Integrative analysis identifies co-dependent gene expression regulation of BRG1 and CHD7 at distal regulatory sites in embryonic stem cells

Pengyi Yang\textsuperscript{1,2,*}, Andrew Oldfield\textsuperscript{2,3,†}, Taiyun Kim\textsuperscript{4,†}, Andrian Yang\textsuperscript{4}, Jean Yee Hwa Yang\textsuperscript{1} and Joshua W. K. Ho\textsuperscript{4,5,*}

\textsuperscript{1}Charles Perkins Centre and School of Mathematics and Statistics, University of Sydney, Camperdown, NSW 2006, Australia, \textsuperscript{2}Systems Biology Group, Epigenetics & Stem Cell Biology Laboratory, National Institute of Environmental Health Sciences, National Institutes of Health, RTP, NC 27709, USA, \textsuperscript{3}Institute of Human Genetics, CNRS UPR 1142, Montpellier, France, \textsuperscript{4}Victor Chang Cardiac Research Institute and \textsuperscript{5}St. Vincent’s Clinical School, University of New South Wales, Darlinghurst, NSW 2010, Australia

\textsuperscript{*}To whom correspondence should be addressed.

\textsuperscript{†}The authors wish it to be known that, in their opinion, the second and third authors should be regarded as joint Second Authors

Associate Editor: Jonathan Wren

Received on November 14, 2016; revised on January 21, 2017; editorial decision on February 7, 2017; accepted on February 8, 2017

Abstract

\textbf{Motivation:} DNA binding proteins such as chromatin remodellers, transcription factors (TFs), histone modifiers and co-factors often bind cooperatively to activate or repress their target genes in a cell type-specific manner. Nonetheless, the precise role of cooperative binding in defining cell-type identity is still largely uncharacterized.

\textbf{Results:} Here, we collected and analyzed 214 public datasets representing chromatin immunoprecipitation followed by sequencing (ChIP-Seq) of 104 DNA binding proteins in embryonic stem cell (ESC) lines. We classified their binding sites into those proximal to gene promoters and those in distal regions, and developed a web resource called Proximal And Distal (PAD) clustering to identify their co-localization at these respective regions. Using this extensive dataset, we discovered an extensive co-localization of BRG1 and CHD7 at distal but not proximal regions. The comparison of co-localization sites to those bound by either BRG1 or CHD7 alone showed an enrichment of ESC master TFs binding and active chromatin architecture at co-localization sites. Most notably, our analysis reveals the co-dependency of BRG1 and CHD7 at distal regions on regulating expression of their common target genes in ESC. This work sheds light on cooperative binding of TF binding proteins in regulating gene expression in ESC, and demonstrates the utility of integrative analysis of a manually curated compendium of genome-wide protein binding profiles in our online resource PAD.

\textbf{Availability and Implementation:} PAD is freely available at http://pad.victorchang.edu.au/ and its source code is available via an open source GPL 3.0 license at https://github.com/VCCRI/PAD/

\textbf{Contact:} pengyi.yang@sydney.edu.au or j.ho@victorchang.edu.au

\textbf{Supplementary information:} Supplementary data are available at \textit{Bioinformatics} online.
1 Introduction

Chromatin remodellers, transcription factors (TFs), histone modifiers and various other co-factors, collectively referred to as DNA binding proteins, play critical roles in establishing and maintaining cell type-specific gene regulatory networks (Chen and Dent, 2014). With the advance of ChIP-Seq, the genome-wide binding profiles of a large number of DNA binding proteins have been mapped in numerous cell types by various research groups and international consortia such as ENCODE (Consortium et al., 2012) and FANTOM (Klein et al., 2005). Through these data and other functional studies, it has become increasingly clear that cooperative binding is a widely utilized mechanism for DNA binding proteins to recognize specific genomic elements and activate or repress their target genes (Spitz and Furlong, 2012). However, the functional extent of such partnerships genome-wide has yet to be fully addressed. Our previous study found that NF-Y, a CCAAT-binding complex, cooperates with different sets of TFs at gene promoters than it does at distal regulatory elements. This dichotomy is essential to regulate, respectively, cell type-invariant and cell type-specific gene expression (Oldfield and Furlong, 2012). One intriguing question is whether such a multi-faceted mechanism is also utilized by other DNA binding proteins to regulate their specific target genes.

In order to dissect different binding roles and investigate their functional consequences of protein cooperative binding, we curated a comprehensive repository of 104 DNA binding proteins whose genome-wide binding profiles were mapped using ChIP-Seq in mouse embryonic stem cell (ESC) lines. We subsequently classified their binding sites according to their distance to gene promoters. Through such a classification, we characterized the proximal and distal binding preference of each of these proteins. We developed a user friendly web resource to access this repository, called proximal and distal (PAD) clustering, and applied this tool to characterize protein co-localization at proximal and distal regions using this large compendium of ESC-specific protein binding profiles. In particular, we observed that many chromatin remodellers show strong preference in binding at distal regions in ESCs. These include BRG1, the ATPase subunit of the chromatin remodeling complex esBAF (Ho et al., 2009), and several members of chromodomain-helicase-DNA-binding (CHD) proteins. While previous studies have reported significant overlaps in binding locations of BRG1 and the CHD family of proteins in various cell types including ESCs (de Dieuleveut et al., 2016; Laurette et al., 2015; Morris et al., 2013), less is known for their functional dependency. Using BRG1 as an anchor against all other DNA binding proteins included in PAD. Our analyses confirmed the substantial binding overlaps between BRG1 and CHD7, and to a lesser extent to CHD4, but revealed a novel aspect that these co-localization sites are predominately found at distal regions but not at promoter proximal regions.

We then focused on characterizing the co-localization sites of BRG1 and CHD7 and compared them to those where either BRG1 or CHD7 bind independently. The comparative analysis revealed sites where BRG1 and CHD7 localize are significantly enriched for the binding of ESC master TFs including NANOG, SOX2 and OCT4 over independent sites. Furthermore, co-localization sites show chromatin features reminiscent of active distal regulatory elements with considerably stronger H3K27ac, H3K4me1 and P300 ChIP-Seq signals and more three dimensional (3D) chromatin interactions supported by HiC-Seq data. The analysis of gene expression data comparing Brg1 knockdown (KD) and Chd7 knockout (KO) with wild-type (WT) ESC samples showed significant up-regulation of genes targeted by BRG1 and CHD7 co-localization sites compared to those targeted by either BRG1 or CHD7 alone. These results demonstrate the utility of PAD and suggest co-dependent gene expression regulation of BRG1 and CHD7 at distal regulatory sites in ESCs, a previously unappreciated functional reliance.

2 Materials and methods

ChIP-Seq data analysis. All ChIP-Seq data analyzed in this study were generated from mouse ESC lines (Supplementary Table S1). Reads from each ChIP-Seq dataset were aligned to the mouse genome (mm9 assembly) using Bowtie version 0.12.8 (Langmead et al., 2009). Only reads that mapped to unique genomic regions with at most two mismatches were retained for follow up analysis. To create UCSC Genome Browser tracks, data were normalized to reads per million (RPM) and plotted as histograms. To define binding sites for each DNA binding factor, aligned reads were processed using SISRs with P < 0.001, a stringent cut-off to ensure high confidence on called binding sites (Jothi et al., 2008). Datasets that have too few called binding sites (<1000) or a suspiciously large number of binding sites (>50000) were removed from further analysis. Manual inspection was then performed to select the most representative data for factors that were profiled by multiple laboratories. This resulted in a total of 214 ChIP-Seq datasets corresponding to 104 DNA binding proteins. Multiple datasets for the same protein were combined by selecting the one that has median number of peaks for subsequent analysis. BRG1-alone sites are defined as BRG1 peaks that have ChIP-Seq signals that at least 5-fold greater than the CHD7 ChIP-Seq signals at the same locus. CHD7-alone sites are determined in a similar fashion. BRG1 + CHD7 sites are those loci where BRG1 and CHD7 peaks overlap.

Classification of proximal and distal binding sites. A binding site was classified as a proximal binding site if the center of the binding site is within 1000 bp to one or more TSS annotated by RefSeqGene. It was classified as a distal binding site otherwise.

RNA Polymerase II (RNA-Pol II) ChIA-PET and Hi-C-Seq data analysis. Chromatin interaction between distal binding sites and promoters were determined based on RNA-Pol II ChIA-PET (Zhang et al., 2013) and Hi-C-Seq (Dixon et al., 2012) datasets in ESCs. A distal binding site was said to interact with a promoter if the 1 Kb region surrounding the center of the binding site and the 1 Kb region surrounding the center of the TSS overlap in the ChIA-PET and/or Hi-C-Seq interaction maps (FDR < 0.05).

Co-localization quantification. Co-localization at either proximal or distal regions were quantified using Jaccard index implemented in Bedtools (Quinlan, 2014). Specifically, the binding sites (either proximal or distal) of two factors were compared to find the total number of intersecting base pairs of all overlapping proximal or distal binding sites and the number of base pairs in union of all proximal or distal binding sites. The Jaccard index is given as: $Jaccard(P_i, P_j) = \frac{\text{Interation} \cup \text{Intersection}}{\text{Union} \cup \text{Intersection}}$, where $\text{Interation}(.)$ is the function that counts the number of base pairs covered by a set of peaks. Note that this implementation will not quantify factors that bind adjacent to each other with no overlapping base pairs.

PAD implementation. PAD clustering was developed in Python 2.7 based on Django web framework. The backend was implemented in Python utilizing python packages and the front end was developed using jQuery and javascript. The input table was implemented using jQuery tables and Bootstrap. Django-crispy-forms were used to wrap the input form to allow aesthetic and simple implementations. Django-multupload was also used to allow user to upload multiple file for submission. PAD uses plotly.js package to
visualizes co-localization matrix as heatmap by generating a background distribution of Jaccard index using a permutation-based approach where called peaks from peak files were permuted 100,000 times. This allows the calculation of log^10 fold change of Jaccard index observed from two peaks files log^10\left(\frac{\text{Jaccard}(P_1, P_2)}{\text{Jaccard}_{\text{permute}} + 0.1}\right). In addition, PAD also uses the permuted data to generate empirical P-values.

**Gene expression analysis.** Data from Brg1 knockdown (KD) and WT ESC samples (in triplicates), measured using Affymetrix Mouse Genome 430 2.0 Array, were downloaded and processed as described in (Ho et al., 2011). Log2 fold change for each gene was then calculated by averaging across triplicates in either Brg1 KD or WT and subtracting log_2 transformed values in Brg1 KD measurement with WT measurement. The same approach as above was applied to process gene expression data measured from Chd7 knock-out and WT ESC samples using Illumina MouseRef-8 v2.0 Expression BeadChip (Schnetz et al., 2010). Gene with the closest TSS to a binding site was defined as the target gene of that binding site. This definition was applied to assign target genes to both proximal and distal binding sites.

### 3 Results

#### 3.1 Binding preference of DNA binding proteins in ESCs

Using PAD, we found that most of the binding sites that are in close proximity to an annotated TSS are distributed within 1 Kb (Fig. 1A). Interestingly, we noticed that proximal binding of ESC master TFs such as NANOG, SOX2 and OCT4 are in general upstream of the TSS whereas core subunits of Polycomb Repressive Complex 2 (PRC2) such as SUZ12, JARID2 and EZH2 generally bind downstream. We found that RNAPII binds around 30bp downstream of the TSS, consistent with the promoter-proximal pausing of RNAPII (Adelman and Lis, 2012).

Subsequently, we classified each protein binding site as either promoter proximal or distal. As expected, promoter mark H3K4me3 is found to be mostly located proximally to TSS whereas enhancer mark H3K4me1 is associated with distal regulatory elements, is found to be mostly located distally (Fig. 1B). Furthermore, TFs that are known to be ubiquitously found at the TSS generally show preference to proximal binding (e.g. TBP, SP1, SP2, SP3, WDR5, C-MYC and N-MYC) whereas TFs that are known to be ESC-specific and occupy distal regulatory elements generally show preference to distal binding (e.g. ESRRB, OCT4, NANOG, TCFCP2L1, OTX2 and SOX2) (Fig. 1B). In agreement with our previous findings, the subunits of NF-Y complex (NFYA, NFYB and NFYC) bind both proximal and distal regions (Fig. 1B).

#### 3.2 Chromatin remodellers bind preferentially at distal regions and interact with TSSs

Through analysing the binding preference of DNA binding proteins included in our repository, we observed that many chromatin remodelers show preferential binding at distal regions in ESCs (Supplementary Fig. S1B). This includes BRG1, the ATPase subunit of the esBAF complex (Ho et al., 2009) and several members of CHD family proteins. Gene set enrichment test against chromatin remodelers defined in (Zhang et al., 2012) using Limma R package (Smyth, 2004) confirmed that they are statistically enriched (P = 0.034) for preferential distal binding compared to other DNA binding proteins (Supplementary Fig. S1A).

Chromatin remodelers are crucial in controlling DNA accessibility and previous studies from several groups have demonstrated that BRG1 and CHD proteins play key roles in pluripotency and/or differentiation of ESCs (Ho et al., 2009, 2011; Schnetz et al., 2010). To gain deeper insight into how they govern the pluripotency network, we chose BRG1 as a representative chromatin remodeler and investigated its genomic binding profile in comparison to those of various histone modifications and ESC master TFs (Supplementary Fig. S1B). We found that both ESC identity genes (e.g. Klf4, E2f1) and differentiation genes (e.g. Otx2) are marked by BRG1 at a nearby distal regulatory region. These distal binding sites are found to co-localize with ESC master TFs at accessible chromatin regions and interact with their target TSS through chromatin looping, supported by RNAPII ChIA-PET data in ESC (Zhang et al., 2013).

#### 3.3 Co-localization of DNA binding proteins at proximal and distal regions with BRG1

We next asked which DNA binding proteins co-localize most frequently at the same distal and proximal regions genome-wide with BRG1 (Supplementary Fig. S2A). We found that CHD7 co-localizes most extensively with BRG1 at distal regions whereas HDAC2...
co-localizes the most at proximal regions. However, after taking the co-localization scores (based on Jaccard index) into account, the co-localization with HDAC2 appears to be non-specific to proximal and distal regions while CHD7 shows a high co-localization score with BRG1 only in distal regions as indicated by the delta Jaccard index in Supplementary Figure S2B. Previous studies have found that BRG1 and CHD7 co-localize substantially in ESCs (Schnetz et al., 2010) and form part of a PBAF chromatin remodeling complex in neural crest derived cells (Laurette et al., 2015). Our results confirm these findings, but additionally reveal the critical detail that BRG1 and CHD7 co-localize predominantly at distal regulatory regions and not proximally to TSSs. By ranking co-localization based on the delta Jaccard index, we also identified several other factors that show differential co-localization with BRG1. These include another chromatin remodeler CHD4, PRC1 subunit RING1B and several ESC master TFs including TCF3, SOX2, STAT3 and NANOG (Supplementary Fig. S2B).

We subsequently utilized PAD to cluster BRG1 with a panel of selected ESC master TFs as well as several factors that show differential co-localization with BRG1. We found that BRG1 has weak co-localization at proximal regions (Supplementary Fig. S2C) even with those factors that appear to be differentially co-localizing with BRG1 proximally at TSS (e.g. N-MYC, C-MYC and WDR5). In contrast, BRG1 co-localizes extensively with CHD7 and several other ESC master TFs at distal regions (Supplementary Fig. S2C). We then focused on comparing BRG1 and CHD7 binding sites (Supplementary Fig. S2D) and defined distal sites that are (i) BRG1-alone with respect to CHD7 (1493 sites), (ii) co-localized between BRG1 and CHD7 (718 sites) and (iii) CHD7-alone with respect to BRG1 (461 sites). For proximal binding, given the limited number of promoter-proximal CHD7 binding sites, we included only (iv) the 618 proximal BRG1 alone binding sites in the subsequent analysis.

3.4 Gene expression regulation depends on BRG1 and CHD7 distal co-localization

To understand the functional significance of the co-localization of BRG1 and CHD7 at distal regions, we next compared the four groups of binding sites defined in the previous section. As expected, distal sites are distinguished from proximal sites by H3K4me3 and H3K4me3. For distal sites, comparable amount of BRG1 or CHD7 ChIP-Seq signals were found at those co-occupied by them and those occupied by either BRG1 or CHD7 alone (Fig. 2A). Remarkably, BRG1 and CHD7 co-occupied distal sites show significantly more ChIP-Seq signals of H3K27ac, P300 and H3K4me1 compared to both the BRG1 and CHD7 alone sites. These co-occupied distal sites correspond to increased ESC master TF binding (i.e. NANOG, SOX2 and OCT4), and relatively more DNase I hypersensitivity. Moreover, we observed significantly more overlap betweenESC super enhancers defined in Whyte et al., 2013 and BRG1/CHD7 co-localization sites compared to their independent sites (Fig. 2B). Together, these data suggest that the co-localization of BRG1 and CHD7 corresponds to more permissive chromatin regions and many of them are at the loci previously defined as super enhancers.

Next, we analyzed the gene expression changes before and after knockdown (KD) of Brg1 (Fig. 2C) (Ho et al., 2009) and before and after knockout (KO) of Chd7 (Supplementary Fig. S3A) (Schnetz et al., 2010). We found that genes that are targeted by distal BRG1 and CHD7 co-localization show significant up-regulation after Brg1 KO (i.e. Brg1 KD versus WT ESC) or Chd7 KO (i.e. Chd7 KO versus WT ESC) compared to those targeted distally by either BRG1 or CHD7 alone or bound by BRG1 proximally. Analysis of Hi-C-Seq data suggests that distal BRG1 and CHD7 co-localization sites also have increased chromatin interactions compared to each of their alone sites (Supplementary Fig. S3B). Together, these results suggest that BRG1 and CHD7 depend on each other to establish an active chromatin architecture and selectively partner with other ESC master TFs to regulate gene expression programs at distal regulatory elements (Supplementary Fig. S3C).

Acknowledgements

We thank R. Jothi for useful discussion and critical comments. We thank G. Hu for discussion on initial design of the analysis and interpretation of results.

Funding

This work is supported in part by funds from University of Sydney, the New South Wales Ministry of Health, the Human Frontier Science Program (BRG00842014), National Health and Medical Research Council (1103271 and 1113338), National Heart Foundation, the Amazon Web Service Cloud Credits for Research Program. Pengyi Yang is supported by the University of Sydney Postdoctoral Research Scheme.

Conflict of Interest: none declared.

References


Co-dependency of BRG1 and CHD7 in ESC gene regulation

Supplementary: Integrative analysis identifies co-dependent gene expression regulation of BRG1 and CHD7 at distal regulatory sites in embryonic stem cells

Pengyi Yang, Andrew Oldfield, Taiyun Kim, Andrian Yang, Jean Yee Hwa Yang, Joshua W.K. Ho

Supplementary Figure 1. Preferential binding of chromatin remodelling proteins at distal regions.

A. DNA binding proteins are ranked by each of their distal and proximal binding sites ratios. Proteins with a ratio of one have equal number of distal and proximal binding sites. Gene set enrichment test implemented in Limma R package was applied to the set of chromatin remodelling proteins defined in AnimalTFDB. chromatin remodelling proteins that are included in the curated repository (i.e. have been ChIP-Seq profiled in ESCs) are highlighted.

B. ChIP-Seq profiles of Brg1 occupancy in ESC at example genes Klf4, E2f1 and Otx2. ChIP-Seq profiles of histone modifications, DNase I hypersensitivity, and various ESC master TFs and co-factors are included for comparison. Chromatin interactions supported by Rnapii ChIP-PET in ESC are highlighted by vertical yellow bars.
Supplementary Figure 2. Co-localisation of DNA binding proteins in ESC with BRG1.

A. Co-localisation, quantified by Jaccard index, of all other DNA binding proteins with respect to Brg1 at distal and proximal regions, respectively. Those that have a Jaccard index above the dashed red lines are highlighted.

B. Selected panels of factors that show the highest differential co-localisation at distal (brown bar) or proximal (green bar) regions with BRG1.

C. PAD clustering of selected ESC master TFs and differentially co-localized factors with BRG1 at proximal and distal regions, respectively.

D. Venn diagrams showing overlaps of proximal and distal binding site, respectively, between BRG1 and CHD7.
Supplementary Figure 3. Functional characterisation of BRG1 and CHD7 co-localisation and alone sites and proposed regulatory model.

A. Expression comparison of Chd7 KO and WT ESCs with respect to genes targeted by distal BRG1 or CHD7 alone sites or their co-localisation sites; or BRG1 proximal sites. A gene is said to be targeted by a site if its TSS is the closest among all genes to that site. Statistical significance is calculated using two-sided Wilcoxon-rank sum test.

B. Number of chromatin interaction (log2 of HiC-Seq counts) found within 1KB of a binding sites. The same four categories of binding sites as in A are compared and two-sided Wilcoxon-rank sum test is used to calculate statistical significance.

C. Schematic model that summarizes the differences between a BRG1-alone (and CHD7-alone) binding site and a BRG1+CHD7 co-localization site.
Supplementary Table 1. List of public ChIP-Seq data sets. List of public ChIP-Seq data sets generated by various groups that passed the quality control filtering and are included in PAD web resource for subsequent data analysis.