

RESEARCH

Open Access

Ligand-dependent dynamics of retinoic acid receptor binding during early neurogenesis

Shaun Mahony^{1†}, Esteban O Mazzone^{2†}, Scott McCuine³, Richard A Young³, Hynek Wichterle², David K Gifford^{1*}

Abstract

Background: Among its many roles in development, retinoic acid determines the anterior-posterior identity of differentiating motor neurons by activating retinoic acid receptor (RAR)-mediated transcription. RAR is thought to bind the genome constitutively, and only induce transcription in the presence of the retinoid ligand. However, little is known about where RAR binds to the genome or how it selects target sites.

Results: We tested the constitutive RAR binding model using the retinoic acid-driven differentiation of mouse embryonic stem cells into differentiated motor neurons. We find that retinoic acid treatment results in widespread changes in RAR genomic binding, including novel binding to genes directly responsible for anterior-posterior specification, as well as the subsequent recruitment of the basal polymerase machinery. Finally, we discovered that the binding of transcription factors at the embryonic stem cell stage can accurately predict where in the genome RAR binds after initial differentiation.

Conclusions: We have characterized a ligand-dependent shift in RAR genomic occupancy at the initiation of neurogenesis. Our data also suggest that enhancers active in pluripotent embryonic stem cells may be preselecting regions that will be activated by RAR during neuronal differentiation.

Background

Cellular competence, fate determination, and differentiation are influenced by the external signals cells receive. While these external signals can take the form of steroid hormones, protein growth factors, or other molecules, their presence is typically communicated by signal-responsive transcription factors (TFs). The effect of a signal on gene expression, and ultimately on cell fate, depends on where such TFs bind to the genome. Therefore, understanding how signal-responsive TFs are integrated into a dynamic cellular context will further our knowledge of the mechanisms guiding the acquisition of specific cellular identities.

In the developing neural tube, retinoid signaling initiates neural differentiation [1], specifies caudal hindbrain and rostral cervical spinal identity [2,3], and controls patterning and differentiation of spinal motor neurons and interneurons [4-6]. Retinoic acid (RA) is the most

commonly used neuralizing agent during *in vitro* embryonic stem (ES) cell differentiation since exposure to it results in a rapid transition from pluripotent embryoid bodies to committed neuronal precursors. The response to RA during neuronal development is mediated by the action of retinoic acid receptor isoforms (collectively abbreviated here as RARs). It has been proposed that RARs are constitutively bound to target sites in the absence of retinoids [7], recruiting co-repressors such as Ncor1 and Ncor2 [8]. In the presence of the retinoid ligand, RAR (often heterodimerized with RXR) recruits co-activators (Ncoa1 and Ncoa2), p300, and core components of the transcriptional machinery [7]. However, the proposed independence of RAR binding from the presence of the ligand has only been confirmed at a small number of sites.

While some characterization of RAR genomic binding has recently been carried out in mouse ES and human breast cancer cell lines [9-11], it is unknown which genes are targeted by RAR during neurogenesis, and how RAR binding targets are selected. Chromatin accessibility and protein cooperativity may both play roles in restricting the cohort of bound locations under a given

* Correspondence: gifford@mit.edu

† Contributed equally

¹Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 32 Vassar Street, Cambridge, MA 02139, USA
Full list of author information is available at the end of the article

set of cellular conditions. For example, in human breast cancer cell lines, RAR binding is highly coincident with the binding of estrogen receptor (ER) α , FoxA1, and Gata3 [10,11], and FoxA1 is required for RAR recruitment [10]. Recent work has demonstrated that TF binding also correlates with nucleosome-free regions [12], certain histone modifications [13-17], and the occupancy of other regulatory proteins [18,19] in the same cellular conditions. It is not known how these relationships extend through developmental time at individual enhancers. Enhancers may be entirely developmental stage-specific, in which case the sites bound by a regulator in one developmental stage should not be coincident with the sites bound by a subsequent stage-specific TF. Alternatively, enhancers may be reused across developmental time, and the occupancy patterns of regulatory proteins or epigenetic markers may anticipate the future binding of newly activated TFs during differentiation [20,21]. Determining the dynamics of RAR binding during early neuronal development may therefore yield insight into the precise temporal response of cells to retinoid signaling and how enhancers are organized to facilitate this response.

In this study, we examine the genome-wide binding of RARs during RA induced differentiation of ES cells into spinal motor neurons [22]. Retinoid signaling initiates the transition from pluripotency to neurogenesis in this model system, and provides rostral-caudal information to developing motor neurons. By profiling the binding of active RAR isoforms in both the presence and absence of retinoid signaling, we observe that only a small subset of sites are constitutively bound. An additional set of sites is bound only in the presence of RA, and the existence of this set provides a convenient opportunity to examine how pre-RA occupied and post-RA occupied sites correlate with the relatively well-characterized regulatory network in mouse ES cells. We find that binding information for ES cell TFs and other regulatory proteins accurately predicts both constitutive and exclusively post-RA RAR binding. The binding of core ES cell regulators is highly correlated with pre-RA bound RAR sites, slightly less correlated with post-RA bound RAR sites, and much less correlated with the binding of other TFs in further differentiated tissues, arguing that the active regulatory network may be one of the most important determinants of TF binding.

Results

RAR ChIP-seq profiles direct genomic interactions during early differentiation

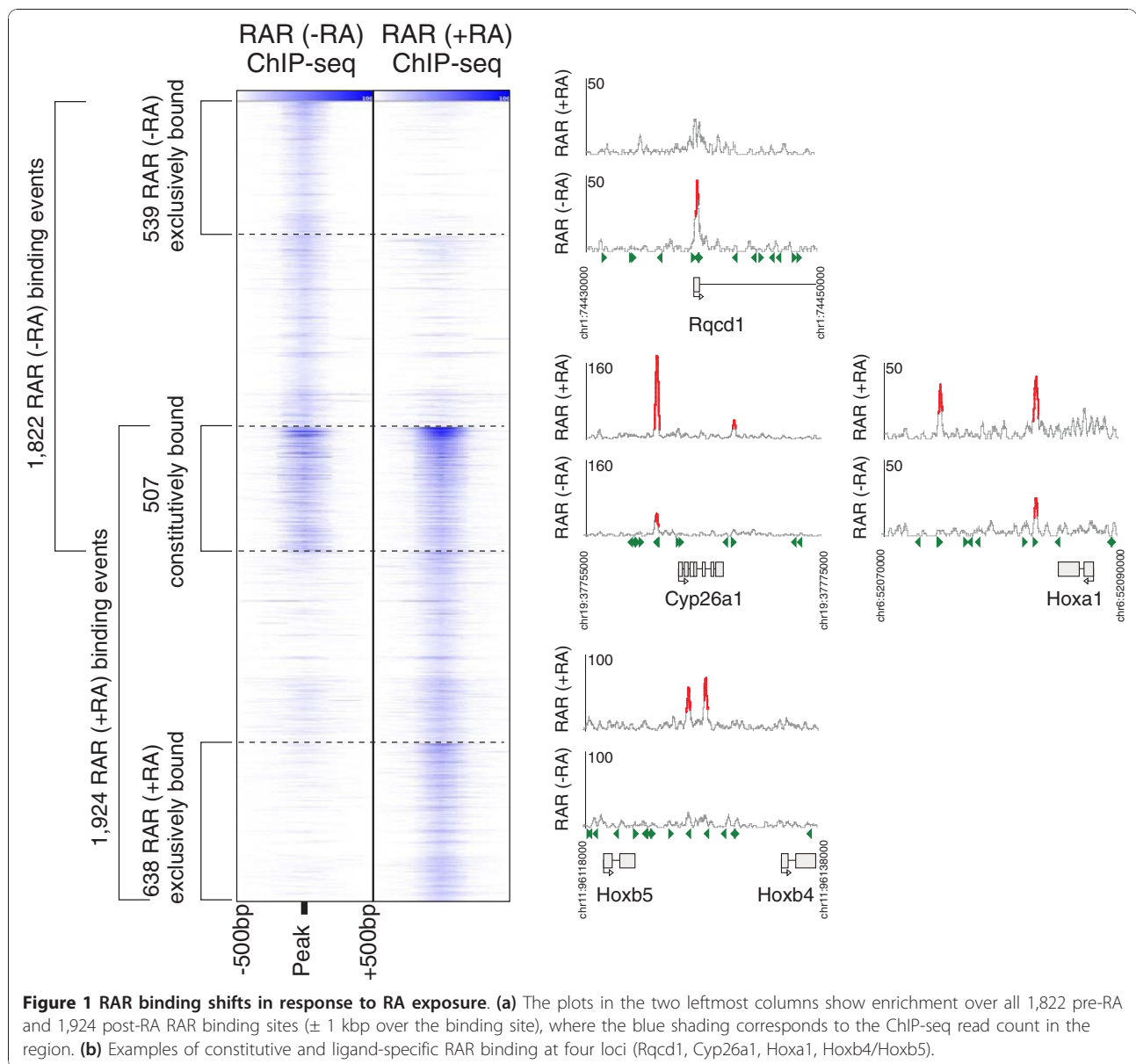
Using a pan-RAR antibody, we profiled the genome-wide occupancy of RAR isoforms in differentiating embryoid bodies after 8 hours of exposure to RA, finding significant ChIP-seq enrichment at 1,924 sites.

We also profiled RAR occupancy in the same developmental stage but in the absence of retinoid signaling, finding 1,822 sites of significant enrichment. A number of previously characterized retinoic acid response elements (RAREs) were observed to be bound in both conditions, including RAREs at *Rarb*, *Hoxa1*, and *Cyp26a1* (Figure 1) [23]. A recent promoter-focused ChIP-chip study of RAR in mouse embryonic stem cells [9] suggested that few RAR binding sites contained 'direct-repeat' hormone response elements. In contrast, we find that high-similarity hormone response element motifs occur at RAR ChIP-enriched sites at a higher rate than that observed in published ChIP-seq studies of other nuclear hormone receptors such as ER α , Esrrb, and Nr5a2 [10,24-26] (Additional file 1). The most frequent motifs at our enriched sites are the direct-repeat motifs with spacers of 5 bp or 2 bp (DR5 and DR2, respectively; Additional file 1), which RAR is known to preferentially bind [23,27]. The binding events with the highest ChIP-enrichment are more likely to contain high-similarity matches to the DR5 and DR2 motifs (Additional file 2), suggesting that many of the most enriched sites represent direct RAR-DNA binding events.

RAR binding shifts in response to RA exposure

In contradiction to the model of RAR constitutively binding to its targets [7], only 507 of the predicted RAR binding events are significantly enriched both in the presence and absence of retinoid exposure, where significant enrichment is defined by our binding event detection methodology (see Materials and methods). Figure 1 presents a clustergram of all sites bound before or after RA exposure, and is arranged according to the pattern of enrichment across both conditions. As the figure indicates, we need to be cautious when determining if a site is bound exclusively in one condition. For instance, some sites display similar enrichment levels across both conditions, but this enrichment level is only deemed significant in one condition (that is, it falls below the significance threshold in the other condition). After further analysis, we define a set of 638 sites that are bound exclusively in the presence of retinoid signaling, as they are not significantly enriched in the absence of RA exposure (compared with control), and their levels of ChIP-seq enrichment are significantly different in the presence and absence of RA (see Materials and methods). Conversely, at least 539 sites are bound only in the absence of retinoid exposure.

Intriguingly, some of the shift in RAR binding sites may be explained by a ligand-dependent shift in RAR's binding preference. Sites bound only in the absence of RA contain more direct repeat motifs with 0-bp or 1-bp spacers than sites bound only in the presence of RA (Additional files 3 and 4). Prior studies have shown that



such motif configurations can be bound by RAR [28,29]. On the other hand, sites bound exclusively in the presence of RA contain more DR5 motifs. These direct repeat motifs are amongst the set of sequence features that have the most significant difference in occurrence frequency between RAR sites bound exclusively in the presence or absence of retinoid signaling (Additional file 5). However, only approximately 14% of exclusively pre-RA sites contain high similarity matches to the DR0 or DR1 motifs, while only 13% of exclusively post-RA sites contain high similarity DR5 motifs. Therefore, a potential shift in RAR's direct binding preference offers only a partial explanation for the observed condition-exclusive binding patterns.

By comparing the relative occurrence of all known TF binding motifs in each condition-exclusive set, we also find that exclusively post-RA sites contain significantly more E-box and ETS-family motifs than exclusively pre-RA sites (Additional file 5). Exclusively post-RA sites also contain more instances of a palindromic motif with consensus sequence 'TCTCGCGAGA'. It is not known which proteins may interact with this motif, although the motif is over-represented in mammalian promoter regions [30], and has recently been characterized as a regulatory sequence [31]. The observation of these over-represented secondary motifs suggests that some of the exclusively post-RA binding sites may occur due to ligand-dependent interactions between RAR and cofactors, or some may

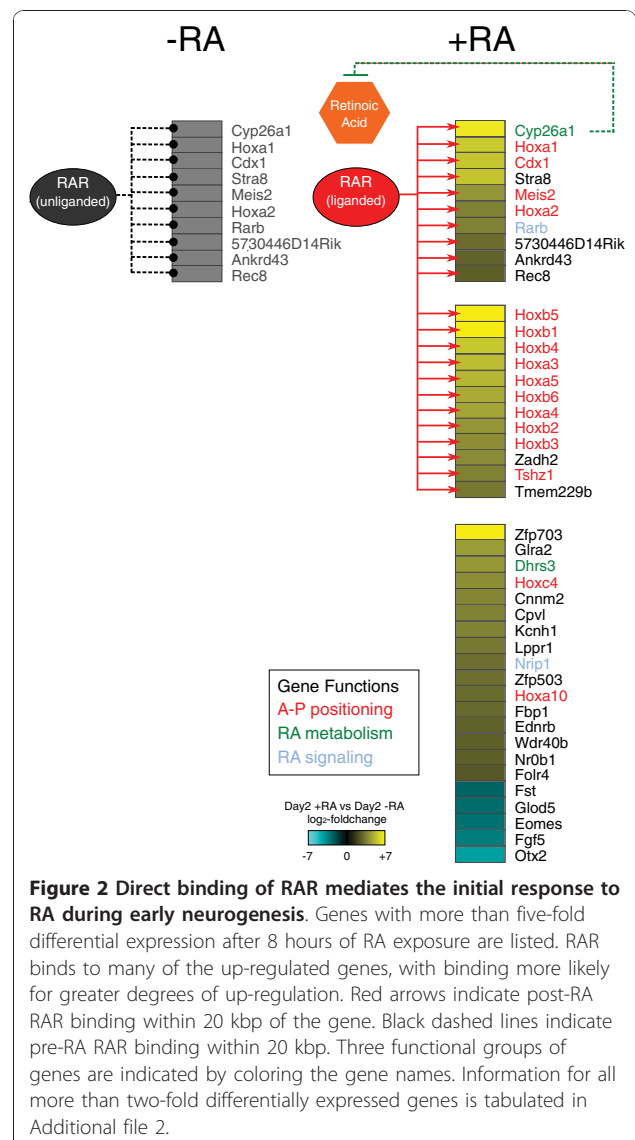
potentially represent indirect binding events caused by enhancer-promoter looping. Most of the motifs with significantly higher relative frequency in the exclusively pre-RA sites are related to DR0 or DR1 patterns.

A compact retinoid response is directly mediated by RAR

In order to determine which RAR binding sites are associated with transcriptional regulation, we characterized the early transcriptional response to retinoid signaling. Despite the dramatic consequences initiated by RA exposure, microarray-based gene expression analysis reveals that only 96 genes are differentially expressed given 8 hours of RA exposure (more than two-fold change, $P < 0.01$; Additional file 6). Of these, 81 genes are up-regulated. The most prevalent theme in the expression response is the acquisition of rostro-caudal identity; 12 anterior Hox genes are significantly up-regulated, along with the Hox co-factors *Meis1*, *Meis2*, *Pbx2*, and other positioning genes such as *Tshz1* and *Cdx1*. While RAR β is up-regulated, the response also attenuates retinoid signaling via the induction of retinoid metabolism genes (*Cyp26a1*, *Dhrs3*, *Rbp1*) and a repressor of RAR, *Nrip1* [32]. Thirty-five significantly up-regulated genes are within 20 kbp of a post-RA RAR binding event, including many of the most differentially expressed genes (Figure 2; Additional file 6). Exclusively post-RA RAR targets are no less associated with differential expression than the constitutively bound targets; while 20 significantly up-regulated genes are nearby constitutively bound RAR sites, 15 up-regulated genes are only bound after RA.

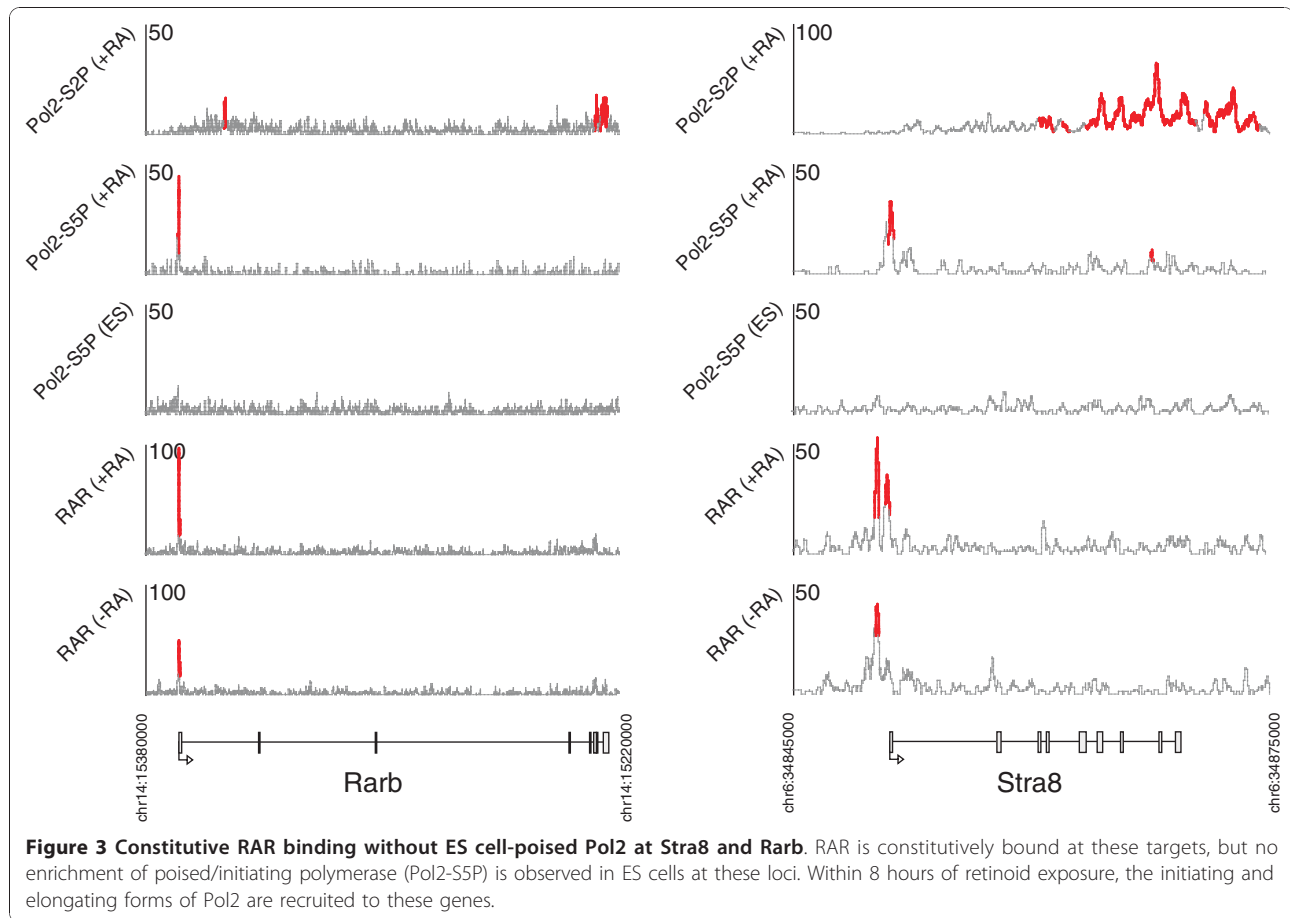
RAR binding is associated with RNA polymerase II initiation

The set of RAR binding sites near differentially expressed genes represents a small proportion of the total complement of post-RA RAR binding sites. It is likely that many other RAR binding sites play regulatory roles during the retinoid response that are not apparent from microarray-based differential expression analysis. We used ChIP-seq to characterize RNA polymerase II (Pol2) initiation (as signified by Pol2 CTD serine 5 phosphorylation, Pol2-S5P [33-35]) and elongation (as signified by Pol2 CTD serine 2 phosphorylation, Pol2-S2P [33-35]) after 8 hours of RA exposure. We identified 3,409 significant Pol2 initiation events, of which 424 were within 5 kbp of post-RA RAR binding events. Of these RAR-associated Pol2-S5P events, 402 (95%) are within 1 kbp of the transcription start sites, or within the gene body, of 269 known genes and non-coding RNAs. Significant enrichment of Pol2-S2P is observed within or at the 3' end of 214 genes (80%) bound by RAR and Pol2-S5P, demonstrating that many of these genes are actively transcribed post-RA (for example, see Figure 3). Therefore, the correlation between RAR



binding and Pol2 initiation and elongation suggests that RAR may play a wider role in driving and maintaining transcription beyond that observed from microarray-based differential expression analysis. We again find no evidence that exclusively post-RA RAR binding sites are less associated with Pol2 initiation than constitutively bound sites; both sets of sites are coincident with Pol2-S5P events at similar rates.

A proposed model of RAR functionality suggests that it acts as a transcriptional repressor in the absence of RA signaling, and becomes an activator after ligand binding [7]. To assess the dynamics of RAR's interactions with Pol2, we compare the post-RA Pol2 ChIP-seq profiles with Pol2-S2P and Pol2-S5P ChIP-seq data from the pluripotent state [36]. Of the 424 RAR-associated Pol2-S5P events characterized post-RA, the majority



(390) are also enriched for Pol2-S5P in the pluripotent state. The pre-RA pattern of RAR binding does not seem to affect the behavior of Pol2 at these sites; both constitutive and exclusively post-RA RAR binding sites are coincident with constitutive Pol2 initiation events at similar rates. From the 214 RAR-bound genes that displayed enrichment for both initiating and elongating Pol2 after RA exposure, 54 (25%) also display evidence of Pol2 elongation in the pluripotent state. Genome-wide, we find a set of only 27 significant Pol2-S5P initiation events that are bound by Pol2 after RA exposure but show no evidence of enrichment in pluripotent cells. Only 11 of these events are near RAR binding events. Surprisingly, this compact set of RAR targets for which Pol2 is not poised in pluripotent cells includes *Hoxa1*, *Cyp26a1*, *Rarb*, and *Stra8* (for example, see Figure 3). Therefore, these critical RA-responsive genes are constitutively bound by RAR, but Pol2 is only recruited to their promoters after RA exposure.

In summary, our examination of potential interactions between RAR and Pol2 before and after retinoid exposure adds complexity to the proposed model of RAR functionality. Only a small set of important retinoid

targets fit the simple model of RAR recruiting Pol2 to the transcription start site only after RA exposure. Many more RAR target genes already have poised Pol2 before retinoid signaling, regardless of whether RAR is constitutively bound. A further set of bound genes is already being actively transcribed before RA exposure.

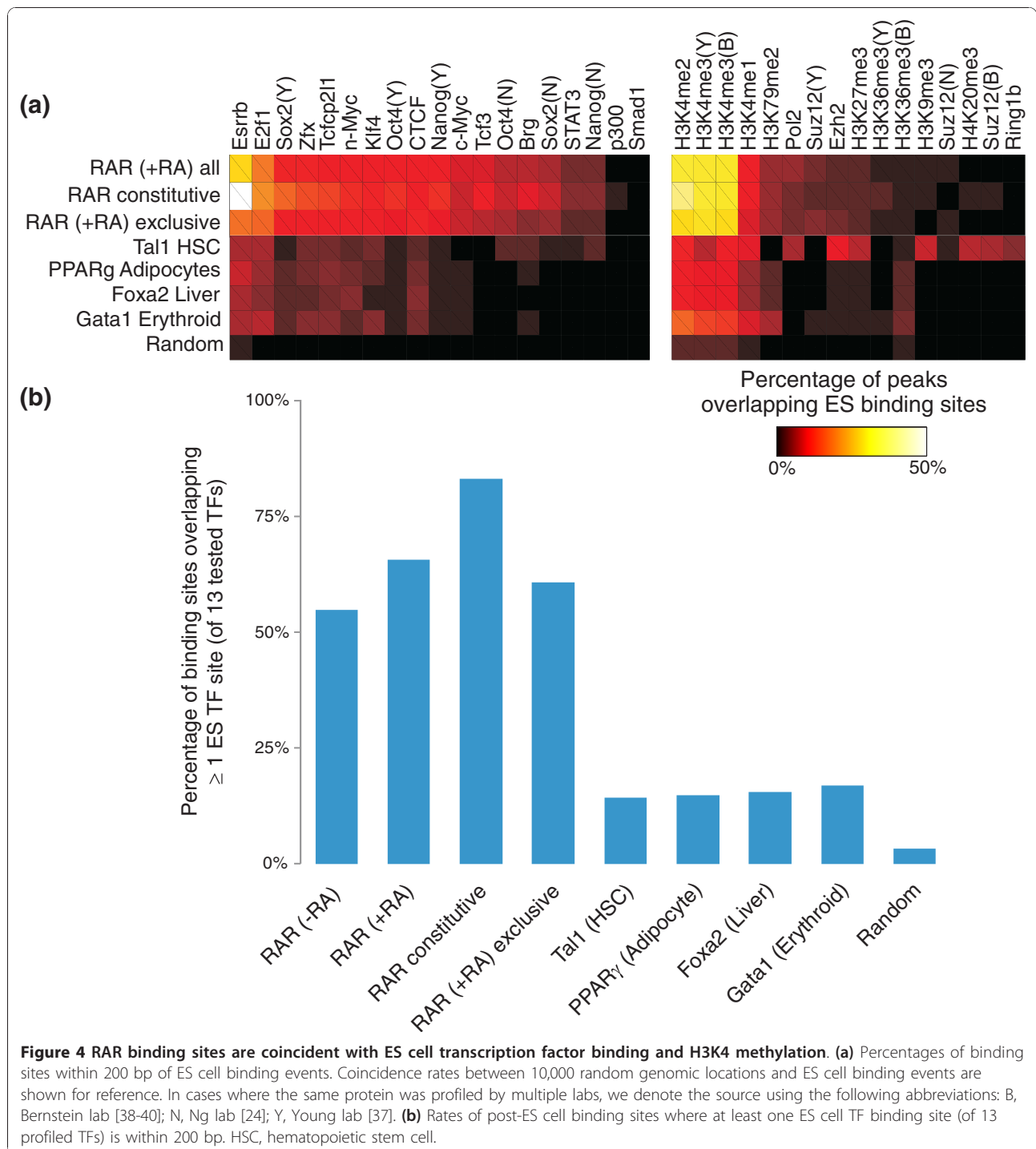
RAR binding is associated with ES cell regulatory state

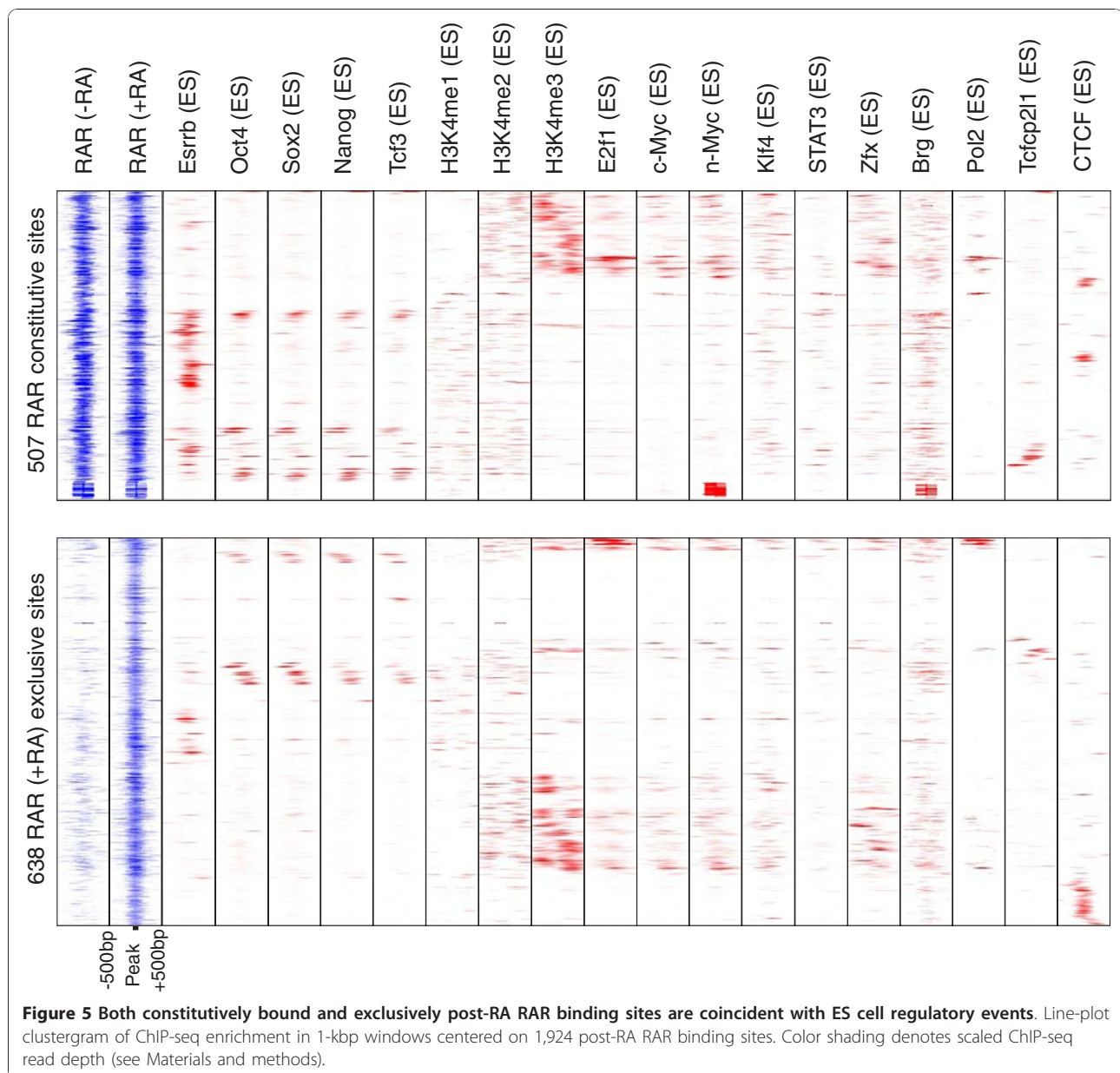
DNA-binding preference alone is not sufficient to explain the specificity of RAR's post-RA genomic occupancy. At least 150,000 high-similarity matches to the DR2 and DR5 motifs do not display significant RAR binding either before or after RA exposure. One possibility is that RAR bound sites are distinguished by their chromatin structure profiles and the occupancy of other regulatory proteins in the surrounding genomic region. To assess the regulatory state of RAR binding sites, we compare constitutively bound sites (by definition occupied both post-RA and in the preceding pluripotent state) to published ChIP-seq data in mouse ES cells, including data for multiple TFs, co-factors, histone modifications, and chromatin modifying proteins [24,37-41].

We observe that the locations of constitutively bound RAR binding sites are highly coincident with the binding sites of many regulatory proteins in ES cells (Figures 4a and 5). While only 3% of randomly selected sites are within 200 bp of at least one ES cell TF binding site, 83% of constitutively bound RAR sites display the same proximity (Figure 4b). Surprisingly, the associations are

not limited to general TFs; many exclusively post-RA RAR sites are coincident with the binding sites of core ES cell state regulators, such as *Esrrb* and *Oct4*.

RAR must recognize the sites bound exclusively post-RA after the established ES cell pluripotent regulatory state has begun to respond to RA exposure. According to the hypothesis that all developmental enhancers are





epigenetically marked at the earliest stages of development [20,21], RAR will bind post-RA to sites that are already bound by other regulators in ES cells. Alternatively, RAR may recognize unbound developmental enhancers that are specific to neuronal fate. We find that 61% of exclusively post-RA RAR binding sites are within 200 bp of at least one known ES cell TF binding site (Figure 4b). Thus, the observed associations between RAR and ES cell TF binding sites suggest that RAR binds to some sites that were bound by stage-specific TFs in the earlier pluripotent state, even at sites to which RAR itself was not bound in that stage. However, the associations between ES cell binding sites and

exclusively post-RA RAR sites are less than those with constitutively bound RAR sites, and thus our observations are not fully consistent with the hypothesis that all developmental enhancers are marked in ES cells.

To further examine the relationships between ES cell regulatory state and later developmental enhancers, we analyzed data from published ChIP-seq experiments performed in unrelated adult or late differentiation cell types: Foxa2 in liver [17], Gata1 in erythroid cells [42], Tal1 in hematopoietic stem cells [43], and peroxisome proliferator activated receptor (PPAR) γ (another nuclear hormone receptor) in adipocyte differentiation [25]. While all of these stage-specific TFs bind to the same

regions as ES cell TFs at a higher rate than expected by chance (Figure 4a), none of them approaches the rate of overlap observed for RAR during early differentiation. Therefore, the relationships between RAR and ES cell TFs do not merely result from all possible enhancers being unveiled by ES cell ChIP-seq data.

ES cell TF binding predicts post-RA RAR binding

The observed relationships between RAR binding and earlier binding events suggest that TF binding information from ES cells can be used to predict where signaling TFs will bind in a proximal developmental state. Predicting if a motif sequence will be bound based on motif similarity alone leads to high rates of additional predictions (Figure 6) [44]; for a motif similarity threshold with which we can correctly predict 500 post-RA bound RAREs, we also predict that approximately 65,000 additional sites should be bound. Recent reports demonstrate the use of co-temporal histone modification ChIP-seq data for predicting TF binding to motif sequences [14,16,45]. We can similarly combine the motif-similarity score with a score based on the sum of normalized read counts from ES cell TF ChIP-seq experiments in 500-bp windows around the sites (see Materials and methods). As shown in Figure 6, this combined score significantly decreases the rate of additional predictions for a given true-positive rate. Using the combined motif and ES cell TF score, we reduce the number of additional predictions 85% (to approximately 9,600) while correctly predicting 500 bound RAREs. We find that ES cell TF binding data outperforms conservation, ES cell p300 ChIP-seq data, and ES cell H3K4

methylation data in predicting which RARE motifs will be bound (Figure 6).

Note that the improvement in predictive performance described above is achieved with a naïve approach that assumes all ES cell TF data sources are equally informative for post-RA RAR binding. We can compare the predictive performance of ES cell TF data sources to that of histone modification information by training a supervised classification technique to classify sites as bound or unbound. Specifically, we trained support vector machines (SVMs) to discriminate between sites that are bound by RAR and a negative set of 10,000 unbound sites. As shown in Table 1, test set SVM performance is highest when making use of all available ES cell data. SVMs trained using the same ES cell data sources perform worse when predicting PPAR γ binding in adipocytes or Foxa2 binding in liver (Table 1).

Interestingly, our SVM results suggest that the ES cell TF binding landscape is more informative than ES cell histone modification data when predicting the genomic locations that are bound by signal-responsive TFs. SVMs that are trained using only ES cell TF binding data offer higher classification performance of bound sites than SVMs that are trained using only ES cell histone modification data. This observation holds true when predicting sites that are only bound by RAR before or after RA exposure.

Discussion

By profiling the dynamics of RAR occupancy at the initiation of neurogenesis, we have characterized a ligand-dependent shift in binding targets. This shift in binding targets is relevant to RAR's role in gene regulation, as both constitutively and exclusively post-RA bound sites are associated to a similar degree with gene expression and polymerase recruitment. Recent analyses of RAR binding profiled genome-scale occupancy only in the presence of retinoids, and thus did not observe a ligand-dependent shift in binding [9-11]. Indeed, on the basis of a small number of ChIP-quantitative PCR experiments, Delacroix *et al.* [9] suggested that most

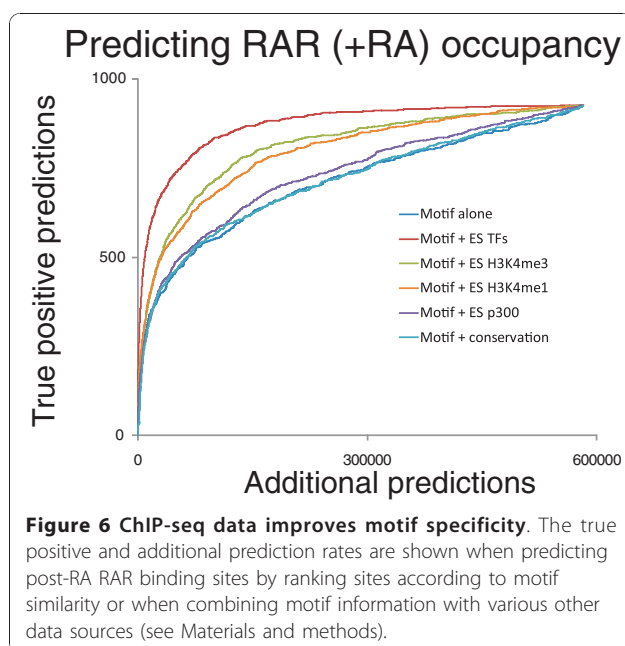


Table 1 Motif occupancy classification performance using ES cell ChIP training data

Binding sites	All ES cell experiments	ES cell TF experiments	ES cell histone modifications
RAR (constitutively bound)	0.96	0.92	0.81
RAR (post-RA exclusively bound)	0.81	0.77	0.73
PPAR γ (adipocytes)	0.62	0.58	0.53
Foxa2 (liver)	0.63	0.56	0.50

Performance is measured as receiver operating characteristic (ROC) area under curves for SVMs trained to discriminate between significant binding sites and randomly selected unbound locations.

RAR binding sites are occupied both in the presence and absence of retinoids.

Some of RAR's shift in binding may be explained by ligand-dependent binding preference or ligand-dependent interactions between RAR and co-activators or co-repressors. In addition, a mixture of RAR isoforms is active at the initiation of neurogenesis, and changes in the composition of this mixture may lead to changes in binding occupancy. For example, RAR β is activated after retinoid exposure, and may have different binding preferences or cofactor interactions from the isoforms active in the absence of RA (RAR γ and RAR α). Preliminary evidence suggests that the pan-RAR antibody has limited affinity for RAR β , as we have not had success using this antibody for CHIP experiments at later points in development when RAR β becomes the dominant isoform (data not shown). However, given the pan-RAR antibody vendor specifications, we cannot exclude the possibility that some of the exclusively post-RA binding sites may be attributed to RAR β binding.

We have also found that the binding sites of RAR after RA signaling are extensively associated with the binding of other regulatory proteins in the temporally preceding pluripotent environment. Furthermore, we have demonstrated that we can accurately predict where RAR will bind in the genome given knowledge of the preceding regulatory state. The apparent dependence of RAR binding on prior cellular state suggests that the response of differentiating cells to external signals may be context and developmental-stage dependent, with some future binding events being potentiated by current genomic occupancy patterns.

The causal relationships underlying the association between RAR binding and the ES cell regulatory network remain unclear, so we can only summarize possible explanations for the observed data. CHIP-seq data from ES cells may provide a read-out of accessible regions of the genome, thereby indicating which regions are amenable to TF binding in that environment. Since the predictive capacity of ES cell regulatory data decreases with temporal distance from ES cell state (Table 1), we do not believe that ES cell CHIP-seq data merely serves as an indicator of all enhancers that may be bound under any condition or cell type. Rather, the regions bound by regulatory proteins in a given developmental stage may be more likely to remain accessible for TF binding in a related future stage. Direct cooperation between RAR and TFs active in ES cells may also account for some coincident binding sites. Of all tested data sources, Esrrb binding in ES cells is the most correlated with RAR occupancy before and after RA exposure. Esrrb is an orphan nuclear receptor that binds to hormone response element motifs. It is therefore possible that Esrrb heterodimerizes or otherwise directly

cooperates with RAR at direct repeat hormone response element (HRE) motifs, facilitating stable binding events before and/or after RA signaling. However, direct interactions between Esrrb and RAR are not required for cooperativity to arise. For example, Esrrb could maintain chromatin accessibility at some direct repeat HREs until RAR binds after retinoid exposure. All of RAR's associations with ES cell core regulators cannot be explained by Esrrb occupancy alone; as shown in Figure 5, many RAR binding sites are associated with the binding of ES cell TFs other than Esrrb.

The observation that RAR binding is correlated with the occupancy of other regulatory proteins is supported by other recent CHIP studies of RAR. Delacroix *et al.* [9] demonstrate cell-type specific RAR occupancy in mouse ES cells and mouse embryonic fibroblasts, which correlates with cell-type-specific H3K4me3 patterns. Both Hua *et al.* [10] and Ross-Innes *et al.* [11] show that RAR and ER α colocalize at many regions in a human breast cancer cell line (MCF-7). Hua *et al.* [10] also demonstrate that many RAR and FoxA1 binding sites coincide in MCF-7 cells, and that RAR binding is decreased at such sites when FoxA1 is knocked down. Therefore, RAR may preferentially bind to RARE motifs that are made accessible by the binding of other TFs or chromatin modifying proteins.

A number of previous studies have demonstrated that certain regulatory information may be used to predict co-temporal TF occupancy. For example, enrichment of p300 [18], H3K4me1 [17,45], H3K4me3 [15,45], and regions of open chromatin (as assayed by DNaseI hypersensitivity [12,46]) have each been correlated with the binding of TFs in ES cells and other tissues. Ours is the first demonstration that regulatory information in a given cell type may be used to predict future TF binding events. Furthermore, the markers examined in the previous studies are typically associated with active enhancers. In our study, we use all available information to predict any RAR binding event, regardless of its association with transcription. Our rationale is that binding events that do not produce co-temporal transcription are not necessarily neutral, especially in the context of differentiation. For example, binding events that do not produce transcription under one set of conditions may disrupt chromatin structure enough to allow different proteins to bind to proximal sites during a future developmental stage.

Conclusions

We have described a compact transcriptional response to RA at the initiation of neurogenesis, which may be potentiated by associations between RAR and earlier regulatory events. As more regulatory data are collected from a greater diversity of cell types and developmental

stages, it will be of interest to further elucidate temporal dependencies between the genomic occupancy of regulatory proteins. Indeed, exploring such temporal networks of binding events may lead to greater understanding of the influences on cell fate during differentiation.

Materials and methods

Cell culture and motor neuron differentiation

ES cells were differentiated as previously described [22]. Briefly, ES cells were trypsinized and seeded at 5×10^5 cells/ml in ANDFK medium (Advanced DMEM/F12: Neurobasal (1:1) medium, 10% knockout-SR, Pen/Strep, 2 mM L-glutamine, and 0.1 mM 2-mercaptoethanol) to initiate formation of embryoid bodies (day 0). Medium was exchanged on days 1, 2 and 5 of differentiation. Patterning of embryoid bodies was induced by supplementing media on day 2 with 1 μ M all-*trans*-RA (Sigma, St. Louis, MO, USA) and 0.5 μ M agonist of hedgehog signaling (SAG, Calbiochem, La Jolla, CA, USA). For ChIP experiments, the same conditions were used but scaled to seed 1×10^7 cells on day 0.

Expression analysis

Total RNA was extracted from ES cells or embryoid bodies using Qiagen RNeasy kit (Qiagen, Valencia, CA, USA). For quantitative PCR analysis, cDNA was synthesized using SuperScript III (Invitrogen, Carlsbad, CA, USA) and amplified using SYBR green brilliant PCR amplification kit (Stratagene, La Jolla, CA, USA) and Mx3000 thermocycler (Stratagene). For GeneChip expression analysis, RNA was amplified using Ovation amplification and labeling kit (NuGen, San Carlos, CA, USA) and hybridized to Affymetrix Mouse Genome 430 2.0 microarrays. Expression microarray experiments were performed in biological triplicate for each analyzed time point. Arrays were scanned using the GeneChip Scanner 3000. Data analysis was carried out using the affymGUI BioConductor package [47]. GC Robust Multi-array Average (GCRMA) normalization [48] was performed across all arrays, followed by linear model fitting using Limma [49]. Differentially expressed genes after 8 hours of RA treatment were defined by ranking all probesets by the moderated *t*-statistic-derived *P*-value (adjusted for multiple testing using Benjamini and Hochberg's method [50]) and setting thresholds of *P* < 0.01 and a fold-change of at least 2. All arrays were submitted to the NIH Gene Expression Omnibus (GEO) database under accession number [GEO:GSE19372].

ChIP-seq protocols

ChIP protocols were adapted from [51]. Descriptions of these protocol modifications have been previously published [52]. Briefly, approximately 6×10^7 cells taken from each developmental time point were cross-linked

using formaldehyde and snap-frozen in liquid nitrogen. Cells were thawed on ice, resuspended in 5 ml lysis buffer 1 (50 mM Hepes-KOH, pH 7.5, 140 mM NaCl, 1 mM EDTA, 10% glycerol, 0.5% NP-40, 0.25% Triton X-100) and mixed on a rotating platform at 4°C for 5 minutes. Samples were spun down for 3 minutes at 3,000 rpm, resuspended in 5 ml lysis buffer 2 (10 mM Tris-HCl, pH 8.0, 200 mM NaCl, 1 mM EDTA, 0.5 mM EGTA), and mixed on a rotating platform for 5 minutes at room temperature. Samples were spun down once more, resuspended in lysis buffer 3 (10 mM Tris-HCl, pH 8.0, 100 mM NaCl, 1 mM EDTA, 0.5 mM EGTA, 0.1% Na-deoxycholate, 0.5% N-lauroylsarcosine) and sonicated using a Misonix 3000 model sonicator to shear cross-linked DNA to an average fragment size of approximately 500 bp. Triton X-100 was added to the lysate after sonication to final concentrations of 1% and the lysate spun down to pellet cell debris. The resulting whole-cell extract supernatant was incubated on a rotating mixer overnight at 4°C with 100 μ l of Dynal Protein G magnetic beads that had been preincubated for 24 hours with 10 μ g of the appropriate antibody in a phosphate-buffered saline/bovine serum albumin solution. Pan-RAR (Santa Cruz Biotechnology, Santa Cruz, CA, USA, sc-773), Pol2-S5P (Abcam, [Cambridge, UK, ab5131]), and Pol2-S2P (Abcam, H5 clone ab24758) antibodies were used for ChIP experiments. After approximately 16 hours of bead-lysate incubation, beads were collected with a Dynal magnet. ChIP samples probing for TF binding were washed with the following regimen, mixing on a rotating mixer at 4°C for 5 minutes per buffer: low-salt buffer (20 mM Tris at pH 8.1, 150 mM NaCl, 2 mM EDTA, 1% Triton X-100, 0.1% SDS), high-salt buffer (20 mM Tris at pH 8.1, 500 mM NaCl, 2 mM EDTA, 1% Triton X-100, 0.1% SDS), LiCl buffer (10 mM Tris at pH 8.1, 250 mM LiCl, 1 mM EDTA, 1% deoxycholate, 1% NP-40), and TE containing 50 mM NaCl. ChIP samples probing for histone and chromatin marks were washed four times with RIPA buffer (50 mM Hepes-KOH, pH 7.6, 500 mM LiCl, 1 mM EDTA, 1% NP-40, 0.7% Na-deoxycholate) and then once with TE containing 50 mM NaCl, again mixing on a rotating mixer at 4°C for 5 minutes per buffer. After the final bead wash, samples were spun down to collect and discard excess wash solution, and bound antibody-protein-DNA fragment complexes were eluted from the beads by incubation in elution buffer at 65°C with occasional vortexing. Cross-links were reversed by overnight incubation at 65°C. Samples were digested with RNase A and Proteinase K to remove proteins and contaminating nucleic acids, and the DNA fragments precipitated with cold ethanol. Purified DNA fragments were processed according to a modified version of the Illumina/Solexa sequencing protocol [53].

Raw sequencing data (FASTQ format) were submitted to the NIH GEO/Sequence Read Archive database under accession number [GEO:GSE19409].

Third-party ES cell ChIP-seq datasets

FASTQ files containing raw sequence and quality information were downloaded from the Short Read Archive [54].

The following published mouse ES cell experimental datasets were used to predict binding occupancy: c-Myc, CTCF, E2f1, Esrrb, Klf4, Nanog, n-Myc, Oct4, p300, Smad1, Sox2, STAT3, Suz12, Tcfcp2l1, Zfx, and green fluorescent protein control as published in Chen *et al.* [24]; Nanog, Oct4, Sox2, Tcf3, Suz12, H3K36me3, H3K4me3, H3K79me2, whole-cell extract (WCE) control as published in Marson *et al.* [37]; RNA-Pol2, H3K27me3, H3K36me3, H3K4me3, H3K9me3, H4K20me3, H3 control, WCE control as published in Mikkelsen *et al.* [38]; Ezh2, Ring1b, Suz12 as published in Ku *et al.* [39]; H3K4me1, H3K4me2 as published in Meissner *et al.* [40]; Brg, IgG control as published in Ho *et al.* [41].

Mouse ES cell Pol2-S5P and Pol2-S2P ChIP-seq experiments as published in Rahl *et al.* [36] were compared to Pol2 phosphorylation data generated by our study.

In addition, ChIP-seq experiments for other nuclear receptors were used in the construction of Additional file 1: mouse ES cell Nr5a2 as published in Heng *et al.* [26]; mouse adipocyte PPAR γ and RXR as published in Nielsen *et al.* [25]; human MCF-7 ER α and RAR α as published in Hua *et al.* [10].

ChIP-seq data analysis

Sequence reads were aligned to the mouse genome (version mm8) using Bowtie [55] version 0.9.9.2 with options -k 2 -best. Only uniquely mapping reads were analyzed further. Multiple hits aligning to the same nucleotide position were discarded above the level expected at a 10^{-7} probability from a per-base Poisson model of the uniquely mappable portion of the mouse genome. In practice, this caps the number of hits that start at the same nucleotide to three in the RAR ChIP-seq experiments.

Binding event detection for RAR, Pol2-S5P, and various published TF ChIP-seq experiments was carried out using a customized methodology that uses statistical significance testing to find regions producing an over-abundance of sequenced reads in the signal experiments compared with the control. The algorithm is run twice across the data. The first pass estimates a scaling factor for control sequencing read depth and a model of the distribution of sequencing read alignment hits around binding events. The second pass applies these parameters to predict a final set of significant events. Before the first pass, the scaling factor is initialized to be the ratio of

total hit counts between the signal and control channels. The binding distribution model is initialized to be an empirical distribution estimated around predicted binding events in Oct4 ChIP-seq data [37] (Additional file 7).

All alignment hits are extended in both 3' and 5' directions, mirroring the observed distribution of hits around binding events. The extension magnitudes are set equal to the positions where the binding model distribution intersects a uniform distribution over the same area (Additional file 7). Control channel hit counts are scaled using the signal-control scaling factor. A sliding window of bin width 50 bp and offset 25 bp is run over the genome. Overlapping extended hit counts are calculated for both the signal and (scaled) control channels.

The background distribution of ChIP-seq hits is modeled as a non-homogenous Poisson process with parameters estimated from the scaled control hit counts. Specifically, the Poisson parameter λ is chosen as the maximum mean overlapping hit count in 50-bp windows of those observed from: i) the entire genome; ii) a 5-kbp window centered on the current location; and iii) a 10-kbp window centered on the current location. The use of this dynamic background model is motivated by the desire to correct local ChIP-seq enrichment biases that appear in the signal and control channels, and is similar to the model employed by MACS [56]. A given bin is denoted as potentially enriched if the overlapping hit count exceeds that expected from the background model at a P -value of 10^{-9} .

P -values for each potentially enriched bin's over-representation in the signal channel over the control are calculated using the binomial distribution CDF [57]. Neighboring regions in the set of potentially enriched regions are merged, and the maximal P -value observed for the constituent bins is attached to the resulting merged region. The P -values are corrected for multiple hypothesis testing using Benjamini and Hochberg's method, and all regions with corrected P -values above 0.001 are discarded. False-discovery rates are estimated by repeating the event discovery procedures after swapping the scaled control channel and the signal channel.

After the first pass, the scaling factor is estimated by carrying out linear regression on the hit counts observed in 10,000-bp windows that are devoid of potentially significant events in both the signal and control channels. The binding model is estimated from enriched regions with P -values $< 10^{-7}$ and signal/control hit count enrichment > 10 . These regions are aligned around the 'peak' location, defined as the position of maximum probability when scanning the current binding model over the region's hit landscape.

The above technique was also used when estimating enriched 'domains' for histone modifications, certain chromatin-associated proteins, and Pol2-S2P. However,

to capture broader domains of enrichment, the bin width was set to 500 bp and the bin offset to 250 bp. The background model was also set to a homogeneous Poisson threshold estimated from the entire genome. The significance threshold was raised to $P < 0.01$ for domain calling.

DNA motif analysis

De novo motif finding was performed in 200-bp windows centered on the 200 top-ranked peaks for each examined CHIP-seq experiment. SOMBRERO [58,59], AlignACE [60], BioProspector [61], and Weeder [62] were each run on the sequences. SOMBRERO was run for all even motif lengths between 8 and 22 bp with default settings apart from a complexity threshold of 0.01. A third-order Markov model of the mouse genome (version mm8) was employed as background, and a prior based on known mammalian TF binding motifs was also used [59]. AlignACE, BioProspector, and Weeder were run with default settings. STAMP [63] was used to cluster the discovered motifs and remove degeneracy in the results. STAMP was used to match two of the non-HRE motifs to the binding preference of Sp1, and to the 'M8' motif reported by Xie *et al.* [30] in a genome-wide scan of promoter sequences.

Log-likelihood scoring thresholds for the discovered DR5 and DR2 motifs were calculated by simulating 1,000,000 100-bp sequences using a third-order Markov model of the mouse genome (mm8 version). The motif scoring thresholds that yield false discovery rates of 1%, 0.5%, and 0.1% in this set of sequences were recorded.

The analysis of HRE motif frequency shown in Additional files 1 and 3 was performed using a model HRE half-site generated by aligning the half-sites in a curated database of confirmed HRE binding sites (NHRscan [64]). An arbitrary log-likelihood threshold of 5.0 was used to find matches to the half-site motif, which has the effect of matching the following half-site sequences that appear in the NHRscan database: TGACCT, TGACCC, TGA ACT, TGTCCCT, TGCCCT, TGAACC, TGGCCT, TGTCCC, TGATCT, TGACCA, TGACCG, TGCCCC, TGTACT, GGACCT, AGACCT, TGGCCC, TGAGCT, TCACCT, TGATCC, TAACCT, TGGACT. When scoring dimers, the same scoring threshold was used for both halves of the site, and the spacer sequence was unpenalized. These criteria are relatively strict; only 44% of confirmed HRE sites in the NHRscan database pass these thresholds.

Comparison of binding site sets

A 200-bp window was used to define coincident locations between post-ES cell binding sites and ES cell binding sites or domains. The expected coincidence rates were calculated using a set of 10,000 randomly chosen genomic locations

that are not located within 500 bp of any of the tested post-ES cell binding sites and also lie within 500-bp windows that are at least 80% uniquely mappable at a 26-mer resolution.

When calculating the rate of binding sites that are within 200 bp of at least one ES cell TF binding site, binding sites from the following 13 ES cell TF CHIP-seq experiments were used: the Young lab experiments for Oct4, Sox2, Nanog and Tcf3, and the Ng lab experiments for CTCF, c-Myc, n-Myc, E2f1, Esrrb, STAT3, Klf4, Smad1, and Zfx.

The clustergrams in Figures 1 and 5 were generated by plotting the overlapping read counts (where reads have been artificially extended to 200 bp) in 1-kbp windows centered on RAR peaks. The ordering of peaks was determined by clustering 50-bp-binned data using Matlab's clustergram function and the optimal leaf-ordering algorithm [65]. Since the various examined CHIP-seq experiments have different dynamic ranges and degrees of typical enrichment, a single color scale is not appropriate for all tracks. The saturation colors in Figure 5 are therefore chosen such that only 0.001% of the 50-bp windows in the genome will have overlapping read counts in excess (excluding those regions that have non-random accumulations of reads in sequenced control channels). The saturation thresholds for the tracks in Figure 5 are as follows: RAR (day 2 -RA) = 35, RAR (day 2 + RA) = 44 reads, H3K4me1 = 31 reads, H3K4me2 = 47 reads, H3K4me3 = 154 reads, Brg = 27 reads, c-Myc = 80 reads, CTCF = 71 reads, E2f1 = 489 reads, Esrrb = 329 reads, Klf4 = 57 reads, n-Myc = 67 reads, p300 = 18 reads, STAT3 = 50 reads, Tcfcp2l1 = 473 reads, Zfx = 92 reads, Nanog = 221 reads, Oct4 = 132 reads, Sox2 = 216 reads, Tcf3 = 125 reads.

Constitutively bound and ligand-dependent RAR binding sites

A post-RA RAR binding site is defined as constitutively bound if it is within 200 bp of a significant binding site estimated in the pre-RA RAR experiment. As outlined in the main text, an RAR binding site is defined as being exclusively bound post-RA if it fulfills the criteria of being: i) significantly enriched post-RA in relation to the WCE control; ii) not significantly enriched pre-RA in relation to the WCE control; and iii) significantly enriched post-RA in relation to the pre-RA signal. The third criterion here entails performing peak-finding analysis for the post-RA RAR CHIP-seq experiment as described above, but substituting the pre-RA experiment for the WCE control. The aim is to ensure that a post-RA binding site does not display any CHIP-seq enrichment before RA exposure; for example, we wish to exclude events that display CHIP-seq enrichment just below the threshold of statistical significance in the pre-RA experiment. Equivalent procedures are carried out

for finding condition-specific pre-RA RAR and Pol2-S5P binding sites.

Motif specificity analysis

The motif specificity analysis presented in Figure 6 is based on genome-wide matches to the DR2 or DR5 motifs. A scoring threshold for these motifs was chosen such that one-third of all RAR (day 2 +RA) peaks had a match to either motif within 100 bp of the peak position. These criteria yield 929 'true-positive' hits to the DR2/5 motifs proximal to 644 peaks. The same scoring thresholds yield over 877,000 matches to the DR2/5 motifs throughout the mouse genome. This set of positions was filtered for those that do not overlap RAR binding events (within 200 bp) and are located within 500-bp windows that are at least 80% uniquely mappable at a 26 bp resolution. Thus, the set of 'additional predictions' contains 582,612 positions. Some of these additional predictions may serve as binding sites for other TFs, or indeed for RAR under different cellular conditions. However, the vast majority are expected to be false positive predictions.

The log-likelihood similarity score to DR2/5 motifs was used to rank the predictions, and the additional prediction rate for each true positive rate is plotted in Figure 6. Next, for various ES ChIP-seq experiments, the number of reads contained in a 500-bp window surrounding each motif match was counted, and these counts were normalized according to the total uniquely mapped read count for the experiment of interest. For each motif match, normalized read counts were summed across two collections of experiments; the ES cell TF collection (the 16 sequence-specific TFs) and the ES cell H3K4 collection (H3K4me1/2/3 from the Bernstein lab, and H3K4me3 from the Young lab). The motif matches were then ranked separately for each collection according to the summed normalized read counts. Average phastCons [66] conservation scores were also calculated for a 20-bp window around each motif match, and this scoring was again used for ranking (50-bp, 100-bp, and 200-bp windows were also tested for generating phastCons scores, without any increase in performance). To generate a combined ranking for two data sources (for example, motif + ES cell TFs), we normalized the ranking such that each motif match was assigned a score between 0 and 1 for each data source. The product of both scores is taken for each motif match, and the matches are again ranked based on this score. Again, the additional prediction rate for each true positive rate is plotted in Figure 6 for each combined score. We note that the described method for combining motif and ES cell data scores is simplistic, and more sophisticated schemas for incorporating knowledge of ES cell ChIP-seq data may attain much greater improvements in motif specificity than presented in Figure 6.

Support vector machine classification of sites

SVMs were implemented using the libSVM R library [67], and trained using default settings (C-classification, RBF kernels). Positive training sets were generated by randomly selecting 500 bound sites from each set of predicted peaks. A negative set of 10,000 randomly chosen sites was also defined. The negative set sites are located within 500-bp windows that are at least 80% uniquely mappable at a 26-bp resolution, and do not overlap any sites bound by the test TFs (that is, RAR, PPAR γ , Foxa2). ChIP-seq read counts were extracted from 500-bp windows surrounding the positive and negative positions. During each SVM training run, 50 positive sites and 50 negative sites were randomly extracted as test data. The SVMs were trained on the remaining data, and predictive performance was tested on the held out data. SVM training was repeated 100 times, and the averages of the resulting receiver operating characteristic (ROC)-area under curves (AUC) values are reported in Table 1.

Additional material

Additional file 1: Supplementary Figure S1. Enrichment of HRE motifs with various configurations and half-site spacer lengths under ChIP-enriched regions for various nuclear hormone receptor TFs [10,24-26]. The bar-charts show the frequencies of peaks containing each HRE motif configuration within 50 bp of the top 20% of binding sites using strict similarity to a model HRE half-site (see Materials and methods).

Additional file 2: Supplementary Figure S2. Higher-ranked RAR binding sites are more likely to contain DR2/5 motifs. The figure shows the cumulative proportion of ranked peaks that contain matches to the DR2 or DR5 motifs at a threshold set using the 0.5% false positive rate.

Additional file 3: Supplementary Figure S3. Enrichment of HRE motifs with various configurations and half-site spacer lengths at the top 100 constitutive and ligand-specific RAR binding sites (as in Additional file 1).

Additional file 4: Supplementary Figure S4. RAR binding shifts in response to retinoic acid exposure. The plots in the two leftmost columns show enrichment over all constitutive and ligand-specific RAR binding sites (\pm 1 kbp over the binding site), where the blue shading corresponds to the ChIP-seq read count in the region. The plots to the right show matches to motifs over the same regions, with three motif similarity thresholds represented by green color shading (0.5% false positive rate motif scoring thresholds).

Additional file 5: Supplementary Table S1. Differential frequencies of motifs in exclusively post-RA binding sites compared with exclusively pre-RA binding sites, and vice versa. Only motifs with a P -value < 0.05 are shown. The motif names have prefixes denoting their source, as follows: T = TRANSFAC, J = Jaspar, X = Xie et al. [30], U = UniProbe.

Additional file 6: Supplementary Table S2. List of 96 differentially expressed genes (> 2-fold, P < 0.01) between day 2 + 8 hours RA and day 2. Tick marks denote the presence of RAR binding sites within 20 kbp of a gene's transcription start site in the presence or absence of RA.

Additional file 7: Supplementary Figure S5. Empirically estimated distribution of Oct4 ChIP-seq hits around predicted peaks. The uniform expectation over the same area is shown as a dashed red line.

Abbreviations

bp: base pair; ChIP: chromatin immunoprecipitation; DR: direct-repeat motif; ER: estrogen receptor; ES: embryonic stem cell; GEO: Gene Expression

Omnibus; HRE: hormone response element; kbp: kilo-base-pair; Pol2: RNA polymerase II; Pol2-S2P: Pol2 CTD serine 2 phosphorylation; Pol2-S5P: Pol2 CTD serine 5 phosphorylation; PPAR: peroxisome proliferator activated receptor; RA: retinoic acid; RAR: retinoic acid receptor; RARE: retinoic acid response element; SVM: support vector machine; TF: transcription factor; WCE: whole-cell extract.

Acknowledgements

EOM is the David and Sylvia Lieb Fellow of the Damon Runyon Cancer Research Foundation (DRG-1937-07). SM, SMCc, and the work were supported by NIH grant P01 NS055923 (DKG, RAY, HW).

Author details

¹Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 32 Vassar Street, Cambridge, MA 02139, USA.

²Departments of Pathology, Neurology, and Neuroscience, Center for Motor Neuron Biology and Disease, Columbia University Medical Center, 630 West 168th St, New York, NY 10032, USA. ³Whitehead Institute for Biomedical Research, Nine Cambridge Center, Cambridge, MA 02142, USA.

Authors' contributions

All authors participated in the design of the study. EOM performed cell culture, motor neuron differentiation, and gene expression profiling. SMCc performed the ChIP-seq experiments. SM performed and performed all computational analyses. SM, EOM, HW and DKG drafted the manuscript. All authors read and approved the final manuscript.

Received: 13 October 2010 Revised: 10 December 2010

Accepted: 13 January 2011 Published: 13 January 2011

References

- Corral RDD, Storey KG: **Opposing FGF and retinoid pathways: a signalling switch that controls differentiation and patterning onset in the extending vertebrate body axis.** *BioEssays* 2004, **26**:857-869.
- Niederreither K, Vermot J, Schuhbauer B, Chambon P, Dollé P: **Retinoic acid synthesis and hindbrain patterning in the mouse embryo.** *Development* 2000, **127**:75-85.
- Liu JP, Laufer E, Jessell TM: **Assigning the positional identity of spinal motor neurons: rostrocaudal patterning of Hox-c expression by FGFs, Gdf11, and retinoids.** *Neuron* 2001, **32**:997-1012.
- Sockanathan S, Jessell TM: **Motor neuron-derived retinoid signaling specifies the subtype identity of spinal motor neurons.** *Cell* 1998, **94**:503-514.
- Novitsch B, Wichterle H, Jessell T, Sockanathan S: **A requirement for retinoic acid-mediated transcriptional activation in ventral neural patterning and motor neuron specification.** *Neuron* 2003, **40**:81-95.
- Pierani A, Brenner-Morton S, Chiang C, Jessell TM: **A sonic hedgehog-independent, retinoid-activated pathway of neurogenesis in the ventral spinal cord.** *Cell* 1999, **97**:903-915.
- Niederreither K, Dollé P: **Retinoic acid in development: towards an integrated view.** *Nat Rev Genet* 2008, **9**:541-553.
- Lazar MA: **Nuclear receptor corepressors.** *Nucl Recept Signal* 2003, **1**:e001.
- Delacroix L, Moutier E, Altobelli G, Legras S, Poch O, Choukallah M, Bertin I, Jost B, Davidson I: **Cell-specific interaction of retinoic acid receptors with target genes in mouse embryonic fibroblasts and embryonic stem cells.** *Mol Cell Biol* 2010, **30**:231-244.
- Hua S, Kittler R, White KP: **Genomic antagonism between retinoic acid and estrogen signaling in breast cancer.** *Cell* 2009, **137**:1259-1271.
- Ross-Innes CS, Stark R, Holmes KA, Schmidt D, Spyrou C, Russell R, Massie CE, Vowler SL, Eldridge M, Carroll JS: **Cooperative interaction between retinoic acid receptor- α and estrogen receptor in breast cancer.** *Genes Dev* 2010, **24**:171-182.
- Sabo PJ, Kuehn MS, Thurman R, Johnson BE, Johnson EM, Cao H, Yu M, Rosenzweig E, Goldy J, Haydock A, Weaver M, Shafer A, Lee K, Neri F, Humbert R, Singer MA, Richmond TA, Dorschner MO, McArthur M, Hawrylycz M, Green RD, Navas PA, Noble WS, Stamatoyannopoulos JA: **Genome-scale mapping of DNase I sensitivity *in vivo* using tiling DNA microarrays.** *Nat Methods* 2006, **3**:511-518.
- Guccione E, Martinato F, Finocchiaro G, Luzi L, Tizzoni L, Dall'Olio V, Zardo G, Nervi C, Bernard L, Amati B: **Myc-binding-site recognition in the human genome is determined by chromatin context.** *Nat Cell Biol* 2006, **8**:764-770.
- Heintzman ND, Hon GC, Hawkins RD, Kheradpour P, Stark A, Harp LF, Ye Z, Lee LK, Stuart RK, Ching CW, Ching KA, Antosiewicz-Bourget JE, Liu H, Zhang X, Green RD, Lobanenkov VV, Stewart R, Thomson JA, Crawford GE, Kellis M, Ren B: **Histone modifications at human enhancers reflect global cell-type-specific gene expression.** *Nature* 2009, **459**:108-112.
- Whittington T, Perkins AC, Bailey TL: **High-throughput chromatin information enables accurate tissue-specific prediction of transcription factor binding sites.** *Nucleic Acids Res* 2009, **37**:14-25.
- Won K, Ren B, Wang W: **Genome-wide prediction of transcription factor binding sites using an integrated model.** *Genome Biol* 2010, **11**:R7.
- Robertson AG, Bilenky M, Tam A, Zhao Y, Zeng T, Thiessen N, Cezard T, Fejes AP, Wederell ED, Cullum R, Euskirchen G, Krzywinski M, Birol I, Snyder M, Hoodless PA, Hirst M, Marra MA, Jones SJ: **Genome-wide relationship between histone H3 lysine 4 mono- and tri-methylation and transcription factor binding.** *Genome Res* 2008, **18**:1906-1917.
- Visel A, Blow MJ, Li Z, Zhang T, Akiyama JA, Holt A, Plajzer-Frick I, Shoukry M, Wright C, Chen F, Afzal V, Ren B, Rubin EM, Pennacchio LA: **ChIP-seq accurately predicts tissue-specific activity of enhancers.** *Nature* 2009, **457**:854-858.
- Yang A, Zhu Z, Kapranov P, McKeon F, Church GM, Gingeras TR, Struhl K: **Relationships between p63 binding, DNA sequence, transcription activity, and biological function in human cells.** *Mol Cell* 2006, **24**:593-602.
- Szutorisz H, Canzonetta C, Georgiou A, Chow C, Tora L, Dillon N: **Formation of an active tissue-specific chromatin domain initiated by epigenetic marking at the embryonic stem cell stage.** *Mol Cell Biol* 2005, **25**:1804-1820.
- Xu J, Watts JA, Pope SD, Gadue P, Kamps M, Plath K, Zaret KS, Smale ST: **Transcriptional competence and the active marking of tissue-specific enhancers by defined transcription factors in embryonic and induced pluripotent stem cells.** *Genes Dev* 2009, **23**:2824-2838.
- Wichterle H, Lieberam I, Porter JA, Jessell TM: **Directed differentiation of embryonic stem cells into motor neurons.** *Cell* 2002, **110**:385-397.
- Balmer JE, Blomhoff R: **A robust characterization of retinoic acid response elements based on a comparison of sites in three species.** *J Steroid Biochem Mol Biol* 2005, **96**:347-354.
- Chen X, Xu H, Yuan P, Fang F, Huss M, Vega VB, Wong E, Orlov YL, Zhang W, Jiang J: **Integration of external signaling pathways with the core transcriptional network in embryonic stem cells.** *Cell* 2008, **133**:1106-1117.
- Nielsen R, Pedersen TA, Hagenbeek D, Moulos P, Siersbæk R, Megens E, Denisov S, Børgesen M, Francoijs K, Mandrup S, Stunnenberg HG: **Genome-wide profiling of PPAR γ :RXR and RNA polymerase II occupancy reveals temporal activation of distinct metabolic pathways and changes in RXR dimer composition during adipogenesis.** *Genes Dev* 2008, **22**:2953-2967.
- Heng JD, Feng B, Han J, Jiang J, Kraus P, Ng J, Orlov YL, Huss M, Yang L, Lufkin T, Lim B, Ng H: **The nuclear receptor Nr5a2 can replace Oct4 in the reprogramming of murine somatic cells to pluripotent cells.** *Cell Stem Cell* 2010, **6**:167-174.
- Näär AM, Boutin J, Lipkin SM, Yu VC, Holloway JM, Glass CK, Rosenfeld MG: **The orientation and spacing of core DNA-binding motifs dictate selective transcriptional responses to three nuclear receptors.** *Cell* 1991, **65**:1267-1279.
- Han K, Song H, Moon I, Augustin R, Moley K, Rogers M, Lim H: **Utilization of DR1 as true RARE in regulating the Ssm, a novel retinoic acid-target gene in the mouse testis.** *J Endocrinol* 2007, **192**:539-551.
- Rastinejad F, Wagner T, Zhao Q, Khorasanizadeh S: **Structure of the RXR-RAR DNA-binding complex on the retinoic acid response element DR1.** *EMBO J* 2000, **19**:1045-1054.
- Xie X, Mikkelsen TS, Gnirke A, Lindblad-Toh K, Kellis M, Lander ES: **Systematic discovery of regulatory motifs in conserved regions of the human genome, including thousands of CTCF insulator sites.** *Proc Natl Acad Sci USA* 2007, **104**:7145-7150.
- Guo G, Bauer S, Hecht J, Schulz MH, Busche A, Robinson PN: **A short ultraconserved sequence drives transcription from an alternate FBN1 promoter.** *Int J Biochem Cell Biol* 2008, **40**:638-650.
- Heim K, White K, Deng D, Tomlinson C, Moore J, Freemantle S, Spinella M: **Selective repression of retinoic acid target genes by RIP140 during**

- induced tumor cell differentiation of pluripotent human embryonal carcinoma cells. *Mol Cancer* 2007, **6**:57.
33. Hirose Y, Ohkuma Y: **Phosphorylation of the C-terminal domain of RNA polymerase II plays central roles in the integrated events of eucaryotic gene expression.** *J Biochem* 2007, **141**:601-608.
34. Palancade B, Bensaude O: **Investigating RNA polymerase II carboxyl-terminal domain (CTD) phosphorylation.** *Eur J Biochem* 2003, **270**:3859-3870.
35. Brookes E, Pombo A: **Modifications of RNA polymerase II are pivotal in regulating gene expression states.** *EMBO Rep* 2009, **10**:1213-1219.
36. Rahl PB, Lin CY, Seila AC, Flynn RA, McCuine S, Burge CB, Sharp PA, Young RA: **c-Myc regulates transcriptional pause release.** *Cell* 2010, **141**:432-445.
37. Marson A, Levine SS, Cole MF, Frampton GM, Brambrink T, Johnstone S, Guenther MG, Johnston WK, Wernig M, Newman J: **Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells.** *Cell* 2008, **134**:521-533.
38. Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, Alvarez P, Brockman W, Kim T, Koche RP, Lee W, Mendenhall E, O'Donovan A, Presser A, Russ C, Xie X, Meissner A, Wernig M, Jaenisch R, Nusbaum C, Lander ES, Bernstein BE: **Genome-wide maps of chromatin state in pluripotent and lineage-committed cells.** *Nature* 2007, **448**:553-560.
39. Ku M, Koche RP, Rheinbay E, Mendenhall EM, Endoh M, Mikkelsen TS, Presser A, Nusbaum C, Xie X, Chi AS, Adli M, Kasif S, Ptaszek LM, Cowan CA, Lander ES, Koseki H, Bernstein BE: **Genomewide analysis of PRC1 and PRC2 occupancy identifies two classes of bivalent domains.** *PLoS Genet* 2008, **4**:e1000242.
40. Meissner A, Mikkelsen TS, Gu H, Wernig M, Hanna J, Sivachenko A, Zhang X, Bernstein BE, Nusbaum C, Jaffe DB, Gnirke A, Jaenisch R, Lander ES: **Genome-scale DNA methylation maps of pluripotent and differentiated cells.** *Nature* 2008, **454**:766-770.
41. Ho L, Jothi R, Ronan JL, Cui K, Zhao K, Crabtree GR: **An embryonic stem cell chromatin remodeling complex, esBAF, is an essential component of the core pluripotency transcriptional network.** *Proc Natl Acad Sci USA* 2009, **106**:5187-5191.
42. Cheng Y, Wu W, Ashok Kumar S, Yu D, Deng W, Tripic T, King DC, Chen K, Zhang Y, Drautz D, Giardine B, Schuster SC, Miller W, Chiaromonte F, Zhang Y, Blobel GA, Weiss MJ, Hardison RC: **Erythroid GATA1 function revealed by genome-wide analysis of transcription factor occupancy, histone modifications, and mRNA expression.** *Genome Res* 2009, **19**:2172-2184.
43. Wilson NK, Miranda-Saavedra D, Kinston S, Bonadies N, Foster SD, Calero-Nieto F, Dawson MA, Donaldson IJ, Dumon S, Frampton J, Janky R, Sun X, Teichmann SA, Bannister AJ, Göttgens B: **The transcriptional program controlled by the stem cell leukemia gene *Scf/Tal1* during early embryonic hematopoietic development.** *Blood* 2009, **113**:5456-5465.
44. Wasserman WW, Sandelin A: **Applied bioinformatics for the identification of regulatory elements.** *Nat Rev Genet* 2004, **5**:276-287.
45. Won K, Chepelev I, Ren B, Wang W: **Prediction of regulatory elements in mammalian genomes using chromatin signatures.** *BMC Bioinformatics* 2008, **9**:547.
46. Follows GA, Dhami P, Göttgens B, Bruce AW, Campbell PJ, Dillon SC, Smith AM, Koch C, Donaldson IJ, Scott MA, Dunham I, Janes ME, Vetriche D, Green AR: **Identifying gene regulatory elements by genomic microarray mapping of DNase hypersensitive sites.** *Genome Res* 2006, **16**:1310-1319.
47. Wettenhall JM, Simpson KM, Satterley K, Smyth GK: **affyLmGUI: a graphical user interface for linear modeling of single channel microarray data.** *Bioinformatics* 2006, **22**:897-899.
48. Wu Z, Irizarry RA, Gentleman R, Martinez-Murillo F, Spencer F: **A model-based background adjustment for oligonucleotide expression arrays.** *J Am Stat Assoc* 2004, **99**:909-917.
49. Smyth GK: **Linear models and empirical Bayes methods for assessing differential expression in microarray experiments.** *Stat Appl Genet Mol Biol* 2004, **3**(Article 3).
50. Benjamini Y, Hochberg Y: **Controlling the false discovery rate: a practical and powerful approach to multiple testing.** *J R Stat Soc Series B (Methodological)* 1995, **57**:289-300.
51. **Young Lab ChIP Protocol.** [http://jura.wi.mit.edu/young_public/hESregulation/ChIP.html].
52. Guenther MG, Lawton LN, Rozovskaia T, Frampton GM, Levine SS, Volkert TL, Croce CM, Nakamura T, Canaani E, Young RA: **Aberrant chromatin at genes encoding stem cell regulators in human mixed-lineage leukemia.** *Genes Dev* 2008, **22**:3403-3408.
53. **Illumina Genome Analyzer.** [http://www.illumina.com/systems/genome_analyzer_ixi.ilmn].
54. Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetverin V, Church DM, Dicuccio M, Edgar R, Federhen S, Feolo M, Geer LY, Helmsberg W, Kapustin Y, Khovayko O, Landsman D, Lipman DJ, Madden TL, Maglott DR, Miller V, Ostell J, Pruitt KD, Schuler GD, Shumway M, Sequeira E, Sherry ST, Sirotkin K, Souvorov A, Starchenko G, Tatusov RL, et al: **Database resources of the National Center for Biotechnology Information.** *Nucleic Acids Res* 2008, **36**:D13-21.
55. Langmead B, Trapnell C, Pop M, Salzberg SL: **Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.** *Genome Biol* 2009, **10**:R25.
56. Zhang Y, Liu T, Meyer C, Eeckhoutte J, Johnson D, Bernstein B, Nusbaum C, Myers R, Brown M, Li W, Liu XS: **Model-based Analysis of ChIP-Seq (MACS).** *Genome Biol* 2008, **9**:R137.
57. Rozowsky J, Euskirchen G, Auerbach RK, Zhang ZD, Gibson T, Bjornson R, Carriero N, Snyder M, Gerstein MB: **PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls.** *Nat Biotechnol* 2009, **27**:66-75.
58. Mahony S, Hendrix D, Golden A, Smith TJ, Rokhsar DS: **Transcription factor binding site identification using the self-organizing map.** *Bioinformatics* 2005, **21**:1807-1814.
59. Mahony S, Golden A, Smith TJ, Benos PV: **Improved detection of DNA motifs using a self-organized clustering of familial binding profiles.** *Bioinformatics* 2005, **21**(Suppl 1):i283-91.
60. Hughes JD, Estep PW, Tavazoie S, Church GM: **Computational identification of cis-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*.** *J Mol Biol* 2000, **296**:1205-1214.
61. Liu X, Brutlag DL, Liu JS: **BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes.** *Pac Symp Biocomput* 2001, 127-138.
62. Pavesi G, Mereghetti P, Mauri G, Pesole G: **Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes.** *Nucleic Acids Res* 2004, **32**:W199-203.
63. Mahony S, Benos PV: **STAMP: a web tool for exploring DNA-binding motif similarities.** *Nucleic Acids Res* 2007, **35**:W253-258.
64. Sandelin A, Wasserman WW: **Prediction of nuclear hormone receptor response elements.** *Mol Endocrinol* 2005, **19**:595-606.
65. Bar-Joseph Z, Demaine ED, Gifford DK, Srebro N, Hamel AM, Jaakkola TS: **K-ary clustering with optimal leaf ordering for gene expression data.** *Bioinformatics* 2003, **19**:1070-1078.
66. Siepel A, Haussler D: **Phylogenetic Hidden Markov Models.** *Statistical Methods in Molecular Evolution* New York: Springer; 2005, 325-351.
67. **LIBSVM: a library for support vector machines.** [http://www.csie.ntu.edu.tw/~cjlin/libsvm/].

doi:10.1186/gb-2011-12-1-r2

Cite this article as: Mahony et al.: Ligand-dependent dynamics of retinoic acid receptor binding during early neurogenesis. *Genome Biology* 2011 **12**:R2.

Submit your next manuscript to BioMed Central and take full advantage of:

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at
www.biomedcentral.com/submit

