (2011). Time dependency of molecular evolutionary rates? Yes and no. *Genome Biol Evol* 3, 1324–1328.
- Suganuma, T., and Workman, J.L. (2008). Crosstalk among Histone Modifications. *Cell* 135, 604–607.
- Suganuma, T., and Workman, J.L. (2011). Signals and combinatorial functions of histone modifications. *Annu Rev Biochem* 80, 473–499.
- Sun, Y., Xu, Y., Roy, K., and Price, B.D. (2007). DNA damage-induced acetylation of lysine 3016 of ATM activates ATM kinase activity. *Mol Cell Biol* 27, 8502–8509.
- Suyama, M., Torrents, D., and Bork, P. (2006). PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res* 34, W609–W612.
- Szklarczyk, D., Franceschini, A., Kuhn, M., Simonovic, M., Roth, A., Minguéz, P., Doerks, T., Stark, M., Müller, J., Bork, P., et al. (2011). The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res* 39, D561–D568.
- Tahiliani, M., Koh, K.P., Shen, Y., Pastor, W.A., Bandukwala, H., Brudno, Y., Agarwal, S., Iyer, L.M., Liu, D.R., Aravind, L., et al. (2009). Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science* 324, 930–935.
- Takahashi, K., and Yamanaka, S. (2006). Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* 126, 663–676.

Takahashi, K., Tanabe, K., Ohnuki, M., Narita, M., Ichisaka, T., Tomoda, K., and Yamanaka, S. (2007). Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell* 131, 861–872.

Tan, C.S.H., Bodenmiller, B., Pasculescu, A., Jovanovic, M., Hengartner, M.O., Jørgensen, C., Bader, G.D., Aebersold, R., Pawson, T., and Linding, R. (2009a). Comparative analysis reveals conserved protein phosphorylation networks implicated in multiple diseases. *Sci Signal* 2, ra39.

Tan, C.S.H., Pasculescu, A., Lim, W.A., Pawson, T., Bader, G.D., and Linding, R. (2009b). Positive selection of tyrosine loss in metazoan evolution. *Science* 325, 1686–1688.

Tan, M., Luo, H., Lee, S., Jin, F., Yang, J.S., Montellier, E., Buchou, T., Cheng, Z., Rousseaux, S., Rajagopal, N., et al. (2011). Identification of 67 histone marks and histone lysine crotonylation as a new type of histone modification. *Cell* 146, 1016–1028.

Tarrant, M.K., and Cole, P.A. (2009). The chemical biology of protein phosphorylation. *Annu Rev Biochem* 78, 797–825.

Tasset, C., Bernoux, M., Jauneau, A., Pouzet, C., Brière, C., Kieffer-Jacquiod, S., Rivas, S., Marco, Y., and Deslandes, L. (2010). Autoacetylation of the *Ralstonia solanacearum* Effector PopP2 Targets a Lysine Residue Essential for RRS1-R-Mediated Immunity in Arabidopsis. *PLoS Pathog* 6, e1001202.

Taverna, S.D., Li, H., Ruthenburg, A.J., Allis, C.D., and Patel, D.J. (2007). How chromatin-binding modules interpret histone modifications: lessons from professional pocket pickers. *Nat Struct Mol Biol* 14, 1025–1040.

Thibodeaux, G.N., and van der Donk, W.A. (2012). An engineered lantipeptide synthetase serves as a general leader peptide-dependent kinase. *Chem. Commun.*

Thomas, P.D., Kejariwal, A., Campbell, M.J., Mi, H., Diemer, K., Guo, N., Ladunga, I., Ulitsky-Lazareva, B., Muruganujan, A., Rabkin, S., et al. (2003). PANTHER: a browsable database of gene products organized by biological function, using curated protein family and subfamily classification. *Nucleic Acids Res* 31, 334–341.

Thomson, M., and Gunawardena, J. (2009). The rational parameterization theorem for multisite post-translational modification systems. *J Theor Biol* 261, 626–636.

Thorsness, P.E., and Koshland, D.E. (1987). Inactivation of isocitrate dehydrogenase by phosphorylation is mediated by the negative charge of the phosphate. *J Biol Chem* 262, 10422–10425.

- Tompa, P. (2005). The interplay between structure and function in intrinsically unstructured proteins. *FEBS Lett* 579, 3346–3354.
- Tompa, P., and Fuxreiter, M. (2008). Fuzzy complexes: polymorphism and structural disorder in protein-protein interactions. *Trends Biochem Sci* 33, 2–8.
- Tompa, P., Fuxreiter, M., Oldfield, C.J., Simon, I., Dunker, A.K., and Uversky, V.N. (2009). Close encounters of the third kind: disordered domains and the interactions of proteins. *BioEssays* 31, 328–335.
- Tóth-Petróczy, A., and Tawfik, D.S. (2011). Slow protein evolutionary rates are dictated by surface-core association. *Proc Natl Acad Sci USA* 108, 11151–11156.
- Treeck, M., Sanders, J.L., Elias, J.E., and Boothroyd, J.C. (2011). The phosphoproteomes of *Plasmodium falciparum* and *Toxoplasma gondii* reveal unusual adaptations within and beyond the parasites' boundaries. *Cell Host Microbe* 10, 410–419.
- Tria, F., Caglioti, E., Loreto, V., and Pagnani, A. (2010). A stochastic local search algorithm for distance-based phylogeny reconstruction. *Mol Biol Evol* 27, 2587–2595.
- Trott, J., Hayashi, K., Surani, A., Babu, M.M., and Martinez Arias, A. (2012). Dissecting ensemble networks in ES cell populations reveals micro-heterogeneity underlying pluripotency. *Mol Biosyst* 8, 744–752.
- Tweedie-Cullen, R.Y., Brunner, A.M., Grossmann, J., Mohanna, S., Sichau, D., Nanni, P., Panse, C., and Mansuy, I.M. (2012). Identification of combinatorial patterns of post-translational modifications on individual histones in the mouse brain. *PLoS ONE* 7, e36980.
- Ubersax, J.A., and Ferrell, J.E. (2007). Mechanisms of specificity in protein phosphorylation. *Nat Rev Mol Cell Biol* 8, 530–541.
- Uddin, M.N., and Kim, J.Y. (2011). Non-cell-autonomous RNA Silencing Spread in Plants. *Botanical Studies* 52, 129–136.
- UniProt Consortium (2010). The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Res* 38, D142–D148.
- UniProt Consortium (2012). Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res* 40, D71–D75.
- Upadhyay, A.K., Horton, J.R., Zhang, X., and Cheng, X. (2011). Coordinated methyl-lysine erasure: structural and functional linkage of a Jumonji demethylase domain and a reader domain. *Curr Opin Struct Biol*.

Vacic, V., and Iakoucheva, L.M. (2012). Disease mutations in disordered regions--exception to the rule? *Mol Biosyst* 8, 27–32.

Vacic, V., Iakoucheva, L.M., and Radivojac, P. (2006). Two Sample Logo: a graphical representation of the differences between two sets of sequence alignments. *Bioinformatics* 22, 1536–1537.

van den Berg, D.L.C., Snoek, T., Mullin, N.P., Yates, A., Bezstarosti, K., Demmers, J., Chambers, I., and Poot, R.A. (2010). An Oct4-centered protein interaction network in embryonic stem cells. *Cell Stem Cell* 6, 369–381.

van Dieck, J., Fernandez-Fernandez, M.R., Veprintsev, D.B., and Fersht, A.R. (2009). Modulation of the oligomerization state of p53 by differential binding of proteins of the S100 family to p53 monomers and tetramers. *J Biol Chem* 284, 13804–13811.

van Noort, V., Seebacher, J., Bader, S., Mohammed, S., Vonkova, I., Betts, M.J., Kühner, S., Kumar, R., Maier, T., O'Flaherty, M., et al. (2012). Cross-talk between phosphorylation and lysine acetylation in a genome-reduced bacterium. *Mol Syst Biol* 8.

Vaquerizas, J.M., Kummerfeld, S.K., Teichmann, S.A., and Luscombe, N.M. (2009). A census of human transcription factors: function, expression and evolution. *Nat Rev Genet* 10, 252–263.

Varki, A. (2006). Nothing in glycobiology makes sense, except in the light of evolution. *Cell* 126, 841–845.

Varki, A. (2010). Colloquium paper: uniquely human evolution of sialic acid genetics and biology. *Proc Natl Acad Sci USA* 107 Suppl 2, 8939–8946.

Varki, A., and Altheide, T.K. (2005). Comparing the human and chimpanzee genomes: searching for needles in a haystack. *Genome Res* 15, 1746–1758.

Varki, A., and Nelson, D.L. (2007). Genomic comparisons of humans and chimpanzees. *Annu. Rev. Anthropol.* 36, 191–209.

Varki, A., Geschwind, D.H., and Eichler, E.E. (2008). Explaining human uniqueness: genome interactions with environment, behaviour and culture. *Nat Rev Genet* 9, 749–763.

Varki, N.M., Strobert, E., Dick, E.J., Benirschke, K., and Varki, A. (2011). Biomedical differences between human and nonhuman hominids: potential roles for uniquely human aspects of sialic acid biology. *Annu Rev Pathol* 6, 365–393.

Veening, J.-W., Smits, W.K., and Kuipers, O.P. (2008). Bistability, epigenetics, and bet-hedging in bacteria. *Annu. Rev. Microbiol.* 62, 193–210.

- Velankar, S., Alhroub, Y., Best, C., Caboche, S., Conroy, M.J., Dana, J.M., Fernandez Montecelo, M.A., van Ginkel, G., Golovin, A., Gore, S.P., et al. (2012). PDBe: Protein Data Bank in Europe. *Nucleic Acids Res* *40*, D445–D452.
- Vermeulen, M., Hubner, N.C., and Mann, M. (2008). High confidence determination of specific protein-protein interactions using quantitative mass spectrometry. *Curr Opin Biotechnol* *19*, 331–337.
- Vilella, A.J., Severin, J., Ureta-Vidal, A., Heng, L., Durbin, R., and Birney, E. (2009). EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res* *19*, 327–335.
- Virdee, S., Kapadnis, P.B., Elliott, T., Lang, K., Madrzak, J., Nguyen, D.P., Riechmann, L., and Chin, J.W. (2011). Traceless and site-specific ubiquitination of recombinant proteins. *J Am Chem Soc* *133*, 10708–10711.
- Vizcaíno, J.A., Côté, R., Reisinger, F., Foster, J.M., Mueller, M., Rameseder, J., Hermjakob, H., and Martens, L. (2009). A guide to the Proteomics Identifications Database proteomics data repository. *Proteomics* *9*, 4276–4283.
- Vogel, C., Abreu, R. de S., Ko, D., Le, S.-Y., Shapiro, B.A., Burns, S.C., Sandhu, D., Boutz, D.R., Marcotte, E.M., and Penalva, L.O. (2010). Sequence signatures and mRNA concentration can explain two-thirds of protein abundance variation in a human cell line. *Mol Syst Biol* *6*, 400.
- Vos, R.A., Caravas, J., Hartmann, K., Jensen, M.A., and Miller, C. (2011). BIO::Phyloinformatics analysis using perl. *BMC Bioinformatics* *12*, 63.
- Vu, T.H., Jirtle, R.L., and Hoffman, A.R. (2006). Cross-species clues of an epigenetic imprinting regulatory code for the *IGF2R* gene. *Cytogenet Genome Res* *113*, 202–208.
- Vucetic, S., Xie, H., Iakoucheva, L.M., Oldfield, C.J., Dunker, A.K., Obradovic, Z., and Uversky, V.N. (2007). Functional anthology of intrinsic disorder. 2. Cellular components, domains, technical terms, developmental processes, and coding sequence diversities correlated with long disordered regions. *J Proteome Res* *6*, 1899–1916.
- Vuzman, D., and Levy, Y. (2010). DNA search efficiency is modulated by charge composition and distribution in the intrinsically disordered tail. *Proc Natl Acad Sci USA* *107*, 21004–21009.
- Vuzman, D., Hoffman, Y., and Levy, Y. (2012). Modulating Protein-DNA Interactions by Post-Translational Modifications at Disordered Regions. *Pac Symp Biocomput* *17*, 188–199.
- Wagner, R.T., and Cooney, A.J. (2009). OCT4: less is more. *Cell Res* *19*, 527–528.

- Wagner, S.A., Beli, P., Weinert, B.T., Nielsen, M.L., Cox, J., Mann, M., and Choudhary, C. (2011). A Proteome-wide, Quantitative Survey of In Vivo Ubiquitylation Sites Reveals Widespread Regulatory Roles. *Mol Cell Proteomics* 10, M111.013284.
- Walsh, C.T., Garneau-Tsodikova, S., and Gatto, G.J. (2005). Protein posttranslational modifications: the chemistry of proteome diversifications. *Angew Chem Int Ed Engl* 44, 7342–7372.
- Wang, E.T., Sandberg, R., Luo, S., Khrebtkova, I., Zhang, L., Mayr, C., Kingsmore, S.F., Schroth, G.P., and Burge, C.B. (2008). Alternative isoform regulation in human tissue transcriptomes. *Nature* 456, 470–476.
- Wang, L., Gural, A., Sun, X.-J., Zhao, X., Perna, F., Huang, G., Hatlen, M.A., Vu, L., Liu, F., Xu, H., et al. (2011a). The leukemogenicity of AML1-ETO is dependent on site-specific lysine acetylation. *Science* 333, 765–769.
- Wang, Q., Zhang, Y., Yang, C., Xiong, H., Lin, Y., Yao, J., Li, H., Xie, L., Zhao, W., Yao, Y., et al. (2010). Acetylation of metabolic enzymes coordinates carbon source utilization and metabolic flux. *Science* 327, 1004–1007.
- Wang, Z., Ding, G., Geistlinger, L., Li, H., Liu, L., Zeng, R., Tateno, Y., and Li, Y. (2011b). Evolution of protein phosphorylation for distinct functional modules in vertebrate genomes. *Mol Biol Evol* 28, 1131–1140.
- Ward, J.J., Sodhi, J.S., McGuffin, L.J., Buxton, B.F., and Jones, D.T. (2004). Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J Mol Biol* 337, 635–645.
- Watanabe, K., and Yokobori, S.-I. (2011). tRNA Modification and Genetic Code Variations in Animal Mitochondria. *J Nucleic Acids* 2011, 623095.
- Weatheritt, R.J., and Gibson, T.J. (2012). Linear motifs: lost in (pre)translation. *Trends Biochem Sci* 37, 333–341.
- Weatheritt, R.J., Davey, N.E., and Gibson, T.J. (2012). Linear motifs confer functional diversity onto splice variants. *Nucleic Acids Res* 40, 7123–7131.
- Webby, C.J., Wolf, A., Gromak, N., Dreger, M., Kramer, H., Kessler, B., Nielsen, M.L., Schmitz, C., Butler, D.S., Yates, J.R., et al. (2009). Jmjd6 catalyses lysyl-hydroxylation of U2AF65, a protein associated with RNA splicing. *Science* 325, 90–93.
- Webster, D.M., Teo, C.F., Sun, Y., Wloga, D., Gay, S., Klonowski, K.D., Wells, L., and Dougan, S.T. (2009). O-GlcNAc modifications regulate cell survival and epiboly during zebrafish development. *BMC Developmental Biology* 2007 7:111 9, 28.

Weinert, B.T., Wagner, S.A., Horn, H., Henriksen, P., Liu, W.R., Olsen, J.V., Jensen, L.J., and Choudhary, C. (2011). Proteome-wide mapping of the *Drosophila* acetylome demonstrates a high degree of conservation of lysine acetylation. *Sci Signal* 4, ra48.

Wells, M., Tidow, H., Rutherford, T.J., Markwick, P., Jensen, M.R., Mylonas, E., Svergun, D.I., Blackledge, M., and Fersht, A.R. (2008). Structure of tumor suppressor p53 and its intrinsically disordered N-terminal transactivation domain. *Proc Natl Acad Sci USA* 105, 5762–5767.

Wend, P., Holland, J.D., Ziebold, U., and Birchmeier, W. (2010). Wnt signaling in stem and cancer stem cells. *Semin. Cell Dev. Biol.* 21, 855–863.

Weternan, M.A., Wilbrink, M., and Geurts van Kessel, A. (1996). Fusion of the transcription factor TFE3 gene to a novel gene, PRCC, in t(X;1)(p11;q21)-positive papillary renal cell carcinomas. *Proc Natl Acad Sci USA* 93, 15294–15298.

Whelan, S., and Goldman, N. (2004). Estimating the frequency of events that cause multiple-nucleotide changes. *Genetics* 167, 2027–2043.

Wickham, H. (2010). A Layered Grammar of Graphics. *J Comput Graph Stat* 19, 3–28.

Wildes, D., and Wells, J.A. (2010). Sampling the N-terminal proteome of human blood. *Proc Natl Acad Sci USA* 107, 4561–4566.

Wilkins, A.D., Lua, R., Erdin, S., Ward, R.M., and Lichtarge, O. (2010). Sequence and structure continuity of evolutionary importance improves protein functional site discovery and annotation. *Protein Sci* 19, 1296–1311.

Wilson, D., Madera, M., Vogel, C., Chothia, C., and Gough, J. (2007). The SUPERFAMILY database in 2007: families and functions. *Nucleic Acids Res* 35, D308–D313.

Wilson, D., Pethica, R., Zhou, Y., Talbot, C., Vogel, C., Madera, M., Chothia, C., and Gough, J. (2009). SUPERFAMILY--sophisticated comparative genomics, data mining, visualization and phylogeny. *Nucleic Acids Res* 37, D380–D386.

Winn, M.D., Ballard, C.C., Cowtan, K.D., Dodson, E.J., Emsley, P., Evans, P.R., Keegan, R.M., Krissinel, E.B., Leslie, A.G.W., McCoy, A., et al. (2011). Overview of the CCP4 suite and current developments. *Acta Crystallogr D Biol Crystallogr* 67, 235–242.

Witze, E.S., Old, W.M., Resing, K.A., and Ahn, N.G. (2007). Mapping protein post-translational modifications with mass spectrometry. *Nat Methods* 4, 798–806.

Wu, L., Zee, B.M., Wang, Y., Garcia, B.A., and Dou, Y. (2011). The RING finger protein MSL2 in the MOF complex is an E3 ubiquitin ligase for H2B K34 and is involved in crosstalk with H3 K4 and K79 methylation. *Mol Cell* 43, 132–144.

Xhemalce, B., and Kouzarides, T. (2010). A chromodomain switch mediated by histone H3 Lys 4 acetylation regulates heterochromatin assembly. *Genes Dev* 24, 647–652.

Xie, H., Vucetic, S., Iakoucheva, L.M., Oldfield, C.J., Dunker, A.K., Obradovic, Z., and Uversky, V.N. (2007a). Functional anthology of intrinsic disorder. 3. Ligands, post-translational modifications, and diseases associated with intrinsically disordered proteins. *J Proteome Res* 6, 1917–1932.

Xie, H., Vucetic, S., Iakoucheva, L.M., Oldfield, C.J., Dunker, A.K., Uversky, V.N., and Obradovic, Z. (2007b). Functional anthology of intrinsic disorder. 1. Biological processes and functions of proteins with long disordered regions. *J Proteome Res* 6, 1882–1898.

Xie, Z., Dai, J., Dai, L., Tan, M., Cheng, Z., Wu, Y., Boeke, J.D., and Zhao, Y. (2012). Lysine succinylation and lysine malonylation in histones. *Mol Cell Proteomics*.

Xiong, B., Lu, S., and Gerton, J.L. (2010). Hos1 Is a Lysine Deacetylase for the Smc3 Subunit of Cohesin. *Curr Biol*.

Xu, G., Paige, J.S., and Jaffrey, S.R. (2010). Global analysis of lysine ubiquitination by ubiquitin remnant immunoaffinity profiling. *Nat Biotechnol* 28, 868–873.

Xue, B., Oldfield, C.J., Van, Y.-Y., Dunker, A.K., and Uversky, V.N. (2012). Protein intrinsic disorder and induced pluripotent stem cells. *Mol Biosyst* 8, 134–150.

Yamzon, J.L., Kokorowski, P., and Koh, C.J. (2008). Stem cells and tissue engineering applications of the genitourinary tract. *Pediatr. Res.* 63, 472–477.

Yanai, I., Benjamin, H., Shmoish, M., Chalifa-Caspi, V., Shklar, M., Ophir, R., Bar-Even, A., Horn-Saban, S., Safran, M., Domany, E., et al. (2005). Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics* 21, 650–659.

Yang, X.-D., Tajkhorshid, E., and Chen, L.-F. (2010). Functional interplay between acetylation and methylation of the RelA subunit of NF-kappaB. *Mol Cell Biol* 30, 2170–2180.

Yap, K.L., and Zhou, M.-M. (2010). Keeping it in the family: diverse histone recognition by conserved structural folds. *Crit Rev Biochem Mol Biol* 45, 488–505.

Yates, J.R., Ruse, C.I., and Nakorchevsky, A. (2009). Proteomics by mass spectrometry: approaches, advances, and applications. *Annu Rev Biomed Eng* 11, 49–79.

Yount, J.S., Moltedo, B., Yang, Y.-Y., Charron, G., Moran, T.M., López, C.B., and Hang, H.C. (2010). Palmitoylome profiling reveals S-palmitoylation-dependent antiviral activity of IFITM3. *Nat Chem Biol* 6, 610–614.

- Yu, B.J., Kim, J.A., Moon, J.H., Ryu, S.E., and Pan, J.-G. (2008). The diversity of lysine-acetylated proteins in *Escherichia coli*. *J. Microbiol. Biotechnol.* *18*, 1529–1536.
- Yu, H., Shao, Y., Gao, L., Zhang, L., Guo, K., Wu, C., Hu, X., and Duan, H. (2012). Acetylation of sphingosine kinase 1 regulates cell growth and cell-cycle progression. *Biochem Biophys Res Commun* *417*, 1242–1247.
- Yuchi, Z., Lau, K., and Van Petegem, F. (2012). Disease mutations in the ryanodine receptor central region: crystal structures of a phosphorylation hot spot domain. *Structure* *20*, 1201–1211.
- Zeng, L., and Zhou, M.-M. (2002). Bromodomain: an acetyl-lysine binding domain. *FEBS Lett* *513*, 124–128.
- Zeng, L.R., Park, C.H., Venu, R.C., Gough, J., and Wang, G.L. (2008). Classification, Expression Pattern, and E3 Ligase Activity Assay of Rice U-Box-Containing Proteins. *Molecular Plant* *1*, 800–815.
- Zhang, K., Chen, Y., Zhang, Z., Tao, S., Zhu, H., and Zhao, Y. (2010a). Unrestrictive identification of non-phosphorylation PTMs in yeast kinases by MS and PTMap. *Proteomics* *10*, 896–903.
- Zhang, Q., Chakravarty, S., Ghersi, D., Zeng, L., Plotnikov, A.N., Sanchez, R., and Zhou, M.-M. (2010b). Biochemical profiling of histone binding selectivity of the yeast bromodomain family. *PLoS ONE* *5*, e8903.
- Zhang, R., and Lin, Y. (2009). DEG 5.0, a database of essential genes in both prokaryotes and eukaryotes. *Nucleic Acids Res* *37*, D455–D458.
- Zhang, T., Wang, S., Lin, Y., Xu, W., Ye, D., Xiong, Y., Zhao, S., and Guan, K.-L. (2012). Acetylation negatively regulates glycogen phosphorylase by recruiting protein phosphatase 1. *Cell Metab* *15*, 75–87.
- Zhang, Z., Tan, M., Xie, Z., Dai, L., Chen, Y., and Zhao, Y. (2010c). Identification of lysine succinylation as a new post-translational modification. *Nat Chem Biol* *7*, 58–63.
- Zhao, S., Xu, W., Jiang, W., Yu, W., Lin, Y., Zhang, T., Yao, J., Zhou, L., Zeng, Y., Li, H., et al. (2010). Regulation of cellular metabolism by protein lysine acetylation. *Science* *327*, 1000–1004.
- Zhou, Q., Brown, J., Kanarek, A., Rajagopal, J., and Melton, D.A. (2008). In vivo reprogramming of adult pancreatic exocrine cells to beta-cells. *Nature* *455*, 627–632.

Zielinska, D., Gnad, F., Wiśniewski, J.R., and Mann, M. (2010). Precision mapping of an in vivo N-glycoproteome reveals rigid topological and sequence constraints. *Cell* 141, 897–907.

Zielinska, D.F., Gnad, F., Jedrusik-Bode, M., Wiśniewski, J.R., and Mann, M. (2009). *Caenorhabditis elegans* has a phosphoproteome atypical for metazoans that is enriched in developmental and sex determination proteins. *J Proteome Res* 8, 4039–4049.