Cell Press Selections

Reprint supplement

Immunoprecipitation

Insights into Protein-Protein and Protein-Nucleic Acid Interaction

CellPress
www.cell.com

We make really good antibodies.

Additional Resources from Elsevier

METHODS:

CLIP: A method for identifying protein–RNA interaction sites in living cells

Jernej Ulea, Kirk Jensen, Aldo Melea, and Robert B. Darnella
Volume 37, Issue 4, December 2005, Pages 376–386 (214)

ChIP-seq: Using high-throughput sequencing to discover protein–DNA interactions

Dominic Schmidt, Michael D. Wilson, Christiana Spyrou, Gordon D. Brown , James Hadfield, and Duncan T. Odom
Volume 48, Issue 3, July 2009, Pages 240–248 (120)

LABORATORY METHODS IN ENZYMOLOGY: PROTEIN PART C

Protein Protocols/Protein Precipitation

Chapter One: TCA Precipitation

Laura Koontz
Pages 3–10

Protein Protocols/Protein Pull-Down Methods

Chapter Two: Coimmunoprecipitation of Proteins from Yeast

Erica Gerace and Danesh Moazed
Pages 13–26

Foreword

We are pleased to introduce the latest edition of *Cell Press Selections*. These editorially curated reprint collections highlight a particular area of life science by bringing together articles from the Cell Press journals. In this selection, we present recent insights into chromatin immunoprecipitation (ChIP) technologies.

Interactions between proteins and DNA are essential to life. These interactions mediate transcription, DNA replication, recombination, and DNA repair, processes central to the biology of every organism.

To begin understanding the crosstalk between DNA and proteins, an early report in 1978 demonstrated chromatin crosslinked with formaldehyde, preserving DNA-protein and protein-protein interactions (Jackson, 1978, *Cell*). A few years later, the Lis and Varshavsky labs added an immunoprecipitation step with specific histone antibodies to capture these protein-DNA complexes crosslinked via UV or formaldehyde, moving this technology forward with their seminal publications describing ChIP (Lis, 1984, *Proc. Natl. Acad. Sci. USA*; Varshavsky, 1988, *Cell*). Following these pioneering reports, ChIP has been extensively developed and refined, with improved variations on this methodology continually arising.

We hope that you will enjoy reading this collection of articles and will visit www.cell.com to find other high-quality research and review articles touting the uses of ChIP technology.

Finally, we are grateful for the generosity of Bethyl Laboratories, who helped to make this reprint collection possible.

For more information about Cell Press Selections:
Gordon Sheffield
Program Director, Cell Press Selections
g.sheffield@cell.com
617-386-2189

Immunoprecipitation

Article

Interactions between JARID2 and Noncoding RNAs Regulate PRC2 Recruitment to Chromatin	*Syuzo Kaneko, Roberto Bonasio, Ricardo Saldaña-Meyer, Takahaki Yoshida, Jinsook Son, Koichiro Nishino, Akihiro Umezawa, and Danny Reinberg*

Resources

Quantitative ChIP-Seq Normalization Reveals Global Modulation of the Epigenome	*David A. Orlando, Mei Wei Chen, Victoria E. Brown, Snehakumari Solanki, Yoon J. Choi, Eric R. Olson, Christian C. Fritz, James E. Bradner, and Matthew G. Guenther*
Determination and Inference of Eukaryotic Transcription Factor Sequence Specificity	*Matthew T. Weirauch, Ally Yang, Mihai Albu, Atina G. Cote, Alejandro Montenegro-Montero, Philipp Drewe, Hamed S. Najafabadi, Samuel A. Lambert, Ishminder Mann, Kate Cook, Hong Zheng, Alejandra Goity, Harm van Bakel, Jean-Claude Lozano, Mary Galli, Mathew G. Lewsey, Eryong Huang, Tuhin Mukherjee, Xiaoting Chen, John S. Reece-Hoyes, Sridhar Govindarajan, Gad Shaulsky, Albertha J.M. Walhout, François-Yves Bouget, Gunnar Ratsch, Luis F. Larrondo, Joseph R. Ecker, and Timothy R. Hughes*
Unambiguous Identification of miRNA:Target Site Interactions by Different Types of Ligation Reactions	*Stefanie Grosswendt, Andrei Filipchyk, Mark Manzano, Filippos Klironomos, Marcel Schilling, Margareta Herzog, Eva Gottwein, and Nikolaus Rajewsky*
Analysis of the Human Endogenous Coregulator Complexome	*Anna Malovannaya, Rainer B. Lanz, Sung Yun Jung, Yaroslava Bulynko, Nguyen T. Le, Doug W. Chan, Chen Ding, Yi Shi, Nur Yucer, Giedre Krenciute, Beom-Jun Kim, Chunshu Li, Rui Chen, Wei Li, Yi Wang, Bert W. O'Malley, and Jun Qin*
Chromatin Immunoprecipitation Indirect Peaks Highlight Long-Range Interactions of Insulator Proteins and Pol II Pausing	*Jun Liang, Laurent Lacroix, Adrien Gamot, Suresh Cuddapah, Sophie Queille, Priscillia Lhoumaud, Pierre Lepetit, Pascal G.P. Martin, Jutta Vogelmann, Franck Court, Magali Hennion, Gaël Micas, Serge Urbach, Olivier Bouchez, Marcelo Nöllmann, Keji Zhao, Eldon Emberly, and Olivier Cuvier*
A High-Throughput Chromatin Immunoprecipitation Approach Reveals Principles of Dynamic Gene Regulation in Mammals	*Manuel Garber, Nir Yosef, Alon Goren, Raktima Raychowdhury, Anne Thielke, Mitchell Guttman, James Robinson, Brian Minie, Nicolas Chevrier, Zohar Itzhaki, Ronnie Blecher-Gonen, Chamutal Bornstein, Daniela Amann-Zalcenstein, Assaf Weiner, Dennis Friedrich, James Meldrim, Oren Ram, Christine Cheng, Andreas Gnirke, Sheila Fisher, Nir Friedman, Bang Wong, Bradley E. Bernstein, Chad Nusbaum, Nir Hacohen, Aviv Regev, and Ido Amit*
CAST-ChIP Maps Cell-Type-Specific Chromatin States in the *Drosophila* Central Nervous System	*Tamás Schauer, Petra C. Schwalie, Ava Handley, Carla E. Margulies, Paul Flicek, and Andreas G. Ladurner*
Dynamic *trans*-Acting Factor Colocalization in Human Cells	*Dan Xie, Alan P. Boyle, Linfeng Wu, Jie Zhai, Trupti Kawli, and Michael Snyder*
Combinatorial Patterning of Chromatin Regulators Uncovered by Genome-wide Location Analysis in Human Cells	*Oren Ram, Alon Goren, Ido Amit, Noam Shoresh, Nir Yosef, Jason Ernst, Manolis Kellis, Melissa Gymrek, Robbyn Issner, Michael Coyne, Timothy Durham, Xiaolan Zhang, Julie Donaghey, Charles B. Epstein, Aviv Regev, and Bradley E. Bernstein*

Molecular Cell
Article

Interactions between JARID2 and Noncoding RNAs Regulate PRC2 Recruitment to Chromatin

Syuzo Kaneko,[1,5] Roberto Bonasio,[1,4,5] Ricardo Saldaña-Meyer,[1] Takahaki Yoshida,[2] Jinsook Son,[1] Koichiro Nishino,[3] Akihiro Umezawa,[2] and Danny Reinberg[1,*]
[1]Howard Hughes Medical Institute and NYU School of Medicine, Department of Molecular Pharmacology and Biochemistry, New York, NY 10016, USA
[2]National Research Institute for Child Health and Development, Department of Reproductive Biology, Tokyo 157-8535, Japan
[3]University of Miyazaki, Faculty of Agriculture, Laboratory of Veterinary Biochemistry and Molecular Biology, Miyazaki 889-2192, Japan
[4]Present address: Department of Cell and Developmental Biology, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA 19104, USA
[5]These authors contributed equally to this work
*Correspondence: Danny.Reinberg@nyumc.org
http://dx.doi.org/10.1016/j.molcel.2013.11.012

SUMMARY

JARID2 is an accessory component of *Polycomb* repressive complex-2 (PRC2) required for the differentiation of embryonic stem cells (ESCs). A role for JARID2 in the recruitment of PRC2 to target genes silenced during differentiation has been put forward, but the molecular details remain unclear. We identified a 30-amino-acid region of JARID2 that mediates interactions with long noncoding RNAs (lncRNAs) and found that the presence of lncRNAs stimulated JARID2-EZH2 interactions in vitro and JARID2-mediated recruitment of PRC2 to chromatin in vivo. Native and crosslinked RNA immunoprecipitations of JARID2 revealed that Meg3 and other lncRNAs from the imprinted *Dlk1-Dio3* locus, an important regulator of development, interacted with PRC2 via JARID2. Lack of *MEG3* expression in human induced pluripotent cells altered the chromatin distribution of JARID2, PRC2, and H3K27me3. Our findings show that lncRNAs facilitate JARID2-PRC2 interactions on chromatin and suggest a mechanism by which lncRNAs contribute to PRC2 recruitment.

INTRODUCTION

Polycomb group (PcG) genes are key epigenetic regulators in multicellular organisms, as they maintain transcriptional repression of lineage-specific genes throughout development, thus contributing to the stability of cell identity (Schwartz and Pirrotta, 2007). All mammalian PcG protein complexes identified so far perform their epigenetic function by acting on chromatin (Lanzuolo and Orlando, 2012); in particular, the *Polycomb* repressive complex 2 (PRC2) is responsible for di- and trimethylation of lysine 27 in histone H3 (H3K27me2/3) (Margueron and Reinberg, 2011), a hallmark of facultative heterochromatin (Trojer and Reinberg, 2007).

One of the outstanding questions regarding mammalian PRC2 function is that of specificity of action: how are certain genes selected for repression while others are unaffected? How can the same molecular machinery silence different genes in different cell lineages? Because none of the core components of PRC2 (EZH2, EED, SUZ12, RBBP4/7) possess a DNA binding domain (Margueron and Reinberg, 2011), it is believed that chromatin targeting must be specified elsewhere, by interactions with DNA-binding factors (Boulay et al., 2012; Kim et al., 2009), preexisting histone methylation (Margueron et al., 2009), chromatin-associated long noncoding RNAs (lncRNAs) (Rinn et al., 2007; Tsai et al., 2010), or a combination thereof (Margueron and Reinberg, 2011).

One essential factor for proper recruitment of PRC2 during the early phases of embryonic stem cell (ESC) differentiation is the Jumonji family, ARID domain-containing protein JARID2 (Landeira et al., 2010; Li et al., 2010; Pasini et al., 2010; Peng et al., 2009; Shen et al., 2009), which is often deleted in chronic myeloid malignancies (Puda et al., 2012). In the absence of JARID2, PRC2 is recruited late and incompletely to its target genes and its enzymatic function is diminished (Li et al., 2010; Son et al., 2013), which results in failure to follow the differentiation program. Although JARID2 target sites are enriched for CGG- and GA-containing sequences (Peng et al., 2009), its DNA binding preferences lack the specificity to explain its distribution on chromatin (Li et al., 2010). Therefore, the nature of the recruitment pathway for JARID2 and the mode by which JARID2 regulates downstream steps of PRC2 assembly and function remain unclear.

Noncoding RNAs have been implicated in the regulation of epigenetic pathways, from early work on the lncRNA Xist in X chromosome inactivation (Brockdorff et al., 1992; Brown et al., 1992) and antisense transcripts in imprinted loci (John and Surani, 1996) to the more recent discovery of HOTAIR (Rinn et al., 2007) and its proposed role as a scaffold for chromatin-modifying "supercomplexes" (Tsai et al., 2010). Mammalian genomes contain thousands of lncRNAs (Guttman et al., 2009), most of which remain functionally uncharacterized. Because of their large size, potential for tertiary structure formation, and ability to form sequence-specific interactions with DNA, lncRNAs appear well suited to exchange information between

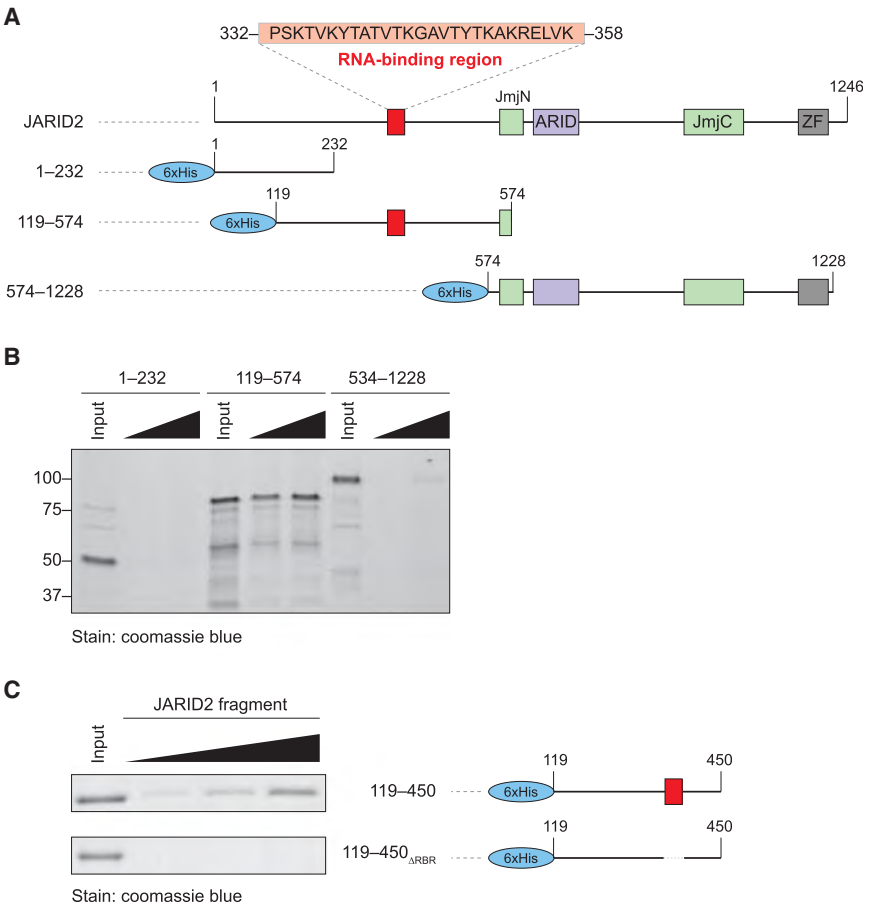


Figure 1. Identification of the RNA-Binding Region of JARID2

(A) Domain organization of human JARID2 and scheme of the 6xHis-fused truncations utilized in the mapping experiments.

(B) In vitro streptavidin pull-down after incubation of increasing concentrations of the indicated JARID2 recombinant fragments with biotinylated $HOTAIR_{1–333}$. Input, 2 μg; titration, 2 and 4 μg.

(C) High-resolution mapping of the residues of JARID2 necessary for RNA binding in vitro. The two indicated fragments (right) were incubated with $HOTAIR_{1–333}$ and assayed as in (B). Input, 2 μg; titration, 1, 2, and 4 μg.

See also Figure S1.

chromatin-modifying complexes and the genomic sequence (Bonasio et al., 2010; Rinn and Chang, 2012). *Polycomb* repressive complex-1 (PRC1), PRC2, and the MLL complex interact with the lncRNAs ANRIL, HOTAIR, and HOTTIP, respectively, and these interactions facilitate their recruitment to chromatin (Rinn et al., 2007; Wang et al., 2011; Yap et al., 2010). However, the molecular details and downstream consequences of these RNA-protein interactions remain poorly understood. For example, the RNA-binding activity of PRC2 has been attributed to both EZH2 (Kaneko et al., 2010; Zhao et al., 2010) and SUZ12 (Kanhere et al., 2010), and, in unbiased analyses, large portions of the transcriptome were reported to bind to PRC2 (Kaneko et al., 2013; Khalil et al., 2009; Zhao et al., 2010), raising the question of how specificity is achieved in vivo.

Here, we show that the PRC2 accessory subunit JARID2 binds to lncRNAs in vivo and in vitro and that its interaction with MEG3, an lncRNA encoded by the imprinted *DLK1–DIO3* locus, is necessary for proper recruitment and assembly of PRC2 at a subset of target genes in pluripotent stem cells.

RESULTS

JARID2 Binds to RNA In Vitro

Despite a requirement for JARID2 during development (Takeuchi et al., 1995) and its key role in PRC2 recruitment and function during differentiation of mouse ESCs (Li et al., 2010; Pasini et al., 2010; Peng et al., 2009; Shen et al., 2009), the mechanisms by which JARID2 is targeted to chromatin and orchestrates PRC2 function remain poorly understood. Given that several PcG and PcG-associated proteins interact with lncRNAs, which in some cases regulate their recruitment to chromatin (Kanhere et al., 2010; Rinn et al., 2007; Yap et al., 2010), and based on our own preliminary observations in vitro (Kaneko et al., 2010), we hypothesized that lncRNAs might also regulate the function of JARID2.

We previously mapped an RNA-binding region (RBR) of EZH2, a core component of PRC2, and found that phosphorylation of a threonine within that region stimulated binding to lncRNAs (Kaneko et al., 2010). We performed similar in vitro RNA-binding assays on JARID2 using a bait spanning nucleotides 1–333 of HOTAIR, a lncRNA that regulates PRC2 function (Tsai et al., 2010), and detected an affinity for RNA within an internal fragment of JARID2, but not in the N-terminal or C-terminal regions (Figures 1A and 1B). Further mapping experiments revealed that the deletion of residues 332–358 resulted in a severe decrease of RNA binding in vitro (Figure 1C; Figures S1A and S1B available online). The sequence spanning these residues is conserved strongly in vertebrates (Figure S1C), but only very weakly with *Drosophila* (Figure S1D), suggesting that JARID2 may have acquired an additional layer of regulation in vertebrates.

JARID2 binds to EZH2, the catalytic component of PRC2 (Margueron and Reinberg, 2011) and stimulates its histone methyltransferase activity (Li et al., 2010; Son et al., 2013). These functions require the same internal fragment that contains the RBR (Figure S1E) but can be uncoupled from the latter, given that deletion of residues 332–358 did not affect the ability of JARID2 to interact with PRC2 in vivo (Figure S1F) or to stimulate its enzymatic activity (Figure S1G), whereas the 349–574 fragment did not bind to RNA (Figure S1A) but retained the ability to interact with nucleosomes (Figure S1E) (Son et al., 2013).

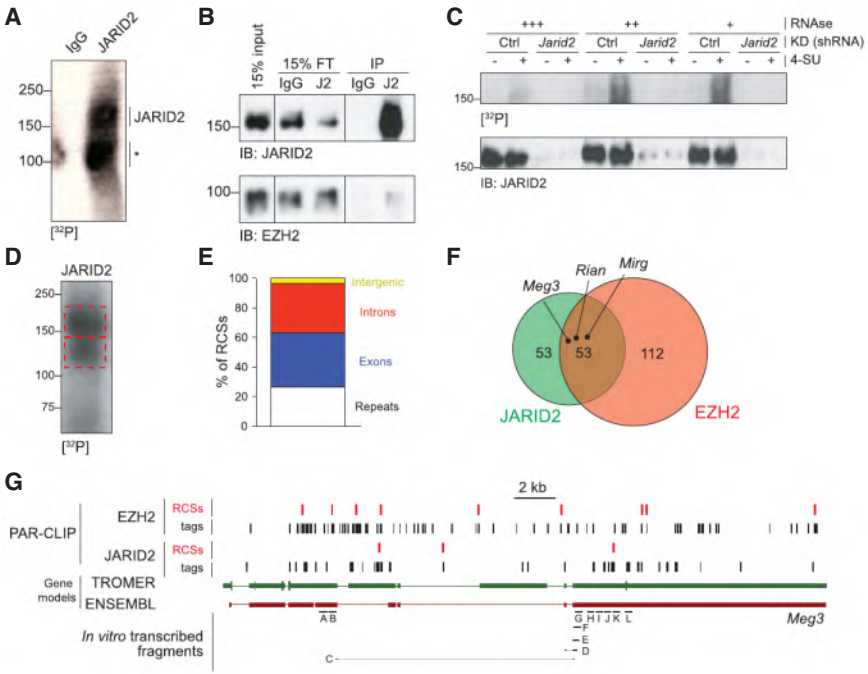


Figure 2. JARID2 and EZH2 Share Interacting lncRNAs In Vivo, Including Meg3

(A) PAR-CLIP with JARID2 antibodies or IgG in E14 ESC cells. The position of full-length JARID2 is indicated. The asterisk marks a presumed degradation product.

(B) Immunoblot on the same material utilized for the autoradiography in (A). J2, anti-JARID2 antibody.

(C) PAR-CLIP (top) and immunoblot for JARID2 (bottom) performed in cells pulsed (+) or not pulsed (−) with 4-SU and stably transfected with an shRNA against *Jarid2* or a control shRNA. Extracts were treated with increasing concentration of a cocktail of DNase-free RNase A and T1.

(D) PAR-CLIP-seq blot for JARID2.

(E) Distribution of JARID2 RCSs identified by PARalyzer in the genome. The stacked columns represent % of total RCSs. "Repeats" include all features listed in the RepMask database.

(F) Venn diagram of lncRNAs containing RCSs for JARID2, EZH2, or both. PAR-CLIP data for EZH2 were taken from GSE49433 (Kaneko et al., 2013).

(G) Genome browser view of JARID2 and EZH2 CLIP tags (black bars) or RCSs identified by PARalyzer (red bars) mapping to the Meg3 lncRNA. Gene models for *Meg3* according to both TROMER and ENSEMBL are shown. Meg3 fragments tested for in vitro binding are indicated at the bottom.

See also Figure S2 and Table S1.

Therefore, although partially overlapping regions of JARID2 are required for these various functions, the only activity that we could uniquely attribute to the 332–358 fragment is that of binding RNA in vitro. Henceforth, we will refer to these residues as the RBR of JARID2 and to the mutant protein lacking these residues as JARID2$_{\Delta RBR}$.

JARID2 and EZH2 Bind to lncRNAs In Vivo, Including Meg3

To determine whether JARID2 makes direct contacts with RNA in vivo, we utilized the photoactivatable-ribonucleoside-enhanced crosslinking and immunoprecipitation (PAR-CLIP) technique, which crosslinks RNA that has incorporated 4-thiouridine (4-SU) to proteins in vivo (Hafner et al., 2010). Consistent with our hypothesis, we detected a strong PAR-CLIP signal in immunoprecipitations (IPs) with our JARID2 antibody (Li et al., 2010) in extracts from embryonic day 14 (E14) ESCs (Figure 2A). These IPs were conducted in presence of 2% lauryldimethylbetaine, a zwitterionic detergent that almost completely abolished the PRC2-JARID2 interaction while preserving antibody reactivity (Figure 2B). Furthermore, the radioactive signal we observed must have originated from RNA crosslinked to JARID2, because it was dependent on the incorporation of 4-SU and was erased by treatment with increasing concentrations of RNase and by *Jarid2* knockdown (Figure 2C).

To identify the RNAs bound to JARID2 in vivo, we excised ^{32}P-labeled bands from the PAR-CLIP membranes (Figure 2D), eluted the crosslinked RNA, and sequenced it. In addition to the major, full-length band, we excised a faster migrating band that also reacted with JARID2 antibodies in WT but not in *Jarid2*$^{-/-}$ *cells* (data not shown). We obtained 90,000–200,000 unique CLIP tags in each replicate and analyzed their distribution using the PARalyzer software, which takes advantage of the T→C transitions caused by 4-SU crosslinking to discriminate signal from noise (Corcoran et al., 2011). PARalyzer identified 9,050 putative RNA-protein contact sites (RCSs) for JARID2, of which ~26% overlapped by more than 50% with repeats (Figure 2E) and were discarded. Among the 2,057 lncRNAs annotated in the mouse genome (ENSEMBL release 67), we identified 106 that contained at least one nonrepetitive RCS for JARID2 (Table S1). This bioinformatic pipeline applied to our previously generated EZH2 PAR-CLIP data (Kaneko et al., 2013) revealed that, in the same ESCs, EZH2 interacted with 165 lncRNAs, of which 53 were in common with JARID2 (Figure 2F; Table S1), including Meg3/Gtl2, Rian, and Mirg, three lncRNAs encoded within the imprinted *Dlk1-Dio3* locus.

We focused our attention on the lncRNA Meg3 (also known as Gtl2), because of previous reports linking it to pluripotency (Stadtfeld et al., 2010), imprinting (da Rocha et al., 2008), and PRC2 function (Zhao et al., 2010). Consistent with our PARalyzer analysis, several tags and RCSs from both JARID2 and EZH2 PAR-CLIP data mapped to *Meg3* (Figure 2G). Although some of the JARID2 CLIP tags mapped to a 5′ region annotated as an intron by ENSEMBL (red track), the existence of an exon in this region is supported by the TROMER database (Benson et al., 2004) (green track), and our own RNA-seq (data not shown). Another cluster of JARID2 CLIP tags mapped to the 3′ exon and accumulated in a region where PARalyzer identified an RCS (Figure 2G). Few CLIP tags mapped to highly expressed genes such as *Nanog* (Figure S2A) or *Gapdh* (Figure S2B), suggesting that the presence of *Meg3* tags reflected direct interactions in vivo.

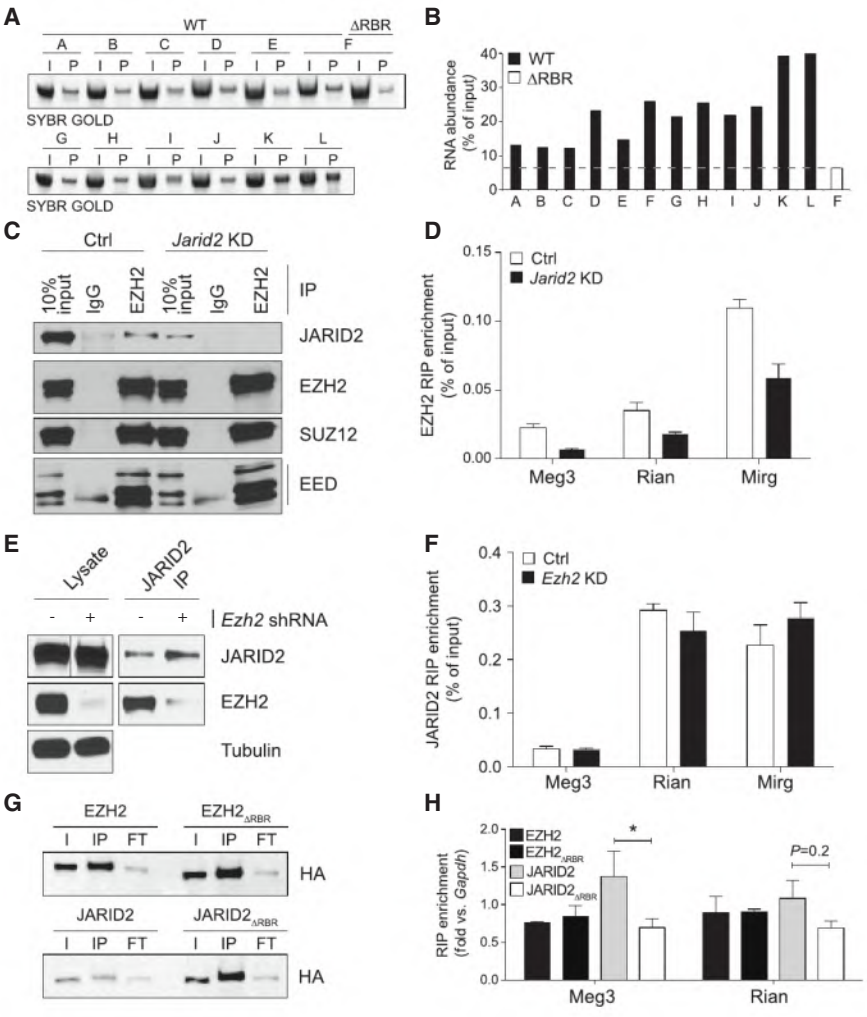


Figure 3. The JARID2 RBR Mediates Interactions of Meg3 with PRC2

(A) In vitro pull-down assay with different fragments of Meg3 using 4 μg of GST-JARID2 119–450 WT or ΔRBR.

(B) Quantification of bands shown in (A).

(C) RIPs for EZH2 were performed in E14 ESCs stably transfected with an shRNA against *Jarid2* (KD) or empty vector (ctrl). Coprecipitated proteins were revealed by western blot. The 10% input and IgG lanes are shown as controls.

(D) qRT-PCR on EZH2 RIPs from control E14 ESCs (white bars) or *Jarid2* knockdown E14 ESCs (black bars). Data are shown as percentage of RIP input. Bars represent the mean of four replicates + SD.

(E and F) As in (C) and (D) but *Ezh2* was knocked down and the RIP were performed with JARID2 antibodies.

(G) Western blots for HA RIPs from nuclear extracts of KH2 transiently transfected with N3-tagged EZH2, $EZH2_{\Delta RBR}$ (top), JARID2, and $JARID2_{\Delta RBR}$ (bottom). I, input; IP, HA immunoprecipitation; FT, flow-through.

(H) qRT-PCR normalized to *Gapdh* levels on HA RIPs described in (E). Bars indicate the mean of three biological replicates + SEM. *p < 0.05 by Mann-Whitney *U* test.

See also Figures S3 and S4.

The RBR of JARID2 Contributes to Meg3 Binding In Vitro and In Vivo

Having identified Meg3 as a JARID2-interacting lncRNA by CLIP, we next tested this interaction in vitro. We performed pull-down assays with in vitro-transcribed fragments of Meg3 (Figure 2G, bottom) obtained from a previously generated clone (Zhao et al., 2010). Although all tested fragments bound to a recombinant JARID2 fragment spanning the RBR (Figure 3A), the interaction was stronger for fragments originating from the 3′ end of the clone (Figure 3B), including one (fragment "K") that spanned the only RCS identified in this region (Figure 2G). We analyzed structural predictions for these fragments but found no obvious similarities among those that bound with higher affinity, except a trend for less stable structures (Figure S3). Importantly, a JARID2 fragment lacking the RBR displayed a pronounced reduction in binding affinity, at least toward the Meg3 fragment tested (fragment "F"; Figures 3A and 3B).

To validate these JARID2-RNA interactions with a technique more quantitative than PAR-CLIP, we resorted to native RNA immunoprecipitations (RIPs) followed by quantitative PCR (qPCR). Consistent with our PAR-CLIP results and previous reports (Zhao et al., 2010), Meg3 was enriched in PRC2 RIPs performed with EZH2 antibodies, and so were two other lncRNAs encoded within the same imprinted locus, Rian and Mirg (Figures 3C and 3D). However, when we depleted JARID2 by small hairpin RNA (shRNA)-mediated knockdown, we observed a considerable decrease in the amounts of these lncRNAs coprecipitating with PRC2 (Figures 3C and 3D), and similar results were obtained with SUZ12 antibodies (Figures S4A and S4B). Importantly, knockdown of *Ezh2* did not affect the ability of JARID2 to bind to Meg3, Rian, or Mirg (Figures 3E and 3F), suggesting that the JARID2-Meg3 interaction makes the largest contribution to the affinity of Meg3 for PRC2.

We then asked whether Meg3 bound to JARID2 via the RBR. To this end, we transiently expressed in mouse ESCs hemagglutinin (HA)-tagged EZH2, EZH2 lacking the previously identified RBR ($EZH2_{\Delta RBR}$) (Kaneko et al., 2010), JARID2, and $JARID2_{\Delta RBR}$ and performed HA RIPs followed by quantitative RT-PCRs (qRT-PCRs). Consistent with the results presented above, the interaction of Meg3 with JARID2 was significantly decreased when its RBR was removed, whereas EZH2 and $EZH2_{\Delta RBR}$ coprecipitated Meg3 with equal efficiencies (Figures 3G and 3H). We detected a similar trend for Rian, although in that case the difference did not reach statistical significance (Figure 3H). Similarly, human JARID2, but not human $JARID2_{\Delta RBR}$, bound to Meg3 in mouse ESCs (Figures S4C and S4D).

These data supported the conclusion that Meg3 interacts with PRC2 mainly through the RBR of JARID2, which led us to speculate that this lncRNA may participate in the function of JARID2 on chromatin.

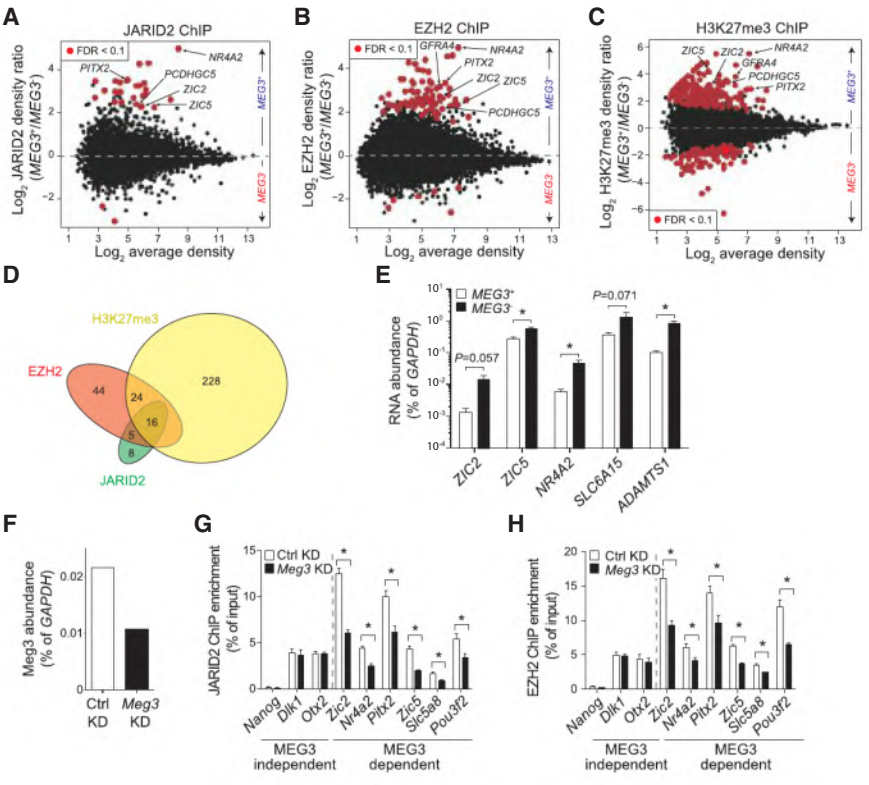


Figure 4. Decreased Occupancy of PRC2 at Some Chromatin Targets in *MEG3*$^{-}$ Cells

(A–C) MA plots for JARID2 (A), EZH2 (B), and H3K27me3 (C) occupancy, as determined by the normalized and input-corrected read densities in *MEG3*$^{+}$ (above dotted line) versus *MEG3*$^{-}$ (below dotted line) hiPSC lines. Each dot represents an ER in common between at least two hiPSC lines. DBRs with an FDR < 0.1 are displayed in red.

(D) Venn diagram for DBRs with FDR < 0.1.

(E) qRT-PCR analysis of PRC2 targets in *MEG3*$^{+}$ and *MEG3*$^{-}$ hiPSCs. Bars represent the mean RNA abundance (as percentage of *GAPDH*) in the five *MEG3*$^{+}$ and three *MEG3*$^{-}$ lines tested. *p < 0.05, as calculated by Mann-Whitney *U* test.

(F) qRT-PCR for *Meg3* 24 hr after transfection of KH2 ESCs with control (white bar) or *Meg3* siRNAs.

(G and H) ChIP-qPCR for JARID2 (G) or EZH2 (H) with primers mapping to PRC2 peaks near the indicated genes in KH2 ESCs treated with control (white bars) or *Meg3* (black bars) siRNAs. Bars represent the mean of three replicates + SEM. *p < 0.05 by Mann-Whitney *U* test.

See also Figures S5–S7 and Tables S2 and S3.

MEG3 Regulates PRC2 Occupancy In *Trans*

In light of the connection between MEG3 and pluripotency (Stadtfeld et al., 2010), we turned our attention to a set of eight human induced pluripotent stem cell (hiPSC) lines that differ greatly in their levels of *MEG3* expression (Nishino et al., 2011), thus offering a natural experimental system in which to study the effects of MEG3 on PRC2 function. We classified these lines into 5 *MEG3*$^{+}$ and 3 *MEG3*$^{-}$ (Figures S5A and S5B) and confirmed that *EZH2*, *SUZ12*, and *JARID2* were expressed at similar levels in *MEG3*$^{+}$ and *MEG3*$^{-}$ lines (Figures S5C–S5E). Importantly, pluripotent markers *OCT3/4* and *NANOG* were also expressed at comparable levels in all lines and at much higher levels than in differentiated cells, such as foreskin fibroblasts (Figures S5F and S5G).

Next, we performed chromatin immunoprecipitation (ChIP) followed by deep sequencing (ChIP-seq) for JARID2, EZH2, and H3K27me3, the product of PRC2 catalysis. We identified enriched regions (ERs) for all three features and compared normalized read densities in *MEG3*$^{+}$ versus *MEG3*$^{-}$ cells. The genome-wide analysis revealed 29, 89, and 268 differentially bound regions (DBRs) with a false discovery rate (FDR) < 0.1 in *MEG3*$^{+}$ versus *MEG3*$^{-}$ cells for JARID2, EZH2, and H3K27me3, respectively (Figures 4A–4C). At most of these *MEG3*-dependent DBRs, PRC2 exhibited stronger binding in the hiPSC lines that expressed *MEG3*, suggesting that the lncRNA had a stimulatory function (Figures 4A–4C; Figures S6A–S6C). Genes near the *MEG3*-dependent DBRs were enriched for functional terms related to the regulation of transcription during embryonic development and differentiation (Table S2), even taking into account the enrichment of developmental and transcription-related terms in the background population comprising all JARID2, EZH2, and H3K27me3 ERs (Table S3).

Of the 29 *MEG3*-dependent DBRs for JARID2, 21 overlapped with EZH2 DBRs (Figure 4D), and of these 16 also overlapped with H3K27me3 DBRs, a fraction much larger than expected by chance alone (p value < 10^{-20}, hypergeometric distribution). Among the regions with lower JARID2 occupancy in the absence of *MEG3* were the loci encoding the transcription factors *ZIC5*, *NR4A2*, *ZIC2*, and *PITX*, as well as the neuronal gene *PCDHGC5* (Figure S6D and data not shown), all of which function in differentiating or differentiated cells and must therefore be silenced in pluripotent stem cells. In all loci tested, the *MEG3*-dependent loss of PRC2 targeting resulted in transcriptional derepression (Figure 4E).

It has been suggested that, in mouse ESCs, Meg3 exerts a *cis*-repressive effect on the adjacent *Dlk1* gene by recruiting PRC2 (Zhao et al., 2010). To determine whether a similar mechanism was conserved in hiPSCs, we analyzed the *DLK1* promoter in *MEG3*$^{+}$ versus *MEG3*$^{-}$ hiPSCs. Unexpectedly, we observed no differences in JARID2 or PRC2 occupancy (data not shown) and no evidence of a reciprocal correlation in the levels of *MEG3* and *DLK1* RNA in these cells (Figure S5H) or with the protein-coding RNA at the other extremity of the imprinted locus, *DIO3* (Figure S5J).

Because the *DLK1-DIO3* locus encodes multiple ncRNAs (*MEG3*, *RIAN*, and *MIRG*), all repressed in *MEG3*$^{-}$ hiPSCs, we examined the effect of altering *MEG3* levels alone on PRC2 localization. We overexpressed *MEG3* (or GFP RNA as a control) in *MEG3*$^{-}$ hiPSCs and analyzed the distribution of JARID2 and EZH2 by ChIP-seq. A caveat of this experiment is that *MEG3* was expressed at levels ~10 times higher than in the average *MEG3*$^{+}$ line (Figure S7A); nonetheless, we observed increased

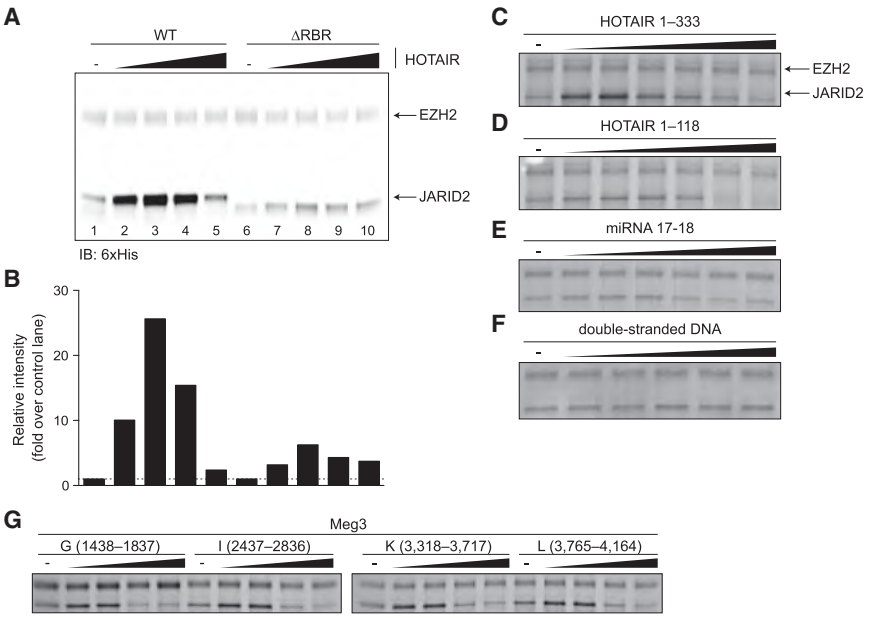


Figure 5. HOTAIR Stimulates EZH2–JARID2 Interactions via the JARID2 RBR

(A) FLAG-6xHis-tagged recombinant EZH2 (20 pmol) was incubated with 6xHis-tagged $JARID2_{119–450}$ WT or ΔRBR (40 pmol) in presence of increasing amounts of $HOTAIR_{1–333}$ (0–12 pmol). Pull-down was performed with anti-FLAG beads and proteins revealed by 6xHis immunoblot.

(B) Densitometric quantification of the signal for JARID2 shown in (A). WT and ΔRBR lanes were normalized each to lane 1 and lane 6 (no HOTAIR control), respectively.

(C–F) In vitro interaction stimulation assays with recombinant 6xHis-tagged $JARID2_{119–574}$ and FLAG-6xHis-tagged full-length EZH2 in presence of increasing concentrations of $HOTAIR_{1–333}$ (C), a smaller 5′ truncation of HOTAIR (D), a microRNA (E), or double-stranded DNA (F). Pull down was performed with anti-FLAG beads and proteins stained with Coomassie blue.

(G) Same as (C)–(F) using Meg3 fragments.

densities of JARID2 at some *MEG3*-dependent DBRs, but not at control targets (Figure S7B). However, we did not detect recovery of JARID2 occupancy at all JARID2 DBRs (Figure S7C), suggesting that the regulation afforded by endogenous MEG3 could not be fully recapitulated by providing large amounts of the lncRNA in *trans*. Interestingly, PRC2 was preferentially depleted from the EZH2 DBRs upon overexpression of *MEG3* (Figure S7D). Although this depletion of EZH2 was unexpected, the fact that the EZH2 DBRs were preferentially depleted by the manipulation of MEG3 levels is consistent with the idea that occupancy at these sites is selectively regulated by this lncRNA.

To further verify that the observed changes in PRC2 occupancy were directly related to the changes in MEG3 levels, we performed transient knockdown of the orthologous lncRNA in mouse ESCs. Despite only partial knockdown efficiency (~50%; Figure 4F), six out of nine DBRs tested lost JARID2 and EZH2 after Meg3 depletion (Figures 4G and 4H), suggesting that the regulation of PRC2 by Meg3 is conserved between human and mouse. PRC2 occupancy at *Plek2*, *Mbnl3*, and *Cdx4* was not affected by *Meg3* knockdown (data not shown), despite the fact that orthologous loci were among the MEG3-dependent DBRs in hiPSCs. However, it is not surprising that subtle differences in gene regulation would exist across this species barrier, especially given that mouse and human stem cells are not equivalent (Ginis et al., 2004).

Together, our ChIP experiments in hiPSCs and mouse ESCs support the conclusion that MEG3 acts in *trans* on PRC2 and JARID2 by facilitating their recruitment to a subset of target genes.

RNA Facilitates JARID2-PRC2 Interactions

Given that loss of MEG3 caused defects in PRC2 recruitment and that several lncRNAs crosslinked to both JARID2 and EZH2 in vivo, we hypothesized that RNA-mediated scaffolding could stabilize JARID2-PRC2 interactions. To test this hypothesis in vitro, we first examined the lncRNA HOTAIR, which functions as a scaffold between the LSD1/CoREST/REST complex and PRC2 (Tsai et al., 2010). Low amounts of HOTAIR stimulated the interaction between recombinant EZH2 and JARID2 fragments in vitro (Figures 5A and 5B, compare lanes 1–3), whereas higher concentrations of the RNA resulted in a return to baseline interaction levels, suggesting the possibility of squelching (Figures 5A and 5B, compare lanes 3–5). This stimulation was considerably reduced for JARID2 fragments that lacked the RBR (Figures 5A and 5B, lanes 6–10) but retained the ability to interact with EZH2 (Figure 5A, compare lanes 1 and 6) or when we replaced the HOTAIR lncRNA fragment with a shorter version, a control pre-microRNA that forms two stem-loops, or double-stranded DNA containing the HOTAIR lncRNA sequence (Figures 5C–5F). Stimulation of binding was also observed when we incubated recombinant EZH2 and JARID2 with those Meg3 fragments that displayed maximum affinity in the in vitro pull-down assay (Figure 5G). Therefore, stimulation of JARID2-EZH2 interactions may be a general mechanism of action for lncRNAs that bind to these proteins.

The JARID2 RBR Stimulates PRC2 Assembly on Chromatin

Having discovered that lncRNAs stimulate JARID2-PRC2 interactions in vivo (Figure 4) and in vitro (Figure 5) and knowing that the RBR was required for the latter (Figures 5A and 5B), we asked whether it was also required for the former. To this end, we overexpressed JARID2 or $JARID2_{\Delta RBR}$ in human foreskin fibroblasts, which express high levels of *HOTAIR* (Rinn et al., 2007), and measured its accumulation at known genomic targets by ChIP-qPCR. Wild-type (WT) and mutant JARID2 were expressed at comparable levels in lentivirally transduced fibroblasts (Figure 6A, top) and did not affect EZH2 levels (Figure 6A, middle). Upon JARID2 overexpression, we observed increased accumulation

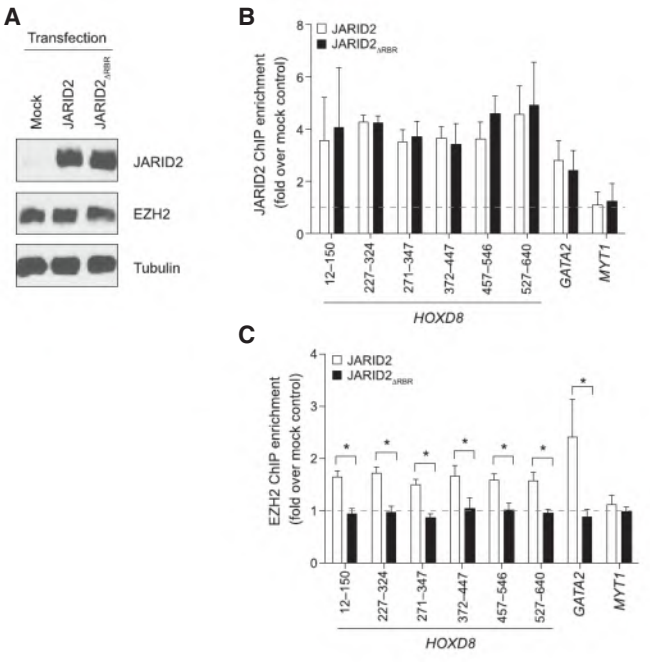


Figure 6. JARID2 Recruits EZH2 to Chromatin in an RBR-Dependent Manner

(A) Western blots for foreskin fibroblasts transduced with lentiviruses expressing JARID2, JARID2$_{\Delta RBR}$, or mock transduced.

(B and C) ChIP-qPCR with antibodies against JARID2 (B) or EZH2 (C) at known PRC2 chromatin targets in foreskin fibroblasts transduced as in (A). Several primer sets are shown for *HOXD8*. The primers for *MYT1* were designed at a distal location, devoid of PRC2, and serve as a negative control. The ChIP enrichment is normalized against that obtained with the same antibodies in mock-transfected control cells (dotted line). Bars represent mean of four technical replicates + SD (B) or three biological replicates + SEM (C). *$p < 0.05$ by Mann-Whitney *U* test.

of the protein at known chromatin targets, such as *HOXD8* and *GATA2*, compared with controls (Figure 6B). JARID2$_{\Delta RBR}$ accumulated at these targets with the same efficiency as the WT protein but, unlike the WT, it was incapable to recruit additional EZH2, which remained at the same levels as observed in the mock-transfected cells (Figure 6C). Given that these targets were already silent in the mock-transfected controls, we did not attempt to detect further repression at the transcriptional level.

These results suggest that RNA-protein interactions via the JARID2 RBR contribute to the recruitment and assembly of PRC2 on chromatin. Interestingly, depletion of HOTAIR in foreskin fibroblasts also impairs PRC2 recruitment at these loci (Rinn et al., 2007; Tsai et al., 2010), suggesting that it may act through interactions with the JARID2 RBR.

The fact that JARID2$_{\Delta RBR}$ is incapable of recruiting EZH2 to target sites is compatible with a model by which lncRNAs modulate JARID2-PRC2 interactions and orchestrate the distribution and activity of PRC2 on chromatin.

DISCUSSION

The results presented above allow us to add JARID2 to the growing list of chromatin-associated proteins that interact with lncRNAs and offer further support to the hypothesis that lncRNAs are a component of the *Polycomb* axis in mammals.

We mapped the JARID2 RBR to an N-terminal fragment that contains no annotated features or domains, despite the fact that, in addition to RNA binding, it is also responsible for interactions with SUZ12, EZH2, nucleosomes, and PRC2 stimulation (Kim et al., 2003; Li et al., 2010; Pasini et al., 2010; Son et al., 2013). Although we assigned these biochemical activities to distinct protein fragments (Figure S1), they do map to adjacent regions, and it is tempting to speculate that this vicinity might reflect a functional crosstalk, by which, for example, RNA binding could stimulate nucleosomal binding and contribute to PRC2 regulation.

Among the lncRNAs identified by PAR-CLIP, we focused our functional analysis on Meg3 because of its connections with ESC pluripotency, imprinting, and PRC2 function (da Rocha et al., 2008; Stadtfeld et al., 2010; Zhao et al., 2010). We demonstrated JARID2-Meg3 interactions using RIP-qPCR, PAR-CLIP, and in vitro pull-down assays. Importantly, we also found RCSs for EZH2 within *Meg3* in our previously published EZH2 PAR-CLIP data set (Kaneko et al., 2013) (Figures 2F and 2G). This is consistent with earlier results obtained by native and UV-cross-linked RIP (Zhao et al., 2010) and supports a model by which Meg3 contacts both JARID2 and EZH2 and stimulates their interaction. Without JARID2 the interaction between Meg3 and PRC2 is much weaker (Figure 3D), suggesting that of the two contact points, the one on JARID2 makes the larger contribution to the affinity for Meg3. We note that by using *Ezh2*$^{-/-}$ cells as a control, Zhao et al. could not have detected this requirement for JARID2, because in absence of EZH2 the IP performed with an anti-EZH2 antibody would not have recovered JARID2.

We envision a model in which some lncRNAs function as scaffold to stimulate assembly of PRC2 at JARID2 target sites (Figure 7A). In addition, the existence of DBRs that lose JARID2 occupancy in absence of MEG3 (Figure 4A) suggests that, at certain sites, lncRNAs might also be required for the initial recruitment of JARID2 (Figure 7B). In both scenarios, the net result of lncRNA action is to increase PRC2 occupancy and H3K27me3 deposition (Figures 7A and 7B). We have demonstrated this model using MEG3 and genomic targets in hiPSCs and mouse ESCs (Figure 4); however, the recovery of other lncRNAs cross-linked to both JARID2 and EZH2 by PAR-CLIP-seq (Figure 2F; Table S1) allows us to speculate that this mode of action might not be limited to MEG3. In fact, the JARID2$_{\Delta RBR}$ mutant does not recruit PRC2 to the *HOX* locus in foreskin fibroblasts, despite the fact that this locus is not MEG3 dependent. We propose that other lncRNAs function as scaffold for JARID2-PRC2 interactions in this setting and we note that foreskin fibroblasts express high levels of *HOTAIR* (Rinn et al., 2007), which is also able to stimulate JARID2-EZH2 interactions in vitro (Figure 5).

Not all lncRNAs bind equally to JARID2 and EZH2. Our reanalysis of EZH2 PAR-CLIP tags (Kaneko et al., 2013) identified a number of lncRNAs not shared with JARID2 (Figure 2F; Table S1) that, likely, regulate PRC2 function in JARID2-independent ways. In addition to lncRNAs, EZH2 binds to a variety of coding transcripts in vivo and in vitro with high affinity but seemingly low specificity (Davidovich et al., 2013; Kaneko et al., 2013). Those interactions appear to constitute a distinct regulatory

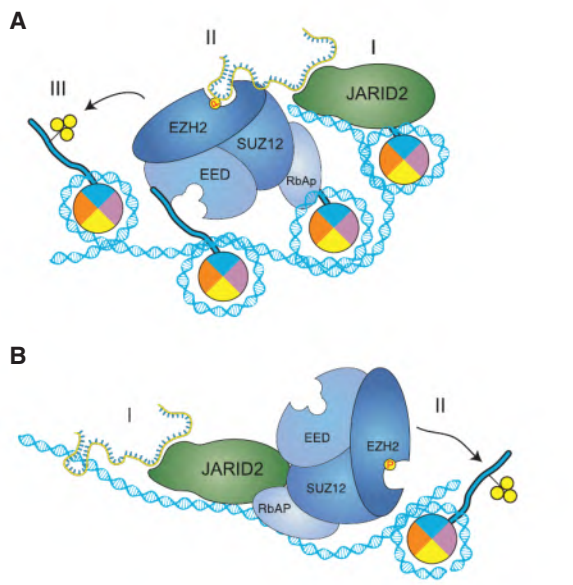


Figure 7. Proposed Model for the Interplay of lncRNAs, JARID2, and PRC2

(A) At some target genes, the presence of JARID2 by itself (I) is not sufficient for maximum PRC2 recruitment, which requires scaffolding by lncRNAs (II). The presence of both JARID2 and lncRNAs stimulates further recruitment and assembly of PRC2 on chromatin, resulting in increased H3K27me3 (III). The structure of lncRNAs bound to JARID2 (and PRC2) remains to be elucidated, and the one shown here is only for the purpose of illustration.

(B) In some cases, lncRNAs might contribute to the initial recruitment of JARID2 to chromatin (I). Because JARID2 also binds PRC2 via protein-protein interactions, this results in increased PRC2 recruitment and H3K27 methylation (II).

mechanism from the one described here, as they occur mostly at promoters of transcribed genes, which have low PRC2 occupancy (Davidovich et al., 2013; Kaneko et al., 2013) and are largely devoid of JARID2 (data not shown). However, we cannot exclude that nascent RNAs could compete with lncRNAs for binding to EZH2, which might help explain their apparent inhibitory function.

Our gain-of-function experiments in hiPSCs confirmed that at least some JARID2 DBRs identified in *MEG3*$^{+}$ versus *MEG3*$^{-}$ hiPSCs can be rescued by supplying MEG3 lncRNA in *trans* to otherwise *MEG3*-negative hiPSCs (Figure S7B). The ectopic expression of MEG3 lncRNAs caused EZH2 depletion rather than increased occupancy at the previously identified MEG3-dependent DBRs (Figure S7D), which was in contrast to our expectations. However, considering that lentiviral transduction resulted in 10-fold higher *MEG3* levels compared to endogenous *MEG3* (Figure S7A), we speculate that such an excess of RNA molecules acted in a dominant negative fashion through squelching, as seen in vitro (Figure 5). It is also possible that the random integration of the lentivirally encoded *MEG3* at ectopical sites within the genome may explain its unphysiological behavior in these experiments, given that the genomic location of lncRNA genes appears to play a pivotal role in their function (Ulitsky et al., 2011). Nonetheless, the fact that MEG3-dependent DBRs were selectively affected confirmed that occupancy at these targets is indeed regulated by MEG3, which was further supported by the results of transient *Meg3* knockdown in mouse ESCs (Figures 4G and 4H).

Genomic imprinting of *MEG3* is unstable in human ESCs, and several hiPSC lines, regardless of their parental cell type, maintain repression at this locus even after continuous passaging (Nishino et al., 2011). This is likely mediated by aberrant DNA hypermethylation (Nishino et al., 2011) (data not shown). Therefore, it is possible that current reprogramming protocols fail to set the appropriate epigenetic state at the *DLK1-DIO3* locus, which is an important consideration given that improper regulation of this imprinted regions leads to developmental abnormalities in mice (Stadtfeld et al., 2010; Takahashi et al., 2009) and humans (Kagami et al., 2008). The mechanistic details of the epigenetic alteration at this imprinted locus during reprogramming remains elusive, but our data suggest that MEG3 interactions with PRC2 might play an important role.

Although our findings suggest a molecular mechanism by which MEG3 contributes to JARID2 and PRC2 function, further investigations are required to address (1) whether and how this mechanism extends to the other lncRNAs identified by PAR-CLIP, and (2) whether and how lncRNAs regulate de novo recruitment of their interacting proteins to distant loci. Our in vitro binding analyses of Meg3 lncRNA suggest that JARID2 does not bind to RNA in a sequence-specific manner, consistent with the fact that the primary sequence of lncRNAs is not well conserved (Ulitsky et al., 2011).

In conclusion, we have demonstrated that JARID2, an essential regulatory component of PRC2 in pluripotent stem cells, contains an RNA-binding region that mediates, at least in part, its interaction with the imprinted lncRNA MEG3. This—and possibly other—RNA interaction contributes to proper recruitment and assembly of PRC2 on target genes and likely plays an important role in orchestrating the epigenetic regulation of gene expression that accompanies the transition from stem cell pluripotency to differentiation.

EXPERIMENTAL PROCEDURES

For information about antibodies, oligonucleotides, and plasmids see Tables S4–S6 and the Supplemental Experimental Procedures.

Cells

HiPSCs were cultured as described previously (Nishino et al., 2011). Cells were harvested at passage 34 (UtE-iPS-4), 41 (UtE-iPS-11), 45 (UtE-iPS-7), 41 (AM-iPS-6), 50 (AM-iPS-8), 39 (UtE-iPS-6), 41 (MRC5-iPS-25), and 90 (Edom-iPS-2). Human foreskin fibroblasts (#PCS-201-010, lot# 58490326; ATCC) were cultured in fibroblast basal medium (ATCC) plus fibroblast growth kit-low serum (ATCC). KH2 ESCs expressing the reverse tetracycline-controlled transactivator (rtTA) (Hochedlinger et al., 2005) were maintained in standard mouse ESC (mESC) culture conditions. KH2 lines expressing *Jarid2* and *Jarid2*$_{\Delta RBR}$ were generated by transfection of the relevant pINTA-N3 construct and selecting with 50 μg/ml Zeocin (Invitrogen). Transgene expression was induced with doxycycline for 24 hr. E14Tg2A.4 mESC lines (E14 mESC) and HEK293 and HEK293T cells were cultured as described previously (Kaneko et al., 2010). *Jarid2* knockdown mESCs were described previously (Li et al., 2010).

In Vitro Binding Assays

Protein purification, synthesis of HOTAIR lncRNA (1–333), and biotinylated RNA pull-down assays were described previously (Kaneko et al., 2010). All

RNA fragments were generated by in vitro transcription; see the Supplemental Information for details.

LncRNA-mediated stimulation of EZH2–JARID2 interactions was assayed by incubating FLAG- and 6xHis-tagged EZH2 (20 pmol) with increasing amounts (0–12 pmol) of lncRNAs in 100 μl of binding buffer (50 mM Tris-HCl, pH 7.9; 100 mM KCl; 0.1% NP-40) for 30 min at 25°C. 6xHis-tagged truncated JARID2 (40 pmol) was added to the reaction and incubated for 30 min at 25°C. Complexes were purified using FLAG-M2 affinity gel (Sigma), after washing with binding buffer.

For JARID2–Meg3 in vitro binding assays, 4 μg of GST-JARID2$_{119–450}$ were incubated with a series of Meg3 fragments (1 μg, 400 nt each; see Table S7). Bound RNAs were purified with glutathione beads, resolved with 7M urea gels, stained with SYBR-gold (Invitrogen), and quantified with an ImageQuant LAS4000 (GE Healthcare Life Sciences).

Knockdowns

For conditional knockdown of *Ezh2* in E14 mESCs, we generated stable clones with an integration of pTRIPZ lentiviral inducible shRNAmir targeting human and mouse *Ezh2* (#RHS4696-99635303; Open Biosystems). Selection was done by puromycin (1 μg/ml) and clones were screened by red fluorescent protein (RFP) expression after 3–4 days of doxycycline (1 μg/ml) induction.

For transient knockdown of *Meg3*, we tested four siRNAs from QIAGEN (SI05169486, SI05169710, SI01060129, and SI01060136) and used the siRNA resulting in the most efficient knockdown (SI05169486) (see Supplemental Information).

ChIP

ChIP from hiPSCs, foreskin fibroblast, and mESCs was performed as described (Kaneko et al., 2007), with minor modification, and libraries were constructed as described (Gao et al., 2012). Briefly, cells were crosslinked with 1% formaldehyde for 10 min and sonicated in ChIP buffer (50 mM Tris-HCl [pH 7.9], 150 mM NaCl, 1% Triton X-100, 0.5% NP-40, 5 mM EDTA [pH 8.0], 1 mM phenylmethanesulfonylfluoride, and protease inhibitors) with a Diagenode Bioruptor. Incubations with antibodies were carried out in an ultrasonic water bath for 30 min at 4°C. Samples were decrosslinked at 65°C for ~16 hr for library construction or 95°C 10 min for ChIP-qPCR. See the Supplemental Information for more details.

RNA Immunoprecipitation

For Figures 3G and 3H, nuclear extracts were obtained using an established protocol (Dignam et al., 1983) with minor modifications to minimize RNase activity, lysates were diluted in RIP buffer (20 mM Tris [pH 7.9]$_{4°C}$, 200 mM KCl, 0.05% IGEPAL CA-630, 10 mM EDTA), cleared by centrifugation at 20,000 × *g* for 10 min, and incubated with depleting amounts of antibody for 3 hr at 4°C. Immunocomplexes were recovered with protein G-coupled dynabeads (Invitrogen) for 1 hr at 4°C. Beads were washed in RIP-W buffer (20 mM Tris [pH 7.9]$_{4°C}$, 200 mM KCl, 0.05% IGEPAL CA-630, 1 mM $MgCl_2$) twice and incubated with 2 U TURBO DNase (Ambion) in 20 μl RIP-W buffer for 10 min at room temperature to avoid DNA bridging artifacts. After two additional washes, RNA was eluted and purified with TRIzol (Invitrogen).

For Figures 2C–2F, RIPs were performed on whole-cell lysates, as described before (Kaneko et al., 2010) (see also Supplemental Information).

ChIP-Seq Analysis

Sequenced reads from ChIP-seq experiments were mapped with BOWTIE using parameters -v2 -m4–best (Langmead et al., 2009). Normalized genome-wide read densities were computed and visualized on the UCSC genome browser. Bound regions (ERs) were identified using MACS 2.09 (Zhang et al., 2008) and default parameters. ERs were associated to gene targets using ChIPpeakAnno and ENSEMBL annotation 67.

PAR-CLIP

ESCs were pulsed with 100 μM 4-SU (Sigma) for 16–24 hr and crosslinked with 400 mJ/cm^2 UVA (365 nm) using a Stratalinker UV crosslinker (Stratagene). Cells were lysed for 10 min at 37°C in CLIP buffer (20 mM HEPES [pH 7.4], 5 mM EDTA, 150 mM NaCl, 2% lauryldimethylbetaine) with protease inhibitors, 20 U/ml Turbo DNase (Life Technologies), and 200 U/ml murine RNase inhibitor (New England Biolabs). IPs were carried out in CLIP buffer for 1 hr at 4°C. When necessary, extracts were treated with RNase A + T1 cocktail (Ambion) for 5′ at 37°C. Immunocomplexes were recovered with protein G-coupled dynabeads for 45 min at 4°C. DNA was removed with Turbo DNase (2 U in 20 μl). Crosslinked RNA was labeled by incubations with 5U Antarctic phosphatase and 5U T4 PNK (both from New England Biolabs) in presence of 10 μCi [γ-^{32}P] ATP (PerkinElmer, MA). Labeled material was resolved on 8% bis-tris gels, transferred to nitrocellulose, and exposed to autoradiography films for 1–24 hr.

For PAR-CLIP-seq, 100 pmol of a 3′-blocked DNA adaptor was ligated to the RNA after dephosphorylation and before 5′ labeling by incubating the beads with T4 RNA ligase 1 (New England Biolabs) for 1 hr at 25°C. After autoradiography, bands were excised and the RNA eluted with proteinase K for 30 minutes at 37°C and proteinase K in 3.5M urea for 30 minutes at 55°C. Custom 5′ adapters were ligated, and the products were size-selected on polyacrylamide or agarose gels, amplified, and sequenced on an Illumina HiSeq 2000.

For the analysis, adapter sequences were removed and reads < 17 nt discarded. The remaining reads were mapped to the mm9 genome using BOWTIE (Langmead et al., 2009), allowing two mismatches and removing duplicates. RCSs were identified with PARalyzer (Corcoran et al., 2011) requiring at least two T→C conversions per RCS. For Figure 2F, RCSs were assigned to lncRNAs in ENSEMBL 67 when they overlapped anywhere within the gene body to account for imprecisions in the annotation of lncRNAs.

RNA Structural Predictions

Structural predictions and minimum free-energy calculations shown in Figure S3 were performed with the Vienna RNA Websuite using default settings (Gruber et al., 2008).

ACCESSION NUMBERS

The ChIP and CLIP sequences reported in this paper have been deposited to the Gene Expression Omnibus (GEO) with the accession number GSE48518. EZH2 PAR-CLIP-seq data used for comparative analyses were taken from GEO accession number GSE49433 (Kaneko et al., 2013).

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, seven figures, and seven tables and can be found with this article online at http://dx.doi.org/10.1016/j.molcel.2013.11.012.

ACKNOWLEDGMENTS

We thank the Genome Technology Center at NYU for help with sequencing, John Rinn and Matthias Stadtfeld for comments on the manuscript, and Varun Narendra for bioinformatic analyses. This work was supported by grants from the National Institutes of Health (GM-64844 and R37-37120) and the Howard Hughes Medical Institute (to D.R.). R.B. was supported by a Helen Hay Whitney Foundation postdoctoral fellowship and by the Helen L. and Martin S. Kimmel Center for Stem Cell Biology postdoctoral fellow award. R.S.M. was supported by a Ph.D. and an international research stay fellowship from CONACyT (213029).

Received: July 10, 2013
Revised: October 2, 2013
Accepted: November 21, 2013
Published: December 26, 2013

REFERENCES

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., and Wheeler, D.L. (2004). GenBank: update. Nucleic Acids Res. *32* (Database issue), D23–D26.

Bonasio, R., Tu, S., and Reinberg, D. (2010). Molecular signals of epigenetic states. Science *330*, 612–616.

Boulay, G., Dubuissez, M., Van Rechem, C., Forget, A., Helin, K., Ayrault, O., and Leprince, D. (2012). Hypermethylated in cancer 1 (HIC1) recruits polycomb repressive complex 2 (PRC2) to a subset of its target genes through interaction with human polycomb-like (hPCL) proteins. J. Biol. Chem. *287*, 10509–10524.

Brockdorff, N., Ashworth, A., Kay, G.F., McCabe, V.M., Norris, D.P., Cooper, P.J., Swift, S., and Rastan, S. (1992). The product of the mouse Xist gene is a 15 kb inactive X-specific transcript containing no conserved ORF and located in the nucleus. Cell *71*, 515–526.

Brown, C.J., Hendrich, B.D., Rupert, J.L., Lafrenière, R.G., Xing, Y., Lawrence, J., and Willard, H.F. (1992). The human XIST gene: analysis of a 17 kb inactive X-specific RNA that contains conserved repeats and is highly localized within the nucleus. Cell *71*, 527–542.

Corcoran, D.L., Georgiev, S., Mukherjee, N., Gottwein, E., Skalsky, R.L., Keene, J.D., and Ohler, U. (2011). PARalyzer: definition of RNA binding sites from PAR-CLIP short-read sequence data. Genome Biol. *12*, R79.

da Rocha, S.T., Edwards, C.A., Ito, M., Ogata, T., and Ferguson-Smith, A.C. (2008). Genomic imprinting at the mammalian Dlk1-Dio3 domain. Trends Genet. *24*, 306–316.

Davidovich, C., Zheng, L., Goodrich, K.J., and Cech, T.R. (2013). Promiscuous RNA binding by Polycomb repressive complex 2. Nat. Struct. Mol. Biol. *20*, 1250–1257.

Dignam, J.D., Lebovitz, R.M., and Roeder, R.G. (1983). Accurate transcription initiation by RNA polymerase II in a soluble extract from isolated mammalian nuclei. Nucleic Acids Res. *11*, 1475–1489.

Gao, Z., Zhang, J., Bonasio, R., Strino, F., Sawai, A., Parisi, F., Kluger, Y., and Reinberg, D. (2012). PCGF homologs, CBX proteins, and RYBP define functionally distinct PRC1 family complexes. Mol. Cell *45*, 344–356.

Ginis, I., Luo, Y., Miura, T., Thies, S., Brandenberger, R., Gerecht-Nir, S., Amit, M., Hoke, A., Carpenter, M.K., Itskovitz-Eldor, J., and Rao, M.S. (2004). Differences between human and mouse embryonic stem cells. Dev. Biol. *269*, 360–380.

Gruber, A.R., Lorenz, R., Bernhart, S.H., Neuböck, R., and Hofacker, I.L. (2008). The Vienna RNA websuite. Nucleic Acids Res. *36* (Web Server issue), W70–W74.

Guttman, M., Amit, I., Garber, M., French, C., Lin, M.F., Feldser, D., Huarte, M., Zuk, O., Carey, B.W., Cassady, J.P., et al. (2009). Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature *458*, 223–227.

Hafner, M., Landthaler, M., Burger, L., Khorshid, M., Hausser, J., Berninger, P., Rothballer, A., Ascano, M., Jr., Jungkamp, A.C., Munschauer, M., et al. (2010). Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. Cell *141*, 129–141.

Hochedlinger, K., Yamada, Y., Beard, C., and Jaenisch, R. (2005). Ectopic expression of Oct-4 blocks progenitor-cell differentiation and causes dysplasia in epithelial tissues. Cell *121*, 465–477.

John, R.M., and Surani, M.A. (1996). Imprinted genes and regulation of gene expression by epigenetic inheritance. Curr. Opin. Cell Biol. *8*, 348–353.

Kagami, M., Sekita, Y., Nishimura, G., Irie, M., Kato, F., Okada, M., Yamamori, S., Kishimoto, H., Nakayama, M., Tanaka, Y., et al. (2008). Deletions and epimutations affecting the human 14q32.2 imprinted region in individuals with paternal and maternal upd(14)-like phenotypes. Nat. Genet. *40*, 237–242.

Kaneko, S., Rozenblatt-Rosen, O., Meyerson, M., and Manley, J.L. (2007). The multifunctional protein p54nrb/PSF recruits the exonuclease XRN2 to facilitate pre-mRNA 3′ processing and transcription termination. Genes Dev. *21*, 1779–1789.

Kaneko, S., Li, G., Son, J., Xu, C.F., Margueron, R., Neubert, T.A., and Reinberg, D. (2010). Phosphorylation of the PRC2 component Ezh2 is cell cycle-regulated and up-regulates its binding to ncRNA. Genes Dev. *24*, 2615–2620.

Kaneko, S., Son, J., Shen, S.S., Reinberg, D., and Bonasio, R. (2013). PRC2 binds active promoters and contacts nascent RNAs in embryonic stem cells. Nat. Struct. Mol. Biol. *20*, 1258–1264.

Kanhere, A., Viiri, K., Araújo, C.C., Rasaiyaah, J., Bouwman, R.D., Whyte, W.A., Pereira, C.F., Brookes, E., Walker, K., Bell, G.W., et al. (2010). Short RNAs are transcribed from repressed polycomb target genes and interact with polycomb repressive complex-2. Mol. Cell *38*, 675–688.

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

Kim, T.G., Kraus, J.C., Chen, J., and Lee, Y. (2003). JUMONJI, a critical factor for cardiac development, functions as a transcriptional repressor. J. Biol. Chem. *278*, 42247–42255.

Kim, H., Kang, K., and Kim, J. (2009). AEBP2 as a potential targeting protein for Polycomb Repression Complex PRC2. Nucleic Acids Res. *37*, 2940–2950.

Landeira, D., Sauer, S., Poot, R., Dvorkina, M., Mazzarella, L., Jørgensen, H.F., Pereira, C.F., Leleu, M., Piccolo, F.M., Spivakov, M., et al. (2010). 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*, 618–624.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. *10*, R25.

Lanzuolo, C., and Orlando, V. (2012). Memories from the polycomb group proteins. Annu. Rev. Genet. *46*, 561–589.

Li, G., Margueron, R., Ku, M., Chambon, P., Bernstein, B.E., and Reinberg, D. (2010). Jarid2 and PRC2, partners in regulating gene expression. Genes Dev. *24*, 368–380.

Margueron, R., and Reinberg, D. (2011). The Polycomb complex PRC2 and its mark in life. Nature *469*, 343–349.

Margueron, R., Justin, N., Ohno, K., Sharpe, M.L., Son, J., Drury, W.J., 3rd, Voigt, P., Martin, S.R., Taylor, W.R., De Marco, V., et al. (2009). Role of the polycomb protein EED in the propagation of repressive histone marks. Nature *461*, 762–767.

Nishino, K., Toyoda, M., Yamazaki-Inoue, M., Fukawatase, Y., Chikazawa, E., Sakaguchi, H., Akutsu, H., and Umezawa, A. (2011). DNA methylation dynamics in human induced pluripotent stem cells over time. PLoS Genet. *7*, e1002085.

Pasini, D., Cloos, P.A., Walfridsson, J., Olsson, L., Bukowski, J.P., Johansen, J.V., Bak, M., Tommerup, N., Rappsilber, J., and Helin, K. (2010). JARID2 regulates binding of the Polycomb repressive complex 2 to target genes in ES cells. Nature *464*, 306–310.

Peng, J.C., Valouev, A., Swigut, T., Zhang, J., Zhao, Y., Sidow, A., and Wysocka, J. (2009). Jarid2/Jumonji coordinates control of PRC2 enzymatic activity and target gene occupancy in pluripotent cells. Cell *139*, 1290–1302.

Puda, A., Milosevic, J.D., Berg, T., Klampfl, T., Harutyunyan, A.S., Gisslinger, B., Rumi, E., Pietra, D., Malcovati, L., Elena, C., et al. (2012). Frequent deletions of JARID2 in leukemic transformation of chronic myeloid malignancies. Am. J. Hematol. *87*, 245–250.

Rinn, J.L., and Chang, H.Y. (2012). Genome regulation by long noncoding RNAs. Annu. Rev. Biochem. *81*, 145–166.

Rinn, J.L., Kertesz, M., Wang, J.K., Squazzo, S.L., Xu, X., Brugmann, S.A., Goodnough, L.H., Helms, J.A., Farnham, P.J., Segal, E., and Chang, H.Y. (2007). Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. Cell *129*, 1311–1323.

Schwartz, Y.B., and Pirrotta, V. (2007). Polycomb silencing mechanisms and the management of genomic programmes. Nat. Rev. Genet. *8*, 9–22.

Shen, X., Kim, W., Fujiwara, Y., Simon, M.D., Liu, Y., Mysliwiec, M.R., Yuan, G.-C., Lee, Y., and Orkin, S.H. (2009). Jumonji modulates polycomb activity and self-renewal versus differentiation of stem cells. Cell *139*, 1303–1314.

Son, J., Shen, S.S., Margueron, R., and Reinberg, D. (2013). Nucleosome binding activities within JARID2 and EZH1 regulate the function of PRC2 on chromatin. Genes Dev., in press. Published online December 15, 2013. gad.225888.113.

Stadtfeld, M., Apostolou, E., Akutsu, H., Fukuda, A., Follett, P., Natesan, S., Kono, T., Shioda, T., and Hochedlinger, K. (2010). Aberrant silencing of imprinted genes on chromosome 12qF1 in mouse induced pluripotent stem cells. Nature *465*, 175–181.

Takahashi, N., Okamoto, A., Kobayashi, R., Shirai, M., Obata, Y., Ogawa, H., Sotomaru, Y., and Kono, T. (2009). Deletion of Gtl2, imprinted non-coding RNA, with its differentially methylated region induces lethal parent-origin-dependent defects in mice. Hum. Mol. Genet. *18*, 1879–1888.

Takeuchi, T., Yamazaki, Y., Katoh-Fukui, Y., Tsuchiya, R., Kondo, S., Motoyama, J., and Higashinakagawa, T. (1995). Gene trap capture of a novel mouse gene, jumonji, required for neural tube formation. Genes Dev. *9*, 1211–1222.

Trojer, P., and Reinberg, D. (2007). Facultative heterochromatin: is there a distinctive molecular signature? Mol. Cell *28*, 1–13.

Tsai, M.C., Manor, O., Wan, Y., Mosammaparast, N., Wang, J.K., Lan, F., Shi, Y., Segal, E., and Chang, H.Y. (2010). Long noncoding RNA as modular scaffold of histone modification complexes. Science *329*, 689–693.

Ulitsky, I., Shkumatava, A., Jan, C.H., Sive, H., and Bartel, D.P. (2011). Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. Cell *147*, 1537–1550.

Wang, K.C., Yang, Y.W., Liu, B., Sanyal, A., Corces-Zimmerman, R., Chen, Y., Lajoie, B.R., Protacio, A., Flynn, R.A., Gupta, R.A., et al. (2011). A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. Nature *472*, 120–124.

Yap, K.L., Li, S., Muñoz-Cabello, A.M., Raguz, S., Zeng, L., Mujtaba, S., Gil, J., Walsh, M.J., and Zhou, M.M. (2010). Molecular interplay of the noncoding RNA ANRIL and methylated histone H3 lysine 27 by polycomb CBX7 in transcriptional silencing of INK4a. Mol. Cell *38*, 662–674.

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W., and Liu, X.S. (2008). Model-based analysis of ChIP-Seq (MACS). Genome Biol. *9*, R137.

Zhao, J., Ohsumi, T.K., Kung, J.T., Ogawa, Y., Grau, D.J., Sarma, K., Song, J.J., Kingston, R.E., Borowsky, M., and Lee, J.T. (2010). Genome-wide identification of polycomb-associated RNAs by RIP-seq. Mol. Cell *40*, 939–953.

Cell Reports
Resource

Quantitative ChIP-Seq Normalization Reveals Global Modulation of the Epigenome

David A. Orlando,[1,*] Mei Wei Chen,[1] Victoria E. Brown,[1] Snehakumari Solanki,[1] Yoon J. Choi,[1] Eric R. Olson,[1] Christian C. Fritz,[1] James E. Bradner,[2,3] and Matthew G. Guenther[1,*]
[1]Syros Pharmaceuticals, 480 Arsenal Street, Watertown, MA 02472, USA
[2]Department of Medical Oncology, Dana-Farber Cancer Institute
[3]Department of Medicine, Harvard Medical School, Boston, MA 02115, USA
*Correspondence: dorlando@syros.com (D.A.O.), mguenther@syros.com (M.G.G.)
http://dx.doi.org/10.1016/j.celrep.2014.10.018

SUMMARY

Epigenomic profiling by chromatin immunoprecipitation coupled with massively parallel DNA sequencing (ChIP-seq) is a prevailing methodology used to investigate chromatin-based regulation in biological systems such as human disease, but the lack of an empirical methodology to enable normalization among experiments has limited the precision and usefulness of this technique. Here, we describe a method called ChIP with reference exogenous genome (ChIP-Rx) that allows one to perform genome-wide quantitative comparisons of histone modification status across cell populations using defined quantities of a reference epigenome. ChIP-Rx enables the discovery and quantification of dynamic epigenomic profiles across mammalian cells that would otherwise remain hidden using traditional normalization methods. We demonstrate the utility of this method for measuring epigenomic changes following chemical perturbations and show how reference normalization of ChIP-seq experiments enables the discovery of disease-relevant changes in histone modification occupancy.

INTRODUCTION

The ability to map genomic occupancy of transcriptional regulators, histone posttranslational modifications, and DNA methylation (epigenomic modifications) has enabled the elucidation of transcriptional mechanisms, genome organization, mapping of functional regulatory elements, and discovery of disease-associated chromatin markers (Badeaux and Shi, 2013; Barski et al., 2007; Lee and Young, 2013; Rivera and Ren, 2013; Zhou et al., 2011). Such targeted and large-scale epigenome mapping efforts have revealed chromatin regulatory proteins that are therapeutic targets for a wide variety of human diseases (Azad et al., 2013; Dawson and Kouzarides, 2012; Deshpande et al., 2012; Wee et al., 2014). Many of these chromatin regulators exhibit cell-type-selective, gene-selective, or disease-relevant effects, creating a critical need to study the chromatin modifications catalyzed by these regulators. Accurate quantification of both global and loci-specific chromatin modifications is needed to allow the discovery and characterization of epigenomic regulators and epigenome-modulating agents.

Traditional ChIP-seq methodologies are not inherently quantitative and therefore do not allow direct comparisons between samples derived from different cell types or between cells that have experienced a perturbation, such as a genomic alteration or chemical treatment. For example, if we employ the traditional reads per million (RPM) ChIP-seq normalization method, a cell population containing chromatin state "A" (a high level of histone posttranslational modification) will appear similar to a cell population containing chromatin state "B," where 50% of the signal has been removed (Figure 1A), because the signal is quantified as a simple percentage of all mapped reads. Moreover, additional variables, such as variations in genome fragmentation, immunoprecipitation efficiency, or other experimental steps, frequently confound analysis. Efforts to correct for these variables have produced in silico normalization strategies, but an empirical method to enable direct and quantitative comparisons among epigenomic ChIP-seq data sets is still lacking (Bardet et al., 2012; Landt et al., 2012; Liang and Keleş, 2012; Liu et al., 2013; Nair et al., 2012). Because of the experimental and analytical restrictions of ChIP-seq, a robust normalization methodology is needed to quantify epigenome differences among varying cell populations, treatments, and genomic states.

RESULTS

Here we present a method, called ChIP with reference exogenous genome (ChIP-Rx), that utilizes a constant amount of reference or "spike-in" epigenome, added on a per-cell basis, to allow direct comparison between two or more ChIP-seq samples (Figure 1B). Analogous methodologies have been applied in areas of gene-expression analysis that have revealed global transcriptional amplification upon normalization and in MethylC-seq, where bisulfate conversion rates have been normalized (Kanno et al., 2006; Krueger et al., 2012; Lin et al., 2012; Lovén et al., 2012; van de Peppel et al., 2003). These advancements have allowed standardization, precision, and a mechanistic

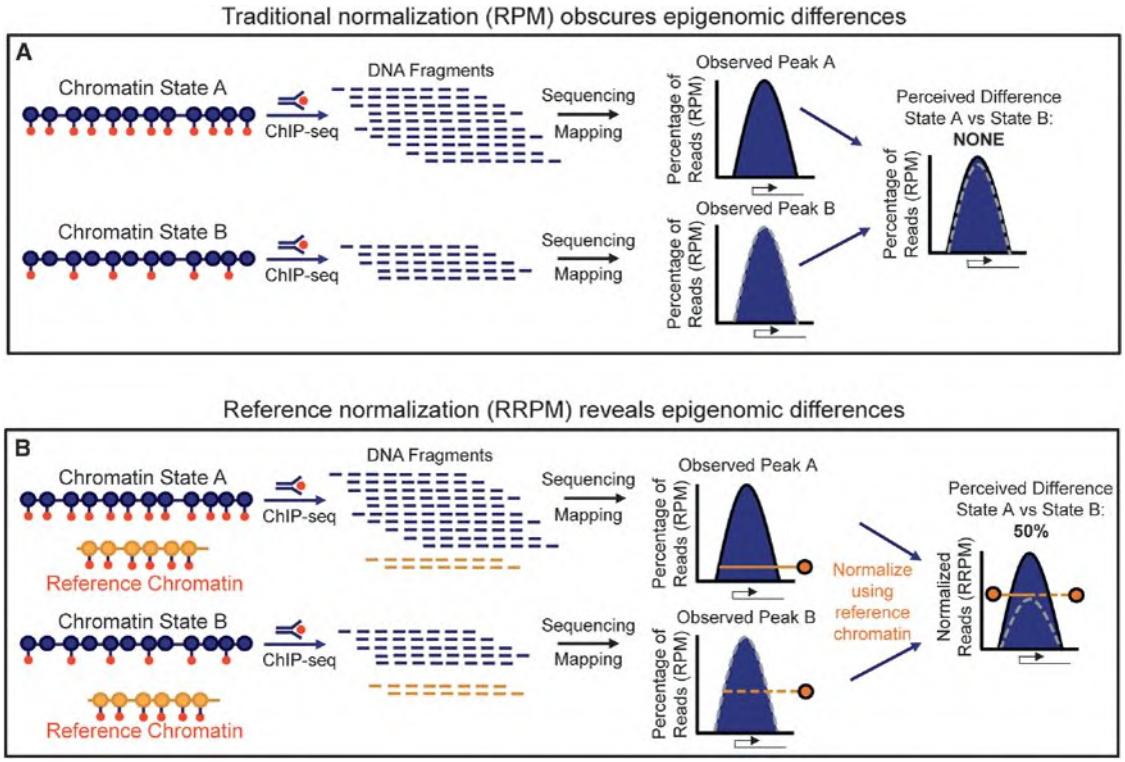


Figure 1. Normalization and Interpretation of ChIP-Seq Data

(A) Schematic representation of a typical ChIP-seq data workflow. Interrogation of a human epigenome (Blue circles, nucleosomes) with a full complement of histone modification (red circles, top) versus an epigenome with a half complement of histone modification (red circles, bottom). ChIP, sequencing, and mapping using reads per million (RPM) reveals ChIP-seq peaks (blue). A comparison of the peaks as a percentage of the total reads reveals little difference.

(B) Schematic representation of a ChIP-seq data workflow with reference genome normalization. Interrogation of a human epigenome (Blue circles, nucleosomes) with a full complement of histone modification (red circles, top) versus an epigenome with a half complement of histone modification (red circles, bottom). A fixed amount of reference epigenome (orange, nucleosomes; red, histone modifications) is added to human cells in each condition. After ChIP, sequencing, and mapping, the ChIP sequence reads are normalized to the percentage of reference genome reads in the sample (reference-adjusted RPM [RRPM]). A comparison of ChIP-seq signals using normalized reads reveals a 50% difference between peaks. This method is called ChIP with reference exogenous genome (ChIP-Rx).

understanding of RNA transcription (van Bakel and Holstege, 2004; Jiang et al., 2011; Li et al., 2013); however, no cell-count-normalized methods have been applied to global correction of histone posttranslational modifications. Since a vast array of histone modifications have been described in eukaryotic cells that play roles in organismal development, maintenance of cell state, differentiation, and disease, including those associated with transcriptional processes, genome organization, DNA repair, and cell-cycle progression (Calo and Wysocka, 2013; Pastor et al., 2013; Rinn and Chang, 2012; Rivera and Ren, 2013; Tan et al., 2011; Tian et al., 2012), a quantitative method for comparing these key marks is needed. We reasoned that the *Drosophila melanogaster* genome would be a desirable exogenous reference for mammalian cells because the *Drosophila* genome is well studied and has a high-quality sequence assembly, there is minimal mapping of the *Drosophila* genome sequence to human or mouse genomes (>0.05%; Table S1; Supplemental Experimental Procedures), *Drosophila* cells are readily available in large quantities, and the *Drosophila* epigenome displays nearly all of the key histone modification marks reported in humans. Moreover, histone proteins are among the most conserved proteins from humans to yeast, indicating that available ChIP-quality antibodies would likely recognize both *Drosophila* and human chromatin (Sullivan et al., 2002; Wolffe and Pruss, 1996).

To determine the impact of mixing interspecies epigenomes, we tested whether the addition of a reference genome (*Drosophila* S2 cells) would inherently affect our ability to detect a histone modification within the test sample (human cells) using ChIP-seq. We compared Jurkat cells alone with Jurkat cells that had been mixed with *Drosophila* cells and analyzed the resulting histone H3 lysine-79 dimethyl (H3K79me2) ChIP-seq profiles (Figure S1; Table S2). We determined that mixing of *Drosophila* and human cells did not induce large-scale changes in the H3K79me2 profiles, as the profiles of these cell populations were highly correlated by total signal as well as enriched loci overlap (Pearson correlation = 0.96; Supplemental Experimental Procedures). Moreover, reads originating from human or *Drosophila* could be separated with 99% accuracy (Supplemental Experimental Procedures). Together, these results indicate that the addition of a reference genome did not impede our ability to detect histone mark occupancy.

We devised an experiment to test our ability to detect changes in histone modification occupancy throughout the human

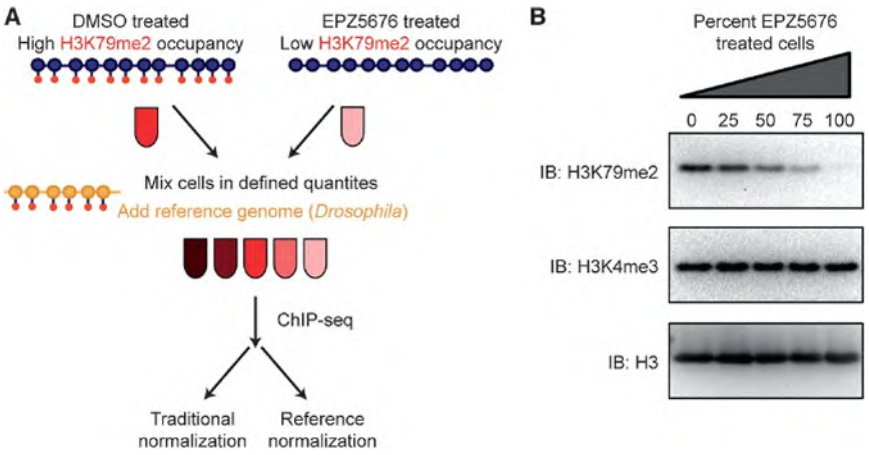


Figure 2. Experimental Design of Differential H3K79me2 Detection

(A) Schematic representation of differential H3K79me2 detection and normalization strategies. Two populations of cells were produced: a human epigenome (blue nucleosomes) with a full complement of H3K79me2 (red circles, top left) and a human epigenome (blue nucleosomes) with depleted H3K79me2 due to EPZ5676 exposure (top right). These cells were mixed in defined proportions in order to allow a dilution of total genomic histone modification (dark red to pink). Cell mixtures were subjected to ChIP-seq in the presence of the reference *Drosophila* epigenome (orange). ChIP-seq signals were calculated based on traditional or *Drosophila*-reference-normalized methods. See also Figure S1.

(B) Western blot validation of H3K79me2 depletion in Jurkat cells. Mixtures of 0%–100% EPZ5676-treated cells (0:100; 25:75; 50:50, 75:25; 100:0 proportions of [DMSO-treated:EPZ5676-treated] cells) were measured by immunoblot (IB) for the presence of H3K79me2, H3K4me3, or total histone H3 (loading control). Treated cells were exposed to 20 μM EPZ5676 for 4 days.

See also Table S1.

epigenome using ChIP-Rx (Figure 2A). We reasoned that an initial test of ChIP-Rx normalization should feature an epigenomic modification that could be readily removed and was not essential for cell viability in the model cell line. Using the selective DOT1L inhibitor EPZ5676 (Daigle et al., 2013), we depleted the Histone H3 lysine-79 dimethyl (H3K79me2) modification from Jurkat cell bulk histones (Figure 2B). The H3K79me2 modification is catalyzed by the DOT1L protein and is associated with the release of paused RNA Polymerase II and licensing of transcriptional elongation (van Leeuwen et al., 2002; Ng et al., 2002; Shanower et al., 2005; Steger et al., 2008). This modification is typically deposited within the 5′ regions of genes and its presence is not critical for Jurkat cell viability (Daigle et al., 2011, 2013; Schübeler et al., 2004; Steger et al., 2008). By mixing untreated cells (full H3K79me2) with EPZ5676-treated cells (H3K79me2 depleted), we created a set of cell populations with defined quantities of H3K79me2, as verified by immunodetection (Figure 2B). To each of these cell populations we added *Drosophila* cells at a ratio of one *Drosophila* cell per two human cells, which provided a constant ''reference'' amount of H3K79me2 per human cell. We performed ChIP-seq from samples consisting of 0:100, 25:75, 50:50, 75:25, and 100:0 proportions of EPZ5676 to DMSO-treated Jurkat cells. We then tested whether traditional ChIP-seq analysis methods would reveal the decrease in human per-cell H3K79me2 occupancy and, if not, whether the addition of the *Drosophila* epigenome would allow detection of H3K79me2 removal.

A key prediction of our normalization method is that as the global level of a histone modification is depleted in human cells, the percentage of total reads mapping to the reference *Drosophila* genome should increase. This is because the constant amount of reference genome added per human cell accounts for a greater percentage of total ChIP DNA fragments as human epitopes are lost (see the ratio of blue to orange DNA fragments in Figure 1B). To test this prediction, we interrogated cells with defined H3K79me2 levels (Figure 2B) by ChIP-Rx to measure genomic H3K79me2 occupancy. As a control, we also measured H3K4me3 occupancy, which is a histone modification that is not appreciably changed within our test cell populations (Figure 2B). As predicted, H3K79me2 depletion in Jurkat cells both reduced ChIP-Rx reads mapping to the human genome and increased reads mapping to the *Drosophila* genome (Figure 3A). We did not observe a similar change in the mapping ratio for H3K4me3 in the same samples, consistent with the finding that H3K4me3 was not preferentially removed from the human genome (Figure 3B). These results demonstrate that a reference genome can internally normalize the read count.

We next used the reference *Drosophila* genome to quantitatively normalize across experiments. To make ChIP-seq data quantitative on a per-cell basis, it is necessary to introduce a reference signal that is constant per cell, from which a normalization factor can be derived. Our ChIP-Rx protocol uses the signal from a fixed amount of *Drosophila* genome per human cell as this reference. We derived a normalization factor (see the Supplemental Experimental Procedures) for each experiment, such that the resulting *Drosophila* signal was equilibrated across all experiments (Table S2; Figure S2). Using traditional RPM normalization, the loci-specific ChIP-seq profiles and metagene profiles for H3K79me2 and H3K4me3 appear unchanged for the majority of samples (Figures 3C–3F), despite evidence that the H3K79me2 modification is progressively depleted (Figure 2B). After normalization with the *Drosophila* reference (normalized reference-adjusted RPM [RRPM]), a striking and gradated decrease in H3K79me2 signal across the samples is evident (Figures 3C, 3E, 3G, and S3). Normalization did not appreciably affect the metagene profiles of the control H3K4me3 experiments (Figures 3D, 3F, 3H, and S3). Repeat experiments produced the same result in all cases: normalization revealed a loss of H3K79me2 across the samples and H3K4me3 profiles were not significantly affected (Figure S4). These results indicate that normalization to a *Drosophila* reference is an effective method for quantitatively comparing multiple experiments and can reveal changes in histone modification that may not be apparent without proper normalization.

Having validated the ChIP-Rx methodology using standardized quantities of H3K79me2, we next tested our ability to

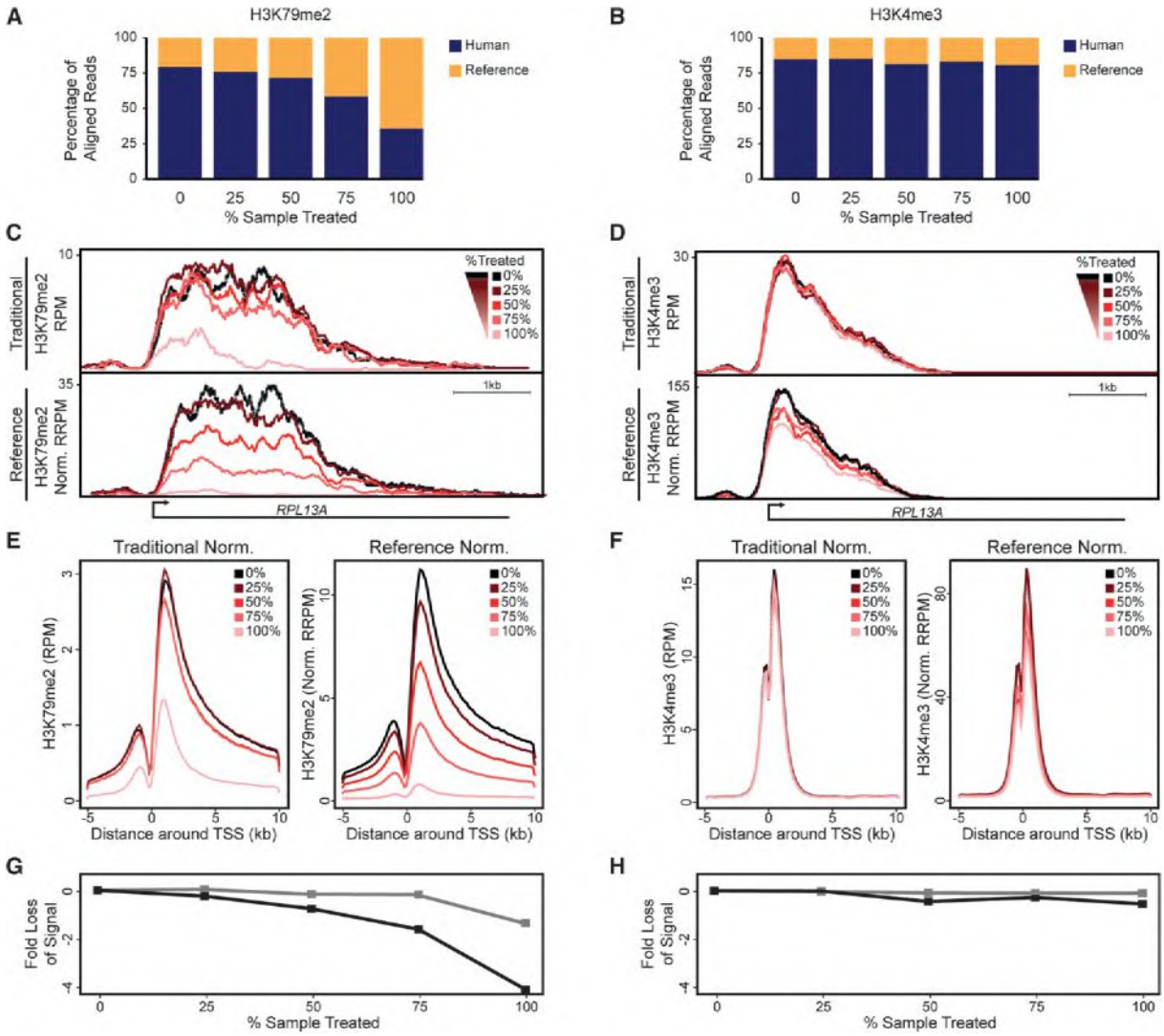


Figure 3. ChIP-Rx Reveals Quantitative Epigenome Changes

(A and B) Percentage of reads aligning to either test (human, blue) or *Drosophila* (reference, orange) genomes after H3K79me2 ChIP-Rx (A) or H3K4me3 ChIP-Rx (B). Samples containing 0%, 25%, 50%, 75%, or 100% EPZ5676 treated Jurkat cells were used as defined in Figure 2B.

(C and D) Sequenced reads from H3K79me2 (C) and H3K4me3 (D) immunoprecipitations at the RPL13A gene locus in traditional reads per million (RPM,top) or reference-adjusted reads per million (RRPM, bottom; see Experimental Procedures). Color indicates the percentage of sample treated with EPZ5676. The gene model is shown below the track.

(E) Meta-gene profile of H3K79me2-occupied genes in Jurkat cells. Meta-gene profiles were produced with traditional RPM (left) or RRPM (right). Color indicates the percentage of Jurkat cell sample treated with EPZ5676 as in Figure 2B. Region −5 to +10 kb around the transcription start site (TSS) is shown. Meta-gene profile was derived from top 5,000 protein-coding genes as defined by total H3K79me2 signal in the 0% treated (untreated with EPZ5676) sample. A meta-gene profile representing all genes is shown in Figure S3.

(F) Meta-gene profile of H3K4me3-occupied genes in Jurkat cells. Meta-gene profiles were produced with traditional RPM (left) or RRPM (right). Color indicates the percentage of Jurkat cell sample treated with EPZ5676 as in Figure 2B. Region −5 to +10 kb around the transcription start site (TSS) is shown. Meta-gene profile was derived from top 5,000 protein-coding genes as defined by total H3K4me3 signal in the 0% treated (untreated with EPZ5676) sample. A meta-gene profile representing all genes is shown in Figure S3.

(G and H) Line graphs display the observed fold-change difference in average meta-gene signal across the −5 to +10 kb window around the TSS for each H3K79me2 (G) or H3K4me3 (H) ChIP sample (x axis) relative to the signal from the 0% treated population using traditional (gray) or reference (black) normalization. See also Figures S2–S4 and Table S2.

normalize between ChIP-seq experiments in a disease-relevant system. MV4;11 acute myelomonocytic leukemia cells are a DOT1L-inhibitor sensitive model of human mixed-lineage-linked leukemia, a disease characterized by reciprocal translocations of the mixed-lineage leukemia (MLL) gene (Daigle et al., 2011, 2013; Deshpande et al., 2012). We treated MV4;11 cells with DOT1L inhibitor and measured changes in H3K79me2 occupancy in the presence or absence of the reference *Drosophila*

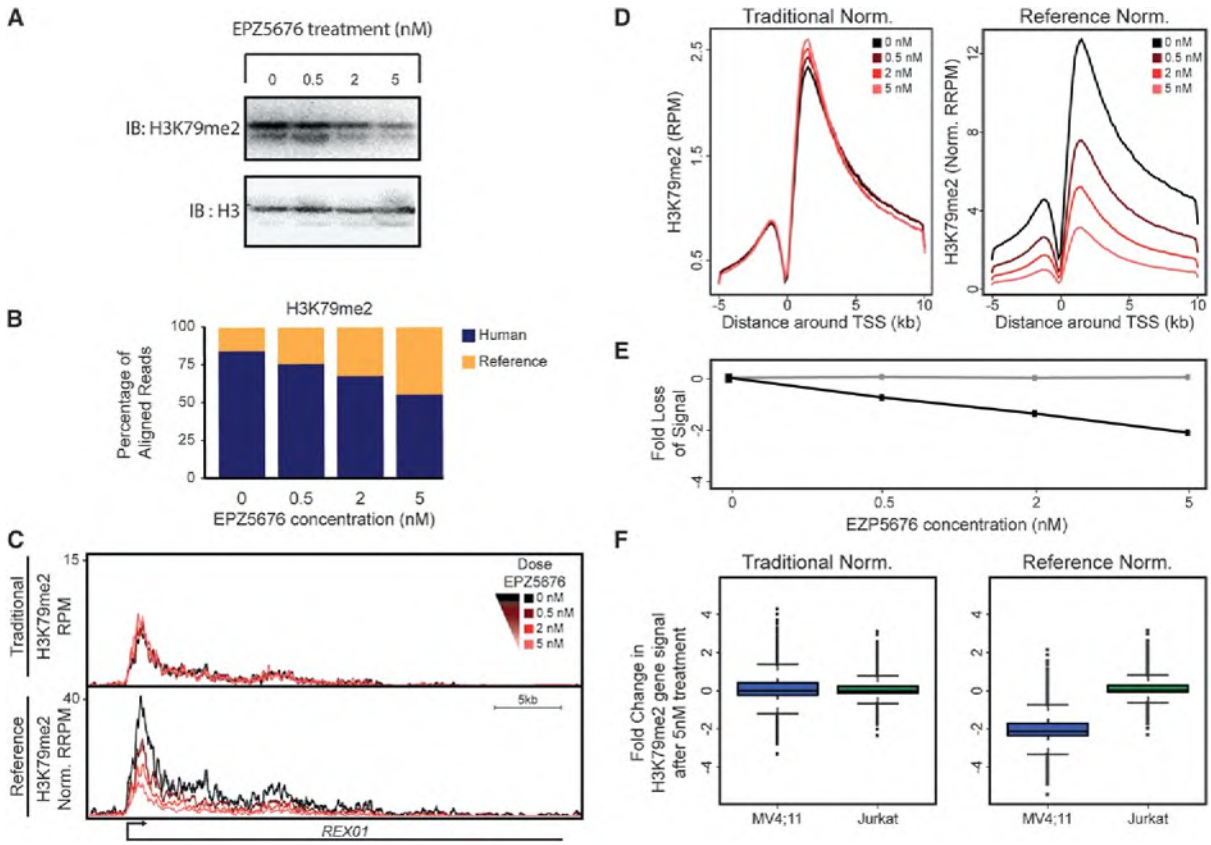


Figure 4. ChIP-Rx Reveals Epigenomic Alterations in Disease Cells that Respond to Drug Treatment

(A) Western blot showing the levels of H3K79me2 in MV4;11 cells after treatment for 4 days with increasing concentrations of EPZ5676.

(B) Percentage of H3K79me2 ChIP-seq reads aligning to either test (human, blue) or *Drosophila* (reference, orange) genomes after H3K79me2 ChIP-Rx from MV4;11 cells treated as in (A).

(C) Sequenced reads from H3K79me2 immunoprecipitations at the REXO1 gene locus in standard RPM (top) or RRPM (bottom) (see Experimental Procedures). Color indicates the concentration of EPZ5676 given to each sample. The gene model is shown below the track.

(D) Meta-gene profile of H3K79me2-occupied genes in MV4;11 cells. Meta-gene profiles were produced with traditional Reads Per Million (RPM, left) or Reference-adjusted Reads Per Million (RRPM, right). Color indicates the concentration of EPZ5676 used in each sample. The region −5 kb to +10 kb around the TSS is shown. Meta-gene profile was derived from top 5,000 protein-coding genes as defined by total H3K79me2 signal in the 0nM treated (untreated with EPZ5676) sample. A meta-gene profile representing all genes is shown in Figure S3.

(E) Line graph displays the observed fold-change difference in average meta-gene signal across the −5 to +10 kb window around the TSS for each H3K79me2 ChIP sample (x axis) relative to the signal from the 0 nM treated population using standard (gray) or reference (black) normalization.

(F) Box plots display the distribution of the observed fold change of H3K79me2 signal −5 kb to +10 kb around the TSS of all genes between the 0 nM and 5 nM treated samples (blue, MV4;11; green, Jurkat) for all genes using traditional (left) or reference-adjusted (right) normalization (see the Supplemental Experimental Procedures).

See also Figures S3 and S5 and Table S2.

epigenome (Figures 4A–4E, S5A, and S5B). EPZ5676 induced a dose-dependent decrease in bulk H3K79me2 (Figure 4A), but this result was masked when we quantified H3K79me2 occupancy using traditional normalization (Figures 4C and 4E). We observed a dose-dependent decrease in H3K79me2 genomic occupancy only after employing reference normalization (Figures 4C–4E). This unmasking of epigenomic effects may be critical for understanding the cell-type-selective effects of small-molecule epigenome modulators. For example, MV4;11 cells exhibit global H3K79me2 depletion at a low dose of EPZ5676, consistent with the known selectivity of the DOT1L inhibitor for leukemic cells carrying MLL translocations, but EPZ5676-insensitive Jurkat cells do not (Figures 4A and S5C; Daigle et al., 2013). Thus, normalizing to a reference exogenous genome rectifies the protein-level measurements and genome occupancy of modified histones, and reveals subtle epigenomic changes that may underlie or predict cellular responses to drugs (Figure 4). These results show that ChIP-Rx enables the discovery of epigenomic changes that can provide insight into disease and inform drug mechanisms.

DISCUSSION

In summary, we have demonstrated that ChIP-Rx allows the discovery and quantification of dynamic epigenomic profiles across mammalian cells that would otherwise remain hidden using

traditional normalization methods. A recent study employed a similar reference strategy for ChIP-seq normalization (Bonhoure et al., 2014); however, our method offers two crucial advantages that allow direct comparative epigenomic analysis. First, our method introduces the reference at the beginning of the experiment, thus normalizing for variation throughout the experiment, including chromatin fragmentation and immunoenrichment, both of which are critical for epitope and genome retrieval (Kidder et al., 2011; Meyer and Liu, 2014; Raha et al., 2010). Second, as was analogously shown for RNA expression correction (Lovén et al., 2012), our method introduces the reference "spike-in" on a per-cell basis as opposed to total chromatin, thus allowing the detection of unidirectional chromatin changes irrespective of variations in ploidy or gross chromatin. Thus, our method provides greater accuracy in determining epigenome changes that occur upon cell perturbation or exposure to small-molecule inhibitors as compared with current methods. Importantly, ChIP-Rx allows for the detection of subtle epigenomic changes, as opposed to qualitative occupancy calls, and thus advances the ChIP-seq methodology from a descriptive, binary readout to one that reveals gradated epigenomic changes. This is particularly important for the dose-ranging characterization of chemical tools and therapeutics targeting chromatin-associated complexes via genome-wide approaches. Application of this methodology to additional model systems, including mouse, rat, and zebrafish (Table S1), as well as additional histone modifications, including repressive (i.e., H3K27me3) and activating (i.e., H3K27ac) histone modifications, will enable far-reaching studies of comparative epigenomics.

We recommend the implementation of ChIP-Rx whenever quantitative or comparative epigenomic changes are under investigation. The method described here will be critical for understanding the global and site-selective epigenomic changes that occur in human disease, during cell-state changes, and especially the action of small-molecule inhibitors of chromatin-modulating proteins.

EXPERIMENTAL PROCEDURES

Human Cell Lines, Growth, and Treatment

Jurkat cells were obtained from ATCC and maintained in RPMI (Life Technologies) supplemented with 10% fetal bovine serum (FBS; Life Technologies) at 5% CO_2 in 37°C. MV4;11 cells were obtained from ATCC and maintained in RPMI (Life Technologies) supplemented with 10% FBS at 5% CO_2 in 37°C.

Jurkat cells were treated with DMSO or EPZ5676 (Selleck Chemicals, catalog number S7062) at 5 nM or 20 μM for 4 days, and MV4;11 cells were treated with DMSO or EPZ5676 at 0.5 nM, 2 nM, or 5 nM for 4 days. Live-cell numbers were quantified using the Countess cell counter (Life Technologies).

At harvest, cells were crosslinked with 1% formaldehyde by addition of 1/10 volume of fresh 11% formaldehyde solution (11% formaldehyde 0.1 M NaCl, 1 mM EDTA, 0.5 mM EGTA, 50 mM HEPES) and incubation at room temperature for 8 min. Crosslinking reactions were quenched with a 1/20 volume of 2.5 M glycine for 1–5 min and cells were pelleted. The cells were then washed three times with ice-cold PBS. Washed cell pellets were flash frozen and stored at −80°C.

Preparation of *Drosophila* S2 Cells

Drosophila S2 cells (ATCC catalog number CRL-1963; Biovest part number OO.763/OO.627) were cultured in Schneider's *Drosophila* media (Life Technologies catalog number 21720-024) supplemented with 10% FBS to attain a density of 0.5–0.6 × 10^6 cells/ml. Cell culture and scale-up to 2 L was performed by Biovest International.

At harvest, cells were crosslinked with 1% formaldehyde by addition of a 1/10 volume of fresh 11% formaldehyde solution (11% formaldehyde 0.1 M NaCl, 1 mM EDTA, 0.5 mM EGTA, 50 mM HEPES) and incubation at room temperature for 8 min. Crosslinking reactions were quenched with a 1/20 volume of 2.5 M glycine for 1–5 min and cells were pelleted. The cells were then washed three times with ice-cold PBS. Washed cell pellets were flash frozen and stored at −80°C at 1 × 10^8 cells per aliquot.

ChIP-Rx

For each ChIP-Rx experiment, a 2:1 ratio of human:*Drosophila* cells was used. This corresponds to 20 million crosslinked human cells and 10 million crosslinked S2 cells (Jurkat experiments) or 15 million crosslinked human cells and 7.5 million crosslinked S2 cells (MV4;11 experiments).

S2 cells were added to human cells at the beginning of the ChIP-Rx workflow (during nuclei isolation). Once *Drosophila* S2 and human cells were combined, the sample was treated as a single ChIP-seq sample throughout the experiment until completion of DNA sequencing.

Briefly, frozen, crosslinked human and *Drosophila* cells were resuspended in parallel in cold Lysis Buffer 1 (140 mM NaCl, 1 mM EDTA, 50 mM HEPES, 10% glycerol, 0.5% NP-40, 0.25% Triton-X-100), incubated 10 min at 4°C, and pelleted. Both human and *Drosophila* cell samples were resuspended in parallel in Lysis Buffer 2 (10 mM TRIS [pH 8.0], 200 mM NaCl, 1 mM EDTA, 0.5 mM EGTA), incubated for 10 min at 4°C, and combined to the desired cell number ratios (two human cells per one *Drosophila* cell). The composite cell nuclei was then pelleted and resuspended in sonication buffer (10 mM TRIS [pH 8.0], 1 mM EDTA, 0.1% SDS).

Composite samples (human + *Drosophila* S2) in sonication buffer were sonicated using a Covaris E220 sonication water bath for 5 min. Sheared chromatin was diluted 1:1 in 2× dilution buffer (300 mM NaCl, 2 mM EDTA, 50 mM TRIS [pH 8.0], 1.5% Triton-X, 0.1% SDS) and incubated with either H3K79me2 (Abcam 3594)- or H3K4me3 (Millipore 07-473)-conjugated Protein G Dynal beads (Invitrogen) overnight (8–16 hr, rotating) at 4°C, and then washed two times with wash buffer 1 (50 mM HEPES, 140 mM NaCl, 1 mM EDTA, 1 mM EGTA, 0.75% Triton-X, 0.1% SDS, 0.05% DOC), two times with high-salt wash buffer (50 mM HEPES, 500 mM NaCl, 1 mM EDTA, 1 mM EGTA, 0.75% Triton-X, 0.1% SDS, 0.05% DOC), and one time with TE-NaCl buffer (10 mM Tris [pH 8.0], 1 mM EDTA, 50 mM NaCl). Samples were eluted from beads for 1 hr at 65°C in elution buffer (50 mM TRIS [pH 8.0], 10 mM EDTA, 1% SDS) and supernatant reverse-crosslinked at 65°C for 6–16 hr. Samples were diluted 1:1 with TE buffer (50 mM TRIS [pH 8.0], 1 mM EDTA) and treated with RNase A (0.2 mg/ml) for 2 hr at 37°C and then Proteinase K (0.2 mg/ml) for 2 hr at 55°C. DNA was isolated by phenol-chloroform extraction and ethanol precipitation. For detailed protocols see the Supplemental Experimental Procedures and Guenther et al. (2008).

Library Construction, Sequencing, and Data Collection

Libraries were constructed with the Illumina Tru-Seq library preparation kit using a target fragment size of 200–400 bp and multiplexing barcodes. Libraries were sequenced using Illumina HiSeq 2000 with single-end reads for 40 cycles. Sequences were demultiplexed and aligned using Bowtie2 against a "genome" that combines the human hg19 genome and the *Drosophila* dm3 genome (see the Supplemental Experimental Procedures). Individual accession numbers and read statistics available in Table S2.

Western Blots

Cells were harvested from all treatment groups and lysed with Triton extraction buffer (PBS containing 0.5% Triton X-100 [v/v], cOmplete Protease Inhibitors [Roche]) for 10 min with rotation. Nuclei were collected and acid extracted with 0.2 N HCl overnight. Histone proteins were collected from the supernatant and immunoblotted for H3K4me3 (Millipore 07-473), H3K79me2 (Abcam 3594), and histone H3 (Abcam 1791).

Determination of the Normalization Factor

A complete description of the basis and derivation of the ChIP-Rx normalization factor is provided in the Supplemental Experimental Procedures. In brief, we derived a normalization constant, α, such that after normalization the signal

per-reference cell (β) is the same across all samples. The total ChIP-seq signal derived from reference cells is simply the count of reads (in millions) aligning to the *Drosophila* genome, which we represent as N_d. Because the percentage of reference cells as a fraction of the total number of cells is constant and we assume that the epigenome of the reference cells does not vary appreciably, we can derive α as

$$\alpha * N_d = \beta$$

Because β is a constant, we can simply rewrite this as

$$\alpha * N_d = 1$$

or

$$\alpha = \frac{1}{N_d},$$

multiplying the read counts by α produces a normalized read count in normalized RRPM.

ACCESSION NUMBERS

The raw sequencing data reported in this work have been deposited in the NCBI Gene Expression Omnibus under accession number GSE60104.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, five figures, and two tables and can be found with this article online at http://dx.doi.org/10.1016/j.celrep.2014.10.018.

AUTHOR CONTRIBUTIONS

D.A.O., M.G.G., J.E.B., C.C.F., and E.R.O. designed and analyzed the research. M.W.C. conducted ChIP-seq and ChIP-Rx experiments. All other experiments were performed by M.W.C., V.E.B., Y.J.C., S.S., and M.G.G. D.A.O. performed the computational analysis. M.G.G. and D.A.O. wrote the manuscript.

ACKNOWLEDGMENTS

We thank Richard Young, Matthew Eaton, Cindy Collins, Charles Lin, Jakob Loven, Tony Lee, Jason Marineau, Michael McKeown, and Peter Rahl for helpful discussions and comments. We also thank Thomas Volkert and the Whitehead Genome Technology Core for high-throughput sequencing. J.E.B. is a founder of Syros Pharmaceuticals. D.A.O., M.W.C., V.E.B., S.S., Y.J.C., E.R.O., C.C.F., and M.G.G. are employees of Syros Pharmaceuticals.

Received: August 19, 2014
Revised: September 24, 2014
Accepted: October 9, 2014
Published: October 30, 2014

REFERENCES

Azad, N., Zahnow, C.A., Rudin, C.M., and Baylin, S.B. (2013). The future of epigenetic therapy in solid tumours—lessons from the past. Nat Rev Clin Oncol *10*, 256–266.

Badeaux, A.I., and Shi, Y. (2013). Emerging roles for chromatin as a signal integration and storage platform. Nat. Rev. Mol. Cell Biol. *14*, 211–224.

Bardet, A.F., He, Q., Zeitlinger, J., and Stark, A. (2012). A computational pipeline for comparative ChIP-seq analyses. Nat. Protoc. *7*, 45–61.

Barski, A., Cuddapah, S., Cui, K., Roh, T.-Y., Schones, D.E., Wang, Z., Wei, G., Chepelev, I., and Zhao, K. (2007). High-resolution profiling of histone methylations in the human genome. Cell *129*, 823–837.

Bonhoure, N., Bounova, G., Bernasconi, D., Praz, V., Lammers, F., Canella, D., Willis, I.M., Herr, W., Hernandez, N., and Delorenzi, M.; CycliX Consortium (2014). Quantifying ChIP-seq data: a spiking method providing an internal reference for sample-to-sample normalization. Genome Res. *24*, 1157–1168.

Calo, E., and Wysocka, J. (2013). Modification of enhancer chromatin: what, how, and why? Mol. Cell *49*, 825–837.

Daigle, S.R., Olhava, E.J., Therkelsen, C.A., Majer, C.R., Sneeringer, C.J., Song, J., Johnston, L.D., Scott, M.P., Smith, J.J., Xiao, Y., et al. (2011). Selective killing of mixed lineage leukemia cells by a potent small-molecule DOT1L inhibitor. Cancer Cell *20*, 53–65.

Daigle, S.R., Olhava, E.J., Therkelsen, C.A., Basavapathruni, A., Jin, L., Boriack-Sjodin, P.A., Allain, C.J., Klaus, C.R., Raimondi, A., Scott, M.P., et al. (2013). Potent inhibition of DOT1L as treatment of MLL-fusion leukemia. Blood *122*, 1017–1025.

Dawson, M.A., and Kouzarides, T. (2012). Cancer epigenetics: from mechanism to therapy. Cell *150*, 12–27.

Deshpande, A.J., Bradner, J., and Armstrong, S.A. (2012). Chromatin modifications as therapeutic targets in MLL-rearranged leukemia. Trends Immunol. *33*, 563–570.

Guenther, M.G., Lawton, L.N., Rozovskaia, T., Frampton, G.M., Levine, S.S., Volkert, T.L., Croce, C.M., Nakamura, T., Canaani, E., and Young, R.A. (2008). Aberrant chromatin at genes encoding stem cell regulators in human mixed-lineage leukemia. Genes Dev. *22*, 3403–3408.

Jiang, L., Schlesinger, F., Davis, C.A., Zhang, Y., Li, R., Salit, M., Gingeras, T.R., and Oliver, B. (2011). Synthetic spike-in standards for RNA-seq experiments. Genome Res. *21*, 1543–1551.

Kanno, J., Aisaki, K.I., Igarashi, K., Nakatsu, N., Ono, A., Kodama, Y., and Nagao, T. (2006). "Per cell" normalization method for mRNA measurement by quantitative PCR and microarrays. BMC Genomics *7*, 64.

Kidder, B.L., Hu, G., and Zhao, K. (2011). ChIP-Seq: technical considerations for obtaining high-quality data. Nat. Immunol. *12*, 918–922.

Krueger, F., Kreck, B., Franke, A., and Andrews, S.R. (2012). DNA methylome analysis using short bisulfite sequencing data. Nat. Methods *9*, 145–151.

Landt, S.G., Marinov, G.K., Kundaje, A., Kheradpour, P., Pauli, F., Batzoglou, S., Bernstein, B.E., Bickel, P., Brown, J.B., Cayting, P., et al. (2012). ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Res. *22*, 1813–1831.

Lee, T.I., and Young, R.A. (2013). Transcriptional regulation and its misregulation in disease. Cell *152*, 1237–1251.

Li, Y., Wang, H., Muffat, J., Cheng, A.W., Orlando, D.A., Lovén, J., Kwok, S.-M., Feldman, D.A., Bateup, H.S., Gao, Q., et al. (2013). Global transcriptional and translational repression in human-embryonic-stem-cell-derived Rett syndrome neurons. Cell Stem Cell *13*, 446–458.

Liang, K., and Keleş, S. (2012). Normalization of ChIP-seq data with control. BMC Bioinformatics *13*, 199.

Lin, C.Y., Lovén, J., Rahl, P.B., Paranal, R.M., Burge, C.B., Bradner, J.E., Lee, T.I., and Young, R.A. (2012). Transcriptional amplification in tumor cells with elevated c-Myc. Cell *151*, 56–67.

Liu, B., Yi, J., Sv, A., Lan, X., Ma, Y., Huang, T.H., Leone, G., and Jin, V.X. (2013). QChIPat: a quantitative method to identify distinct binding patterns for two biological ChIP-seq samples in different experimental conditions. BMC Genomics *14* (*Suppl 8*), S3.

Lovén, J., Orlando, D.A., Sigova, A.A., Lin, C.Y., Rahl, P.B., Burge, C.B., Levens, D.L., Lee, T.I., and Young, R.A. (2012). Revisiting global gene expression analysis. Cell *151*, 476–482.

Meyer, C.A., and Liu, X.S. (2014). Identifying and mitigating bias in next-generation sequencing methods for chromatin biology. Nat. Rev. Genet. *15*, 709–721.

Nair, N.U., Sahu, A.D., Bucher, P., and Moret, B.M.E. (2012). ChIPnorm: a statistical method for normalizing and identifying differential regions in histone modification ChIP-seq libraries. PLoS ONE *7*, e39573.

Ng, H.H., Feng, Q., Wang, H., Erdjument-Bromage, H., Tempst, P., Zhang, Y., and Struhl, K. (2002). Lysine methylation within the globular domain of histone

H3 by Dot1 is important for telomeric silencing and Sir protein association. Genes Dev. *16*, 1518–1527.

Pastor, W.A., Aravind, L., and Rao, A. (2013). TETonic shift: biological roles of TET proteins in DNA demethylation and transcription. Nat. Rev. Mol. Cell Biol. *14*, 341–356.

Raha, D., Hong, M., and Snyder, M. (2010). ChIP-Seq: a method for global identification of regulatory elements in the genome. Curr. Protoc. Mol. Biol. *Chapter 21*, 1–14.

Rinn, J.L., and Chang, H.Y. (2012). Genome regulation by long noncoding RNAs. Annu. Rev. Biochem. *81*, 145–166.

Rivera, C.M., and Ren, B. (2013). Mapping human epigenomes. Cell *155*, 39–55.

Schübeler, D., MacAlpine, D.M., Scalzo, D., Wirbelauer, C., Kooperberg, C., van Leeuwen, F., Gottschling, D.E., O'Neill, L.P., Turner, B.M., Delrow, J., et al. (2004). The histone modification pattern of active genes revealed through genome-wide chromatin analysis of a higher eukaryote. Genes Dev. *18*, 1263–1271.

Shanower, G.A., Muller, M., Blanton, J.L., Honti, V., Gyurkovics, H., and Schedl, P. (2005). Characterization of the grappa gene, the Drosophila histone H3 lysine 79 methyltransferase. Genetics *169*, 173–184.

Steger, D.J., Lefterova, M.I., Ying, L., Stonestrom, A.J., Schupp, M., Zhuo, D., Vakoc, A.L., Kim, J.-E., Chen, J., Lazar, M.A., et al. (2008). DOT1L/KMT4 recruitment and H3K79 methylation are ubiquitously coupled with gene transcription in mammalian cells. Mol. Cell. Biol. *28*, 2825–2839.

Sullivan, S., Sink, D.W., Trout, K.L., Makalowska, I., Taylor, P.M., Baxevanis, A.D., and Landsman, D. (2002). The Histone Database. Nucleic Acids Res. *30*, 341–342.

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

Tian, Z., Tolić, N., Zhao, R., Moore, R.J., Hengel, S.M., Robinson, E.W., Stenoien, D.L., Wu, S., Smith, R.D., and Paša-Tolić, L. (2012). Enhanced top-down characterization of histone post-translational modifications. Genome Biol. *13*, R86.

van Bakel, H., and Holstege, F.C.P. (2004). In control: systematic assessment of microarray performance. EMBO Rep. *5*, 964–969.

van de Peppel, J., Kemmeren, P., van Bakel, H., Radonjic, M., van Leenen, D., and Holstege, F.C.P. (2003). Monitoring global messenger RNA changes in externally controlled microarray experiments. EMBO Rep. *4*, 387–393.

van Leeuwen, F., Gafken, P.R., and Gottschling, D.E. (2002). Dot1p modulates silencing in yeast by methylation of the nucleosome core. Cell *109*, 745–756.

Wee, S., Dhanak, D., Li, H., Armstrong, S.A., Copeland, R.A., Sims, R., Baylin, S.B., Liu, X.S., and Schweizer, L. (2014). Targeting epigenetic regulators for cancer therapy. Ann. N Y Acad. Sci. *1309*, 30–36.

Wolffe, A.P., and Pruss, D. (1996). Hanging on to histones. Chromatin. Curr. Biol. *6*, 234–237.

Zhou, V.W., Goren, A., and Bernstein, B.E. (2011). Charting histone modifications and the functional organization of mammalian genomes. Nat. Rev. Genet. *12*, 7–18.

Resource

Cell

Determination and Inference of Eukaryotic Transcription Factor Sequence Specificity

Matthew T. Weirauch,[1,2,15] Ally Yang,[2,15] Mihai Albu,[2] Atina G. Cote,[2] Alejandro Montenegro-Montero,[3] Philipp Drewe,[4] Hamed S. Najafabadi,[2] Samuel A. Lambert,[5] Ishminder Mann,[2] Kate Cook,[5] Hong Zheng,[2] Alejandra Goity,[3] Harm van Bakel,[2,6] Jean-Claude Lozano,[7] Mary Galli,[8] Mathew G. Lewsey,[8,9] Eryong Huang,[10] Tuhin Mukherjee,[11] Xiaoting Chen,[11] John S. Reece-Hoyes,[12] Sridhar Govindarajan,[13] Gad Shaulsky,[10] Albertha J.M. Walhout,[12] François-Yves Bouget,[7] Gunnar Ratsch,[4] Luis F. Larrondo,[3] Joseph R. Ecker,[8,9,14] and Timothy R. Hughes[2,5,*]

[1]Center for Autoimmune Genomics and Etiology (CAGE) and Divisions of Biomedical Informatics and Developmental Biology, Cincinnati Children's Hospital Medical Center, Cincinnati, OH 45229, USA

[2]Banting and Best Department of Medical Research and Donnelly Centre, University of Toronto, Toronto ON M5S 3E1, Canada

[3]Departamento de Genética Molecular y Microbiología, Facultad de Ciencias Biológicas, Pontificia Universidad Católica de Chile, Santiago 8331150, Chile

[4]Computational Biology Center, Sloan-Kettering Institute, New York, NY 10065, USA

[5]Department of Molecular Genetics, University of Toronto, Toronto ON M5S 1A8, Canada

[6]Icahn Institute for Genomics and Multiscale Biology, Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York City, NY 10029, USA

[7]Sorbonne Universités, UPMC Univ Paris 06, CNRS UMR 7621, CNRS, Laboratoire d'Océanographie Microbienne, Observatoire Océanologique, F-66650 Banyuls/mer, France

[8]Genomic Analysis Laboratory, The Salk Institute for Biological Studies, La Jolla, CA 92037, USA

[9]Plant Biology Laboratory, The Salk Institute for Biological Studies, La Jolla, CA 92037, USA

[10]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX 77030, USA

[11]Department of Electronic and Computing Systems, University of Cincinnati, Cincinnati, OH 45221, USA

[12]Program in Systems Biology, University of Massachusetts Medical School, Worcester, MA 01655, USA

[13]DNA2.0 Inc., Menlo Park, CA 94025, USA

[14]Howard Hughes Medical Institute, The Salk Institute for Biological Studies, La Jolla, CA 92037, USA

[15]Co-first author

*Correspondence: t.hughes@utoronto.ca

http://dx.doi.org/10.1016/j.cell.2014.08.009

SUMMARY

Transcription factor (TF) DNA sequence preferences direct their regulatory activity, but are currently known for only ~1% of eukaryotic TFs. Broadly sampling DNA-binding domain (DBD) types from multiple eukaryotic clades, we determined DNA sequence preferences for >1,000 TFs encompassing 54 different DBD classes from 131 diverse eukaryotes. We find that closely related DBDs almost always have very similar DNA sequence preferences, enabling inference of motifs for ~34% of the ~170,000 known or predicted eukaryotic TFs. Sequences matching both measured and inferred motifs are enriched in chromatin immunoprecipitation sequencing (ChIP-seq) peaks and upstream of transcription start sites in diverse eukaryotic lineages. SNPs defining expression quantitative trait loci in *Arabidopsis* promoters are also enriched for predicted TF binding sites. Importantly, our motif "library" can be used to identify specific TFs whose binding may be altered by human disease risk alleles. These data present a powerful resource for mapping transcriptional networks across eukaryotes.

INTRODUCTION

Transcription factor (TF) sequence specificities, typically represented as "motifs," are the primary mechanism by which cells recognize genomic features and regulate genes. Eukaryotic genomes contain dozens to thousands of TFs encoding at least one of the >80 known types of sequence-specific DNA-binding domains (DBDs) (Weirauch and Hughes, 2011). Yet, even in well-studied organisms, many TFs have unknown DNA sequence preference (de Boer and Hughes, 2012; Zhu et al., 2011), and there are virtually no experimental DNA binding data for TFs in the vast majority of eukaryotes. Moreover, even for the best-studied classes of DBDs, accurate prediction of DNA sequence preferences remains very difficult (Christensen et al., 2012; Persikov and Singh, 2014), despite the fact that identification of "recognition codes" that relate amino acid (AA) sequences to preferred DNA sequences has been a long-standing goal in the study of TFs (De Masi et al., 2011; Desjarlais and Berg, 1992; Seeman et al., 1976). These deficits represent a fundamental limitation in our ability to

analyze and interpret the function and evolution of DNA sequences.

The sequence preferences of TFs can be characterized systematically both in vivo (Odom, 2011) and in vitro (Jolma and Taipale, 2011; Stormo and Zhao, 2010). The most prevalent method for in vivo analysis is currently chromatin immunoprecipitation sequencing (ChIP-seq) (Barski and Zhao, 2009; Park, 2009), but ChIP does not inherently measure relative preference of a TF to individual sequences and may not identify correct TF motifs due to complicating factors such as chromatin structure and partner proteins (Gordân et al., 2009; Li et al., 2011; Liu et al., 2006; Yan et al., 2013). In contrast, it is relatively straightforward to derive motifs from all of the common methods for in vitro analysis of TF sequence specificity, including protein binding microarrays (PBMs), bacterial 1-hybrid (B1H), and high-throughput in vitro selection (HT-SELEX) (Stormo and Zhao, 2010), all of which have been applied to hundreds of proteins (Berger et al., 2008; Enuameh et al., 2013; Jolma et al., 2013; Noyes et al., 2008).

Previous large-scale studies have reported that proteins with similar DBD sequences tend to bind very similar DNA sequences, even when they are from distantly related species (e.g., fly and human). This observation is important because it suggests that the sequence preferences of TFs may be broadly inferred from data for only a small subset of TFs (Alleyne et al., 2009; Berger et al., 2008; Bernard et al., 2012; Noyes et al., 2008). However, these analyses have utilized data for only a handful of DBD classes and species and they contrast with numerous demonstrations that mutation of one or a few critical DBD AAs can alter the sequence preferences of a TF (Aggarwal et al., 2010; Cook et al., 1994; De Masi et al., 2011; Mathias et al., 2001; Noyes et al., 2008), which suggest that prediction of DNA binding preferences by homology should be highly error-prone. To our knowledge, rigorous and exhaustive analyses of the accuracy and limitations of inference approaches to predicting TF DNA-binding motifs using DBD sequences has not been done.

Here, we determined the DNA sequence preferences for >1,000 carefully-selected TFs from 131 species, representing all major eukaryotic clades and encompassing 54 DBD classes. We show that, in general, sequence preferences can be accurately inferred by overall DBD AA identity, suggesting that mutations that dramatically impact sequence specificity are relatively rare. By identifying distinct confidence thresholds for each individual DBD class (i.e., levels of protein sequence identity above which motifs can be assumed to be identical between two proteins), we infer sequence preferences for roughly one-third of all known eukaryotic TFs, based on experimental data for fewer than 2% of them. Cross-validation indicates that ~89% of predicted sequence preferences are as accurate as experimental replicates of the same TF. We demonstrate the functional relevance and utility of both known and inferred motifs by showing that they coincide with ChIP-seq binding peak sequences, are enriched in the promoter regions of diverse eukaryotes, and significantly overlap eQTLs in *Arabidopsis*. We also demonstrate how our data can be used to predict the specific TFs whose binding would be altered by a human disease risk allele. To house the data and the resulting inferences, we have created the Cis-BP database (catalog of inferred sequence binding preferences), freely available at http://cisbp.ccbr.utoronto.ca.

RESULTS

PBM Data for >1,000 Diverse Eukaryotic TFs

We sought to examine the relationship between the DBD AA sequence identity and the DNA sequence specificity of any two proteins and to simultaneously broadly survey the sequence specificity of eukaryotic TFs. To identify TFs, we first scanned the AA sequences for each of 81 different types of DBDs for which there is an available Pfam model (Weirauch and Hughes, 2011), using known or predicted proteins in 290 sequenced eukaryote genomes. We identified a total of 166,851 putative TFs that fit these criteria. From these, we selected 2,913 individual TFs to analyze, using several different criteria aimed at achieving the goals of our study, including a relatively even balance among DBD classes and species, a survey of different levels of sequence identity among proteins, and a deeper focus on several model organisms and abundant DBD classes (see Experimental Procedures). Figure S1, available online, depicts the overall scheme.

We analyzed each TF using PBM assays, following procedures previously described (Berger et al., 2006; Weirauch et al., 2013). The PBM technique can be summarized as follows: a GST-tagged DNA-binding protein is "hybridized" to a double-stranded DNA microarray, and subsequent addition of a fluorescently tagged antibody reveals the DNA sequences that the protein has bound and to what degree. Each PBM contains a diverse set of ~41,000 35-mer probes, designed such that all possible 10-mers are present once and only once; thus, all nonpalindromic 8-mers are present 32 times, allowing for a robust and unbiased assessment of sequence preference to all possible 8-mers. Values for individual 8-mers are typically given as both *E* scores (that represent relative rank of intensities and range from −0.5 to + 0.5) (Berger et al., 2006) and *Z* scores (that scale approximately with binding affinity) (Badis et al., 2009). PBMs also allow derivation of position weight matrices (PWMs) up to 14 bases wide (Badis et al., 2009; Berger et al., 2006; Mintseris and Eisen, 2006; Weirauch et al., 2013), as well as International Union of Pure and Applied Chemistry (IUPAC) consensus sequences (Ray et al., 2013). PWMs can be represented as sequence logos (Schneider and Stephens, 1990) that are typically taken as synonymous with "motifs." Here, we report 8-mer scores, PWMs, and consensus sequences, and use whichever is best suited to individual analyses. For example, while motifs are well suited for visualization of sequence preferences, and convenient for scanning longer sequences, they can underestimate reproducibility of experiments due to the introduction of uncertainties in the process of PWM derivation (Weirauch et al., 2013; Zhao and Stormo, 2011).

We analyzed each of the 2,913 proteins using two different PBM arrays, designated "ME" and "HK" after their designers (Lam et al., 2011). Of these, 1,032 (encompassing 1,017 different TFs) yielded data that satisfy our stringent success criteria, including *E* scores >0.45 on both arrays, and agreement in both 8-mer data and motifs between the two arrays (Berger et al., 2008; Weirauch et al., 2013) (see Experimental Procedures). The distribution of the 1,032 proteins over 131 species and 54 DBD classes is summarized in Figure 1A. Many values in this matrix are zero because the majority of DBD classes are

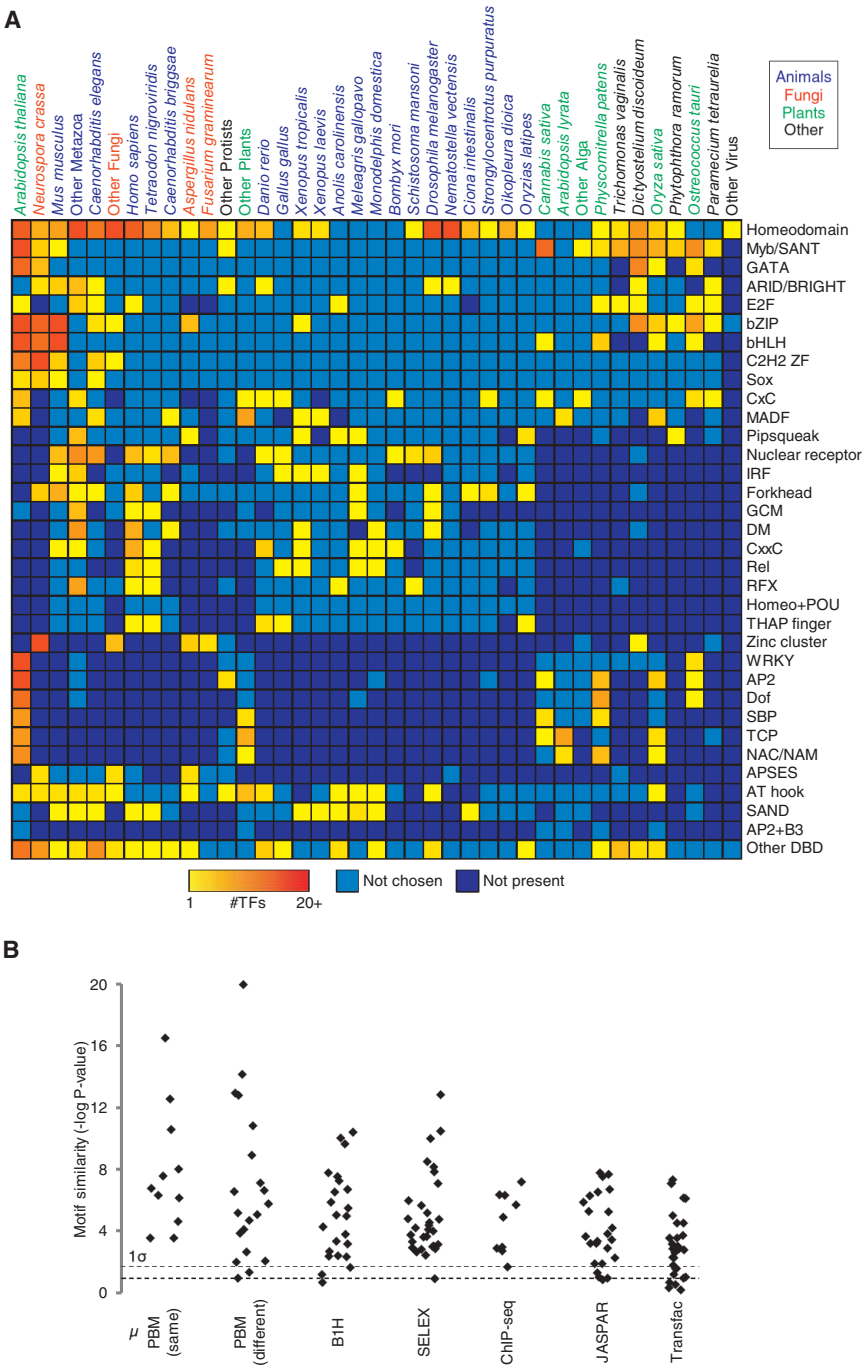


Figure 1. Overview of the Motif Data Set

(A) TFs characterized in this study by species and DBD class. TFs with multiple DBD classes are indicated with a "+" (e.g., AP2+B3). DBD classes and species containing fewer than five members are grouped into "Other." Species are ordered by the total number of TFs with characterized motifs. (B) PBM-derived motifs are similar to previously characterized motifs. We compared new PBM-derived motifs to previously determined motifs for the same TF. p values were calculated using the TomTom PWM similarity tool (Tanaka et al., 2011), with Euclidean distance and default parameter settings. Dashed lines indicate mean (bottom) and mean + 1 SD (top) of p values obtained from 10,000 randomly selected PWM pairs. "PBM (same)" and "PBM (dif)" indicate PBMs from other studies performed using the same, or different array designs as this study, respectively.

See also Figure S1 and Tables S1, S2, and S6.

only present in certain subsets of species (e.g., plant-specific TFs) (Weirauch and Hughes, 2011). PBM failures may be due to any of several causes, including protein misfolding, requirement for cofactors, or bona fide lack of sequence-specific DNA binding activity.

For 123 of the 1,017 examined TFs, there are previously described motifs. In most cases, the new motifs we obtained are highly similar to motifs compiled from the literature (JASPAR and Transfac), derived from PBMs in other studies, derived from other technologies (B1H and HT-SELEX), or derived from ChIP-seq experiments (Figure 1B; sources provided as Table S1; full motif comparisons provided as Table S2). Importantly, a large majority of the proteins we analyzed (894/1,017) had no previous binding data, and among these, roughly half yielded a motif that is highly different from any previously known motif (see Experimental Procedures). For several DBD classes (CG-1, CxC, GRAS, LOB, and Storekeeper) we characterized sequence specificity for the first time. CxC domains, for example, span a wide range of organisms encompassing plants, animals, and protists and recognize variations of a unique, largely conserved TTTCGAAA motif.

Inference of TF Sequence Specificity Using Degree of Identity in the DNA-Binding Domain

We next used the PBM data to ask how well the percent protein sequence identity in the DBD between two proteins correlates with similarity in their DNA sequence preferences. As a measure of DNA sequence preference similarity, we calculated the overlap in high-scoring 8-mers (e.g., *E* score >0.45) (see Experimental Procedures), as the *E* scores are on a uniform scale that facilitates direct comparison. The boxplots shown in Figure 2 and Data S1 illustrate that there are different characteristic relationships between DBD identity and 8-mer overlap for different DBD classes. However, virtually all of the plots display a sigmoidal appearance, such that a particular threshold defines the %AA identity over which the majority (75% or more) of protein pairs have very similar DNA binding preferences (i.e., are

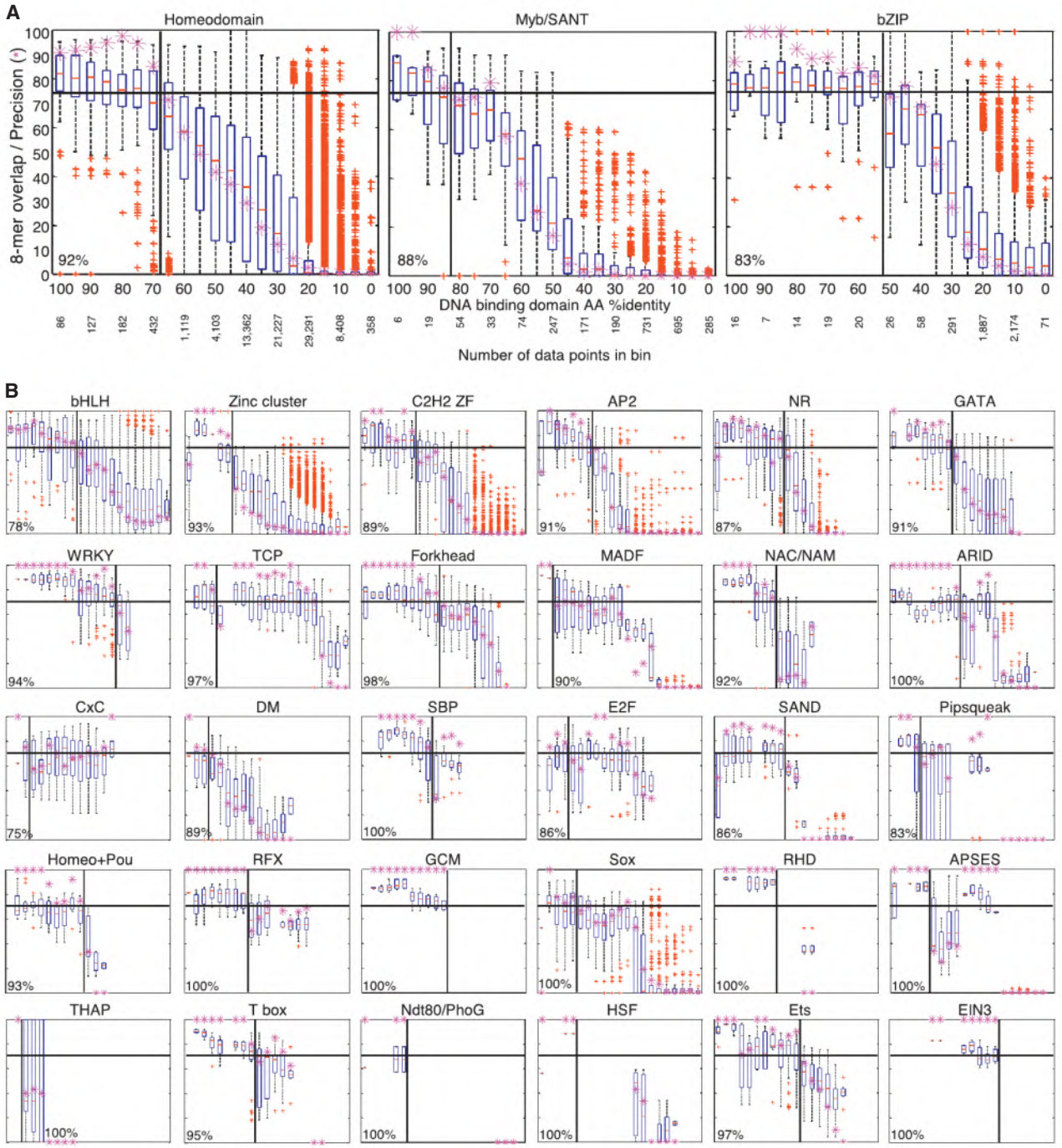


Figure 2. Motif Inference Thresholds by DBD Class

(A) Relationship between similarity in DBD AA sequence and DNA sequence preferences. Boxplots depict the relationship between the %ID of aligned AAs and % of shared 8-mer DNA sequences with *E* scores exceeding 0.45, for the three DBD classes with the most PBMs in this study. Red lines indicate the median value of each bin. Edges of the blue boxes indicate the 25th and 75th percentiles. Whiskers extend to the most extreme data points not considered outliers. Red crosses indicate outliers. %ID bins range from 0 to 100, of size 10, in increments of five. Bottom: number of DBD pairs in each bin. Pink asterisks indicate the precision of the corresponding bin (i.e., the fraction of protein pairs with 8-mer similarity at least as high as the 25th percentile of replicates). Horizontal line indicates the 75% precision line used to choose the inference threshold. Vertical lines indicate AA %ID threshold (i.e., the point before the pink asterisks drop below the horizontal line). Percentage in lower left corner indicates cross validation success rate.

(B) Relationship for all DBD classes. Boxplots for all DBD classes for which we could establish an inference threshold, depicted as in (A). DBD classes are ordered by the number of TFs characterized in this study.

NR, nuclear receptor. See also Figures S2 and S6.

at least as similar as the 25th percentile of replicates of the same protein—see Experimental Procedures) (Figure 2). For homeodomains, this threshold is 70% AA identity (Figure 2A), almost exactly what we previously reported (65%) (Berger et al., 2008).

Strikingly, such a threshold can be drawn for virtually every DBD class (Figure 2 and Data S1). Moreover, exactly the same trends and thresholds are observed in a leave-one-out cross validation prediction framework, in which the high-scoring 8-mers are predicted to be identical to those of the protein with the closest %AA identity (see Experimental Procedures). Using this framework, we estimate that ~89% of predicted sequence preferences (hereafter, "inferences") above the thresholds shown in Figure 2 are as accurate as experimental replicates (Data S1), with similar precision obtained whether inferences are derived from only orthologs (88%) or only paralogs (90%). The distribution of DBD AA similarities among the 1,017 proteins we analyzed is nearly identical to the distribution among all sequenced eukaryotes (data not shown), so the same numbers can be expected for de novo predictions. Furthermore, comparisons among PWMs (i.e., motifs) confirm these thresholds, despite the potential variation contributed by motif derivation (Weirauch et al., 2013; Zhao and Stormo, 2011). To illustrate this concept, Figure 3 depicts PBM-derived motifs obtained for the Myb/SANT family, grouped according to the relationship of their DBD AA sequences. Darkly shaded regions, which invariably contain nearly indistinguishable motifs, indicate groups of TFs with DBD similarity exceeding the 87.5% threshold for this family.

We conclude that TFs with DBD AA sequence identity above these thresholds will typically have very similar sequence specificity. As previously proposed, this presents a simple approach for broadly predicting the sequence preferences of TFs (Alleyne et al., 2009; Berger et al., 2008; Bernard et al., 2012; Noyes et al., 2008): the 8-mer data, motifs, and consensus sequences can be directly transferred from the nearest protein above the threshold for which data are available. The data presented here encompass 37 DBD classes with sufficient data (i.e., comparisons in each of the bins in the box plots) to produce thresholds, together representing ~85% of all known eukaryotic TFs. Until more data are available, we propose that a threshold of 70% AA identity (the mean, median, and mode across all DBD classes) can be applied to the remaining DBD classes (boxplots in Figure S2 show the aggregate of all of the classes with no threshold).

Cis-BP: A Catalog of Direct and Inferred Sequence Binding Preferences

Using the DBD identity thresholds generated above, we globally assigned probable motifs (and other sequence preference data, if available) to TFs from 290 eukaryotic genomes. To do this, we supplemented the PBM data collected in this study with 4,234 additional published motifs (674 from PBMs and 3,560 from other sources—see Table S1) derived from 1,850 different proteins, mainly in human (623 proteins), mouse (409), fly (316), or yeast (216). Altogether, there are experimentally determined motifs for only 1.7% (2,750) of the 166,851 unique TFs within these genomes. Using the thresholds determined here, it is possible to infer a motif for roughly one-third of all TFs encoded by sequenced eukaryotes, bringing the total (known + inferred) to 57,165. Figures 4A and 4B summarize the coverage by DBD class and by species (see Table S3 for all DBD classes and species). Lineages that benefit most from the inference scheme include vertebrates, plants, fungi, and insects, which contain many orthologs conserved in the model species analyzed most heavily. For example, the motif collection for zebrafish (*Danio rerio*), which largely consists of inferred motifs, is as complete as that of mouse and human (Figure 4B).

To facilitate use of the known and inferred motifs by the scientific community, we created a database called catalogue of inferred sequence preferences of DNA-binding proteins (Cis-BP) (http://cisbp.ccbr.utoronto.ca). In addition to the new experimental data reported here, Cis-BP contains comprehensive 8-mer binding scores, position weight matrices, and IUPAC consensus motifs from publicly available sources (JASPAR (Portales-Casamar et al., 2010), Transfac (public data only) (Matys et al., 2006), FlyFactorSurvey (Zhu et al., 2011), FactorBook (Wang et al., 2013), and data from 674 PBM experiments taken from other studies (compiled in UniPROBE) (Newburger and Bulyk, 2009).

In-Vitro-Derived Motifs Predict ChIP-Seq Peaks

To examine the relationship between in-vitro-defined motifs (both measured and inferred) and in vivo binding sites, we asked how well the motifs in Cis-BP predict in vivo binding, based on ENCODE human ChIP-seq data (see Experimental Procedures). To gauge the ability of a motif to discriminate between real ChIP peaks (positives) and peak sequences permuted using an algorithm that maintains all dinucleotide frequencies (negatives), we used the area under the receiver operating characteristic (AUROC) summary statistic, in which perfect discrimination between positives and negatives scores 1.00 and random guessing scores 0.50. Nearly all (111) of the 114 PBM-derived motifs achieved AUROC scores exceeding the 0.50 random expectation level in at least one ChIP data set (Figure 5A and Table S4). Strikingly, over one-third of the motifs (43 of 114), in a variety of DBD classes and cell types, achieve AUROCs exceeding 0.90 in at least one ChIP data set, including GABPA in H1-hESC cells (Ets family, AUROC = 0.99), USF2 in GM12878 cells (bHLH, 0.97), and FOS in K562 cells (bZIP, 0.97). Motifs inferred from in vitro data from related proteins achieve AUROC values similar to those obtained using in vitro motifs from exactly the same protein that was ChIPped (Figure 5A). Overall, PBM-derived motifs are equally accurate as those derived from other sources; the mean AUROCs across the 19 TFs with at least one PBM-derived motif and one Transfac motif are 0.834 and 0.827, respectively (Figure 5B and Table S4), and 0.826 versus 0.831 for the 14 TFs with both PBM and HT-SELEX-derived motifs (Figure 5C and Table S4). The small number of cases in which motifs derived from these other sources performed better than PBM motifs appear to correspond to multimeric binding that was not detected in the PBM assays. From this analysis, we conclude that there is generally no fundamental discrepancy between motifs obtained from different assays and between the in vivo and in vitro sequence preferences of TFs.

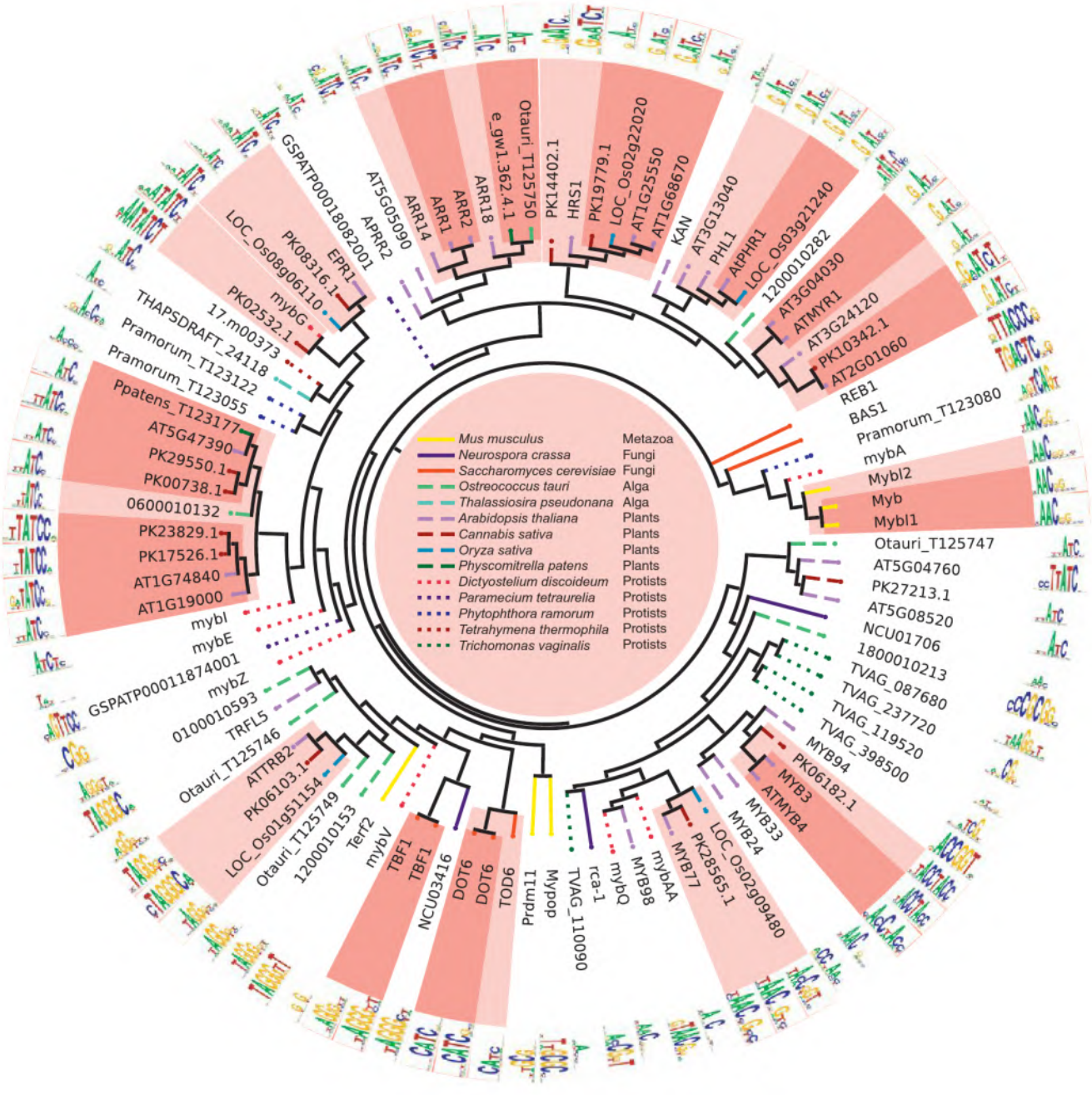


Figure 3. Overview of Myb/SANT Family Motifs

PBM-derived motifs from the Myb/SANT family (84 from this study, 13 from other studies) are shown. Tree reflects the percent of identical AAs after alignment. Dark shading, 87.5% AA identity (standard inference threshold); light shading, >70% AA identity (relaxed inference threshold). TBF1 and DOT6 each have two motifs because they were examined in two different studies.

Motifs Display Positional Bias in Promoters across Eukaryotes

Enrichment and positional bias of TF motifs in promoters has been reported in several model species and human (FitzGerald et al., 2004; Lee et al., 2007; Mariño-Ramírez et al., 2004; Ohler et al., 2002), and in some cases TFs appear to be involved in determining promoter identity (de Boer et al., 2014; Megraw et al., 2009). However, it is unknown whether this is a general property of eukaryotic promoters. We found that genomes from vertebrates, plants, fungi, and protists all displayed enrichment of binding sites for their TFs just upstream of the TSS, relative to permuted sequences and also relative to unrelated (control) motifs taken from other lineages (Figure 6). The enrichment involved both directly determined and inferred motifs

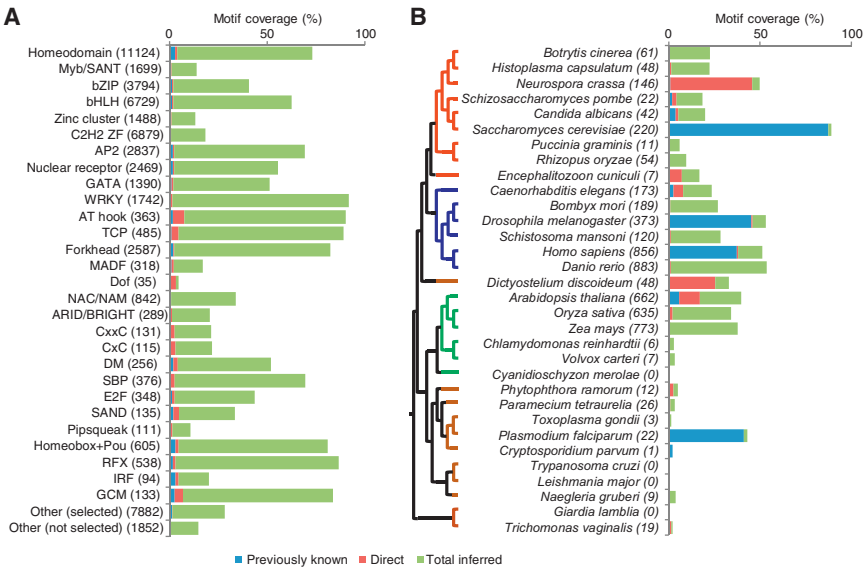


Figure 4. TF Motif Coverage

TFs with multiple protein isoforms are counted as a single gene.

(A) Motif coverage by DBD class. DBD classes sorted top to bottom by number of TFs characterized in this study. Those with fewer than eight proteins characterized in this study are grouped into "Other." "Other (selected)" indicates DBD classes selected for characterization in this study. "Other (not selected)" indicates DBD classes not characterized here. "Direct" includes those experimentally characterized in this study, but not previously known. "Total inferred" excludes those experimentally characterized in this or previous studies.

(B) Motif coverage by species. Tree at left, phylogenetic relationships between organisms (Baldauf et al., 2000).

See also Table S3.

(Figure 6). Strikingly, the TF classes with motifs enriched in promoters differed between lineages of organisms (Figure S3), indicating that these trends either arose independently in different lineages, or have evolved considerably in most of them. Furthermore, we found little to no correspondence between overall motif enrichment and motif sequence composition (i.e., overall GC-content) (Table S5), and the enrichment was observed whether the promoters were characteristically GC-rich or AT-rich (i.e., predicted to be nucleosome favoring or disfavoring, respectively). We observed similar trends using a reduced set of nonredundant motifs for each organism (Figure S4), showing that the phenomenon is not due to expansion of a small number of TFs families that bind promoters. These observations indicate that, across eukaryotes, the region just upstream of the TSS typically is enriched for binding sites that may function in either promoter definition or gene regulation.

Arabidopsis Expression Quantitative Trait Loci SNPs Are Enriched for TF Binding Sites

Identifying causal genetic variants and their mechanisms of action is a fundamental challenge in association studies. As our data greatly expand the motif collection in the model plant *Arabidopsis*, we examined a data set of expression quantitative trait loci (eQTLs) defined from *Arabidopsis* genomic sequences and matched seedling RNA sequencing (RNA-seq) taken from 19 strains (Gan et al., 2011). Among highly significant eQTLs found within 1 kb upstream of a TSS, there was a striking enrichment (>3-fold for the SNPs with strongest association) for overlap with TF motif matches (Figure 7A), strongly suggesting that these SNPs impact transcription factor binding. As an example, Figures 7B and 7C show a SNP in the promoter of the *AT5G47250* gene, in which loss of a potential binding site for the VNI2 TF in the "A" allele correlates almost perfectly with a dramatic increase in *AT5G47250*'s expression level, consistent with the well-characterized role of VNI2 as a transcriptional repressor (Yamaguchi et al., 2010; Yang et al., 2011). Both *VNI2* and *AT5G47250* were associated with the plant defense response to the pathogenic oomycete *H. arabidopsidis* in a recent genome-wide association study examining 107 different phenotypes (Atwell et al., 2010). Taken together, these results suggest that VNI2 represses the expression of *AT5G47250* in a pathogen response context, a mechanism dependent on the specific SNP present in the *AT5G47250* promoter region. We expect that the Cis-BP motif collection will be useful for similar analyses in other organisms: eQTL analysis can be performed in virtually any species for which there are multiple strains (or individuals) and does not require well-developed genetic systems. In fact, in this analysis, a similar level of enrichment was observed using only a set of 65 motifs inferred from organisms other than *Arabidopsis thaliana* (Figure S5), suggesting that the motif data need not be derived from the organism in question (see also below).

Identification of TFs Affected by Disease-Associated Genetic Variants

Recent analyses indicate that between 85% and 93% of disease- and trait-associated variants are located in noncoding regions (Hindorff et al., 2009; Maurano et al., 2012), suggesting that many might alter TF binding events. However, the identification of specific TFs whose binding might be affected by a given variant remains a challenging and laborious task. We devised a system that utilizes all available PBM data to produce a ranked list of human TFs whose binding might be affected by any given genetic variant (see Experimental Procedures). To examine the utility of this system for analyzing human disease-associated variants, we collected a set of 15 SNPs whose alleles have been experimentally demonstrated to affect the binding of a specific TF. One of these SNPs affects two TFs, bringing the total of analyzed TF/SNP pairs to 16. Strikingly, in ten of these 16 cases, this procedure ranked the correct TF (or a highly related TF from the same DBD class) in the top five and often number one (Figure 7D and Data S1). We note that most of the novel high-ranking TFs we identify have likely never been experimentally examined and thus might also represent bona fide cases.

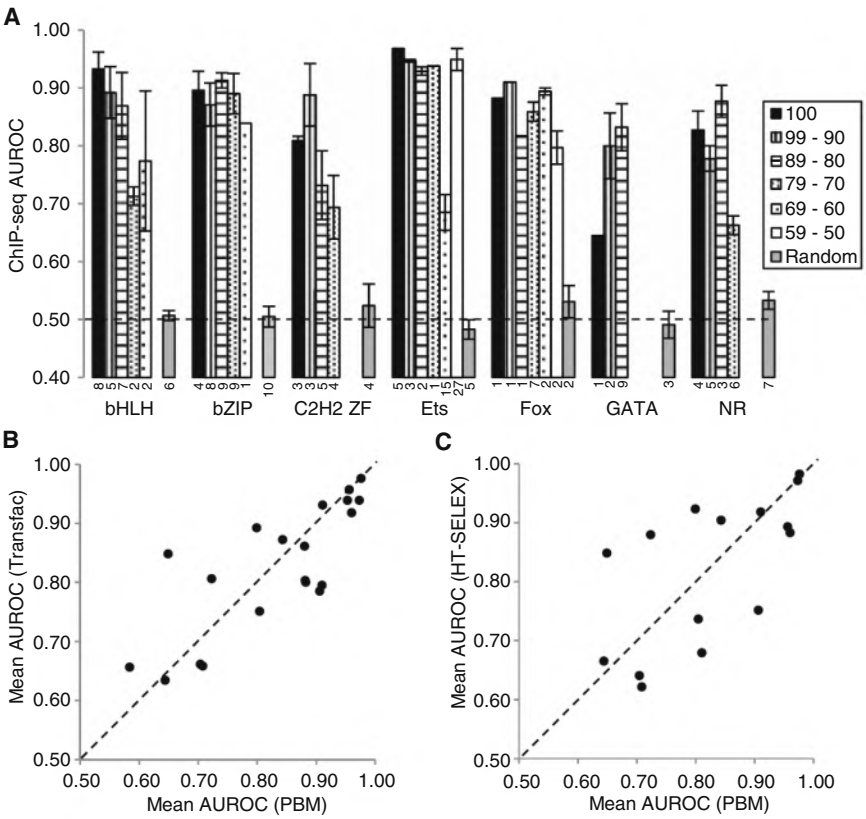


Figure 5. PBM-Derived Motifs Identify In Vivo TF Binding Locations

(A) AUROC analysis, showing ability of directly determined and inferred motifs to distinguish ChIP-seq peak sequences from scrambled sequences. We identified TFs with available ENCODE ChIP-seq data that also have PBM data available either for that TF, or for related TFs (based on the inference threshold for the DBD class). We then gauged the ability of the PBM-derived motifs to distinguish real ChIP peaks from scrambled sequences (maintaining all dinucleotide frequencies) using the AUROC (see Experimental Procedures). For each DBD class, results are binned by DBD %AA ID (key indicated at upper right). Numbers below each bar indicate the count in each bin. Error bars indicate SE. "Random" indicates results obtained with a randomly assigned, unrelated TF motif. Fox, Forkhead box. NR, nuclear receptor. Figure S7 shows results obtained using an alternative null model.

(B) Comparison of AUROC for PBM-derived motifs and literature-derived motifs. We identified TFs with ENCODE ChIP-seq experimental data that also have both Transfac and PBM-derived motifs available. For each TF, we calculated the best AUROC obtained by any PBM or any Transfac motif on any of the ENCODE cell line ChIP experiments for that TF. For TFs with multiple motifs from the same source, the plot shows the mean AUROC across the motifs.

(C) PBM-derived motifs versus HT-SELEX-derived motifs. Same as for (B), but including only TFs with motifs available both from PBMs and a recent HT-SELEX study (Jolma et al., 2013).

See also Table S4.

In one example, a recent study implicated rs554219 in estrogen-receptor-positive breast cancer tumors and used a series of experiments to predict and eventually establish that this SNP causes the differential binding of two Ets family TFs, ELK4 and GABPA (French et al., 2013). Because rs554219 does not overlap ChIP-seq binding peaks from ENCODE or other sources, ELK4 and GABPA were identified using a series of EMSA competition experiments involving known TF binding sites, a laborious process involving substantial guesswork. Our automated computational procedure correctly ranked ELK4 number one and GABPA number two out of all human TFs (their EMSA data demonstrate strong and moderate differential binding of ELK4 and GABPA, respectively). The ELK4 prediction is based on an inference from the pufferfish *Tetraodon nigroviridis* ELK4 protein (81% AA ID to human ELK4), further illustrating the utility of cross-species inferences. Moreover, differential binding of ELK4 to the alleles of rs554219 can also be predicted using data inferred from *Drosophila melanogaster* Ets21C (57% AA ID to human ELK4) and *Caenorhabditis elegans* F19F10.1 (only 33% AA ID to ELK4). Thus, our system (available at http://cisbp.ccbr.utoronto.ca/TFTools.php) can accurately predict the specific TFs whose binding is affected by risk alleles of disease-associated SNPs, even when using PBM data inferred from distantly related organisms.

DISCUSSION

We anticipate that the new data collected here—as well as the inferred motifs across eukaryotes—will be an invaluable resource and knowledge base for functional genomics and analysis of gene regulation. In addition to the data itself, the Cis-BP database also contains web-based interfaces to tools for scanning DNA sequences for putative motifs, reporting the TFs with motifs similar to a given motif, predicting the motif recognized by a given TF based on its DBD AA sequence, and identifying TFs that will bind differentially to two different DNA sequences (e.g., disease risk and nonrisk alleles). Cis-BP will enable dissection of regulatory mechanisms from expression data even in the vast majority of species for which there are currently no genetic tools.

The analyses here present a strategy to rapidly populate TF motif collections across the eukaryotes. As we previously proposed for RNA-binding proteins (Ray et al., 2013), targeting members of the largest groups of uncharacterized proteins for experimental analysis will allow the largest number of inferred motifs to be obtained. Motif inference based on DBD identity alone is only a first approximation, but it is remarkably cost effective: the analyses described here indicate that motifs can be inferred for 34% of all TFs, using data from only 1.7%. We extrapolate that 1,032 additional successful experiments

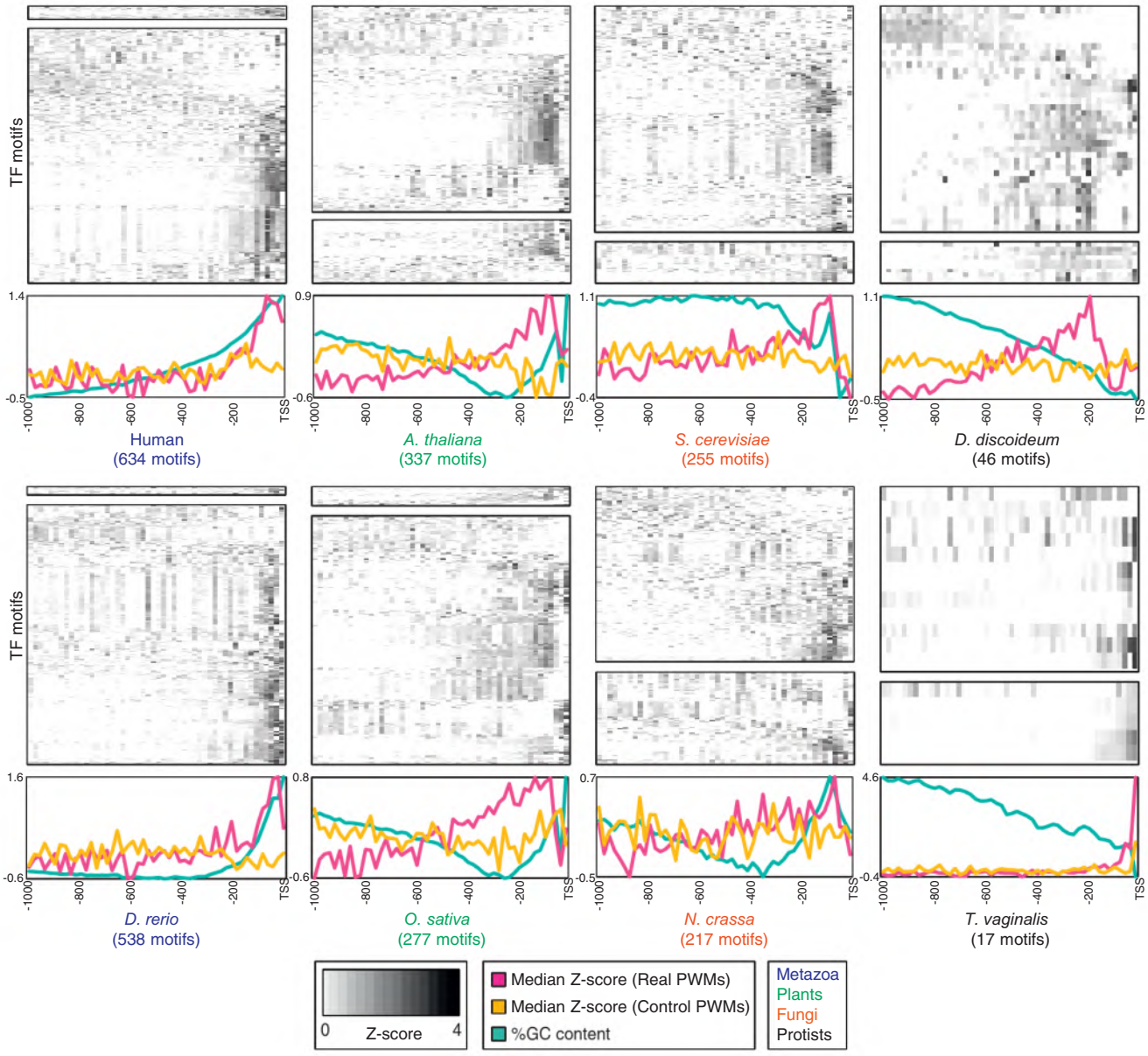


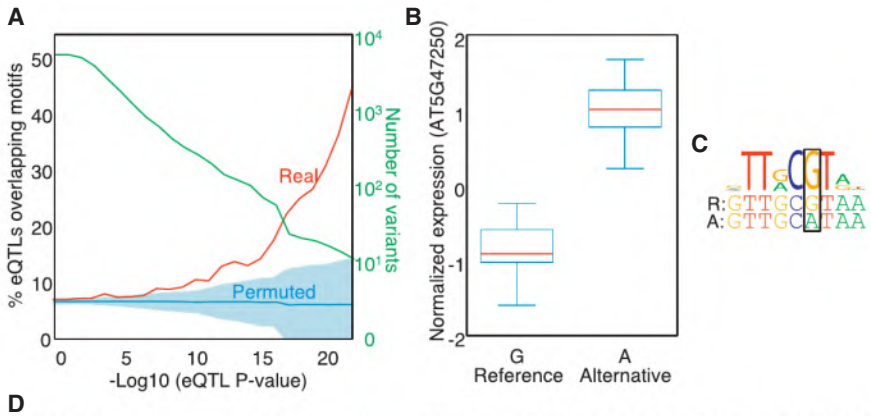
Figure 6. Positional Bias of Motif Matches in Eukaryotic Promoters

PBM-derived PWMs (direct, top heatmap; inferred, bottom heatmap) scored in 20 bp bins, normalized to dinucleotide-permuted controls, averaged across all promoters, and displayed as *Z* scores (see Experimental Procedures). Each row in the heatmap corresponds to one PWM. Rows were clustered using hierarchical clustering (Pearson correlation, average linkage). Summary plots at the bottom indicate the median *Z* score, taken across all PWMs from the indicated species ("Real PWMs"), or across a set of PWMs from unrelated lineages ("Control PWMs") (see Experimental Procedures).

See also Table S5 and Figures S3 and S4.

(<0.5% of all TFs and a number identical to that in this study)—one from each of the largest groups of orthologs and paralogs with no known motifs—would increase coverage across the eukaryotes from 34% to 48%.

The inference scheme described here relies on the high degree of conservation among DBDs. Indeed, our analyses confirm the "deep homology" that has been described for metazoan developmental processes and the TFs that regulate them (e.g., homeodomains) (Berger et al., 2008; Carroll, 2008; Noyes et al., 2008) and furthermore indicate that deep homology is a property of the sequence preferences of many TFs in all eukaryotic kingdoms. Our initial analyses (data not shown) suggest that many motifs likely date to the base of metazoans, land plants, angiosperms (flowering plants), or euteleostomi (bony vertebrates), consistent with well-established TF expansions in these lineages (de Mendoza et al., 2013; Weirauch and Hughes, 2011).



D

Disease	SNP	Known TF(s)	Event	Best rank (related)	Best rank (exact)	DBD class
Breast cancer	rs554219	ELK4,GABPA	Loss	1,2	1,2	Ets
COPD	rs1542725	SP3	Loss	1	N/A	C2H2 ZF
Alpha Thalassemia	SNP195	GATA1	Loss	1	N/A	GATA
Prostate cancer	rs1859961	FOXA1	Loss	1	N/A	Forkhead
Breast cancer	rs7895676	CEBPB	Gain	1	4	bZIP
Colorectal cancer	rs6983267	TCF7L2	Loss	2	3	Sox
Breast cancer	rs35054928	E2F1	Loss	3	N/A	E2F
Lupus	rs57095329	ETS1	Loss	3	-	Ets
Myocardial infarction	rs12740374	CEBPA	Loss	5	7	bZIP

Figure 7. Overlap of Predicted TF Binding Sites with *cis*-eQTLs

(A) Number and percentage of *Arabidopsis cis*-eQTLs overlapping motifs, as a function of eQTL significance. Shaded region indicates one standard deviation in the expected distribution (see Experimental Procedures).

(B) A *cis*-eQTL affecting the expression of the *AT5G47250* gene. Boxplots (see Figure 2 legend) indicate the median normalized gene expression level for each allele of the *cis*-eQTL. "Reference" indicates the allele present in the *Arabidopsis* reference genome assembly.

(C) The same *cis*-eQTL "breaks" a putative binding site for the VNI2 transcriptional repressor. Sequence logo depicts the DNA-binding motif we obtained for VNI2. Sequences below indicate the reference (top) and alternative (bottom) alleles of the *cis*-eQTL SNP (boxed) and its flanking bases.

(D) Prediction of human TF binding events altered by disease risk alleles. We created a method for using PBM data to predict TFs whose binding is affected by disease-associated genetic variants and applied it to 16 known examples. Shown here are the ten cases in which we ranked the correct TF (column labeled "exact") or a highly related TF from the same DBD class (column labeled "related") within the top five TFs. The "Event" column indicates whether the risk allele results in a "Loss" or "Gain" of binding of the TF. N/A, indicates that PBM data are not available for the corresponding TF; –, indicates that the TF did not receive a rank because both alleles had an *E* score >0.45.

See also Figure S5.

Despite widespread conservation, TF repertories do change over evolutionary time, and these changes likely shape eukaryotic evolution (de Mendoza et al., 2013). TFs that tend to diversify, such as the large metazoan C2H2 class (Stubbs et al., 2011), will present an ongoing challenge to a complete characterization of eukaryotic TF motifs. In other lineages, other classes of TFs have expanded and diversified, including the nuclear receptor class in *C. elegans* (Maglich et al., 2001), the zinc cluster/GAL4 class in fungi (Shelest, 2008), and several classes in plants (Lang et al., 2010). The data described here confirm that the sequence specificities of at least some of these factors have also diversified. Mapping recognition codes represents an alternative approach to more complete cataloguing of TF motifs, and the data presented here provide many new examples for the study of TF-DNA recognition. Exceptions to the simple AA similarity rules described here should also be informative regarding mechanisms of sequence recognition, because they will identify AA residues critical for DNA sequence specificity (De Masi et al., 2011; Noyes et al., 2008). Ongoing efforts to further the collective knowledge of TF binding specificities will greatly advance our understanding of TF-DNA interactions, as well as our ability to interpret the function of DNA sequences, including understanding the functional impact of natural genetic variants in human and other species.

EXPERIMENTAL PROCEDURES

Full details are provided in the Extended Experimental Procedures.

Data Availability

PBM data are available in the Gene Expression Omnibus (GEO) database under accession number GSE53348. PBM data, clone information, and other data from analyses carried out in this study are available on the project web site: http://hugheslab.ccbr.utoronto.ca/supplementary-data/CisBP/. Additional data (including 8-mer scores, PWMs, sequence logos, and information on TFs) are found on the CisBP web server (http://cisbp.ccbr.utoronto.ca/).

Selection, Cloning, and PBM Analysis of TFs

We compiled the predicted proteomes of 290 eukaryotic organisms from a variety of sources and supplemented them with an additional 49 known TFs from organisms without fully sequenced genomes. We scanned all protein sequences for putative DNA-binding domains (DBDs) using the 81 Pfam (Finn et al., 2010) models listed in Weirauch and Hughes (2011) and the HMMER tool (Eddy, 2009). Each protein was classified into a family based on its DBDs and their order in the protein sequence. We selected 2,913 individual TFs to analyze, using several different criteria, including a relatively even balance among DBD classes and species, a survey of different levels of sequence identity among proteins, and a deeper focus on several model organisms and abundant DBD classes. For most constructs, we designed primers to clone the region encompassing all DBDs plus the 50 flanking endogenous AAs on either side (or until the termini of the protein) by conventional PCR methods into one of a panel of T7-GST vectors for expression in *E. coli* (referred to hereafter as "plasmid constructs"). PBM laboratory methods were identical to those described in Lam et al. (2011) and Weirauch et al. (2013). Each plasmid was analyzed in duplicate on two different arrays with differing probe sequences (denoted "ME" and "HK"). Calculation of 8-mer *Z* and *E* scores was performed as previously described (Berger et al., 2006). To obtain a single representative motif for each protein, we used a procedure similar to a recent study from our group in which we generated motifs for each array using four different algorithms and chose the best-performing single motif based on cross-replicate array evaluations (Weirauch et al., 2013).

Inference Scheme

We established a separate inference threshold for each DBD class. We first aligned the DBD sequences of all constructs within a DBD class using clustalOmega (Sievers et al., 2011). We then calculated the AA %ID for all construct pairs (i.e., the number of identical AAs in the alignments). Within each DBD class, we grouped all PBM construct pairs into bins, based on AA %ID. We used overlapping bins of size 10, ranging from 0 to 100, increasing by 5. We calculated the precision of each bin by comparing the DNA sequence preferences obtained from all characterized protein pairs contained in the bin. We quantified the similarity of the DNA sequence preferences of two proteins as the fraction of shared high-scoring 8-mers. We considered a prediction to be correct only if this fraction exceeded the value obtained at the 25th percentile of experimental replicates (i.e., the fraction of shared 8-mers between ME and HK arrays for the same protein). The proportion of predictions scored as correct (i.e., precision) for each bin of each DBD class is shown as magenta stars in Figure 2 and Data S1.

We chose inference thresholds for each DBD class based on the precision scores of each AA %ID bin. Because we used the 25th percentile threshold to define precision, we would expect a precision of 0.75 or higher in each AA %ID bin. We therefore chose an inference threshold for each DBD class by identifying the final AA %ID bin before precision drops below 0.75 (vertical bars in Figure 2). Similar thresholds were obtained regardless of the *E* and *Z* score 8-mer thresholds used and also regardless of the replicate overlap percentile considered (i.e., 25th percentile, requiring 0.75 precision or 20th percentile, requiring 0.80 precision) (Figure S6). The final threshold for a DBD class was chosen as the median threshold across the eight 8-mer similarity measures (see Figure S6 and Data S1). We found this scheme to be appropriate for most DBD classes (all of which are depicted in Figure 2). For three DBD classes (IRF, CXXC zinc fingers, and Dof zinc fingers), we could not establish a threshold—these therefore received a threshold of 100%. We used a threshold of 40% for AT-hook TFs, which recognize AT-rich sequences, based on manual inspection of the data (see Data S1). For the remaining classes, with suggestive but insufficient data, we chose a threshold of 70%, which is the mean, median, and mode threshold across all DBD classes.

We used the AA %ID of all pairs of proteins to infer motifs, 8-mer scores, and consensus sequences within each DBD class by simple transfer (i.e., aligning the DBD sequences of all proteins and all constructs in a given DBD class, as described above, and calculating the AA %ID of each protein with each construct). We also evaluated the effectiveness of our inference scheme in a leave-one-out cross validation framework, in which the PBM data for each characterized protein was held out and compared to the PBM data of its nearest neighbor (i.e., the characterized protein with highest AA %ID), using a similar scoring scheme to that used to calculate the precisions.

Comparison to ChIP-Seq Data

We calculated AUROC scores on real and permuted (maintaining dinucleotide frequencies) ChIP-seq peak sequences following Weirauch et al. (2013). We obtained ENCODE consortium human ChIP-seq data from the UCSC Genome Browser (Rosenbloom et al., 2012). For each ChIP experiment, we extracted the top 500 scoring peak region sequences and scored them (and the permuted sequences) using all direct and inferred PWM models for the given TF. For each PWM/experiment pair, we then calculated the AUROC using these sets of 500 positives and 500 negatives. Similar results were obtained using a negative set consisting of ChIP-bound peaks from unrelated TFs with matched GC-content (Figure S7).

Positional Bias of Motifs in Eukaryotic Promoters

We obtained the 1,000 bases upstream of transcription start sites (or, if unavailable, translation start sites) and scored the PWMs of each organism at each position. We then placed the resulting scores into 20 bp bins, summed the scores for each bin, and took the average across all promoters for the given species for each bin. To correct for mono- and dinucleotide biases, we also scored shuffled promoter sequences, which were created by shuffling the sequences within each 20 bp bin (while maintaining dinucleotide frequencies). For each PWM, we then calculated the ratio of each bin's real score relative to the score of the shuffled sequence. The resulting ratios were then normalized across all bins for the given PWM using a standard *Z* score transformation.

We also calculated *Z* scores for a negative control set of TF PWMs for each organism, consisting of a collection of random motifs from species in other clades that were unrelated to any PWM from the given species.

Arabidopsis eQTL Analysis

We used a publicly available data set (Gan et al., 2011) containing genome-wide RNA-seq variance-stabilized expression levels (Huber et al., 2002) taken from 19 strains of seedling *Arabidopsis thaliana* and matching genome sequences. We identified matches to each *A. thaliana* PWM within the 1,000 bases upstream of each TSS. We calculated the percentage of genetic variants that affect these putative binding sites, as a function of the *cis*-eQTL p value of the variant (red line, Figure 7). We also created a null distribution (blue line and blue shaded region, Figure 7) to exclude the possibility that the observed percentages might solely be due to the higher density of TF binding sites in promoter regions.

Human Disease SNP/TF Analysis

We devised a system for utilizing our collection of PBM data to identify candidate human TFs whose binding might be affected by the allelic sequences of genetic variants. In this system, we score each variant (along with its flanking genomic bases) using 8-mer *E* scores taken from the 3,132 PBM experiments contained in our database. For each PBM experiment, we identify the highest scoring 8-mer *E* score attained by any of the risk allele sequences (E_{risk}) and the highest attained by any nonrisk allele ($E_{nonrisk}$). We then identify all PBM experiments where only one of E_{risk} and $E_{nonrisk}$ has an *E* score value exceeding 0.45 (values above this threshold will likely be strongly bound by the given TF (Berger et al., 2008)) and map these experiments to human using the inference scheme. This procedure thus produces a ranked list of human TFs whose binding is likely to be affected by the alleles of a given SNP (e.g., strongly binding to one allele but not binding to the other).

ACCESSION NUMBERS

The Gene Expression Omnibus (GEO) accession number for the PBM data reported in this paper is GSE53348.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Extended Experimental Procedures, seven figures, six tables, and one data file and can be found with this article online at http://dx.doi.org/10.1016/j.cell.2014.08.009.

AUTHOR CONTRIBUTIONS

M.T.W. and T.R.H. conceived of the study, oversaw analyses, and wrote the manuscript. M.T.W. performed most of the computational analyses. A.Y. performed and oversaw most of the experimental analyses. M.T.W. and M.A. developed and implemented the database and web server. A.G.C., A.M.M., I.M., H.Z., A.G., J.C.L., M.G., M.G.L., E.H., and J.R.H. performed additional experiments. H.S.N., M.A., H.vB., and X.C. provided additional computational support. P.D. performed the *Arabidopsis* eQTL analysis of Figure 7. S.L. and M.A. performed the motif location enrichment analysis in Figure 6. K.C. performed motif analyses related to Figure 3. T.M. performed analysis for the ChIP-seq comparisons in Figure 5. S.G., G.S., A.J.M.W., F.Y.B., G.R., L.F.L., and J.R.E. developed and oversaw experimental and computational analyses. All authors provided critical feedback on the manuscript, with extensive contributions from A.M.M., H.vB., J.R.H., A.J.M.W., L.F.L., M.T.W., and T.R.H.

ACKNOWLEDGMENTS

We thank the following people for experimental and computational support on this project: Shaheynoor Talukder, Andrew Vorobyov, Anton van der Ven, Wilfred de Vega, Nicole Park, C. Alexander Valencia, Geanany Rasanathan, Yogesh Hooda, Sanie Mnaimneh, Kenneth Chu, Oliver Boright, Jerry Li, Agnieszka Janska, Esther Chan, Trevis Alleyne, Emily Stoakes, Oliver Stegle, Todd Riley, and Harmen Bussemaker. We are grateful to Martha Bulyk, Jussi

Taipale, Raphael Kopan, Artem Barski, Ian Lewkowich, Brenda Andrews, and Charlie Boone for helpful conversations and advice, and the many investigators who generously contributed template DNA or RNA used in PCR cloning of DBDs (listed in Table S6). This work was supported by grants from the Canadian Institutes of Health Research to T.R.H. (MOP-77721 and MOP-111007). M.T.W. was supported by fellowships from CIHR and the Canadian Institute for Advanced Research (CIFAR) Junior Fellows Genetic Networks Program. T.R.H. is a CIFAR scholar. E.H. and G.S. were supported by a grant from the NIH/NICHD P01 HD39691. J.S.R.-H. and A.J.M.W. were funded by NIH GM082971. M.G.L. was supported by an EU Marie Curie International Outgoing Fellowship (project 252475). Research in the J.R.E. laboratory is supported by the National Science Foundation (NSF) (grant MCB-1024999), the Howard Hughes Medical Institute, and the Gordon and Betty Moore Foundation (GBMF 3034). J.R.E. is an HHMI-GBMF Investigator. A.M.M., A.G., and L.F.L. were supported by Millennium Nucleus for Fungal Integrative and Synthetic Biology (NC120043) and Fondo Nacional de Desarrollo Científico y Tecnológico (FONDECYT 1131030), both to L.F.L.

Received: January 8, 2014
Revised: April 3, 2014
Accepted: August 6, 2014
Published: September 11, 2014

REFERENCES

Aggarwal, P., Das Gupta, M., Joseph, A.P., Chatterjee, N., Srinivasan, N., and Nath, U. (2010). Identification of specific DNA binding residues in the TCP family of transcription factors in Arabidopsis. Plant Cell *22*, 1174–1189.

Alleyne, T.M., Peña-Castillo, L., Badis, G., Talukder, S., Berger, M.F., Gehrke, A.R., Philippakis, A.A., Bulyk, M.L., Morris, Q.D., and Hughes, T.R. (2009). Predicting the binding preference of transcription factors to individual DNA k-mers. Bioinformatics *25*, 1012–1018.

Atwell, S., Huang, Y.S., Vilhjálmsson, B.J., Willems, G., Horton, M., Li, Y., Meng, D., Platt, A., Tarone, A.M., Hu, T.T., et al. (2010). Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. Nature *465*, 627–631.

Badis, G., Berger, M.F., Philippakis, A.A., Talukder, S., Gehrke, A.R., Jaeger, S.A., Chan, E.T., Metzler, G., Vedenko, A., Chen, X., et al. (2009). Diversity and complexity in DNA recognition by transcription factors. Science *324*, 1720–1723.

Baldauf, S.L., Roger, A.J., Wenk-Siefert, I., and Doolittle, W.F. (2000). A kingdom-level phylogeny of eukaryotes based on combined protein data. Science *290*, 972–977.

Barski, A., and Zhao, K. (2009). Genomic location analysis by ChIP-Seq. J. Cell. Biochem. *107*, 11–18.

Berger, M.F., Philippakis, A.A., Qureshi, A.M., He, F.S., Estep, P.W., 3rd, and Bulyk, M.L. (2006). Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. Nat. Biotechnol. *24*, 1429–1435.

Berger, M.F., Badis, G., Gehrke, A.R., Talukder, S., Philippakis, A.A., Peña-Castillo, L., Alleyne, T.M., Mnaimneh, S., Botvinnik, O.B., Chan, E.T., et al. (2008). Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences. Cell *133*, 1266–1276.

Bernard, B., Thorsson, V., Rovira, H., and Shmulevich, I. (2012). Increasing coverage of transcription factor position weight matrices through domain-level homology. PLoS ONE *7*, e42779.

Carroll, S.B. (2008). Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. Cell *134*, 25–36.

Christensen, R.G., Enuameh, M.S., Noyes, M.B., Brodsky, M.H., Wolfe, S.A., and Stormo, G.D. (2012). Recognition models to predict DNA-binding specificities of homeodomain proteins. Bioinformatics *28*, i84–i89.

Cook, W.J., Mosley, S.P., Audino, D.C., Mullaney, D.L., Rovelli, A., Stewart, G., and Denis, C.L. (1994). Mutations in the zinc-finger region of the yeast regulatory protein ADR1 affect both DNA binding and transcriptional activation. J. Biol. Chem. *269*, 9374–9379.

de Boer, C.G., and Hughes, T.R. (2012). YeTFaSCo: a database of evaluated yeast transcription factor sequence specificities. Nucleic Acids Res. *40*, D169–D179.

de Boer, C.G., van Bakel, H., Tsui, K., Li, J., Morris, Q.D., Nislow, C., Greenblatt, J.F., and Hughes, T.R. (2014). A unified model for yeast transcript definition. Genome Res. *24*, 154–166.

De Masi, F., Grove, C.A., Vedenko, A., Alibés, A., Gisselbrecht, S.S., Serrano, L., Bulyk, M.L., and Walhout, A.J. (2011). Using a structural and logics systems approach to infer bHLH-DNA binding specificity determinants. Nucleic Acids Res. *39*, 4553–4563.

de Mendoza, A., Sebé-Pedrós, A., Šestak, M.S., Matejcic, M., Torruella, G., Domazet-Loso, T., and Ruiz-Trillo, I. (2013). Transcription factor evolution in eukaryotes and the assembly of the regulatory toolkit in multicellular lineages. Proc. Natl. Acad. Sci. USA *110*, E4858–E4866.

Desjarlais, J.R., and Berg, J.M. (1992). Toward rules relating zinc finger protein sequences and DNA binding site preferences. Proc. Natl. Acad. Sci. USA *89*, 7345–7349.

Eddy, S.R. (2009). A new generation of homology search tools based on probabilistic inference. Genome Inform. *23*, 205–211.

Enuameh, M.S., Asriyan, Y., Richards, A., Christensen, R.G., Hall, V.L., Kazemian, M., Zhu, C., Pham, H., Cheng, Q., Blatti, C., et al. (2013). Global analysis of Drosophila Cys_2-His_2 zinc finger proteins reveals a multitude of novel recognition motifs and binding determinants. Genome Res. *23*, 928–940.

Finn, R.D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J.E., Gavin, O.L., Gunasekaran, P., Ceric, G., Forslund, K., et al. (2010). The Pfam protein families database. Nucleic Acids Res. *38*, D211–D222.

FitzGerald, P.C., Shlyakhtenko, A., Mir, A.A., and Vinson, C. (2004). Clustering of DNA sequences in human promoters. Genome Res. *14*, 1562–1574.

French, J.D., Ghoussaini, M., Edwards, S.L., Meyer, K.B., Michailidou, K., Ahmed, S., Khan, S., Maranian, M.J., O'Reilly, M., Hillman, K.M., et al.; GENICA Network; kConFab Investigators (2013). Functional variants at the 11q13 risk locus for breast cancer regulate cyclin D1 expression through long-range enhancers. Am. J. Hum. Genet. *92*, 489–503.

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

Gordân, R., Hartemink, A.J., and Bulyk, M.L. (2009). Distinguishing direct versus indirect transcription factor-DNA interactions. Genome Res. *19*, 2090–2100.

Hindorff, L.A., Sethupathy, P., Junkins, H.A., Ramos, E.M., Mehta, J.P., Collins, F.S., and Manolio, T.A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc. Natl. Acad. Sci. USA *106*, 9362–9367.

Huber, W., von Heydebreck, A., Sültmann, H., Poustka, A., and Vingron, M. (2002). Variance stabilization applied to microarray data calibration and to the quantification of differential expression. Bioinformatics *18* (*Suppl 1*), S96–S104.

Jolma, A., and Taipale, J. (2011). Methods for Analysis of Transcription Factor DNA-Binding Specificity In Vitro. Subcell. Biochem. *52*, 155–173.

Jolma, A., Yan, J., Whitington, T., Toivonen, J., Nitta, K.R., Rastas, P., Morgunova, E., Enge, M., Taipale, M., Wei, G., et al. (2013). DNA-binding specificities of human transcription factors. Cell *152*, 327–339.

Lam, K.N., van Bakel, H., Cote, A.G., van der Ven, A., and Hughes, T.R. (2011). Sequence specificity is obtained from the majority of modular C2H2 zinc-finger arrays. Nucleic Acids Res. *39*, 4680–4690.

Lang, D., Weiche, B., Timmerhaus, G., Richardt, S., Riaño-Pachón, D.M., Corrêa, L.G., Reski, R., Mueller-Roeber, B., and Rensing, S.A. (2010). Genome-wide phylogenetic comparative analysis of plant transcriptional regulation: a timeline of loss, gain, expansion, and correlation with complexity. Genome Biol. Evol. *2*, 488–503.

Lee, W., Tillo, D., Bray, N., Morse, R.H., Davis, R.W., Hughes, T.R., and Nislow, C. (2007). A high-resolution atlas of nucleosome occupancy in yeast. Nat. Genet. *39*, 1235–1244.

Li, X.Y., Thomas, S., Sabo, P.J., Eisen, M.B., Stamatoyannopoulos, J.A., and Biggin, M.D. (2011). The role of chromatin accessibility in directing the widespread, overlapping patterns of Drosophila transcription factor binding. Genome Biol. *12*, R34.

Liu, X., Lee, C.K., Granek, J.A., Clarke, N.D., and Lieb, J.D. (2006). Whole-genome comparison of Leu3 binding in vitro and in vivo reveals the importance of nucleosome occupancy in target site selection. Genome Res. *16*, 1517–1528.

Maglich, J.M., Sluder, A., Guan, X., Shi, Y., McKee, D.D., Carrick, K., Kamdar, K., Willson, T.M., and Moore, J.T. (2001). Comparison of complete nuclear receptor sets from the human, Caenorhabditis elegans and Drosophila genomes. Genome Biol. *2*, RESEARCH0029.

Mariño-Ramírez, L., Spouge, J.L., Kanga, G.C., and Landsman, D. (2004). Statistical analysis of over-represented words in human promoter sequences. Nucleic Acids Res. *32*, 949–958.

Mathias, J.R., Zhong, H., Jin, Y., and Vershon, A.K. (2001). Altering the DNA-binding specificity of the yeast Matalpha 2 homeodomain protein. J. Biol. Chem. *276*, 32696–32703.

Matys, V., Kel-Margoulis, O.V., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., Reuter, I., Chekmenev, D., Krull, M., Hornischer, K., et al. (2006). TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. Nucleic Acids Res. *34*, D108–D110.

Maurano, M.T., Humbert, R., Rynes, E., Thurman, R.E., Haugen, E., Wang, H., Reynolds, A.P., Sandstrom, R., Qu, H., Brody, J., et al. (2012). Systematic localization of common disease-associated variation in regulatory DNA. Science *337*, 1190–1195.

Megraw, M., Pereira, F., Jensen, S.T., Ohler, U., and Hatzigeorgiou, A.G. (2009). A transcription factor affinity-based code for mammalian transcription initiation. Genome Res. *19*, 644–656.

Mintseris, J., and Eisen, M.B. (2006). Design of a combinatorial DNA microarray for protein-DNA interaction studies. BMC Bioinformatics *7*, 429.

Newburger, D.E., and Bulyk, M.L. (2009). UniPROBE: an online database of protein binding microarray data on protein-DNA interactions. Nucleic Acids Res. *37*, D77–D82.

Noyes, M.B., Christensen, R.G., Wakabayashi, A., Stormo, G.D., Brodsky, M.H., and Wolfe, S.A. (2008). Analysis of homeodomain specificities allows the family-wide prediction of preferred recognition sites. Cell *133*, 1277–1289.

Odom, D.T. (2011). Identification of Transcription Factor-DNA Interactions In Vivo. Subcell. Biochem. *52*, 175–191.

Ohler, U., Liao, G.C., Niemann, H., and Rubin, G.M. (2002). Computational analysis of core promoters in the Drosophila genome. Genome Biol. *3*, RESEARCH0087.

Park, P.J. (2009). ChIP-seq: advantages and challenges of a maturing technology. Nat. Rev. Genet. *10*, 669–680.

Persikov, A.V., and Singh, M. (2014). De novo prediction of DNA-binding specificities for Cys2His2 zinc finger proteins. Nucleic Acids Res. *42*, 97–108.

Portales-Casamar, E., Thongjuea, S., Kwon, A.T., Arenillas, D., Zhao, X., Valen, E., Yusuf, D., Lenhard, B., Wasserman, W.W., and Sandelin, A. (2010). JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. Nucleic Acids Res. *38*, D105–D110.

Ray, D., Kazan, H., Cook, K.B., Weirauch, M.T., Najafabadi, H.S., Li, X., Gueroussov, S., Albu, M., Zheng, H., Yang, A., et al. (2013). A compendium of RNA-binding motifs for decoding gene regulation. Nature *499*, 172–177.

Rosenbloom, K.R., Dreszer, T.R., Long, J.C., Malladi, V.S., Sloan, C.A., Raney, B.J., Cline, M.S., Karolchik, D., Barber, G.P., Clawson, H., et al. (2012). ENCODE whole-genome data in the UCSC Genome Browser: update 2012. Nucleic Acids Res. *40*, D912–D917.

Schneider, T.D., and Stephens, R.M. (1990). Sequence logos: a new way to display consensus sequences. Nucleic Acids Res. *18*, 6097–6100.

Seeman, N.C., Rosenberg, J.M., and Rich, A. (1976). Sequence-specific recognition of double helical nucleic acids by proteins. Proc. Natl. Acad. Sci. USA *73*, 804–808.

Shelest, E. (2008). Transcription factors in fungi. FEMS Microbiol. Lett. *286*, 145–151.

Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., et al. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol. Syst. Biol. *7*, 539.

Stormo, G.D., and Zhao, Y. (2010). Determining the specificity of protein-DNA interactions. Nat. Rev. Genet. *11*, 751–760.

Stubbs, L., Sun, Y., and Caetano-Anolles, D. (2011). Function and Evolution of C2H2 Zinc Finger Arrays. Subcell. Biochem. *52*, 75–94.

Tanaka, E., Bailey, T., Grant, C.E., Noble, W.S., and Keich, U. (2011). Improved similarity scores for comparing motifs. Bioinformatics *27*, 1603–1609.

Wang, J., Zhuang, J., Iyer, S., Lin, X.Y., Greven, M.C., Kim, B.H., Moore, J., Pierce, B.G., Dong, X., Virgil, D., et al. (2013). Factorbook.org: a Wiki-based database for transcription factor-binding data generated by the ENCODE consortium. Nucleic Acids Res. *41*, D171–D176.

Weirauch, M.T., and Hughes, T.R. (2011). A catalogue of eukaryotic transcription factor types, their evolutionary origin, and species distribution. Subcell. Biochem. *52*, 25–73.

Weirauch, M.T., Cote, A., Norel, R., Annala, M., Zhao, Y., Riley, T.R., Saez-Rodriguez, J., Cokelaer, T., Vedenko, A., Talukder, S., et al.; DREAM5 Consortium (2013). Evaluation of methods for modeling transcription factor sequence specificity. Nat. Biotechnol. *31*, 126–134.

Yamaguchi, M., Ohtani, M., Mitsuda, N., Kubo, M., Ohme-Takagi, M., Fukuda, H., and Demura, T. (2010). VND-INTERACTING2, a NAC domain transcription factor, negatively regulates xylem vessel formation in Arabidopsis. Plant Cell *22*, 1249–1263.

Yan, J., Enge, M., Whitington, T., Dave, K., Liu, J., Sur, I., Schmierer, B., Jolma, A., Kivioja, T., Taipale, M., and Taipale, J. (2013). Transcription factor binding in human cells occurs in dense clusters formed around cohesin anchor sites. Cell *154*, 801–813.

Yang, S.D., Seo, P.J., Yoon, H.K., and Park, C.M. (2011). The Arabidopsis NAC transcription factor VNI2 integrates abscisic acid signals into leaf senescence via the COR/RD genes. Plant Cell *23*, 2155–2168.

Zhao, Y., and Stormo, G.D. (2011). Quantitative analysis demonstrates most transcription factors require only simple models of specificity. Nat. Biotechnol. *29*, 480–483.

Zhu, L.J., Christensen, R.G., Kazemian, M., Hull, C.J., Enuameh, M.S., Basciotta, M.D., Brasefield, J.A., Zhu, C., Asriyan, Y., Lapointe, D.S., et al. (2011). FlyFactorSurvey: a database of Drosophila transcription factor binding specificities determined using the bacterial one-hybrid system. Nucleic Acids Res. *39*, D111–D117.

CellPress

Molecular Cell
Resource

Unambiguous Identification of miRNA:Target Site Interactions by Different Types of Ligation Reactions

Stefanie Grosswendt,[1,3] Andrei Filipchyk,[1,3] Mark Manzano,[2] Filippos Klironomos,[1] Marcel Schilling,[1] Margareta Herzog,[1] Eva Gottwein,[2] and Nikolaus Rajewsky[1,*]

[1]Systems Biology of Gene Regulatory Elements, Max-Delbrück-Center for Molecular Medicine, Robert-Rössle-Strasse 10, 13125 Berlin, Germany

[2]Department of Microbiology-Immunology, Feinberg School of Medicine, Northwestern University, 303 East Chicago Avenue, Chicago, IL 60611, USA

[3]Co-first authors

*Correspondence: rajewsky@mdc-berlin.de

http://dx.doi.org/10.1016/j.molcel.2014.03.049

SUMMARY

To exert regulatory function, miRNAs guide Argonaute (AGO) proteins to partially complementary sites on target RNAs. Crosslinking and immunoprecipitation (CLIP) assays are state-of-the-art to map AGO binding sites, but assigning the targeting miRNA to these sites relies on bioinformatics predictions and is therefore indirect. To directly and unambiguously identify miRNA:target site interactions, we modified our CLIP methodology in *C. elegans* to experimentally ligate miRNAs to their target sites. Unexpectedly, ligation reactions also occurred in the absence of the exogenous ligase. Our in vivo data set and reanalysis of published mammalian AGO-CLIP data for miRNA-chimeras yielded ~17,000 miRNA:target site interactions. Analysis of interactions and extensive experimental validation of chimera-discovered targets of viral miRNAs suggest that our strategy identifies canonical, noncanonical, and nonconserved miRNA:targets. About 80% of miRNA interactions have perfect or partial seed complementarity. In summary, analysis of miRNA:target chimeras enables the systematic, context-specific, in vivo discovery of miRNA binding.

INTRODUCTION

miRNAs associate with Argonaute (AGO) proteins to guide the RNA-induced silencing complex (RISC) to transcripts and thereby repress protein production of target mRNAs (Baek et al., 2008; Bartel, 2009; Fabian et al., 2010; Selbach et al., 2008). Consequently, the biological role of a miRNA is mainly specified by its set of targets. Identifying miRNA targets remains challenging because, in animals, a miRNA typically has hundreds of direct targets under negative selection (Brennecke et al., 2005; Krek et al., 2005; Lewis et al., 2005; Xie et al., 2005), and target recognition occurs through only partial sequence complementarity (Bartel, 2009; Rajewsky, 2006). Of particular importance to target recognition is the miRNA seed sequence, i.e., nucleotides (nt) 2–7 from the 5′ end of the miRNA (Bartel, 2009; Lewis et al., 2005; Lai, 2002; Rajewsky, 2006). Perfect complementarity to the seed is often found to be fundamental for binding and a regulatory response. However, in addition to these canonical binding sites, numerous noncanonical miRNA target sites have been reported (Bagga et al., 2005; Chi et al., 2012; Didiano and Hobert, 2006; Helwak et al., 2013; Lal et al., 2009; Shin et al., 2010; Vella et al., 2004). The base-pairing patterns for noncanonical targets are not well understood due to difficulties in their identification.

Conventional approaches to identify miRNA targets commonly aim to detect perfect seed matches in 3′ UTRs, often by incorporating additional information, such as conservation, accessibility, and expression of 3′ UTR sequences. Despite the general success of these methods, they do not take into account context specificity, such as binding sites masked by other RNA binding proteins (RBPs) or tertiary structure constraints, and are not effective at identifying noncanonical or nonconserved sites. Moreover, false positive rates are often high unless specificity is boosted at the expense of sensitivity.

Recently, crosslinking and immunoprecipitation (CLIP) methods (Chi et al., 2009; Hafner et al., 2010; Lebedeva et al., 2011) have identified AGO binding sites at a transcriptome-wide scale, generating context-dependent AGO binding maps. These data per se do not reveal the identity of the miRNA(s) bound to a certain site. Therefore, tools have been developed to computationally predict which miRNAs are bound at which AGO sites (Erhard et al., 2013; Khorshid et al., 2013; Liu et al., 2013; Majoros et al., 2013). However, assumptions must be made, such as which miRNAs are loaded into AGO, how miRNAs recognize targets, and regarding the validity of biophysical energy or hidden Markov or logistic models. It is therefore still difficult to confidently and unambiguously assign which miRNA was bound to a certain site, especially for sites containing none or several different seed matches. The identification of miRNA:targets can be further complicated by sequence similarities between miRNAs. For example, viral miRNAs can share seed

sequences with human miRNAs and thus can interfere with human miRNA binding in infected cells (Gottwein et al., 2011; 2007; Manzano et al., 2013; Skalsky et al. 2007; 2012; Zhao et al., 2011).

We set out to complement existing approaches by experimental miRNA:target identification. We recently developed iPAR-CLIP, a biochemical method to generate in vivo maps of binding sites for RBPs in *C. elegans* (Jungkamp et al., 2011). Here, we used iPAR-CLIP and mapped 29,000 unique AGO binding sites in the worm, improving resolution and depth of previous studies (Zisoulis et al., 2010). Additionally, we experimentally ligated miRNAs to their binding sites. Our method is similar, but not identical, to the CLASH protocol recently applied in a human cell line (Helwak et al., 2013). Sequencing and computational analysis of these chimeras revealed thousands of miRNA:targets in *C. elegans*. Unexpectedly, we also detected thousands of chimeras when no ligase was added, indicating endogenous RNA ligase activity in standard CLIP assays. Indeed, by reanalyzing sequencing data from published AGO-CLIP experiments, we succeeded in compiling >13,000 human and mouse miRNA:targets.

We present multiple lines of evidence that these miRNA:targets have features expected from functional interactions. We also tested the functionality of miRNA:targets for viral miRNAs. With the exception of a viral miRNA with poor targeting proficiency (Garcia et al., 2011; Manzano et al., 2013), we confirmed regulation for 87% of tested sites, including noncanonical and nonconserved sites. Computational analyses of our chimera-identified miRNA:targets suggest that ~80% of these interactions have statistically highly significant perfect or imperfect (1 nt mismatch) complementarity to the miRNA seed (nt 2-7). Our data further suggest that mismatches in the seed occur predominantly at positions 2 and 7. Thus, AGO-CLIP sequencing data contain chimeric reads that enable the identification of endogenous, context-specific miRNA:targets on a transcriptome-wide scale. These data allow insights into principles by which miRNA recognize target sites.

RESULTS

Identification of 3,600 miRNA:Targets in *C. elegans*

We adapted our iPAR-CLIP protocol (Jungkamp et al., 2011) to ligate miRNAs to target sites in *C. elegans* (Figure 1A). Briefly, worms incorporated photoreactive 4-thiouridine nucleosides (4sU) into their RNA, which crosslinks to bound proteins during UV irradiation. After homogenization, the lysate was treated with RNase T1. Argonaute ALG-1 was immunoprecipitated, and bound RNAs were treated again with RNase, recovered under stringent conditions, and deep sequenced (Figures S1A and S1B available online; Experimental Procedures). For the miRNA:target ligations, we added T4 RNA ligase to immunopurified and washed AGO complexes. To prevent circularization, we prepared the RNA ends, leaving the 3′ end of target sites blocked (Figure S1C). Thus, T4 RNA ligase solely connects the 3′ hydroxyl (3′ OH) of full-length miRNAs with the 5′ ends of target RNA fragments.

In addition to two standard AGO iPAR-CLIP samples, we generated two biologically and technically independent replicates of ligation samples and control samples. Control samples were generated without the addition of a ligase. Bioinformatic analysis (Experimental Procedures) of the standard AGO iPAR-CLIP, the ligation, and the control samples identified the presence of a total of 13.5 million nonchimeric reads, which had an average read length of 32 nt and mapped uniquely to the transcriptome. In PAR-CLIP, the frequency of T-to-C conversions in reads reflects the RNA-protein crosslink efficacy. This frequency was very high (14:1) in nonchimeric reads compared to all other possible nucleotide changes.

These data defined a high-resolution AGO binding map (Figure S2A) that includes 2,286 *C. elegans* AGO sites previously identified (Zisoulis et al., 2010: 4,806 unique target sites in 3,093 genes, average length 122 nt; present study: unique 29,000 sites in 8,339 genes, average length 42 nt; Table S3). Our bioinformatics analyses (Experimental Procedures) revealed the presence of thousands of miRNA-chimeric reads in the ligation samples (Figure 1B). When mapping the target sequence in chimeras to the transcriptome, the vast majority mapped to 3′ UTRs. Moreover, almost all target sites fell precisely into AGO binding sites (Figure S1D). Consequently, we mapped chimera target sequences directly to AGO sites. This increased the sensitivity of target recovery due to the smaller search space (Experimental Procedures). In total, we identified 3,627 miRNA:targets for *C. elegans* (Table S3), 677 of which were supported by more than one chimeric sequence read. Interactions supported by one read showed essentially the same features as interactions recovered by >1 read (see below; Figures S1E and S1F).

Control Samples Also Contain miRNA:Target Chimeras

Unexpectedly, we detected substantial numbers of chimeras in our control samples as well as in our AGO iPAR-CLIP samples. These samples have not been treated with T4 RNA ligase to generate chimeras (Figure 1C). While ligation samples had miRNA:target chimeras containing complete or truncated miRNAs, nearly all chimeras detected in control and AGO iPAR-CLIP samples contained 3′-truncated miRNA sequences (Figures 1C, S1G, and S1H).

Truncated miRNAs were strongly enriched in a guanine immediately upstream of the cleavage sites in miRNAs and target RNAs (Figure 1D). RNase T1 cuts with high preference after guanines, strongly suggesting that RNase T1 produced the ends used as substrates for this ligation reaction. The ligation activity required for ligation of 2′,3′-cyclic P (can convert into 3′P) is present in eukaryotic cell lysates, as previously reported (Filipowicz et al., 1983; Martinez et al., 2002; Perkins et al., 1985) (Figure 1E).

Could RNAs in the lysate randomly ligate to AGO-loaded miRNAs? Bacteria are the food source for *C. elegans*. Therefore, iPAR-CLIP sequencing data usually contain a substantial fraction of bacterial sequences (~30%). In contrast, less than 2% of recovered iPAR-CLIP chimeras contained *C. elegans* miRNAs together with bacterial sequences, indicating that ligation of random RNA fragments to AGO-loaded miRNAs occurred only rarely. Moreover, miRNA:targets from different samples were highly overlapping (Figure S1I). The majority (76%) of interactions derived from chimeras with complete miRNAs were also identified from chimeras with truncated miRNAs (Figure S1J),

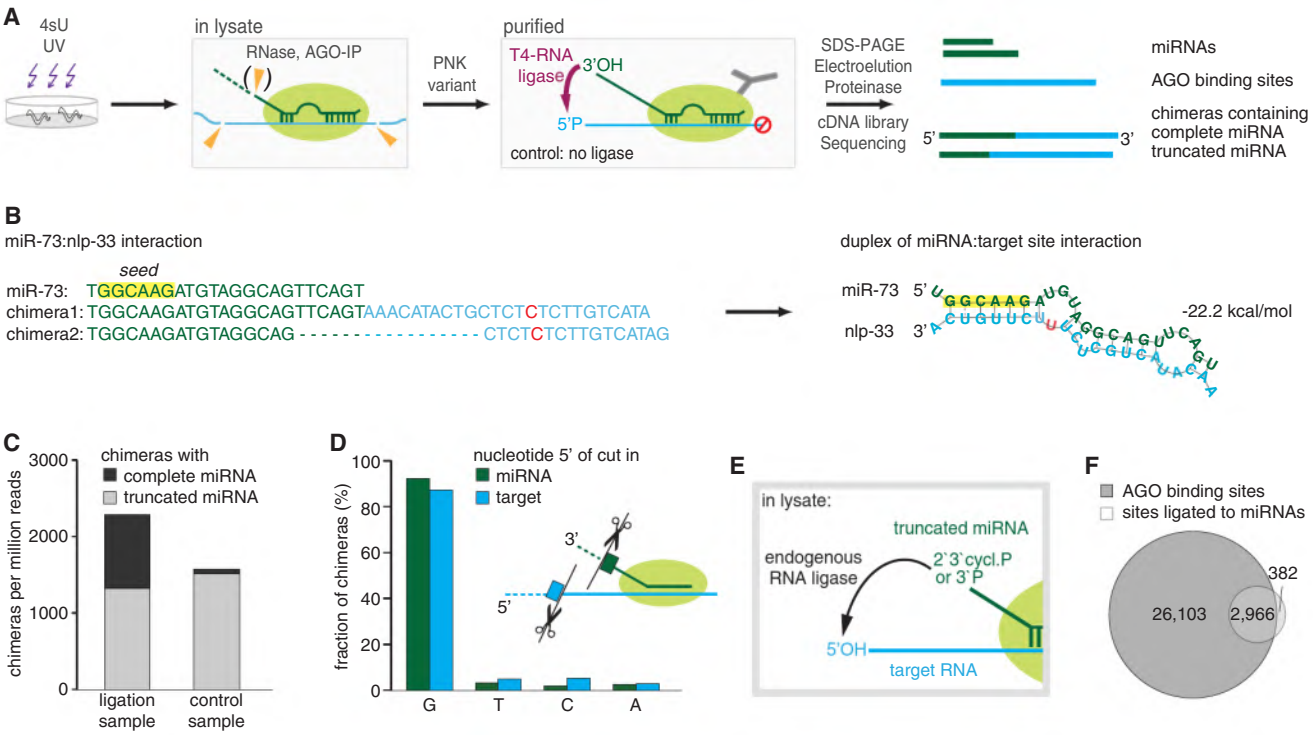


Figure 1. Generation of miRNA:Target Chimeras via Different Types of Ligations in *C. elegans*

(A) *C. elegans* RNA labeled with photoreactive nucleoside 4-thiouridines (4sU) is crosslinked to bound proteins in vivo. After homogenization of worms, the lysate is treated with RNase T1. Some miRNAs are shortened, and others remain complete. Following immunoprecipitation (IP) and washing of AGO, crosslinked RNA is phosphorylated by a PNK variant (leaves 3′ ends blocked) and treated with T4 RNA ligase, which ligates the 3′-hydroxyl end of complete miRNAs to bound RNA fragments. Crosslinked RNA is recovered and deep sequenced. Computational analysis detects sequence reads of miRNAs and AGO binding sites, along with chimeric reads containing miRNAs connected to their targets.

(B) Example of a miRNA interaction recovered from chimeric reads. Predicted reconstruction of the miRNA:target duplex. Green, miRNA sequence; blue, target sequence; red, T-to-C conversion.

(C) Data from the ligation sample contain chimeras with 3′ truncated (length of miRNA sequence $\geq$ 13 nt) and with complete miRNAs. A comparable fraction of chimeras with truncated miRNAs was also found in a control sample, to which no ligase was added to generate chimeras.

(D) miRNA and target ends involved in the ligations of the control sample are highly enriched in an upstream G, suggesting that RNase T1 generated the ends used for this type of ligation.

(E) Truncated miRNAs are ligated by the ligase activity of the lysate during IP.

(F) The majority (89%) of chimera-derived miRNA target sites (mapped to the transcriptome) overlap with AGO binding sites generated from nonchimeric reads. See also Figure S1.

suggesting the recovery of in vivo interactions for both ligation reactions. Together, these findings prompted us to systematically analyze the ~3,600 identified *C. elegans* miRNA:target site interactions jointly for known features of miRNA targeting.

C. elegans miRNA:Targets Recovered from Chimeras Have Features Characteristic of miRNA Binding

C. elegans miRNA:target chimeras mapped mainly to 3′ UTRs and coding sequences (Figure S2A), similar to AGO binding sites. Also, the sequences ligated to miRNAs are bona fide AGO sites since 89% of mapped chimeras overlapped by at least 80% of their length with AGO binding sites, which is highly statistically significant (p ~0, Figure 1F). Consistently, targets in miRNA-chimeras had a T-to-C conversion rate of 84%, enriched 20-fold over any other type of nucleotide change (Figure S2B). This crosslink-specific signal indicates that sequences ligated to complete or truncated miRNAs were bound by AGO.

We next analyzed if targets found in miRNA chimeras have sequence complementarities to the ligated miRNA. We screened for perfect (non-G:U) complementarity to miRNA nt 2–7 (seed), complementarity to miRNA nt 2–7 containing one mismatched or bulged nucleotide, and complementarity to miRNA nt 2–8 containing two mismatches. Together, these modes were detected in ~80% of interactions and are highly significantly enriched compared to random controls (Figure 2A).

We analyzed miRNA:target hybridization with RNAhybrid (Rehmsmeier et al., 2004). The median free energy was lower (by 3.3 kcal/mol or ~2–3 hybridized nt) for miRNA:targets compared to controls (Figure S2C). Base pairing was as expected for miRNA (Wee et al., 2012; Khorshid et al., 2013): clearly preferred in the seed while reduced at positions 9, 10, and 11 (Figure 2B).

As for interactions with perfect seed matches, analysis of miRNA:targets without detected seed complementarities

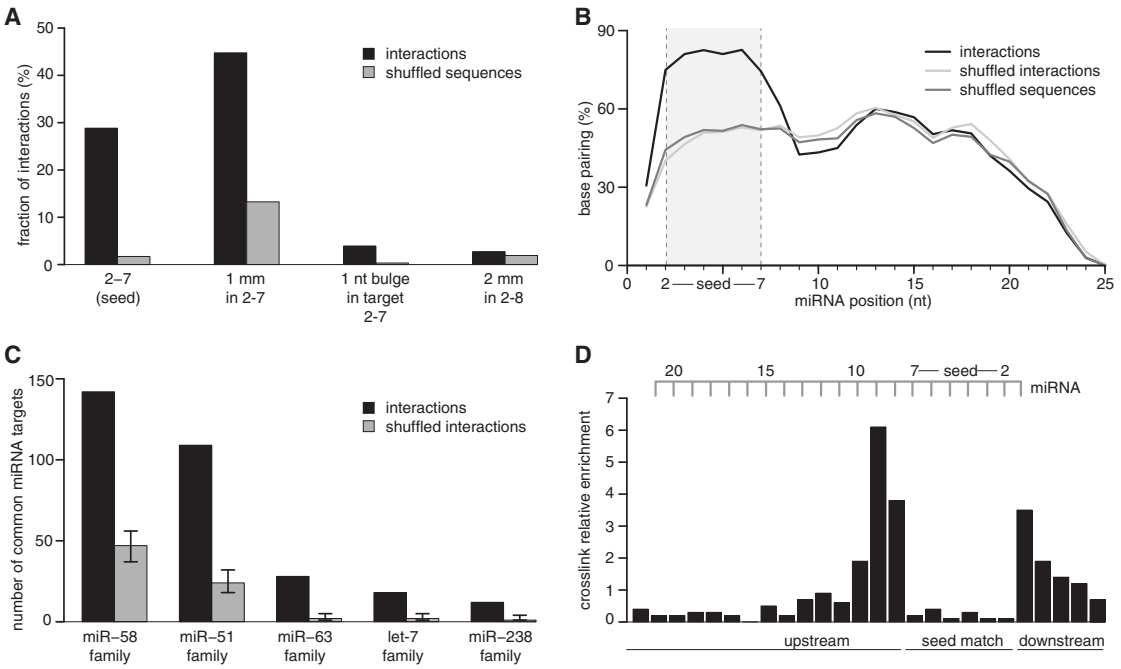


Figure 2. *C. elegans* miRNA:Targets (3,600) Derived from Chimeras Reflect Endogenous miRNA Targeting

(A) Target RNAs were analyzed for complementarity to the seed region of their ligated miRNAs. Approximately 80% of interactions possess the tested complementarities. Shuffled sequences (dinucleotides in target sequences are permuted) served as control. mm, mismatch. Mismatches were broadly distributed over all types of nucleotides, including G:U.

(B) Hybridization profile summarized over all interactions. The predicted frequency of a miRNA position to be base paired is plotted along the miRNA length. Duplex structures of miRNA:targets were predicted by RNAhybrid, allowing G:U pairing. Shuffled sequences (dinucleotides in target sequences are permuted) and shuffled interactions (targets are swapped between miRNAs) served as control.

(C) Target sites derived from miRNA-chimeras are found ligated to miRNAs of the same family more often than expected by chance ($p < 0.0001$). Shuffling target sites between miRNA families served as control.

(D) Local frequency of crosslink-induced T-to-C conversions in target RNAs from interactions with a perfect 2–7 seed match (normalized to local thymidine frequency). Nucleotides hybridized to the seed of the miRNA are strongly indisposed to crosslink with the protein. See also Figure S2.

revealed direct evidence for crosslinking (T-to-C conversions, Figure S2D). However, their binding free energy was decreased less (Figure S2E), and hybridization profiles did not indicate enriched base pairing within the seed region (Figure S2F).

Since miRNA family members have the same seed, they are expected to share some of their targets, and indeed, target sites were ligated to members of the same miRNA family much more often than expected by chance. Thus, chimeras can capture multiple endogenous miRNA targeting events at the same target site (Figure 2C, $p < 0.0001$).

An increased T-to-C conversion frequency directly upstream of seed matches was previously reported for AGO PAR-CLIP data (Hafner et al., 2010; Kishore et al., 2011). Similarly, miRNA:targets with a perfect or imperfect seed match showed clear conversion patterns, with the highest number of conversions at the second position upstream and a strong depletion within seed matches (Figures 2D and S2G). Interestingly, miRNAs found in ligation products had a 3-fold lower T-to-C conversion rate compared to nonligated miRNAs (Figure S2H). This is because noncrosslinked miRNAs tend to be lost under the denaturing conditions of protein purification, while ligated, non-crosslinked miRNAs can pass purification due to their covalent connection to AGO.

Published Mammalian AGO-CLIP Data Contain Ligated miRNA:Targets

If, in our *C. elegans* experiments, ligated miRNA:targets are generated through a ligation activity naturally present in the lysate, then existing AGO-CLIP data should also contain miRNA:targets. Therefore, we searched for miRNA-chimeras in published AGO-CLIP data sets across several model systems and CLIP methods, such as HITS-CLIP (high-throughput sequencing of RNA isolated by crosslinking immunoprecipitation; Chi et al., 2009) and PAR-CLIP (Photo-activatable-ribonucleoside-enhanced crosslinking and immunoprecipitation; Hafner et al., 2010) (Table 1; Figure S6; Experimental Procedures). Our pipeline confidently detected miRNA sequences of 13 nt and longer and mapped targets as short as 16 nt with an estimated false discovery rate (FDR) < 5%. This sensitivity was essential for analysis since AGO-CLIP sequences from published studies were typically short (16-36 nt).

In total, we recovered ~11,000 human miRNA:targets, ~2,000 mouse miRNA:targets, ~500 for Kaposi's sarcoma-associated herpesvirus (KSHV) miRNAs, and ~300 for Epstein-Barr virus miRNAs. As expected, most chimeras (>80%) contained 3′-truncated miRNAs. Numbers of miRNA:targets varied strongly