Genome Binding and Gene Regulation by Stem Cell Transcription Factors

Johannes Hendrik Brandsma

Genome Binding and Gene Regulation by Stem Cell Transcription Factors

Johannes Hendrik Brandsma

Colofon

ISBN: 978-94-6299-289-4
Cover design: J.H. Brandsma
Image from ©iStockphoto.com/llebbid
Lay-out: J.H. Brandsma
Printed by Ridderprint BV, The Netherlands

The studies presented in this thesis were conducted at the department of Cell Biology, Erasmus Medical Centre, Rotterdam and financially supported by NWO ALW open grant: 821.02.004

Genome Binding and Gene Regulation by Stem Cell Transcription Factors

Binding aan het genoom en genregulatie door stamceltranscriptiefactoren

Proefschrift

Ter verkrijging van de graad van doctor aan de
Erasmus Universiteit Rotterdam
op gezag van de rector magnificus
Prof.dr H.A.P. Pols
en volgens besluit van het College voor Promoties

De openbare verdediging zal plaatsvinden op
dinsdag 16 februari 2016 om 15:30

door

Johannes Hendrik Brandsma

geboren te Heerenveen

Erasmus University Rotterdam

Promotiecommissie

Promotor:	Prof.dr. F.G. Grosveld
Overige leden:	Prof.dr. J.N.J. Philipsen Prof.dr. J.H. Gribnau Prof.dr. W.L. de Laat
Copromotor:	Dr. R.A. Poot

“At every step, he felt as a man might feel who, after admiring the smooth, cheerful motion of a boat on the water, actually gets into the boat himself. He saw that apart from having to sit steadily in the boat without rocking, he also had to keep in mind, without forgetting for a moment where he was going, that there was water beneath his feet, that he had to row, that his unaccustomed hands hurt, and that it was easy only when you looked at it, but that doing it, though it made you very happy, was very hard.”

Leo Tolstoj, Anna Karanina , Bantam classic reissue (English translation), 2006

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

ASD	autism spectrum disorders
bp	basepair
CCHS	Congenital Central Hypoventilation Syndrome
ChIP	Chromatin Immunoprecipitation
ChIP-MS	ChIP followed by Mass spectrometry
ChIP-qPCR	ChIP followed by quantitative PCR
ChIP-seq	ChIP followed massively parallel DNA sequencing
DNA	Deoxyribonucleic acid
E	Embryonic day
emPAI	Exponentially modified protein abundance index
ESC	Embryonic Stem Cell
GO	Gene ontology
GWAS	Genome-wide association study
ID	Intellectual disability
IP	Immunoprecipitation
IP-MS	Immunoprecipitation followed by mass-spectrometry
iPSC	Induced pluripotent stem cells
kDa / kD	kilo Dalton
KD	Knock Down
KO	Knock Out
LOF/LoF	loss of function
MD	Mental disorders
mRNA	messenger RNA
NSC	Neural Stem Cell
Oct4	Octamer-binding transcription factor 4
PCR	polymerase chain reaction
PolII	RNA Polymerase II
RNA	Ribonucleic acid
SNP	Single nucleotide polymorphism
SO-motif	Sox2-Oct4 composite DNA binding motif
Sox2	SRY (sex determining region Y)-box 2
STARR-seq	Self-transcribing active regulatory region sequencing
SZ	Schizophrenia
TAD	topological associated domain
TSS	Transcription Start Site
wt	wild type

OUTLINE

Nearly all cells of an individual organism contain the same genome. However, each cell type transcribes a different set of genes due to the presence of different sets of cell type-specific transcription factors. Such transcription factors bind to regulatory regions such as promoters and enhancers and regulate their activity in gene transcription. Transcription factors interact with each other and form tissue-specific transcription factor networks. Identification of genome binding and gene regulation by transcription factors will enhance our understanding of how transcription factors specify cell types.

Chapter 1 serves as a general introduction into eukaryotic transcription regulation and describes the role of enhancers and transcription factors. Chapters 2 – 4 describe the experimental work of this thesis. **Chapter 2** describes Chromatin Immunoprecipitation followed by Mass Spectrometry (ChIP-MS), a technique used to identify transcription factors and other genome binding proteins that bind to promoters, enhancers or heterochromatin. **Chapter 2** specifically describes the genome-binding pattern of transcription- and pluripotency factor Dppa2, which we identify by ChIP-MS and ChIP-seq to bind to promoters outside the classical pluripotency network. **Chapter 3** describes a transcription factor interaction network of over 200 proteins in neural stem cells, which was assembled by the identification of interaction partners of four mental disorder-associated transcription factors. **Chapter 4** describes work on the genome-wide localization of Oct4, Sox2 and Nanog in embryonic stem cells and investigates the genome localization of these transcription factors upon depletion of Oct4 or Sox2 from embryonic stem cells. The last chapter, **Chapter 5**, summarizes the results from Chapter 2-4 and provides additional discussion to these chapters, concerning its implications and future directions.

Chapter1

Introduction

INTRODUCTION

The most abundant class of cis-regulatory elements are enhancers[1]. Around half a million putative enhancers have been identified in the human genome[2,3], vastly surpassing the number of human protein-coding genes. The number of active enhancers in a single cell type ranges in the ten thousands[2]. Enhancer activity is cell-type specific and controls cell identity by allowing the regulation of gene expression[1,2,4-6]. Enhancers are cis-acting regulatory sequences that stimulate transcription of genes by interacting with their promoters[7] and enhancer-promoter looping is a prerequisite for transcription[8]. Enhancers contain DNA motifs that allow the specific binding of certain transcription factors and are often bound by multiple transcription factors[9]. Transcription factors recruit factors that modify chromatin accessibility and recruit co-activators such as P300 and the Mediator complex to the enhancer. The Mediator complex interacts with transcription factors that are bound to the enhancer and with the pre-initiation complex and RNA polymerase II (PolII) that are bound to the promoter through chromatin looping and as such driving transcription (see also below)[10]. As enhancers play an important role in gene regulation, it can be expected that malfunction in enhancer-mediated gene regulation might be the underlying cause for several genetic diseases. Enhancers and other cis-regulatory regions harbor common trait and disease variants[6,11,12]. Single-nucleotide polymorphisms (SNPs) identified in Genome-wide association studies (GWAS) were found to be particularity enriched on enhancers and transcription start sites (TSSs) for several cell types[2]. Many more mutations in enhancers are expected to be found with the recent onset of next generation sequencing technology. This chapter aims to summarize these recent findings surrounding enhancers and to discuss the role of enhancers in disease. First the genome-wide techniques that were utilized to increase our understanding of enhancers are summarized. Then the current understanding of transcriptional enhancers and their role in the transcriptional regulatory landscape and development is discussed. Finally, the aberrant behavior of enhancers during disease is discussed and illustrated with examples from the literature.

Identification and characterization of enhancers using NGS techniques

DNA is wrapped around nucleosomes. Nucleosomes are formed by an octamer of histone proteins, that is wrapped by 147 base pairs of DNA. The octamer is formed by two heterodimers consisting of H3 and H4 histones flanked by a H2A- and H2B histone. The N-terminal sequences of H3 and H4 histones, also called the histone tails, protrude from the nucleosome core and are post-translationally modified (i.e. 'histone modifications'). Histone modifications mark distinct regions of the genome and may affect chromatin structure and may attract proteins such as transcription factors and co-activators (reviewed in [13]). Several techniques have been developed that can be used to identify and characterize enhancers genome-wide (reviewed in [14], see table 1) and the leading technique in this respect, Chromatin Immunoprecipitation followed by deep sequencing (ChIP-seq), targets histone modifications that mark enhancers. Enhancers are marked by H3 mono-methylation at Lysine 4 (H3K4me1) and for genome-wide identification of enhancers ChIP-seq can be utilized by targeting enhancer marker H3K4me1[5,15]. In combination with a ChIP-seq for acetylation of H3K27 (H3K27ac) active enhancers can be identified[15]. ChIP-seq for these histone modifications can be conveniently performed from 1 million cells. Several specialist approaches have been

developed for histone ChIP-seqs on reduced numbers of cells ranging from 100,000[16], 10,000[17], 5,000[18] to 1,000 cells[18,19]. ChIP-seq for co-activator P300[20] or subunits of the co-activator complex Mediator[6] have also been used to identify enhancers. Other approaches constitute of multiple ChIP-seqs against different key transcription factors of a cell type and identifying so-called multi transcription factor binding loci[9,21]. DNaseI hyper sensitivity assay followed by deep sequencing (DNase-seq) identifies the open regions of the genome, which includes enhancers, but also other genetic regulatory elements such as promoters, silencers and insulators[22]. Another technique that assesses open chromatin regions are transposase-accessible chromatin using sequencing (ATAC-seq)[23] and formaldehyde-assisted isolation of regulatory elements (FAIRE-seq)[24].

Method	Characteriztion	Remarks
ChIP-seq against H3K4me1	Genomic location	Does not distinguish between active and poised enhancers
ChIP-seq against H3K27ac	Genomic location of active enhancers	H3K27ac also marks promoters
ChIP-seq against P300	Genomic location of active enhancers	P300 also binds to promoters and is sometimes present on poised enhancers
ChIP-seq against Mediator	Genomic location of active enhancers	Mediator is also present on promoters
ChIP-seq against key transcription factors	Genomic location	Does not distinguish between active and poised enhancers. Only identifies a subset of all enhancers
DnaseI Hypersensitivity Sequencing, FAIRE-seq, ATAC-seq	Genomic location of open regions	Does not distinguis between enhancers and other regulatory elements
RNA-seq and GRO-seq	Genomic location and activity	Not all enhancers produce eRNAs
HiC, CaptureC, T2C	Genomic Interaction	Often lacks the resolution to map interactions between proximal enhancers
ChIA-PET	Genomic location and interaction	
STARR-seq	Activity	Measures potential activity of genomic sequences.
ChIP-MS against H3K4me1	Enhancer binding proteins	Does not characterize individual enhancers
ChIP-MS against H3K27ac	Active enhancer binding proteins	Does not characterize individual enhancers

Table1: Overview of OMICS methods to characterize Enhancers

Enhancers can also be identified via RNA-seq by detection of so-called enhancer RNAs (eRNAs, see below). An interesting RNA-seq variant is Global run-on sequencing (GRO-seq)[25], which instead of quantifying global transcript levels, specifically assesses ongoing transcription in the nucleus. An advantage to GRO-seq is that it can be used to identify eRNAs and simultaneously assess transcriptional activity of nearby genes[26]. Chromosome Conformation Capture (3C) and evolved 3C-variants, form another very important category of techniques for characterizing enhancers and other cis-regulatory regions. These techniques enable the identification of chromatin domains that are in close proximity (i.e. 'interaction of chromatin')[27] and can be used to confirm the interaction of an enhancer with its promoter. HiC is the only genome-wide variant within the 3C family and can be used to map all chromatin-chromatin interactions with an average resolution of 10.5 kb[28]. Two recently published 3C techniques known as Capture-C[29] and T2C[30] are not genome-wide, but offer an improved average resolution of 2 – 6 kb[30] and do not rely on ultra-deep sequencing[30]. Another technique that combines ChIP with 3C technology is called 'chromatin interaction analysis by paired-end tag sequencing' (ChIA-

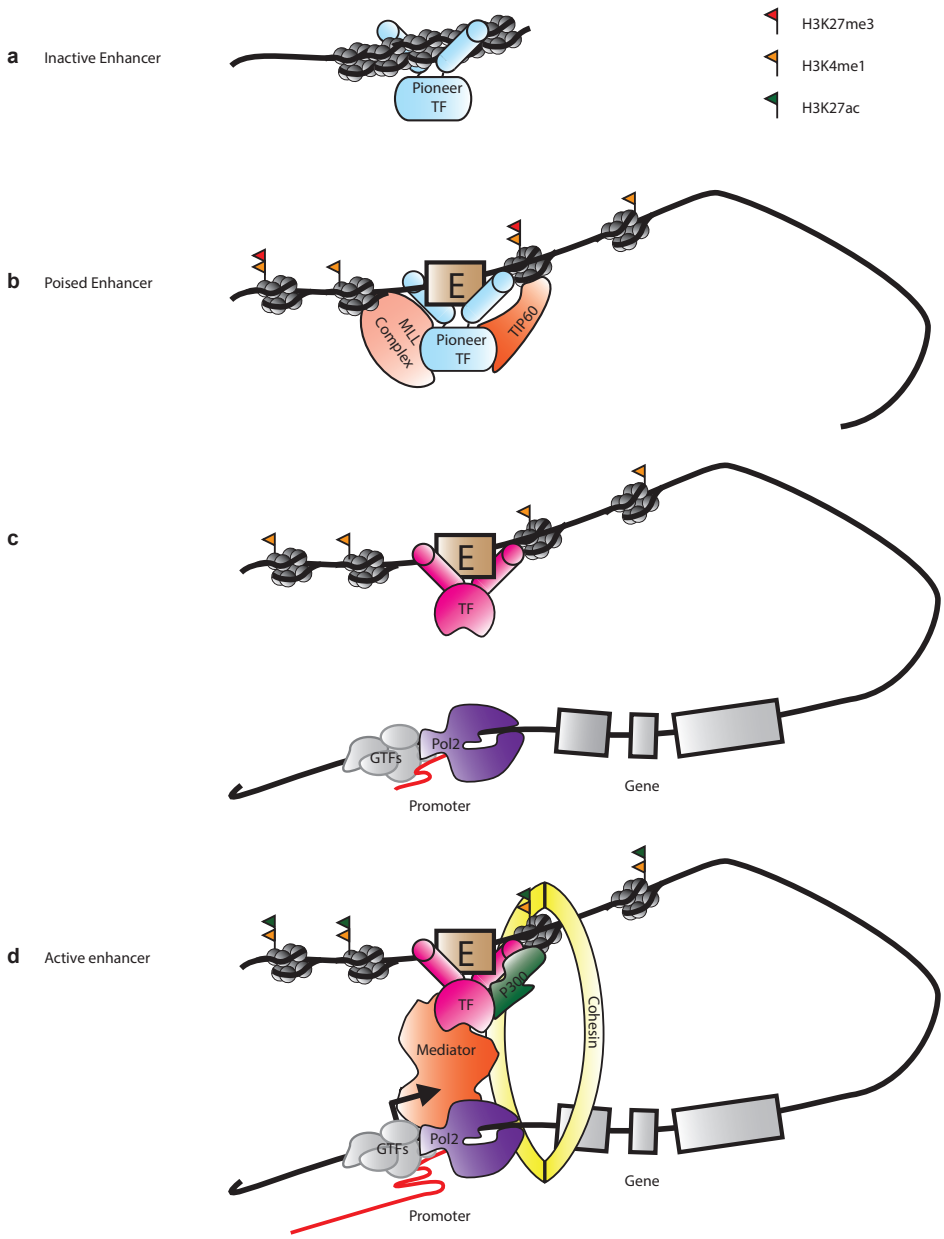


Figure 1: Model of enhancer activation in Eukaryotes. (a) A DNA motif in a genomic region with closed unmarked chromatin is bound by a pioneer transcription factor. **(b)** The Pioneer transcription factor attracts chromatin modifying complexes like e.g. the MLL complex that monomethylates H3K4. H3K4me1 attracts the TIP60 complex that deposits H2A.Z, which leads to open and more accessible chromatin. Poised enhancers are sometimes also marked by H3K27me3. **(c)** The increased accessibility of the chromatin allows other transcription factors to recognize their DNA binding motifs and bind the genome. The promoter may already be bound by general transcription factors of the pre-initiation complex and Polymerase II. **(d)** These transcription factors recruit co-activators like p300 that acetylates H3K27 and the Mediator complex

Legend continues on the bottom of the next page

PET). ChIA-PET is targeted against (e.g.) a transcription factor. This allows identification of the binding regions of the transcription factor, but also identifies regions that are interacting with the binding region[31]. Self-transcribing active regulatory region sequencing (STARR-seq) assesses the transcriptional potential of enhancers. In this high throughput technique, potential enhancers are cloned into a construct downstream of a minimal promoter. Activation of transcription by the potential enhancers results in transcription of the enhancer itself. After a cell line has been transfected with the constructs, RNA is isolated and analyzed for over-represented potential enhancers[32]. To identify enhancer-bound proteins, Chromatin Immunoprecipitation followed by Mass Spectrometry (ChIP-MS) for H3K4me1 can be utilized (see Chapter 2).

Establishment and activation of enhancers

Enhancers can be separated in three categories. Inactive enhancers, poised enhancers and active enhancers (figure 1). In contrast to poised and active enhancers, inactive enhancers cannot be distinguished from the general chromatin landscape by e.g. chromatin marks or transcription factor binding[33]. Poised and active enhancers are both characterized by H3K4me1[5,34], whereas active enhancers also exhibit H3K27ac[35,36]. In addition, poised enhancers are often also enriched for H3K27me3[37] (figure 1). In an exemplary study using STARR-seq, a third of the potential active enhancer sequences were shown to be silenced by H3K27me3, but were able to drive transcription outside their genomic environment[32] . Poised and active enhancers are equally likely to interact with the TSS[28].

Pioneer transcription factors play an important role in establishing poised enhancers. Pioneer transcription factors are a special class of transcription factors, which are able to engage their target sites in closed and silent chromatin lacking any evident histone marks making it accessible for other chromatin binding proteins[38] (reviewed in[39]). However, not all types of closed chromatin are permissive to pioneer factors. Heterochromatin consisting of H3K9me2 or H3K9me3 impedes pioneer transcription factor binding[38,39]. H3K27me3 marked silenced chromatin is probably more permissive, but this remains to be investigated. A study on a pioneer transcription factor PU.1 in macrophages and B cells demonstrated that PU.1 initiates nucleosome remodeling, which increases the accessibility to other TFs and eventually facilitates mono-methylation of H3K4[40]. Based on this and other data[41], a general model was proposed in which pioneer transcription factors make unmarked closed chromatin accessible, allowing other transcription factors to bind to the chromatin and attract chromatin remodelers such as the MLL complex, which catalyzes the mono-methylation of H3K4[39,40] (figure 1). In *Drosophila* it was shown that H3K27ac signal always coincides with H3K4me1 signal on enhancers, while H3K4me1 does occur without H3K27ac[36]. This suggests that the poised state probably always precedes the active state in enhancers. Other studies in mouse and human do report H3K27ac apart from H3K4me1 on enhancers[35,37,42], but it was suggested that this is due to under-sampling of the H3K4me1 signal[36]. The biological function of H3K4me1 is still

(Figure 1. Legend continues from previous page)
that bridges regulatory signals to the promoter and stimulates initiation or elongation of transcription by Polymerase II. Cohesin acts to stabilize the enhancer-promoter loop. E: Enhancer. TF: Transcription factor. GTFs: General transcription factors. PolII: RNA Polymerase II. Black line: DNA, Red line: RNA. Flags indicate presence of histone modification: H3K27me3 (red), H3K4me1 (orange), H3K27ac (green).

under debate, but has been suggested to specifically attract the TIP60/p400 complex that catalyzes deposition and acetylation of the H2A.Z, which leads to increased accessibility of enhancer chromatin[43].

Poised enhancers allow for rapid activation of gene expression. Upon IFN-γ stimulation in HeLa cells, STAT1 bound to unmarked genomic regions and to poised enhancers marked by H3K4me1. The genes associated with STAT1 bound poised enhancers showed increased expression within 30 minutes after IFN-γ stimulation, while genes associated with unmarked STAT1 binding sites did not[5]. H3K4me1 may also allow for rapid reactivation of enhancers. Macrophages stimulated with LPS, activated inactive enhancers. When the macrophages were no longer stimulated with LPS, H3K27ac was rapidly lost, while H3K4me1 was sustained. Renewed LPS stimulation resulted in faster induction of gene expression than observed with the initial LPS stimulation[44].

Poised enhancers can also be bound and activated by transcription factors that are expressed in later development or differentiation. For example, the enhancer for the λ5 gene is bound by Sox2 and Foxd3 and kept in a poised in embryonic stem cells (ESCs). Later in pro-B cells this enhancer is bound by Sox4 and the λ5 gene is highly expressed[45]. Another example is the enhancer for the *Alb1* gene, which is also kept in a permissive state by Foxd3. Upon mesodermal differentiation the *Alb1* enhancer is bound by FoxA1 and FoxA2, eventually activating transcription of the *Alb1* gene in hepatocytes[46]. In the case of ESC to neural differentiation, Sox2 binds enhancers in ESCs that are later activated by Sox3 in neural progenitor cells. Besides that, Sox3 also preselects for enhancers in neural progenitor cells that are later activated by Sox11 in differentiating neurons[47]. Transcription factors such as these are able to recognize their DNA sequence motif in open chromatin and directly bind the DNA, often by interacting with and attracting other transcription factors, co-activators and chromatin associated complexes (figure 1).

One such co-activator that is attracted by transcription factors to active enhancers is the Mediator complex. The Mediator complex bridges regulatory signals of DNA binding transcription factors on enhancers to PolII bound to the promoters as part of the pre initiation complex and as such affects the initiation, pausing and elongation of PolII(reviewed in[10]) (figure 1). The binding of transcription factors to Mediator evokes conformational changes in the Mediator complex[48], which facilitates interaction with PolII[49] and triggers activation of stalled PolII[50]. Mediator subunit CDK8 has been speculated to be part of Mediator after pausing of PolII, attracting CDK9 as part of the super elongation complex, which in turn phosphorylates PolII, promoting release of PolII and transcription elongation[10]. Mediator also stabilizes promoter-enhancer loops by stimulating deposition of cohesin, which may form Cohesin ring-like structures between different DNA elements[51] like observed between sister chromatids[52].

CTCF together with Cohesin is also required for promoter-enhancer looping[53,54]. CTCF binds the genome via its specific DNA binding motif [55]. The orientation of CTCF binding motifs on the genome determine the direction of chromatin looping[56]. Two opposite oriented CTCF binding motifs with CTCF bound in opposite topology are believed to establish chromatin loops that facilitate interaction between promoters and enhancers

via cohesin deposition[56]. Transcription itself is not required for the establishment of enhancer-promoter loops[57].

P300 and CBP are also co-activators that are attracted by transcription factors to enhancers (figure 1), and over 400 known transcription factors interact physically and functionally with P300[58,59]. P300 and CBP catalyze acetylation of H3K27[60,61]. Like H3K27ac[35,36], P300 binding is also used as a specific marker for active enhancers[62]. However, not all P300 binding sites are active enhancers and binding of P300 to enhancers does not always lead to acetylation of H3K27 as a subset of P300-bound poised enhancers lacks H3K27ac and/or are marked by H3K27me3[37]. This suggests that recruitment and the enzymatic activity of P300 are separately regulated[43].

H3K27ac is recognized by bromodomain containing proteins. P300 and other histone acetyl transferases contain bromodomains and therefore amplify their own binding[43]. Also some subunits of the TFIID complex contain bromodomains and although these and other proteins are initially probably recruited through bromodomain-independent mechanisms, recognition of acetylated lysines may stabilize and amplify activating signals[43]. Also Brd4 and other members of the BRD protein family recognize H3K27ac. Brd4 has been shown to assist in transcription elongation with PolII, but this is not dependent on the bromodomain[63]. A suggested role of histone acetylation is its potential to affect chromatin structure and interactions, because of its charge-nullifying ability and thus might regulate enhancer-promoter communication[43].

Transcription factors cooperate in genome binding

Transcription factors play an important role in establishing and activating enhancers (see above). It has been estimated that there are around 2600 DNA binding transcription factors encoded by the human genome[64], of which 200 – 300 are expressed in each cell type[65]. Many of these transcription factors interact[66,67] and form cell- and tissue-specific transcription factor networks, in which tissue-specific transcription factors interact with ambiguously expressed transcription factors[66].

DNA binding transcription factors often specifically bind to a specific sequence on the genome, also called the DNA binding motif. Depending on the expression of an interaction partner, a transcription factor can be targeted to different genomic targets. For example, in ESCs Oct4 (Pou5f1) and Sox2 interact and cooperatively bind to a DNA motif that consists of joint motif between a known POU binding and Sox binding motif[68] (see chapter 4), maintaining pluripotency of ESCs[69]. When Oct4 partners up with other transcription factors it leads to differentiation. Increased expression of Sox17, as seen in the inner cell mass during mesodermal differentiation, invokes Oct4 to switch interaction partners from Sox2 to Sox17 and thereby targets genes that trigger endodermal differentiation[70]. Expression of Otx2, as is observed when ESCs exit pluripotency, leads Oct4 to bind Otx2 and results in recruitment of Oct4 by Otx2 leading to the activation of enhancers associated with neural differentiation[71]. Also Sox2 is directed to different targets depending on its interaction partner. Besides that Sox2 interacts with Oct4 in ESCs (see above), it also interacts with Nanog in ESCs and this regulates ESC self-renewal[72]. During development, Sox2 forms a DNA binding complex with Pax6 that regulates initiation of lens development for the eye[73]. Sox2 also induces hair cell fate in

the cochlea with Eya1 and Six1[74] and is involved in many more interactions[75].

As mentioned above, transcription factors recruit co-regulators and chromatin modifying proteins. The effect of binding of a transcription factor on chromatin or transcription is often determined by the other proteins it attracts. For example, Oct4 has been described as a transcriptional activator, with 89% of its target genes being down regulated upon knock-down[76]. Correspondingly, Oct4 interacts with, amongst others, the subunits of the activating BAF complex[77]. The interface that Oct4 most likely uses for interaction with BAF subunit Smarca4 was found to be essential for iPS reprogramming[78]. However Oct4 is also associated with silenced genes[79] and was found to interact with subunits of the Polycomb repressive complex1 (PRC1) and Nucleosome remodeling and deacetylase (NuRD)[77,80] and to recruit the H3K9 methyltransferase Eset to silence trophoblast-associated genes[81-83].

Redundancy within and between enhancers

When addressing the role of enhancers in disease, it is important to consider the redundancy in enhancers regulating the same gene(s). It has been proposed that besides the varying levels of redundancy observed between (homologous) genes, there is also varying redundancy between regulatory elements[84]. While some enhancers are essential, others display varying degrees of redundancy (reviewed in[85]). An early publication describes closely situated enhancer elements driving expression of *Sgs4* in *Drosophila*, where each combination of two elements was sufficient to direct expression, but none of the three could act alone[86]. However, apparently redundant enhancers are often well conserved and should be studied in the context of development[85]. For example, the *Svb* gene is essential for development in *Drosophila*. The *Svb* gene is regulated by a primary essential enhancer and two secondary enhancers. The secondary enhancers (also called shadow enhancers) play only a minimal role in optimal culturing conditions, but are essential when more environmental and genetic variability is introduced[87]. In a related study it was proposed that a gene is often regulated by multiple enhancers, because multiple enhancers have more chance to form successful promoter-enhancer interactions, resulting in a higher robustness of target gene expression[85,88]. Essential and redundant enhancers both exist, but it is difficult to estimate which of the two categories is more prominent, in part due to publication bias against negative results[85]. However, akin to redundancy between genes, varying degrees of redundancy between enhancers will determine whether point mutations result in abnormalities and disease.

Enhancer RNAs

Another interesting feature of active enhancers is PolII mediated bi-directional transcription at the enhancer site, resulting in the non-coding RNA class enhancer RNAs (eRNAs). Transcription of eRNAs was shown to correlate with transcription of nearby active genes [89]. Knock-down of the eRNA associated with P53 enhancers resulted in reduced expression of surrounding genes. This demonstrated that eRNA is more than just a by-product of transcriptional activation[90]. Targeted RNA interference using siRNA against several eRNAs suggested a functional role for eRNAs in transcription of target genes and promoter-enhancer looping[91].

Enhancers in the context of the three dimensional structure of the genome

The genome is organized into topologically associated domains (TADs), which are self-interacting regions defined by regulatory elements called insulators[92]. TADs in mouse and human have a median sizes of around 800 kb[92]. TADs are found to be highly conserved between different cell types and also between species[5,92]. Interaction between genomic regions (for example enhancer to promoter) are generally restricted to genomic regions within a TAD[92,93] and expression patterns of genes within TADs correlate[93]. TADs are defined by insulator elements[92] and these insulators limit interactions from one TAD to another. CTCF binds these insulator elements via its specific DNA binding motif[55] and is required for the enhancer blocking activity of insulators[94]. Global depletion of CTCF increases interactions between neighboring TADs, while reducing interactions within a TAD[95].

Within a TAD, genomic distance is not a restriction for interaction. Enhancers can be as distant to their gene target as 1 Mb, exemplified by an enhancer of the Sonic Hedgehog gene[96]. It is therefore important to understand that the real physical distance that two regions have to bridge in order to interact is not directly represented by their linear genomic distance. Instead, linear genomic distance between promoter and enhancer should be perceived in the context of genomic organization, where a TAD organizes promoters and enhancers such that the physical distance between enhancer and promoter is decreased and interaction becomes possible . Distal elements often do not interact with the closest transcription start site[97-99] and are instead often located hundreds of kb away from their targets. The exact range in which enhancers act varies amongst different studies and is likely to be influenced by the different methods used and the resolution these methods offer. HiC found that only 25% of enhancer promoter pairs are within a 50 kb range and that 57% span 100 kb or larger[28]. Another HiC study found the median distance between interacting regions to be around 120 kb genomic distance[98]. A ChIA-PET study of PolII, for two cell lines found a median distance between promoter and enhancers of 57 kb and 37 kb respectively (calculated from Table S5 of [99]), with many promoter-enhancer interactions also spanning distances of 150 kb – 200 kb, but only a few surpassing several megabases[99]. ChIA-PET of enhancer-binding oestrogen-receptor-alpha found 86% of all duplex interactions (2 regions interact) to be spanning distances within 100 kb, with less than 1% bridging distances beyond 1Mb, while complex interactions (2+ regions interact) were often found to span more than 100 kb[31].

The majority of promoters and cis-regulatory regions interact with multiple genomic regions[98,99]. The enhancers of the locus control region of the β-globin gene have been suggested to collaborate by aggregating into a single active chromatin hub[100], where the enhancers from the locus control region form loops with the nearby globin genes[101]. Deletion of individual enhancer elements of the locus control region resulted in only modest reduction of β-globin gene expression, but deletion of the entire locus control region resulted in a dramatic reduced expression of the β-globin gene[102] (reviewed in[100]). Several enhancers can interact with a single promoter, but also promoters interact with promoters[99]. Genes corresponding to interacting promoters often have correlated expression[99] and might be part of chromatin hubs as described for the locus control region of the β-globin gene or transcription factories. However, a subclass of interacting promoters display enhancer-like chromatin modifications and act as enhancers driving

the expression of other promoters[32,99].

The evolving enhancer

Studies across different eukaryotic species revealed that enhancers underlie evolutionary changes between species[103-106] and that enhancers are, in contrast to promoters, not well conserved between species[103]. A possible explanation is that a gene often has one promoter, while its expression can be driven by various enhancers with different levels of redundancy (see above). I.e. mutations in enhancers have less chance of being deleterious than mutations in promoters, allowing for more evolutionary flexibility in enhancers.

Enhancers involved in early embryogenesis and organ ontogenesis tend to be better conserved than enhancers involved in later tissue differentiation, which is reflected by the morphology of species during the different stages of embryonic development[107,108]. For example, conserved enhancers amongst species in liver were found near liver-specific genes[103]. A study across 20 mammalian species revealed that recently evolved enhancers dominate the regulatory landscape[103]. The majority of recently evolved enhancers derived from the adaptation of ancestral DNA[103,109], frequently ancestral enhancers that are adapted to a new function[109]. Only a minority derived from expansion of repeat elements[103].

A study on two enhancers regulating the Krüppel transcription factor in three different *Drosophila* species showed that, although total expression levels of Krüppel were conserved, the relative portion of total expression driven by either of the two enhancers varied across the species[110]. In other words, enhancer activity lost by one enhancer, is gained by the other enhancer. In fact, as evolutionary distance between species increases, such compensatory changes between enhancers are increasingly observed[106], suggesting that this flexibility amongst enhancers leads to variation upon which evolution can act[106,110].

An appealing example that illustrates recent adaptation of existing enhancers in human evolution, is an enhancer regulating Lactase-phlorizin hydrolase (*LPH*). *LPH* is for most humans specifically expressed during infancy and early childhood, allowing digestion of lactose from breast milk. Persistent expression of *LPH* (lactase persistence) during adulthood enables lifelong digestion of lactose. Although this adaptation convergently evolved for several heterogeneously distributed populations, the different 5 underlying SNPs found so far are all located in the same enhancer region 14 kb upstream of *LPH* and increase the enhancer activity of this region[111-114]. One of these SNPs, 13910T, completely associates with lactase persistence in the European population and creates a strong binding site for transcription factor Oct1 resulting in maintained *LPH* enhancer activity through adult life. In contrast, in the non-lactase persistence associated variant 13910C Oct1 only binds the enhancer efficiently when weaning age-specific *HNF1α* is also expressed[115].

Super-enhancers

A subset of enhancers was recently described as so called super-enhancers in a publication by Whyte et al.,[6]. Super-enhancers are domains that consist of clusters of

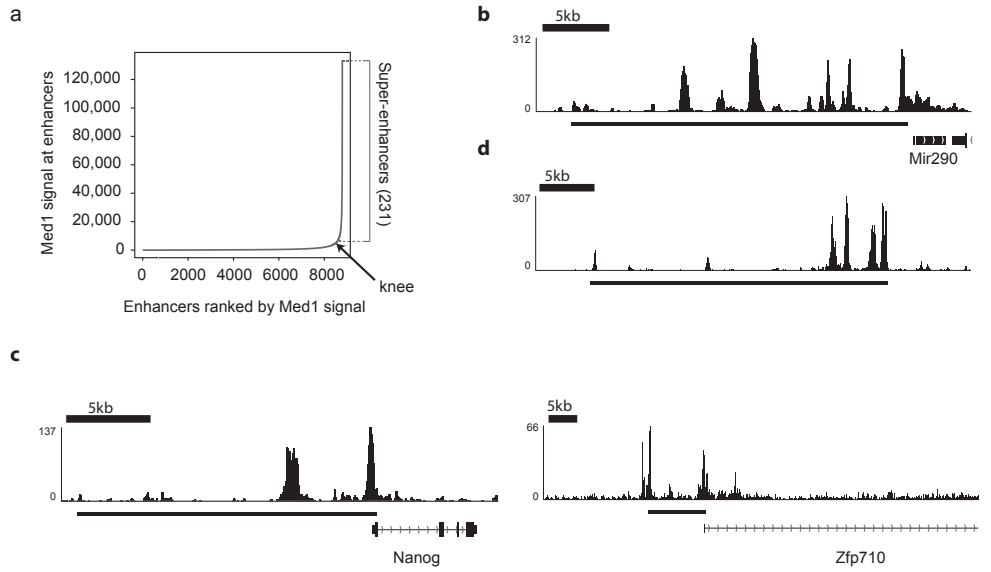


Figure 2. Super enhancers. (a) Distribution of Mediator ChIP-seq signal (total reads) across joint enhancers from Whyte et al.,[6]. Figure adapted from Figure 1C in Whyte et al.,[6]. Arrow indicates the knee in the curve, which was used as a cut-off between super-enhancers and regular enhancers. **(b -d)**. Binding profile of the super-enhancer as determined in Whyte et al.,[6], y-axis displays sequence reads. Black bars indicate the region determined as super-enhancer by Whyte et al.,[6]. **(b)** Binding profile of super-enhancer near *Mir290-295*. **(c)** Binding profile of super-enhancers near the *Nanog* and *Zfp710* genes, that display Med1 ChIP-seq signal on the promoter. **(d)** Binding profile of a super-enhancer downstream of *Sox2*, that contains large regions that display little ChIP-seq signal.

enhancers densely occupied by master regulator transcription factors and transcriptional co-activator Mediator and are often found near genes that define cell identity[6]. Super-enhancers are associated with genes that control cell identity[6,116] and super-enhancers at key oncogenes are acquired in cancer[117].

To determine super-enhancers in a given cell type, Mediator1 (Med1) ChIP-seq signal or ChIP-seq signal of other enhancer-associated proteins and histone modifications was used to determine 'regular' enhancers. Regular enhancers within 12.5 kb were joined into a single region. These joint regions and non-joined enhancers were ranked for their cumulative ChIP-seq signal. All the enhancers beyond the knee in the curve were considered super-enhancers (figure 2a) and contain an exceptionally high density of Mediator. The 'knee of the curve' and the 12.5 kb cutoffs do not appear to be based on a biological rationale[118]. Citing the methods Whyte et al.,[6]: "*We first scaled the data such that the x and y axis were from 0-1. We then found the x axis point for which a line with a slope of 1 was tangent to the curve. We define enhancers above this point to be super-enhancers, and enhancers below that point to be typical enhancers".* The knee of the curve as cut-off between regular and super-enhancers is sensitive to outliers. Mediator signal is also found on promoters and this is ignored by the original manuscript[6], where in some cases super-enhancers also contained Mediator signal that is located on the promoter (Figure 2b). Super-enhancers often consists of several -often putative- enhancers that are separated by stretches of up to 12.5 kb that often contain very little ChIP-seq signal (Figure 2b,c).

Diseases caused by ectopic enhancer binding

Variants that affect enhancers and are associated with disease are found on enhancers[2,119], in transcription factor DNA binding domains[119,120] and protein-protein interaction domains[120]. Mutations that result in the disruption of the genomic environment of the enhancer have also been associated to genetic disease and resulted in different interactions of the enhancer with its targets (see below). Together these variants and mutations have the potential to affect normal enhancer function (figure 3). Considerations for each category are summarized in Table 2.

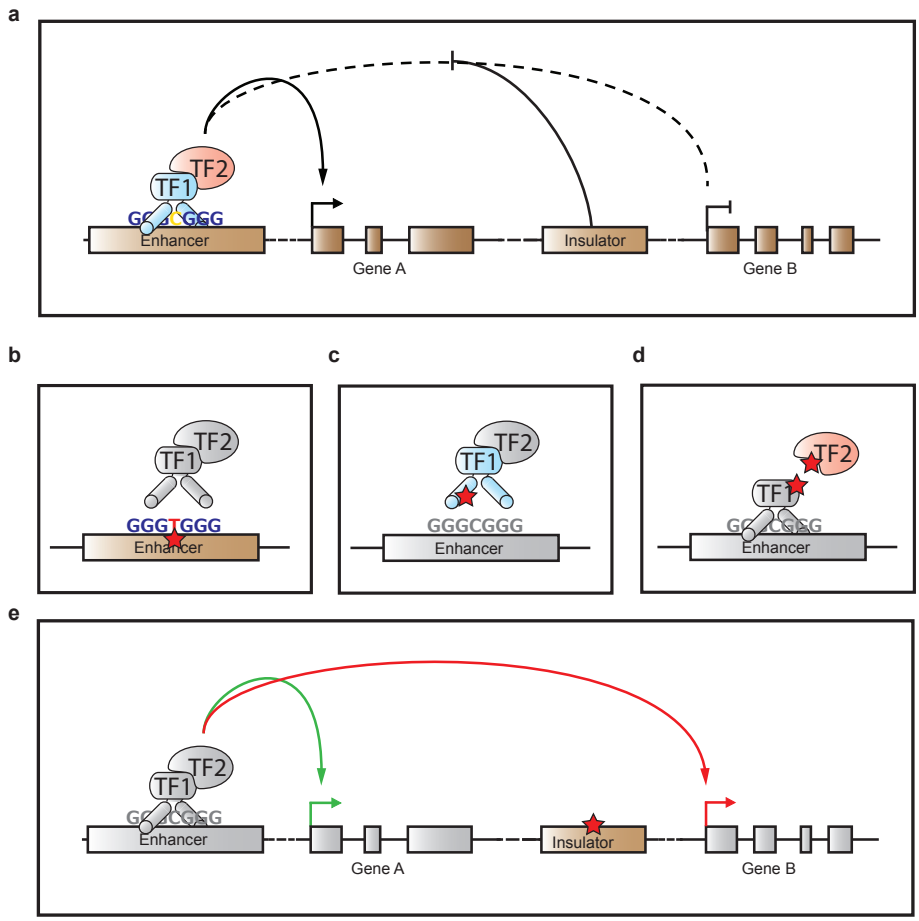


Figure 3: Overview of mutations that affect gene regulation by enhancers and cause disease. **(a)** Normal enhancer function. A transcription factor (TF1) binds by its DNA binding domain to a specific DNA motif on an enhancer and recruits another transcription factor (TF2). Together these transcription factors are required for enhancer-activated transcription of Gene A. The insulator element prevents contact between the enhancer and the promoter of gene B. **(b-e)** Mutations (red star) affecting enhancer function and causing misregulation of genes. **(b)** A mutation in the DNA binding motif in the enhancer element impedes binding of a transcription factor. **(c)** A mutation in the DNA binding domain of the transcription factor changes DNA binding specificity or abolishes DNA binding. **(d)** A mutation in the protein-protein interaction interface of the transcription factor impedes binding of other transcription factors. **(e)** Mutations that affect the position of an insulator element relative to the enhancer.

Category	effect	direct/indirect	affected genes
Enhancer element	ectopic binding of enhancer element by transcription factors	direct	locally
Genomic environment of enhancer	ectopic regulation of genes	indirect	locally
Transcription Factor: DNA binding domain	Loss of binding to enhancer elements by transcription factor	indirect	globally
Transcription Factor: Protein-protein interface	Loss of recruitment of co-activating transcription factors	indirect	globally

Table2: Categories of mutations that affect enhancer function

Mutations in enhancers

Enhancer function can be disrupted by deletion of the enhancer, leading to disease. For example, a genomic deletion of the enhancer that regulates the Sclerostin gene was implicated in Van Buchem disease[121]. A deletion of a secondary enhancer to *ATOH7* was implicated in a local heritable form of blindness amongst Kurds living in Northern Iran[122].

A wide range of DNA binding motifs facilitate binding of specific transcription factors to enhancers[123] (also see above) and enhancer function can be disrupted by point mutations in DNA binding motifs on enhancers. For example, a restless legs syndrome-associated SNP was demonstrated to reduce activity of the enhancers associated with the *MEIS1* locus. This SNP caused reduced binding of transcription factor CREB1 to the enhancer[124].

Besides reducing binding of transcription factors to enhancers, mutations in enhancer can also lead to increased binding of transcription factors. A nucleotide duplication inside an E-Box motif was suggested to create a novel binding site for Lef1, underlying cleft lip and palate in a small Brazilian family[125].

Variants associated with cancer risk have also been identified in enhancers regulating oncogenes. A study characterizing enhancers in colon cancer genome suggested that changes in the enhancers in colon cancer cells drives a specific transcriptional program that promotes colon carcinogenesis and found thousands of enhancer loci variants enriched for cancer genetic risk variants[126]. Individual variants in cancer have also been characterized. For example, a SNP located in the enhancer of *TOX3* is associated with breast cancer. This SNP causes increased binding of pioneer factor FOXA1 to the enhancer, resulting in the recruitment of the TLE1 repressor, diminishing enhancer activity and thereby decreasing expression of the *TOX3* gene[127]. In another locus a SNP associated with prostate cancer risk was also found to increase FOXA1 binding and to increase androgen responsiveness leading to increased enhancer activity[128]. Another cancer variant found is a colorectal cancer risk SNP in a TCF7L2 binding site on an enhancer required for tumorigenesis[129]. This SNP results in increased binding of the TCF7L2 transcription factor and increased luciferase activity of the enhancer interacting with the promoter of the proto-oncogene *MYC*[130]. Similar results with the same SNP were demonstrated in prostate cancer cells[131]. Also in colorectal cancer, a SNP in another enhancer of *MYC* impeded binding of the β-catenin transcription factor, resulting in increased expression of MYC[132].

In summary, variation and mutations in enhancers have been reported in relation to

disease. Point mutations in DNA binding motifs can result in increase or decrease of transcription factor binding or enhancers can be deleted entirely, impeding enhancer function and causing misregulation of genes leading to disease. Enrichment of disease-associated SNPs in enhancers[2] suggests that enhancers are often the underlying cause of human disease.

Mutations in DNA binding domains of transcription factors

Mutations in DNA binding domains may influence binding of transcription factors to enhancers, impeding enhancer function. For example, two mutations in the DNA binding homeo-domain of LHX4 were found in patients with pituitary hormone deficiency[133]. Both mutations prevent LHX4 from binding enhancers and lead to loss of gene activation[119,133]. Several deletions and missense mutations in the DNA binding domain of the CBFA1 transcription factor abolished DNA binding and are associated with cleidocranial dysplasia which is characterized by skelatal abnormalities[134,135]. Mutations in the DNA binding homeo-domain of repressing transcription factor ARX are associated with various syndromes all characterized by intellectual disability. These mutations lead to loss of enhancer binding by ARX and loss of gene repression[136].

Mutations in the DNA binding domain of transcription factors may not necessarily lead to complete loss of binding. For example, Klf1 is a hematopoietic C2H2 zinc finger transcription factor that plays a role in erythroid gene expression . The Nan-mutant mouse contains an amino acid substitution E339D in the second zinc finger of Klf1. While the wild type KLF1 binds to GC-box motif containing element GGGG[C/T]GGGG , the Nan-KLF1 mutant only binds GGGGCGGGG. This results in misregulation of Klf1 bound genes with GGGGTGGGG elements and severe hemolytic anemia in *Nan* heterozygous mutant mice[137].

Gain of genomic binding was also suggested to be a disease causing mechanism associated with mutations in DNA binding domains[119,120], but is supposedly rare[120]. In summary, mutations in the DNA binding domains of transcription factors can cause disease by abrogating genomic binding entirely or by causing binding to a different (sub) set of genomic targets (see also[119]). This may in turn lead to ectopic enhancer function resulting in misregulation of genes.

Mutations that result in disruption of the genomic environment of the enhancer

Mutations affecting the genomic environment surrounding enhancers can disrupt the activation of enhancer targets. Enhancers can be targeted to the wrong promoters when their TADs are disrupted. For example, families with rare limb malformations were shown to contain genomic rearrangements. These genomic rearrangements misplaced a cluster of enhancers relative to an insulator/boundary element of a TAD. This resulted in inappropriate interactions with promoters otherwise insulated from interaction with these enhancers, resulting in expression of genes involved in limb formation causing limb malformation[138]. Misplacement of enhancers has also been described to cause cancer. A frequently observed chromosomal rearrangement associated with acute myeloid leukemia causes a distal *GATA2* enhancer to ectopically activate the proto-oncogene *EVI1*[139].

Mutations that introduce a novel regulatory element can disrupt enhancer regulation. A SNP between α-globin genes and their upstream regulatory elements created a promoter-like element, which interfered with the normal activation of all downstream α-globin genes, causing the genetic disease α-thalassemia[140].

Mutations affecting CTCF and Cohesin might also influence the enhancer landscape and its interaction. Mutations in several Cohesin subunits cause Cornelia de Lange syndrome and this disease has been suggested to be the direct result of enhancer-mediated processes[141]. Although disruption of individual CTCF binding sites in relation to disease remains to be reported, it is tempting to speculate that displacement of CTCF by a mutation in its binding site might locally lead to inappropriate gene activation by enhancers and disease. Inversion of CTCF binding sites was shown to influence the specificity of promoter-enhancer interaction[56] (see also above) and might one day also be discovered to underlie disease.

REFERENCES

1. Shen, Y. et al. A map of the cis-regulatory sequences in the mouse genome. Nature 488, 116-120 (2012).
2. Consortium, E.P. An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57-74 (2012).
3. Thurman, R.E. et al. The accessible chromatin landscape of the human genome. Nature 489, 75-82 (2012).
4. Pennacchio, L.A. et al. In vivo enhancer analysis of human conserved non-coding sequences. Nature 444, 499-502 (2006).
5. Heintzman, N.D. et al. Histone modifications at human enhancers reflect global cell-type-specific gene expression. Nature 459, 108-12 (2009).
6. Whyte, W.A. et al. Master transcription factors and mediator establish super-enhancers at key cell identity genes. Cell 153, 307-19 (2013).
7. Banerji, J., Rusconi, S. & Schaffner, W. Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. Cell 27, 299-308 (1981).
8. Deng, W. et al. Controlling long-range genomic interactions at a native locus by targeted tethering of a looping factor. Cell 149, 1233-44 (2012).
9. Chen, X. et al. Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. Cell 133, 1106-17 (2008).
10. Allen, B.L. & Taatjes, D.J. The Mediator complex: a central integrator of transcription. Nat Rev Mol Cell Biol 16, 155-66 (2015).
11. Maurano, M.T. et al. Systematic localization of common disease-associated variation in regulatory DNA. Science 337, 1190-5 (2012).
12. Parker, S.C. et al. Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants. Proc Natl Acad Sci U S A 110, 17921-6 (2013).
13. Suganuma, T. & Workman, J.L. Signals and combinatorial functions of histone modifications. Annu Rev Biochem 80, 473-99 (2011).
14. Maston, G.A., Landt, S.G., Snyder, M. & Green, M.R. Characterization of enhancer function from genome-wide analyses. Annu Rev Genomics Hum Genet 13, 29-57 (2012).
15. Ernst, J. et al. Mapping and analysis of chromatin state dynamics in nine human cell types. Nature 473, 43-9 (2011).
16. Gilfillan, G.D. et al. Limitations and possibilities of low cell number ChIP-seq. BMC Genomics 13, 645 (2012).
17. Adli, M., Zhu, J. & Bernstein, B.E. Genome-wide chromatin maps derived from limited numbers of hematopoietic progenitors. Nat Methods 7, 615-8 (2010).
18. Shankaranarayanan, P. et al. Single-tube linear DNA amplification (LinDA) for robust ChIP-seq. Nat Methods 8, 565-7 (2011).

19. Shen, J. et al. H3K4me3 epigenomic landscape derived from ChIP-Seq of 1,000 mouse early embryonic cells. Cell Res 25, 143-7 (2015).
20. Ghisletti, S. et al. Identification and characterization of enhancers controlling the inflammatory gene expression program in macrophages. Immunity 32, 317-28 (2010).
21. He, A., Kong, S.W., Ma, Q. & Pu, W.T. Co-occupancy by multiple cardiac transcription factors identifies transcriptional enhancers active in heart. Proc Natl Acad Sci U S A 108, 5632-7 (2011).
22. Boyle, A.P. et al. High-resolution mapping and characterization of open chromatin across the genome. Cell 132, 311-22 (2008).
23. Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y. & Greenleaf, W.J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat Methods 10, 1213-8 (2013).
24. Gaulton, K.J. et al. A map of open chromatin in human pancreatic islets. Nat Genet 42, 255-9 (2010).
25. Core, L.J., Waterfall, J.J. & Lis, J.T. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science 322, 1845-8 (2008).
26. Wang, D. et al. Reprogramming transcription by distinct classes of enhancers functionally defined by eRNA. Nature 474, 390-4 (2011).
27. Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. Capturing chromosome conformation. Science 295, 1306-11 (2002).
28. Jin, F. et al. A high-resolution map of the three-dimensional chromatin interactome in human cells. Nature 503, 290-4 (2013).
29. Hughes, J.R. et al. Analysis of hundreds of cis-regulatory landscapes at high resolution in a single, high-throughput experiment. Nat Genet 46, 205-12 (2014).
30. Kolovos, P. et al. Targeted Chromatin Capture (T2C): a novel high resolution high throughput method to detect genomic interactions and regulatory elements. Epigenetics Chromatin 7, 10 (2014).
31. Fullwood, M.J. et al. An oestrogen-receptor-alpha-bound human chromatin interactome. Nature 462, 58-64 (2009).
32. Arnold, C.D. et al. Genome-wide quantitative enhancer activity maps identified by STARR-seq. Science 339, 1074-7 (2013).
33. Heinz, S., Romanoski, C.E., Benner, C. & Glass, C.K. The selection and function of cell type-specific enhancers. Nat Rev Mol Cell Biol 16, 144-54 (2015).
34. Heintzman, N.D. et al. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nat Genet 39, 311-8 (2007).
35. Creyghton, M.P. et al. Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proc Natl Acad Sci U S A 107, 21931-6 (2010).
36. Bonn, S. et al. Tissue-specific analysis of chromatin state identifies temporal signatures of enhancer activity during embryonic development. Nat Genet 44, 148-56 (2012).
37. Rada-Iglesias, A. et al. A unique chromatin signature uncovers early developmental enhancers in humans. Nature 470, 279-83 (2011).
38. Soufi, A., Donahue, G. & Zaret, K.S. Facilitators and impediments of the pluripotency reprogramming factors' initial engagement with the genome. Cell 151, 994-1004 (2012).
39. Iwafuchi-Doi, M. & Zaret, K.S. Pioneer transcription factors in cell reprogramming. Genes Dev 28, 2679-92 (2014).
40. Heinz, S. et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell 38, 576-89 (2010).
41. Li, Z. et al. Foxa2 and H2A.Z mediate nucleosome depletion during embryonic stem cell differentiation. Cell 151, 1608-16 (2012).
42. Wang, Z. et al. Combinatorial patterns of histone acetylations and methylations in the human genome. Nat Genet 40, 897-903 (2008).
43. Calo, E. & Wysocka, J. Modification of enhancer chromatin: what, how, and why? Mol Cell 49, 825-37 (2013).
44. Ostuni, R. et al. Latent enhancers activated by stimulation in differentiated cells. Cell 152, 157-71 (2013).
45. Liber, D. et al. Epigenetic priming of a pre-B cell-specific enhancer through binding of Sox2 and Foxd3 at the ESC stage. Cell Stem Cell 7, 114-26 (2010).
46. Xu, J. et al. Transcriptional competence and the active marking of tissue-specific enhancers by defined transcription factors in embryonic and induced pluripotent stem cells. Genes Dev 23, 2824-38 (2009).

47. Bergsland, M. et al. Sequentially acting Sox transcription factors in neural lineage development. Genes Dev 25, 2453-64 (2011).
48. Taatjes, D.J., Naar, A.M., Andel, F., 3rd, Nogales, E. & Tjian, R. Structure, function, and activator-induced conformations of the CRSP coactivator. Science 295, 1058-62 (2002).
49. Bernecky, C. & Taatjes, D.J. Activator-mediator binding stabilizes RNA polymerase II orientation within the human mediator-RNA polymerase II-TFIIF assembly. J Mol Biol 417, 387-94 (2012).
50. Meyer, K.D., Lin, S.C., Bernecky, C., Gao, Y. & Taatjes, D.J. p53 activates transcription by directing structural shifts in Mediator. Nat Struct Mol Biol 17, 753-60 (2010).
51. Kagey, M.H. et al. Mediator and cohesin connect gene expression and chromatin architecture. Nature 467, 430-5 (2010).
52. Gruber, S., Haering, C.H. & Nasmyth, K. Chromosomal cohesin forms a ring. Cell 112, 765-77 (2003).
53. Guo, Y. et al. CTCF/cohesin-mediated DNA looping is required for protocadherin alpha promoter choice. Proc Natl Acad Sci U S A 109, 21081-6 (2012).
54. Monahan, K. et al. Role of CCCTC binding factor (CTCF) and cohesin in the generation of single-cell diversity of protocadherin-alpha gene expression. Proc Natl Acad Sci U S A 109, 9125-30 (2012).
55. Lobanenkov, V.V. et al. A novel sequence-specific DNA binding protein which interacts with three regularly spaced direct repeats of the CCCTC-motif in the 5'-flanking sequence of the chicken c-myc gene. Oncogene 5, 1743-53 (1990).
56. Guo, Y. et al. CRISPR Inversion of CTCF Sites Alters Genome Topology and Enhancer/Promoter Function. Cell 162, 900-10 (2015).
57. Palstra, R.J. et al. Maintenance of long-range DNA interactions after inhibition of ongoing RNA polymerase II transcription. PLoS One 3, e1661 (2008).
58. Kasper, L.H. & Brindle, P.K. Mammalian gene expression program resiliency: the roles of multiple coactivator mechanisms in hypoxia-responsive transcription. Cell Cycle 5, 142-6 (2006).
59. Hospital, S.J.C.s.R. CBP-P300 interactome 10-23-09.xls. (ed. https://www.stjude.org/SJFile/CBP-p300%20interactome%2010-23-09.xls) (2009).
60. Goodman, R.H. & Smolik, S. CBP/p300 in cell growth, transformation, and development. Genes Dev 14, 1553-77 (2000).
61. Holmqvist, P.H. & Mannervik, M. Genomic occupancy of the transcriptional co-activators p300 and CBP. Transcription 4, 18-23 (2013).
62. Visel, A. et al. ChIP-seq accurately predicts tissue-specific activity of enhancers. Nature 457, 854-8 (2009).
63. Kanno, T. et al. BRD4 assists elongation of both coding and enhancer RNAs by interacting with acetylated histones. Nat Struct Mol Biol 21, 1047-57 (2014).
64. Babu, M.M., Luscombe, N.M., Aravind, L., Gerstein, M. & Teichmann, S.A. Structure and evolution of transcriptional regulatory networks. Curr Opin Struct Biol 14, 283-91 (2004).
65. Vaquerizas, J.M., Kummerfeld, S.K., Teichmann, S.A. & Luscombe, N.M. A census of human transcription factors: function, expression and evolution. Nat Rev Genet 10, 252-63 (2009).
66. Ravasi, T. et al. An atlas of combinatorial transcriptional regulation in mouse and man. Cell 140, 744-52 (2010).
67. Jolma, A. et al. DNA-dependent formation of transcription factor pairs alters their binding specificity. Nature 527, 384-8 (2015).
68. Remenyi, A. et al. Crystal structure of a POU/HMG/DNA ternary complex suggests differential assembly of Oct4 and Sox2 on two enhancers. Genes Dev 17, 2048-59 (2003).
69. Niwa, H. How is pluripotency determined and maintained? Development 134, 635-46 (2007).
70. Aksoy, I. et al. Oct4 switches partnering from Sox2 to Sox17 to reinterpret the enhancer code and specify endoderm. EMBO J 32, 938-53 (2013).
71. Yang, S.H. et al. Otx2 and Oct4 drive early enhancer activation during embryonic stem cell transition from naive pluripotency. Cell Rep 7, 1968-81 (2014).
72. Gagliardi, A. et al. A direct physical interaction between Nanog and Sox2 regulates embryonic stem cell self-renewal. EMBO J 32, 2231-47 (2013).
73. Kamachi, Y., Uchikawa, M., Tanouchi, A., Sekido, R. & Kondoh, H. Pax6 and SOX2 form a co-DNA-binding partner complex that regulates initiation of lens development. Genes Dev 15, 1272-86 (2001).
74. Ahmed, M. et al. Eya1-Six1 interaction is sufficient to induce hair cell fate in the cochlea by activating Atoh1 expression in cooperation with Sox2. Dev Cell 22, 377-90 (2012).
75. Kondoh, H. & Kamachi, Y. SOX-partner code for cell specification: Regulatory target selection and

underlying molecular mechanisms. Int J Biochem Cell Biol 42, 391-9 (2010).
76. Sharov, A.A. et al. Identification of Pou5f1, Sox2, and Nanog downstream target genes with statistical confidence by applying a novel algorithm to time course microarray and genome-wide chromatin immunoprecipitation data. BMC Genomics 9, 269 (2008).
77. van den Berg, D.L. et al. An Oct4-centered protein interaction network in embryonic stem cells. Cell Stem Cell 6, 369-81 (2010).
78. Esch, D. et al. A unique Oct4 interface is crucial for reprogramming to pluripotency. Nat Cell Biol 15, 295-301 (2013).
79. Boyer, L.A. et al. Core transcriptional regulatory circuitry in human embryonic stem cells. Cell 122, 947-56 (2005).
80. Liang, J. et al. Nanog and Oct4 associate with unique transcriptional repression complexes in embryonic stem cells. Nat Cell Biol 10, 731-9 (2008).
81. Yuan, P. et al. Eset partners with Oct4 to restrict extraembryonic trophoblast lineage potential in embryonic stem cells. Genes Dev 23, 2507-20 (2009).
82. Lohmann, F. et al. KMT1E mediated H3K9 methylation is required for the maintenance of embryonic stem cells by repressing trophectoderm differentiation. Stem Cells 28, 201-12 (2010).
83. Yeap, L.S., Hayashi, K. & Surani, M.A. ERG-associated protein with SET domain (ESET)-Oct4 interaction regulates pluripotency and represses the trophectoderm lineage. Epigenetics Chromatin 2, 12 (2009).
84. Ahituv, N. et al. Deletion of ultraconserved elements yields viable mice. PLoS Biol 5, e234 (2007).
85. Barolo, S. Shadow enhancers: frequently asked questions about distributed cis-regulatory information and enhancer redundancy. Bioessays 34, 135-41 (2012).
86. Jongens, T.A., Fowler, T., Shermoen, A.W. & Beckendorf, S.K. Functional redundancy in the tissue-specific enhancer of the Drosophila Sgs-4 gene. EMBO J 7, 2559-67 (1988).
87. Frankel, N. et al. Phenotypic robustness conferred by apparently redundant transcriptional enhancers. Nature 466, 490-3 (2010).
88. Perry, M.W., Boettiger, A.N. & Levine, M. Multiple enhancers ensure precision of gap gene-expression patterns in the Drosophila embryo. Proc Natl Acad Sci U S A 108, 13570-5 (2011).
89. Kim, T.K. et al. Widespread transcription at neuronal activity-regulated enhancers. Nature 465, 182-7 (2010).
90. Melo, C.A. et al. eRNAs are required for p53-dependent enhancer activity and gene transcription. Mol Cell 49, 524-35 (2013).
91. Li, W. et al. Functional roles of enhancer RNAs for oestrogen-dependent transcriptional activation. Nature 498, 516-20 (2013).
92. Dixon, J.R. et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature 485, 376-80 (2012).
93. Nora, E.P. et al. Spatial partitioning of the regulatory landscape of the X-inactivation centre. Nature 485, 381-5 (2012).
94. Bell, A.C., West, A.G. & Felsenfeld, G. The protein CTCF is required for the enhancer blocking activity of vertebrate insulators. Cell 98, 387-96 (1999).
95. Zuin, J. et al. Cohesin and CTCF differentially affect chromatin architecture and gene expression in human cells. Proc Natl Acad Sci U S A 111, 996-1001 (2014).
96. Lettice, L.A. et al. A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. Hum Mol Genet 12, 1725-35 (2003).
97. Kieffer-Kwon, K.R. et al. Interactome maps of mouse gene regulatory domains reveal basic principles of transcriptional regulation. Cell 155, 1507-20 (2013).
98. Sanyal, A., Lajoie, B.R., Jain, G. & Dekker, J. The long-range interaction landscape of gene promoters. Nature 489, 109-13 (2012).
99. Li, G. et al. Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. Cell 148, 84-98 (2012).
100. de Laat, W. & Duboule, D. Topology of mammalian developmental enhancers and their regulatory landscapes. Nature 502, 499-506 (2013).
101. Patrinos, G.P. et al. Multiple interactions between regulatory regions are required to stabilize an active chromatin hub. Genes Dev 18, 1495-509 (2004).
102. Bender, M.A. et al. Targeted deletion of 5'HS1 and 5'HS4 of the beta-globin locus control region reveals additive activity of the DNaseI hypersensitive sites. Blood 98, 2022-7 (2001).
103. Villar, D. et al. Enhancer evolution across 20 mammalian species. Cell 160, 554-66 (2015).

104. Jones, F.C. et al. The genomic basis of adaptive evolution in threespine sticklebacks. Nature 484, 55-61 (2012).
105. Prescott, S.L. et al. Enhancer Divergence and cis-Regulatory Evolution in the Human and Chimp Neural Crest. Cell 163, 68-83 (2015).
106. Arnold, C.D. et al. Quantitative genome-wide enhancer activity maps for five Drosophila species show functional enhancer conservation and turnover during cis-regulatory evolution. Nat Genet 46, 685-92 (2014).
107. Nord, A.S. et al. Rapid and pervasive changes in genome-wide enhancer usage during mammalian development. Cell 155, 1521-31 (2013).
108. Stergachis, A.B. et al. Developmental fate and cellular maturity encoded in human regulatory DNA landscapes. Cell 154, 888-903 (2013).
109. Cotney, J. et al. The evolution of lineage-specific regulatory activities in the human embryonic limb. Cell 154, 185-96 (2013).
110. Wunderlich, Z. et al. Kruppel Expression Levels Are Maintained through Compensatory Evolution of Shadow Enhancers. Cell Rep 12, 1740-7 (2015).
111. Tishkoff, S.A. et al. Convergent adaptation of human lactase persistence in Africa and Europe. Nat Genet 39, 31-40 (2007).
112. Jones, B.L. et al. Diversity of lactase persistence alleles in Ethiopia: signature of a soft selective sweep. Am J Hum Genet 93, 538-44 (2013).
113. Troelsen, J.T., Olsen, J., Moller, J. & Sjostrom, H. An upstream polymorphism associated with lactase persistence has increased enhancer activity. Gastroenterology 125, 1686-94 (2003).
114. Enattah, N.S. et al. Identification of a variant associated with adult-type hypolactasia. Nat Genet 30, 233-7 (2002).
115. Lewinsky, R.H. et al. T-13910 DNA variant associated with lactase persistence interacts with Oct-1 and stimulates lactase promoter activity in vitro. Hum Mol Genet 14, 3945-53 (2005).
116. Loven, J. et al. Selective inhibition of tumor oncogenes by disruption of super-enhancers. Cell 153, 320-34 (2013).
117. Hnisz, D. et al. Super-enhancers in the control of cell identity and disease. Cell 155, 934-47 (2013).
118. Pott, S. & Lieb, J.D. What are super-enhancers? Nat Genet 47, 8-12 (2015).
119. Fuxman Bass, J.I. et al. Human gene-centered transcription factor networks for enhancers and disease variants. Cell 161, 661-73 (2015).
120. Sahni, N. et al. Widespread macromolecular interaction perturbations in human genetic disorders. Cell 161, 647-60 (2015).
121. Loots, G.G. et al. Genomic deletion of a long-range bone enhancer misregulates sclerostin in Van Buchem disease. Genome Res 15, 928-35 (2005).
122. Ghiasvand, N.M. et al. Deletion of a remote enhancer near ATOH7 disrupts retinal neurogenesis, causing NCRNA disease. Nat Neurosci 14, 578-86 (2011).
123. Bryne, J.C. et al. JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. Nucleic Acids Res 36, D102-6 (2008).
124. Spieler, D. et al. Restless legs syndrome-associated intronic common variant in Meis1 alters enhancer function in the developing telencephalon. Genome Res 24, 592-603 (2014).
125. Fakhouri, W.D. et al. An etiologic regulatory mutation in IRF6 with loss- and gain-of-function effects. Hum Mol Genet 23, 2711-20 (2014).
126. Akhtar-Zaidi, B. et al. Epigenomic enhancer profiling defines a signature of colon cancer. Science 336, 736-9 (2012).
127. Cowper-Sal lari, R. et al. Breast cancer risk-associated SNPs modulate the affinity of chromatin for FOXA1 and alter gene expression. Nat Genet 44, 1191-8 (2012).
128. Jia, L. et al. Functional enhancers at the gene-poor 8q24 cancer-linked locus. PLoS Genet 5, e1000597 (2009).
129. Sur, I.K. et al. Mice lacking a Myc enhancer that includes human SNP rs6983267 are resistant to intestinal tumors. Science 338, 1360-3 (2012).
130. Pomerantz, M.M. et al. The 8q24 cancer risk variant rs6983267 shows long-range interaction with MYC in colorectal cancer. Nat Genet 41, 882-4 (2009).
131. Wasserman, N.F., Aneas, I. & Nobrega, M.A. An 8q24 gene desert variant associated with prostate cancer risk confers differential in vivo activity to a MYC enhancer. Genome Res 20, 1191-7 (2010).
132. Wright, J.B., Brown, S.J. & Cole, M.D. Upregulation of c-MYC in cis through a large chromatin loop linked to a cancer risk-associated single-nucleotide polymorphism in colorectal cancer cells. Mol Cell

Biol 30, 1411-20 (2010).
133. Pfaeffle, R.W. et al. Three novel missense mutations within the LHX4 gene are associated with variable pituitary hormone deficiencies. J Clin Endocrinol Metab 93, 1062-71 (2008).
134. Lee, B. et al. Missense mutations abolishing DNA binding of the osteoblast-specific transcription factor OSF2/CBFA1 in cleidocranial dysplasia. Nat Genet 16, 307-10 (1997).
135. Mundlos, S. et al. Mutations involving the transcription factor CBFA1 cause cleidocranial dysplasia. Cell 89, 773-9 (1997).
136. Shoubridge, C., Tan, M.H., Seiboth, G. & Gecz, J. ARX homeodomain mutations abolish DNA binding and lead to a loss of transcriptional repression. Hum Mol Genet 21, 1639-47 (2012).
137. Siatecka, M. et al. Severe anemia in the Nan mutant mouse caused by sequence-selective disruption of erythroid Kruppel-like factor. Proc Natl Acad Sci U S A 107, 15151-6 (2010).
138. Lupianez, D.G. et al. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. Cell 161, 1012-25 (2015).
139. Groschel, S. et al. A single oncogenic enhancer rearrangement causes concomitant EVI1 and GATA2 deregulation in leukemia. Cell 157, 369-81 (2014).
140. De Gobbi, M. et al. A regulatory SNP causes a human genetic disease by creating a new transcriptional promoter. Science 312, 1215-7 (2006).
141. Herz, H.M., Hu, D. & Shilatifard, A. Enhancer malfunction in cancer. Mol Cell 53, 859-66 (2014).

Chapter2

Proteins that bind regulatory regions identified by histone modification chromatin immunoprecipitations and mass spectrometry

Erik Engelen[1]*, **Johannes H. Brandsma**[1]*, Maaike J. Moen[1], Luca Signorile[1], Dick H. W. Dekkers[2], Jeroen Demmers[2], Christel E.M. Kockx[3], Zehila Ozgür[3], Wilfred F. J. van IJcken[3], Debbie L.C. van den Berg[1,4], Raymond A. Poot[1]

[1] Department of Cell Biology, Erasmus MC, Wytemaweg 80, 3015 CN Rotterdam, The Netherlands
[2] Proteomics Center, Erasmus MC, Wytemaweg 80, 3015 CN, Rotterdam, The Netherlands
[3] Center for Biomics, Erasmus MC, Wytemaweg 80, 3015 CN, Rotterdam, The Netherlands
[4] The Francis Crick Institute, Mill Hill Laboratory, The Ridgeway, London NW7 1AA, United Kingdom

*** These authors contributed equally to this work**

Nature Communications 6, 7155 (2015)

ABSTRACT

The locations of transcriptional enhancers and promoters were recently mapped in many mammalian cell types. Proteins that bind those regulatory regions can determine cell identity but have not been systematically identified. Here we purify native enhancers, promoters or heterochromatin from embryonic stem cells by chromatin immunoprecipitations (ChIP) for characteristic histone modifications and identify associated proteins using mass spectrometry (MS). 239 factors are identified and predicted to bind enhancers or promoters with different levels of activity, or heterochromatin. Published genome-wide data indicate a high accuracy of location prediction by ChIP-MS. A quarter of the identified factors are important for pluripotency and includes Oct4, Esrrb, Klf5, Mycn and Dppa2, factors that drive reprogramming to pluripotent stem cells. We determined the genome-wide binding sites of Dppa2 and find that Dppa2 operates outside the classical pluripotency network. Our ChIP-MS method provides a detailed read-out of the transcriptional landscape representative of the investigated cell type.

INTRODUCTION

A mammalian genome supports the generation of the hundreds of different cell types in an organism. These cell types display distinct gene expression profiles as a direct consequence of differences in the activation state of their gene promoters and distal cis-regulatory elements called transcriptional enhancers. The ENCODE project has generated a wealth of data on the genome-wide chromatin landscape of many different mouse and human cell types[1,2]. In particular, the genome-wide identification of regulatory regions such as transcriptional enhancers and promoters and their state of activity has the potential to increase our understanding of how cell type identity is acquired and maintained. From reprogramming experiments it has become increasingly clear that the identity of cells is to a large extent determined by transcription factors, which bind enhancers and promoters[3-6]. It is therefore of interest to purify native transcriptional enhancers and promoters of a given cell type and identify the proteins that bind to these regulatory regions. Here we performed chromatin immunoprecipitations for histone modifications associated with promoters, enhancers or heterochromatin in mouse embryonic stem cells (ESCs) and identified the proteins present in the different precipitated fractions by mass spectrometry, a method that we named ChIP-MS.

Our ChIP-MS experiments identified 239 factors that we could predict to bind to promoters, enhancers or heterochromatin. Among these factors are subunits of several chromatin modifying complexes and proteins that play a role in different aspects of transcriptional regulation. We also find key ESC transcription factors such as Oct4, Esrrb, Dppa2 and Klf5 that are not only important for maintaining ESC self-renewal but also facilitate the reprogramming of somatic cells to induced pluripotent stem cells (iPSCs)[4,7-9]. Genome-wide data sets were available for 28 ChIP-MS-identified factors and correlated well with the ChIP-MS-based predictions for these factors, suggesting a high level of accuracy of location prediction by ChIP-MS.

For many of the detected factors, the genome-wide localization has not yet been

determined and our ChIP-MS results provide the first evidence of their binding preference for a particular type of regulatory DNA. To illustrate that ChIP-MS can identify factors with an interesting genome-wide location, we determined the genome-wide binding sites of pluripotency marker and reprogramming factor Dppa2. We show that Dppa2 is not part of the classical pluripotency transcriptional network and that Dppa2 target genes reach full activation only later in development.

RESULTS

ChIP-MS rationale and procedure.

Transcriptional enhancers and promoters can be recognized by the chemical modifications of their associated histones, especially histone H3. Promoters of transcribed genes were found to contain histone H3 tri-methylated at lysine 4 (H3K4me3)[10,11] and the level of their activity correlates with the level of H3K27 acetylation (H3K27ac) present[12,13]. Enhancers contain histone H3 mono-methylated at lysine 4 (H3K4me1)[14] and active enhancers can be recognized by the presence of the H3K27ac mark[15,16]. Inactive (hetero)chromatin is marked by H3K9me3[17]. The presence or absence of these and other chromatin marks was used to postulate fifteen different chromatin regions in the mammalian genome, including promoters and enhancers with different levels of activity[13].

We anticipated that ChIP for H3K4me3 would precipitate active promoters and ChIP for H3K4me1 would precipitate enhancers. ChIP for H3K27ac would preferentially precipitate the most active promoters and enhancers, whereas ChIP for H3K9me3 would precipitate heterochromatin. Accordingly, we performed large scale ChIPs in biological duplicate for H3K4me3, H3K4me1, H3K27ac or H3K9me3, and for GFP as a control, in mouse embryonic stem cells (ESCs, Fig. 1a,b). Crosslinking of the chromatin was performed with Disuccinimidyl glutarate (DSG), a protein-protein crosslinker, followed by standard formaldehyde crosslinking, to increase the crosslinking efficiency of genome-bound factors to the chromatin[18-20]. ChIP wash steps were performed in low adherence tubes to increase protein yield and reduce background[20,21]. Bound protein factors were de-crosslinked and eluted by prolonged heating in protein denaturing conditions, separated on an SDS-polyacrylamide gel, tryptic peptides isolated and analyzed by mass spectrometry. A representative protein gel showed the (unresolved) histones precipitated with each histone modification antibody but not with the GFP control (Fig. 1c). Analysis by Western blot revealed that comparable amounts of chromatin were precipitated in the different histone-modification ChIPs, as indicated by the total content of histone H3 (Fig. 1d). ChIP against H3K4me1, H3K4me3 or H3K27ac precipitated chromatin with these respective histone modifications (Fig. 1d). Minor amounts of H3K4me1 were observed in the H3K4me3 ChIP and vice-versa. This is to be expected as H3K4me1 is present at low levels around active promoters[15]. H3K9me3 ChIP precipitated H3K9me3-marked chromatin but no significant amounts of the other histone modifications (Fig. 1d). We conclude that the histone modification ChIPs efficiently precipitated the intended chromatin fractions.

Subsequently, we tested whether our modified ChIP protocol, that we use for ChIP-MS, still precipitated the intended genomic regions, as compared to conventional ChIP. DNA

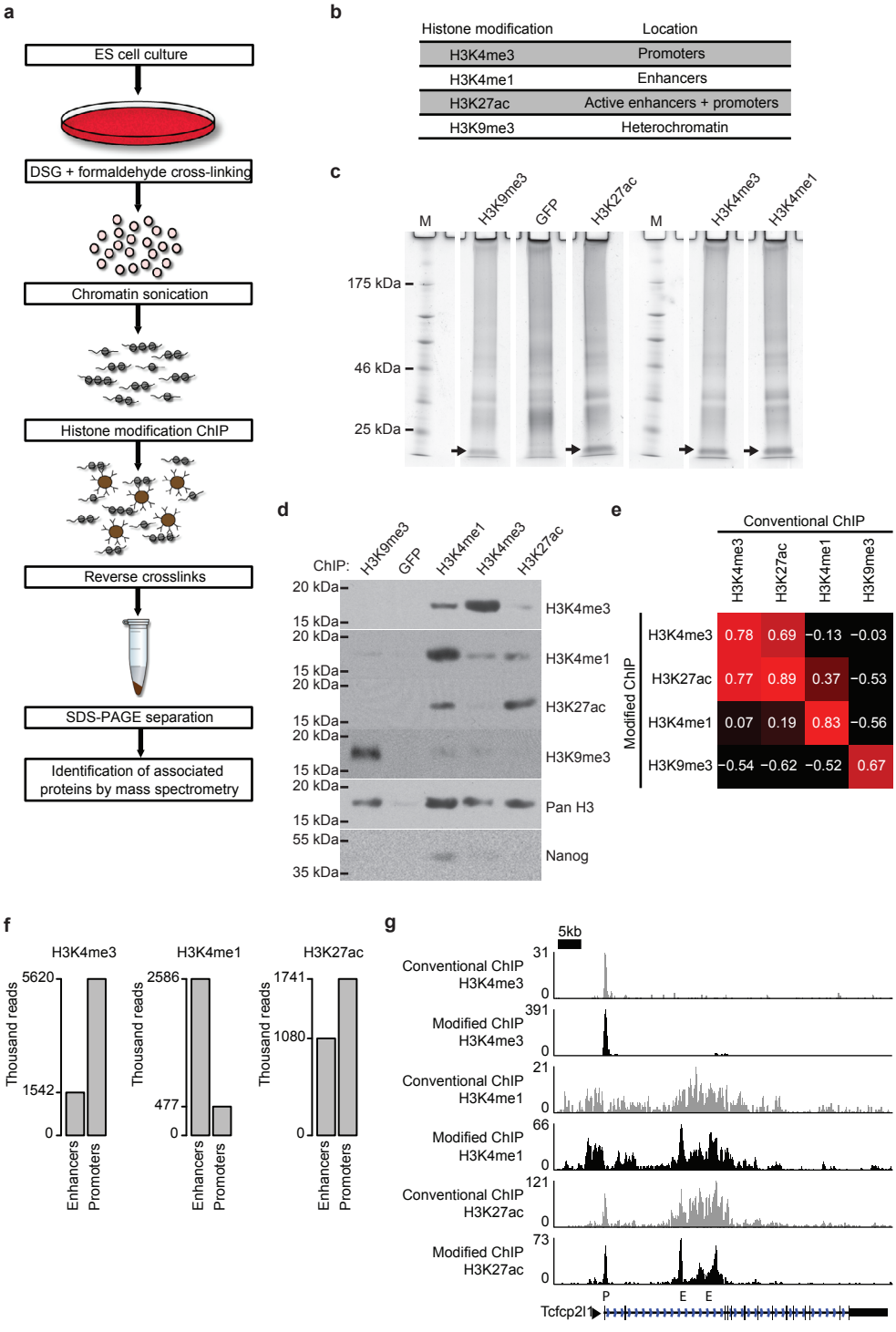


Figure1.(Legendatthebottomofthenextpage)

precipitated by modified ChIP for the different histone modifications was sequenced (ChIP-seq) and mapped to the genome. DNA precipitated by modified ChIP correlated well with the corresponding published conventional ChIP-seq for all four histone modifications (Fig. 1e). Modified H3K4me3 ChIP predominantly precipitated promoters and modified H3K4me1 ChIP precipitated predominantly enhancers, as intended (Fig. 1f). Examples of histone modification tracks around pluripotency genes *Tcfcp2l1* (Fig. 1g) and *Nanog* (Supplementary Fig. 1) show high similarity between our modified ChIP and conventional ChIP. We conclude that the inclusion of additional crosslinker DSG has not significantly altered the genomic regions precipitated by our ChIP protocol, as compared to conventional ChIP.

Prediction of genome localization of identified factors.

We analyzed the different precipitated chromatin fractions and GFP-control fractions by mass spectrometry for an unbiased identification of the protein factors present in each fraction. We identified 249 factors that have at least a 3-fold difference in Exponentially Modified Protein Abundance Index (emPAI) score, a measure for the amount of protein present[22], in the ChIPs for one histone modification compared to the ChIPs for one or more of the other histone modifications. Included factors should have no or very low presence (more than five-fold lower emPAI score) in any of the GFP control ChIPs (Supplementary Tables 1 and 2, and Methods). These two selection steps were included to exclude proteins that bind to chromatin indiscriminately of the tested histone modifications, or are background of the ChIP-MS procedure, respectively. Of the 249 factors, 10 factors were only present in the H3K27ac fraction, which does not discriminate between promoters and enhancers, leaving 239 factors for which we could predict their binding to promoters, enhancers or heterochromatin. We assigned to identified factors the locations “promoter”, “enhancer” and “heterochromatin” according to the fraction (H3K4me3, H3K4me1 and H3K9me3, respectively) in which they have the highest emPAI value (Supplementary Tables 1, 2 and 3, and Methods). This annotation is not absolute, as factors can be present in more than one location, but it does provide clarity and facilitates a more systematic validation with published genome-wide localization data (see below).

We also indicated the presence of a factor in the H3K27Ac fractions by calculating the ratio of its average emPAI value in the H3K27ac fractions over its H3K4me3 emPAI score or its H3K4me1 emPAI score, whichever one is the highest, a ratio that we call the H3K27ac ratio. Presenting the ChIP-MS association of a factor with the H3K27ac modification in this way compensates for the considerable differences in ChIP-MS detection levels for

Figure 1. Outline and initial validation of the ChIP-MS protocol. (a) Flowchart of the ChIP-MS protocol. **(b)** Histone modifications used in ChIP-MS and their predominant location on the genome. **(c)** Representative 10% polyacrylamide gel with proteins from ChIPs for the indicated histone modifications and the GFP control ChIP. Arrows indicate unresolved histones in the histone modification ChIPs, which are absent in the GFP control ChIP. Molecular weight markers are depicted by M. **(d)** Western blot analyses of the histone modification content, histone content and the presence of Nanog in the immunoprecipitated chromatin fractions. Different ChIPs are indicated at the top, antibodies used for the different Western blot analyses are on the right. **(e)** Correlation between DNA regions precipitated by modified ChIP and conventional ChIP for H3K4me3, H3K27ac, H3K4me1 or H3K9me3. **(f)** Overlap of DNA precipitated with modified ChIP for H3K4me3, H3K4me1 or H3K27ac with promoters and enhancers. Number of ChIP-seq reads overlapping with promoters or enhancers is indicated. **(g)** ChIP-seq tracks for modified ChIP or conventional ChIP for H3K4me3, H3K4me1 or H3K27ac around pluripotency gene Tcfcp2l1. Sequence reads were plotted relative to chromosomal position. Genome location of Tcfcp2l1 is shown, scale bar indicates 5 kb of genome. P indicates promoter, E indicates putative enhancer.

different proteins (Supplementary Table 1) and is therefore more informative than its H3K27ac emPAI value per se. Small emPAI values for H3K4me3 or H3K4me1 increase the uncertainty of the H3K27ac ratio value. We used the H3K27ac ratio as a predictor for the level of H3K27ac, and thereby the activity, of the promoters or enhancers bound by the factor (Supplementary Table 1).

For a visual representation (Fig. 2), identified factors and complexes were allocated according to their predicted binding to promoters, enhancers or heterochromatin. The calculated H3K27ac ratios were used to position predicted enhancer-binding factors on the upper horizontal axis reflecting the level of activity of bound enhancers or position predicted promoter binders on the right vertical axis reflecting the level of activity of bound promoters (Fig. 2).

An early indication that the predictions from our ChIP-MS experiments were valid came from the “promoter” prediction of all 5 identified RNApol2 subunits and 7 identified TFIID subunits (Supplementary Table 3). The H3K9 methyltransferase Suv3-9 binds pericentric heterochromatin[23] and was indeed observed solely in the H3K9me3 fraction (“heterochromatin” prediction, Fig. 2 and Supplementary Table 1). We identified a large number of subunits of established chromatin modifying complexes including the BAF complex, Sin3 complex and MLL complex (Fig. 2 and Supplementary Table 3). Strikingly, the localization prediction for different subunits within the same complex was nearly 100% identical (Supplementary Table 3), indicating a high level of consistency in the predictions.

Among the ChIP-MS identified factors with the highest H3K27ac ratio, predicting binding to highly active regulatory regions, were chromatin factors of the BET family; Brd2, 3 and 4 (Fig. 2 and Supplementary Table 1). BET family members were shown to bind to hyperacetylated chromatin[24]. Brd4 was recently identified as a key factor in the marking and functional maintenance of exceptionally large and active “super enhancers”, which regulate the expression of cell fate determining genes[25,26]. ChIP-MS classified Brd4 to bind predominantly to enhancers and Brd2 and 3 to bind promoters. Using ChIP-seq, Brd2 and Brd3 were indeed shown to bind promoters, whereas Brd4 was also present at enhancers[27]. The finding that from all ChIP-MS detected proteins, several BET family members are among the proteins with the highest H3K27ac ratio, validates the use of the H3K27ac ratio as an indicator of the level of activity of bound promoters or enhancers.

Many studies have identified factors that are important for maintaining ESC pluripotency[28-30] or factors that reprogram somatic cells towards ESC-like induced pluripotent cells (iPSCs)[4,6,7]. We found that more than a quarter of our ChIP-MS identified factors (63 out of 239) have a role in pluripotency acquisition or maintenance (Supplementary Table 1). ChIP-MS-identified factors included established reprogramming factors Oct4, Esrrb, Klf5 and Mycn (Fig. 2 and Supplementary Table 1), which as part of a 3-4 factor mix, reprogram somatic cells to iPSCs[4,7,8]. Our ChIP-MS data predicted that Oct4, Esrrb and Klf5 bind predominantly to enhancers and Mycn predominantly binds to promoters, in agreement with published genome localization data[3,31]. Nanog, another well-known pluripotency factor, is difficult to detect by mass spectrometry[21,32] and was indeed not identified by ChIP-MS. Western blot analysis of our ChIP-MS samples showed that

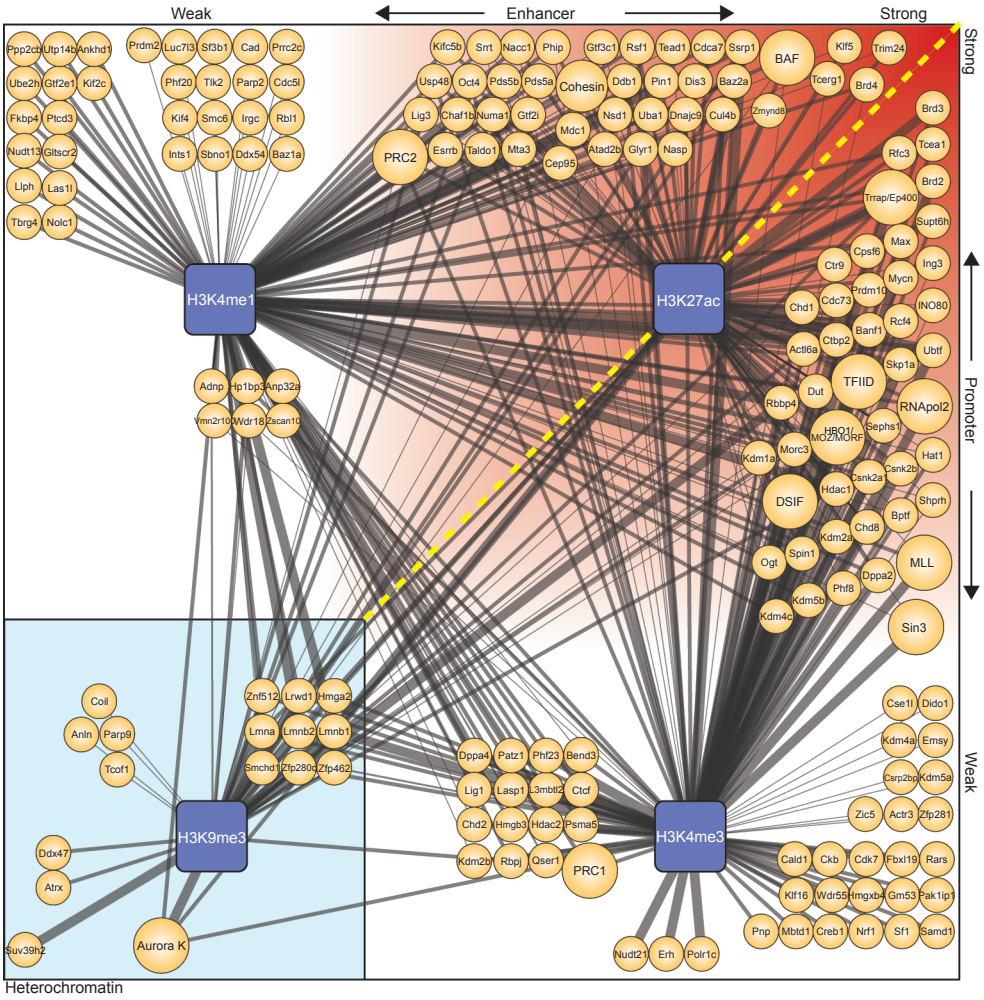


Figure 2. ChIP-MS predicted locations of identified factors and complexes. Visual representation of factors (small orange circles) and complexes (large orange circles) identified by ChIP-MS for four different histone modifications (blue squares). Thickness of the edges indicates average emPAI score of a factor or complex in histone modification ChIP. Factors and complexes are positioned according to their ChIP-MS location prediction. To the left of the yellow dashed line are predicted enhancer binders, positioned horizontally from weak activity enhancers (left) to strong activity enhancers (right) according to their H3K27ac ratio. To the right of the yellow dashed line are predicted promoter binders positioned vertically from weak activity promoters (bottom) to strong activity promoters (top) according to their H3K27ac ratio. In the left bottom square are factors and complexes predicted to bind heterochromatin.

Nanog was present in the H3K4me1 fraction (Fig. 1d), suggesting it binds to enhancers, in agreement with published data[3].

Estimation of ChIP-MS prediction accuracy.

Our list of 239 ChIP-MS assigned factors includes 28 factors for which the genome-wide binding sites have been determined in mouse ESCs by ChIP-seq (Fig. 3a), which provides an opportunity to probe the accuracy of our localization prediction. In a first

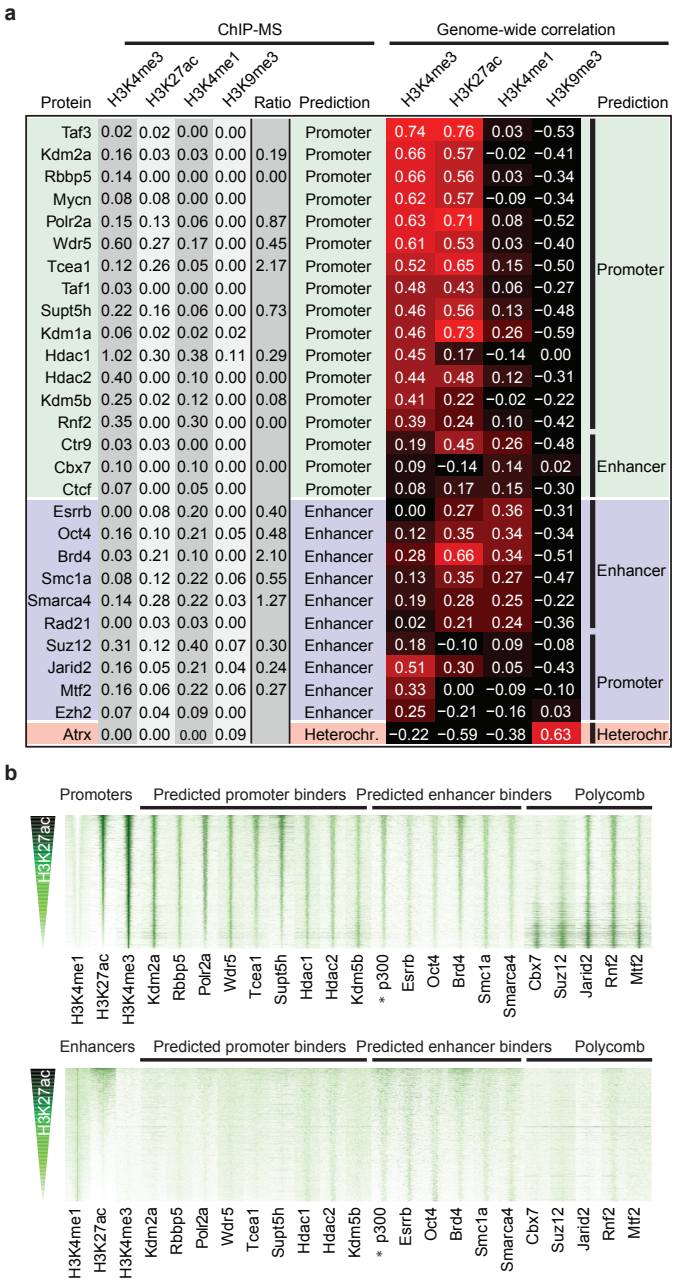


Figure 3. Validation of ChIP-MS predictions with published genome-wide location information. **(a)** Comparison of location prediction by ChIP-MS with location prediction by correlation of genome-wide binding sites with the indicated histone modifications on the genome. Protein factors for which genome-wide locations in mouse ESCs are determined by ChIP-seq are listed on the left, according to their ChIP-MS prediction as promoter binder (top, green panel), enhancer binder (middle, blue panel), or heterochromatin binder (bottom, red panel). Indicated in columns from left to right are: protein factor, its average emPAI values in the different histone modification ChIPs, its H3K27ac ratio (if highest

Legend continues on the bottom of the next page

analysis, we compared our ChIP-MS-based predictions (Fig. 3a) with location predictions derived from the correlation of factor binding sites with the different histone marks on the genome (Fig. 3a). Of the 17 factors predicted by ChIP-MS to be promoter-associated, 14 factors (82%) were indeed most associated on the genome with the active promoter mark H3K4me3 (Fig. 3a). In the case of Ctr9, Ctcf and Cbx7, the ChIP-MS prediction was not conform the location prediction by genome-wide correlation. However, in these cases both the ChIP-MS values and genome-wide correlations for H3K4me3 and H3K4me1 (which differentiates between "promoter" and "enhancer" prediction) were very similar (Fig. 3a). Moreover, the correlation with any of the four tested histone marks was low for Ctcf and Cbx7.

Of the 10 factors predicted by ChIP-MS to be predominantly associated with enhancers, 6 had indeed the highest association on the genome with enhancer mark H3K4me1 (Fig. 3a). These include Oct4 and Esrrb, two key pluripotency and reprogramming transcription factors, and Smarca4 (Brg1), the catalytic subunit of the SWI-SNF chromatin modifying complex (Fig. 3a). All the wrongly annotated factors (Suz12, Jarid2, Mtf2, Ezh2) are members of the Polycomb family of repressor proteins, which had "Enhancer" ChIP-MS predictions but were assigned "Promoter" by the genome-wide correlation of their binding sites (Fig. 3a). Polycomb factors are abundant in ESCs and often bind broadly at relatively inactive promoters with similarly low levels of H3K4me3 and H3K4me1 (see below)[33,34]. Both histone marks are indeed faithfully detected by ChIP-MS but in the above cases this leads to wrong predictions, albeit by small margins in the H3K4me3 and H3K4me1 ChIP-MS values (Fig. 3a). Atrx was correctly assigned by ChIP-MS to bind H3K9me3-containing heterochromatin (Fig. 3a). We conclude from this analysis that ChIP-MS predicted factor location to promoters, enhancers or heterochromatin with high accuracy, with the few false identifications in the expected grey areas.

Subsequently, we assessed whether the H3K27ac ratio of a factor correlates with the association of its binding sites with H3K27ac on the genome (Fig. 3a). From the factors with genome-wide location information (Fig. 3a), we took factors with a highest emPAI value of 0.1 or higher, to be well above the detection limit of our ChIP-MS experiments. We calculated for each of these factors their H3K27ac ratio (Fig. 3a). These ratios were compared to the correlation of genome-wide binding of these factors with H3K27ac-marked regions (Fig. 3a). Indeed, promoter-predicted factors Tcea1, Polr2a and Supt5h have the highest H3K27ac ratios and have relatively high correlations with H3K27ac on the genome (Fig. 3a). Promoter-predicted factors Hdac1 and 2, Kdm5b, Rnf2 and Cbx7 have low H3K27ac ratios and are factors with relatively low genome-wide associations

(Figure 3. Legend continues from previous page)
emPAI value ≥ 0.1), ChIP-MS location prediction, correlation of genome-wide binding sites with the indicated histone modifications and location prediction by highest correlation with a histone modification, according to Figure 1b. **(b)** Binding of selected protein factors to promoters and enhancers in mouse ESCs. Heatmaps of 12913 promoters (upper panel) or 30564 enhancers (lower panel), centered on H3K4me3 signal (Promoters) or H3K4me1 signal (Enhancers), ranked on H3K27ac content from top to bottom. Displayed is 8 kb around the center of the promoter or enhancer. Normalized ChIP-seq reads representing the level of H3K4me1, H3K27ac and H3K4me3 histone modifications are indicated in the first three lanes. Normalized ChIP-seq reads representing relative binding intensity to promoters (upper panel) and enhancers (lower panel) of protein factors from Figure 3a (highest emPAI value ≥ 0.1) are displayed in lanes 4-12 and 14-20. Factors are arranged according to binding prediction or Polycomb factor identity. *p300 was not predicted by ChIP-MS but its genome-wide location was included in lane 13 for comparison.

with H3K27ac (Fig. 3a). Kdm2a and Rbbp5 have low H3K27ac ratios but still have high genome-associations with H3K27ac. Hence, in these two cases the H3K27ac ratios do not correlate well with H3K27ac association on the genome. Brd4 is the enhancer-predicted factor with the highest H3K27ac ratio and indeed has the highest correlation with H3K27ac on the genome (Fig. 3a). Oct4, Esrrb and Smc1a have intermediate H3K27ac ratios and intermediate genome-wide association with H3K27ac. Smarca4 has a high H3K27ac ratio, which in this case was not a good predictor, as Smarca4 has an intermediate level of genome-wide association with H3K27ac. We conclude from our above analyses that H3K27ac ratios provide a good, albeit not flawless, indication of the level of localization to H3K27ac-marked regions on the genome.

In a second approach to assess the accuracy of ChIP-MS, we investigated the presence of the above factors at promoters and enhancers in mouse ESCs. We assigned promoters as being present at the start of a gene and containing H3K4me3. Enhancers were assigned by their H3K4me1 content in the absence of H3K4me3. Promoters and enhancers were ranked top to bottom by their H3K27ac content (Fig. 3b). Promoter-predicted factors Kdm2a, Rbbp5, Polr2a, Wdr5, Tcea1, Supt5h and Kdm5b were indeed observed to only bind promoters and not enhancers (Fig. 3b). Hdac1 and Hdac2 predominantly bound to promoters but also showed binding to enhancers. Enhancer-predicted factors Esrrb, Oct4, Brd4 and Smc1a all showed strong binding to enhancers (Fig. 3b). These factors also showed binding to promoters to varying degrees. For comparison, we also included published genome-wide localization data of archetypal enhancer binder p300 in Figure 3b. Remarkably, p300 showed binding to enhancers and promoters (Fig. 3b), as previously observed in human ESCs[33]. Polycomb factors Rnf2, Jarid2, Suz12, Cbx7 and Mtf2 bind promoters with no correlation (Rnf2, Jarid2) or an anti-correlation (Cbx7, Suz12 and Mtf2) for H3K4me3 and H3K27ac content (Fig. 3b). The broad binding of Polycomb factors to promoters with relatively low H3K4me3 levels would explain their similar ChIP-MS values for H3K4me3 and H3K4me1 (Fig. 3a) and the associated ChIP-MS prediction uncertainties. The above analysis suggests that ChIP-MS-mediated prediction of binding to promoters or enhancers has high accuracy, with Polycomb factors again being an exception.

Dppa2 is not part of the classical ESC pluripotency network.

For many of the ChIP-MS detected factors the genome localization has not yet been determined by genome-wide ChIP and our experiments provide the first information on their genome binding preferences. As an example that ChIP-MS can identify factors with an unusual and therefore interesting genomic distribution, we focused on Dppa2 (Developmental PluriPotency Associated 2). Dppa2 is a member of family that also contains Dppa3 (Stella) and Dppa4, which all harbor a SAP DNA binding domain. The genome-wide binding sites of members of this family have not been determined so far. Dppa2 is exclusively expressed in the inner cell mass of the early embryo and later in the developing germ line and in cells derived from these tissues, such as embryonic stem cells and primordial germ cells[35]. Furthermore, Dppa2 expression was identified as an early marker for successful reprogramming towards iPSCs[9]. Dppa2 knockout ESCs have a slower proliferation rate and Dppa2 knockout mice die after birth from respiratory defects[36]. Recently it was shown that Dppa2, in combination with Lin28, Sall4 and Esrrb, drives the reprogramming of fibroblasts into iPSCs[9]. We selected Dppa2 because its

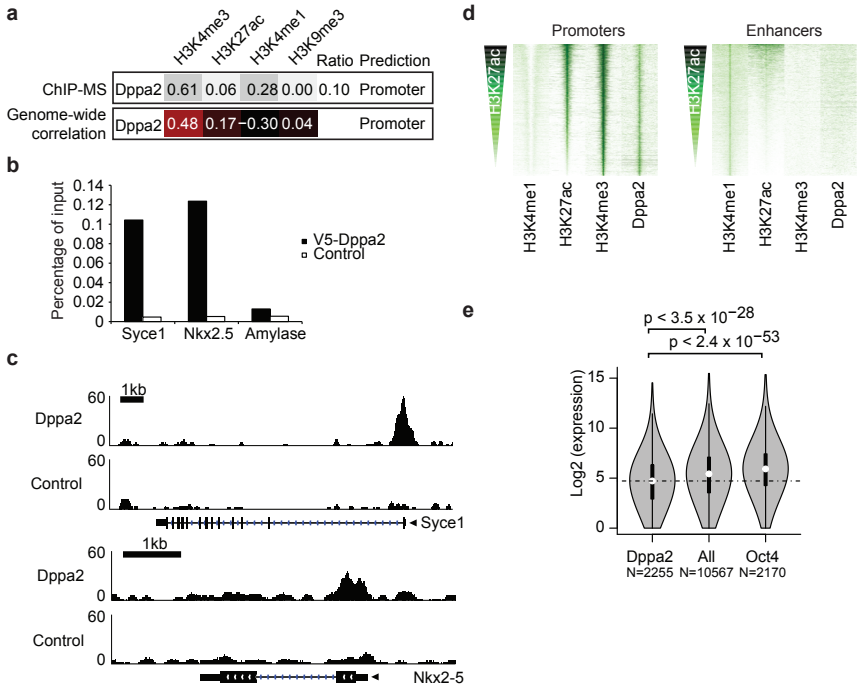
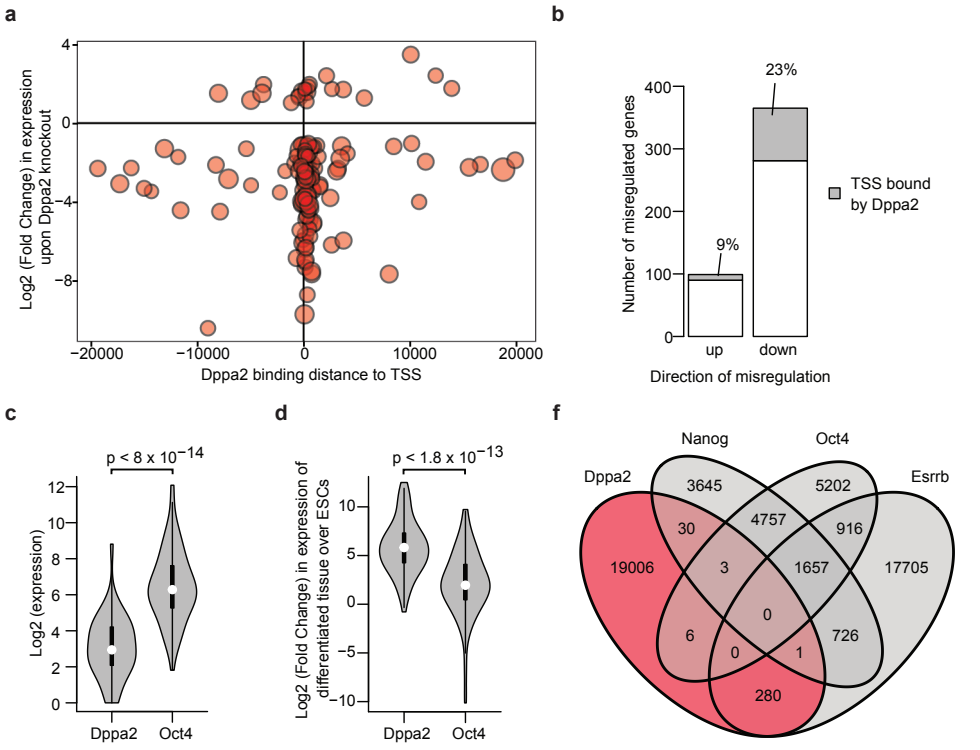


Figure 4. Analyses of genome-wide binding sites of Dppa2. (a) Comparison of location prediction for Dppa2 by ChIP-MS with location prediction by the correlation of identified Dppa2 genome-wide binding sites with the indicated histone modifications on the genome. Indicated in the upper panel from left to right are: Dppa2 average emPAI values in the different histone modification ChIPs, its H3K27ac ratio and ChIP-MS location prediction. Indicated in the lower panel from left to right are: the correlation of Dppa2 genome-wide binding sites with the indicated histone modifications and location prediction by highest correlation with a histone modification, according to Figure 1b. **(b)** Binding of Dppa2 to the promoters of the indicated genes, detected by anti-V5 ChIP on V5-Dppa2 expressing ESCs or control ESCs. Precipitated DNA for the indicated genes is shown as percentage of input, the Amylase gene is used as a negative control region. **(c)** Localization of Dppa2 on the promoter of Syce1 (upper panel) or Nkx2-5 (lower panel). Sequence reads from anti-V5 ChIP-seq on V5-Dppa2 expressing ESCs (Dppa2) or control ESCs (Control) were plotted relative to chromosomal position. Genome locations of Syce1 gene (upper panel) and Nkx2-5 gene (lower panel) are shown, scale bars indicate 1 kb of genome. **(d)** Binding of Dppa2 to promoters and enhancers in mouse ESCs. Heatmaps of 12913 promoters (left panel) or 30564 enhancers (right panel), centered on H3K4me3 signal (Promoters) or H3K4me1 signal (Enhancers), ranked on H3K27ac content from top to bottom. Displayed is 8 kb around the center of the promoter or enhancer. Normalized ChIP-seq reads representing the level of H3K4me1, H3K27ac and H3K4me3 histone modifications are indicated in the first three lanes. Normalized V5-Dppa2 ChIP-seq reads representing the relative binding intensity of Dppa2 to promoters (left panel) and enhancers (right panel) are displayed in the fourth lane of each panel. **(e)** Distribution of absolute expression levels of H3K4me3 marked genes in mouse ESCs that (from left to right) are bound at the promoter by Dppa2, all genes, and bound within 20kb around the promoter by Oct4. Shown is a violin plot where the white dot indicates the median and the thick black bar indicates 50% of the genes. Log2 value of the absolute expression, derived from published RNAseq data, the number of genes in each category and p-values by Mann-Whitney test are indicated.

ChIP-MS profile was unusual for a pluripotency-inducing factor. Dppa2 had the highest emPAI score for H3K4me3 but a low H3K27ac ratio, suggesting it binds predominantly to promoters with low activity (Fig. 4a). This is different for other pluripotency-inducing factors, such as Oct4 and Esrrb, which predominantly bind moderately active enhancers (Fig. 3b). To identify the genome-wide binding sites of Dppa2, we established an ESC line that expressed V5-tagged Dppa2 and we performed anti-V5 Dppa2 ChIP and



e

Dppa2 target genes

Tissue / Cells	N	FE
Testis	21	2.20
Epidermis	4	4.50
Lens	4	2.60
Kidney	4	2.30
Stomach	3	5.60
Nucleus accumbens	3	3.80
Lung	3	2.90
T.cells CD8.	2	8.80
Retinal pigment epithelium	2	2.50
Mast cells IgE	2	1.70
Heart	2	1.50
Mast cells	2	1.50
Skeletal muscle	2	1.00
Liver	2	0.79
Ciliary bodies	1	3.70
Osteoblast day5	1	3.00
Macrophage bone marrow 2h LPS	1	2.80
X3T3.L1	1	2.60
C2C12	1	2.50
Ovary	1	2.50

Oct4 target genes

Tissue / Cells	N	FE
Embryonic stem cell line Bruce4	17	4.90
Embryonic stem cell line V26 2	8	3.60
Intestine small	7	3.60
Macrophage bone marrow 6hr LPS	6	3.40
Skeletal muscle	6	2.00
Liver	5	1.30
Salivary gland	4	3.90
NK cells	4	1.80
Dorsal root ganglia	4	1.60
Testis	4	0.29
Ovary	3	5.00
Umbilical cord	3	4.80
T cells CD4	2	4.10
Uterus	2	4.00
Osteoblast day14	2	3.90
C2C12	2	3.40
Prostate	2	1.70
Mast cells IgE	2	1.20
Neuro2a	2	1.10
Adipose brown	2	1.00

g

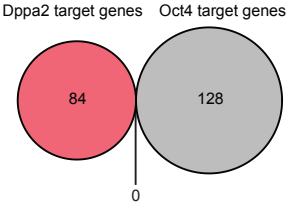


Figure 5. Dppa2 target genes and their overlap with the pluripotency network. (a) Bubble plot indicating the positions of Dppa2 binding sites relative to the transcription start site (TSS) on genes that are either up-regulated ≥ 2 fold (upper part) or down-regulated ≥ 2 fold (lower part) upon Dppa2 gene knockout in mouse ESCs. Log2 of the fold change in expression upon Dppa2 KO is indicated on the Y-axis. Distance of Dppa2 binding sites from the transcription start site is indicated on the X-axis. Size of

Legend continues on the bottom of the next page

sequenced the precipitated genomic DNA (ChIP-seq). We verified that V5-Dppa2 binds to the promoters of *Syce1* and *Nkx2-5* (Fig. 4b,c), the only known Dppa2 genomic binding sites[36], which suggested that our V5-Dppa2 ChIP identified bona-fide Dppa2 binding sites. Dppa2 binding sites had the highest genome-wide association with H3K4me3, whereas the association with H3K27ac (Fig. 4a) is lower than that of Oct4 and Esrrb (Fig. 3a). Subsequently, we investigated the presence of Dppa2 at promoters and enhancers in mouse ESCs. We found that Dppa2 binds promoters but is absent from enhancer regions (Fig. 4d). Interestingly, Dppa2 promoter binding displayed no correlation with H3K27ac content (Fig. 4d), a binding pattern that otherwise was only observed with the repressor Kdm5b and the Polycomb repressor proteins (Fig. 3b). These analyses suggest that our Dppa2 ChIP-MS location prediction was correct and that the binding pattern of Dppa2 is different compared to other reprogramming factors.

As expected, Dppa2 binds the promoters of genes with a lower median expression than other H3K4me3-marked promoters and Oct4-bound genes in ESCs (Fig. 4e and Supplementary Table 4). The gene expression profile of Dppa2 knockout ESCs was recently determined and it was observed that far more genes were down-regulated than up-regulated, compared to wild-type ESCs[36]. Dppa2 binds the promoters of nearly a quarter of the down-regulated genes but much less to promoters of the up-regulated genes (Fig. 5a,b). This suggests that Dppa2 maintains the expression of its putative target genes (Supplementary Table 5) by binding at their promoter. In contrast, Oct4 maintains the expression of its target genes by binding mostly outside promoters (Supplementary Fig. 2). We find that the median expression of Dppa2 target genes is 10-fold lower than the median expression of Oct4 target genes in ESCs (Fig. 5c). Nearly all Dppa2 target genes are higher expressed in tissues other than ESCs (Fig. 5d and Supplementary Table 5). The largest minority of Dppa2 target genes is highest expressed in testes, but many Dppa2 target genes have their highest expression in other tissues (Fig. 5e and Supplementary Table 5). For Oct4, the pattern is very different, as many Oct4 target genes have their highest expression in ESCs (Fig. 5d,e). Considering that Dppa2 and Oct4 appear to regulate different sets of genes, we determined the overlap in genome-wide binding sites between classical pluripotency factors such as Oct4, Nanog

(Figure 5. Legend continues from previous page)
the bubbles correlates with fold difference of Dppa2 ChIP peak over control. **(b)** Bar diagram showing the total number of up-regulated and down-regulated genes upon Dppa2 knockout and the number of these genes bound by Dppa2 within 1 kb from the TSS (grey areas). The number of Dppa2 bound genes is also indicated as a percentage of the total number of up-regulated or down-regulated genes. **(c)** Distribution of absolute expression levels in mouse ESCs of Dppa2 target genes and Oct4 target genes. Dppa2 target genes are bound by Dppa2 within 1 kb of the TSS and ≥ 2-fold down-regulated upon Dppa2 knockout, Oct4 target genes are bound by Oct4 within 20 kb of the TSS and ≥ 2 fold down-regulated after 24 hrs of Oct4 depletion. Shown is a violin plot where the white dot indicates the median and the thick black bar indicates 50% of the genes. Log2 of the absolute expression, derived from published RNAseq data and p-value by Mann-Whitney test are indicated. **(d)** Distribution of the fold change in expression of Dppa2 target genes and Oct4 target genes in the differentiated tissue with the highest expression versus expression in mouse ESCs. Shown is a violin plot where the white dot indicates the median and the thick black bar indicates 50% of the genes. Log2 of the fold change in expression, derived from published RNAseq data and p-value by Mann-Whitney test are indicated. **(e)** Lists of tissues or cells where Dppa2 target genes (left panel) or Oct4 target genes (right panel) are highest expressed. Tissue or cells, the number of genes (N) and fold enrichment (FE) of a tissue/ cell type within Dppa2 target genes or Oct4 target genes are indicated. The 20 tissues/ cells with the highest number of gene overlap and fold enrichment are shown. **(f)** Venn diagram showing the overlap of genomic binding sites in mouse ESCs of Dppa2, Nanog, Oct4 and Esrrb. **(g)** Venn diagram showing the lack of overlap of Dppa2 target genes and Oct4 target genes.

and Esrrb, and Dppa2. Whereas Oct4, Nanog and Esrrb showed an extensive overlap in binding sites, as previously reported[3], Dppa2 showed little overlap with these factors (Fig. 5f). In particular, the overlap of Dppa2 with Oct4 and Nanog was nearly absent. Moreover, there was no overlap between Dppa2 target genes and Oct4 target genes (Fig. 5g). We conclude that Dppa2 regulates a set of genes that is separate from the set of genes regulated by the classical pluripotency transcription factors, and not ESC-specific in its expression.

DISCUSSION

We describe here ChIP-MS, a method to predict the binding of factors to enhancers or promoters. Our experimental set-up is straightforward and does not rely on metabolic labeling for quantification by mass spectrometry. Nevertheless, using emPAI[22] and the simple rule that the ChIP fraction in which a factor has the highest emPAI score decides its binding prediction, we achieved a remarkably high percentage of correct predictions when comparing to published ChIP-seq data. We employed crosslinking with protein-protein crosslinker DSG, followed by standard crosslinking with formaldehyde. We previously used this “double crosslinking” procedure to improve ChIP efficiency[19-21]. Aside from our extended crosslinking procedure, our ChIP-MS procedure is based on standard ChIP protocols and mass spectrometry procedures, which should facilitate its application in other cell types.

A number of studies have been performed to screen for protein factors that bind to individual histone modifications by using modified histone peptides[37-39] or *in vitro* assembled modified nucleosomes[40], or identify protein factors that bind native chromatin harboring specific histone modifications by conventional ChIP combined with mass spectroscopy[41]. Binding to different tri-methylated lysines was assessed in these studies, but binding to enhancer marks or activity marks, such as the H3K4me1- or H3K27ac-marked chromatin that we interrogated with ChIP-MS, has not been addressed yet. Previous studies identified a number of ubiquitously expressed factors that we also observed in our ChIPs. However, our ChIP-MS procedure is sufficiently sensitive to also detect the sequence-specific transcription factors, such as Oct4, Esrrb, and Klf5, that determine ESC identity. These factors will be different in other cell types, which makes our procedure highly suitable to study the changing spectrum of regulatory region-associated factors during cell differentiation.

The list of proteins for which we predict genome localization by ChIP-MS (Fig. 2 and Supplementary Table 1) contains factors that play a role in a number of cellular processes, including all levels of transcriptional regulation and chromatin organization. ChIP-MS detected chromatin modifying complexes with clear location predictions, such as the BAF complex (enhancer), Aurora kinase complex (heterochromatin), Trrap complex (promoter), MLL complex (promoter) and Sin3 complex (promoter). The annotation of subunits of the Polycomb complexes PRC1 and PRC2 was more ambiguous, as PRC1 and PRC2 bind broad areas around inactive promoters marked to a similar extent by H3K4me3 and H3K4me1, leading to false “enhancer” predictions for several PRC subunits. Fortunately, Polycomb factors are well characterized[34,42] and would therefore be easy to recognize and treated with caution in any ChIP-MS prediction list.

Figure 2 ranks detected factors by the activity of their bound promoters or enhancers. To our knowledge, such an analysis has not been performed yet and provides a read-out on an important criterion for a large set of factors in ESCs. The ranking is based on the H3K27ac ratio; the ratio of the H3K27ac emPAI score over the highest of the H3K4me3 emPAI score or H3K4me1 emPAI score, which we anticipated would be the more informative value than the H3K27 emPAI score per se, as it compensates for the considerable differences in ChIP-MS detection levels for different factors. Perhaps the clearest indication that the H3K27ac ratio performs well in differentiating factors by the activity of their bound regions is that out of the nearly 240 factors detected by ChIP-MS, 3 members of the family of BET proteins, including Brd4, are among the proteins with the highest H3K27ac ratios. Brd4 was recently identified as a functional component and marker of "super enhancers"[25,26], arguably the most active enhancers in the genome of a cell. The ranking of the chromatin modifying complexes follows common sense. The activating BAF chromatin remodeling complex and Trrap histone acetylase complex have higher H3K27ac ratios than the Sin3 repression complex and the PRC1 and PRC2 complexes. The good performance of ChIP-MS on factors with a known genome-wide location suggests that also the localization predicted for the many factors without genome-wide ChIP data will in most cases be accurate. Our data set, graphically represented in Figure 2, therefore provides valuable new information on the transcriptional network in ESCs. Importantly, ChIP-MS can detect factors with an unusual, and therefore interesting, genome localization that can then be further investigated, Dppa2 being an example.

We performed our ChIP-MS experiments in ESCs. The establishment and maintenance of pluripotency, as well as the exit from pluripotency, is intensely studied in ESCs and several large data sets of relevant factors for the above processes are available. We find that a quarter (63 factors) of the ChIP-MS detected factors contributes to maintaining pluripotency. ChIP-MS detected Oct4, Esrrb, Klf5, Dppa2 and Mycn, factors which, as part of a 3-4 factor mix, reprogram somatic cells to iPSCs[4,7,8]. Intriguingly, ChIP-MS predicts that these factors do not all bind to the same type of locations on the genome. Oct4 and Esrrb were predicted to bind moderately active enhancers, whereas Klf5 binds to highly active enhancers. Mycn and Dppa2 were predicted to bind to high and low activity promoters, respectively. This suggests a division of labor between the different factors in the reprogramming process.

We found that Dppa2 had a different ChIP-MS profile compared to other pluripotency factors and accordingly we determined its genome-wide binding sites by ChIP-seq. Indeed, Dppa2 turned out to be an unusual pluripotency factor. Dppa2 binding sites and target genes do not overlap with Oct4, suggesting that Dppa2 is not part of the classical pluripotency circuit. Furthermore, Dppa2 target genes were found to be much lower expressed than Oct4 target genes in ESCs and higher expressed later in development. It is an intriguing question how Dppa2 can be an early marker and factor for reprogramming to iPSCs, as was recently shown[9], without actually regulating ESC-specific genes. Dppa2 was proposed as a factor that binds target genes to maintain an active chromatin structure and facilitate their later expression[36]. This epigenetic marking hypothesis is consistent with the expression pattern of the identified Dppa2 target genes. However, we did not find that genes bound by Dppa2 in ESCs were preferentially down-regulated in

Dppa2 knockout lungs, using a published gene expression set[36].

In conclusion, we established here a method to annotate factors to enhancers and promoters with different activities. ChIP-MS is straightforward in its set-up, which should facilitate its application to other cell types and growth conditions, provided sufficient cell quantities can be obtained. We showed that ChIP-MS data add to our knowledge and understanding of the transcriptional circuitry that determines cell identity.

METHODS

Cell Lines and constructs.

Mouse embryonic stem cell line CGR8 was grown on gelatin-coated dishes without feeders in GMEM (Glasgow Minimum Essential Medium) supplemented with LIF (leukemia inhibitory factor), 15% FBS, 0.25% sodium bicarbonate, 1mM glutamine, 1mM sodium pyruvate, non-essential aminoacids, 50μM β-mercaptoethanol and penicillin/streptomycin, as previously described[19]. The coding sequence for Dppa2 was amplified from mouse ES cell cDNA and cloned with an N-terminal V5-tag into a pPyCAG driven expression vector. CGR8 cells were transfected with the V5-Dppa2 expression vector using Lipofectamine 2000 (Invitrogen), clones were selected with 1μg/ml puromycin (Sigma) and stable expression of V5-tagged Dppa2 tested by Western blot analysis with anti-V5 antibody (1:2000) (Invitrogen).

ChIP-MS procedure.

For each histone modification ChIP, 300x10^6 ESCs were used. For chromatin preparation, cells were washed on plate three times with PBS and incubated with 2 mM Disuccinimidyl glutarate (DSG, Thermo Scientific) in PBS for 45 min at room temperature. Subsequently, ESCs were washed in PBS three times, 0.1 volume of 11% formaldehyde (Merck) in 50 mM Hepes-KOH pH 7.5, 100 mM NaCl, 1 mM EDTA, 0.5 mM EGTA was added, mixed and incubated for 12 min at room temperature, washed two times in 4 ˚C PBS and collected by centrifugation. All subsequent steps were performed on ice with pre-cooled buffers. Cell lysis was performed as described[43]. In brief, cells were collected and resuspended in LB1 (50 mM Hepes-KOH, pH 7.5, 140 mM NaCl, 1 mM EDTA, 10% glycerol, 0.5% NP-40, 0.25% Triton X-100). After 10 minutes incubation, cells were pelleted by centrifugation and resuspended in LB2 (10 mM Tris-HCl, pH 8.0, 200mM NaCl, 1mM EDTA, 0.5mM EGTA). After 10 min incubation, cells were pelleted and resuspended in 3 ml freshly prepared LB3 (10 mM Tris-HCl, pH 8.0, 100 mM NaCl, 1 mM EDTA, 0.5 mM EGTA, 0.1% Na-deoxycholate, 0.5% N-lauroylsarcosine) and sonicated on a Soniprep 150 (MSE), 27 cycles 15 s on, 45 s off on amplitude 7. After sonication, enriched DNA fragment size was confirmed to be between 200 and 1000 bp. 300 x 10^6 ESCs yielded approximately 10 mg chromatin (as measured by DNA content).

Antibodies used in the different histone modification ChIPs or GFP control ChIP are against H3K4me1 (ab8895, Abcam), H3K4me3 (ab8580, Abcam), H3K27Ac (ab4729, Abcam), H3K9me3 (ab8898, Abcam) and GFP (sc8334, Santa Cruz Biotechnology). To prevent immunoglobin elution and subsequent interference with the mass spectrometry analysis, 50 μg antibodies were crosslinked to 500 μl Protein A magnetic bead solution (15 mg beads, Life Technologies) with Dimethyl Pimelimidate (Sigma). Cross-linked

antibody-bead complexes were equilibrated in LB3 buffer and subsequently blocked with 0.5 mg/ml BSA (New England Biolabs) and 0.2 mg/ml sonicated salmon sperm DNA (Stratagene) for one hour. The antibody-bead mixture was rotated overnight with approximately 10 mg chromatin at 4°C. Beads were transferred to 1.5 ml no stick tubes (Alpha laboratories) and washed five times 5 minutes in RIPA buffer (50 mM Hepes-KOH pH 7.6, 500 mM LiCl, 1mM EDTA, 1% NP-40, 0.7% Na-deoxycholate). After washing, the beads were boiled for 35 min at 95 °C in 2x SDS sample buffer (100 mM Tris-HCl pH 6.8, 200 mM DTT, 4% SDS, 20% Glycerol, 0.2% Bromophenol blue) and supernatant was transferred to a fresh tube. ChIP-MS samples were run on 10% precast SDS-PAGE gels (NuPage Invitrogen) and stained with colloidal coomassie stain (Invitrogen). Gel lanes were sliced, in-gel digested with trypsine to yield peptides and proteins identified by analyses on an LQT-Orbitrap mass spectrometer (Thermo), as described[21]. For Western blot analyses, ChIP samples were separated on a 4-12% polyacrylamide gel (Novex) and nitrocellulose blots probed with antibodies against the used histone modifications (see above, 1:500 dilution, pan histone H3 antibody (Abcam 1791, 1:1000 dilution) and Nanog (Cosmo Bio Ltd., 1:2000 dilution). Non-cropped versions of the Western blot panels in Figure 1d can be found in Supplementary Fig. 3.

ChIP-MS inclusion and prediction criteria.

Two independent ChIPs were performed for each tested histone modification and for GFP, as control ChIPs, and analyzed by mass spectrometry. For inclusion into the ChIP-MS list of identified proteins (Supplementary Tables 1 and 2), factors needed to be identified by mass spectrometry with a Mascot score of 50 or higher in at least one histone modification ChIP. A Mascot score of at least 45 in any of the other ChIPs was annotated in the ChIP-MS list. In case of Mascot scores between 45 and 60, individual peptide MS/MS spectra were checked manually and interpreted as valid identifications or discarded. In addition, the Mascot program was used to determine the Mascot peptide significance threshold ($P < 0.05$) in the ChIP-MS samples. Significance thresholds were H3K4me3 (Experiment 1); Mascot score 28, H3K4me3 (Experiment 2); Mascot score 28, H3K4me1 (Experiment 1); Mascot score 28, H3K4me1 (Experiment 2); Mascot score 28, H3K27ac (Experiment 1); Mascot score 28, H3K27ac (Experiment 2); Mascot score 28, H3K9me3 (Experiment 1); Mascot score 28, H3K9me3 (Experiment 2); Mascot score 29, GFP (Experiment 1); Mascot score 28, GFP (Experiment 2); Mascot score 28. For inclusion into the ChIP-MS list (Supplementary Tables 1 and 2), protein identifications had to be based on peptides with a Mascot score at or above the Mascot peptide significance threshold of the sample in which the peptides were observed. For assessment of the quantity of the identified proteins in the ChIP-MS samples we used emPAI a calculation method based on the number of peptide spectra identified by MS, normalized for the number of peptides that theoretically should be identifiable for that protein[22] (to compensate for large proteins, which likely have more MS peptides). Inclusion into the ChIP-MS list required an at least 5-fold higher average emPAI score in the ChIPs for at least one histone modification compared to any of the anti-GFP control ChIPs. Inclusion into the ChIP-MS list further required a factor to have an at least 3-fold higher average emPAI score in the two ChIPs for one histone modification compared to the two ChIPs for one or more of the other histone modifications, to exclude factors that bind chromatin indiscriminately of the tested histone modifications. Cytoskeletal and cytoplasmic proteins were excluded. Average emPAI scores were calculated from the two independent ChIP-MS experiments.

Localisation prediction was according to the following criteria; highest average emPAI score in H3K4me3 ChIP samples gives “promoter” prediction, highest average emPAI score in the H3K4me1 ChIP samples gives “enhancer” prediction, highest average emPAI score in the H3K9me3 ChIP samples gives “heterochromatin” prediction. In case average emPAI scores for the H3K4me3 and H3K4me1 ChIP samples were equal, the prediction was “promoter”. The H3K27ac ratio of a factor is defined as the ratio of its average H3K27ac emPAI score over the average H3K4me3 emPAI score or average H3K4me1 emPAI score, whichever one is the highest.

Chromatin immunoprecipitation and sequencing.
Anti-V5 ChIPs were performed as described[21]. For V5-Dppa2 ChIP, CGR8 ESCs stably expressing V5-Dppa2 (see above) were used, for the control ChIP, the parental CGR8 parental ESC line was used. For each ChIP, 100 x 10^6 ESCs were used. Precipitated DNA was analyzed by quantitative PCR or used for library generation followed by next generation sequencing on an Illumina Genome analyzer, as described[20].

Data Analysis.
Sequences with low complexity that are unlikely to map uniquely to the genome were removed from the Dppa2 ChIP-seq, modified ChIP-seq experiments for the used histone modifications and published ChIP-seq datasets (Supplementary Table 6), using prinseq-lite with the dust method with 7 as threshold[44]. The remaining sequences with a Phred score <70 were mapped to the mm9 reference genome using Bowtie[45] v0.12.7, where we used a seed length of 36 in which we allowed a maximum of 2 mismatches. If a read had multiple alignments only the best matching read was reported. Duplicated reads were removed. MACS[46] v1.4.2 was used for peak calling of Esrrb, Nanog, Oct4, Polr2a, P300, H3K4me1 and H3K4me3 using default settings. ChIP-seq datasets with multiple replicates were merged. For peak calling the Polr2a, P300, H3K4me1, H3K4me3 ChIP-seq datasets, the sequenced input was used as control. For peak calling Esrrb, Nanog and Oct4, the GFP ChIP was use as a control (Supplementary Table 6). For Dppa2, peak calling, we provided MACS1.4.2 with a shift size of 75 base pairs. Peaks with a p-value ≤ 1 x 10^{-10} were retained for Dppa2, Oct4, Nanog, Polr2a, H3K4me3, H3K4me1 and sites with a p-value ≤ 2 x 10^{-10} for Esrrb and P300. For Dppa2, only peaks with at least 100 aligned reads were retained. Dppa2, Oct4, Nanog and Esrrb peaks were considered to overlap if their peak summit was within 125 base pairs of each other. Venn diagrams, violin plots and bubble plots were created in R using the VennDiagram, vioplot and ggplots2 packages, respectively. To calculate the overlap between the mapped reads from the modified ChIP-seq experiments for H3K4me3, H3K4me1, H3K27ac and promoters or enhancers. Promoters were defined as the regions from -1 kb to +1 kb of the summits of a significant RNAPol2 binding sites (Polr2a) within 1 kb of a transcription start site (TSS). Enhancers were defined as the regions from -1 kb to +1 kb of the summits of a significant P300 binding sites that were not within 1kb of a TSS. The sequencing profiles of the conventional ChIP and modified ChIP for the used histone modifications and the Dppa2 ChIP-seq experiments (Fig. 1g, Fig. 4c and Supplementary Fig. 1) were created in the IGV browser[47].

Genome-wide correlation.
To calculate the genome wide correlation between the published histone modification

ChIP-seq datasets and the protein factors, we divided the entire genome in bins of 1000 base pairs and calculated Reads Per Million (RPM) for all bins in all datasets. The input was subtracted. For each histone modification and protein factor we selected the 4000 bins with the highest RPM. A unified list was created for each individual protein factor, containing the selected bins of the four used histone modifications and that of the protein factor itself. The Spearman correlation coefficients of the protein factor with the different histone modifications were calculated from this list. To calculate the correlations between conventional and modified ChIP-seq experiments for the used histone modifications, we used a unified list that included the 4000 bins with the highest RPM for each conventional and modified ChIP-seq experiment.

Heatmaps.
To assign promoter regions, H3K4me3 peak summits, as determined by MACS (see above) were required to be within 1 kb range of a TSS, resulting in 12913 promoter regions. To assign enhancer regions, H3K4me1 peak summits were filtered against the presence of H3K4me3 signal in a region from -4.1 kb and + 4.1 kb around the H3K4me1 peak summit, resulting in 30564 enhancer regions. Promoters were sorted for number of H3K27ac reads present in the central 2 kb of the promoter region. Enhancers were sorted for the number of H3K27ac reads in the central 8.2 kb of the enhancer region, to also include broad enhancers. For both promoters and enhancers we displayed a region from -4.1 kb to +4.1 kb around the peak summits, divided into 51 bins of 160 bp each. Promoter and enhancer heatmaps for each protein factor or histone modification were normalized by calculating the RPM based on the sum of all reads found in the displayed promoter and enhancer region for that factor.

Expression of Dppa2 and Oct4 target genes.
Dppa2-bound genes contained a MACS-called Dppa2 peak (see above) within 1 kb from the TSS. Dppa2 target genes were defined as Dppa2-bound genes with at least 2-fold difference in expression in Dppa2 knockout ESCs, compared to wild-type ESCs and an adjusted p-value of ≤ 0.10, in a Dppa2 knockout microarray dataset[31]. The GEO2R script, as provided by the authors on the Gene Expression Omnibus (GEO), was used to calculate fold change in expression and adjusted p-value for each probe. Multiple probes to the same gene were aggregated by taking the average fold change. Oct4-bound genes contained a MACS-called Oct4 peak (see above) within 20 kb from the TSS. Oct4 target genes were defined as Oct4-bound genes with at least 2-fold difference in expression between 24 hrs and 0 hrs after Oct4 knockdown[48] in ZHBTc4 ESCs that have their only intact Oct4 gene under doxycycline control. The used microarray dataset[48] was already normalized by the authors. A published RNA sequencing dataset consisting of two replicates[49] was used to calculate the mean expression of Dppa2- or Oct4-bound genes and Dppa2- or Oct4-target genes. Both replicates were mapped against mouse reference NCBIM37.67 using Tophat[50] v2.0.11 with default settings and a segment length of 20. The aligned exon reads were counted and normalized using Bioconductor DESeq2 package in R. Replicates were normalized by dividing the counts by their sizefactors. The expression level per gene was calculated by taking the average of both replicates and calculating the reads per kb for each gene. To calculate the fold change of Dppa2- and Oct4-target genes in differentiated tissues over ESCs, we used the BioGPS mouse MOE430 Gene Atlas[51]. The same database was used to determine the tissue or cell line

in which the Dppa2 and Oct4 target genes were highest expressed.

Acknowledgements
We thank Harmen van der Werken for advice on bioinformatics analyses and Erik-Jan Rijkers for proteomics data formatting. We thank Sjaak Philipsen, Derk ten Berge, Danny Huylebroeck, Joost Gribnau and Frank Grosveld for suggestions that improved the manuscript. E.E. and J.H.B were supported by ALW-open program grants from the Netherlands Organisation for Scientific Research (NWO), M.M. by a grant from the Erasmus MC Stem Cell Institute, D.L.C. v.d. B. and R.A.P. by a network grant from the Netherlands Institute of Regenerative Medicine (NIRM) and D.L.C. v.d. B. by a FEBS long term fellowship and R.A.P. by a grant from the Dutch government to the Netherlands Institute for Regenerative Medicine (NIRM, grant No. FES0908).

Author contributions
E.E. performed the ChIP-MS and Dppa2 experiments and analyzed the data. J.H.B. performed the modified ChIP-seq experiments and all bioinformatics analyses. M.J.M. performed the Western analyses and cloned Dppa2 cDNA into an expression vector. L.S. cloned Dppa2 cDNA into an expression vector. D.H.W.D and J.D. performed the mass spectrometry analyses. C.E.M.K., Z.O. and W.F.J.v.IJ. performed Illumina sequencing of ChIP material. D.L.C.v.d.B. conceived the study and performed pilot experiments, R.A.P. conceived the study and designed experiments. R.A.P., E.E., J.H.B and D.L.C.v.d.B. wrote the manuscript.

Author information
The authors declare no competing interests. ChIP sequencing data are available through the Gene Expression Omnibus (NCBI), accession code GSE58113.

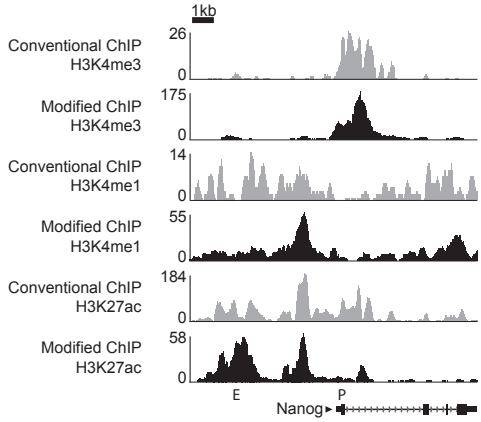
REFERENCES

1. Mouse, E.C. et al. An encyclopedia of mouse DNA elements (Mouse ENCODE). Genome Biol 13, 418 (2012).
2. Consortium, E.P. et al. An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57-74 (2012).
3. Chen, X. et al. Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. Cell 133, 1106-17 (2008).
4. Takahashi, K. & Yamanaka, S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. Cell 126, 663-76 (2006).
5. Soufi, A., Donahue, G. & Zaret, K.S. Facilitators and impediments of the pluripotency reprogramming factors' initial engagement with the genome. Cell 151, 994-1004 (2012).
6. Buganim, Y., Faddah, D.A. & Jaenisch, R. Mechanisms and models of somatic cell reprogramming. Nat Rev Genet 14, 427-39 (2013).
7. Nakagawa, M. et al. Generation of induced pluripotent stem cells without Myc from mouse and human fibroblasts. Nat Biotechnol 26, 101-6 (2008).
8. Feng, B. et al. Reprogramming of fibroblasts into induced pluripotent stem cells with orphan nuclear receptor Esrrb. Nat Cell Biol 11, 197-203 (2009).
9. Buganim, Y. et al. Single-cell expression analyses during cellular reprogramming reveal an early stochastic and a late hierarchic phase. Cell 150, 1209-22 (2012).
10. Bernstein, B.E. et al. Genomic maps and comparative analysis of histone modifications in human and mouse. Cell 120, 169-81 (2005).
11. Pokholok, D.K. et al. Genome-wide map of nucleosome acetylation and methylation in yeast. Cell 122, 517-27 (2005).

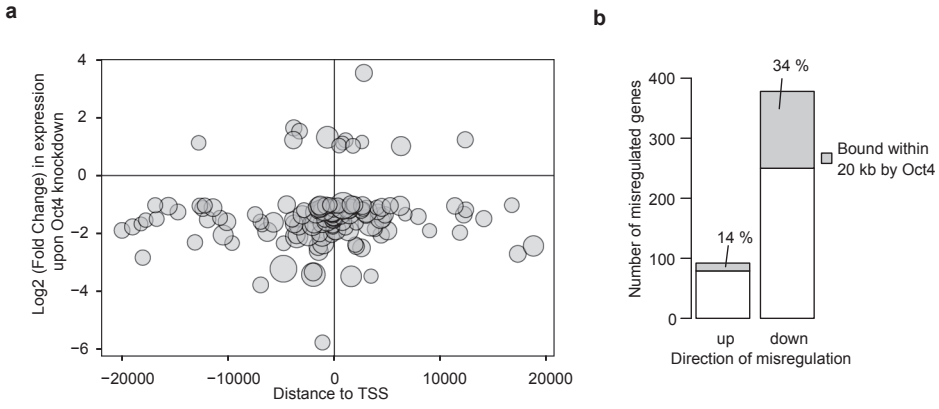
12. Hawkins, R.D. et al. Dynamic chromatin states in human ES cells reveal potential regulatory sequences and genes involved in pluripotency. Cell Res 21, 1393-409 (2011).
13. Ernst, J. et al. Mapping and analysis of chromatin state dynamics in nine human cell types. Nature 473, 43-9 (2011).
14. Heintzman, N.D. et al. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nat Genet 39, 311-8 (2007).
15. Creyghton, M.P. et al. Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proc Natl Acad Sci U S A 107, 21931-6 (2010).
16. Rada-Iglesias, A. et al. A unique chromatin signature uncovers early developmental enhancers in humans. Nature 470, 279-83 (2011).
17. Peters, A.H. et al. Partitioning and plasticity of repressive histone methylation states in mammalian chromatin. Mol Cell 12, 1577-89 (2003).
18. Nowak, D.E., Tian, B. & Brasier, A.R. Two-step cross-linking method for identification of NF-kappaB gene network by chromatin immunoprecipitation. Biotechniques 39, 715-25 (2005).
19. van den Berg, D.L. et al. Estrogen-related receptor beta interacts with Oct4 to positively regulate Nanog gene expression. Mol Cell Biol 28, 5986-95 (2008).
20. Engelen, E. et al. Sox2 cooperates with Chd7 to regulate genes that are mutated in human syndromes. Nat Genet 43, 607-11 (2011).
21. van den Berg, D.L. et al. An Oct4-centered protein interaction network in embryonic stem cells. Cell Stem Cell 6, 369-81 (2010).
22. Ishihama, Y. et al. Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein. Mol Cell Proteomics 4, 1265-72 (2005).
23. Aagaard, L. et al. Functional mammalian homologues of the Drosophila PEV-modifier Su(var)3-9 encode centromere-associated proteins which complex with the heterochromatin component M31. Embo J 18, 1923-38 (1999).
24. Filippakopoulos, P. et al. Histone recognition and large-scale structural analysis of the human bromodomain family. Cell 149, 214-31 (2012).
25. Whyte, W.A. et al. Master transcription factors and mediator establish super-enhancers at key cell identity genes. Cell 153, 307-19 (2013).
26. Hnisz, D. et al. Super-enhancers in the control of cell identity and disease. Cell 155, 934-47 (2013).
27. Anders, L. et al. Genome-wide localization of small molecules. Nat Biotechnol 32, 92-6 (2014).
28. Ivanova, N. et al. Dissecting self-renewal in stem cells with RNA interference. Nature 442, 533-8 (2006).
29. Hu, G. et al. A genome-wide RNAi screen identifies a new transcriptional module required for self-renewal. Genes Dev 23, 837-48 (2009).
30. Chia, N.Y. et al. A genome-wide RNAi screen reveals determinants of human embryonic stem cell identity. Nature 468, 316-20 (2010).
31. Jiang, J. et al. A core Klf circuitry regulates self-renewal of embryonic stem cells. Nat Cell Biol 10, 353-60 (2008).
32. Wang, J. et al. A protein interaction network for pluripotency of embryonic stem cells. Nature 444, 364-8 (2006).
33. Ram, O. et al. Combinatorial patterning of chromatin regulators uncovered by genome-wide location analysis in human cells. Cell 147, 1628-39 (2011).
34. Schuettengruber, B. & Cavalli, G. Recruitment of polycomb group complexes and their role in the dynamic regulation of cell fate choice. Development 136, 3531-42 (2009).
35. Bortvin, A. et al. Incomplete reactivation of Oct4-related genes in mouse embryos cloned from somatic nuclei. Development 130, 1673-80 (2003).
36. Nakamura, T., Nakagawa, M., Ichisaka, T., Shiota, A. & Yamanaka, S. Essential roles of ECAT15-2/Dppa2 in functional lung development. Mol Cell Biol 31, 4366-78 (2011).
37. Bartke, T. et al. Nucleosome-interacting proteins regulated by DNA and histone methylation. Cell 143, 470-84 (2010).
38. Vermeulen, M. et al. Quantitative interaction proteomics and genome-wide profiling of epigenetic histone marks and their readers. Cell 142, 967-80 (2010).
39. Eberl, H.C., Spruijt, C.G., Kelstrup, C.D., Vermeulen, M. & Mann, M. A map of general and specialized chromatin readers in mouse tissues generated by label-free interaction proteomics. Mol Cell 49, 368-78 (2013).

40. Nikolov, M. et al. Chromatin affinity purification and quantitative mass spectrometry defining the interactome of histone modification patterns. Mol Cell Proteomics 10, M110 005371 (2011).
41. Soldi, M. & Bonaldi, T. The proteomic investigation of chromatin functional domains reveals novel synergisms among distinct heterochromatin components. Mol Cell Proteomics 12, 764-80 (2013).
42. Aloia, L., Di Stefano, B. & Di Croce, L. Polycomb complexes in stem cells and embryonic development. Development 140, 2525-34 (2013).
43. Boyer, L.A. et al. Core transcriptional regulatory circuitry in human embryonic stem cells. Cell 122, 947-56 (2005).
44. Schmieder, R. & Edwards, R. Quality control and preprocessing of metagenomic datasets. Bioinformatics 27, 863-4 (2011).
45. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10, R25 (2009).
46. Zhang, Y. et al. Model-based analysis of ChIP-Seq (MACS). Genome Biol 9, R137 (2008).
47. Robinson, J.T. et al. Integrative genomics viewer. Nat Biotechnol 29, 24-6 (2011).
48. Sharov, A.A. et al. Identification of Pou5f1, Sox2, and Nanog downstream target genes with statistical confidence by applying a novel algorithm to time course microarray and genome-wide chromatin immunoprecipitation data. BMC Genomics 9, 269 (2008).
49. Stadler, M.B. et al. DNA-binding factors shape the mouse methylome at distal regulatory regions. Nature 480, 490-5 (2011).
50. Kim, D. et al. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol 14, R36 (2013).
51. Lattin, J.E. et al. Expression analysis of G Protein-Coupled Receptors in mouse macrophages. Immunome Res 4, 5 (2008).

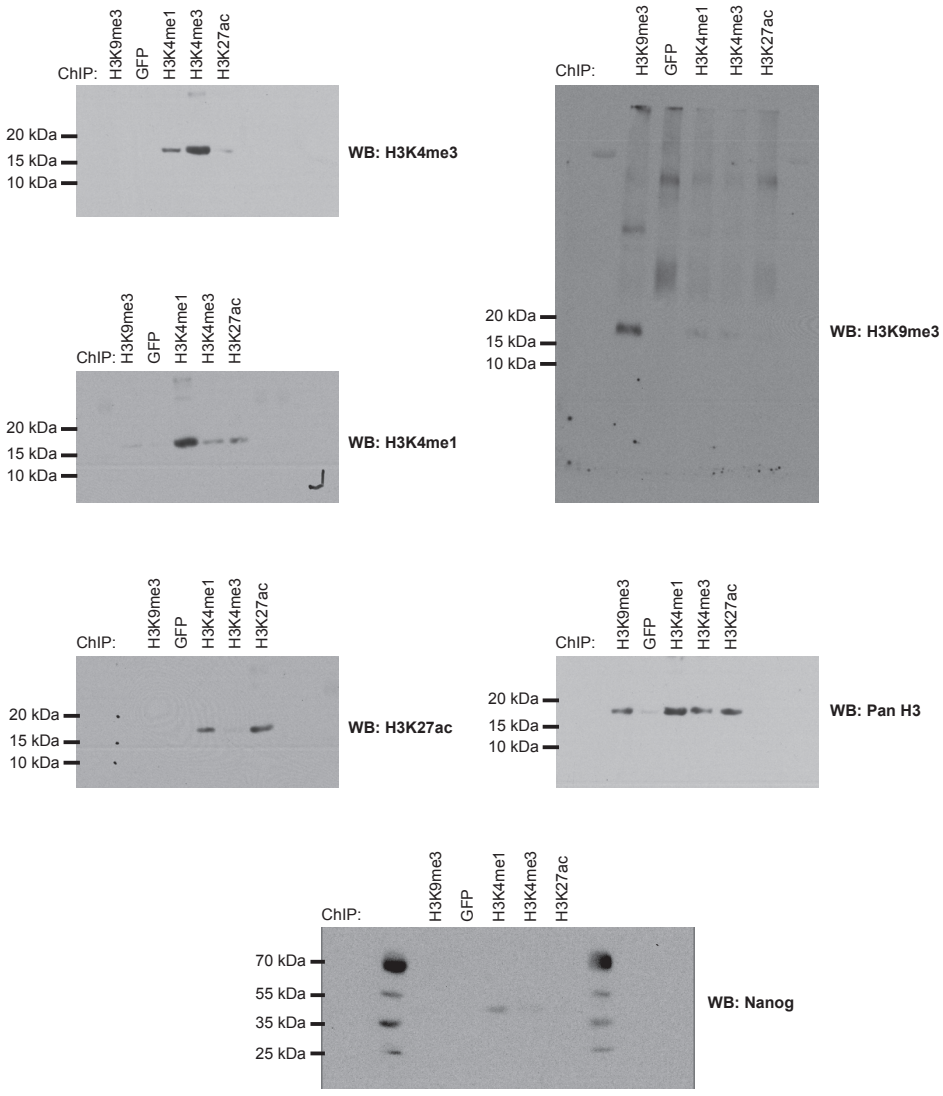
SUPPLEMENTARY INFORMATION



Supplementary Figure 1. Modified ChIP-seq tracks around pluripotency gene Nanog. ChIP-seq tracks for modified or conventional ChIP for H3K4me3, H3K4me1 or H3K27ac around pluripotency gene *Nanog*. Sequence reads were plotted relative to chromosomal position. Genome location of *Nanog* is shown, scale bar indicates 1 kb of genome. P indicates Promoter, E indicates putative enhancer.



Supplementary Figure 2. Position of Oct4 binding sites on Oct4 target genes. (a) Bubble plot indicating the positions of Oct4 binding sites relative to the transcription start site (TSS) on genes that are either up-regulated ≥ 2 fold (upper part) or down-regulated ≥ 2 fold (lower part) after 24 h of Oct4 depletion in mouse ESCs. Log2 of the fold change in expression is indicated on the Y-axis, Oct4 binding distance from TSS (in base pairs) is indicated on the X-axis. Size of the bubbles correlates with fold difference of Oct4 ChIP peak over control. **(b)** Bar diagram showing the total number of up-regulated and down-regulated genes upon Oct4 depletion and the number of these genes bound by Oct4 within 20 kb from the TSS (grey areas). The number of Oct4-bound genes is also indicated as a percentage of the total number of up-regulated or down-regulated genes.



Supplementary Figure 3. Western Blots (full scans) for ChIP-MS samples (as in Figure 1d). Western blots for histone modifications and Nanog for ChIP-MS samples. ChIP-MS samples (top), molecular weight markers (left) and antibody used for Western blot (right) are indicated.

Supplementary Table 1. ChIP-MS identified proteins, predictions and phenotypes

Protein	Complex[a]	H3K4me3 avrg emPAI	H3K27ac avrg emPAI	H3K4me1 avrg emPAI	H3K9me3 avrg emPAI	GFP avrg emPAI	H3K27ac ratio	Prediction[b]	Pluripotency phenotype[c]
Actl6a	BAF + INO80 + Trrap/Ep400	0,29	0,16	0,12	0,08		0,55	Promoter	-
Actr3	Arp2/3	0,04					0	Promoter	-
Actr5			0,03					Unclear	-
Adnp				0,03	0,02		0	Enhancer	-
Ankhd1				0,10			0	Enhancer	-
Anln					0,03			Heterochromatin	-
Anp32a	SET	0,17		0,18			0	Enhancer	-
Arid1a	BAF		0,07	0,06			1,17	Enhancer	yes
Arid4a	Sin3	0,12					0	Promoter	-
Ash2l	MLL	0,25					0	Promoter	yes
Atad2b			0,03	0,04	0,04		0,75	Enhancer	-
Atrx					0,09			Heterochromatin	-
Aurkb	Aurora K	0,22	0,10	0,10	0,65			Heterochromatin	yes
Banf1		0,66	0,47	0,63			0,71	Promoter	yes
Bap18	MLL	0,63					0	Promoter	-
Baz1a	ACF			0,01			0	Enhancer	-
Baz2a	NoRC	0,06	0,22	0,20	0,07		1,10	Enhancer	-
Bend3		0,24		0,04			0	Promoter	-
Bptf	NURF	0,08	0,02	0,02			0,25	Promoter	yes
Brd1	MOZ/MORF	0,03	0,05	0,10			0,50	Enhancer	-
Brd2		0,37	0,75	0,18			2,03	Promoter	-
Brd3		0,12	0,32				2,67	Promoter	-
Brd4		0,03	0,21	0,10			2,10	Enhancer	-
Brms1	Sin3	0,14					0	Promoter	-
Brms1l	Sin3	0,47					0	Promoter	-
Brpf1	MOZ/MORF	0,16					0	Promoter	-
Brpf3	MOZ/MORF	0,03					0	Promoter	-
Cad				0,04			0	Enhancer	-
Cald1		0,06					0	Promoter	-
Cbx7	PRC1	0,10		0,10			0	Promoter	yes
Cdc5l	Prp19			0,04			0	Enhancer	-
Cdc73	PAF1	0,10	0,10				1,00	Promoter	yes
Cdca7			0,09	0,09			1,00	Enhancer	-
Cdca8	Aurora K	0,28	0,13	0,24	1,17			Heterochromatin	-
Cdk7	TFIIH	0,05					0	Promoter	-
Cdk9	P-TEFb		0,09					Unclear	yes
Cep95			0,02	0,04			0,50	Enhancer	-
Chaf1b	CAF1		0,23	0,52			0,44	Enhancer	yes
Chd1		0,26	0,24	0,14	0,02		0,92	Promoter	yes
Chd2		0,05		0,03			0	Promoter	-
Chd8		0,15	0,03				0,20	Promoter	-
Ckb		0,14					0	Promoter	-
Coil					0,03			Heterochromatin	-
Cpsf6	CFIm	0,07	0,07				1,00	Promoter	-
Creb1		0,11					0	Promoter	-
Cse1l		0,04					0	Promoter	-
Csnk2a1		0,89	0,27	0,21			0,30	Promoter	-
Csnk2b		0,82	0,26	0,08			0,32	Promoter	-
Csrp2bp	ATAC	0,04					0	Promoter	-
Ctbp2		0,46	0,30	0,43	0,04		0,65	Promoter	yes
Ctcf		0,07		0,05			0	Promoter	-
Ctr9	PAF1	0,03	0,03				1,00	Promoter	yes
Cul4b	DDB1/Cul4 ubiquitin ligase	0,17	0,18	0,18			1,00	Enhancer	-
Ddb1	DDB1/Cul4 ubiquitin ligase	0,19	0,24	0,29			0,83	Enhancer	-
Ddx47					0,15			Heterochromatin	-
Ddx54				0,04			0	Enhancer	-
Dido1		0,02					0	Promoter	-
Dis3			0,05	0,05			1,00	Enhancer	-
Dmap1	Trrap/Ep400	0,07		0,04			0	Promoter	yes
Dnajc9		0,07	0,14	0,14			1,00	Enhancer	-
Dpf2	BAF		0,22	0,12			1,83	Enhancer	-
Dppa2		0,61	0,06	0,29			0,10	Promoter	yes
Dppa4		0,58		0,39			0	Promoter	yes
Dpy30	MLL	0,18					0	Promoter	yes
Dut		0,41	0,18	0,18			0,44	Promoter	-
Emsy		0,02					0	Promoter	-
Ep300			0,05					Unclear	yes
Ep400	Trrap/Ep400		0,02					Unclear	yes

Protein	Complex[a]	H3K4me3 avrg emPAI	H3K27ac avrg emPAI	H3K4me1 avrg emPAI	H3K9me3 avrg emPAI	GFP avrg emPAI	H3K27ac ratio	Prediction[b]	Pluripotency phenotype[c]
Erh		0,37					0	Promoter	-
Esrrb			0,08	0,20			0,40	Enhancer	yes
Ezh2	PRC2	0,07	0,04	0,09			0,44	Enhancer	yes
Fam60a	Sin3	0,24					0	Promoter	-
Fbxl19	SCF ubiquitin ligase	0,05					0	Promoter	-
Fkbp4				0,07			0	Enhancer	-
Gltscr2				0,07			0	Enhancer	-
Glyr1		0,19	0,24	0,26	0,07		0,92	Enhancer	-
Gm53		0,19					0	Promoter	-
Gtf2e1	TFIIE			0,08			0	Enhancer	-
Gtf2i			0,04	0,07			0,57	Enhancer	-
Gtf3c1	TFIIIC	0,01	0,01	0,02			0,50	Enhancer	-
Hat1		0,11	0,04	0,04			0,36	Promoter	-
Hcfc1	MLL	0,05					0	Promoter	yes
Hdac1	NuRD + Sin3 + REST	1,02	0,30	0,38	0,11		0,29	Promoter	yes
Hdac2	NuRD + Sin3 + REST	0,40		0,10			0	Promoter	-
Hdgfrp2			0,05					Unclear	-
Hmga2		0,14	0,14		0,28			Heterochromatin	-
Hmgb3		0,11		0,1			0	Promoter	-
Hmgxb4		0,20					0	Promoter	-
Hp1bp3		0,26		0,38	0,23		0	Enhancer	-
Incenp	Aurora K	0,06			0,26			Heterochromatin	-
Ing1	Sin3	0,06					0	Promoter	-
Ing2	Sin3	0,11					0	Promoter	-
Ing3	Trrap/Ep400	0,04	0,04				1,00	Promoter	-
Ing4	Trrap/Ep400	0,14		0,07			0	Promoter	-
Ing5	MOZ/MORF	0,58	0,13	0,20			0,22	Promoter	yes
Ino80	INO80	0,05	0,05				1,00	Promoter	yes
Ints1	Integrator			0,02			0	Enhancer	-
Irgc				0,04			0	Enhancer	-
Jarid2	PRC2	0,16	0,05	0,21	0,04		0,24	Enhancer	yes
Kdm1a	BHC	0,06	0,02	0,02	0,02		0,33	Promoter	yes
Kdm2a	SCF ubiquitin ligase	0,16	0,03	0,03			0,19	Promoter	-
Kdm2b	PRC1	0,09		0,01			0	Promoter	yes
Kdm4a		0,03					0	Promoter	-
Kdm4c		0,65	0,03	0,13			0,05	Promoter	yes
Kdm5a		0,02					0	Promoter	-
Kdm5b		0,25	0,02	0,12			0,08	Promoter	yes
Kif2c				0,05			0	Enhancer	-
Kif4				0,03			0	Enhancer	-
Kifc5b			0,07	0,15			0,47	Enhancer	-
Klf16		0,16					0	Promoter	-
Klf5			0,07	0,04			1,75	Enhancer	yes
L3mbtl2		0,08		0,07			0	Promoter	yes
Las1l	5FMC			0,06			0	Enhancer	-
Lasp1		0,29		0,13			0	Promoter	-
Lig1		0,08		0,02			0	Promoter	-
Lig3			0,02	0,05			0,40	Enhancer	-
Llph				0,13			0	Enhancer	-
Lmna		0,13	0,08	0,08	0,48			Heterochromatin	-
Lmnb1		0,99	1,60	1,83	3,15	0,37		Heterochromatin	-
Lmnb2		0,12	0,24	0,30	0,98			Heterochromatin	-
Lrwd1	ORC	0,14		0,11	0,35			Heterochromatin	-
Lsm2	U6 SnRNP		0,18					Unclear	-
Luc7l3				0,04			0	Enhancer	-
Max		0,11	0,11	0,10			1,00	Promoter	yes
Mbtd1		0,06					0	Promoter	-
Mdc1			0,03	0,05			0,60	Enhancer	-
Meaf6	Trrap/Ep400	0,51	0,19	0,19			0,37	Promoter	-
Men1	MLL	0,32	0,03	0,15			0,09	Promoter	-
Mll2	MLL	0,21		0,02		0,01	0,00	Promoter	-
Morc2a			0,05					Unclear	-
Morc3		0,70	0,24	0,24	0,02		0,34	Promoter	-
Mta3	NuRD	0,24	0,14	0,27			0,52	Enhancer	-
Mtf2	PRC2	0,16	0,06	0,22	0,06		0,27	Enhancer	yes
Mycn		0,08	0,08				1,00	Promoter	yes
Myst2	HBO1	0,56	0,36	0,39	0,06		0,64	Promoter	yes
Myst3	MOZ	0,08					0	Promoter	yes
Myst4	MORF	0,04					0	Promoter	-
Nacc1		0,04	0,07	0,13			0,54	Enhancer	yes
Nasp			0,02	0,02			1,00	Enhancer	-
Nfrkb			0,03					Unclear	yes

Protein	Complex[a]	H3K4me3 avrg emPAI	H3K27ac avrg emPAI	H3K4me1 avrg emPAI	H3K9me3 avrg emPAI	GFP avrg emPAI	H3K27ac ratio	Prediction[b]	Pluripotency phenotype[c]
Nolc1				0,05			0	Enhancer	-
Nrf1		0,08					0	Promoter	-
Nsd1		0,01	0,04	0,05			0,80	Enhancer	-
Nudt13				0,10			0	Enhancer	-
Nudt21	CFIm	0,48					0	Promoter	-
Numa1		0,22	0,15	0,29	0,10		0,52	Enhancer	-
Oct4		0,16	0,10	0,21	0,05		0,48	Enhancer	yes
Ogt		0,15	0,02	0,05			0,13	Promoter	yes
Pak1ip1		0,19					0	Promoter	-
Parp2				0,03			0	Enhancer	-
Parp9					0,04			Heterochromatin	-
Patz1		0,05		0,05			0	Promoter	-
Pbrm1	BAF	0,02	0,11	0,11			1,00	Enhancer	-
Pds5a	Wapl		0,03	0,05			0,60	Enhancer	-
Pds5b	Wapl	0,02	0,01	0,02			0,50	Enhancer	-
Phc1	PRC1			0,04			0	Enhancer	-
Phf16	HBO1	0,04					0	Promoter	-
Phf20				0,02			0	Enhancer	yes
Phf23		0,53		0,09			0	Promoter	yes
Phf8		0,23	0,02				0,09	Promoter	-
Phip		0,04	0,07	0,13			0,54	Enhancer	-
Pin1			0,11	0,11			1,00	Enhancer	yes
Pnp		0,20					0	Promoter	-
Polr1c	Pol I	0,3					0	Promoter	-
Polr2a	Pol II	0,15	0,13	0,06			0,87	Promoter	-
Polr2b	Pol II	0,43	0,33	0,07			0,77	Promoter	-
Polr2c	Pol II	0,57	0,16	0,11			0,28	Promoter	-
Polr2e	Pol II	0,85	0,24	0,08			0,28	Promoter	-
Polr2g	Pol II	0,22					0	Promoter	-
Ppp2cb				0,11			0	Enhancer	-
Prdm10		0,03	0,03				1,00	Promoter	-
Prdm2				0,02			0	Enhancer	-
Prrc2c				0,01			0	Enhancer	-
Psma5	Proteasome	0,07		0,07			0	Promoter	-
Ptcd3				0,12			0	Enhancer	-
Qser1		0,03		0,01			0	Promoter	-
Rad21	Cohesin		0,03	0,03			1,00	Enhancer	yes
Rars		0,05					0	Promoter	-
Rbbp4	NuRD/CAF1	1,51	0,62	1,09	0,37	0,12	0,41	Promoter	-
Rbbp5	MLL	0,14					0	Promoter	yes
Rbl1				0,03			0	Enhancer	-
Rbpj		0,30		0,20	0,14		0	Promoter	-
Rfc3		0,05	0,10				2,00	Promoter	-
Rfc4		0,20	0,15	0,19			0,75	Promoter	-
Rnf2	PRC1	0,35		0,30			0	Promoter	yes
Rsf1	RSF	0,05	0,11	0,12			0,92	Enhancer	-
Ruvbl2	Trrap/Ep400	1,36	1,91	0,41	0,16		1,40	Promoter	yes
Samd1		0,26					0	Promoter	-
Sap130	Sin3	0,18					0	Promoter	-
Sap30	Sin3	0,29					0	Promoter	-
Sap30l	Sin3	0,09					0	Promoter	-
Sbno1				0,01			0	Enhancer	-
Sephs1		0,14	0,05				0,36	Promoter	-
Sf1		0,14					0	Promoter	-
Sf3b1				0,04			0	Enhancer	-
Shprh		0,04	0,01	0,03			0,25	Promoter	-
Sin3a	Sin3	1,34	0,15	0,30			0,11	Promoter	yes
Sin3b	Sin3	0,09					0	Promoter	-
Skp1a	SCF ubiquitin ligase	0,45	0,23	0,10			0,51	Promoter	-
Smarca4	BAF	0,14	0,28	0,22	0,03		1,27	Enhancer	yes
Smarcb1	BAF		0,09	0,05			1,80	Enhancer	yes
Smarcc1	BAF	0,10	0,46	0,31	0,07		1,48	Enhancer	yes
Smarcd1	BAF		0,50	0,35			1,43	Enhancer	-
Smc1a	Cohesin	0,08	0,12	0,22	0,06		0,55	Enhancer	yes
Smc6	Smc5/Smc6			0,03			0	Enhancer	yes
Smchd1			0,02	0,03	0,04			Heterochromatin	-
Spin1		0,72	0,13	0,13			0,18	Promoter	-
Srrt		0,04	0,04	0,08			0,50	Enhancer	-
Ssrp1	FACT	0,68	0,93	0,73	0,29	0,13	1,27	Enhancer	-
Suds3	Sin3	0,41					0	Promoter	-
Supt4h2	DSIF	0,34					0	Promoter	yes
Supt5h	DSIF	0,22	0,16	0,06			0,73	Promoter	-

Protein	Complex[a]	H3K4me3 avrg emPAI	H3K27ac avrg emPAI	H3K4me1 avrg emPAI	H3K9me3 avrg emPAI	GFP avrg emPAI	H3K27ac ratio	Prediction[b]	Pluripotency phenotype[c]
Supt6h		0,07	0,09				1,29	Promoter	-
Suv39h2					0,31			Heterochromatin	yes
Suz12	PRC2	0,31	0,12	0,40	0,07		0,30	Enhancer	yes
Taf1	TFIID	0,03					0	Promoter	yes
Taf2	TFIID	0,06	0,02				0,33	Promoter	yes
Taf3	TFIID	0,02	0,02				1,00	Promoter	yes
Taf4a	TFIID	0,11					0	Promoter	yes
Taf5	TFIID	0,05					0	Promoter	yes
Taf6	TFIID	0,20	0,14				0,70	Promoter	yes
Taf7	TFIID	0,16	0,11				0,69	Promoter	yes
Taldo1			0,05	0,11			0,45	Enhancer	-
Tbrg4				0,09			0	Enhancer	-
Tcea1	TFIIS	0,12	0,26	0,05			2,17	Promoter	-
Tcerg1		0,02	0,03	0,02			1,50	Enhancer	-
Tcof1					0,02			Heterochromatin	-
Tead1			0,04	0,04			1,00	Enhancer	-
Tlk2				0,06			0	Enhancer	-
Trim24			0,25	0,04			6,25	Enhancer	-
Trim33			0,08					Unclear	-
Trp53			0,06					Unclear	-
Trrap	Trrap/Ep400	0,05	0,05	0,01			1,00	Promoter	yes
Uba1			0,03	0,03			1,00	Enhancer	-
Ube2h				0,24			0	Enhancer	-
Ubtf		0,79	0,44	0,43			0,56	Promoter	-
Usp48		0,15	0,08	0,21	0,05		0,38	Enhancer	-
Utp14b				0,07			0	Enhancer	-
Vmn2r100		0,02		0,06			0	Enhancer	-
Wdr18	5FMC	0,04		0,12	0,08		0	Enhancer	-
Wdr5	MLL	0,60	0,27	0,17			0,45	Promoter	yes
Wdr55		0,05					0	Promoter	-
Zfp280c		0,05			0,22			Heterochromatin	-
Zfp281		0,04					0	Promoter	yes
Zfp462		0,01			0,02			Heterochromatin	-
Zic5		0,03					0	Promoter	-
Zmynd8	Integrator		0,05	0,04			1,25	Enhancer	-
Znf512		0,13		0,24	0,25			Heterochromatin	-
Zscan10		0,04		0,09			0	Enhancer	yes

[a] Subunit of indicated protein complex

[b] Prediction of genome localization based on our ChIP-MS criteria

[c] ESC pluripotency phenotype (references in supplementary material)

Supplementary Table 2. ChIP-MS identified proteins

	H3K4me3								H3K27ac								H3K4me1								H3K9me3								GFP							
	Experiment 1				Experiment 2				Experiment 1				Experiment 2				Experiment 1				Experiment 2				Experiment 1				Experiment 2				Experiment 1				Experiment 2			
Protein	Mascot	emPAI	Unique pept.[a]	Pept. count[b]	Mascot	emPAI	Unique pept.[a]	Pept. count[b]	Mascot	emPAI	Unique pept.[a]	Pept. count[b]	Mascot	emPAI	Unique pept.[a]	Pept. count[b]	Mascot	emPAI	Unique pept.[a]	Pept. count[b]	Mascot	emPAI	Unique pept.[a]	Pept. count[b]	Mascot	emPAI	Unique pept.[a]	Pept. count[b]	Mascot	emPAI	Unique pept.[a]	Pept. count[b]	Mascot	emPAI	Unique pept.[a]	Pept. count[b]	Mascot	emPAI	Unique pept.[a]	Pept. count[b]
Actl6a	305	0.57	6	25					122	0.16	2	9	81	0.16	2	5	191	0.24	3	9					91	0.16	2	4												
Actr3	56	0.08	1	2																																				
Actr5													79	0.05	1	1																								
Adnp																	97	0.06	3	3					66	0.03	1	1												
Ankhd1																	111	0.19	3	5																				
Anln																													63	0.06	2	2								
Anp32a	75	0.17	2	3	47	0.17	1	2													74	0.36	2	3																
Arid1a									170	0.05	4	4	235	0.08	5	5	202	0.07	5	6	241	0.05	3	3																
Arid4a	79	0.24	2	4																																				
Ash2l	195	0.35	5	15	77	0.14	2	3																																
Atad2b									89	0.06	2	2					101	0.04	2	2	80	0.04	2	2					152	0.07	3	4								
Atrx																									259	0.1	7	11	264	0.07	6	7								
Aurkb					238	0.44	4	18					69	0.2	2	2	70	0.19	2	4					276	0.72	6	14	252	0.57	5	17								
Banf1	107	0.93	2	6	59	0.39	1	2	78	0.93	2	2					103	0.87	2	5	71	0.39	1	3																
Bap18	166	0.79	3	30	119	0.47	2	8																																
Baz1a																	55	0.02	1	2																				
Baz2a	57	0.04	2	2	239	0.07	5	7	363	0.21	10	19	579	0.23	11	23	526	0.21	12	20	468	0.19	9	24	191	0.09	5	8	110	0.04	2	7								
Bend3	196	0.26	5	7	206	0.21	5	5													77	0.08	2	2																
Bptf	422	0.11	10	17	173	0.04	4	10	102	0.03	3	3					114	0.02	2	3	115	0.02	2	2																
Brd1	70	0.06	2	5					57	0.03	1	2	88	0.06	2	2	156	0.08	3	7	142	0.11	4	9																
Brd2	471	0.51	10	45	273	0.23	5	11	527	0.57	11	30	750	0.93	14	53	191	0.17	4	6	231	0.18	4	15																
Brd3	119	0.15	3	5	81	0.09	2	4	278	0.44	7	14	174	0.2	4	11																								
Brd4	72	0.05	2	5					313	0.23	9	15	323	0.18	6	33					132	0.19	3	13																
Brms1	81	0.28	2	2																																				
Brms1l	179	0.6	5	5	177	0.33	3	5																																

	H3K4me3						H3K27ac						H3K4me1						H3K9me3						GFP					
	Experiment 1			Experiment 2			Experiment 1			Experiment 2			Experiment 1			Experiment 2			Experiment 1			Experiment 2			Experiment 1			Experiment 2		
Protein	Mascot	emPAI	Unique pept.[a]	Pept. count[b]	Mascot	emPAI	Unique pept.[a]	Pept. count[b]	Mascot	emPAI	Unique pept.[a]	Pept. count[b]	Mascot	emPAI	Unique pept.[a]	Pept. count[b]	Mascot	emPAI	Unique pept.[a]	Pept. count[b]	Mascot	emPAI	Unique pept.[a]	Pept. count[b]	Mascot	emPAI	Unique pept.[a]	Pept. count[b]	Mascot	emPAI	Unique pept.[a]	Pept. count[b]
Brpf1	233	0.23	7	11	91	0.08	3	3																								
Brpf3					86	0.05	2	5																								
Cad													204	0.08	6	9														
Cald1	61	0.12	2	2																										
Cbx7	50	0.19	1	4									56	0.19	1	1														
Cdc5l													58	0.08	2	2														
Cdc73	82	0.2	3	3					84	0.13	2	2																		
Cdca7									64	0.18	1	2	51	0.18	1	4														
Cdca8					191	0.56	4	20	62	0.25	2	3	56	0.25	2	2	344	1.17	7	24	392	1.17	7	40						
Cdk7					51	0.1	1	1																						
Cdk9									61	0.18	2	2																		
Cep95									45	0.04	1	14	68	0.07	2	8														
Chaf1b									45	0.45	1	1	109	1.04	2	2														
Chd1	672	0.34	15	48	382	0.18	8	29	239	0.16	8	14	212	0.11	6	11	323	0.16	7	28	66	0.04	2	2						
Chd2	177	0.09	4	22									96	0.05	3	7														
Chd8	516	0.23	14	33	199	0.07	4	12																						
Ckb	66	0.18	2	4	60	0.09	1	1																						
Coil																					62	0.06	1	3						
Cpsf6					88	0.13	2	13	131	0.13	2	14																		
Creb1	60	0.22	2	2																										
Cse1l					72	0.07	2	64																						
Csnk2a1	383	1.4	10	19	115	0.37	3	6	106	0.27	3	4	58	0.17	2	3														
Csnk2b	238	1.32	6	14	80	0.32	2	3	135	0.52	3	5	51	0.15	1	2														
Csrp2bp	66	0.08	2	9																										
Ctbp2	452	0.67	7	68	154	0.24	4	8	229	0.44	5	6	309	0.51	6	9					60	0.08	1	2						
Ctcf	87	0.14	3	5									89	0.09	2	4														
Ctr9	108	0.06	2	2					91	0.06	2	2																		

	H3K4me3								H3K27ac								H3K4me1								H3K9me3								GFP							
	Experiment 1				Experiment 2				Experiment 1				Experiment 2				Experiment 1				Experiment 2				Experiment 1				Experiment 2				Experiment 1				Experiment 2			
Protein	Mascot	emPAI	Unique pept.[a]	Pept. count[b]	Mascot	emPAI	Unique pept.[a]	Pept. count[b]	Mascot	emPAI	Unique pept.[a]	Pept. count[b]	Mascot	emPAI	Unique pept.[a]	Pept. count[b]	Mascot	emPAI	Unique pept.[a]	Pept. count[b]	Mascot	emPAI	Unique pept.[a]	Pept. count[b]	Mascot	emPAI	Unique pept.[a]	Pept. count[b]	Mascot	emPAI	Unique pept.[a]	Pept. count[b]	Mascot	emPAI	Unique pept.[a]	Pept. count[b]	Mascot	emPAI	Unique pept.[a]	Pept. count[b]
Cul4b	262	0.26	7	9	79	0.07	2	2	239	0.22	6	8	195	0.14	4	4	229	0.13	4	8	275	0.22	7	11																
Ddb1	349	0.22	8	10	198	0.15	5	5	339	0.29	9	9	317	0.19	6	7	414	0.31	10	15	339	0.26	8	14																
Ddx47																									81	0.15	2	2	103	0.15	2	2								
Ddx54																	68	0.07	2	2																				
Dido1	120	0.04	3	12																																				
Dis3													92	0.1	3	3					84	0.1	3	3																
Dmap1	100	0.14	2	3													69	0.07	1	1																				
Dnajc9					48	0.13	1	3					89	0.27	2	5					137	0.27	2	4																
Dpf2									189	0.36	4	4	57	0.08	1	6	127	0.15	2	2	66	0.08	1	2																
Dppa2	255	0.86	6	33	157	0.36	3	13					69	0.11	1	9	85	0.21	2	6	152	0.36	3	19																
Dppa4	231	0.92	7	17	75	0.24	2	22									59	0.23	2	2	134	0.54	3	21																
Dpy30	50	0.36	1	4																																				
Dut	117	0.64	3	8	58	0.18	1	4	49	0.18	1	2	48	0.18	1	3	53	0.17	2	6	69	0.18	1	5																
Emsy	56	0.03	1	1																																				
Ep300									147	0.06	4	4	135	0.04	3	8																								
Ep400													101	0.03	3	3																								
Erh	71	0.74	2	2																																				
Esrrb									45	0.08	1	1	48	0.08	1	1	139	0.23	3	6	112	0.16	2	4																
Ezh2	99	0.09	2	5	73	0.04	1	2	47	0.04	1	2	98	0.04	1	2	79	0.04	1	2	141	0.14	3	4																
Fam60a	75	0.48	2	2																																				
Fbxl19					75	0.1	2	4																																
Fkbp4																	62	0.14	2	2																				
Gltscr2																	153	0.13	2	2																				
Glyr1	136	0.19	3	6	147	0.19	2	9	181	0.34	4	6	111	0.13	2	4	292	0.32	5	16	172	0.19	3	11	97	0.13	2	2												
Gm53					59	0.37	2	3																																
Gtf2e1																	76	0.15	2	4																				
Gtf2i									70	0.07	2	3					68	0.06	2	4	102	0.07	3	3																

	H3K4me3							H3K27ac							H3K4me1							H3K9me3							GFP						
	Experiment 1					Experiment 2			Experiment 1					Experiment 2			Experiment 1					Experiment 2			Experiment 1					Experiment 2					
Protein	Mascot	emPAI	Unique pept.[a]	Pept. count[b]		Mascot	emPAI	Unique pept.[a]	Pept. count[b]		Mascot	emPAI	Unique pept.[a]	Pept. count[b]		Mascot	emPAI	Unique pept.[a]	Pept. count[b]		Mascot	emPAI	Unique pept.[a]	Pept. count[b]		Mascot	emPAI	Unique pept.[a]	Pept. count[b]		Mascot	emPAI	Unique pept.[a]	Pept. count[b]	
Gtf3c1	55	0.02	1	1		53	0.02	1	1							97	0.03	2	4															
Hat1	88	0.15	2	4		62	0.07	1	2							63	0.07	1	6															
Hcfc1	137	0.09	4	7																														
Hdac1	453	1.65	9	75		239	0.38	5	24		170	0.3	4	9		244	0.48	6	23		87	0.14	2	7		60	0.07	1	3					
Hdac2	222	0.79	5	52																														
Hdgfrp2											67	0.1	2	2		114	0.2	3	6														
Hmga2						82	0.28	1	7		56	0.28	1	1							52	0.28	1	6		87	0.28	1	7					
Hmgb3	52	0.21	1	6												63	0.2	1	6														
Hmgxb4	60	0.39	2	2																														
Hp1bp3	45	0.12	1	3		104	0.39	3	6							142	0.51	4	7		132	0.19	4	8		172	0.27	4	11					
Incenp						152	0.11	3	9												270	0.28	7	14		317	0.24	6	20					
Ing1	52	0.12	1	1																													
Ing2	57	0.11	1	15		57	0.11	1	6																									
Ing3						70	0.08	1	1		74	0.08	1	1																				
Ing4	60	0.28	2	2																	48	0.13	1	1										
Ing5	168	0.87	4	22		71	0.29	1	2		63	0.13	1	1		72	0.27	2	3		73	0.13	1	2										
Ino80	147	0.06	3	5		79	0.04	3	3																									
Ints1																69	0.03	2	17															
Irgc																68	0.08	2	12															
Jarid2	283	0.17	8	13		309	0.14	5	11		128	0.05	2	4		281	0.19	7	11		70	0.03	2	2		104	0.05	2	3					
Kdm1a	94	0.08	2	2		91	0.04	1	1		57	0.04	1	1		86	0.04	1	1		53	0.04	1	1										
Kdm2a	435	0.24	9	43		158	0.08	3	15		66	0.06	2	2		62	0.03	1	3															
Kdm2b	267	0.16	6	13		71	0.02	2	5							64	0.02	1	3															
Kdm4a	67	0.06	2	3																														
Kdm4c	861	0.87	18	86		647	0.43	11	48		119	0.06	2	4		270	0.2	6	16															
Kdm5a	61	0.04	2	2																														
Kdm5b	519	0.3	13	38		424	0.2	9	16		107	0.04	2	3		347	0.18	8	18															

	H3K4me3								H3K27ac								H3K4me1								H3K9me3								GFP							
	Experiment 1						Experiment 2						Experiment 1						Experiment 2						Experiment 1						Experiment 2						Experiment 1						Experiment 2	
Protein	Mascot		emPAI		Unique pept.[a]		Pept. count[b]		Mascot		emPAI		Unique pept.[a]		Pept. count[b]		Mascot		emPAI		Unique pept.[a]		Pept. count[b]		Mascot		emPAI		Unique pept.[a]		Pept. count[b]		Mascot		emPAI		Unique pept.[a]		Pept. count[b]	
Kif2c									70		0.09		2		2																									
Kif4									103		0.05		2		2																									
Kifc5b									126		0.16		4		5				108		0.14		2		3															
Klf16	100		0.31		2		8																																	
Klf5									53		0.07		1		1				72		0.07		1		1															
L3mbtl2	94		0.15		3		3										111		0.14		3		3																	
Las1l																	106		0.12		3		3																	
Lasp1																			57		0.26		1		2															
Lig1	147		0.15		4		4										55		0.03		1		1																	
Lig3									67		0.03		1		1				115		0.1		3		3															
Lin9																	53		0.25		1		1																	
Lmna	96		0.1		2		3		130		0.16		3		4		155		0.15		3		5				451		0.55		9		12		290		0.41		6		8
Lmnb1	424		0.72		10		29		805		1.95		19		39		1278		2.93		23		61		644		0.72		10		29		1483		4.34		28		88		1025		1.95		15		78		7
Lmnb2									170		0.24		4		6		297		0.42		6		13		173		0.17		4		6		701		1.24		15		22		435		0.71		9		30		
Lrwd1									220		0.28		5		13		59		0.05		1		2		93		0.16		3		3		163		0.28		5		9		224		0.42		6		21		
Lsm2																	63		0.36		1		1																	
Luc7l3																	55		0.07		1		1																	
Max	56		0.21		1		9		45		0.21		1		1		72		0.2		1		1																	
Mbtd1	108		0.11		2		3																																	
Mdc1									143		0.06		3		10		100		0.06		3		3		96		0.04		2		6									
Meaf6	130		0.63		3		6		117		0.38		2		4										104		0.38		2		3									
Men1	333		0.45		7		24		106		0.18		3		8		50		0.06		1		2		232		0.25		4		16									
Mll2	575		0.23		16		35		518		0.18		10		21										146		0.04		3		8									
Morc2a									107		0.1		4		4																									
Morc3	865		0.91		21		64		225		0.18		6		8		437		0.3		8		22		141		0.1		3		4								59		0.03		1		2		
Mta3	141		0.28		4		6		171		0.28		5		5										189		0.26		4		4									
Mtf2	74		0.12		2		3		52		0.06		1		1		45		0.06		1		1		123		0.18		3		3		161		0.26		4		6		62		0.12		2		2		

	H3K4me3								H3K27ac								H3K4me1								H3K9me3								GFP							
	Experiment 1				Experiment 2				Experiment 1				Experiment 2				Experiment 1				Experiment 2				Experiment 1				Experiment 2				Experiment 1				Experiment 2			
Protein	Mascot	emPAI	Unique pept.[a]	Pept. count[b]	Mascot	emPAI	Unique pept.[a]	Pept. count[b]	Mascot	emPAI	Unique pept.[a]	Pept. count[b]	Mascot	emPAI	Unique pept.[a]	Pept. count[b]	Mascot	emPAI	Unique pept.[a]	Pept. count[b]	Mascot	emPAI	Unique pept.[a]	Pept. count[b]	Mascot	emPAI	Unique pept.[a]	Pept. count[b]	Mascot	emPAI	Unique pept.[a]	Pept. count[b]	Mascot	emPAI	Unique pept.[a]	Pept. count[b]	Mascot	emPAI	Unique pept.[a]	Pept. count[b]
Mycn	66	0.16	2	3					66	0.16	2	3																												
Myst2	327	0.75	9	30					201	0.36	6	7	296	0.36	5	11													89	0.11	2	2								
Myst3	293	0.1	6	11	167	0.05	3	4																																
Myst4	70	0.04	2	2	81	0.04	2	3																																
Nacc1					47	0.07	1	2					111	0.13	2	2	130	0.19	3	5	52	0.07	1	1																
Nasp													58	0.04	1	1					46	0.04	1	1																
Nfrkb									58	0.05	2	2																												
Nolc1																	57	0.1	2	4																				
Nrf1	66	0.15	2	2																																				
Nsd1	62	0.02	2	3					138	0.05	4	4	96	0.02	2	3	190	0.05	4	9	163	0.04	3	11																
Nudt13																	62	0.19	2	3																				
Nudt21	123	0.96	4	6																																				
Numa1	402	0.13	9	25	339	0.3	6	17	431	0.18	11	17	457	0.11	7	17	685	0.19	13	31	666	0.39	11	27	284	0.1	6	10	105	0.09	2	5								
Oct4	94	0.32	3	6					66	0.2	2	2					170	0.42	4	7					51	0.1	1	2												
Ogt	208	0.2	6	6	183	0.1	3	3					46	0.03	1	1	178	0.1	3	3																				
Pak1ip1																	134	0.37	4	4																				
Parp2																	86	0.05	2	4																				
Parp9																									58	0.08	2	2												
Patz1	107	0.1	3	10													74	0.09	2	3																				
Pbrm1	66	0.04	2	3					220	0.11	6	8	174	0.11	5	9	146	0.08	4	6	218	0.14	6	11																
Pds5a													99	0.05	2	5	162	0.09	4	6																				
Pds5b					78	0.04	2	2					56	0.02	1	1					83	0.04	2	2																
Phc1																					80	0.07	2	2																
Phf16					69	0.08	2	5																																
Phf20																					60	0.03	1	1																
Phf23	148	0.4	2	19	196	0.66	3	24													56	0.18	1	6																
Phf8	345	0.33	7	24	144	0.13	3	8	51	0.04	1	1																												

	H3K4me3								H3K27ac								H3K4me1								H3K9me3								GFP							
	Experiment 1				Experiment 2				Experiment 1				Experiment 2				Experiment 1				Experiment 2				Experiment 1				Experiment 2				Experiment 1				Experiment 2			
Protein	Mascot	emPAI	Unique pept.[a]	Pept. count[b]	Mascot	emPAI	Unique pept.[a]	Pept. count[b]	Mascot	emPAI	Unique pept.[a]	Pept. count[b]	Mascot	emPAI	Unique pept.[a]	Pept. count[b]	Mascot	emPAI	Unique pept.[a]	Pept. count[b]	Mascot	emPAI	Unique pept.[a]	Pept. count[b]	Mascot	emPAI	Unique pept.[a]	Pept. count[b]	Mascot	emPAI	Unique pept.[a]	Pept. count[b]	Mascot	emPAI	Unique pept.[a]	Pept. count[b]	Mascot	emPAI	Unique pept.[a]	Pept. count[b]
Phip	152	0.05	4	7	63	0.02	1	1	134	0.07	4	4	187	0.07	4	6	248	0.1	7	9	295	0.15	7	17																
Pin1													57	0.21	1	3					48	0.21	1	2																
Pnp	129	0.39	3	5																																				
Polr1c	174	0.6	5	8																																				
Polr2a	460	0.18	11	21	321	0.11	6	9	333	0.12	8	9	498	0.14	9	15	129	0.05	3	3	159	0.07	4	5																
Polr2b	709	0.67	18	54	258	0.18	6	14	407	0.34	11	17	421	0.31	10	17	63	0.05	2	2	148	0.08	3	4																
Polr2c	308	0.93	7	16	147	0.21	2	2					158	0.32	3	5					100	0.21	2	2																
Polr2e	252	1.37	6	12	100	0.33	2	2	57	0.33	2	2	71	0.15	1	2					55	0.15	2	2																
Polr2g	114	0.44	2	2																																				
Ppp2cb																					77	0.22	2	2																
Prdm10	63	0.06	2	2									87	0.06	2	5																								
Prdm2																					66	0.04	2	4																
Prrc2c																	65	0.02	2	2																				
Psma5	79	0.14	1	1													45	0.13	1	1																				
Ptcd3																	235	0.24	5	7																				
Qser1	135	0.06	4	4	46	0.02	1	1																																
Rad21													57	0.05	1	1	51	0.05	1	1																				
Rars																	89	0.09	2	2																				
Rbbp4	697	2.02	14	108	476	1	9	71	288	0.62	7	13	304	0.62	6	28	507	1.18	12	25	488	1	9	48	186	0.41	5	8	147	0.23	4	5								
Rbbp5	144	0.27	4	9																																				
Rbl1																	62	0.06	2	3																				
Rbpj	169	0.3	4	8	142	0.3	4	9													222	0.39	5	14					61	0.14	2	2								
Rfc3					71	0.09	1	1					74	0.19	2	3																								
Rfc4	94	0.3	3	4	52	0.09	1	1	103	0.3	3	3					47	0.08	1	1	137	0.3	4	4																
Rnf2	212	0.6	5	12	52	0.1	1	1													176	0.6	4	18																
Rsf1	81	0.07	2	3	76	0.02	1	4	192	0.09	4	7	279	0.12	5	16	127	0.09	3	9	325	0.14	6	19																
Ruvbl2	729	2.3	14	92	240	0.42	5	7	497	1.02	9	28	716	2.8	15	24	291	0.49	6	10	170	0.32	4	4	130	0.32	4	4												

	H3K4me3								H3K27ac								H3K4me1								H3K9me3								GFP							
	Experiment 1				Experiment 2				Experiment 1				Experiment 2				Experiment 1				Experiment 2				Experiment 1				Experiment 2				Experiment 1				Experiment 2			
Protein	Mascot	emPAI	Unique pept.[a]	Pept. count[b]	Mascot	emPAI	Unique pept.[a]	Pept. count[b]	Mascot	emPAI	Unique pept.[a]	Pept. count[b]	Mascot	emPAI	Unique pept.[a]	Pept. count[b]	Mascot	emPAI	Unique pept.[a]	Pept. count[b]	Mascot	emPAI	Unique pept.[a]	Pept. count[b]	Mascot	emPAI	Unique pept.[a]	Pept. count[b]	Mascot	emPAI	Unique pept.[a]	Pept. count[b]	Mascot	emPAI	Unique pept.[a]	Pept. count[b]	Mascot	emPAI	Unique pept.[a]	Pept. count[b]
Samd1	206	0.3	4	22	139	0.22	3	9																																
Sap130	215	0.22	5	14	183	0.14	4	5																																
Sap30	136	0.57	3	9																																				
Sap30l	53	0.18	1	2																																				
Sbno1																	59	0.02	2	3																				
Sephs1	83	0.28	2	3					47	0.09	1	1																												
Sf1	62	0.27	2	2																																				
Sf3b1													150	0.07	3	4																								
Shprh	84	0.04	2	2	75	0.04	2	2	46	0.02	1	2					118	0.06	3	6																				
Sin3a	1717	1.85	36	143	1192	0.82	22	64	278	0.16	5	9	321	0.21	8	8	695	0.38	12	19																				
Sin3b	115	0.09	3	4	118	0.09	3	4																																
Skp1a	110	0.45	2	11	71	0.45	2	5	61	0.45	2	3					61	0.2	1	2																				
Smarca4	430	0.22	9	15	106	0.06	3	5	530	0.25	11	18	742	0.3	7	33	510	0.16	8	13	580	0.27	11	29	56	0.02	1	1	57	0.04	2	2								
Smarcb1									72	0.18	2	2									63	0.09	1	3																
Smarcc1	131	0.09	4	5	137	0.11	3	4	495	0.47	12	15	535	0.45	10	20	317	0.25	9	10	471	0.36	8	15	51	0.03	1	1	98	0.11	3	3								
Smarcd1									280	0.54	7	7	240	0.45	6	8	152	0.34	4	5	197	0.36	5	5																
Smc1a	126	0.11	4	4	56	0.05	2	2	240	0.19	7	7	66	0.05	2	2	452	0.3	11	12	180	0.13	4	6	141	0.11	4	4												
Smc6																	82	0.05	2	2																				
Smchd1									78	0.03	2	2					137	0.05	3	5					119	0.05	3	4	51	0.02	1	2								
Spin1	248	1.31	6	45	62	0.13	1	17	46	0.13	1	4	77	0.13	1	8	48	0.13	1	4	58	0.13	1	9																
Srrt	61	0.07	2	2					91	0.07	3	3					130	0.11	3	3	83	0.04	1	1																
Ssrp1	582	0.93	12	66	474	0.42	7	54	585	0.93	13	39	687	0.93	11	65	685	0.88	13	41	511	0.57	8	88	219	0.22	4	14	296	0.35	6	25	75	0.11	2	3	150	0.15	3	7
Suds3	234	0.6	6	8	105	0.21	2	2																																
Supt4h2	77	0.67	2	4																																				
Supt5h	386	0.31	10	13	265	0.13	4	9	191	0.16	5	5	281	0.16	5	13	72	0.06	2	2	155	0.06	2	4																
Supt6h	207	0.12	6	7	90	0.02	1	1	177	0.1	5	5	249	0.08	4	4																								
Suv39h2																									118	0.22	4	7	167	0.39	4	7								

		H3K4me3						H3K27ac						H3K4me1						H3K9me3						GFP					
		Experiment 1				Experiment 2				Experiment 1				Experiment 2				Experiment 1				Experiment 2				Experiment 1				Experiment 2		
Protein		Mascot	emPAI	Unique pept.[a]	Pept. count[b]	Mascot	emPAI	Unique pept.[a]	Pept. count[b]	Mascot	emPAI	Unique pept.[a]	Pept. count[b]	Mascot	emPAI	Unique pept.[a]	Pept. count[b]	Mascot	emPAI	Unique pept.[a]	Pept. count[b]	Mascot	emPAI	Unique pept.[a]	Pept. count[b]	Mascot	emPAI	Unique pept.[a]	Pept. count[b]	Mascot	emPAI	Unique pept.[a]	Pept. count[b]
Suz12		229	0.37	6	11	251	0.25	5	11	73	0.09	2	2	193	0.14	4	6	294	0.29	6	10	455	0.5	8	19	144	0.14	3	6				
Taf1		100	0.05	3	3																												
Taf2		149	0.12	4	5					47	0.03	1	1									50	0.08	1	1								
Taf3		79	0.04	1	7					68	0.04	1	1									47	0.04	1	1								
Taf4a		144	0.22	4	4																	90	0.07	2	2								
Taf5		65	0.09	2	2																												
Taf6		230	0.29	5	6	136	0.11	2	3	97	0.11	3	3	168	0.17	3	3					52	0.01	1	1								
Taf7		84	0.2	2	5	70	0.11	1	1	48	0.11	1	1	59	0.11	1	2					83	0.06	2	2								
Taldo1														50	0.1	1	1	55	0.21	2	2												
Tbrg4																		67	0.17	2	3												
Tcea1		78	0.23	2	6					133	0.52	4	4					47	0.1	1	2												
Tcerg1						57	0.03	1	1					125	0.06	2	3					63	0.03	1	2								
Tcof1																										75	0.03	1	1				
Tead1														56	0.08	1	2					50	0.08	1	1								
Tlk2																		67	0.08	2	3	47	0.04	1	1								
Trim24										265	0.24	7	7	343	0.25	7	8					90	0.07	2	2								
Trim33										123	0.09	3	3	131	0.06	2	3																
Trp53														52	0.11	1	1																
Trrap		204	0.06	5	5	148	0.04	4	4	98	0.04	3	3	314	0.05	6	9					52	0.01	1	1								
Uba1														64	0.06	2	2					83	0.06	2	2								
Ube2h																		56	0.47	2	3												
Ubtf		732	1.07	19	73	457	0.5	9	34	374	0.5	10	14	359	0.38	8	18	342	0.35	8	14	457	0.5	10	28								
Usp48		195	0.13	4	11	162	0.16	3	7					143	0.16	3	7	457	0.29	10	19	223	0.13	4	18	53	0.05	1	2				
Utp14b																		110	0.13	3	5												
Vmn2r100						59	0.04	2	3													63	0.12	2	3								
Wdr18		52	0.08	1	1													119	0.24	3	3					68	0.08	2	2	49	0.08	1	2
Wdr5		202	0.57	4	22	170	0.62	5	12	86	0.21	2	3	133	0.33	3	10					115	0.33	3	4								

	H3K4me3								H3K27ac								H3K4me1								H3K9me3								GFP							
	Experiment 1				Experiment 2				Experiment 1				Experiment 2				Experiment 1				Experiment 2				Experiment 1				Experiment 2				Experiment 1				Experiment 2			
Protein	Mascot	emPAI	Unique pept.[a]	Pept. count[b]	Mascot	emPAI	Unique pept.[a]	Pept. count[b]	Mascot	emPAI	Unique pept.[a]	Pept. count[b]	Mascot	emPAI	Unique pept.[a]	Pept. count[b]	Mascot	emPAI	Unique pept.[a]	Pept. count[b]	Mascot	emPAI	Unique pept.[a]	Pept. count[b]	Mascot	emPAI	Unique pept.[a]	Pept. count[b]	Mascot	emPAI	Unique pept.[a]	Pept. count[b]	Mascot	emPAI	Unique pept.[a]	Pept. count[b]	Mascot	emPAI	Unique pept.[a]	Pept. count[b]
Wdr55					57	0.09	2	2																																
Zfp280c					116	0.09	4	5																	161	0.24	4	11	190	0.19	4	10								
Zfp281	74	0.08	2	2																																				
Zfp462																	53	0.01	1	1					68	0.03	2	2												
Zic5	64	0.06	1	1																																				
Zmynd8									57	0.03	2	2	101	0.06	2	6					55	0.07	1	7																
Znf512					153	0.25	4	10									160	0.3	5	7	91	0.18	3	3	191	0.32	5	13	161	0.18	3	14								
Zscan10	60	0.08	2	2																	93	0.17	2	2																

[a] Number of unique, non-redundant peptides identified for the specified protein

[b] Number of peptide spectra counted for the specified protein

Supplementary Table 3. ChIP-MS identified complex subunits and predictions

Complex / subunits	H3K4me3 avrg emPAI	H3K27ac avrg emPAI	H3K4me1 avrg emPAI	H3K9me3 avrg emPAI	GFP avrg emPAI	H3K27ac ratio	Prediction[a]
BAF complex	0,04	0,25	0,17	0,01		1,43	Enhancer
Arid1a (Baf250a)		0,07	0,06				Enhancer
Dpf2 (Baf45d)		0,22	0,12				Enhancer
Pbrm1 (Baf180)	0,02	0,11	0,11				Enhancer
Smarca4 (Brg1)	0,14	0,28	0,22	0,03			Enhancer
Smarcb1 (Baf47)		0,09	0,05				Enhancer
Smarcc1 (Baf155)	0,10	0,46	0,31	0,07			Enhancer
Smarcd1 (Baf60a)		0,50	0,35				Enhancer
Sin3 complex	0,32	0,01	0,03			0,04	Promoter
Sin3a	1,34	0,15	0,30				Promoter
Sin3b	0,09						Promoter
Arid4a (Rbbp1)	0,12						Promoter
Brms1	0,14						Promoter
Brms1l	0,47						Promoter
Fam60a	0,24						Promoter
Ing1	0,06						Promoter
Ing2	0,11						Promoter
Sap130	0,18						Promoter
Sap30	0,29						Promoter
Sap30l	0,09						Promoter
Suds3	0,41						Promoter
PRC1 complex	0,15		0,14			0	Promoter
Cbx7	0,10		0,10				Promoter
Phc1			0,04				Enhancer
Rnf2	0,35		0,30				Promoter
PRC2 complex	0,17	0,07	0,23	0,04		0,29	Enhancer
Ezh2	0,07	0,04	0,09				Enhancer
Jarid2	0,16	0,05	0,21	0,04			Enhancer
Mtf2	0,16	0,06	0,22	0,06			Enhancer
Suz12	0,31	0,12	0,40	0,07			Enhancer
TFIID complex	0,09	0,04				0,47	Promoter
Taf1	0,03						Promoter
Taf2	0,06	0,02					Promoter
Taf3	0,02	0,02					Promoter
Taf4a	0,11						Promoter
Taf5	0,05						Promoter
Taf6	0,20	0,14					Promoter
Taf7	0,16	0,11					Promoter
Aurora Kinase complex	0,19	0,08	0,11	0,69			Heterochromatin
Aurkb	0,22	0,10	0,10	0,65			Heterochromatin
Cdca8	0,28	0,13	0,24	1,17			Heterochromatin
Incenp	0,06			0,26			Heterochromatin

Complex / subunits	H3K4me3 avrg emPAI	H3K27ac avrg emPAI	H3K4me1 avrg emPAI	H3K9me3 avrg emPAI	GFP avrg emPAI	H3K27ac ratio	Prediction[a]
Pol II complex	0,44	0,17	0,06			0,39	Promoter
Polr2a	0,15	0,13	0,06				Promoter
Polr2b	0,43	0,33	0,07				Promoter
Polr2c	0,57	0,16	0,11				Promoter
Polr2e	0,85	0,24	0,08				Promoter
Polr2g	0,22						Promoter
MLL complex	0,34	0,04	0,05			0,13	Promoter
Mll2	0,21		0,02		0,01		Promoter
Ash2l	0,25						Promoter
Bap18	0,63						Promoter
Dpy30	0,18						Promoter
Hcfc1	0,05						Promoter
Men1	0,32	0,03	0,15				Promoter
Rbbp5	0,14						Promoter
Wdr5	0,60	0,27	0,17				Promoter
Trrap/Ep400 complex	0,30	0,39	0,09	0,03	0,00	1,33	Promoter
Trrap	0,05	0,05	0,01				Promoter
Ep400		0,02					Unclear
Dmap1	0,07		0,04				Promoter
Ruvbl2	1,36	1,91	0,41	0,16			Promoter
HBO1/MOZ/MORF complex	0,20	0,07	0,09	0,01		0,34	Promoter
Myst2	0,56	0,36	0,39	0,06			Promoter
Myst3	0,08						Promoter
Myst4	0,04						Promoter
Brd1	0,03	0,05	0,10				Enhancer
Brpf1	0,16						Promoter
Brpf3	0,03						Promoter
Meaf6	0,51	0,19	0,19				Promoter
Ing4	0,14		0,07				Promoter
Ing5	0,58	0,13	0,20				Promoter
Phf16	0,04						Promoter
Cohesin complex	0,04	0,07	0,12	0,03		0,60	Enhancer
Rad21		0,03	0,03				Enhancer
Smc1a	0,08	0,12	0,22	0,06			Enhancer
DSIF complex	0,28	0,08	0,03			0,29	Promoter
Supt4h2	0,34						Promoter
Supt5h	0,22	0,16	0,06				Promoter

[a] Prediction of genome localization based on our ChIP-MS criteria

Supplementary Table 4. Dppa2 target genes

Gene symbol	Ensembl gene ID	Tissue/cells with highest expression	FC over Embryonic stem cells
Pdgfb	ENSMUSG00000000489	Macrophage bone marrow 2hr LPS	83.7
Gnmt	ENSMUSG00000002769	Liver	3404.7
Gstm5	ENSMUSG00000004032	Testis	247.1
Taf7l	ENSMUSG00000009596	Placenta	633.1
Slc47a1	ENSMUSG00000010122	Kidney	64.5
Dazl	ENSMUSG00000010592	Testis	27.1
Mov10l1	ENSMUSG00000015365	Testis	26.4
Cd83	ENSMUSG00000015396	Macrophage peri LPS thio 1hrs	2432.4
Nkx2-5	ENSMUSG00000015579	Heart	26.8
Nuak1	ENSMUSG00000020032	MEF	46.2
Sycp3	ENSMUSG00000020059	Testis	19.2
Fgf22	ENSMUSG00000020327	Epidermis	1.8
Nipal4	ENSMUSG00000020411	Epidermis	56.4
Myocd	ENSMUSG00000020542	Umbilical cord	121.9
Cmpk2	ENSMUSG00000020638	Macrophage peri LPS thio 7hrs	2480.9
Rasgrf2	ENSMUSG00000021708	Hypothalamus	160.9
Ddx4	ENSMUSG00000021758	Testis	13.1
Galnt14	ENSMUSG00000024064	Kidney	11.1
Lipo1	ENSMUSG00000024766	Lacrimal gland	33.4
Rin1	ENSMUSG00000024883	Nucleus accumbens	19.1
Tdrd1	ENSMUSG00000025081	Testis	82.6
Syce1	ENSMUSG00000025480	Testis	24.4
1500015O10Rik	ENSMUSG00000026051	Osteoblast day21	2756
Tnfrsf11a	ENSMUSG00000026321	RAW 264 7	79.7
Cybrd1	ENSMUSG00000027015	Stomach	58.2
Bfsp1	ENSMUSG00000027420	Lens	5785.3
Adad1	ENSMUSG00000027719	Testis	1618.9
Sycp1	ENSMUSG00000027855	Testis	108.1
Hormad1	ENSMUSG00000028109	Not Available	Not available
Spaca1	ENSMUSG00000028264	Testis	1382.4
Slc10a4	ENSMUSG00000029219	Not Available	Not available
Figla	ENSMUSG00000030001	Lens	1.7
Mesp2	ENSMUSG00000030543	Testis	3.3
Chst15	ENSMUSG00000030930	Mast cells IgE	73.9
Cryab	ENSMUSG00000032060	Lens	157.1
Itga11	ENSMUSG00000032243	Not Available	Not available
Kank1	ENSMUSG00000032702	Epidermis	61.2
Ptprm	ENSMUSG00000033278	Lung	118.1
Cdcp1	ENSMUSG00000035498	Cornea	36.9
C530008M17Rik	ENSMUSG00000036377	Testis	50.3
Kcna6	ENSMUSG00000038077	Cerebral cortex prefrontal	26.2
Spon1	ENSMUSG00000038156	Lens	391.8
Mpp6	ENSMUSG00000038388	Mast cells IgE	3.6

Gene symbol	Ensembl gene ID	Tissue/cells with highest expression	FC over Embryonic stem cells
Tspyl5	ENSMUSG00000038984	Nucleus accumbens	98
D1Pas1	ENSMUSG00000039224	Testis	6.5
Dusp26	ENSMUSG00000039661	Dorsal root ganglia	95.3
Fkbp6	ENSMUSG00000040013	Testis	12.6
Sec1	ENSMUSG00000040364	Testis	21.3
Mael	ENSMUSG00000040629	Testis	89.3
Rsph6a	ENSMUSG00000040866	Testis	262.4
Prima1	ENSMUSG00000041669	Neuro2a	9.9
Uggt2	ENSMUSG00000042104	Testis	9.2
Abcb4	ENSMUSG00000042476	Liver	190.9
Disc1	ENSMUSG00000043051	Not Available	Not available
Npy5r	ENSMUSG00000044014	Nucleus accumbens	11
Pcsk9	ENSMUSG00000044254	mIMCD.3	20.9
Mettl24	ENSMUSG00000045555	Lung	11.5
Olig3	ENSMUSG00000045591	Thymocyte DP CD4.CD8.	2
She	ENSMUSG00000046280	Lung	19.1
Gm11554	ENSMUSG00000048294	Embryonic stem cells Bruce4+ V26, average	1
Zfp697	ENSMUSG00000050064	Kidney	40.5
Tacstd2	ENSMUSG00000051397	Epidermis	640.2
Kcnf1	ENSMUSG00000051726	Olfactory bulb	263.7
Fes	ENSMUSG00000053158	Mast cells	99.2
Sh2d4a	ENSMUSG00000053886	Stomach	106.8
Kbtbd13	ENSMUSG00000054978	Skeletal muscle	29.6
H2-Q5	ENSMUSG00000055413	T.cells CD8.	783.5
Gulp1	ENSMUSG00000056870	Retinal pigment epithelium	87.5
Unc13d	ENSMUSG00000057948	Mast cells	73.8
Kl	ENSMUSG00000058488	Kidney	1051.7
Kcng2	ENSMUSG00000059852	Heart	80.1
Nkpd1	ENSMUSG00000060621	Stomach	8.2
Rnf217	ENSMUSG00000063760	C2C12	25.6
Zar1	ENSMUSG00000063935	Ovary	4.1
Dmrtc1c2	ENSMUSG00000067561	Not Available	Not available
4930444P10Rik	ENSMUSG00000067795	Testis	249.5
Slc25a31	ENSMUSG00000069041	Testis	27.1
Atxn1l	ENSMUSG00000069895	Ciliary bodies	3.9
Hsf5	ENSMUSG00000070345	Testis	303.5
Zfp783	ENSMUSG00000072653	Retinal pigment epithelium	22.4
Klhl40	ENSMUSG00000074001	Skeletal muscle	253
Tspyl3	ENSMUSG00000074671	T.cells CD8.	38.2
Chst14	ENSMUSG00000074916	X3T3.L1	9.7
Vgll3	ENSMUSG00000091243	Osteoblast day5	16.7

Supplementary Table 5: Accession numbers of studies used for genome-wide correlations

Histone modifications	GEO dataset accession number
H3K27ac	GSE24165
H3K4me1	GSE24165
H3K4me3	GSE24165
H3K9me3	GSE12241
Input	GSE24165

Protein	GEO dataset accession number
Atrx	GSE22162
Brd4	GSE36561
Cbx7	GSE42466
Ctcf	GSE49847
Ctr9	GSE20530
Esrrb	GSE11431
Ezh2	GSE49178
GFP	GSE11431
Hdac1	GSE27844
Hdac2	GSE27844
Jarid2	GSE19708
Kdm1a	GSE27844
Kdm2a	GSE21202
Kdm5b	GSE31968
Mtf2	GSE16526
Mycn	GSE11431
Nanog	GSE44286
Oct4	GSE44286
P300	GSE49847
Polr2a	GSE49847
Rad21	GSE33346
Rbbp5	GSE22934
Rnf2	GSE26680
Smarca4	GSE14344
Smc1a	GSE22557
Supt5h	GSE20485
Suz12	GSE48122
Taf1	GSE31270
Taf3	GSE30959
Wdr5	GSE22934

Protein	Bioproject accession number
Tcea1	PRJEB2674

Supplementary Dataset
Excel table Available at:
http://www.nature.com/ncomms/2015/150520/ncomms8155/extref/ncomms8155-s2.xlsx

References for pluripotency phenotype of factors identified by ChIP-MS

Arid1a[1], Ash2l[2], Aurkb[3], Banf1[4], Bptf1[5], Cbx7[6], Cdc73[7], Cdk9[8], Chaf1b[9], Chd1[10], Ctbp2[11], Ctr9[8,12], Dmap1[13], Dppa2[14], Dppa4[14], Dpy30[15], Ep300[16], Ep400[13], Esrrb[17], Ezh2[18], Hcfc1[19], Hdac1[20], Ing5[8], Ino80[19], Jarid2[21], Kdm1a[22], Kdm2b[23], Kdm4c[24], Kdm5b[25], Klf5[26], L3mbtl2[27], Max[28], Mtf2[29], Mycn[30], Myst2[31], Myst3[31], Nacc1[32], Nfrkb[19], Oct4[33], Ogt[34], Phf20[35], Phf23[12], Pin1[36], , Rad21[37], Rbbp5[15], Rnf2[8,13,37], Ruvbl2[13], Sin3a[38], Smarca4[39], Smarcb1[40], Smarcc1[13], Smc1a[8,13], Smc6[12,13], Supt4h2[13], Suv39h2[19], Suz12[41], Taf1[42], Taf2[42], Taf4a[42], Taf5[42], Taf6[42], Taf3[42,43] , Taf7[19], Tpr[13,19], Trrap[13], Wdr5[44], Zfp281[32,45], Zscan10[46].

1. Gao, X. et al. ES cell pluripotency and germ-layer formation require the SWI/SNF chromatin remodeling component BAF250a. *Proc Natl Acad Sci U S A* **105**, 6656-61 (2008).
2. Wan, M. et al. The trithorax group protein Ash2l is essential for pluripotency and maintaining open chromatin in embryonic stem cells. *J Biol Chem* **288**, 5039-48 (2013).
3. Lee, D.F. et al. Regulation of embryonic and induced pluripotency by aurora kinase-p53 signaling. *Cell Stem Cell* **11**, 179-94 (2012).
4. Cox, J.L. et al. Banf1 is required to maintain the self-renewal of both mouse and human embryonic stem cells. *J Cell Sci* **124**, 2654-65 (2011).
5. Landry, J. et al. Essential role of chromatin remodeling protein Bptf in early mouse embryos and embryonic stem cells. *PLoS Genet* **4**, e1000241 (2008).
6. O'Loghlen, A. et al. MicroRNA regulation of Cbx7 mediates a switch of Polycomb orthologs during ESC differentiation. *Cell Stem Cell* **10**, 33-46 (2012).
7. Wang, P. et al. Parafibromin, a component of the human PAF complex, regulates growth factors and is required for embryonic development and survival in adult mice. *Mol Cell Biol* **28**, 2930-40 (2008).
8. Hu, G. et al. A genome-wide RNAi screen identifies a new transcriptional module required for self-renewal. *Genes Dev* **23**, 837-48 (2009).
9. Bilodeau, S., Kagey, M.H., Frampton, G.M., Rahl, P.B. & Young, R.A. SetDB1 contributes to repression of genes encoding developmental regulators and maintenance of ES cell state. *Genes Dev* **23**, 2484-9 (2009).
10. Gaspar-Maia, A. et al. Chd1 regulates open chromatin and pluripotency of embryonic stem cells. *Nature* **460**, 863-8 (2009).
11. Tarleton, H.P. & Lemischka, I.R. Delayed differentiation in embryonic stem cells and mesodermal progenitors in the absence of CtBP2. *Mech Dev* **127**, 107-19 (2010).
12. Ding, L. et al. A genome-scale RNAi screen for Oct4 modulators defines a role of the Paf1 complex for embryonic stem cell identity. *Cell Stem Cell* **4**, 403-15 (2009).
13. Fazzio, T.G., Huff, J.T. & Panning, B. An RNAi screen of chromatin proteins identifies Tip60-p400 as a regulator of embryonic stem cell identity. *Cell* **134**, 162-74 (2008).
14. Maldonado-Saldivia, J. et al. Dppa2 and Dppa4 are closely linked SAP motif genes restricted to pluripotent cells and the germ line. *Stem Cells* **25**, 19-28 (2007).
15. Jiang, H. et al. Role for Dpy-30 in ES cell-fate specification by regulation of H3K4 methylation within bivalent domains. *Cell* **144**, 513-25 (2011).
16. Zhong, X. & Jin, Y. Critical roles of coactivator p300 in mouse embryonic stem cell differentiation and Nanog expression. *J Biol Chem* **284**, 9168-75 (2009).
17. Ivanova, N. et al. Dissecting self-renewal in stem cells with RNA interference. *Nature* **442**, 533-8 (2006).
18. Shen, X. et al. EZH1 mediates methylation on histone H3 lysine 27 and complements EZH2 in maintaining stem cell identity and executing pluripotency. *Mol Cell* **32**, 491-502 (2008).
19. Chia, N.Y. et al. A genome-wide RNAi screen reveals determinants of human embryonic stem cell identity. *Nature* **468**, 316-20 (2010).
20. Dovey, O.M., Foster, C.T. & Cowley, S.M. Histone deacetylase 1 (HDAC1), but not HDAC2, controls embryonic stem cell differentiation. *Proc Natl Acad Sci U S A* **107**, 8242-7 (2010).
21. Shen, X. et al. Jumonji modulates polycomb activity and self-renewal versus differentiation of stem cells. *Cell* **139**, 1303-14 (2009).
22. Whyte, W.A. et al. Enhancer decommissioning by LSD1 during embryonic stem cell differentiation. *Nature* **482**, 221-5 (2012).
23. He, J. et al. Kdm2b maintains murine embryonic stem cell status by recruiting PRC1 complex to CpG

islands of developmental genes. *Nat Cell Biol* **15**, 373-84 (2013).
24. Loh, Y.H., Zhang, W., Chen, X., George, J. & Ng, H.H. Jmjd1a and Jmjd2c histone H3 Lys 9 demethylases regulate self-renewal in embryonic stem cells. *Genes Dev* **21**, 2545-57 (2007).
25. Xie, L. et al. KDM5B regulates embryonic stem cell self-renewal and represses cryptic intragenic transcription. *EMBO J* **30**, 1473-84 (2011).
26. Ema, M. et al. Kruppel-like factor 5 is essential for blastocyst development and the normal self-renewal of mouse ESCs. *Cell Stem Cell* **3**, 555-67 (2008).
27. Qin, J. et al. The polycomb group protein L3mbtl2 assembles an atypical PRC1-family complex that is essential in pluripotent stem cells and early development. *Cell Stem Cell* **11**, 319-32 (2012).
28. Hishida, T. et al. Indefinite self-renewal of ESCs through Myc/Max transcriptional complex-independent mechanisms. *Cell Stem Cell* **9**, 37-49 (2011).
29. Walker, E. et al. Polycomb-like 2 associates with PRC2 and regulates transcriptional networks during mouse embryonic stem cell self-renewal and differentiation. *Cell Stem Cell* **6**, 153-66 (2010).
30. Chappell, J., Sun, Y., Singh, A. & Dalton, S. MYC/MAX control ERK signaling and pluripotency by regulation of dual-specificity phosphatases 2 and 7. *Genes Dev* **27**, 725-33 (2013).
31. Kagey, M.H. et al. Mediator and cohesin connect gene expression and chromatin architecture. *Nature* **467**, 430-5 (2010).
32. Wang, J. et al. A protein interaction network for pluripotency of embryonic stem cells. *Nature* **444**, 364-8 (2006).
33. Niwa, H., Miyazaki, J. & Smith, A.G. Quantitative expression of Oct-3/4 defines differentiation, dedifferentiation or self-renewal of ES cells. *Nat Genet* **24**, 372-6 (2000).
34. Shafi, R. et al. The O-GlcNAc transferase gene resides on the X chromosome and is essential for embryonic stem cell viability and mouse ontogeny. *Proc Natl Acad Sci U S A* **97**, 5735-9 (2000).
35. Zhao, W. et al. Jmjd3 inhibits reprogramming by upregulating expression of INK4a/Arf and targeting PHF20 for ubiquitination. *Cell* **152**, 1037-50 (2013).
36. Nishi, M. et al. A distinct role for Pin1 in the induction and maintenance of pluripotency. *J Biol Chem* **286**, 11593-603 (2011).
37. Nitzsche, A. et al. RAD21 cooperates with pluripotency transcription factors in the maintenance of embryonic stem cell identity. *PLoS One* **6**, e19470 (2011).
38. McDonel, P., Demmers, J., Tan, D.W., Watt, F. & Hendrich, B.D. Sin3a is essential for the genome integrity and viability of pluripotent cells. *Dev Biol* **363**, 62-73 (2012).
39. Kidder, B.L., Palmer, S. & Knott, J.G. SWI/SNF-Brg1 regulates self-renewal and occupies core pluripotency-related genes in embryonic stem cells. *Stem Cells* **27**, 317-28 (2009).
40. You, J.S. et al. SNF5 is an essential executor of epigenetic regulation during differentiation. *PLoS Genet* **9**, e1003459 (2013).
41. Pasini, D., Bracken, A.P., Hansen, J.B., Capillo, M. & Helin, K. The polycomb group protein Suz12 is required for embryonic stem cell differentiation. *Mol Cell Biol* **27**, 3769-79 (2007).
42. Pijnappel, W.W. et al. A central role for TFIID in the pluripotent transcription circuitry. *Nature* **495**, 516-9 (2013).
43. Liu, Z., Scannell, D.R., Eisen, M.B. & Tjian, R. Control of embryonic stem cell lineage commitment by core promoter factor, TAF3. *Cell* **146**, 720-31 (2011).
44. Ang, Y.S. et al. Wdr5 mediates self-renewal and reprogramming via the embryonic stem cell core transcriptional network. *Cell* **145**, 183-97 (2011).
45. Fidalgo, M. et al. Zfp281 functions as a transcriptional repressor for pluripotency of mouse embryonic stem cells. *Stem Cells* **29**, 1705-16 (2011).
46. Yu, H.B., Kunarso, G., Hong, F.H. & Stanton, L.W. Zfp206, Oct4, and Sox2 are integrated components of a transcriptional regulatory network in embryonic stem cells. *J Biol Chem* **284**, 31327-35 (2009).

Chapter 3

An interaction network of mental disorder proteins in neural stem cells

Maaike J. Moen[1], Hieab H.H. Adams[1]*, **Johannes H. Brandsma**[1]*, Dick H.W. Dekkers[2], Umut Akinci[1], Sofia Karkampouna[1†], Christel E.M. Kockx[3], Zeliha Ozgür[3], Wilfred F.J. van IJcken[3], Jeroen Demmers[2], Raymond A. Poot[1]

[1] Department of Cell Biology, Erasmus MC, Wytemaweg 80, 3015 CN Rotterdam, The Netherlands
[2] Center for Proteomics, Erasmus MC, The Netherlands
[3] Center for Biomics, Erasmus MC, The Netherlands
*** These authors contributed equally to this work**
[†] Present address: Department of Molecular Cell Biology, Leiden University Medical Center, Leiden, The Netherlands

Manuscript submitted

ABSTRACT

Mental disorders (MDs) such as intellectual disability (ID), autism spectrum disorders (ASD) and schizophrenia have a strong genetic component. A substantial fraction of MD-associated genes encode transcriptional regulators. Here, we biochemically purified several transcriptional regulators from neural stem cells to identify a protein interaction network containing over 200 proteins, including 68 MD-associated proteins and 52 proteins encoded by evolutionarily constrained genes. Our network shows the molecular connections between established MD proteins and provides a discovery tool for novel MD genes. Network proteins preferentially co-localize on the genome and cooperate in disease-relevant gene regulation. Our results suggest that the observed transcriptional regulators associated with ID, ASD or schizophrenia are part of a transcriptional network in neural stem cells. We find that more severe mutations in network proteins are associated with MDs that include lower IQ, suggesting that the level of disruption of a shared transcriptional network correlates with cognitive dysfunction.

INTRODUCTION

Mental disorders (MDs) include neurodevelopmental disorders such as intellectual disability (ID) and autism spectrum disorders (ASD), as well as psychiatric disorders such as schizophrenia[1]. ID, ASD and schizophrenia were shown to have a strong genetic component[2-4]. Recently, many *de novo* gene mutations associated with these MDs have been identified by high-throughput sequencing approaches[5-9]. A substantial fraction of MD-associated mutations are in genes encoding proteins involved in transcriptional regulation and chromatin modification[5-7,10,11]. For example, out of 40 genes that were recently found to be *de novo* mutated in multiple ASD patients[6,7], which are therefore strong ASD gene candidates, 22 genes encode transcription factors or chromatin modifiers. It is often unclear to what extent MD-associated transcriptional regulators act together in the same gene regulatory networks and molecular pathways. Such information is important to appreciate the level of shared etiology within a clinically-defined MD. If cooperating transcriptional regulators are associated with different MDs, it may indicate a molecular relation between these MDs. One important predictor of cooperation between transcriptional regulators is their physical interaction. We and others previously showed that interacting transcriptional regulators in embryonic stem cells and neural stem cells co-localize on the genome and depend on each other for genome-binding[12-15], suggesting their cooperation in gene regulation. Indeed, for Sox2 and its interactor Chd7, we could show that they cooperate to regulate genes relevant for the malformations associated with their respective syndromes in humans[15]. Here, we purified several transcription factors from neural stem cells (NSCs) and identified their interaction partners by mass spectrometry. The starting transcription factors Tcf4, Olig2, Npas3 and Sox2 were selected based on their high expression in NSCs, suggesting their relevance for NSC biology. Indeed, a function in NSC biology was shown for Olig2[16], Npas3[17] and Sox2[18]. Tcf4, Olig2, Npas3 and Sox2 are in different degrees connected to mental disorders. *TCF4* haploinsufficiency causes Pitt Hopkins syndrome, which features severe ID, lack of speech, microcephaly and breathing abnormalities[19,20]. Moreover, several SNPs in the *TCF4* locus are genetic risk factors for developing schizophrenia[4,21]. *OLIG2* is triploid in Down syndrome patients. Restoring diploid gene dose for *Olig2* and *Olig1* in a mouse

model for Down syndrome showed recovery of the normal balance of inhibitory and excitatory neuronal activity[22]. *NPAS3* mutations co-segregate with schizophrenia in two families[23,24]. *SOX2* mutations cause an Anophthalmia syndrome with associated cognitive defects in about half of the cases[25,26]. The resulting interaction network of the starting four transcription factors and their interaction partners contains 206 proteins. We find that the network contains many proteins mutated in patients with ID, ASD or schizophrenia, as well as proteins encoded by evolutionarily constrained genes. We provide evidence that transcriptional regulators associated with ID, ASD or schizophrenia can be part of the same transcription network and that within this network mutation-severity strongly correlates with the level of cognitive dysfunction.

RESULTS

Identification of a protein interaction network in NSCs

We recently improved the FLAG-tag affinity protocol to purify transcription factors and their interacting proteins with high efficiency and low background[12,15]. The accuracy of interaction partner identification by this protocol was extensively validated by independent immunoprecipitations[12,15]. Importantly, many identified interactions were shown by us (Table S1) and others (Table S2) to be biologically relevant and uncover novel functions of the target protein or provide insight into the molecular cause of malformations associated with human syndromes[15]. Here, we applied this protocol to purify transcription factors Tcf4, Olig2 and Npas3 from mouse NSCs in which they are highly expressed (Table S3). NSC lines with stable expression of FLAG-tagged Tcf4, Olig2 or Npas3 were grown to large scale and two or three independent purifications of the FLAG-tagged proteins were performed. Interacting proteins, identified by mass spectrometry, present in at least two purifications of the target protein were included (Tables S4, S5 and S6, see Experimental Procedures for inclusion criteria) and combined with the interaction partners of previously purified Sox2[15] (Table S7) into a protein interaction network of 206 proteins and 401 protein-protein interactions (Fig. 1a, Table S8). The interaction network contains multiple chromatin modifying complexes, such as NuRD, SWI-SNF and Ncor, and transcription factors such as Rfx3, Sall3, Olig2 and Sox2 that interact with all four purified transcription factors (Fig. 1a). However, other identified interaction partners were found to be specific for one purified transcription factor, such as Ascl1, Neurod1, Kdm6a (Tcf4), Satb2, Yy1, Adnp (Olig2), Arnt2, Lmo7 (Npas3) and Xpo4 (Sox2) (Fig. 1a). Our network contains far more interaction partners than can be obtained from protein-protein interaction databases. For example, the four transcription factors that seeded our network of 206 proteins give a database-derived interaction network of only 29 proteins (Fig. S1).

The interaction network is enriched for proteins associated with ID, ASD or schizophrenia and proteins encoding evolutionarily constrained genes

We investigated whether the proteins in our interaction network are associated with ID, ASD and/or schizophrenia. Using a list of 528 known ID genes[5], which are mutated in at least 5 ID patients, we find 26 network proteins encoded by ID genes (Fig. 1a, Table 1). Taking into account only NSC-expressed genes, 6.4 network proteins encoded by ID genes would be expected in a random overlap, enrichment p-value $4.2*10^{-9}$. A recent exome sequencing study identified *de novo* loss-of-function (LOF; frameshift, nonsense, splice-site) mutations or missense mutations in 1584 genes specifically in patients

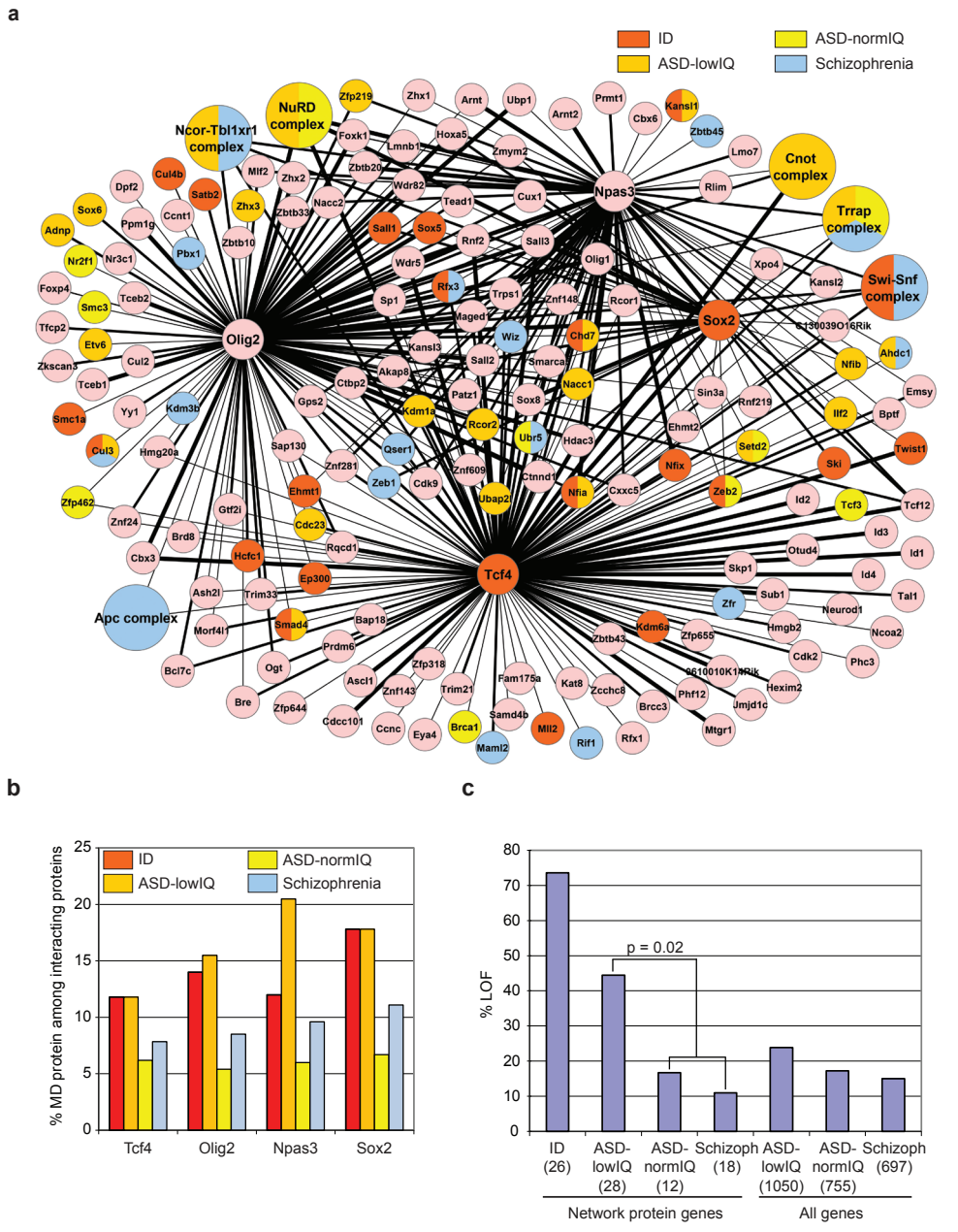


Figure 1. Protein interaction network in neural stem cells. (a) Interaction network representing proteins present in two or more purifications of FLAG-tagged Tcf4, Olig2, Npas3 or Sox2 from neural stem cells. Protein complexes are larger circles, thickness of the edges (black lines) gives an indication of average protein quantity in samples of FLAG-tagged transcription factor with thickest edges; emPAI ≥ 0.6, medium thick edge; emPAI<0.6 and ≥0.2, thin edge; emPAI <0.2. Red color indicates network protein or protein complex subunit(s) encoded by a known ID gene. Orange, yellow and blue color indicate de novo mutation(s) in patients with ASD-lowIQ, ASD-normIQ and schizophrenia, respectively. **(b)** Percentage MD-associated proteins among interaction partners of Tcf4, Olig2, Npas3 and Sox2. MD-categories ID, ASD-

Legend continues on the bottom of the next page

with an ASD including lower IQ (≤90, ASD-lowIQ) or including normal IQ (>90, ASD-normIQ)[6]. The interaction network contains 36 proteins encoded by such putative ASD-associated genes with in total 42 mutations (Fig. 1a, Table 1). A random overlap with the network, corrected by dnenrich[9] for gene length, sequence context and expression in neural stem cells, would expect 31.5 of such mutations in the network (enrichment p-value 0.04). We curated a list of 662 genes with published LOF or missense *de novo* mutations in schizophrenia patients (Table S9) and find 18 network proteins encoded by putative schizophrenia-associated genes with 18 mutations (Fig. 1a, Table 1)**,** where the corrected expectation would be an overlap of 12.2 mutations (p-value 0.07). In total, the network contains 68 proteins associated with ID, ASD and/or schizophrenia and these 68 MD-associated proteins have 260 interactions with other proteins in the network (Fig. 1a, Tables 1, S4, S5, S6 and S7). We identified 47 interactions between ID-associated network proteins and network proteins associated with ASD or schizophrenia (Fig. 1a, Table S10)**.**

As the starting transcription factors (Tcf4, Olig2, Npas3, Sox2) have different associations with MDs, one may expect differences in the percentage of proteins associated with ID, ASD or SZ in their respective interactomes. However, we find no evidence a strong bias in any of the interactomes (Fig. 1b). For example, Tcf4, a transcription factor with strong links to ID, does not have a higher percentage of ID proteins or other categories of MD proteins in its interactome, as compared to the other starting transcription factors (Fig. 1b). Together, this shows that the network as a whole is enriched for MD proteins but the different categories of MD proteins appear homogenously distributed in the network and independent of the currently known strength of MD-association of the starting transcription factor.

The observed enrichments for MD proteins may suggest that our network has an above-average content of proteins encoded by as yet undiscovered MD genes. To find an additional indication of such potential, we overlapped our network with a recently reported list of 1003 genes that are significantly devoid of missense variants in the human population and are likely to be evolutionarily constrained[27] (Fig. 2a). Mutations in these genes were suggested to more likely cause disease, including ASD[27], and indeed we find that genes mutated in patients with ID, ASD-lowIQ, ASD-normIQ or schizophrenia are between 2- and 4-fold enriched for constrained genes (Fig. 2b). We find that a quarter (52 proteins) of the proteins in our interaction network are encoded by constrained genes (Fig. 2a,b, Table S8). An NSC expression-corrected random expectation for this overlap is 12.2 proteins, p-value $1.9*10^{-18}$. Proteins encoded by constrained genes are still enriched in the network after removal of MD-associated proteins (21 observed, NSC expression-corrected expectation 7.9, p-value $6.7*10^{-5}$, Fig. 2b). Remarkably, in each of the four MD categories around 50% of the network proteins with mutations in MD patients are encoded by constrained genes, 10-fold higher than a random overlap (Fig.

(Figure 1. Legend continues from previous page)
lowIQ, ASD-normIQ and schizophrenia are indicated by red, orange, yellow and blue color, respectively. **(c)** Percentage loss-of-function (LOF) mutations in genes mutated in patients with the indicated MD. Network protein genes have the equivalent mouse protein present in the interaction network. Total number of mutations in each category is between brackets, except for ID genes; the number of genes with the equivalent mouse protein in the interaction network is between brackets. p-value of difference LOF ASD-lowIQ vs. LOF ASD-normIQ + LOF schizophrenia was calculated by Fisher's exact test.

Table 1. Network proteins with mutations in the equivalent human gene in mental disorder patients. Human equivalent genes of network proteins with mutations in patients with the indicated mental disorders are listed. Predominant type of gene mutations, LOF or missense, in ID patients is listed. Type and number (if more than one) of gene mutations in patients with ASD-lowIQ, patients with ASD-normIQ or patients with schizophrenia are listed. Evolutionarily constrained genes are indicated. Network proteins with a missense mutation in the human equivalent gene in patients with ASD-normIQ or schizophrenia and a LOF mutation in patients with ID or ASD-lowIQ are marked by >. The opposite pattern is not observed.

Human homologue of network protein	ID	ASD-lowIQ	ASD-normIQ	Schizophrenia	Constrained gene	Severity mutation vs. IQ in MD
TCF4	LOF				X	
ZEB2	LOF		missense		X	>
SMC1A	missense				X	
SATB2	LOF				X	
SOX2	LOF				X	
CHD7	LOF	missense			X	
RFX3	LOF			missense		>
HCFC1	missense				X	
CUL3	LOF	LOF		missense	X	>
SALL1	LOF					
MLL2	LOF				X	
EHMT1	LOF					
SOX5	LOF					
KANSL1	LOF	missense				
EP300	LOF			missense		>
TWIST1	LOF					
KDM6A	LOF					
NFIA	LOF	LOF				
SKI	missense					
NFIX	LOF				X	
SMAD4	missense	missense				
ARID1A	LOF					
SMARCE1	missense					
SMARCB1	missense				X	
SMARCA4	missense				X	
CUL4B	LOF					
ADNP		LOF 2x				
AHDC1		LOF, missense		missense	X	>
SETD2		LOF	missense			>
TBL1XR1		LOF, missense				
UBAP2L		LOF			X	
UBR5			LOF	missense	X	
BRCA1			LOF			
CNOT3		LOF			X	
ILF2		LOF				
NFIB		LOF				
CDC23		LOF				
NACC1		LOF			X	
SOX6		missense				
ZNF219		missense			X	
TRRAP		missense 2x	missense	missense	X	
CHD4		missense			X	
EP400			missense		X	
NCOR1		missense			X	
NR2F1			missense		X	
ZHX3		missense				
ZNF462			missense		X	
KDM1A		missense			X	
RUVBL1			missense			
HDAC1		missense				
MBD2			missense			
CNOT1		missense			X	
RCOR2		missense				
SMC3			missense		X	
TCF3			missense			
ETV6		missense				
ZEB1				LOF		
SMARCC2				LOF	X	
ANAPC5				missense		
WIZ				missense	X	
KDM3B				missense	X	
ZFR				missense		
MAML2				missense		
QSER1				missense		
RIF1				missense		
NCOR2				missense		
ZBTB45				missense	X	
PBX1				missense		

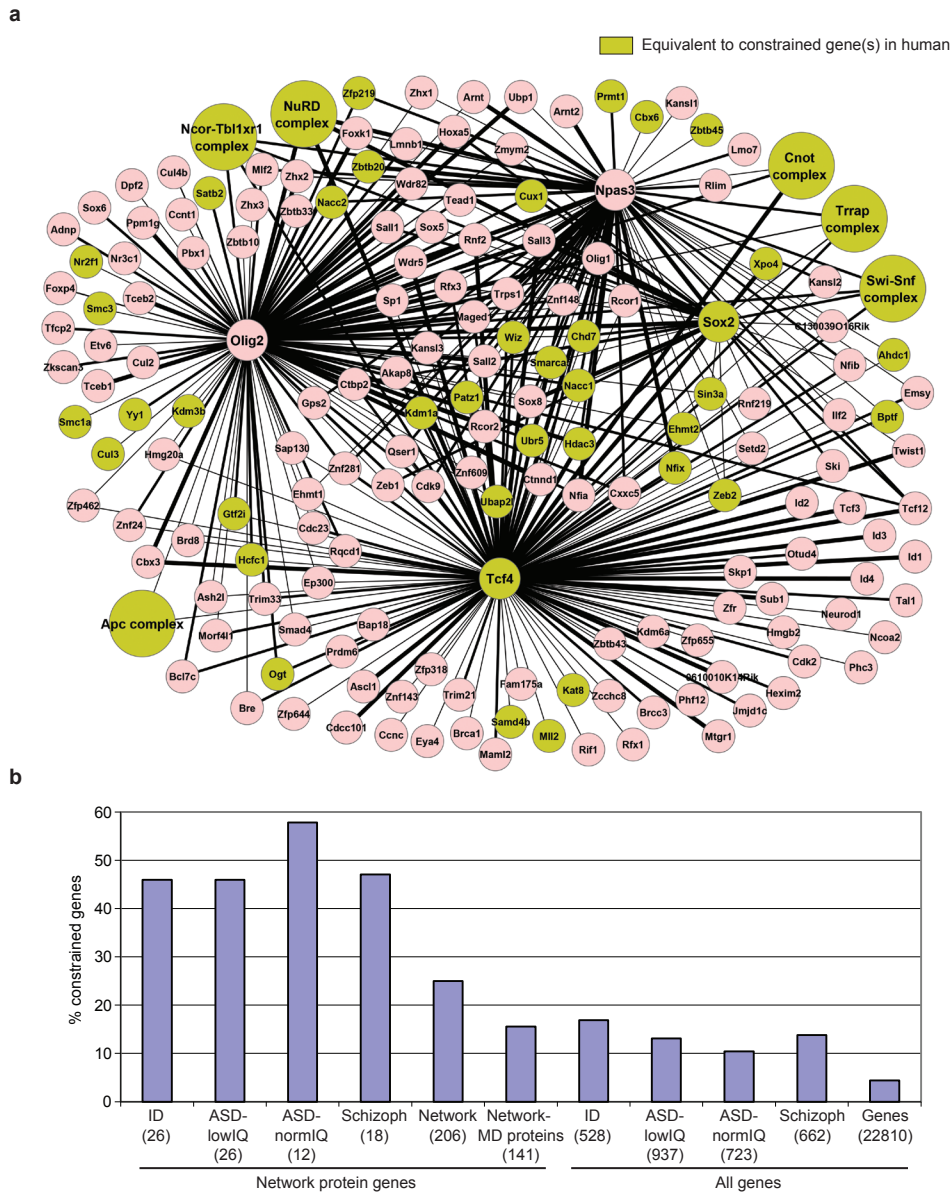


Figure 2. Overlap protein interaction network with constrained human genes. (a) Protein interaction network in neural stem cells. Green color indicates network protein or protein complex subunit(s) encoded by constrained gene(s) in humans. **(b)** Percentage overlap of the indicated categories with constrained genes. Between brackets is the number of human genes equivalent to network proteins in each category (Network protein genes) or total number of genes in each category (All genes).

2b, Table S8). The observed enrichments suggest that network proteins, in particular those encoded by constrained genes, would be good candidates for mutation screening in patients to identify novel MD genes.

Mutation severity in network proteins correlates with cognition levels in the associated MD

Cognitive ability is more affected in patients with ID or ASD-lowIQ than in patients with ASD-normIQ or schizophrenia. LOF mutations, on average, affect gene function more severely than missense mutations. We calculated the percentage of LOF mutations in human genes equivalent to network proteins in the four MD categories. Mutations that cause ID (often as part of ID syndromes) are predominantly LOF mutations for the majority (67%) of the network proteins associated with ID. 43% of the network protein mutations in patients with ASD-lowIQ are LOF mutations, whereas only 17% of the network protein mutations in ASD-normIQ patients and 11% of the network protein mutations in schizophrenia patients are LOF mutations (Fig. 1c, Table 1). Percentages LOF mutations in the complete gene datasets are less distinct between ASD-lowIQ (24%), ASD-normIQ (18%) and schizophrenia (15%, Fig. 1c). One explanation for the exaggerated differences between MD categories in percentages LOF in the network could be that mutations in network protein genes more likely contribute to the MD, as suggested by the high overlap of mutated network proteins with evolutionarily constrained genes (Fig. 2b). In this scenario, the total sets of genes in the different MD categories (Fig. 1c) would contain higher frequencies of non-contributing mutations, which do not have a severity bias. We also find a mutation bias in network proteins with multiple *de novo* mutations, associated with different MDs; six network proteins have missense mutations in patients with ASD-normIQ or schizophrenia and LOF mutations in patients with ID or ASD-lowIQ, whereas the opposite pattern is not observed (Table 1). Together this suggests that particularly in network proteins, the severity of mutations increases in MDs that include lower IQ.

Network transcription factors preferentially co-localize on the genome and cooperate in disease-relevant gene regulation

Having identified an interaction network of MD-related proteins, we investigated whether network proteins preferentially overlap in their genome-wide binding sites, as a proxy for their cooperation in gene regulation[13,28]. We determined the genome-wide binding sites in NSCs for network transcription factors Tcf4, Olig2, Smad4 and Npas3 by ChIP-seq and added published data for Chd7, Sox2, Ascl1 and p300. We also included published data on Brn2 and Max, two transcription factors with NSC expression levels similar to the tested network transcription factors (Table S3) but not part of our interaction network. We find that overlaps in binding sites between transcription factors within the network are on average higher than with Max or Brn2 (Fig. 3a,b). Overlaps above 30% are only observed within the network (Fig. 3a,b) and overlaps above 35% are only observed between network transcription factors that interact with each other (Fig. 3a,c and Fig. 1a).

We subsequently explored whether the interaction network can provide gene regulatory explanations for disease overlap. We started with Tcf4 and Ascl1, two interacting proteins with the highest observed overlap in binding sites (50%, Fig. 3a). Mutations in *ASCL1* cause Congenital Central Hypoventilation Syndrome (CCHS,[29]), which includes breathing abnormalities and Hirschprung disease, two features also observed in Pitt Hopkins

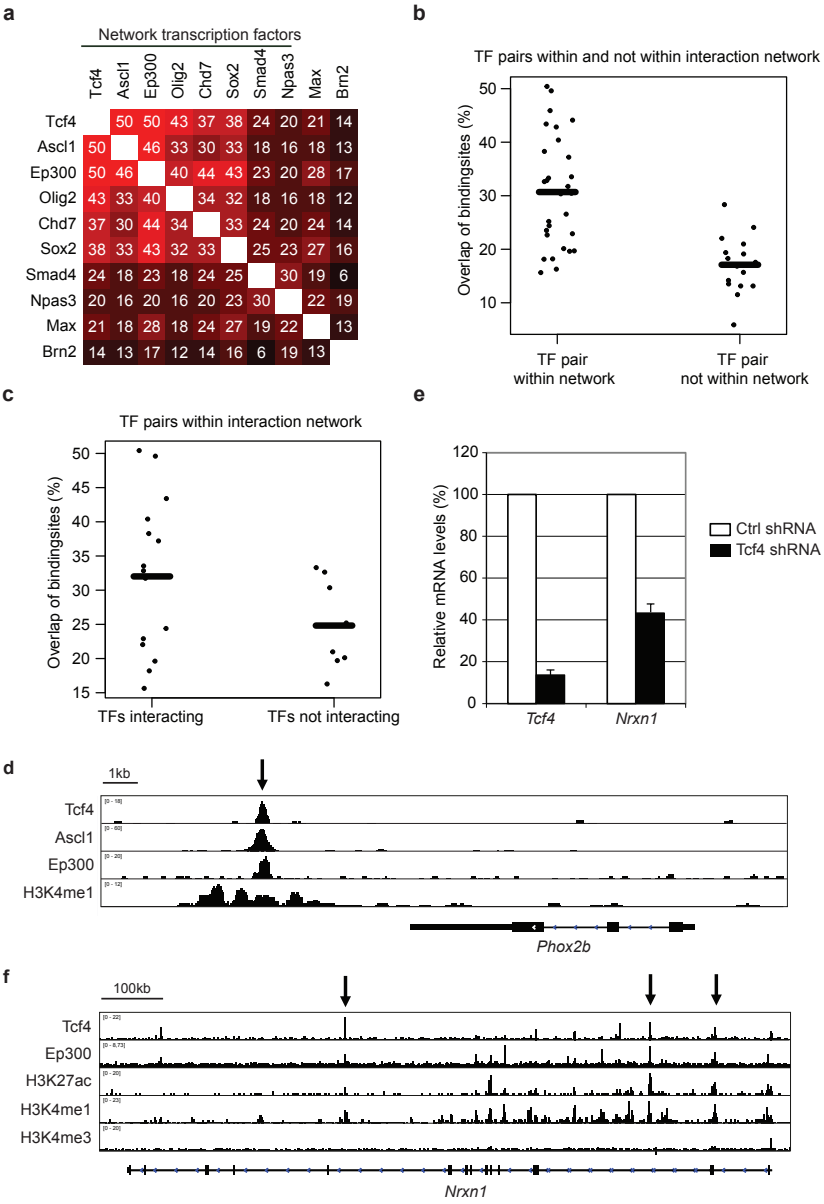


Figure 3. Network transcription factor co-localization on the genome of neural stem cells. (a) Percentage overlap of genome-wide binding sites of pairs of transcription factors (TF pairs). Network transcription factors and percentage overlap of TF pair are indicated. **(b)** Transcription factors are both in the network (left) or one is in the network and one is not (right). Each TF pair is indicated by a black dot, average overlap in each category is indicated by black bar. **(c)** Transcription factors are interacting (left) or not interacting (right). Each TF pair is indicated by a black dot, average overlap in each category is indicated by black bar. **(d and f)** Binding site profile of indicated transcription factors and indicated histone modification profiles at *Phox2b* **(d)** or *Nrxn1* **(f)**. Ep300, H3K4me1 mark enhancers, H3K27ac marks active enhancers, arrows mark transcription factor co-localization. **(e)** Relative mRNA levels by RT-PCR of indicated genes in NSCs treated with the indicated shRNAs, SEM of three independent experiments is indicated.

syndrome. We find that Tcf4 and Ascl1 bind the only transcriptional enhancer near the *Phox2b* gene (Fig. 3d), a gene mutated in 60% of CCHS patients[30] but not expressed in our NSCs (Table S3). Mutations in *TCF4* or *ASCL1* could therefore affect activation of *PHOX2B* during autonomous nervous system development and may thereby cause the overlapping phenotype.

We performed RNAi-mediated knock-down of Tcf4 in NSCs, shortly followed by RNA sequencing. Identified Tcf4 target genes (Table S11) include 71 ID genes, 210 genes *de novo* mutated in ASD patients and 85 genes *de novo* mutated in schizophrenia patients and include well known MD genes *Foxp2*, *Shank3* and *Syngap1* (Table S12). We find that Tcf4 maintains the expression of *Nrxn1* and binds to several active enhancers in the *Nrxn1* gene (Fig. 3e,f and Table S11). Patients with compound heterozygous mutations in *NRXN1* suffer from Pitt Hopkins-like syndrome[31]. Regulation of *Nrxn1* by Tcf4 provides a mechanistic explanation for the strong phenotypic overlap in patients with mutations in any of these two genes. Tcf4 and its interactor Sox2 regulate and colocalize on ID genes *Gpr56*, *Tgfbr2* and *Gli2* (Fig. S2a-d). Disruption of the regulation of *GPR56* by RFX proteins causes cerebral cortex patterning defects and ID[32]. Rfx proteins interact with Tcf4 and Sox2 (Fig. 1a), suggesting their cooperative regulation of *Gpr56*.

Target genes activated by Tcf4 are enriched for genes related to mitotic chromosome segregation (Fig. 4a), a process often affected in primary microcephaly[33]. Indeed, Tcf4 target genes include 6 of the 13 known primary microcephaly genes (Table S12, Fig. 4b) and Tcf4 also regulates *Cenpj* and *Cdk5rap2* (Fig. 4b). Moreover, Tcf4 protein interacts with 10 transcription factors associated with microcephaly, including Smad4 (Fig. 1a and 4c). Smad4, like Tcf4, regulates primary microcephaly genes (Fig. 4d). Tcf4 binds together with Smad4, Sox2, Chd7 and Ep300 to active enhancers at primary microcephaly genes *Mcph1* and *Wdr62* (Fig. 4e). In conclusion, we identified a regulatory network related to microcephaly, which may explain the association of this feature with the participating proteins.

DISCUSSION

A transcription factor interaction network in NSCs enriched for MD-associated proteins

Here we describe the first transcription factor interaction network in a neural system. We rightly anticipated that purifying transcription factors and their associated proteins from NSCs would yield a network enriched for proteins associated with cognitive disorders. Thereby we provide a description of the molecular environment of such proteins, often for the first time, in a cell type highly relevant for neurodevelopment and its diseases. Protein interaction networks containing MD-proteins have been inferred from protein-protein interaction databases[7,8,34-36]. However, we show that the interaction depth of such databases is much less than what we find here experimentally. We carried out our studies in mouse NSCs, as the necessary scale of our proteomics and ChIP-seq experiments would be difficult to perform using human NSCs. Nevertheless, a recent comprehensive comparison of transcriptional networks and transcription factor target genes in mouse and human shows high inter-species conservation[37], making our work relevant for the human situation.

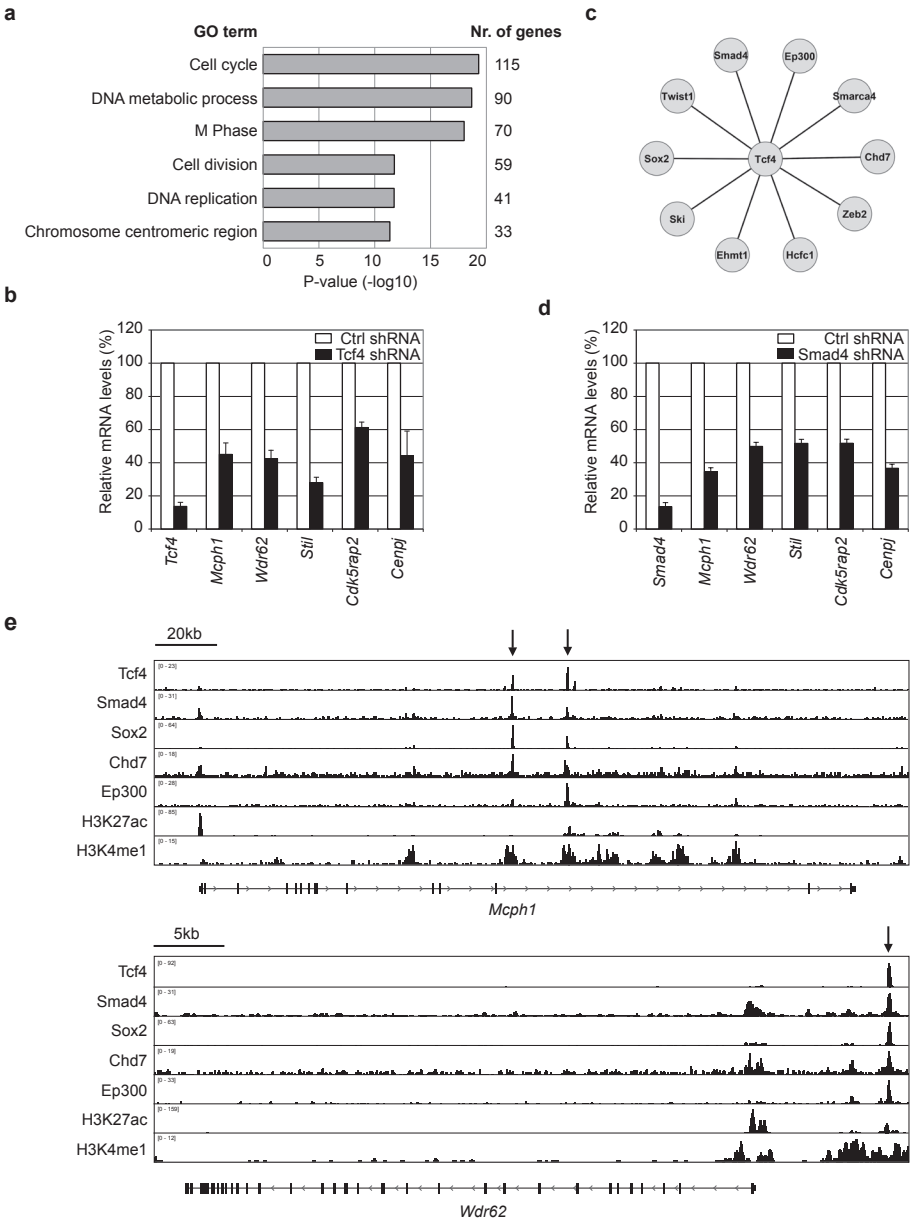


Figure 4. Gene regulation by Tcf4 and interaction partners (a) Gene Ontology analysis of putative Tcf4 target genes. Number of Tcf4 target genes and p-value of each category is indicated. **(b and d)** Regulation of primary microcephaly genes by Tcf4 **(b)** or by Smad4 **(d)**, Relative mRNA levels by RT-PCR of indicated genes in NSCs treated with the indicated shRNAs, SEM of three independent experiments is indicated. **(c)** Interaction network of Tcf4 with network transcription factors that are associated with microcephaly. **(e)** Binding site profile of indicated transcription factors and indicated histone modification profiles at *Mcph1* (upper panel) or *Wdr62* (lower panel). *Angpt2*, internal to Mcph1 and not regulated by Tcf4, is not indicated. Ep300, H3K4me1 mark enhancers, H3K27ac marks active enhancers, arrows mark transcription factor co-localization.

The network enrichment was highest for a set of established ID-associated proteins[5]. Enrichments were lower for ASD-associated and schizophrenia-associated proteins, probably as they were taken from sets of *de novo* mutated proteins where causality was less certain. Protein mutations in these sets that do not contribute to MD would not be enriched in the network but would increase the expected mutation score, reducing overall enrichments. The network is also highly enriched for proteins encoded by evolutionarily constrained genes in humans, a set of genes that is more frequently mutated in disease, including ASD[27]. The enrichments for MD-proteins and constrained genes would suggest that the network has an above-average content of yet to be discovered MD proteins. Indeed, recently three additional bona-fide ID genes were discovered, *RLIM*, *ZBTB20* and *JMJD1C*[38-41], that are encoding network proteins and can be added to 26 ID proteins in the network from the overlap with the 2014 ID gene list[5].

Although MD proteins are enriched in the network, we do not find a strong bias in the percentages of proteins associated with ID, ASD or schizophrenia between the interactomes of the starting proteins Tcf4, Olig2, Npas3 and Sox2. Tcf4 may be considered to have the strongest link to cognition with TCF4 haploinsufficiency causing severe ID in all cases, as part of Pitt Hopkins syndrome in the several hundred patients known[19,20]. SOX2 haploinsufficiency causes ID in only half of the tens of known patients and the ID appears less severe than in Pitt Hopkins syndrome[25,26]. NPAS3 mutations were suggested to cause schizophrenia in two small families[23,24] whereas OLIG2 overexpression in Down Syndrome patients was suggested to contribute to its neurological phenotype[22]. Although it is difficult to draw conclusions from four starting proteins, our interaction network argues against pockets of interacting proteins with strong cognition links. Our data are more compatible with a network with an overall strong relevance for cognitive ability, where individual nodes have different mutation sensitivities.

Genome co-localization and cooperation in gene regulation by proteins in the network

Transcription factors that interact are more likely to cooperate in gene regulation. Another proxy for cooperation in gene regulation is co-localization on the genome[13,28]. For example, we previously showed that Sox2 and its interaction partner Chd7 have a high binding site overlap on the genome and indeed Sox2 and Chd7 have a large overlap (50%) in regulated genes[15]. Interestingly, some of the Sox-Chd7 target genes likely explain the overlap in features between SOX2 syndrome and CHARGE syndrome, caused by CHD7 mutations[15]. Accordingly, to have an additional indication that network proteins preferentially cooperate in gene regulation, we showed that 7 network transcription factors have, on average, more overlap in binding sites with each other than with two transcription factors, Brn2 and Max, that are expressed in NSCs but are not part of the network. In further analogy to Sox2 and Chd7, we find that Tcf4 and Ascl1, the two transcription factors with the highest binding site overlap, also have a strong overlap in their respective human phenotypes. Tcf4 binds with a number of its interaction partners to primary microcephaly genes and (at least) Tcf4 and Smad4 also regulate such genes, providing an explanation for their shared microcephaly phenotype. In conclusion, our interaction network shows features of a transcriptional network, where proteins can cooperate to regulate disease-relevant genes.

Mutation severity in network proteins correlates with IQ levels in the associated MD

We wondered why in our interaction network some mutations cause ID, whereas others cause ASD or schizophrenia, MDs without obligatory loss of IQ. One hypothesis would be that more severe mutations disrupt the network more and have a worse outcome for IQ levels. Mutation severity within a network of interacting proteins has not been analyzed yet in relation to cognition levels. A recent data analysis[9] using *de novo* mutations across all genes in severe ID patients (IQ<50)[42,43], ASD patients[44,45] and schizophrenia patients[9] showed that %LOF increased from 15% in schizophrenia patients, to 17% in ASD patients, to 24% in severe ID patients. %LOF rates in our interaction network of transcriptional regulators show a much stronger increase going from schizophrenia (11% LOF) and ASD-normIQ (17% LOF), two MDs without IQ loss, to ASD-lowIQ (43% LOF) in which IQ is smaller than 90. In particular, %LOF for network proteins associated ASD-lowIQ is nearly twice the %LOF value in the set of *de novo* mutations in severe ID patients[42,43], a group with a more severe cognitive deficit. In addition, in network proteins with multiple mutations across different MDs, mutation severity follows MD severity. We argued above that the enhanced correlation of mutation severity to IQ levels in the network maybe caused in part by mutations often being in network proteins encoded by constrained genes and therefore more likely to be causal. The exaggerated relation within network proteins of mutation severity with IQ loss is consistent with a scenario where the level of disruption of a shared transcriptional network correlates with the level of cognitive dysfunction in the associated MDs. Our interaction network contains only a fraction of the total number of transcriptional regulators believed to be associated with ID, ASD or schizophrenia, but there is no a priori reason to assume that its principles do not apply to a larger network of MD-related transcriptional regulators in NSCs. A common underlying transcriptional network is in line with the significant co-morbidity observed between ID, ASD and schizophrenia[46,47].

METHODS

Transcription factor purification from NSCs and interaction partner identification

NS-5 NSCs were derived from 46C embryonic stem cells[48] and cultured, as described[49] and regularly tested for mycoplasma contamination and for authenticity by expressed NSC markers Pax6, Sox2 and Nestin[15]. NSC lines with stable expression of FLAG-V5-tagged Tcf4, Olig2 or Npas3 were created by electroporation with pCAG promoter-driven plasmids containing the appropriate cDNAs and puromycin selection for individual clones[15]. FLAG-tagged Tcf4, Olig2 or Npas3 were each purified from 1.5 ml nuclear extract, equivalent to $2*10^8$ NSCs, by FLAG-affinity purification, as described[12,15]. Two or three independent purifications of each FLAG-tagged protein from separate NSC cultures and control purifications from separate parental NSC cultures were performed by the same experimentor(s). Identification of proteins by mass spectrometry was as described[12]. Peptide spectra of purifications of Tcf4, Olig2, Npas3 and previous purifications of Sox2[15] were searched against UniProt release 2012-11 for protein identification. Interaction partner identification criteria are as described[12]. In short, a protein is included as interaction partner of a FLAG-tagged transcription factor if present in at least two of its purifications with a Mascot score of 50 or higher and at least 3-fold enriched by Mascot score over control purifications. Interaction network graphics were

made with Cytoscape[50].

Chromatin immunoprecipitations (ChIP)

1.5*10[8] NSCs were used per ChIP. For Olig2 ChIP, NSCs were washed 3 times with PBS, crosslinked with 1/10 volume of fresh 11% buffered formaldehyde solution for 12 min., quenched with 1/20 volume of 2.5M Glycine for 5 min, washed with ice-cold PBS and cell pellets frozen with N2(l) and resuspended and washed 2 times in ice-cold cell lysis buffer (10mM Tris-Cl pH 7.5, 10mM NaCl, 3 mM MgCl2, 0.5% NP40). Cell pellets were resuspended in lysis buffer with 1mM CaCl2 and 4% NP40 and sonicated, as described [51]. ChIP was performed, as described[52] using 15mg of Olig2 antibody (AB9610, Millipore) or rabbit IgG for the control ChIP. For Tcf4 ChIP, FLAG-V5-Tcf4 expressing NSCs were crosslinked with 2 mM disuccinimidyl glutarate (DSG, Thermo Fisher Scientific) and 1% formaldehyde, nuclei isolated, chromatin prepared and ChIP performed, as described[51,53] with 20ml V5-antibody agarose beads (Sigma). DNA was eluted from the V5-beads, as described[53]. NSCs not expressing FLAG-V5-Tcf4 were used as a control. For Npas3 ChIP, NSCs were crosslinked with DSG and formaldehyde and ChIP performed as described[51,53] with 15mg of Npas3 antibody (HPA002892, Sigma) or rabbit IgG as control, and 60ml prot-G beads (GE healthcare), without crosslinking the antibody to the beads. Smad4 ChIP was on NSCs crosslinked with DSG and formaldehyde[51,53], with 15mg of Smad4 antibody (R&D Systems, AF2097) or goat IgG and 150ml prot-G Dynabead solution (Life Technologies), without crosslinking the antibody to the beads. ChIP DNA library preparation and ChIP sequencing on Illumina GAII or HiSeq2500 platforms was performed at the Erasmus MC Center for Biomics, as described[54].

Gene regulation experiments

pSuper-puro constructs encoding Tcf4 shRNA, Tcf4 shRNA2, Smad4 shRNA or Sox2 shRNA were electroporated into NSCs, as described[15], puromycin (2mg ml^{-1}) was added after 18 hrs and NSCs were harvested for analyses at 44 hrs after electroporation. Three independent electroporations were performed per condition. For RNA-seq of untreated NSCs and NSCs transfected with Tcf4 shRNA construct or control shRNA (Dharmacon) construct, poly(A) RNA was isolated using the RNeasy kit (Qiagen), tested for quality with the Bioanalyzer and prepared using the TruSeq RNA sample prep kit v2, as described[55]. RNA-seq was performed at the Erasmus MC Center for Biomics on a HiSeq2500 sequencer (Illumina) according to manufacturer's instructions. RNA samples were sequenced for 36 bp.

Data analysis

Known ID genes (528 genes), mutated in 5 or more ID patients, are from[5]. Genes with *de novo* mutations in patients with ASD with lower IQ (≤90) and patients with ASD with normal IQ (>90) are from [6]. Likely-gene-disrupting (LGD) mutations[6] were classified as LOF mutations. Genes with LOF or missense mutations in siblings were removed, resulting in a list of 1584 genes. Males with ASD and lower IQ (≤90) and females with ASD (which nearly always have lower IQ[6,56]) were classified as ASD with lower IQ (ASD-lowIQ). Males with ASD and normal IQ (>90) were classified as ASD-normIQ. Genes with *de novo* LOF and missense mutations in schizophrenia patients (662 genes) were derived from[8,9,57,58]. Frameshift mutations, nonsense mutations and splice site mutations (within 2 nucleotides from the splice donor site or splice acceptor site[9]) were taken as

LOF mutations.

LOF mutations were calculated as a percentage of all *de novo* coding mutations (LOF + missense) in patients with ASD-lowIQ, ASD-normIQ or schizophrenia, either in genes with the equivalent mouse protein in the interaction network (Network protein genes) or in the total data sets (All genes). For known ID genes with the equivalent mouse protein in the interaction network, the type of mutation in the majority of patients was assessed per gene and assigned as LOF or missense. The percentage of network ID genes with predominant LOF mutations is represented.

Evolutionarily constrained genes, significantly devoid of coding variants in the human population (1003 genes) are from[27]. Enrichments and enrichment P-values of *de novo* mutations in ASD patients or schizophrenia patients in human genes equivalent to network proteins (206 proteins/ human genes) were calculated by dnenrich[9], corrected for gene length, sequence context and expression of the mouse homologue in our neural stem cells. Equivalent human genes of network proteins were provided to the dnenrich program as 'Gene set' and human equivalents of genes expressed in mouse NSCs were provided as 'Background list'. A gene was regarded as expressed in our NSCs, if its expression was above or equal to that of *Zeb1* (0.127 RPKM in our RNAseq data set), which is the network protein with the lowest mRNA expression in our NSCs. Enrichment P-values for network proteins encoding known ID genes or constrained genes are obtained from two-sided binomial tests.

13 human primary microcephaly genes (MCPH1, WDR62, ASPM, CASC5, CENPJ, CENPE, CDK5RAP1, CEP135, CEP152, STIL, CDK6, ZNF533, PHC1) are known[33,59], which were overlapped with Tcf4 target genes (see below). Microcephaly genes were retrieved from the Online Mendelian Inheritance in Man (OMIM) (http://www.omim.org) database by scoring for genes in which mutations cause human monogenic conditions or syndromes that include microcephaly.

ChIP-seq datasets were processed and mapped to the NCBIM37.61 (mm9) reference genome, as described[51]. Published ChIP-seq datasets for Ascl1, Sox2, Brn2, H3K4me1 and H3K27ac in NSCs were retrieved from the Gene Expression Omnibus with accession codes GSE48336, GSE35496, GSE11172 and GSE24164[60-63]. Published ChIP-seq datasets for Ep300 and Max in NSCs were retrieved from European Nucleotide Archive with accession codes ERP002084 and ERP004644[16,64]. MACS 1.4.2 was used for peak calling and for the generation of binding profiles[65] using default settings and the corresponding control ChIP as a control dataset. The 5000 most significant peaks (genome-wide binding sites) for each transcription factor were used to determine the percentage of overlap between two transcription factors. Two binding sites were considered overlapping if their summits were within 250bp. The corresponding figures were generated using R. ChIP-seq tracks were generated in the IGV browser[66].

RNA-seq was mapped against mouse reference NCBIM37.67 (mm9) using Tophat50 v2.0.11 with default settings and a segment length of 20. The aligned exon reads were normalized and differential expression was calculated using Bioconductor DESeq2 package in R[67]. Tcf4 target genes were defined as having at least a 1.5 fold change in

expression (adjusted p-value ≤ 0.01) upon Tcf4 knock-down, at least one significant Tcf4 binding site (p-value ≤ $1*10^{-10}$) within 100 kb of its transcription start site and at least 0.1 RPKM expression in untreated NSCs. DAVID was used for Gene ontology analysis[68].

The database-derived protein interaction network was composed as the sum of mouse and human interactions derived from the StringDB database[69], that were experimentally verified with high confidence scores, and interactions derived from the HPRD database[70]. Interactions of Tcf4 with Ctnnb1, Pias4, Sumo-1 and Ubc1 were discarded as these are with Tcf7l2 (also known as T-cell factor 4, Tcf4). HPRD database-generated Sox2 interactor Rps27a (ribosomal protein) was not based on interaction with Sox2 and discarded. HPRD database-generated Sox2 interactor Fgf4 was discarded as this was an interaction between Sox2 and *Fgf4* gene DNA, not Fgf4 protein.

Author contributions

M.J.M. performed plasmid construction, protein purifications, ChIP experiments and gene regulation experiments. H.H.H.A. performed plasmid construction, protein purifications and proteomics data categorization. J.H.B. processed ChIP-seq and RNA-seq data and performed bioinformatics analyses. D.H.W.D and J.D. performed the mass spectrometry analyses. U.A. and S.K. performed plasmid construction and protein purifications. C.E.M.K., Z.O. and W.F.J.van IJ. performed labeling and Illumina sequencing of ChIP and RNA material. R.A.P. conceived the study and designed experiments. R.A.P. wrote the manuscript with help from co-authors.

Acknowledgments

We thank Richard Festenstein, Frank Grosveld and Danny Huylebroeck for comments on the manuscript and Erik Engelen, Mike Dekker and Marti Quevedo-Calero for technical assistance. M.J.M. was supported by a grant from the Erasmus MC Stem Cell Institute, J.H.B. was supported by an ALW-open program grant (No 821.02.004) from the Netherlands Organisation for Scientific Research (NWO) and R.A.P. by a grant from the Dutch government to the Netherlands Institute for Regenerative Medicine (NIRM, grant No. FES0908) and the DevRepair (P7/07) IAP-VII network. D.H.D and J.D were funded by The Netherlands Proteomics Centre (Project Number 184.032.201), financed by NWO.

Author information

ChIP sequencing and RNA sequencing data are available through the Gene Expression Omnibus (NCBI), accession code GSE70872 . The authors do not declare competing financial interests.

REFERENCES

1. Association, A.P. DSM5 diagnostic and statistical manual of mental disorders. *APA, Arlington* (2013).
2. Ropers, H.H. Genetics of early onset cognitive impairment. *Annu Rev Genomics Hum Genet* **11**, 161-87 (2010).
3. Mefford, H.C., Batshaw, M.L. & Hoffman, E.P. Genomics, intellectual disability, and autism. *N Engl J Med* **366**, 733-43 (2012).
4. Schizophrenia Working Group of the Psychiatric Genomics, C. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421-7 (2014).

5. Gilissen, C. *et al.* Genome sequencing identifies major causes of severe intellectual disability. *Nature* **511**, 344-7 (2014).
6. Iossifov, I. *et al.* The contribution of de novo coding mutations to autism spectrum disorder. *Nature* **515**, 216-21 (2014).
7. De Rubeis, S. *et al.* Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature* **515**, 209-15 (2014).
8. Gulsuner, S. *et al.* Spatial and temporal mapping of de novo mutations in schizophrenia to a fetal prefrontal cortical network. *Cell* **154**, 518-29 (2013).
9. Fromer, M. *et al.* De novo mutations in schizophrenia implicate synaptic networks. *Nature* **506**, 179-84 (2014).
10. Sahin, M. & Sur, M. Genes, circuits, and precision therapies for autism and related neurodevelopmental disorders. *Science* **350**(2015).
11. Ronan, J.L., Wu, W. & Crabtree, G.R. From neural development to cognition: unexpected roles for chromatin. *Nat Rev Genet* **14**, 347-59 (2013).
12. van den Berg, D.L. *et al.* An Oct4-centered protein interaction network in embryonic stem cells. *Cell Stem Cell* **6**, 369-81 (2010).
13. Chen, X. *et al.* Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell* **133**, 1106-17 (2008).
14. Kim, J., Chu, J., Shen, X., Wang, J. & Orkin, S.H. An extended transcriptional network for pluripotency of embryonic stem cells. *Cell* **132**, 1049-61 (2008).
15. Engelen, E. *et al.* Sox2 cooperates with Chd7 to regulate genes that are mutated in human syndromes. *Nat Genet* **43**, 607-11 (2011).
16. Mateo, J.L. *et al.* Characterization of the neural stem cell gene regulatory network identifies OLIG2 as a multifunctional regulator of self-renewal. *Genome Res* **25**, 41-56 (2015).
17. Pieper, A.A. *et al.* The neuronal PAS domain protein 3 transcription factor controls FGF-mediated adult hippocampal neurogenesis in mice. *Proc Natl Acad Sci U S A* **102**, 14052-7 (2005).
18. Graham, V., Khudyakov, J., Ellis, P. & Pevny, L. SOX2 functions to maintain neural progenitor identity. *Neuron* **39**, 749-65 (2003).
19. Zweier, C. *et al.* Haploinsufficiency of TCF4 causes syndromal mental retardation with intermittent hyperventilation (Pitt-Hopkins syndrome). *Am J Hum Genet* **80**, 994-1001 (2007).
20. Amiel, J. *et al.* Mutations in TCF4, encoding a class I basic helix-loop-helix transcription factor, are responsible for Pitt-Hopkins syndrome, a severe epileptic encephalopathy associated with autonomic dysfunction. *Am J Hum Genet* **80**, 988-93 (2007).
21. Stefansson, H. *et al.* Common variants conferring risk of schizophrenia. *Nature* **460**, 744-7 (2009).
22. Chakrabarti, L. *et al.* Olig1 and Olig2 triplication causes developmental brain defects in Down syndrome. *Nat Neurosci* **13**, 927-34 (2010).
23. Kamnasaran, D., Muir, W.J., Ferguson-Smith, M.A. & Cox, D.W. Disruption of the neuronal PAS3 gene in a family affected with schizophrenia. *J Med Genet* **40**, 325-32 (2003).
24. Yu, L. *et al.* A mutation in NPAS3 segregates with mental illness in a small family. *Mol Psychiatry* **19**, 7-8 (2014).
25. Ragge, N.K. *et al.* SOX2 anophthalmia syndrome. *Am J Med Genet A* **135**, 1-7; discussion 8 (2005).
26. Bakrania, P. *et al.* SOX2 anophthalmia syndrome: 12 new cases demonstrating broader phenotype and high frequency of large gene deletions. *Br J Ophthalmol* **91**, 1471-6 (2007).
27. Samocha, K.E. *et al.* A framework for the interpretation of de novo mutation in human disease. *Nat Genet* **46**, 944-50 (2014).
28. Ouyang, Z., Zhou, Q. & Wong, W.H. ChIP-Seq of transcription factors predicts absolute and differential gene expression in embryonic stem cells. *Proc Natl Acad Sci U S A* **106**, 21521-6 (2009).
29. de Pontual, L. *et al.* Noradrenergic neuronal development is impaired by mutation of the proneural HASH-1 gene in congenital central hypoventilation syndrome (Ondine's curse). *Hum Mol Genet* **12**, 3173-80 (2003).
30. Amiel, J. *et al.* Polyalanine expansion and frameshift mutations of the paired-like homeobox gene PHOX2B in congenital central hypoventilation syndrome. *Nat Genet* **33**, 459-61 (2003).
31. Zweier, C. *et al.* CNTNAP2 and NRXN1 are mutated in autosomal-recessive Pitt-Hopkins-like mental retardation and determine the level of a common synaptic protein in Drosophila. *Am J Hum Genet* **85**, 655-66 (2009).
32. Bae, B.I. *et al.* Evolutionarily dynamic alternative splicing of GPR56 regulates regional cerebral cortical patterning. *Science* **343**, 764-8 (2014).

33. Faheem, M. *et al.* Molecular genetics of human primary microcephaly: an overview. *BMC Med Genomics* **8 Suppl 1**, S4 (2015).
34. Neale, B.M. *et al.* Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature* **485**, 242-5 (2012).
35. O'Roak, B.J. *et al.* Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature* **485**, 246-50 (2012).
36. Hormozdiari, F., Penn, O., Borenstein, E. & Eichler, E.E. The discovery of integrated gene networks for autism and related disorders. *Genome Res* **25**, 142-54 (2015).
37. Yue, F. *et al.* A comparative encyclopedia of DNA elements in the mouse genome. *Nature* **515**, 355-64 (2014).
38. Tonne, E. *et al.* Syndromic X-linked intellectual disability segregating with a missense variant in RLIM. *Eur J Hum Genet* **23**, 1652-6 (2015).
39. Hu, H. *et al.* X-exome sequencing of 405 unresolved families identifies seven novel intellectual disability genes. *Mol Psychiatry* (2015).
40. Cordeddu, V. *et al.* Mutations in ZBTB20 cause Primrose syndrome. *Nat Genet* **46**, 815-7 (2014).
41. Saez, M.A. *et al.* Mutations in JMJD1C are involved in Rett syndrome and intellectual disability. *Genet Med* (2015).
42. de Ligt, J. *et al.* Diagnostic exome sequencing in persons with severe intellectual disability. *N Engl J Med* **367**, 1921-9 (2012).
43. Rauch, A. *et al.* Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study. *Lancet* **380**, 1674-82 (2012).
44. Iossifov, I. *et al.* De novo gene disruptions in children on the autistic spectrum. *Neuron* **74**, 285-99 (2012).
45. Sanders, S.J. *et al.* De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* **485**, 237-41 (2012).
46. Kohane, I.S. *et al.* The co-morbidity burden of children and young adults with autism spectrum disorders. *PLoS ONE* **7**, e33224 (2012).
47. Desai, S. *et al.* Changes in psychiatric comorbidity during early postsurgical period in patients operated for medically refractory epilepsy--a MINI-based follow-up study. *Epilepsy Behav* **32**, 29-33 (2014).
48. Ying, Q.L., Stavridis, M., Griffiths, D., Li, M. & Smith, A. Conversion of embryonic stem cells into neuroectodermal precursors in adherent monoculture. *Nat Biotechnol* **21**, 183-6 (2003).
49. Conti, L. *et al.* Niche-independent symmetrical self-renewal of a mammalian tissue stem cell. *PLoS Biol* **3**, e283 (2005).
50. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13**, 2498-504 (2003).
51. Engelen, E. *et al.* Proteins that bind regulatory regions identified by histone modification chromatin immunoprecipitations and mass spectrometry. *Nat Commun* **6**, 7155 (2015).
52. Bernstein, B.E. *et al.* Genomic maps and comparative analysis of histone modifications in human and mouse. *Cell* **120**, 169-81 (2005).
53. Boyer, L.A. *et al.* Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell* **122**, 947-56 (2005).
54. Soler, E. *et al.* The genome-wide dynamics of the binding of Ldb1 complexes during erythroid differentiation. *Genes Dev* **24**, 277-89 (2010).
55. Meinders, M. *et al.* Sp1/Sp3 transcription factors regulate hallmarks of megakaryocyte maturation and platelet formation and function. *Blood* **125**, 1957-67 (2015).
56. Newschaffer, C.J. *et al.* The epidemiology of autism spectrum disorders. *Annu Rev Public Health* **28**, 235-58 (2007).
57. Xu, B. *et al.* De novo gene mutations highlight patterns of genetic and neural complexity in schizophrenia. *Nat Genet* **44**, 1365-9 (2012).
58. Guipponi, M. *et al.* Exome sequencing in 53 sporadic cases of schizophrenia identifies 18 putative candidate genes. *PLoS ONE* **9**, e112745 (2014).
59. Mirzaa, G.M. *et al.* Mutations in CENPE define a novel kinetochore-centromeric mechanism for microcephalic primordial dwarfism. *Hum Genet* **133**, 1023-39 (2014).
60. Webb, A.E. *et al.* FOXO3 shares common targets with ASCL1 genome-wide and inhibits ASCL1-dependent neurogenesis. *Cell Rep* **4**, 477-91 (2013).
61. Lodato, M.A. *et al.* SOX2 co-occupies distal enhancer elements with distinct POU factors in ESCs

and NPCs to specify cell state. *PLoS Genet* **9**, e1003288 (2013).
62. Meissner, A. *et al.* Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature* **454**, 766-70 (2008).
63. Creyghton, M.P. *et al.* Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc Natl Acad Sci U S A* **107**, 21931-6 (2010).
64. Martynoga, B. *et al.* Epigenomic enhancer annotation reveals a key role for NFIX in neural stem cell quiescence. *Genes Dev* **27**, 1769-86 (2013).
65. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**, R137 (2008).
66. Robinson, J.T. *et al.* Integrative genomics viewer. *Nat Biotechnol* **29**, 24-6 (2011).
67. Love, M.I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15**, 550 (2014).
68. Huang da, W., Sherman, B.T. & Lempicki, R.A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* **4**, 44-57 (2009).
69. Szklarczyk, D. *et al.* The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res* **39**, D561-8 (2011).
70. Prasad, T.S., Kandasamy, K. & Pandey, A. Human Protein Reference Database and Human Proteinpedia as discovery tools for systems biology. *Methods Mol Biol* **577**, 67-79 (2009).

SUPPLEMENTAL FIGURES

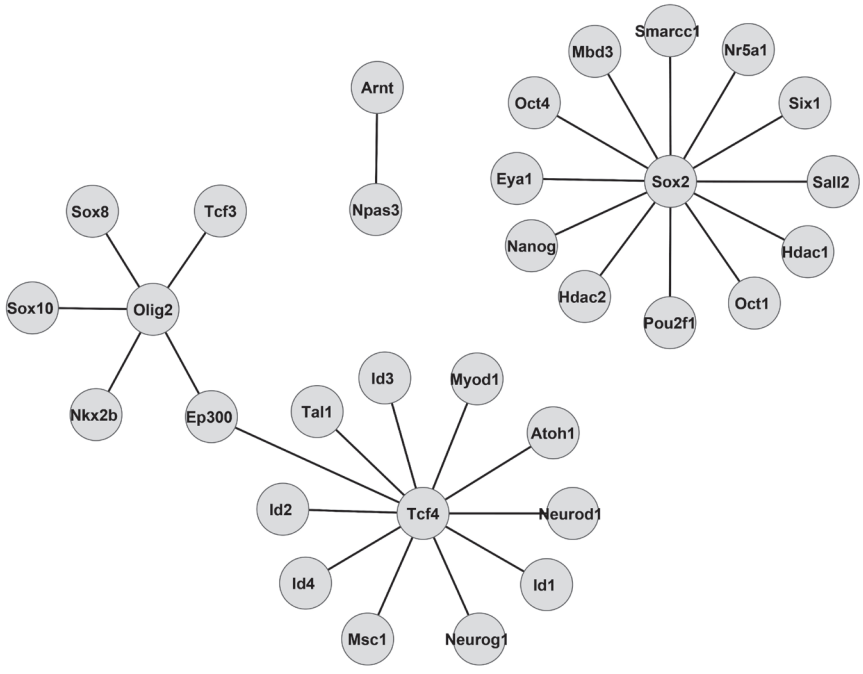


Figure S1. Protein interaction network generated with database information, related to Figure 1. Database-derived interaction network was derived from the sum of mouse and human experimentally-verified interactions of Tcf4, Olig2, Npas3 and Sox2 from the StringDB database and the Human Protein Interaction Database.

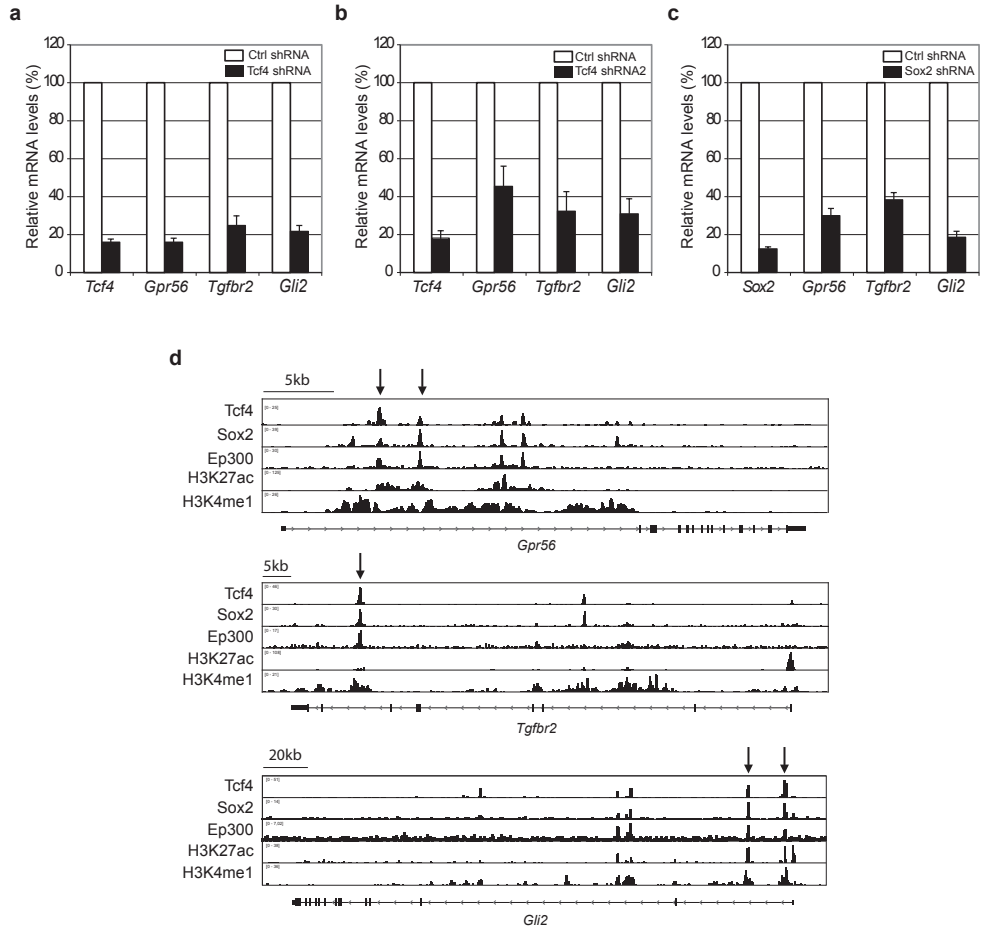


Figure S2. Gene regulation by Tcf4 and Sox2, related to Figure 4. (a-c) Relative mRNA levels by RT-PCR of indicated genes in NSCs treated with the indicated shRNAs. **(b)** Tcf4 shRNA2 represents an alternative Tcf4 shRNA. SEM of three independent experiments is indicated. **(d)** Binding site profile of indicated transcription factors and indicated histone modification profiles at *Gpr56* (upper panel), *Tgfbr2* (middle panel), *Gli2* (bottom panel). Ep300, H3K4me1 mark enhancers, H3K27ac marks active enhancers, arrows mark transcription factor co-localization.

SUPPLEMENTAL TABLES

Our purification protocol 1 has been used by us[1-7] and others[8] to purify proteins and identify their interaction partners by mass spectrometry. The biological relevance of many of these identified protein-protein interactions was determined and is listed in Table S1, if the biological relevance was determined by us (R. Poot is author), or listed in Table S2 if the biological relevance was determined independently by others. Note that in Table S2, all interactions, except the last one (Rnf12-Rex1), were identified by us but validation and determination of their biological relevance was performed independently by others.

Table S1. Previously identified interactions by our purification protocol and their biological relevance, determined by us, related to Figure 1

Interaction	Biological relevance	Reference
Jarid2-PRC2 complex	Transcriptional priming of PRC2 target genes	6
Pcl2-PRC2 complex	PRC2 recruitment to inactive X chromosome and PRC2 target genes	7
Sox2-Chd7	Sox2 and Chd7 cooperate in the regulation of common set of target genes, including disease-associated genes	5
Rybp-PRC1 complex	PRC1 recruitment to target genes in an H3K27me3-independent pathway	4
Nanog-Sox2	Nanog-Sox2 interaction is important for ES cell self-renewal	3
Pax6-Brg1 complex	Pax6-Brg1 complex regulates genes in adult neural stem cells to potentiate neurogenesis	2

Table S2. Previously identified interactions by our purification protocol and their biological relevance, determined independently by others, related to Figure 1

Interaction	Biological relevance	References
Esrrb-Dax1	Dax1 regulates the activity of Esrrb as a transcriptional activator	Biological relevance: 9, interaction identification: 10, 1
Oct4-Sall4	Sall4 stimulates transcriptional activation by Oct4	Biological relevance: 11, interaction identification: 10, 1, 12
Oct4-Wdr5	Oct4-directed Wdr5 binding to promoters of Oct4 target genes promotes the H3K4me3 mark and gene activation	Biological relevance: 13, interaction identification: 1
Esrrb-Ncoa3	Ncoa3 physically links Esrrb to RNA polymerase 2	Biological relevance: 14, interaction identification: 1
Oct4-Ogt	O-GlcNac modification of threonine 228 on Oct4 facilitates the activity of Oct4 in ESC maintenance and reprogramming to iPSCs	Biological relevance: 15, interaction identification: 1, 12
Oct4-SWI-SNF-NuRD complex	Single amino acid changes in Oct4 reduce its binding to SWI-SNF and NuRD complex and abolish its activity in reprogramming to iPSCs	Biological relevance: 16, interaction identification: 1, 12
Rnf12-Rex1	Rnf12 targets Rex1 for degradation to allow for X-chromosome inactivation	Biological relevance and interaction identification: 8

Table S3. Relative mRNA levels of indicated genes in used neural stem cells from our RNA-seq data, related to Figures 1, 3 and 4. a: SD of 3 independent RNA samples is indicated

Gene	Relative expression (rkpm)[a]
Tcf4	18.7 ± 0.5
Olig2	183.7 ± 11.2
Npas3	5.2 ± 0.3
Sox2	48.8 ± 3.2
Smad4	15.2 ± 0.3
Chd7	7.3 ± 0.1
Ascl1	8.3 ± 0.4
Ep300	7.8 ± 0.4
Brn2 (Pou3f2)	6.7 ± 2.0
Max	19.0 ± 0.3
Phox2b	0.1 ± 0.04
Mean gene	17.0
Median gene	0.8

Tables S4-S7: Interacting proteins of Tcf4, Olig2, Npas3 and Sox2, as identified by mass spectrometry analysis of purified protein samples related to Figure 1. Purified FLAG-tagged transcription factor, name of interacting protein, accession number, emPAI score (a relative and semi-quantitative measure of molar amount of protein 17, Mascot score (a measure of correct protein identification) and number of identified unique peptides and independent experiments are indicated. Proteins are categorized by their stable complex, or being a transcription factor (have a link to transcriptional regulation in the Uniprot protein database

(For corresponding tables see the next 9 pages)

Table S4. Interacting proteins of Tcf4 as identified by mass spectrometry analysis of purified protein sample

			Experiment 1			Experiment 2			Experiment 3		
Protein	**Accession**	**Average emPAI**	**EmPAI**	**Mascot**	**Uniq. pept.[a]**	**EmPAI**	**Mascot**	**Uniq. pept.[a]**	**Empai**	**Mascot**	**Uniq. pept.[a]**
Tcf4	Q91YV0	30.48	21.73	2256	31	38.38	2389	29	31.33	2374	29
Apc complex											
Anapc1	P53995	0.20	0.47	1019	24	0.10	251	7	0.02	61	1
Anapc5	Q8BTZ4	0.19	0.54	464	9	0.04	57	1	0.00	0	0
Anapc7	Q9WVM3	0.14	0.37	204	5	0.06	51	1	0.00	0	0
Cnot complex											
Cnot8	Q9D8X5	0.40	0.70	221	5	0.49	193	4	0.00	0	0
Cnot1	B7ZWL1	0.37	0.84	2117	44	0.19	635	14	0.09	340	8
Cnot2	Q8C5L3	0.13	0.27	265	4	0.00	0	0	0.12	73	2
Cnot10	Q8BH15	0.07	0.14	115	3	0.04	72	1	0.04	71	1
Ncor-Tbl1xr1 complex											
Tbl1xr1	Q8BHJ5	2.96	3.97	1226	19	2.97	1185	17	1.93	976	15
Tbl1x	Q9QXE7	1.71	2.29	860	15	1.71	934	15	1.14	619	11
Ncor2	F6Z4B2	0.40	0.49	1428	32	0.31	923	20	0.41	1155	26
Ncor1	E9Q8K6	0.29	0.38	1149	24	0.21	696	14	0.28	888	18
NuRD complex											
Mbd3	D3YTR5	4.04	4.22	631	12	3.96	621	11	3.94	574	11
Gatad2b	Q8VHR5	1.55	2.74	1156	20	1.05	613	11	0.85	604	11
Chd4 (Mi2-beta)	E9QAS5	0.82	1.22	2181	48	0.57	1185	26	0.67	1369	29
Mta1	F8WHY8	0.82	1.07	794	15	0.66	673	12	0.73	777	13
Mbd2	Q9Z2E1	0.49	0.63	245	6	0.47	225	5	0.36	128	4
Gatad2a	E9QMN5	0.45	0.80	623	11	0.28	301	5	0.28	234	5
Spalt proteins											
Sall3	Q62255	0.75	1.06	1316	24	0.66	1096	18	0.54	992	16
Sall2	F7ARK3	0.17	0.25	318	7	0.16	338	5	0.09	272	3
Swi-Snf complex											
Actl6a	Q9Z2N8	1.21	1.12	626	10	1.49	672	11	1.01	601	10
Smarcc2	Q3UID0	0.36	0.42	641	12	0.42	726	12	0.24	304	2
Smarcd1	Q61466	0.32	0.45	340	7	0.33	226	5	0.19	132	2
Smarce1	O54941	0.25	0.26	107	2	0.33	210	4	0.15	105	1
Smarca4	G3UX35	0.15	0.22	429	10	0.15	280	6	0.07	141	3
Arid1a	A2BH40	0.08	0.16	425	9	0.04	162	3	0.03	96	2
Trrap complex											
Ep400	F6R9G0	0.56	0.85	786	18	0.62	438	10	0.21	216	5
Ing3	Q8VEK6	0.32	0.48	176	5	0.40	156	3	0.08	70	1
Trrap	E9PZA7	0.19	0.30	1492	32	0.17	837	20	0.10	481	12
Brd8	Q8R3B7	0.12	0.16	141	4	0.14	222	4	0.07	128	2
Kat5	Q8CHK4	0.10	0.00	0	0	0.19	83	3	0.12	85	2
Transcription factors											
Id4	P41139	4.13	6.71	490	8	2.85	422	6	2.84	334	5
Cbfa2t2	O70374	3.82	6.43	1459	26	3.13	1125	20	1.89	931	17
Twist1	P26687	3.43	3.46	413	6	3.77	423	6	3.07	284	5
Id1	A2AHY3	2.95	3.70	512	7	2.58	416	6	2.57	443	6
Tcf12	Q61286	2.83	3.38	1489	26	2.48	1452	23	2.62	1492	24
Ccdc101	Q9DA08	2.75	3.95	782	13	2.32	730	11	1.99	636	10
Olig1	Q9JKN5	2.63	3.16	581	9	2.37	590	9	2.37	550	9
Id2	P41136	1.48	2.16	278	5	1.38	189	4	0.91	165	3
Rqcd1	Q9JKY0	1.46	2.53	633	13	1.20	365	8	0.64	273	6
Bap18	Q9DCT6	1.07	1.63	247	4	0.93	183	3	0.64	126	3
Wdr5	F6Q3W0	0.91	0.96	452	8	1.06	412	8	0.72	354	6
Tcf3	E9PWE5	0.86	1.61	719	16	0.49	374	8	0.49	398	8
0610010K14Rik	D3Z687	0.86	1.01	247	4	0.93	183	3	0.64	126	3
Hdac3	Q3UM33	0.85	1.08	486	9	0.86	390	8	0.62	340	6
Id3	P41133	0.79	1.82	257	4	0.28	97	1	0.28	91	1
Sox2	Q60I23	0.75	1.30	384	6	0.48	165	4	0.47	163	4
Sub1	P11031	0.73	1.61	192	4	0.57	95	2	0.00	0	0
Ascl1	Q02067	0.68	0.33	110	2	1.23	252	5	0.49	134	3
Cbx3	P23198	0.65	0.96	223	4	0.61	105	2	0.37	80	2
Rnf2	Q9CQJ4	0.63	0.93	295	6	0.42	189	4	0.55	224	5
Kdm1a	A3KG93	0.59	0.71	712	14	0.48	698	11	0.59	745	13

			Experiment 1			Experiment 2			Experiment 3		
Protein	Accession	Average emPAI	EmPAI	Mascot	Uniq. pept.[a]	EmPAI	Mascot	Uniq. pept.[a]	Empai	Mascot	Uniq. pept.[a]
Yeats4	Q9CR11	0.56	0.95	215	4	0.46	147	3	0.28	59	2
Nfia	B1AUB6	0.55	0.82	432	8	0.28	177	4	0.55	363	7
Smarca5	Q91ZW3	0.54	0.66	777	17	0.65	874	16	0.32	564	10
Trps1	Q925H1	0.53	0.79	1205	23	0.46	866	16	0.33	671	12
Mta3	E9Q794	0.52	0.73	471	9	0.33	293	5	0.49	380	6
Trim33	E9QME5	0.50	0.91	1228	21	0.31	647	10	0.27	583	9
Znf148	Q61624	0.45	0.69	736	13	0.41	373	7	0.25	311	6
Morf4l1	P60762	0.43	0.41	270	4	0.50	231	4	0.38	183	3
Wiz	F6ZBR8	0.41	0.51	589	12	0.47	546	12	0.25	360	7
Cdc23	G3X8W7	0.41	0.82	566	11	0.29	167	5	0.11	61	2
Nfix	D3YZ00	0.40	0.71	332	8	0.28	207	4	0.21	153	3
Sox8	Q04886	0.39	0.54	227	6	0.22	102	3	0.40	175	4
Gps2	Q921N8	0.37	0.48	170	4	0.32	105	2	0.31	94	2
Hexim2	Q3TVI4	0.37	0.44	92	2	0.26	50	1	0.41	82	2
Jmjd1c	G3UYW3	0.36	0.59	1625	36	0.23	803	17	0.26	904	18
Hmgb2	P30681	0.35	0.55	145	3	0.51	95	3	0.00	0	0
Prdm6	Q3UZD5	0.35	0.64	397	9	0.36	167	4	0.05	49	1
Zkscan3	Q91VW9	0.33	0.33	270	5	0.41	110	3	0.26	84	2
Trim21	Q3U7K7	0.29	0.39	218	6	0.36	214	4	0.13	121	2
Hoxa5	P09021	0.29	0.62	231	4	0.00	0	0	0.25	105	2
Bcl7c	O08664	0.29	0.00	0	0	0.53	110	3	0.33	90	2
Tal1	A2AD40	0.27	0.27	75	1	0.34	164	3	0.21	105	2
Maml2	F6U238	0.27	0.59	642	13	0.10	151	3	0.13	183	4
Sin3a	Q60520	0.27	0.42	470	13	0.20	415	8	0.18	261	7
Znf281	Q99LI5	0.26	0.45	514	10	0.19	251	5	0.15	220	4
Cdk9	Q99J95	0.26	0.52	261	5	0.00	0	0	0.26	88	3
Ctbp2	E9Q0T4	0.25	0.44	243	5	0.16	81	2	0.16	62	2
Nfib	A2ADI3	0.24	0.49	328	7	0.00	0	0	0.24	180	4
Ilf2	Q9CXY6	0.24	0.26	108	3	0.17	126	3	0.28	128	3
Gatad1	Q920S3	0.24	0.45	116	3	0.26	92	2	0.00	0	0
Ogt	Q8CGY8	0.23	0.49	631	13	0.12	200	4	0.09	153	3
Rfx3	G5E890	0.23	0.41	367	7	0.14	86	2	0.13	146	3
Emsy	Q8BMB0	0.22	0.38	647	12	0.16	294	6	0.13	253	5
Bptf	A2A654	0.22	0.20	812	18	0.25	1194	21	0.21	1112	19
Cdk2	P97377	0.21	0.44	184	4	0.19	62	2	0.00	0	0
Bre	Q8K3W0	0.21	0.28	116	3	0.26	123	3	0.08	47	1
Rcor2	Q8C796	0.21	0.36	254	5	0.26	121	3	0.00	0	0
Cxxc5	Q91WA4	0.20	0.38	130	3	0.00	0	0	0.22	57	2
Rcor1	Q8CFE3	0.20	0.23	132	3	0.29	142	3	0.07	97	2
Skp1	Q9WTX5	0.19	0.20	70	1	0.19	78	1	0.19	74	1
Ehmt2	A2CG77	0.19	0.08	139	4	0.35	426	10	0.13	181	5
Gtf2i	Q9ESZ8	0.18	0.47	403	11	0.07	76	2	0.00	0	0
Zeb2	Q9R0G7	0.17	0.17	253	7	0.25	326	8	0.10	132	4
Ncoa2	Q61026	0.17	0.37	659	13	0.00	0	0	0.14	290	6
Ehmt1	E9Q5A3	0.16	0.13	301	5	0.24	408	8	0.12	227	4
Znf143	O70230	0.15	0.30	196	5	0.10	69	2	0.05	41	1
Znf24	Q91VN1	0.15	0.18	125	3	0.00	0	0	0.27	98	3
Kat8	Q9D1P2	0.14	0.00	0	0	0.21	113	3	0.21	134	3
Zbtb43	G3X9N4	0.13	0.00	0	0	0.20	59	1	0.20	57	1
Eya4	Q8BY78	0.13	0.24	183	4	0.11	110	2	0.05	72	1
Ash2l	E9PU93	0.13	0.21	164	3	0.19	115	3	0.00	0	0
Hcfc1	Q61191	0.13	0.33	768	18	0.03	80	2	0.03	72	2
Zfp644	E9QA22	0.13	0.19	328	8	0.15	369	6	0.05	76	2
Maged1	Q9QYH6	0.13	0.08	340	7	0.04	92	1	0.26	134	2
Neurod1	Q60867	0.12	0.19	100	2	0.18	86	2	0.00	0	0
Chd7	A2AJK6	0.12	0.11	447	10	0.18	693	16	0.07	292	7
Ccnc	Q3UXL9	0.11	0.12	61	1	0.11	54	1	0.11	53	1
Nacc1	Q7TSZ8	0.11	0.13	149	3	0.13	114	2	0.06	43	1
C130039O16Rik	E9Q2I4	0.09	0.16	206	5	0.06	115	2	0.06	111	2
Hmg20a	Q9DC33	0.09	0.00	0	0	0.18	85	2	0.09	65	1
Brca1	P48754	0.08	0.20	538	11	0.05	141	4	0.00	0	0
Smad4	P97471	0.08	0.19	142	3	0.00	0	0	0.06	63	1
Znf609	Q8BZ47	0.08	0.07	158	3	0.09	157	5	0.07	139	3
Ski	B1AUF1	0.08	0.14	96	3	0.09	61	2	0.00	0	0
Kansl3	A2RSY1	0.08	0.12	147	3	0.07	122	2	0.04	95	1
Zfp462	B1AWL2	0.08	0.12	337	9	0.05	147	4	0.06	205	6
Zfp655	Q9CZP3	0.08	0.12	62	2	0.11	81	2	0.00	0	0
Rfx1	P48377	0.07	0.15	170	4	0.00	0	0	0.07	100	2

			Experiment 1			Experiment 2			Experiment 3		
Protein	**Accession**	**Average emPAI**	**EmPAI**	**Mascot**	**Uniq. pept.[a]**	**EmPAI**	**Mascot**	**Uniq. pept.[a]**	**Empai**	**Mascot**	**Uniq. pept.[a]**
Ep300	E9PYJ8	0.07	0.19	573	13	0.00	0	0	0.03	65	2
Kansl2	F8WJE3	0.07	0.15	103	2	0.00	0	0	0.07	69	1
Phc3	D3YY34	0.07	0.11	145	3	0.10	146	3	0.00	0	0
Ubr5	Q80TP3	0.07	0.18	567	15	0.03	0	0	0.00	168	3
Setd2	E9Q5F9	0.07	0.12	314	9	0.05	171	4	0.04	64	1
Zfr	O88532	0.07	0.17	274	6	0.03	66	1	0.00	0	0
Sp1	G3X8Q0	0.06	0.09	124	2	0.04	49	1	0.04	84	1
Samd4b	Q80XS6	0.05	0.10	88	2	0.00	0	0	0.05	58	1
Patz1	Q5NBY9	0.05	0.10	77	2	0.00	0	0	0.05	79	1
Phf12	Q5SPL2	0.04	0.10	116	3	0.03	63	1	0.00	0	0
Sap130	J3QNK5	0.04	0.07	104	2	0.06	64	2	0.00	0	0
Zfp318	F6R9J2	0.03	0.09	122	3	0.00	0	0	0.01	55	1
Kdm6a	O70546	0.03	0.02	64	1	0.07	165	3	0.00	0	0
Mll2	Q6PDK2	0.01	0.02	62	2	0.01	53	1	0.01	48	1
Other											
Ctnnd1	E9Q903	2.11	3.92	2165	35	1.12	1082	20	1.30	1268	21
Akap8	Q9DBR0	0.45	0.68	578	11	0.42	470	8	0.25	317	5
Ubap2l	Q80X50	0.31	0.45	546	12	0.19	271	6	0.30	425	9
Zcchc8	Q9CYA6	0.28	0.45	342	7	0.19	152	3	0.19	176	4
Brcc3	P46737	0.27	0.37	124	3	0.22	93	2	0.22	74	2
Otud4	B2RRE7	0.23	0.19	223	6	0.39	666	12	0.12	215	4
Qser1	A2BIE1	0.16	0.29	655	14	0.14	392	7	0.04	121	2
Rnf219	Q8K2Y0	0.15	0.31	268	6	0.00	0	0	0.13	84	3
Fam175a	Q8BPZ8	0.14	0.26	103	3	0.16	116	2	0.00	0	0
Rif1	Q6PR54	0.14	0.23	750	15	0.14	447	10	0.05	172	4
Ahdc1	Q6PAL7	0.09	0.07	162	3	0.08	229	4	0.11	269	5

[a] unique peptides

Table S5. : Interacting proteins of Olig2 as identified by mass spectrometry analysis of purified protein sample

			Experiment 1			Experiment 2			Experiment 3		
Protein	**Accession**	**Average emPAI**	**EmPAI**	**Mascot**	**Uniq. pept.[a]**	**EmPAI**	**Mascot**	**Uniq. pept.[a]**	**Empai**	**Mascot**	**Uniq. pept.[a]**
Olig2	Q9EQW6	93.96	48.21	1244	17	225.74	1683	23	7.92	885	12
Apc complex											
Anapc4	Q91W96	0.09	0.20	68	2	0.08	149	4	0.00	0	0
Anapc7	Q9WVM3	0.06	0.13	99	2	0.06	85	1	0.00	0	0
Anapc1	P53995	0.03	0.03	71	2	0.05	134	3	0.00	0	0
Cnot complex											
Cnot1	B7ZWL1	0.15	0.18	712	13	0.26	901	20	0.00	0	0
Cnot3	F6YBV1	0.09	0.09	74	2	0.19	64	2	0.00	0	0
Cnot2	Q8C5L3	0.08	0.06	73	2	0.12	83	2	0.07	117	3
Cnot10	Q8BH15	0.07	0.13	131	3	0.08	81	2	0.00	0	0
Ncor-Tbl1xr1 complex											
Tbl1xr1	Q8BHJ5	1.13	1.31	580	11	1.60	825	13	0.47	373	6
Tbl1x	Q9QXE7	0.66	0.70	447	8	0.90	675	10	0.37	329	5
Ncor2	F6Z4B2	0.38	0.56	1611	36	0.41	1398	29	0.17	786	15
Ncor1	E9Q8K6	0.11	0.13	494	10	0.15	625	11	0.04	273	7
NuRD complex											
Mbd3	D3YTR5	3.11	3.70	675	12	4.54	657	11	1.08	503	8
Gatad2b	Q8VHR5	2.93	4.18	1396	23	3.21	1307	23	1.41	976	16
Gatad2a	E9QMN5	1.88	2.86	1188	22	2.15	1134	20	0.62	901	12
Chd4 (Mi2-beta)	E9QAS5	0.73	0.64	1227	29	1.04	1943	42	0.52	1568	32
Mbd2	Q9Z2E1	0.17	0.26	163	3	0.26	172	3	0.00	0	0
Spalt proteins											
Sall3	Q62255	1.12	1.38	1612	29	1.44	1562	28	0.55	1148	19
Sall1	Q6P5E3	0.70	0.65	821	16	1.20	1536	28	0.26	755	12
Sall2	F7ARK3	0.65	0.82	920	17	0.93	1090	17	0.19	384	6
Swi-Snf complex											
Actl6a	Q9Z2N8	0.84	1.16	603	10	1.01	535	9	0.35	186	3
Smarce1	O54941	0.46	0.43	272	5	0.78	380	7	0.17	108	3
Smarcc1	P97496	0.39	0.43	527	12	0.68	775	17	0.07	307	7
Smarcc2	Q3UID0	0.32	0.42	653	11	0.54	795	13	0.00	0	0
Smarcd1	Q61466	0.30	0.06	65	1	0.78	398	9	0.06	71	1
Smarcb1	F6U415	0.21	0.16	78	2	0.46	214	5	0.00	0	0
Arid1a	A2BH40	0.01	0.01	80	1	0.01	80	1	0.00	0	0
Trrap complex											
Ruvbl2	Q9WTM5	3.09	3.53	1205	18	4.87	1391	21	0.88	464	10
Ruvbl1	P60122	1.73	1.54	648	12	3.32	1084	18	0.33	251	6
Brd8	Q8R3B7	0.13	0.24	259	6	0.15	205	4	0.00	0	0
Trrap	E9PZA7	0.08	0.08	466	10	0.12	679	15	0.03	343	10
Transcription factors											
Olig1	Q9JKN5	11.83	11.85	950	14	22.47	1165	16	1.18	411	6
Hdac2	P70288	2.97	3.26	758	13	4.74	860	14	0.91	611	11
Cbx3	P23198	2.46	3.14	353	6	3.83	447	7	0.40	132	2
Hoxa5	P09021	1.39	1.78	392	7	2.11	478	9	0.27	153	2
Rnf2	Q9CQJ4	1.02	1.64	411	9	1.41	340	7	0.00	0	0
Ubp1	Q811S7	0.93	1.44	767	13	1.30	797	12	0.06	65	1
Tceb2	P62869	0.92	1.09	115	3	1.68	198	4	0.00	0	0
Mta3	E9Q794	0.90	1.23	598	13	1.22	671	13	0.24	378	6
Tceb1	P83940	0.88	1.16	104	3	1.16	175	3	0.31	129	1
Sox2	Q60I23	0.84	1.18	390	7	0.97	299	6	0.37	162	3
Kdm1a	A3KG93	0.76	1.11	1069	22	1.04	975	20	0.12	225	4
Trps1	Q925H1	0.67	0.99	1369	29	0.90	1328	26	0.11	336	8
Sox5	B2KFM6	0.64	1.14	745	19	0.79	505	13	0.00	0	0
Hdac3	Q3UM33	0.61	0.85	428	9	0.98	520	9	0.00	0	0
Satb2	Q8VI24	0.56	0.71	443	11	0.97	648	14	0.00	0	0
Tfcp2	Q9ERA0	0.52	0.60	379	7	0.91	508	9	0.06	76	1
Tcf12	Q61286	0.52	0.86	658	14	0.49	449	9	0.21	325	6

			Experiment 1			Experiment 2			Experiment 3		
Protein	Accession	Average emPAI	EmPAI	Mascot	Uniq. pept.[a]	EmPAI	Mascot	Uniq. pept.[a]	Empai	Mascot	Uniq. pept.[a]
Maged1	Q9QYH6	0.52	0.81	686	15	0.74	743	15	0.00	0	0
Zfp219	Q6IQX8	0.51	0.67	593	11	0.75	629	12	0.10	140	2
Zhx2	Q8C0C0	0.50	0.55	830	17	0.94	582	14	0.00	0	0
Nfia	B1AUB6	0.48	0.75	387	8	0.55	421	7	0.15	66	1
Wiz	F6ZBR8	0.46	0.57	533	12	0.67	659	14	0.15	222	5
Wdr5	F6Q3W0	0.43	0.57	255	5	0.72	266	6	0.00	0	0
Nacc1	Q7TSZ8	0.42	0.60	366	8	0.60	426	8	0.07	107	1
Cux1	H3BK24	0.40	0.61	1155	22	0.51	988	20	0.09	287	6
Foxk1	P42128	0.40	0.56	561	12	0.64	448	9	0.00	0	0
Znf148	Q61624	0.40	0.26	286	6	0.90	986	16	0.04	87	1
Ogt	Q8CGY8	0.40	0.58	795	16	0.58	802	16	0.03	107	1
Rqcd1	Q9JKY0	0.37	0.64	206	5	0.48	226	4	0.00	0	0
Bcl7c	O08664	0.36	0.33	56	2	0.76	130	4	0.00	0	0
Ctbp2	E9Q0T4	0.34	0.61	304	7	0.41	237	5	0.00	0	0
Hcfc1	Q61191	0.33	0.52	1108	24	0.37	919	19	0.09	325	5
Rfx3	G5E890	0.32	0.52	514	10	0.44	339	6	0.00	0	0
Znf24	Q91VN1	0.32	0.48	235	5	0.48	238	5	0.00	0	0
Tcf4	Q91YV0	0.32	0.39	325	7	0.28	265	4	0.29	272	5
Yy1	Q00899	0.30	0.35	140	4	0.56	280	6	0.00	0	0
Zbtb20	Q8K0L9	0.30	0.33	286	6	0.57	368	7	0.00	0	0
Gps2	Q921N8	0.30	0.31	115	3	0.58	198	5	0.00	0	0
Nacc2	Q9DCM7	0.29	0.17	144	3	0.69	547	10	0.00	0	0
Cdk9	Q99J95	0.28	0.48	247	5	0.26	185	3	0.09	57	1
Sox6	E9PUW0	0.26	0.42	328	9	0.36	282	8	0.00	0	0
Pbx1	P41778	0.25	0.43	219	5	0.33	240	4	0.00	0	0
Znf281	Q99LI5	0.25	0.42	386	9	0.28	324	6	0.04	86	2
Cxxc5	Q91WA4	0.23	0.35	112	3	0.35	135	3	0.00	0	0
Gtf2i	Q9ESZ8	0.23	0.22	238	6	0.48	526	12	0.00	0	0
Trim33	E9QME5	0.23	0.24	483	8	0.42	575	12	0.03	69	1
Rcor2	Q8C796	0.23	0.26	157	4	0.42	237	6	0.00	0	0
Prmt1	D3Z0A2	0.22	0.44	175	4	0.09	43	1	0.13	47	1
Tead1	E9Q387	0.22	0.33	162	4	0.33	152	4	0.00	0	0
Adnp	Q9Z103	0.22	0.34	395	10	0.31	536	11	0.00	0	0
Zkscan3	Q91VW9	0.19	0.24	278	7	0.05	110	4	0.28	84	2
Nr2f1	Q32NY6	0.19	0.24	87	2	0.33	199	5	0.00	0	0
Etv6	E9Q8J8	0.18	0.34	129	4	0.20	114	3	0.00	0	0
Sin3a	Q60520	0.17	0.26	490	10	0.26	495	10	0.00	0	0
Sox8	Q04886	0.15	0.32	75	2	0.14	82	2	0.00	0	0
Ash2l	E9PU93	0.15	0.32	215	5	0.12	71	2	0.00	0	0
Ccnt1	Q9QWV9	0.14	0.29	291	6	0.09	91	2	0.05	46	1
Zeb1	H3BLP5	0.14	0.18	274	6	0.25	264	5	0.00	0	0
Zbtb33	Q8BN78	0.14	0.16	117	3	0.22	202	4	0.05	68	1
Cul3	Q9JLV5	0.14	0.26	258	6	0.16	159	4	0.00	0	0
Wdr82	Q8BFQ4	0.14	0.21	104	3	0.21	92	2	0.00	0	0
Ubr5	Q80TP3	0.13	0.23	551	12	0.14	984	18	0.02	237	6
Hmg20a	Q9DC33	0.12	0.18	95	2	0.18	67	2	0.00	0	0
Rcor1	Q8CFE3	0.12	0.07	53	1	0.29	221	4	0.00	0	0
Morf4l1	P60762	0.12	0.18	67	2	0.17	73	2	0.00	0	0
Arnt	E9QLT6	0.11	0.21	220	5	0.13	147	3	0.00	0	0
Dpf2	F8WIP7	0.10	0.07	79	1	0.24	188	3	0.00	0	0
Kdm3b	Q6ZPY7	0.10	0.17	361	9	0.13	286	7	0.00	0	0
Zmym2	Q9CU65	0.09	0.11	193	5	0.16	334	7	0.00	0	0
Zhx1	P70121	0.09	0.15	210	4	0.11	167	4	0.00	0	0
Ehmt1	E9Q5A3	0.08	0.12	192	4	0.12	240	5	0.00	0	0
Smad4	P97471	0.08	0.12	89	2	0.12	112	2	0.00	0	0
Cdc23	G3X8W7	0.07	0.11	114	2	0.11	108	2	0.00	0	0
Foxp4	D3Z726	0.07	0.10	61	2	0.10	60	2	0.00	0	0
Kansl3	A2RSY1	0.06	0.04	105	1	0.15	231	4	0.00	0	0
Sp1	G3X8Q0	0.06	0.09	174	2	0.09	154	2	0.00	0	0
Zhx3	Q8C0Q2	0.06	0.07	129	2	0.10	216	4	0.00	0	0
Bre	Q8K3W0	0.05	0.08	81	1	0.08	61	1	0.00	0	0
Nr3c1	P06537	0.05	0.08	64	2	0.08	144	3	0.00	0	0
Ehmt2	A2CG77	0.05	0.05	82	2	0.10	239	4	0.00	0	0
Zfp462	B1AWL2	0.05	0.04	116	3	0.11	348	9	0.00	0	0

			Experiment 1			Experiment 2			Experiment 3		
Protein	**Accession**	**Average emPAI**	**EmPAI**	**Mascot**	**Uniq. pept.[a]**	**EmPAI**	**Mascot**	**Uniq. pept.[a]**	**Empai**	**Mascot**	**Uniq. pept.[a]**
Patz1	Q5NBY9	0.03	0.05	57	1	0.05	66	1	0.00	0	0
Znf609	Q8BZ47	0.03	0.07	112	3	0.02	66	1	0.00	0	0
Smarca5	Q91ZW3	0.03	0.00	0	0	0.06	74	2	0.03	80	2
Ep300	E9PYJ8	0.03	0.05	176	4	0.03	65	2	0.00	0	0
Chd7	A2AJK6	0.02	0.00	0	0	0.03	118	3	0.04	254	4
Sap130	J3QNK5	0.02	0.03	69	1	0.03	55	1	0.00	0	0
Other											
Ppm1g	Q61074	0.88	0.49	343	6	1.78	959	16	0.36	407	5
Ctnnd1	E9Q903	0.85	1.33	1369	26	1.10	1245	22	0.11	336	8
Akap8	Q9DBR0	0.55	0.77	604	13	0.77	605	13	0.10	94	2
Lmnb1	P14733	0.50	0.66	472	10	0.83	487	12	0.00	0	0
Mlf2	Q99KX1	0.29	0.27	78	2	0.61	121	3	0.00	0	0
Cul2	Q9D4H8	0.23	0.26	436	9	0.42	217	7	0.00	0	0
Ubap2l	Q80X50	0.14	0.16	244	5	0.26	363	7	0.00	0	0
Zbtb10	E9Q8X5	0.13	0.33	315	8	0.07	69	2	0.00	0	0
Cul4b	E9PXY1	0.12	0.24	131	4	0.13	283	7	0.00	0	0
Qser1	A2BIE1	0.11	0.14	320	7	0.16	435	8	0.02	119	2
Smc3	Q9CW03	0.05	0.10	140	4	0.05	79	2	0.00	0	0
Smc1a	Q9CU62	0.04	0.05	89	2	0.05	78	2	0.03	63	1

[a] unique peptides

Table S6. : Interacting proteins of Npas3 as identified by mass spectrometry analysis of purified protein sample

			Experiment 1			Experiment 2		
Protein	**Accession**	**Average emPAI**	**EmPAI**	**Mascot**	**Uniq. pept.[a]**	**EmPAI**	**Mascot**	**Uniq. pept.[a]**
Npas3	F8VQB2	9.78	10.55	2679	37	9	2570	38
Cnot complex								
Cnot1	B7ZWL1	0.11	0.19	539	13	0.03	67	2
Ncor-Tbl1xr1 complex								
Tbl1xr1	Q8BHJ5	2.00	2.38	1080	15	1.62	711	12
Tbl1x	Q9QXE7	1.14	1.41	763	12	0.87	406	9
Ncor2	F6Z4B2	0.64	0.87	2084	45	0.41	1270	28
Ncor1	E9Q8K6	0.35	0.58	1891	35	0.11	473	9
NuRD complex								
Gatad2a	E9QMN5	1.02	0.80	567	11	1.23	781	15
Mbd3	D3YTR5	0.87	1.16	398	7	0.57	174	4
Chd4 (Mi2-beta)	E9QAS5	0.75	0.88	1795	39	0.62	1300	32
Spalt proteins								
Sall3	Q62255	0.94	1.01	1328	21	0.86	1143	21
Sall2	F7ARK3	0.47	0.47	655	11	0.47	605	10
Swi-Snf complex								
Smarcb1	F6U415	1.06	1.93	128	2	0.18	83	2
Actl6a	Q9Z2N8	0.82	0.82	495	8	0.82	420	8
Trrap complex								
Ruvbl2	Q9WTM5	2.06	2.80	1097	17	1.32	787	13
Ep400	F6R9G0	0.68	1.28	842	20	0.08	247	7
Trrap	E9PZA7	0.19	0.16	800	19	0.22	913	22
Transcription factors								
Arnt2	Q61324	7.56	7.75	2182	29	7.36	2133	28
Arnt	E9QLT6	3.89	3.61	1774	26	4.17	1852	30
Sox2	Q60I23	1.43	1.30	261	5	1.55	282	6
Hdac1	O09106	1.11	2.21	747	13	0.00	421	8
Olig1	Q9JKN5	1.08	0.68	205	4	1.48	279	6
Nfia	B1AUB6	1.05	1.38	525	11	0.71	288	6
Zbtb20	Q8K0L9	0.81	0.85	671	12	0.77	576	11
Trps1	Q925H1	0.78	0.89	1308	24	0.66	988	20
Nfix	D3YZ00	0.78	0.95	484	10	0.60	363	7
Wdr5	F6Q3W0	0.62	0.62	314	6	0.62	210	5
Nacc1	Q7TSZ8	0.56	0.66	373	8	0.46	279	6
Znf148	Q61624	0.49	0.69	571	11	0.28	266	6
Mta3	E9Q794	0.49	0.53	352	7	0.44	320	6
Rnf2	Q9CQJ4	0.49	0.76	231	5	0.21	65	2
Cxxc5	Q91WA4	0.46	0.54	168	4	0.38	117	3
Ilf2	Q9CXY6	0.39	0.39	109	3	0.39	122	3
Nfib	A2ADI3	0.39	0.58	341	8	0.19	167	4
Zfp219	Q6IQX8	0.38	0.38	336	7	0.38	272	6
Nacc2	Q9DCM7	0.37	0.49	334	7	0.25	190	4
Chd7	A2AJK6	0.36	0.40	1480	31	0.31	1133	25
Hdac3	Q3UM33	0.35	0.44	251	5	0.25	116	3
Rlim	Q9WTV7	0.35	0.38	340	6	0.31	222	5
Cdk9	Q99J95	0.34	0.39	184	4	0.28	135	3
Rcor1	Q8CFE3	0.33	0.51	296	6	0.15	84	2
Wdr82	Q8BFQ4	0.33	0.35	175	3	0.31	88	2
Wiz	F6ZBR8	0.33	0.46	574	11	0.19	191	4
Kdm1a	A3KG93	0.31	0.36	486	8	0.26	302	6
Tcf12	Q61286	0.30	0.33	254	6	0.27	194	5
Rcor2	Q8C796	0.29	0.45	275	6	0.13	77	2
Prmt1	D3Z0A2	0.28	0.34	103	3	0.21	85	2
Kansl2	F8WJE3	0.27	0.39	204	5	0.15	104	2
Znf281	Q99LI5	0.25	0.25	369	7	0.25	293	6
Sox8	Q04886	0.25	0.34	90	2	0.15	75	2
Olig2	Q9EQW6	0.24	0.24	102	2	0.24	63	2

			Experiment 1			Experiment 2		
Protein	Accession	Average emPAI	EmPAI	Mascot	Uniq. pept.[a]	EmPAI	Mascot	Uniq. pept.[a]
Ski	B1AUF1	0.24	0.43	411	8	0.05	58	1
Smarca5	Q91ZW3	0.24	0.31	439	9	0.16	231	5
Foxk1	P42128	0.22	0.12	119	3	0.32	103	3
Gps2	Q921N8	0.22	0.22	101	2	0.22	91	2
Sin3a	Q60520	0.19	0.22	408	8	0.16	236	6
Bptf	A2A654	0.18	0.28	979	22	0.08	262	7
Ehmt2	A2CG77	0.18	0.27	359	9	0.08	150	4
Znf609	Q8BZ47	0.18	0.30	581	11	0.05	95	2
Ubp1	Q811S7	0.17	0.20	186	3	0.13	83	2
Maged1	Q9QYH6	0.17	0.09	286	7	0.24	76	1
Emsy	Q8BMB0	0.16	0.17	350	7	0.14	216	5
Zhx2	Q8C0C0	0.15	0.12	79	2	0.17	133	3
Cux1	H3BK24	0.14	0.19	369	8	0.09	155	4
Kansl3	A2RSY1	0.14	0.16	145	4	0.12	151	3
Patz1	Q5NBY9	0.13	0.10	80	2	0.15	109	3
Ehmt1	E9Q5A3	0.12	0.19	325	7	0.05	94	2
Zhx1	P70121	0.12	0.12	112	3	0.12	88	2
Rfx3	G5E890	0.12	0.09	119	2	0.14	125	3
Setd2	E9Q5F9	0.12	0.15	477	11	0.08	272	7
Zeb2	Q9R0G7	0.11	0.17	242	6	0.05	61	2
Kansl1	A2A5Y4	0.10	0.16	230	5	0.03	68	2
Cbx6	Q9DBY5	0.09	0.09	81	1	0.09	73	1
C130039O16Rik	E9Q2I4	0.08	0.13	170	4	0.03	58	1
Sp1	G3X8Q0	0.05	0.05	115	1	0.05	97	1
Ubr5	Q80TP3	0.05	0.04	220	5	0.05	136	3
Other								
Mlf2	Q99KX1	0.38	0.29	126	2	0.46	93	2
Ctnnd1	E9Q903	0.36	0.37	458	9	0.34	363	8
Akap8	Q9DBR0	0.30	0.33	315	6	0.27	259	6
Qser1	A2BIE1	0.28	0.45	989	19	0.10	176	5
Lmo7	E9PYF4	0.26	0.21	318	7	0.31	448	10
Ubap2l	Q80X50	0.19	0.28	358	7	0.10	134	3
Zbtb45	Q52KG4	0.14	0.14	108	2	0.14	104	2
Lmnb1	P14733	0.11	0.11	128	2	0.11	82	2
Ahdc1	Q6PAL7	0.06	0.07	112	3	0.04	90	2

[a] unique peptides

Table S7. : Interacting proteins of Sox2

			Experiment 1			Experiment 2		
Protein	**Accession**	**Average emPAI**	**EmPAI**	**Mascot**	**Uniq. pept.[a]**	**EmPAI**	**Mascot**	**Uniq. pept.[a]**
Sox2	Q60I23	1.72	1.3	327	5	2.14	531	5
Cnot complex								
Cnot1	B7ZWL1	0.05	0.04	156	3	0.06	245	4
Ncor-Tbl1xr1 complex								
Tbl1xr1	Q8BHJ5	0.43	0.38	248	5	0.47	406	4
Ncor2	F6Z4B2	0.31	0.17	712	13	0.45	1624	28
Hdac3	Q3UM33	0.25	0.16	162	3	0.34	252	4
Sin3a	Q60520	0.08	0.03	63	2	0.13	287	5
NuRD complex								
Hdac1	O09106	1.79	0.68	589	9	2.9	1029	14
Hdac2	P70288	1.46	0.91	535	9	2	797	11
Gatad2b	Q8VHR5	1.44	1.04	743	13	1.84	1310	15
Mta2	Q9R190	1.12	0.86	1005	15	1.37	1140	16
Mta1	F8WHY8	1.05	0.72	832	12	1.37	1274	18
Mbd3	D3YTR5	0.95	0.73	318	4	1.16	503	7
Gatad2a	E9QMN5	0.66	0.31	394	8	1	906	12
Chd4 (Mi2-beta)	E9QAS5	0.60	0.49	1307	24	0.7	1932	32
Spalt proteins								
Sall3	Q62255	1.15	0.4	774	13	1.89	2366	30
Sall2	F7ARK3	0.59	0.27	407	6	0.9	1362	16
Sall1	Q6P5E3	0.42	0.17	345	5	0.67	1296	17
Swi-Snf complex								
Smarcb1 (Ini1)	F6U415	0.71	0.71	67	1	0.71	72	1
Actl6a	Q9Z2N8	0.57	0.57	464	6	0.57	434	6
Smarcc2 (Baf170)	Q3UID0	0.33	0.16	338	5	0.5	1058	13
Smarcc1 (Baf155)	P97496	0.20	0.18	323	5	0.22	403	6
Smarcd1 (Baf60a)	Q61466	0.17	0.13	101	2	0.2	198	3
Smarce1	O54941	0.08	0.08	55	1	0.08	72	1
Trrap complex								
Ruvbl1	P60122	0.89	0.24	214	3	1.53	980	12
Ruvbl2	Q9WTM5	0.82	0.32	265	4	1.32	833	12
Trrap	E9PZA7	0.04	0.04	207	5	0.04	193	4
Transcription factors								
Sox8	Q04886	0.62	0.33	217	4	0.91	264	5
Cux1	H3BK24	0.48	0.22	580	10	0.74	1920	25
Tcf12	Q61286	0.37	0.05	58	1	0.69	820	11
Hoxa5	P09021	0.36	0.44	219	3	0.27	200	2
Ctbp2	E9Q0T4	0.33	0.4	128	2	0.26	199	3
Chd7	A2AJK6	0.31	0.2	1047	19	0.42	2175	30
Zeb1	H3BLP5	0.22	0.15	200	3	0.29	407	6
Tead1	E9Q387	0.17	0.08	91	2	0.26	203	3
Nacc1	Q7TSZ8	0.17	0.21	227	3	0.13	84	2
Trps1	Q925H1	0.16	0.03	97	2	0.29	748	10
Nfib	A2ADI3	0.16	0.12	138	2	0.19	200	4
Rfx3	G5E890	0.14	0.04	131	2	0.24	347	4
Twist1	P26687	0.12	0.12	118	2	0.12	72	1
Zmym2	Q9CU65	0.11	0.02	65	1	0.2	487	8
Sox5	B2KFM6	0.10	0.1	91	2	0.1	124	2
Smarca5	Q91ZW3	0.10	0.13	205	5	0.06	63	2
Zeb2	Q9R0G7	0.07	0.05	151	2	0.08	147	3
C130039O16Rik	E9Q2I4	0.06	0.06	113	2	0.06	197	2
Other								
Xpo4	Q9ESJ0	1.20	0.69	1344	22	1.71	2109	27
Rnf219	Q8K2Y0	0.07	0.05	96	3	0.09	114	2

[a] unique peptides

Table S8. Network proteins and overlap with MD-genes and constrained human genes (supplied as xlsx file), related to Figure 1. Network proteins are alphabetically listed and overlaps with human genes de novo mutated in patients with ID[18], ASD-lowIQ[19], ASD-normIQ[19] or schizophrenia[20-23] are indicated. Overlaps with evolutionary constrained human genes[24] are indicated.

Table S9. Genes with de novo mutations in schizophrenia patients from literature (supplied as xlsx file), related to Figure 1. Human genes with de novo loss of function (LOF) and missense mutations were derived from the studies of[20-22] and[23]. Genes are ordered alphabetically. Type of mutation and source publication are indicated. Frameshift mutations, nonsense mutations and splice site mutations (within 2 nucleotides from the splice donor site or splice acceptor site[20]) were taken as LOF mutations. Genes appear more than once in the list if multiple mutations were identified.

Table S10. Protein interactions between network factors associated with ID and network factors associated with ASD and/or schizophrenia, related to Figure 1. Interacting network proteins (protein A and protein B) are listed and their overlaps with human genes de novo mutated in patients with ID[18], ASD-lowIQ[19], ASD-normIQ[19] or schizophrenia[20-23] are indicated.

	Protein A	Protein B	MD protein A	MD protein B
1	Tcf4	Zeb2	ID	ID, ASD-normIQ
2	Tcf4	Chd7	ID	ID, ASD-lowIQ
3	Tcf4	Rfx3	ID	ID, schizophrenia
4	Tcf4	Nfia	ID	ID, ASD-lowIQ
5	Tcf4	Ahdc1	ID	ASD-lowIQ, schizophrenia
6	Tcf4	Setd2	ID	ASD-lowIQ, ASD-normIQ
7	Tcf4	Tbl1xr1	ID	ASD-lowIQ
8	Tcf4	Ubap2l	ID	ASD-lowIQ
9	Tcf4	Ubr5	ID	ASD-normIQ
10	Tcf4	Brca1	ID	ASD-normIQ
11	Tcf4	Ilf2	ID	ASD-lowIQ
12	Tcf4	Nfib	ID	ASD-lowIQ
13	Tcf4	Cdc23	ID	ASD-lowIQ
14	Tcf4	Nacc1	ID	ASD-lowIQ
15	Tcf4	Trrap	ID	ASD-lowIQ, ASD-normIQ, schizophrenia
16	Tcf4	Chd4	ID	ASD-lowIQ
17	Tcf4	Ep400	ID	ASD-normIQ
18	Tcf4	Ncor1	ID	ASD-lowIQ
19	Tcf4	Zfp462	ID	ASD-normIQ
20	Tcf4	Kdm1a	ID	ASD-lowIQ
21	Tcf4	Mbd2	ID	ASD-normIQ
22	Tcf4	Cnot1	ID	ASD-lowIQ
23	Tcf4	Rcor2	ID	ASD-lowIQ
24	Tcf4	Tcf3	ID	ASD-normIQ
25	Tcf4	Zeb1	ID	schizophrenia
26	Tcf4	Anapc5	ID	schizophrenia
27	Tcf4	Wiz	ID	schizophrenia
28	Tcf4	Zfr	ID	schizophrenia
29	Tcf4	Maml2	ID	schizophrenia
30	Tcf4	Qser1	ID	schizophrenia
31	Tcf4	Rif1	ID	schizophrenia
32	Tcf4	Ncor2	ID	schizophrenia
33	Tcf4	Smarcc2	ID	schizophrenia
34	Sox2	Zeb2	ID	ID, ASD-normIQ
35	Sox2	Chd7	ID	ID, ASD-lowIQ
36	Sox2	Rfx3	ID	ID, schizophrenia
37	Sox2	Tbl1xr1	ID	ASD-lowIQ
38	Sox2	Nfib	ID	ASD-lowIQ
39	Sox2	Nacc1	ID	ASD-lowIQ
40	Sox2	Trrap	ID	ASD-lowIQ, ASD-normIQ, schizophrenia
41	Sox2	Chd4	ID	ASD-lowIQ
42	Sox2	Ruvbl1	ID	ASD-normIQ
43	Sox2	Hdac1	ID	ASD-lowIQ
44	Sox2	Cnot1	ID	ASD-lowIQ
45	Sox2	Zeb1	ID	schizophrenia
46	Sox2	Smarcc2	ID	schizophrenia
47	Sox2	Ncor2	ID	schizophrenia

Table S11. Tcf4 target genes and Tcf4 binding sites (supplied as xlsx file), related to Figure 4. First tab: Tcf4 target genes. Total number of Tcf4 binding sites within 100 kb of the transcription start site (TSS) is indicated. Second tab: Tcf4 binding sites in NSCs, determined by ChIP-seq. Indicated are Chromosome (Chr), Start, end and summit of Tcf4 sequence read peak, the number of reads per peak and -10*log10(P-value), determined by MACS 1.4.2.

Table S12. Tcf4 target gene overlap with genes associated with ID, ASD, schizophrenia or primary microcephaly (supplied as xlsx file), related to Figure 4 . Tcf4 target genes (Table S11) overlapping with ID genes[18], genes de novo mutated in patients with ASD[19], genes de novo mutated in schizophrenia patients (Table S9) or genes mutated in patients with primary microcephaly are included and indicated accordingly.

SUPPLEMENTAL REFERENCES

1. van den Berg, D.L. et al. An Oct4-centered protein interaction network in embryonic stem cells. Cell Stem Cell 6, 369-81 (2010).
2. Ninkovic, J. et al. The BAF complex interacts with Pax6 in adult neural progenitors to establish a neurogenic cross-regulatory transcriptional network. Cell Stem Cell 13, 403-18 (2013).
3. Gagliardi, A. et al. A direct physical interaction between Nanog and Sox2 regulates embryonic stem cell self-renewal. EMBO J 32, 2231-47 (2013).
4. Tavares, L. et al. RYBP-PRC1 complexes mediate H2A ubiquitylation at polycomb target sites independently of PRC2 and H3K27me3. Cell 148, 664-78 (2012).
5. Engelen, E. et al. Sox2 cooperates with Chd7 to regulate genes that are mutated in human syndromes. Nat Genet 43, 607-11 (2011).
6. Landeira, D. et al. Jarid2 is a PRC2 component in embryonic stem cells required for multi-lineage differentiation and recruitment of PRC1 and RNA Polymerase II to developmental regulators. Nat Cell Biol 12, 618-24 (2010).
7. Casanova, M. et al. Polycomblike 2 facilitates the recruitment of PRC2 Polycomb group complexes to the inactive X chromosome and to target loci in embryonic stem cells. Development 138, 1471-82 (2011).
8. Gontan, C. et al. RNF12 initiates X-chromosome inactivation by targeting REX1 for degradation. Nature 485, 386-90 (2012).
9. Uranishi, K., Akagi, T., Sun, C., Koide, H. & Yokota, T. Dax1 associates with Esrrb and regulates its function in embryonic stem cells. Mol Cell Biol 33, 2056-66 (2013).
10. Wang, J. et al. A protein interaction network for pluripotency of embryonic stem cells. Nature 444, 364-8 (2006).
11. Tanimura, N., Saito, M., Ebisuya, M., Nishida, E. & Ishikawa, F. Stemness-related factor Sall4 interacts with transcription factors Oct-3/4 and Sox2 and occupies Oct-Sox elements in mouse embryonic stem cells. J Biol Chem 288, 5027-38 (2013).
12. Pardo, M. et al. An expanded Oct4 interaction network: implications for stem cell biology, development, and disease. Cell Stem Cell 6, 382-95 (2010).
13. Ang, Y.S. et al. Wdr5 mediates self-renewal and reprogramming via the embryonic stem cell core transcriptional network. Cell 145, 183-97 (2011).
14. Percharde, M. et al. Ncoa3 functions as an essential Esrrb coactivator to sustain embryonic stem cell self-renewal and reprogramming. Genes Dev 26, 2286-98 (2012).
15. Jang, H. et al. O-GlcNAc regulates pluripotency and reprogramming by directly acting on core components of the pluripotency network. Cell Stem Cell 11, 62-74 (2012).
16. Esch, D. et al. A unique Oct4 interface is crucial for reprogramming to pluripotency. Nat Cell Biol 15, 295-301 (2013).
17. Ishihama, Y. et al. Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein. Mol Cell Proteomics 4, 1265-72 (2005).
18. Gilissen, C. et al. Genome sequencing identifies major causes of severe intellectual disability. Nature 511, 344-7 (2014).
19. Iossifov, I. et al. The contribution of de novo coding mutations to autism spectrum disorder. Nature 515, 216-21 (2014).
20. Fromer, M. et al. De novo mutations in schizophrenia implicate synaptic networks. Nature 506, 179-84 (2014).

21. Xu, B. et al. De novo gene mutations highlight patterns of genetic and neural complexity in schizophrenia. Nat Genet 44, 1365-9 (2012).
22. Gulsuner, S. et al. Spatial and temporal mapping of de novo mutations in schizophrenia to a fetal prefrontal cortical network. Cell 154, 518-29 (2013).
23. Guipponi, M. et al. Exome sequencing in 53 sporadic cases of schizophrenia identifies 18 putative candidate genes. PLoS ONE 9, e112745 (2014).
24. Samocha, K.E. et al. A framework for the interpretation of de novo mutation in human disease. Nat Genet 46, 944-50 (2014).

Chapter4

Interdependency of Oct4, Sox2 and Nanog localization on the Embryonic Stem Cell Genome

Johannes H. Brandsma*[1], Debbie L.C. van den Berg*[1,2], Florian Halbritter[3], Mike Dekker[1], Zeliha Ozgür[4], Christel E.M. Kockx[4], Wilfred F.J. van IJcken[4], Simon R. Tomlinson[3], Raymond A. Poot[1]

[1] Department of Cell Biology, Erasmus MC, Wytemaweg 80, 3015 CN Rotterdam, The Netherlands
[2] The Francis Crick Institute, Mill Hill Laboratory, The Ridgeway, London NW7 1AA, United Kingdom
[3] MRC Centre for Regenerative Medicine, Institute for Stem Cell Research, School of Biological Sciences, University of Edinburgh, Edinburgh, EH9 3FF, United Kingdom
[4] Center for Biomics, Erasmus MC, Wytemaweg 80, 3015 CN, Rotterdam, The Netherlands

*** These authors contributed equally to this work**

Work in progress

ABSTRACT
Sox2, Nanog and Oct4 are prominent transcription factors in the regulation of pluripotency of Embryonic Stem Cells (ESCs). Sox2, Nanog and Oct4 often co-localize on the ESC genome, together with other transcription factors. Sox2 and Oct4 interact and were shown to bind cooperatively to some of their genomic binding sites. Nanog and Sox2 were recently shown to have a strong protein-protein interaction. To elucidate the interdependency of Oct4, Sox2 and Nanog genome binding, we performed ChIP-seq experiments for these three transcription factors upon acute in vivo depletion of Oct4 or Sox2. We show that on approximately half the Oct4-Sox2-Nanog sites, binding of Sox2 or Nanog depends on Oct4 presence. These sites harbor the classical composite Sox-Oct motif. Interestingly, changes in the binding of Sox2 and Nanog upon Oct4 depletion are highly correlated. Sox2 and Nanog also bind independently of Oct4 to sites containing a Sox motif. Together this suggest that on Oct4-Sox2-Nanog sites, Oct4 recruits Sox2 together with Nanog. This hierarchy of recruitment matches the reported Oct4-Sox2 and Nanog-Sox2 protein-protein interactions.

INTRODUCTION

Oct4, Sox2 and Nanog are transcription factors that form part of the transcriptional regulatory circuit in Embryonic Stem Cells (ESCs) that maintains pluripotency and promotes self-renewal [1,2]. Oct4, Sox2 and Nanog co-localize to the same regulatory regions in ESCs[1,3]. While the affinity of the protein-protein interaction between Oct4 and Sox2 is low, this interaction is required to cooperatively bind to the genome and activate gene transcription[4]. Sox2 and Nanog also have high affinity interaction and this interaction is important to maintain the self-renewal in ESCs[5]. The composite Sox2-Oct4 motif (SO-motif) was previously discovered to be enriched for Oct4, Sox2 or Nanog binding sites in ESCs[1,3,6] and is a composite motif consisting of a motif associated with HMG domain transcription factors, like Sox family transcription factors, and a motif associated with POU domain transcription factors. While the SO-motif was found as the main motif in Oct4 and Sox2 binding sites in ESC[1,3,6], efficient cooperative binding of Sox2 and Oct4 to the SO-motif was never investigated genome-wide and has only been demonstrated for an enhancer site near UTF1[4,7] and mutant FGF4 enhancers[4]. To investigate this genome-wide, we depleted Oct4 and Sox2 in the doxycycline inducible Oct4-null ESC line ZHBTc4 and Sox2-null ESC line 2TS22c, respectively [8,9] and performed chromatin immunoprecipitation (ChIP) followed by massive parallel DNA sequencing (ChIP-seq) to determine the genome-wide binding sites of Sox2, Oct4 and Nanog after depletion of Oct4 or Sox2. We find that Oct4 and Sox2 require each other to bind efficiently to genome-wide binding sites containing the SO-motif. However, binding sites where Oct4 depends on Sox2 for co-localization do not necessarily overlap with binding sites where Sox2 depends on Oct4. We also find that that loss of Sox2 and Nanog binding upon depletion of Oct4 highly correlates and that Sox2 and Nanog also bind the genome independent of Oct4.

RESULTS

Oct4 dependent recruitment of Sox2 and Nanog

To characterize the dependency of Sox2 and Nanog on Oct4 in binding the ESC genome,

ChIP-seq experiments were performed against Oct4, Sox2 and Nanog before (from here on called wt-Oct4) and after 16 hours of Oct4 depletion in the doxycycline inducible Oct4-null ESC line ZHBTc4[8,9] (from here on called ΔOct4). The 16 hour time point was chosen, because it has a minimal effect on Sox2 and Nanog transcript levels, while Oct4 transcript and protein levels are reduced to almost zero[9]. Western Blotting confirmed that the Oct4 protein was no longer detected after 16 hours in the depleted condition, while Sox2 and Nanog proteins levels remained unchanged (Figure S1). To confirm that Oct4 depletion also resulted in reduced binding of Oct4 to the genome, Oct4 binding sites before and after Oct4 depletion were compared and over 90 percent of all genome-wide Oct4 binding sites show at least 2-fold reduced binding signal (Figure 1a). As an example, Oct4 binding sites around the *Nanog* and *Utf1* genes are shown (Figure 1b). The ZHBTc4 ESC line has previously been used to investigate the effect of Oct4 depletion upon gene expression in ESCs[8]. We find no correlation upon Oct4 depletion between the fold-change reduction in

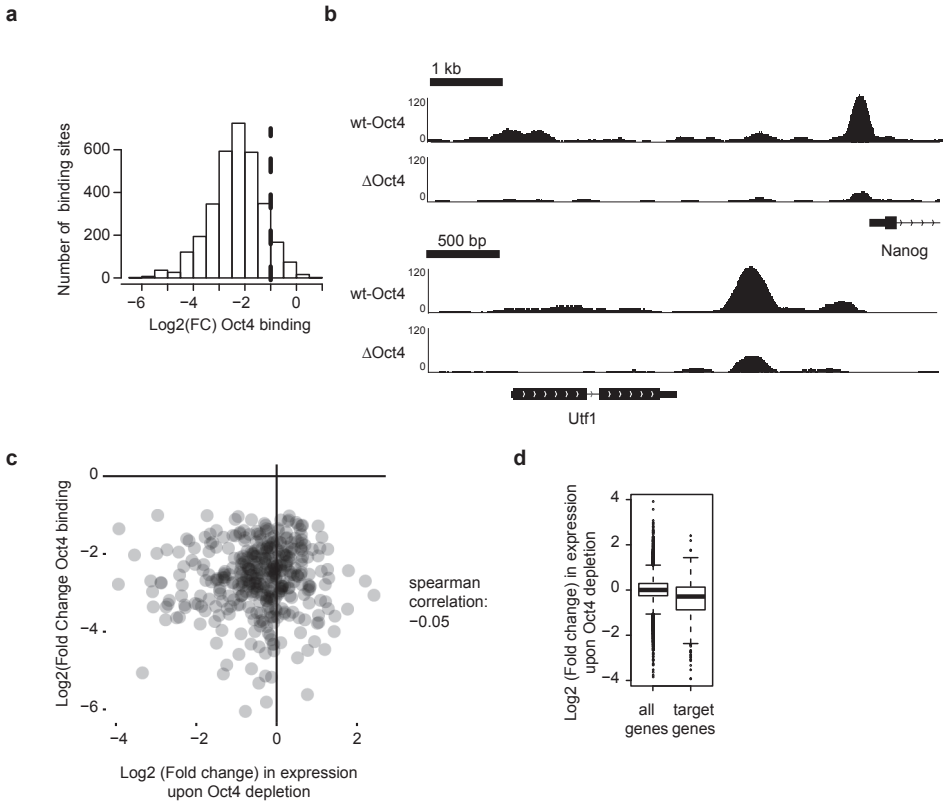


Figure1. (a) A histogram showing change in binding signal of Oct4 before and after depletion of Oct4. Fold change (FC) was calculated by dividing the observed binding signal in the Oct4 binding sites before depletion, by the observed binding signal after depletion. Left of the dotted line are the binding sites that are differentially bound by FC ≥ 0.5. **(b)** Oct4 binding profile of enhancers near Utf1 and Nanog genes before (wt-Oct4) and after (ΔOct4) 16 hours depletion of Oct4. **(c)** Scatterplot showing the correlation between change of Oct4 genome binding to its binding sites and expression perturbation of nearby genes upon Oct4 depletion. **(d)** Boxplot showing the perturbation of gene expression of all murine genes and Oct4 target genes upon Oct4 depletion in ESCs.

Oct4 genome binding and the fold-change reduction in expression of genes annotated to Oct4 binding sites (Figure 1c). However, genes that are annotated to Oct4 binding sites generally do show reduced expression upon Oct4 depletion (Figure 1d).

To assess the effect of Oct4 depletion on Sox2 and Nanog genome binding, the significant binding sites of Sox2 and Nanog in the wt-Oct4 and ΔOct4 conditions were determined. More than half of the significant Oct4 binding sites overlap with binding sites of either Sox2 or Nanog (Figure 2a). Upon depletion of Oct4, these sites show reduced average binding

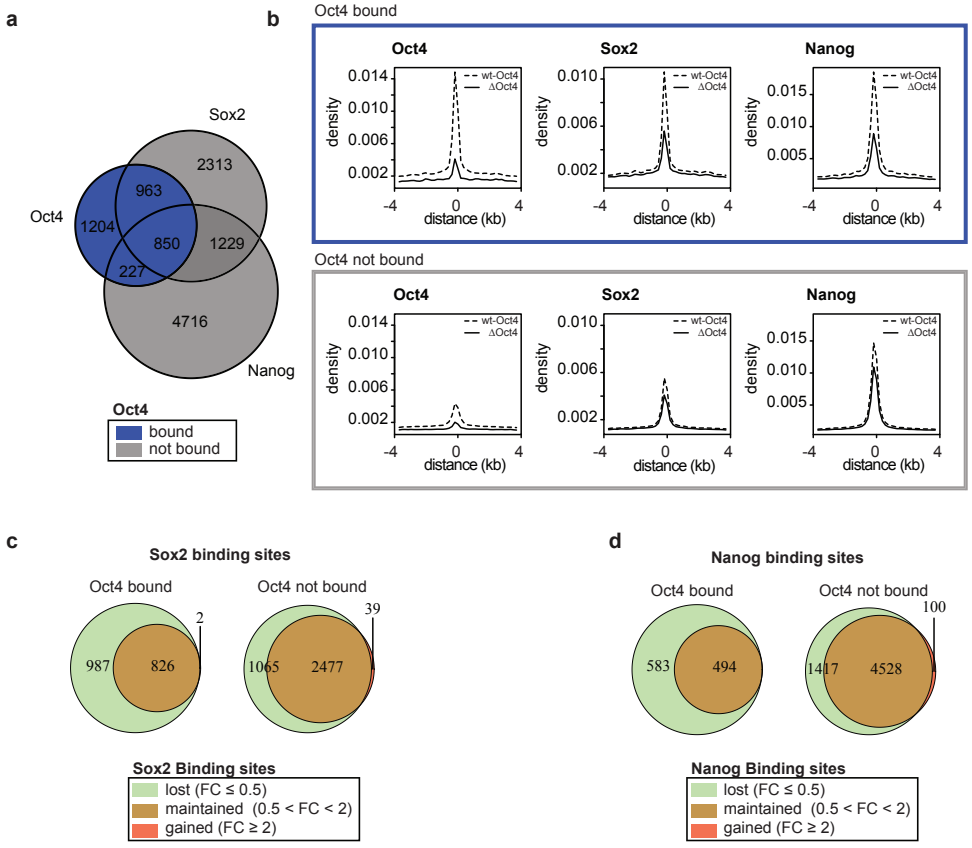


Figure 2. (a) A Venn diagram showing the overlap of significant binding sites between Oct4, Sox2 and Nanog. In blue are binding sites that are significantly for Oct4. In grey are significant binding sites of Sox2 and/or Nanog that are not significant Oct4 binding sites. **(b)** Density plots showing the average ChIP-seq binding density before and after depletion of Oct4 for Oct4, Sox2 and Nanog. Upper panel (blue) shows the average density of Oct4, Sox2 or Nanog for binding sites that have significant ChIP-seq signal for Oct4. Lower panel shows the average density of Oct4, Sox2 or Nanog for binding sites that do not have a significant ChIP-seq signal for Oct4. Interrupted line and uninterrupted line show the densities before and after depletion of Oct4, respectively. **(c)** Venn diagrams showing the overlap of Sox2 binding sites before and after depletion of Oct4. Left Venn diagram shows Sox2 binding sites that are also significant Oct4 binding sites. Right Venn diagram shows Sox2 binding sites that are not significant Oct4 binding sites. **(d)** Venn diagrams showing the overlap of Nanog binding sites before and after depletion of Oct4. Left Venn diagram shows Nanog binding sites that are also significant Oct4 binding sites. Right Venn diagram shows Nanog binding sites that are not significant Oct4 binding sites.

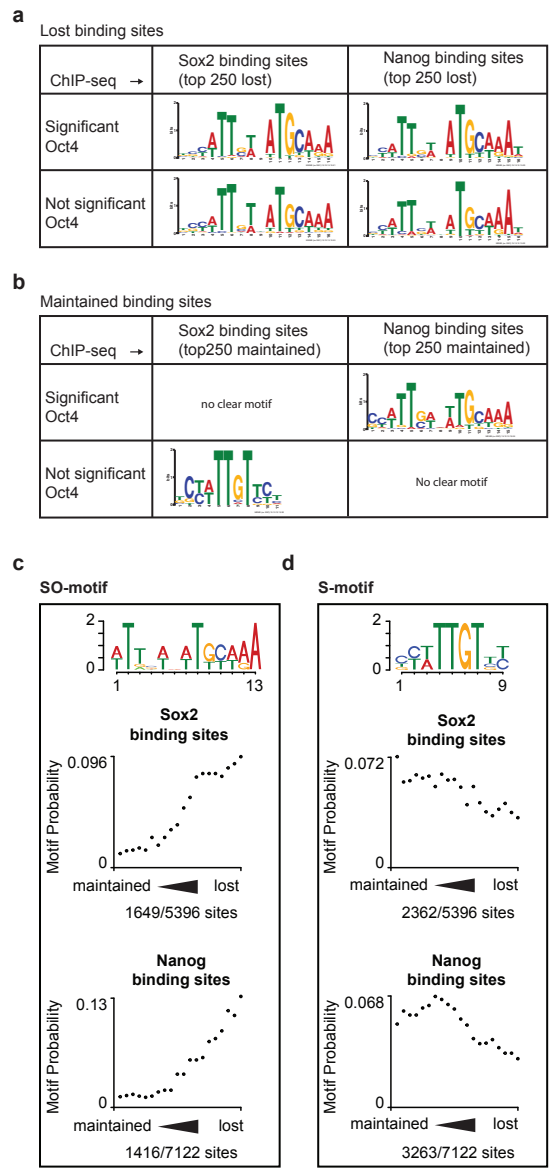


Figure 3. (a) DNA binding motifs found in binding sites that are lost upon Oct4 depletion. **(b)** DNA binding motifs found in binding sites that are maintained upon Oct4 depletion. **(c,d)** Probabilities of SO-motif **(c)** and S-motif **(d)** ranging from maintained to lost Sox2 and Nanog binding sites. Graphs were generated by sorting Sox2 or Nanog binding sites from 'maintained' to 'lost' for their respective change in ChIP-seq signal and divided into 20 bins. The binding sites in each bin that contained at least one of either non-overlapping motifs were counted and the probability was calculated based on all 20 bins. The total counted binding sites containing a motif for each category are indicated below the y-axis.

of Sox2 and Nanog (Figure 2b, upper panel). However, Sox2 and Nanog genomic sites that are not significant Oct4 binding sites show less reduction in the average genomic binding of Sox2 and Nanog (Figure 2b, lower panel). Upon depletion of Oct4 in ZHBTc4 ESCs, 2052 out of 5355 Sox2 binding sites and 2000 out of 7022 Nanog binding sites show 2-fold or more reduced binding (from here on 'lost' binding sites). When we specifically look at Sox2 binding sites that also have significant Oct4 binding, 987 out of 1813 binding sites are lost (Figure 2c). Also 1065 out of 3542 Sox2 binding sites without significant Oct4 binding are lost (Figure 2c). 583 out of 1077 Nanog binding sites with significant Oct4 are lost and 1417 out of 5945 Nanog binding sites without significant Oct4 binding are lost (Figure 2d). De-novo motif discovery was performed on Sox2 and Nanog binding sites with and without significantly bound Oct4 that were lost or maintained for Sox2 and Nanog binding (Figure 3a). Binding sites of Sox2 or Nanog that are lost upon Oct4 depletion are enriched for the composite SO-motif, irrespective of significant Oct4 binding. Therefore we assumed that these binding sites are bound by Oct4, but that we do not find significant Oct4 binding because of the sensitivity of our Oct4 ChIP-seq dataset. The reduced but still present Oct4 signal observed in Sox2 and Nanog sites that are not significantly bound by Oct4 seems to support this assumption (Figure 2b, lower panel). For maintained

binding sites of Sox2 the canonical Sox motif (S-motif) was discovered (Figure 3b). For maintained Nanog binding sites a less pronounced SO-motif was found (Figure 3b). The probability of finding the SO-motif is greater in binding sites that show reduced Sox2 or Nanog binding signal, than in binding sites where Sox2 and Nanog are maintained (Figure 3c). The inverse is true for the S-motif, which is increasingly found in binding sites that are maintained for Sox2 and Nanog upon depletion of Oct4 (Figure 3d).

A strong correlation between the change in binding of Sox2 and Nanog upon Oct4 depletion can be observed (Figure 4a,b). The Sox2 and Nanog binding sites that do not show reduced binding are often not signifcant Oct4 binding sites (Figure 4c). As examples, binding sites of Oct4, Sox2 and Nanog near *Utf1*, *Rest* and *Mib2* are shown (Figure 4d). As can be observed in Figure 4d, arrowheads 2, Sox2 and Nanog binding sites without significant co-localizing Oct4 are less or not affected by depletion of Oct4, whereas Sox2 and Nanog binding sites with significant co-localizing Oct4 show reduced binding of Sox2 and Nanog (Figure 2b and 4d, arrowheads 1). Furthermore, the enhancer marked in Figure 4d by arrowhead 1 near the *Utf1* gene shows reduced binding of Sox2 and Nanog and harbors a motif that resembles the SO-motif, while the enhancer marked by arrowhead 2 near *Mib2* shows maintained binding of Sox2 and Nanog and harbors a S-motif. In addition, the expression of genes near lost Oct4, Sox2 and/or Nanog binding sites that contain the SO-motif are marginally, but significantly, down-regulated upon Oct4 depletion, whereas genes near maintained Sox2 and/or Nanog binding that contain the S-motif are not (Figure4e). Gene ontology using DAVID[10] was performed on genes near maintained and lost binding sites of Sox2 (Table S1a) and Nanog (TableS1b). Genes near Nanog binding sites that are lost upon Oct4 depletion are enriched for GO terms that are related to transcriptional regulation and embryonic morphogenesis, whereas genes near Nanog binding sites that are maintained after depletion are also enriched for organ developmental and/or morphogenesis genes for the heart, neural tube and the respiratory system (TableS1ab). For Sox2 the most notable change in GO terms is the loss of the embryonic morphogenesis GO term (TableS1ab).

Sox2 dependent recruitment of Oct4

To assess to which extent Oct4 is dependent on Sox2 for its localization to the genome, we performed ChIP-seq against Oct4 before (from here on wt-Sox2) and after 24h depletion of Sox2 (from here on ΔSox2) in the doxycycline inducible Sox2-null 2ts-22c ESC line[9]. Western blotting shows that Sox2 is no longer detected after 24 hours of depletion, while Oct4 and Nanog levels seem unaffected (Figure S2). ChIP against Sox2 followed by quantitative PCR (ChIP-qPCR) also shows reduced binding of Sox2 to known binding sites in the ΔSox2 condition (Figure S3a). ChIP-qPCR of Nanog shows moderately reduced binding of Nanog to the -2kb enhancers of the Pou5f1 gene (figure S3b). We also attempted to determine genome-wide binding sites for Nanog under Sox2-depletion conditions but ChIP-seq against Nanog failed (results not shown).

Upon depletion of Sox2 in 2TS22C, 468 out of 1276 Oct4 (wt-Sox2) binding sites are lost and show at least 2-fold reduction in Oct4 binding signal (Figure 5a). De-novo motif discovery was performed on lost and maintained Oct4 (ΔSox2) binding sites. The SO-motif was discovered in both the lost and maintained Oct4 (ΔSox2) binding sites (Figure 5b). We also found that 1325 new Oct4 binding sites are gained upon Sox2 depletion

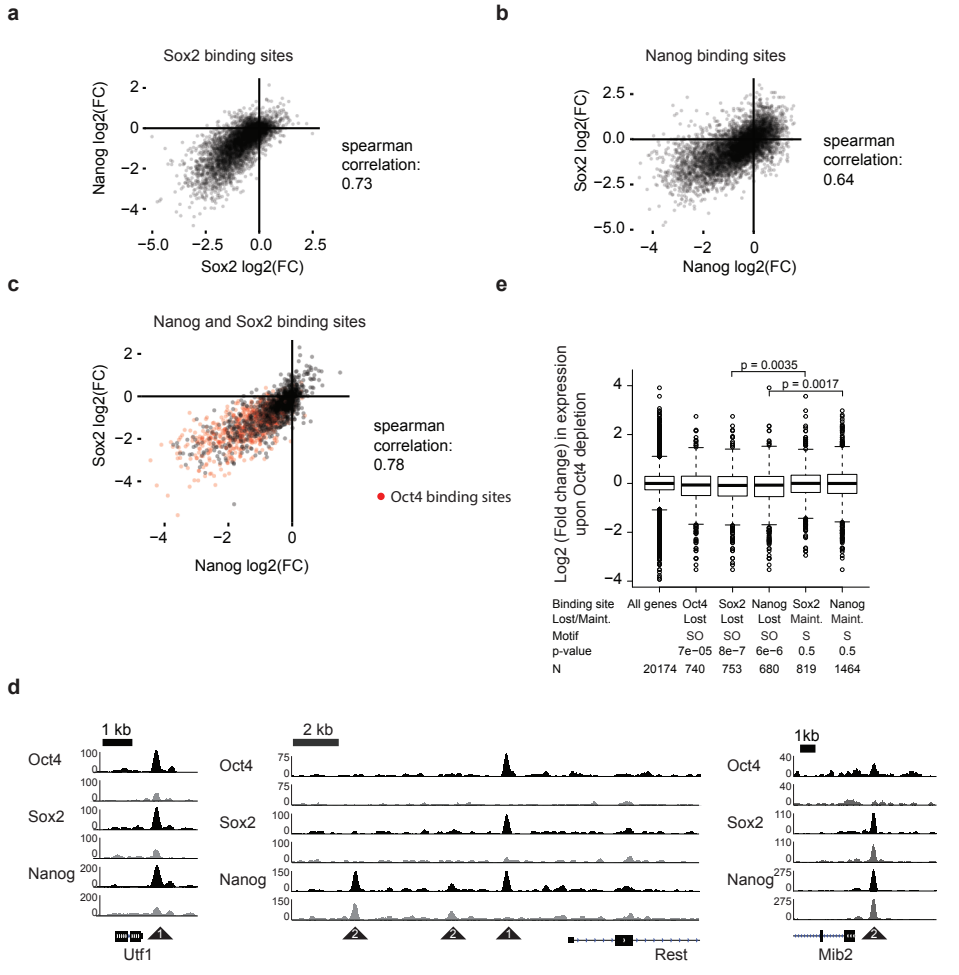


Figure 4. (a) Scatterplot showing the correlation between change in ChIP-seq signal upon Oct4 depletion for Sox2 and Nanog on Sox2 binding sites. Each dot represents a significant binding site for Sox2. The position along the x-axis of each dot represents the change in Sox2 binding signal. The position along the y-axis of each dot represent the change in Nanog binding signal. **(b)** Scatterplot showing the correlation between change in ChIP-seq signal upon Oct4 depletion for Sox2 and Nanog on Nanog binding sites. Each dot represents a significant binding site for Nanog. The position along the x-axis of each dot represents the change in Nanog binding signal. The position along the y-axis of each dot represents the change in Sox2 binding signal. **(c)** Scatter plot showing the correlation between change in ChIP-seq signal upon Oct4 depletion for Sox2 and Nanog on binding sites that are both Sox2 and Nanog binding sites. Each dot represents a significant Sox2-Nanog binding site. The position along the x-axis of each dot represents the change in Nanog binding signal. The position along the y-axis of each dot represents the change in Sox2 binding signal. Sox2-Nanog binding sites that are also significant Oct4 binding sites are marked in red. **(d)** Oct4, Sox2 and Nanog binding profiles of enhancers near Utf1, Rest and Mib2 genes before (black) and after (grey) depletion of Oct4. Arrowheads with '1' indicate binding sites where Sox2 and/or Nanog ChIP-seq signal is lost and arrowheads with '2' that are maintained. **(e)** Boxplots showing fold change expression upon Oct4 depletion for gene subsets. The gene subsets are defined by annotation of significant ChIP-seq binding sites for Oct4, Sox2 or Nanog (ChIP-seq) to their nearest genes, whether transcription factor binding is lost or maintained (Maint.) and whether the binding site contains a SO-motif or S-motif. P-values were calculated with the Mann-Whitney U test. P-values at the bottom of the graph were calculated by comparing the set with 'All genes'. 'All genes' are all genes tested in the Oct4 depletion microarrays.

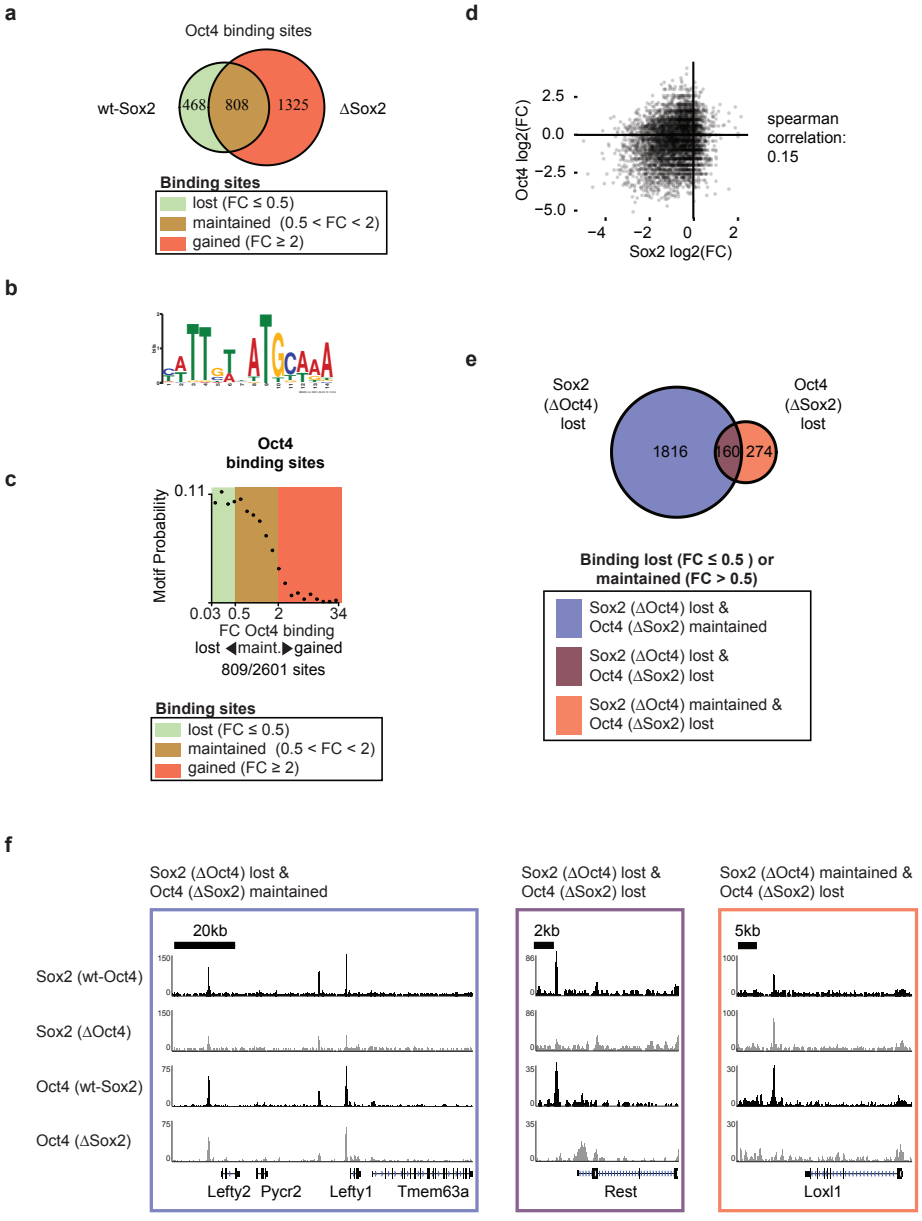


Figure 5. (a) A Venn diagram showing the overlap of Oct4 binding sites before and after depletion of Sox2. **(b)** The SO-motif that was discovered in lost and maintained Oct4 (ΔSox2) binding sites. **(c)** Probabilities of SO-motif ranging from lost to gained Oct4 (ΔSox2) binding sites. Graphs were generated by sorting Oct4 binding sites from most 'lost' to 'gained' and divided into 20 bins. The binding sites in each bin that contained at least one of either non-overlapping motifs were counted and the probability was calculated based on all 20 bins. The total counted binding sites containing the SO motif is indicated below the y-axis. **(d)** Scatterplot showing the correlation between change in ChIP-seq signal for Sox2 (ΔOct4) and Oct4 (ΔSox2) binding sites. Each dot represents a binding site that is significant for Sox2 and/or Oct4. The position on the x-axis represent the change in Sox2 binding signal upon Oct4 depletion. The position on the y-axis represents the change of Oct4 binding signal upon Sox2 depletion.

Legend continues on the bottom of the next page

(Figure 5a). No motifs related to the SO-motif were discovered in the gained Oct4 (ΔSox2) binding sites, this includes the standalone Oct4 or Sox2 motifs. However, we did find the Klf-family motif (GGGGG*GGGG). We also find that Oct4 (ΔSox2) gained peaks mainly localize to active promoters, whereas disappearing peaks mainly localize to enhancers (results not shown, see discussion). The probability of finding the SO-motif is the greatest in lost Oct4 (ΔSox2) binding sites and is decreasingly found in maintained Oct4 (ΔSox2) binding sites and virtually absent in gained Oct4 (ΔSox2) binding sites (Figure 5c).
We wanted to see whether correlation exists between the loss of binding of Sox2 under ΔOct4 conditions and loss of binding of Oct4 under ΔSox2 conditions. To investigate this, we unified the Sox2 and Oct4 binding sites of the ΔOct4 and ΔSox2 experiments, respectively. There appears to be no correlation between sites where there is reduced binding of Sox2 observed upon Oct4 depletion and sites where there is reduced binding of Oct4 upon Sox2 depletion (Figure 5d). Next we determined which binding sites showed reduced binding of Sox2 (ΔOct4) only, Oct4 (ΔSox2) only and which binding sites showed reduced binding for both Sox2 (ΔOct4) and Oct4 (ΔSox2). It appears that 160 sites show at least 2-fold reduced binding for both factors, whereas 1816 and 274 sites are affected in either Sox2 (ΔOct4) or Oct4 (ΔSox2), respectively (Figure 5e). As examples ChIP-seq binding profiles for all three categories are shown Figure 5f.

DISCUSSION

We used ChIP-seq to determine the genome-wide binding sites of Sox2 and Nanog before and after depletion of Oct4. We also determined the genome-wide binding sites of Oct4 before and after depletion of Sox2. Our data shows for the first time genome-wide that Oct4 and Sox2 depend on each other to bind binding sites that contain the SO-motif. The SO-motif plays a critical role in ESC pluripotency as it is present in the enhancers of many ESC identity transcription factors, such as Oct4, Sox2, Nanog, Utf1 and Rest[3,7,11,12] and is bound by many ESC transcription factors[3]. Except for the depleted proteins, we ensured that protein levels of the investigated transcription factors were at wild type levels. This is critical, as transcription of *Oct4*, *Sox2* and *Nanog* is regulated by Oct4 and Sox2[11-13] and small changes in Oct4 or Sox2 protein levels induce differentiation[8,14].
We observed a high correlation between the change in binding of Sox2 and change in binding of Nanog upon depletion of Oct4 (Figure 4a-c), suggesting that genome binding of Nanog and Sox2 are strongly interrelated. Oct4 and Nanog were not identified in each other's interactome[5,15] and Sox2 was found in the interactome of Oct4[15] and Nanog[5]. Together, this suggests that Oct4 directly recruits Sox2 to SO binding sites and that Nanog is recruited to these sites via Sox2.

Oct4 and Sox2 binding is not reciprocally reduced on all SO binding sites upon Sox2 and Oct4 depletion, respectively. Of the Oct4 (ΔSox2) binding sites that are lost we find 274

(Figure 5. Legend continues from previous page)
(e) A Venn diagram showing the overlap between 2 fold reduced Sox2 (ΔOct4) and 2-fold reduced Oct4 (ΔSox2) binding sites. (f) Binding profiles of Oct4 and Sox2 before and after depletion of Sox2 and Oct4, respectively. From left to right: 1) Enhancers near Lefty1 and Lefty2 where Sox2 (ΔOct4) binding is lost, but Oct4 (ΔSox2) is maintained. 2) Enhancer near Rest, where binding signal is lost for both Sox2 (ΔOct4) and Oct4 (ΔSox2). 3). Enhancer near Loxl1, where Oct4 (ΔSox2) binding is lost, but Sox2 (ΔOct4) is maintained. Color of the border corresponds to the color of the different subset in Figure 3e.

out of 434 do not overlap with lost Sox2 (ΔOct4) binding sites. In addition, many Sox2 and Nanog binding sites do not overlap with Oct4 on the genome and are maintained upon Oct4 depletion. The S-motif was discovered in the maintained Sox2 (ΔOct4) binding sites (Figure 3b) and we find that binding sites of Sox2 and Nanog that are maintained upon Oct4 depletion are increasingly enriched for the S-motif (Figure 3d). The S-motif has previously been found for the Sox2-K57E variant[16], which is a Sox2 mutant that shows reduced cooperativity with Oct4 in genome binding[17]. Enrichment of the S-motif in maintained Sox2 (ΔOct4) and Nanog (ΔOct4) binding sites may suggest that Oct4 independent binding of Sox2 and Nanog is facilitated by a high affinity motif for Sox2. Sox2 and Nanog were also suggested to co-bind a joint motif that also contains a Sox2-submotif[5,18], but this motif was not discovered in our maintained Sox2 (ΔOct4) or Nanog (ΔOct4) binding sites. Additionally, cooperation with other transcription factors than Oct4, Sox2 or Nanog may also rescue or facilitate binding to maintained binding sites. Oct4 binds by itself via homodimerization, as was described for an enhancer near the *OPN* gene[19], but we did not discover a separate Oct4 motif. Upon Sox2 depletion, we also found many new significant binding sites for Oct4 (Figure 5a). It is unclear if these Oct4 binding sites represent true Oct4 binding sites or whether they are an artifact of the lack of precipitated binding sites leading to over-sequencing of open regions. We do not find any motifs in the gained Oct4 (ΔSox2) binding sites that suggest direct binding of Oct4 to these sites.

We find that genes with annotated binding sites for Nanog that are maintained upon depletion of Oct4, are enriched for genes that are involved in later development such as respiratory system, neural tube and heart development (TableS1b). It is likely that Nanog contributes to the inactive state of these genes by attracting repressive complexes. The Nanog interactome contains subunits of several complexes that catalyze deposition of repressive histone modifications[5]. Indeed Nanog was shown to recruit interaction partners to repress Gata4 and Gata6 expression to prevent ESC differentiation[20]. In addition, Nanog inhibits the pro-differentiation activities of NFκB by binding to the NFκB protein[21] and mutant ESCs that do not express Nanog can be maintained, but are prone to differentiation[22].

Thus far Oct4 and Sox2 were demonstrated to cooperatively bind to an enhancer site near *UTF1*[4,7] and mutant *FGF4* enhancers[4]. We demonstrate that genome-wide co-localization of Oct4 and Sox2 to the ESC genome is interdependent on binding sites that contain the SO-motif.

MATERIALS AND METHODS

Cell culture

Mouse embryonic stem cell lines ZHBTc4 and 2TS22C were grown without feeders on gelatin coated dishes. ZHBTc4 cells were grown in Glasgow Minimum Essential Medium (GMEM) supplemented with leukemia inhibitory factor (LIF), 15% fetal bovine serum, 0.25% sodium bicarbonate, Gibco MEM non-essential amino acids, 1mM sodium pyruvate, penicillin/streptomycin, 1 mM glutamine, 50μM and β-mercaptoethanol, as previously described[23]. 2TS22C were grown in GMEM supplemented with LIF, 10% fetal bovine serum, Gibco MEM non-essential amino acids, 1mM sodium pyruvate and

β-mercaptoethanol.

Western Analysis of Sox2/Oct4 depletions

Whole cell extracts of the ZHBTc4 and 2TS22C cells under different conditions were analyzed by western blot using 1:1000 Sox2 (Santa Cruz sc-5279),1:2000 Nanog (Cosmo Bio REC-RCAB0002P-F), 1:1000 Oct4 (Santa Cruz sc-5279) , 1:1000 laminin (Santa Cruz 1:1 mixture of sc-6216 and sc-6217) or 1:2000 VCP (Santa Cruz sc-20799) antibodies.

Chromatin immunoprecipitations

ChIPs for Oct4 depletion conditions were performed on 90 million cells per condition in the ZHBTc4 cell line. To deplete Oct4, doxycycline was added 16 hours before harvesting. Control cells and Oct4 depleted cells were harvested simultaneously. ChIPs for Sox2 depletion conditions were performed on 80-90 million cells per condition in the 2TS22C cell line. To deplete Sox2, doxycycline was added 24 hours prior to harvesting. Control cells and Sox2 depleted cells were harvested simultaneously. ChIP was essentially performed as previously described[24], with slight modifications. Briefly, cells were fixed for 10 minutes in formaldehyde solution (50 mM HEPES-KOH pH7.6, 100 mM NaCl, 1mM EDTA pH8.0, 0.5 mM EGTA pH8.0, 11% formaldehyde). Glycine was added to a final concentration of 125 mM and incubation was continues for 10 minutes at room temperature. Cells were washed twice in ice cold PBS containing Roche Complete EDTA-free proteinase inhibitor cocktail (CEF). Cells were harvested and washed thrice in lysis buffer (10 mM Tris pH7.5, 10 mM NaCl, 0.5% NP40, CEF). The pellet was resuspended in 300 µl SDS lysis buffer (10 mM Tris pH8.0, 150 mM NaCl, 1 mM EDTA pH8.0, 1% SDS, CEF) per 80 µl of pellet. Each 800 µl of sample was sonicated on the MSE Soniprep 150, 15 cycles 15 s on, 45 s off, amplitude 7. After sonication, DNA fragment size was confirmed to be between 200 and 500 bp. The sonicated chromatin was 10 times diluted with ChIP dilution buffer (20 mM Tris pH8.0, 150 mM NaCl, 1 mM EDTA, 1% Trition X-100, CEF). Magnetic Protein G Dynabeads were 1 hour blocked in ChIP dilution buffer containing 0.5 mg/ml BSA and 0.2 mg/ml sonicated salmon sperm DNA, washed thrice in ChIP dilution buffer and kept in ChIP dilution buffer. To pre-clear the chromatin, each condition was incubated for 30 minutes with 75 ul of blocked Protein G dynabead solution. Antibodies used in different conditions were GFP (Santa Cruz sc-8334), Sox2 (Santa Cruz sc-5279), Nanog (Cosmo Bio REC-RCAB0002P-F) or Oct4 (Santa Cruz sc-5279). 10 – 15 µg of antibody was added to the chromatin per condition and incubated over night at 4 ^{o}C. The following morning 10 µl of blocked beads solution was added for each 1 µg of antibody and incubated for 1 hour. Supernatant was discarded and beads were washed with 3 times low salt buffer (20 mM Tris-HCl pH 8.0, 150 mM NaCl, 2 mM EDTA, 1% Triton X-100, 0.1% SDS), 1 time high salt buffer (20 mM Tris-HCl pH 8.0, 500 mM NaCl, 2 mM EDTA, 0.1% SDS, 1% Triton X-100), 1 time LiCl buffer (10 mM Tris-HCl, 250 mM LiCl, 1 mM EDTA, 0.5% Sodium deoxycholate, 0.5% NP40) and 1 time with TE buffer containing 50 mM NaCl. Immunoprecipitated chromatin was eluted by resuspension of the beads in 250 µl elution buffer and 15' incubation at 65 ^{o}C. Beads were discarded and the samples were incubated for 4-8 hours at 65 ^{o}C. Samples were diluted with 1 volume of TE buffer and treated with 0.2 mg/ml of proteinase K for 2 hours at 45 ^{o}C. DNA was recovered by two phenol-chloroform extractions. For each condition 10 ng of DNA was used for library generation, followed by next-generation sequencing on an Ilumina

Genome analyzer, as previously[25]. For quantitative PCR analysis the following primers were used: Amylase (CTCCTTGTACGGGTTGGT and AATGATGTGCACAGCTGAA.), Nanog -5 kb (GTCCCCGCTCCTTTTCAGCACTAACCATAC and CGGTTTGAATAGGGAGGAGGGCGTCT), Pou5f1 -2 kb (TTGAACTGTGGTGGAGAGTGCT and TGCACCTTTGTTATGCATCTGCCG), Rest -3 kb (CTCCCCTGGACAATAGCTTC and CGTCCTTCATTTCCTCAGTG).

Data analysis

Mapping of sequence data was performed with minor modifications as previously described[25],. Briefly, sequence reads with low complexity that are unlikely to map uniquely to the genome were removed from the dataset using prinseq-lite with the dust method with 7 as threshold[26]. Bases on 5' and 3' end of the reads with a quality score below 28 were trimmed using prinseq-lite. Trimmed reads were required to have a minimum length of 20 bases. The remaining sequences with a Phred score <70 were mapped to the mm9 (Ensembl NCBIM37.67) reference genome using Bowtie v0.12.7, where a seed length of 36 was used and in which a maximum of 2 mismatches were allowed[27]. If a read had multiple alignments, only the best matching read was reported. Reads mapping to regions not assembled into chromosomes and duplicated reads were removed. SISSRs v1.4 with option –t was used to identify binding sites for all ChIP-seq datasets with the GFP ChIP-seq as background control data set[28]. For the ChIP-seqs from ZHBTc4 and 2TS22c cell line, SISSRs was run with the option –w 10 and –w 50, respectively. In addition Oct4, Sox2 and Nanog ChIP-seqs in ZHBTc4 were analyzed with the option –p on 0.1, 0.2 and 0.1, respectively and the 2TS22c Oct4 ChIP-seqs with option –p 0.005. If centers of two binding regions reported by SISSRs were 25 bp or less apart, they were considered 1 binding region. When calculating the overlap of binding regions between two different ChIP-seq datasets, binding regions were considered to overlap if their centers were 25 bp or less apart. Except the Oct4 (wt-Oct4) and Oct4 (ΔOct4) ChIP-seq dataset, all ΔOct4 and ΔSox2 tracks were normalized to their respective wt-Oct4 and wt-Sox2 controls by correcting for sequence depth. Fold-changes between binding sites in the depletion and wt conditions were based on normalized reads counted in a 201 bp region centered on the summit of a binding site. Scatterplots were created in R using the ggplot2 library. For calculating the overlaps of binding sites between wt-Oct4 and ΔOct4 conditions, a unified list of binding sites was created. If the fold-change in normalized reads between both conditions was smaller or equal to 0.5-fold, a binding region was considered lost. If it was between 2-fold and 0.5-fold, it was considered maintained and if it was bigger or equal to 2-fold it was considered gained upon depletion of Oct4. Venn diagrams were created in R using the VennDiagram library. MEME using the method 'zoops' was used for motif calling[29]. Motif densities were calculated by sorting regions from low to high fold change upon Oct4 or Sox2 depletion. These sorted regions were divided over 20 bins. The SO-motifs of Figure 3a were generalized to [AT]T[TGA][GCTA][TA][GCTA][AT]T[GT][CT][TA][AG]A and the S-motif of Figure 3b was generalized to [AG][AGC]ACAA[AT][GA][ACG] to calculate the motif densities. The number of occurrences of the investigated motif in this bin was counted and divided by the sum of all counts in all bins. If a motif was matched more than once in 1 binding region, then it was only counted once. SO-motif and S-motif were not allowed to overlap. The published and already normalized expression data for the ZHBTc4 cell line was used as such[8]. To determine the target genes of Oct4 (for Figure 1d) and those of maintained and lost Sox2 and Nanog

binding sites (for TableS1), the ClosestGene method described by Sikora-Wohlfeld et al. 2014 was used[30]. For Gene Ontology, DAVID 6.7[31] was used with GOTERM_BP_FAT, GOTERM_CC_FA and GOTERM_MF_FAT. The sequencing profiles were created in the IGV browser [32] and edited for appearance in Adobe Illustrator.

REFERENCES

1. Boyer, L.A. et al. Core transcriptional regulatory circuitry in human embryonic stem cells. Cell 122, 947-56 (2005).
2. Ivanova, N. et al. Dissecting self-renewal in stem cells with RNA interference. Nature 442, 533-8 (2006).
3. Chen, X. et al. Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. Cell 133, 1106-17 (2008).
4. Ambrosetti, D.C., Basilico, C. & Dailey, L. 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, 6321-9 (1997).
5. Gagliardi, A. et al. A direct physical interaction between Nanog and Sox2 regulates embryonic stem cell self-renewal. EMBO J 32, 2231-47 (2013).
6. Loh, Y.H. et al. The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. Nat Genet 38, 431-40 (2006).
7. Remenyi, A. et al. Crystal structure of a POU/HMG/DNA ternary complex suggests differential assembly of Oct4 and Sox2 on two enhancers. Genes Dev 17, 2048-59 (2003).
8. Niwa, H., Miyazaki, J. & Smith, A.G. Quantitative expression of Oct-3/4 defines differentiation, dedifferentiation or self-renewal of ES cells. Nat Genet 24, 372-6 (2000).
9. Masui, S. et al. Pluripotency governed by Sox2 via regulation of Oct3/4 expression in mouse embryonic stem cells. Nat Cell Biol 9, 625-35 (2007).
10. Dennis, G., Jr. et al. DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biology 4, P3 (2003).
11. Kuroda, T. et al. Octamer and Sox elements are required for transcriptional cis regulation of Nanog gene expression. Mol Cell Biol 25, 2475-85 (2005).
12. Chew, J.L. et al. Reciprocal transcriptional regulation of Pou5f1 and Sox2 via the Oct4/Sox2 complex in embryonic stem cells. Mol Cell Biol 25, 6031-46 (2005).
13. Rodda, D.J. et al. Transcriptional regulation of nanog by OCT4 and SOX2. J Biol Chem 280, 24731-7 (2005).
14. Kopp, J.L., Ormsbee, B.D., Desler, M. & Rizzino, A. Small increases in the level of Sox2 trigger the differentiation of mouse embryonic stem cells. Stem Cells 26, 903-11 (2008).
15. van den Berg, D.L. et al. An Oct4-centered protein interaction network in embryonic stem cells. Cell Stem Cell 6, 369-81 (2010).
16. Merino, F. et al. Structural basis for the SOX-dependent genomic redistribution of OCT4 in stem cell differentiation. Structure 22, 1274-86 (2014).
17. Ng, C.K. et al. Deciphering the Sox-Oct partner code by quantitative cooperativity measurements. Nucleic Acids Res 40, 4933-41 (2012).
18. Hutchins, A.P. et al. Co-motif discovery identifies an Esrrb-Sox2-DNA ternary complex as a mediator of transcriptional differences between mouse embryonic and epiblast stem cells. Stem Cells 31, 269-81 (2013).
19. Botquin, V. et al. New POU dimer configuration mediates antagonistic control of an osteopontin preimplantation enhancer by Oct-4 and Sox-2. Genes Dev 12, 2073-90 (1998).
20. Wang, J. et al. A protein interaction network for pluripotency of embryonic stem cells. Nature 444, 364-8 (2006).
21. Torres, J. & Watt, F.M. Nanog maintains pluripotency of mouse embryonic stem cells by inhibiting NFkappaB and cooperating with Stat3. Nat Cell Biol 10, 194-201 (2008).
22. Chambers, I. et al. Nanog safeguards pluripotency and mediates germline development. Nature 450, 1230-4 (2007).
23. van den Berg, D.L. et al. Estrogen-related receptor beta interacts with Oct4 to positively regulate

Nanog gene expression. Mol Cell Biol 28, 5986-95 (2008).
24. Bernstein, B.E. et al. Genomic maps and comparative analysis of histone modifications in human and mouse. Cell 120, 169-81 (2005).
25. Engelen, E. et al. Proteins that bind regulatory regions identified by histone modification chromatin immunoprecipitations and mass spectrometry. Nat Commun 6, 7155 (2015).
26. Schmieder, R. & Edwards, R. Quality control and preprocessing of metagenomic datasets. Bioinformatics 27, 863-4 (2011).
27. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology 10, R25 (2009).
28. Jothi, R., Cuddapah, S., Barski, A., Cui, K. & Zhao, K. Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data. Nucleic Acids Res 36, 5221-31 (2008).
29. Bailey, T.L. & Elkan, C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc Int Conf Intell Syst Mol Biol 2, 28-36 (1994).
30. Sikora-Wohlfeld, W., Ackermann, M., Christodoulou, E.G., Singaravelu, K. & Beyer, A. Assessing computational methods for transcription factor target gene identification based on ChIP-seq data. PLoS Comput Biol 9, e1003342 (2013).
31. Huang da, W., Sherman, B.T. & Lempicki, R.A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 4, 44-57 (2009).
32. Robinson, J.T. et al. Integrative genomics viewer. Nat Biotechnol 29, 24-6 (2011).

SUPPLEMENTARY INFORMATION

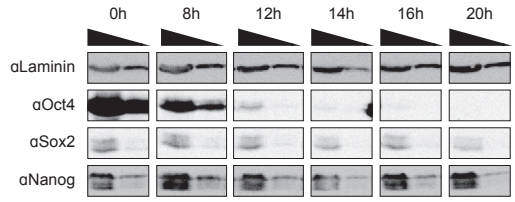


Figure S1. Whole cell extract of ZHBTc4 after 0, 8, 12, 14, 16 or 20 hours of doxycycline induced depletion of Oct4 analyzed by western blot with the indicated antibodies.

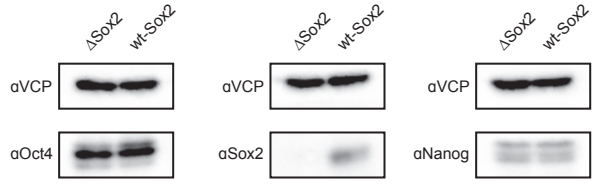


Figure S2. Whole cell extract of 2TS22c after 24 hours of doxycycline induced depletion of Sox2 analyzed by western blot with the indicated antibodies.

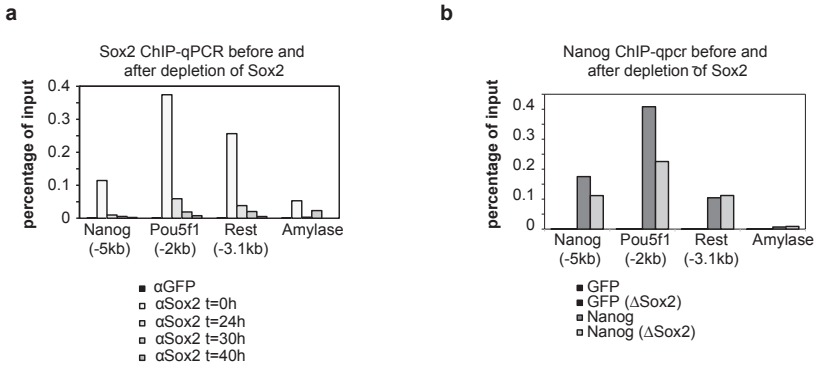


Figure S3. (a) ChIP-qPCR showing binding of Sox2 for different time points after induction of Sox2 depletion. Tested genomic sites are upstream of Nanog, Pou5f1 and Rest. Amylase is a negative control. The distance of the binding to the TSS is indicated between brackets. Time is indicated in hours. The ChIP against GFP was used as mock control and was performed at t=0. **(b)** ChIP-qPCR showing binding of Nanog to the genome upstream of Nanog, Pou5f1 and Rest. Amylase is a negative control. The distance of the binding to the TSS is indicated between brackets. A ChIP against GFP was used as mock control.

a

Lost Sox2 binding sites upon Oct4 KD

Term	Count	FE	adj. p	FDR
regulation of RNA metabolic process	60	2.0	2.5E-04	2.6E-04
regulation of transcription, DNA-dependent	59	2.0	1.7E-04	3.6E-04
regulation of transcription	77	1.7	4.2E-04	1.3E-03
transcription	64	1.8	1.0E-03	4.1E-03
embryonic morphogenesis	23	3.2	1.0E-03	5.2E-03

Lost Nanog binding sites upon Oct4 KD

Term	Count	FE	adj. p	FDR
chordate embryonic development	22	2.8	7.7E-02	8.6E-02
embryonic morphogenesis	20	2.9	4.1E-02	9.0E-02
embryonic development ending in birth or egg hatching	22	2.7	3.0E-02	9.8E-02
transcription	57	1.7	2.4E-02	1.0E-01
regulation of transcription	67	1.6	2.6E-02	1.4E-01

b

Maintained Sox2 binding sites upon Oct4 KD

Term	Count	FE	adj. p	FDR
protein-DNA complex	9	10.0	5.9E-04	3.6E-03
regulation of transcription, DNA-dependent	41	2.2	3.6E-03	4.8E-03
regulation of RNA metabolic process	41	2.1	2.7E-03	7.0E-03
regulation of transcription	52	1.8	6.4E-03	2.5E-02
chromosome organization	18	3.4	5.6E-03	3.0E-02

Maintained Nanog binding sites upon Oct4 KD

Term	Count	FE	adj. p	FDR
tube development	20	3.9	1.3E-03	1.4E-03
heart development	18	4.2	1.2E-03	2.4E-03
embryonic morphogenesis	23	3.3	9.0E-04	2.8E-03
tube morphogenesis	14	4.2	1.1E-02	4.4E-02
respiratory system development	12	5.0	8.5E-03	4.4E-02

Table S1. (a) Gene ontology analysis of putative Sox2 or Nanog targets before depletion of Oct4. **(b)** Gene ontology analysis of putative Sox2 or Nanog targets after depletion of Oct4.

Chapter5

General Discussion

The diversity of cell types in a single multi-cellular organism is enormous, especially when we consider that each cell contains an identical genome. Transcriptional regulation by different cell type-specific sets of transcription factors ensures that different cell types transcribe different sets of genes. Transcription factors regulate gene expression by binding to different functional elements of the genome such as promoters and enhancers. In **Chapter 2** of this thesis we developed a technique that we named ChIP-MS to identify transcription factors and other proteins that bind to promoters, enhancers and heterochromatin in ESCs and identify Dppa2 as specifically binding low activity promoters and that Dppa2 acts outside the classical pluripotency network. Besides that transcription factors bind to the genome, they also interact with other transcription factors and proteins to regulate transcription and form so-called transcriptional networks. In **Chapter 3** we assemble a transcriptional network of over 200 proteins including 69 proteins associated with mental disorders by identifying the interactors of four transcription factors in NSCs. Interacting transcription factors can cooperatively bind to the genome, where interaction influences the genomic target specificity. In **Chapter 4** preliminary data is presented that shows that genome wide localization of Oct4, Sox2 and Nanog is influenced by the presence of Oct4 or Sox2.

The role of Dppa2 as a regulator of transcription.

In the ChIP-MS experiments of **Chapter 2,** Dppa2 was found to have a high emPAI score for H3K4me3, but was relatively low on H3K27ac ChIP-MS in ESCs. This suggested that Dppa2 binds to promoters of genes with low transcriptional activity in ESCs. This observation was confirmed by ChIP-seq of Dppa2, where Dppa2 binding sites identified were intersected with published expression data on Dppa2 KO mice derived ESCs (Dppa2KO-ESCs) and expression data of wild type ESCs. Dppa2 bound genes are mainly downregulated upon Dppa2 KO (**Chapter 2 figure 5ab**), hence Dppa2 is most likely a transcriptional activator. However, Dppa2 binds promoters of genes that are lowly expressed in ESCs **(Chapter 2 figure 4e, 5cd**). What is the role of Dppa2 in the marginal activation of these genes? Increased CpG dinucleotide methylation of the of Dppa2 bound *Nkx2-5* and *Syce1* promoters in Dppa2KO-ESCs suggests that binding of Dppa2 protects these promoters against DNA methylation and dimethylation of H3K9[1]. Family member Dppa3 has been described to bind H3K9me2 in early embryogenesis and to protect DNA against demethylation[2]. As Dppa2 binding of promoters seems to protect against methylation of H3K9[1], it is unlikely that Dppa2 binds to H3K9me2. Even though Dppa2 is co-immunoprecipitated with H3K4me3 marked chromatin (**Chapter2**), it unlikely that Dppa2 directly binds to this histone mark, because Dppa2 mainly binds to promoters of low activity and H3K4me3 is also present on highly active promoters. With the same reasoning direct binding to H3K27ac can also be excluded. A histone modification that is negatively correlated with the activity of H3K4me3 marked promoters is H3K27me3 and marks poised promoters[3] and poised enhancers[4]. To confirm that Dppa2 does not bind H3K9me2 and whether it possibly binds to H3K27me3, enrichments for these and other histone modifications on ESCs Dppa2 bound promoters were calculated using published datasets[5-8] (Figure 1). It indeed appears that Dppa2-bound promoters in ESCs are enriched for H3K27me3 and depleted for other heterochromatin marks such as H3K9me2 and H3K9me3. To confirm that Dppa2 indeed binds H3K27me3 additional experiments have to be performed, such as for example a ChIP-MS against H3K27me3.

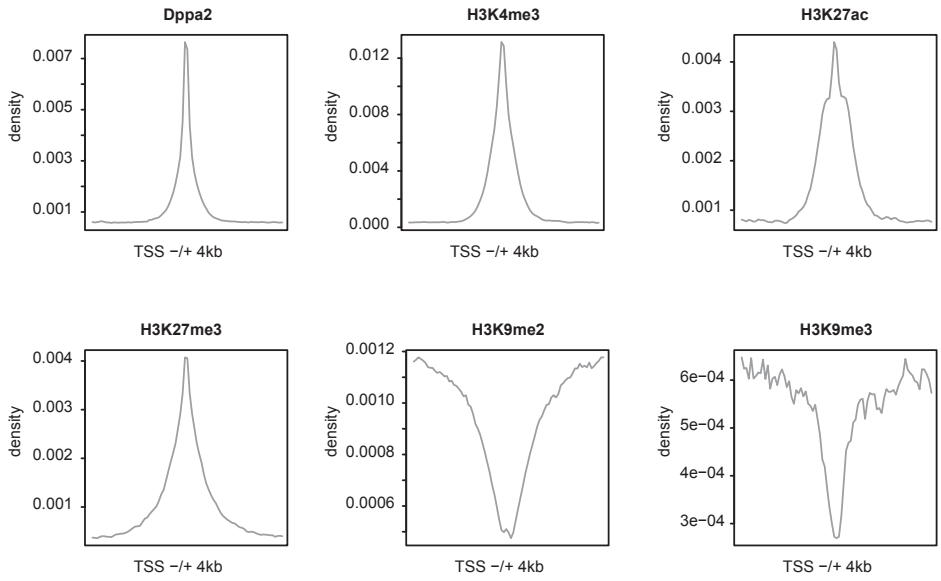


Figure 1. Density plots showing the average ChIP-seq density from indicated ChIP-seq datasets on significant Dppa2-bound promoters. Dppa2-bound promoters are enriched for H3K4me3, H3K27ac and H3K27me3, while being depleted for H3K9me3 and H3K9me2.

The ChIP-MS data also suggests that Dppa4, like Dppa2, localizes to low activity promoters (**Chapter2, Supplementary Table1**). Co-immunoprecipitations in mouse and human showed that Dppa2/DPPA2 interacts with Dppa4/DPPA4[1,9]. The ChIP-MS predicted genome localization of Dppa4 to active chromatin is also in agreement with published data[10] . However, Dppa2 is a transcriptional activator (**chapter 2**), whereas DPPA4 acts as a transcriptional repressor[9]. An antagonistic function of Dppa2 and Dppa4 is also suggested by the partial rescue of Dppa2 KO mice by KO of Dppa4, as the Dppa2/Dppa4 double KO resulted in phenotype less severe than that of the Dppa2 KO[1].

Dppa2 and Dppa4 interact[1,9] and are both expressed in ESCs. Hence it would be interesting to determine which of them is present in excess in ESCs and how protein levels of either Dppa2 or Dppa4 would affect gene activation and repression. Also, if Dppa2 and Dppa4 co-bind a promoter, will it be repressed or activated? If we consider the methylation of Dppa2-bound promoters upon Dppa2 KO[1] and downregulation of expression, we may speculate that Dppa4 in absence of Dppa2 may be involved in attracting DNA-methylating proteins and that this is prevented in presence of Dppa2.

To further elucidate the role of Dppa2 in gene regulation we purified Dppa2 in ESCs to identify potential interactors of Dppa2, using the FLAG-tag affinity protocol, that was also used in **Chapter 3**. However, using this purification protocol for Dppa2 was inefficient, because Dppa2 sticks tightly to the insoluble chromatin fraction (results not shown). The existing protocol was modified. NaCl was added to buffer A up to a concentration of 500mM and prior to the dounce homogenization, the scraped cells were shortly sonicated. Dppa2 was purified using the modified protocol and interactors were identified

using mass spectrometry. The most prominent interactor identified is Dppa4 and was described to interact with Dppa2[1,9].

Dppa3 (also known as Stella or PGC7) protects the maternal genome against demethylation[11,12] and this may hint to a role of Dppa2 and Dppa4 in maternal and paternal expression. In addition, in the gonads of male and female mice Oct4, Dppa4 and Dppa2 undergo downregulation starting from E12.5[13]. While in females downregulation of Dppa4 preceded downregulation of Dppa2 (and Oct4) by several hours, Dppa4 is downregulated much slower than Dppa2 (and Oct4) in males[13]. In contrast to female gonads, Dppa4 expression appears to be maintained in male gonads[13]. These observations and apparent antagonistic roles in transcriptional activation may suggest that Dppa2 and Dppa4 are involved in imprinting.

The role of Dppa2 in pluripotent embryonic stem cells and mouse development

Dppa2 KO mice showed impaired alveolar formation in late lung development and these mice died around birth with respiratory defects[1]. However, Dppa2 expression in mouse was detected in ESCs and ESC-related cells such as GV-stage oocytes, blastocysts, E12.5 primordial germ cells and embryonal carcinoma cells[13], but not in mouse lung[1]. Therefore it was suggested that Dppa2 plays a role in development by leaving epigenetic marks in earlier developmental stages[1]. On first sight this seems to be supported by finding H3K27me3 enriched on Dppa2 promoters and by the data from **Chapter 2 figures 4 and 5**, but although *Nkx2-5* and *Syce1* are misregulated in Dppa2KO lung[1], Dppa2 ESC bound genes as determined in **Chapter 2** were not found enriched amongst misregulated genes in Dppa2 KO lung and *Nkx2-5* and *Syce1* were not differentially methylated in Dppa2-KO lung[1]. Hypothetically Dppa2 might just leave epigenetic marks to a few master regulator genes of lung development and this might explain the lack of enrichment of ESC Dppa2 target genes in Dppa2 KO lung. However, it is unlikely, that a protein that binds over 3000 promoters in ESCs and that supposedly leaves epigenetic marks in ESCs for later development , would only affect a couple of lung master regulatory transcription factors that affect late lung development. Dppa2 KO mice show increased lethality around birth and are found at Mendelian ratios until at least E16.5[1], arguing against major developmental defects in other organs. Another possibility is that Dppa2 is more widely expressed than what is believed. Dppa2 was found to be expressed in porcine lung and was speculated to be expressed in adult progenitor cells of tissues[14], but no additional evidence was provided in support to this speculation.

Are different mental disorders caused by disruption of a shared transcriptional network?

In **Chapter 3** transcription factors were purified from neural stem cells to identify their interaction partners by mass spectrometry and to assemble a protein interaction network. It was explored whether this network can provide gene regulatory explanations for an overlap between intellectual disability, autism spectrum disorders and schizophrenia. The logic followed in this chapter and other work[15] is as follows: 1) If a transcription factor is mutated and causes a (mental) disorder, the genes that this transcription factor regulates underlie the symptoms of this disorder. 2) If multiple (mental) disorders share symptoms and are caused by multiple transcription factors respectively, the shared symptoms are a product of these transcription factors regulating the same gene(s).

The data from **Chapter 3** suggest co-regulation of target genes by Tcf4 and Ascl1, as they interact and co-localize to the genome (**Chapter 3, figure 1 & 2a**). Ascl1 and Tcf4 both activate genes that are highly enriched for cell division regulating genes (**Chapter 3, figure 3a**, [16]), also suggesting that Ascl1 and Tcf4 co-regulate genes. Tcf4 and Ascl1 bind the only transcriptional enhancer near *Phox2b* in NSCs (**Chapter 3 figure 2d**). *PHOX2B* is a gene that is mutated in 60% of patients with Congenital Central Hypoventilation Syndrome (CCHS), which is characterized by breathing abnormalities[17]. Mutations in *ASCL1* are also observed in patients with CCHS[18]. Pitt Hopkins also features breathing abnormalities and is caused by mutations that cause haploinsufficiency of TCF4[19,20], where heterozygous missense and null mutations impair interaction of TCF4 with ASCL1[20]. Heterozygous knock out mice of *Phox2b* indeed showed respiratory dysfunction that partly modeled the breathing phenotype observed in CCHS[21]. Where both CCHS and Pitt Hopkins are featured by periods of reduced breathing, hypoventilation and apnoea respectively, patients with Pitt Hopkins also suffer from hyperventilation[22]. It remains to be proven that the Tcf4-Mash1 bound enhancer indeed regulates *Phox2b* and that misregulation of *PHOX2B* can underlie apnoea in Pitt Hopkins syndrome.

In **Chapter 3** we identified *Nrxn1* as a Tcf4 target gene, as it showed reduced expression upon Tcf4 KD and enhancers in the vicinity of *Nrxn1* were bound by Tcf4. Mutations in *NRXN1* and *CNTNAP2* have been found in patients with Pitt Hopkins-like syndrome and it was speculated that *NRXN1* and *CNTNAP2* are under transcriptional control of TCF4[23]. Regulation of *NRXN1* by TCF4 would explain the similar phenotypes observed between patients with mutations in *NRXN1* or *TCF4*. Interestingly, hyperventilation was also observed in all patients with *NRXN1* mutations[23], suggesting that the hyperventilation in Pitt Hopkins (like) syndrome is downstream of NRXN1 and not directly downstream of PHOX2B (see above). However, as PHOX2B is also a transcription factor, it might also regulate transcription of *NRXN1*, but this remains subject to further investigation.

Gene annotation of transcription factor binding sites

In **chapters 2, 3 and 4** of this thesis binding sites of transcription factors were determined by ChIP-seq and annotated to genes. Binding sites that localize to gene promoters are fairly straightforward annotated to genes, but this is not the case for binding sites that localize to (putative) enhancers. In this thesis three different gene annotation strategies were employed. In **Chapter 2** binding sites were annotated to genes with the nearest TSS, excluding those that are beyond 20 kb distance from the TSS. This methodology of annotation is not ideal as data from chromatin conformation capture-derived technologies suggests that the majority of enhancers do not interact with the nearest promoter[24-26]. It is therefore likely that annotation to the nearest TSS misses many valid target genes and in addition will include false positive genes that are closest to a binding site, but are not regulated by this transcription factor (most often as they are inactive). In **Chapter 2** the latter was overcome by only including genes of which the expression is significantly perturbed upon knock out or knock down of the transcription factor.

In **Chapter 3** Tcf4 target genes in NSCs were required to contain 1 or more binding sites within 100 kb distance from its TSS and to have a significant perturbed expression upon knock down of Tcf4. This is different to **Chapter 2**, as binding sites are now annotated to genes instead of genes to binding sites. For example, in **Chapter 2** a binding site that

is in between two genes will be annotated to the closest gene, even if the distance of both genes to the binding site is not substantially different. **In Chapter 3** this binding site is annotated to both genes. The disadvantage is obvious, as it will introduce more false positive target genes. In **Chapter 3** this is partly overcome by including gene expression perturbation data upon Tcf4 KD.

Sikora-Wohlfeld et al., (2013)[27] summarized the different strategies used in literature to identify transcription factor target genes using ChIP-seq data only (Figure 2). Performance of the different strategies was assessed by using gene expression datasets in amongst other ESCs, consistency of target gene prediction between different ChIP-seq datasets (for the same transcription factor) and functional homogeneity of predicted target genes. The best performing methods according to these parameters was what they called the 'ClosestGene' method. This method assigns, like in **Chapter 2**, the binding sites to the closest gene. In addition, the ClosestGene method scores the genes based on the distance between its TSS and the binding sites annotated to this gene, where these distances are normalized for the transcription factor specific genomic distribution. For each gene a score is calculated, which is the sum of the score of all binding sites that are annotated to this gene. Determining which genes are transcription factor regulated is based on this gene score, but still involves an arbitrary cutoff, which are the 500 highest scoring genes according to the ClosestGene method[27]. In **Chapter 4** the ClosestGene method according to Sikora-Wohlfeld was used to accrue a gene target list for Gene Ontology of Sox2 and Nanog binding sites that were maintained or lost upon depletion of Oct4. Other methods discussed by Sikora-Wohlfeld et al. perform according to them less well than their ClosestGene method and included scoring based on: peak intensities only, peak intensities and distance from the TSS, normalization for random binding events, distance without normalization for genomic distribution of the binding sites and determining the presence of a binding site in a set window around the TSS of a gene[27]. Despite all these methods, it is, I think, difficult to assign target genes for a transcription factor solely based on ChIP-seq data, because ChIP-seq does not contain information

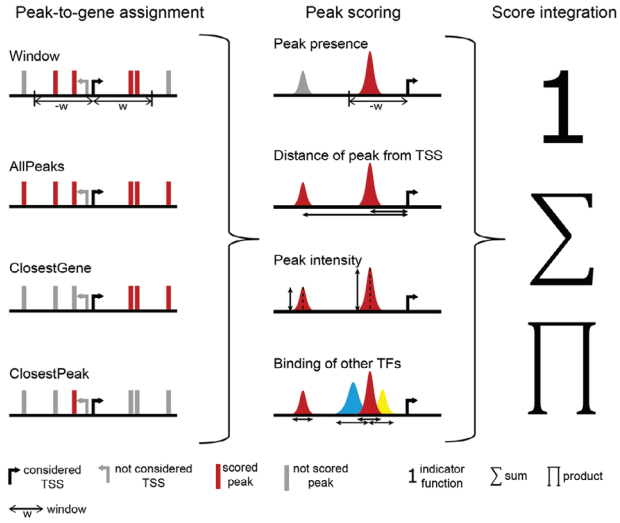


Figure 2. "Overview of the scoring procedures" from Figure 1 of Sikora-Wohlfeld et al., (2013)[27]. Corresponding legend from[27]: "Target gene scoring consists of three steps: (1) peak-to-gene assignment, (2) peak scoring and (3) integration of individual peak scores. Black arrow indicates transcription start site (TSS) of the gene that is to be scored. Grey arrow indicates a TSS of another gene (currently not scored). Red color indicates peaks that are assigned to the evaluated (black) gene; grey peaks are not assigned to this TSS by the given peak-to-gene assignment method. Blue and yellow peaks are peaks of other TFs that might be used to score the functionality of binding sites".

about gene regulation or genomic interaction. To overcome false positives in **Chapter 2 and 3,** we combined ChIP-seq data with gene expression perturbation data. For the identification of target genes corresponding to enhancers, one would ideally combine ChIP-seq with gene-expression perturbation data and chromatin interaction data derived from 5C, ChIA-PET or related technologies, where 5C has a limited resolution and will miss many interaction of enhancers that are within 10 kb of a TSS (see also **Chapter 1**).

Cooperative binding of Oct4, Sox2 and Nanog

In **Chapter 4** the effect of acute Oct4 and Sox2 depletion on the genome localization of Sox2 and Nanog was investigated. It was demonstrated that Sox2 and Nanog require Oct4 for genome wide recruitment to binding sites with the composite Sox2-Oct4 motif. Change in genome binding of Sox2 and Nanog upon Oct4 depletion appeared to be highly correlated and Sox2 and Nanog bind a subset of binding sites enriched for the Sox2 motif independent of Oct4. Together this suggests that Oct4 recruits Sox2 together with Nanog on Oct4-Sox2-Nanog binding sites. It was attempted to compare published Oct4 depletion gene expression perturbation data of genes annotated to maintained binding sites with the S-motif with genes annotated to lost binding sites with the composite Sox2-Oct4 motif, but only modest differences in gene expression perturbation were found (**Chapter 4, Figure 4e**). Lack of accuracy in annotation of lost and maintained binding sites based on ChIP-seq data only (see above) may have obscured the results of this comparison. To further elucidate the relationship between depletion of Oct4 and Sox2 and transcriptional activity of enhancers where transcription factor binding of Oct4, Sox2 and Nanog are lost or maintained, STARR-seq could be performed. STARR-seq is a high throughput technique that assesses enhancer activity of DNA-regulatory region library in a cell type of choice (see chapter 1 and[28]). The regulatory region library would be assembled using precipitated DNA from the ChIPs against Oct4, Sox2 and Nanog as performed in Chapter 4. Using this library STARR-seq would be performed in the doxycycline-inducible Oct4-null ESC line ZHBTc4[29] and Sox2-null[30] ESC lines and possibly in the Nanog-null mutant cell line[31] to assess enhancer activity upon depletion of Oct4, Sox2 or Nanog. Integrating this STARR-seq data set with the binding perturbation data from Chapter 4, followed by de-novo motif discovery, may result in identification of binding motifs of other transcription factors and would possibly provide an explanation for the both activating an repressive functions associated with Oct4 , Sox2 and Nanog.

REFERENCES

1. Nakamura, T., Nakagawa, M., Ichisaka, T., Shiota, A. & Yamanaka, S. Essential roles of ECAT15-2/Dppa2 in functional lung development. Mol Cell Biol 31, 4366-78 (2011).
2. Nakamura, T. et al. PGC7 binds histone H3K9me2 to protect against conversion of 5mC to 5hmC in early embryos. Nature 486, 415-9 (2012).
3. Ernst, J. et al. Mapping and analysis of chromatin state dynamics in nine human cell types. Nature 473, 43-9 (2011).
4. Rada-Iglesias, A. et al. A unique chromatin signature uncovers early developmental enhancers in humans. Nature 470, 279-83 (2011).
5. Creyghton, M.P. et al. Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proc Natl Acad Sci U S A 107, 21931-6 (2010).
6. Liu, N. et al. Recognition of H3K9 methylation by GLP is required for efficient establishment of H3K9 methylation, rapid target gene repression, and mouse viability. Genes Dev 29, 379-93 (2015).
7. Mikkelsen, T.S. et al. Genome-wide maps of chromatin state in pluripotent and lineage-committed

cells. Nature 448, 553-60 (2007).
8. Das, P.P. et al. Distinct and combinatorial functions of Jmjd2b/Kdm4b and Jmjd2c/Kdm4c in mouse embryonic stem cell identity. Mol Cell 53, 32-48 (2014).
9. Tung, P.Y., Varlakhanova, N.V. & Knoepfler, P.S. Identification of DPPA4 and DPPA2 as a novel family of pluripotency-related oncogenes. Stem Cells 31, 2330-42 (2013).
10. Masaki, H., Nishida, T., Kitajima, S., Asahina, K. & Teraoka, H. Developmental pluripotency-associated 4 (DPPA4) localized in active chromatin inhibits mouse embryonic stem cell differentiation into a primitive ectoderm lineage. J Biol Chem 282, 33034-42 (2007).
11. Nakamura, T. et al. PGC7/Stella protects against DNA demethylation in early embryogenesis. Nat Cell Biol 9, 64-71 (2007).
12. Wossidlo, M. et al. 5-Hydroxymethylcytosine in the mammalian zygote is linked with epigenetic reprogramming. Nat Commun 2, 241 (2011).
13. Maldonado-Saldivia, J. et al. Dppa2 and Dppa4 are closely linked SAP motif genes restricted to pluripotent cells and the germ line. Stem Cells 25, 19-28 (2007).
14. Lee, E. et al. Analysis of nuclear reprogramming in cloned miniature pig embryos by expression of Oct-4 and Oct-4 related genes. Biochem Biophys Res Commun 348, 1419-28 (2006).
15. Engelen, E. et al. Sox2 cooperates with Chd7 to regulate genes that are mutated in human syndromes. Nat Genet 43, 607-11 (2011).
16. Castro, D.S. et al. A novel function of the proneural factor Ascl1 in progenitor proliferation identified by genome-wide characterization of its targets. Genes Dev 25, 930-45 (2011).
17. Amiel, J. et al. Polyalanine expansion and frameshift mutations of the paired-like homeobox gene PHOX2B in congenital central hypoventilation syndrome. Nat Genet 33, 459-61 (2003).
18. de Pontual, L. et al. Noradrenergic neuronal development is impaired by mutation of the proneural HASH-1 gene in congenital central hypoventilation syndrome (Ondine's curse). Hum Mol Genet 12, 3173-80 (2003).
19. Amiel, J. et al. Mutations in TCF4, encoding a class I basic helix-loop-helix transcription factor, are responsible for Pitt-Hopkins syndrome, a severe epileptic encephalopathy associated with autonomic dysfunction. Am J Hum Genet 80, 988-93 (2007).
20. Zweier, C. et al. Haploinsufficiency of TCF4 causes syndromal mental retardation with intermittent hyperventilation (Pitt-Hopkins syndrome). Am J Hum Genet 80, 994-1001 (2007).
21. Dauger, S. et al. Phox2b controls the development of peripheral chemoreceptors and afferent visceral pathways. Development 130, 6635-42 (2003).
22. Whalen, S. et al. Novel comprehensive diagnostic strategy in Pitt-Hopkins syndrome: clinical score and further delineation of the TCF4 mutational spectrum. Hum Mutat 33, 64-72 (2012).
23. Zweier, C. et al. CNTNAP2 and NRXN1 are mutated in autosomal-recessive Pitt-Hopkins-like mental retardation and determine the level of a common synaptic protein in Drosophila. Am J Hum Genet 85, 655-66 (2009).
24. Kieffer-Kwon, K.R. et al. Interactome maps of mouse gene regulatory domains reveal basic principles of transcriptional regulation. Cell 155, 1507-20 (2013).
25. Sanyal, A., Lajoie, B.R., Jain, G. & Dekker, J. The long-range interaction landscape of gene promoters. Nature 489, 109-13 (2012).
26. Li, G. et al. Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. Cell 148, 84-98 (2012).
27. Sikora-Wohlfeld, W., Ackermann, M., Christodoulou, E.G., Singaravelu, K. & Beyer, A. Assessing computational methods for transcription factor target gene identification based on ChIP-seq data. PLoS Comput Biol 9, e1003342 (2013).
28. Arnold, C.D. et al. Genome-wide quantitative enhancer activity maps identified by STARR-seq. Science 339, 1074-7 (2013).
29. Niwa, H., Miyazaki, J. & Smith, A.G. Quantitative expression of Oct-3/4 defines differentiation, dedifferentiation or self-renewal of ES cells. Nat Genet 24, 372-6 (2000).
30. Masui, S. et al. Pluripotency governed by Sox2 via regulation of Oct3/4 expression in mouse embryonic stem cells. Nat Cell Biol 9, 625-35 (2007).
31. Chambers, I. et al. Nanog safeguards pluripotency and mediates germline development. Nature 450, 1230-4 (2007).

Addendum

SUMMARY

Cell type-specific sets of transcription factors control cell identity by implementing cell type specific gene expression. Transcription factors bind to regulatory regions such as promoters and enhancers on the genome. Transcription factors regulate gene expression, often together with other transcription factors. In this thesis the genome localization and interaction partners of transcription factors are studied in Embryonic Stem Cells (ESCs) and Neural Stem Cells (NSCs), leading to an improved understanding of their molecular environment and mode of action.

Proteins that bind to regulatory regions such as promoters and enhancers have not been identified systemically. In Chapter 2 active enhancers, promoters or heterochromatin are purified from ESCs by Chromatin Immunoprecipitations for specific histone modifications and co-purified proteins are identified by mass spectrometry, a method we name ChIP-MS. ChIP-MS identifies 239 proteins that are predicted to localize to promoters or enhancers with different levels of activity, or heterochromatin. These localization predictions are verified using published genome-wide datasets that were available for 28 ChIP-MS identified proteins and indicate a high accuracy of genome localization prediction by ChIP-MS. 63 out of 239 ChIP-MS identified proteins are important for pluripotency and include Oct4, Esrrb, Klf5, Mycn and Dppa2, which are reprogramming factors for induced pluripotent stem cells. We select Dppa2 for further investigation, because ChIP-MS data suggests that Dppa2 binds to promoters with low activity, which is unusual for a pluripotency-inducing factor. By a ChIP-seq of V5 tagged Dppa2, we confirm that the ChIP-MS data correctly identifies Dppa2 as a transcription factor that binds to promoters with low activity. Dppa2 is specifically expressed in ESCs or related pluripotent stem cells and Chapter 2 shows that Dppa2 is most likely a transcriptional activator, but that nearly all Dppa2 target genes have their highest expression in other tissues than ESCs. In addition, Dppa2 binding sites show no overlap with the binding sites of pluripotency factors Oct4, Nanog and Esrrb and we therefore conclude that Dppa2 is not part of the classical pluripotency network. Chapter 5 shows that in addition to histone modification H3K4me3 and H3K27ac, Dppa2 bound promoters are also enriched for H3K27me3.

In Chapter 3 the interacting proteins of four transcription factors in mouse NSCs are identified and assembled into a network of over 200 proteins, which is the first transcription factor interaction network in a neural system and includes 68 proteins that are associated with mental disorders. The network is highly enriched for proteins encoded by evolutionarily constrained genes in human. Evolutionarily constrained genes are more frequently mutated in disease, including Autism Spectrum Disorders and this suggests that the network probably contains undiscovered Mental Disorder proteins and can be used as a discovery tool for novel mental disorder genes and to establish the molecular connections for known mental disorder proteins. Chapter 3 also shows that proteins that interact or share a network are also more likely to overlap on the genome. Tcf4 and Ascl1 have a strong overlap in human phenotypes and we show that they interact, have strong overlaps in genomic binding sites and share target genes, amongst which *Phox2b*. Activation of *PHOX2B* during development is likely affected by mutations in TCF4 and ASCL1 and this would explain the overlapping abnormal breathing phenotypes observed in Pit Hopkins syndrome and Congenital Central Hypoventilation Syndrome, that can be

caused by tcf4 and ascl1, respectively.

Transcription factors can cooperatively bind to the genome and regulate gene expression. Transcription factors Oct4, Sox2 and Nanog maintain pluripotency and self-renewal of ESCs and often co-localize on the ESC genome, but cooperative binding of Sox2 and Oct4 was never investigated genome-wide. Chapter 4 shows that genome-wide binding of Oct4 and Sox2 to the ESC genome is interdependent on binding sites that contain the composite Sox2-Oct4-motif, but that binding of Oct4 or Sox2 is not reciprocally reduced upon depletion of Sox2 or Oct4, respectively. It further more shows that Oct4 directly recruits Sox2 and that Nanog is recruited via Sox2 to binding sites that contain the composite Sox2-Oct4 motif and that Sox2 and Nanog also bind the ESC genome independent Oct4 on binding sites that contain the Sox2-motif only.

Finally, Chapter 5 provides additional discussion to the previous chapters and puts the findings of these chapters in broader context and discusses new directions for further research.

SAMENVATTING

Celtype-specifieke transcriptiefactoren beheersen de identiteit van cellen door het bewerkstellen van celtype-specifieke genexpressie. Transcriptiefactoren binden aan regulerende regio's op het genoom, zoals promotoren en enhancers. Transcriptiefactoren reguleren genexpressie, vaak in combinatie met andere transcriptiefactoren. In dit proefschrift worden de lokalisatie op het genoom en de interacterende eiwitpartners van transcriptiefactoren in embryonale stamcellen (ES cellen) en neurale stamcellen (NS cellen) nader onderzocht en dit lijdt tot een verbeterd inzicht van de moleculaire omgeving en werking van transcriptiefactoren.

Eiwitten die binden aan regulerende regio's op het genoom zijn nog niet systematisch geïdentificeerd. In hoofdstuk 2 worden enhancers, promotoren en heterochromatine opgezuiverd uit ES cellen doormiddel van Chromatine-immunoprecipitaties van bepaalde histonmodificaties. Vervolgens wordt de identiteit van co-gepurificeerde eiwitten vastgesteld doormiddel van massaspectrometrie. Deze methode noemen wij ChIP-MS. Doormiddel van ChIP-MS voorspellen we voor 239 eiwitten dat deze binden aan promotoren of enhancers van verschillende activiteit of heterochromatine. Deze lokalisatievoorspellingen zijn getoetst met behulp van gepubliceerde genoombrede datasets die voor 28 eiwitten beschikbaar waren en hieruit blijkt een hoge nauwkeurigheid van de uit ChIP-MS verkregen lokalisatievoorspellingen. 63 van de 239 doormiddel van ChIP-MS geïdentificeerde eiwitten zijn belangrijk voor pluripotentie van ES cellen, inclusief Oct4, Esrrb, Klf5 en Dppa2 die behoren tot de transcriptiefactoren die het herprogrammeren van somatische cellen naar pluripotente stamcellen mogelijk maken. Dppa2 is nader onderzocht, omdat ChIP-MS erop duidt dat Dppa2 aan minder actieve promotoren bindt en dit is ongebruikelijk voor transcriptiefactoren die kunnen herprogrammeren. Doormiddel van een ChIP-seq van Dppa2 bevestigen we dat Dppa2 inderdaad aan minder actieve promotoren bindt. Dppa2 wordt specifiek in ES cellen of gerelateerde pluripotente stamcellen tot expressie gebracht. Hoofdstuk 2 laat zien dat Dppa2 waarschijnlijk een transcriptionele activator is, maar ook dat bijna alle Dppa2 gereguleerde genen het hoogst tot expressie komen in andere weefsels dan ES cellen. Ook overlappen Dppa2 bindingsplekken op het genoom niet met die van pluripotentiefactoren zoals Oct4, Nanog en Essrb, waardoor we de conclusie trekken dat Dppa2 niet deel uit maakt van het klassieke pluripotentienetwerk in ES cellen. Hoofdstuk 5 laat zien dat door Dppa2 gebonden promotoren naast verrijkt te zijn voor histonmodificaties H3K4me3 en H3K27ac, ook verrijkt zijn voor H3K27me3.

In hoofdstuk 3 worden de interacterende eiwitten van vier transcriptiefactoren in muizen NS cellen geïdentificeerd en samengevoegd tot een netwerk van meer dan 200 eiwitten. Dit is het eerste interactienetwerk van transcriptiefactoren in een neuraal celtype en bevat 68 eiwitten die in verband worden gebracht met psychische aandoeningen. Het netwerk is verrijkt voor eiwitten die gecodeerd worden door in de mens evolutionair invariante genen. Mutaties in evolutionair invariante genen leiden vaak tot ziektes, waaronder Autisme spectrum stoornissen. Dit doet vermoeden dat het beschreven netwerk waarschijnlijk nieuwe genen bevat die in verband kunnen worden gebracht met psychische aandoeningen en dat het netwerk dus als hulpmiddel kan worden gebruikt voor het ontdekken van zulke genen. Daarnaast kan het ook worden gebruikt

voor het vaststellen van moleculaire verbanden tussen bestaande genen die in verband worden gebracht met psychiatrische aandoeningen. Hoofdstuk 3 laat ook zien dat het voor eiwitten die interacteren of onderdeel uitmaken van het netwerk waarschijnlijker is dat ze ook overlap vertonen in hun bindingsplaatsen op het genoom. Tcf4 en Ascl1 hebben een grote overlap van fenotypen in de mens en hoofdstuk 3 laat zien dat ze met elkaar interacteren en hun bindingsplaatsen op het genoom een grote overlap vertonen waaronder bij het ziekte-relevante *Phox2b* gen. Mutaties in TCF4 en ASCL1 hebben waarschijnlijk invloed op de activatie van *PHOX2B* tijdens de ontwikkeling en dit zou de abnormale overlappende ademhalingsfenotypen verklaren die worden waargenomen in Pit Hopkins syndroom en aangeboren centraal hypoventilatiesyndroom en die respectievelijk veroorzaakt kunnen worden door tcf4 en ascl1.

Transcriptie factoren kunnen coöperatief binden aan het genoom en zo genexpressie gezamenlijk reguleren. Transcriptiefactoren Oct4, Sox2 en Nanog behouden de pluripotentie en zelf-vernieuwing van ES cellen en binden vaak op dezelfde bindingsplaatsen op het genoom, maar het coöperatief binden van Oct4, Sox2 en Nanog is nog nooit genoombreed onderzocht. Hoofdstuk 4 laat zien dat genoombrede binding van Oct4 and Sox2 wederzijds afhankelijk is op bindingsplaatsen die het samengestelde Sox2-Oct4 bindingsmotief bevatten, maar dat binding van Oct4 en Sox2 niet in dezelfde mate worden beïnvloed door depletie van respectievelijk Sox2 of Oct4. Het laat daarnaast zien dat Oct4 direct Sox2 rekruteert naar bindingsplaatsen met het samengestelde Sox2-Oct4 motief en dat Nanog door Sox2 mede wordt gerekruteerd. Daarnaast zijn Sox2 en Nanog ook onafhankelijk van Oct4 in staat om te binden aan bindingsplaatsen die een Sox2 motief bevatten.

Tenslotte vindt er hoofdstuk 5 aanvullende discussie plaats met betrekking tot de voorgaande hoofdstukken, plaatst het de bevinden van deze hoofdstukken in een bredere context en bediscussieert het de mogelijkheden voor vervolgonderzoek.

CURRICULUM VITAE

PERSONAL DETAILS

Name	Johannes Hendrik Brandsma
Date of birth	29 April 1986
Place of birth:	Heerenveen, The Netherlands

EDUCATION

2011-2016	PhD program at Erasmus Medical Centre Department of Cell Biology, Erasmus Medical Centre, Rotterdam, The Netherlands
2009-2001	Master of Science in Molecular Biology & Biotechnology University of Groningen, Groningen, The Netherlands
2005-2008	Bachelor of Science in Molecular Biology University of Groningen, Groningen, The Netherlands
1998-2005	VWO Natuur & Gezondheid RSG Tromp Meesters, Steenwijk, The Netherlands

RESEARCH

2011 – 2016	PhD research Department of Cell Biology, Erasmus MC, Rotterdam, The Netherlands (Prof.dr. F.G. Grosveld and Dr. R.A. Poot)
2011	MSc research project Department of Medical Biology, University Medical Centre Groningen, The Netherlands (Prof.dr. M.G. Rots)
2010	MSc research project Department of Microbiology, University of Groningen, Groningen, The Netherlands (Prof.dr. A.J.M Driessen)

PUBLICATIONS

Engelen E*,**Brandsma JH***,Moen MJ, Signorile L, Dekkers DHW, Demmers J, Kockx CEM, Ozgür Z, Van IJcken WFJ, Van den Berg DL, Poot RA. Proteins that bind regulatory regions identified by histone modification chromatin immunoprecipitations and mass spectrometry. Nat Commun 6, 7155 (2015)

Benjamin D et **Social Science Genetic Association Consortium**. Education-associated SNPs are enriched for brain function and disorders. (under revision)

Moen MJ, Adams HH*, **Brandsma JH***, Dekkers DHW, Akinci U, Karkampouna S, Kockx CEM, Ozgür Z, Van IJcken WFJ, Demmers J, Poot RA. A protein interaction network of mental disorder factors in neural stem cells. (submitted)

Wang W, Xu L, **Brandsma JH**, Wang Y, Hakim MS, Zhou X, Yin Y, Fuhler GM, Van der Laan LJW, C. Van der Woude J, Sprengers D, Metselaar HJ, Smits R, Poot RA, Peppelenbosch MP, Pan Q. Convergent Transcription of Interferon-stimulated Genes by TNF-α and Interferon-α Augments Their Antiviral Activity against HCV and HEV (under revision).

Wang W, Wang Y, Zhou X, Yin Y, Xu L, Debing Y, Carrillo EH, **Brandsma JH,** Sprengers D, Poot RA, Metselaar HJ, Smits R, Berkhout B, Neyts J, Peppelenbosch MP, Pan Q PKCα/AP-1 cascade directly drives transcription of interferon-stimulated genes and exerts broad antiviral activity (submitted)

(* equal author contribution)

PHD PORTFOLIO

Name student	Johannes H. Brandsma
Erasmus MC Department	Cell biology
Research school	Graduate School MGC
PhD Period	September 2011 – February 2016
Promoter	Frank G. Grosveld
Co-promoter	Raymond A. Poot

PhD training

Courses

2011 Biochemistry and Biophysics (Rotterdam)
2011 Safely working in the laboratory (Leiden)
2012 Cell and Developmental Biology (Rotterdam)
2012 Genetics (Rotterdam)
2012 Next Generation Sequencing data analysis (Rotterdam)
2012 Analysis of microarray gene expression data using R/BioC and web tools (Rotterdam)
2013 Scripting for life science researchers (Leiden)
2013 Literature course (Rotterdam)
2014 RNA-seq workshop (Oslo, Norway)

Workshops, Symposia and Conferences

2011 1st Chromatin Symposium: Chromatin Changes in Differentiation and Malignancies, Giessen, Germany
2012 19th MGC PhD workshop, Dusseldorf, Germany
2012 NIRM/ISD conference Stem Cells, Development and Regulation, Amsterdam, The Netherlands
2013 3rd Winter School of the Collaborative Research Centre TRR81, Kleinwalsertal, Austria (oral presentation)
2013 SBBCD-BVCOB: Experimental models of human diseases, Luik, Belgium
2013 20th MGC PhD workshop, Luxemburg, Luxemburg (poster presentation)
2013 23rd MGC Symposium, Rotterdam, The Netherlands
2014 4th Winter School of the Collaborative Research Centre TRR81, Kleinwalsertal, Austria (oral presentation)
2014 4th IUAP DevRepair meeting, Brussel, Belgium (oral presentation)
2014 21st MGC PhD workshop, Munster, Germany (oral presentation)
2014 24th MGC Symposium, Rotterdam, The Netherlands
2014 5th IUAP DevRepair meeting, Rotterdam, The Netherlands
2015 5th Winter School of the Collaborative Research Centre TRR81, Kleinwalsertal, Austria (oral presentation)
2015 8th annual meeting of the DSSCR (oral presentation), Utrecht, The Netherlands
2015 FASEB: Transcription, Chromatin, and Epigenetics, Palm Beach, Florida, USA (poster presentation)
2015 3rd Chromatin Symposium: Chromatin Changes in Differentiation and Malignancies, Marburg, Germany (poster presentation)

Additional Activities

2013 Junior science Program, Erasmus MC, 3 high school students

DANKWOORD

Although only my name is on the cover of this thesis, it also includes the hard work of many others and I would like to use the next few pages to especially thank them and others for their hard work, support, advice, collegiality and friendship. I attribute this thesis to them.

First, I thank my co-promoter and daily supervisor Raymond. I value the many conversations we had, in which you always kept your calm, even if I lost mine. Your feedback and advice resulted in me making better decisions, whether this concerned the latest version of a figure I squeezed out of our data, experiments, a presentation, writing this thesis or career advice. I thank you for giving me the opportunity to work in your group, I learned a lot and had a good time. I wish you all the best!

My promoter Frank, Thank you for being my promoter and allowing me to pursue my doctoral research in your (now former) department of Cell Biology. Your critical questions during the Monday morning meetings, Monthly PhD meetings and winter school in Kleinwalsertal were the only questions I truly feared.

The members of the reading committee, Sjaak Philipsen, Wouter de Laat and Joost Gribnau. Thank you for being in my reading committee and for promptly reading my thesis and for the provided feedback. Sjaak, I also want to thank you for organizing the winterschool in Kleinwalsertal, which were definitely (the) highlights of my PhD! Joost, thank you (and also Raymond) for organizing the seminar series 'Frontiers in Science in the Low Countries'.

I also thank the other members of my committee, Robert Hofstra, Danny Huylebroeck and Harmen van de Werken. Danny, I always appreciated your feedback. I really enjoyed the biannual Belgium beer tastings and thank you for inviting me to speak at the 4th IUAP DevRepair meeting in Brussel. Harmen, I thank you for the many interesting discussions we had during lunch. To desperation of the others in our lunch group, this often concerned our endeavors in bioinformatics.

My Paranimfen Maaike and Agnese, it is a honor to have you both standing at my side. Agnese, although our work together has not made it into my thesis, I really enjoyed working with you. I hope the fact that my paranimfen only have X sex chromosomes, partly compensates the whole male defense committee. Sorry Agnese, one more that adds to the shameful bias of your list. Maaike, we went through all the trials and tribulations of the PhD together. I especially thank you for the work that is part of Chapter 3 of this thesis and for often taking the initiative in organizing social activities, such as Sinterklaas, for our lab. Maaike, you are next!

I want to thank (former) Lab706 members: Erik, thank you for the enormous work you did for Chapter 2. Debbie, although most of the contact we had was via email, you contributed enormously to this thesis with experiments and ideas from the time you were part of Raymond's lab. Also thanks to Umut, Mike, Luca, Hieab and Sofia for their contributions to the different chapters of this thesis. Martití! I now transfer the responsibility of Flaming Mexican game instructor to you. Burn a beard (I also settle for a pair of eyebrows) for me this year at Kleinwalsertal! I also would like to thank the people of the Proteomics Centre, especially Jeroen Demmers and Dick Dekkers, for doing the mass spectrometry in Chapter 2 and 3. I

also thank the Biomics Centre, especially Zeliha and Wilfred for sequencing all our data that is part of Chapter 2-4. I thank Florian Halbritter and Simon Tomlinson for their pilot analysis of some ChIP-seq data in Chapter 4.

Absolutely essential for the realization of this thesis was learning to do data analysis of Next Generation Sequencing data. This would not have been possible without online communities such as, for example, biostars.org and seqanswers.com. Thank you to all members of these communities that ever asked or answered questions about programming or data analysis.

Then there are the many people whom I like to thank for other contributions such as collaborations, suggestions, questions, scientific discussions, chit-chat or for the social aspects surrounding science. I thank members of Sjaak's lab, with whom I often broke bread and I got to know at least as good as the members of my own lab: Ileana, we started together and I had the honor of being there at the end as your paranimf. Unfortunately you are not able to be my paranimf, but I know that you are in spirit! Maria, my favorite office neighbor, until you looked for higher-up places (10th floor). Thankfully you left your papoetsia for me to remember you by. Sylvia, thank you for your Rotterdam spirit in our multiculti work environment and for warning me against the bureaucratic pitfalls of buying a house. dr. Divine, Pavlos, Thamar and Nynke. Ernie in 706 for all the experimental wisdom, enjoy your approaching retirement! Neighbors at 702 and 710: Anita, Andrea, Guillaume, Petros, Dubi, Robert-Jan, Rik, Alex, Rien and Michael, thank you all for a pleasant work environment. Many thanks to the upstairs people of Cell biology: Niels, Gert-Jan, Kerstin, Nesrin, Derk, Dorota, Johannes, Reinier, Chris, Thomas and Jessica I hope to see you again in Budapest someday, Mihaela thanks for teaching me all about (culturing) pericytes. Thanks to the party people of developmental biology: Cheryl, Fabrizia, Willy, Federica, Hegias, Ruben, Friedemann for our (intoxicated) discussions that you never remembered the next day. Aristea my areola will never be the same again. I thank the activity/party committee (R.I.P.) for organizing borrels and parties. Thank you to Marike, Jasperina and Bep, the secretaries of Cell Biology, for assisting and always being helpful with the administrative procedures (especially those surrounding the defense). I also thank all the other staff that is or was once part of Theme Biomedical Sciences and that keeps the cluster going.

Pap en Mam, zonder jullie aanmoediging, steun en toeverlaat was ik nooit gekomen waar ik nu ben. Ook dank ik mijn broers en zus bij wie ik in altijd terecht kan.

Als laatste, maar zeker niet de minste bedank ik mijn vrouw Lídia. Je hebt mij "je mannetje" door alle dalen heen geholpen en me altijd van een liefdevol thuis voorzien waar ik de teleurstelling van tegenvallende resultaten van me af kon zetten. Ik ben niet altijd even makkelijk geweest in de afgelopen vier jaar en ik ben heel gelukkig met jouw onvoorwaardelijke steun en grenzeloze liefde.

Szeretlek, kismamám.

Grtz,

Johan