RAR/RXR binding dynamics distinguishes pluripotency from differentiation associated cis-regulatory elements

Supplementary methods

Contents

1 Primary analysis of sequencing data

1.1 Description and nomenclature of the samples

All samples are from murine F9 cell line cultures, at various time points after induction of differentiation by RA. Sample types and time are summarized in the following table:

Sample ID	Sample type	Time	Nb replicates
wce-wt	WCE	0h	1
chIP-rar-wt	ChIP RAR	0h	1
chIP-rxr-wt	ChIP RXR	0h	1
mRNA-wt	mRNA	0h	4
chIP-rar-2h	ChIP RAR	2h	1
chIP-rxr-2h	ChIP RXR	2h	1
mRNA-ra-6h	mRNA	6h	2
mRNA-12h	mRNA	12h	2
chIP-rar-24h	ChIP RAR	24h	1
chIP-rxr-24h	ChIP RXR	24h	1
mRNA-ra-24h	mRNA	24h	2
mRNA-dmso-24h	mRNA	24h	1
chIP-rar-48h	ChIP RAR	48h	1
chIP-rxr-48h	ChIP RXR	48h	1
mRNA-ra-48h	mRNA	48h	2
mRNA-dmso-48h	mRNA	48h	1

Starting from untreated cells (`*-wt` samples), the time course was performed in parallel on RA treated cells (`chIP-*-*h` and `mRNA-ra-*h` samples) and on untreated cells (`mRNA-dmso-*` samples). The time course on untreated cells serves as a control of the stability of the initial cell population, and were used as replicates of the `mRNA-wt` sample.

Time points for mRNA-seq and ChIP-seq samples are not meant to match since cis-regulation is known to have a delayed effect on mRNA levels, and the delay may vary from one transcript to another [13].

1.2 ChIP-seq samples

1.2.1 Alignment of sequenced reads

Reads from ChIP-seq samples were aligned using Bowtie [11] version 0.12.9 in MAQ-like policy. We used options `-n 3`, `-e 70` and `-m 1` to discard reads that mapped at more than one location in the genome. We used the UCSC Genome Browser[1] to produce visualizations of the aligned reads along the chromosomes and gene annotations. To this end, the SAM files produced by Bowtie were converted to BAM format with `samtools`[2], which can be directly displayed in the Genome Browser.

1.2.2 Determination of candidate binding regions

Each ChIP-seq sample yields a collection of peaks and it is not straightforward to follow the time course of a single peak: one would have to match peaks between samples (their position may vary, some may appear/disappear, etc ...). Our approach is to use peaks from

[1] `http://genome.ucsc.edu/cgi-bin/hgGateway`

[2] `http://samtools.sourceforge.net/`

Sample ID	#Total reads	#Total alignments	#Aligned reads (%)
wce-wt	21768081	11995208	11995208 (0.55)
chIP-rar-wt	33110281	10330904	10330904 (0.31)
chIP-rar-2h	22259819	10579331	10579331 (0.48)
chIP-rar-24h	33314035	6814684	6814684 (0.20)
chIP-rar-48h	31760038	10391493	10391493 (0.33)
chIP-rxr-wt	32576694	8928011	8928011 (0.27)
chIP-rxr-2h	31513920	15063235	15063235 (0.48)
chIP-rxr-24h	33770496	10748544	10748544 (0.32)
chIP-rxr-48h	33528842	11268656	11268656 (0.34)

Table 1: Alignment results for ChIP-seq samples.

all samples together to determine a fixed set of genomic regions where we can detect binding events. These binding regions are considered as a genomic feature (like genes) and do not depend on one particular condition anymore, so one can study their coverage by sequencing data across several conditions.

More formally, what we will thereafter call *candidate binding regions* are the genomic locations defined as follows:

1. call peaks for each sample, with a really permissive threshold
2. consider the set of positions on the genome which are the summit of a peak in at least one condition
3. define a link between any two summits that are closer than 100bp
4. compute connected components of this graph
5. define, for each connected component, a *binding region*, as a location of 500bp wide centered at the mean of the summits in the connected component

For the first step, we used MACS [18] version 1.4.2 with default options, except the p-value threshold (`--pvalue`) which was set to 10^{-3}. For all chIP samples, we took `wce-wt` as a control. MACS is pretty good at indicating the summit of a peak, it is shown in [18] that it is strongly correlated with the occurrence of the expected sequence motif.

With the above procedure, we obtained a total of 192002 candidate binding regions. Given that the peak calling step was performed with a very permissive threshold, it was necessary to select a subset of *bona fide* binding regions. For this purpose we established two criteria that were used as filters on this initial set. They are described in the following paragraph.

1.3 Filtering candidate binding regions

Significant read enrichment with respect to control We expect that real binding regions will display significantly more reads in at least one ChIP sample than in the control

(WCE). In order to assess this, MACS uses a simple Poisson test taking as a null hypothesis the intensity in control and flanking regions of the peak. However this approach is problematic when the size of the library differs between treatment and control (or between treatments); various *ad hoc* normalization strategies have been proposed (notably in MACS) but we preferred a more principled approach and resorted to the Poisson Margin Test [10], that was specifically designed to address this issue. Alternatively, we could also have used tests based on the negative binomial distribution, like in DESeq [3]. However the estimation of the over-dispersion parameter is arguably not reliable enough in the absence of replicates (though it is technically feasible with DESeq), and we decided to use a simpler model (Poisson-based count distribution), granted that the Poisson Margin Test can be considered pretty conservative.

Each candidate regions was assigned the number of reads it contained, counting only one read if several fell at the same location. For each we tested for a significant difference between RAR binding and control signals at at least one time point. We decided to keep a region if for at least one time point, the test led to a (Benjamini-Hochberg adjusted) p-value smaller than 0.001. We ended with 32850 significantly enriched regions.

In the two remaining filters, each region is assigned its lowest p-value throughout the time serie (that is a score representing the biggest difference between chIP and control during the time course).

Reproducibility between RAR and RXR ChIP samples RAR is known to interact with DNA only in a complex with RXR, while RXR has a lot more possible partners in the NHR family. As a result, it is expected that each RAR binding region is also an RXR binding region (but the converse is not necessarily true). In order to verify that, we represented on Suppl. Fig. 1 the proportion among RAR-positive candidate regions of RXR-positive regions as a function of the unadjusted p-value threshold. We decided to finally keep regions with at most 10^{-7} unadjusted p-value, provided:

- they are not located on a mitochondrial chromosome
- they do not belong to the blacklist[3] published in the context of the ENCODE project [6].

These last operations provided the 13791 regions used in our paper.

1.4 mRNA-seq samples

1.4.1 Alignment of sequenced reads

The reads from mRNA-seq samples were aligned using `Tophat` [15] version 2.0.7 associated with `Bowtie` [11] version 0.12.9. `Tophat` was run with default options.The aligned reads were assigned to genes by `htseq-count` [4], run with default options on the version 63 of the Ensembl mouse annotation [4].

[3]This list is available for download at `https://sites.google.com/site/anshulkundaje/projects/blacklists`

[4]Available at `ftp://ftp.ensembl.org/pub/release-63/gtf/mus_musculus/`

1.4.2 Detection of differentially expressed genes

We sought to detect genes that are differentially expressed in at least one condition (time point) with respect to the untreated cells. As explained in the paragraph 1.1, the samples `mRNA-wt`, `mRNA-dmso-24h` and `mRNA-dmso-48h` were considered as replicates for the untreated condition.

The tests for differential expression were performed using `DESeq` [3] version 1.12.0. A gene was declared modulated if it displayed a significant difference between any two time-points (the cutoff was fixed at 5.10^{-2} of (Bonferroni-Hochberg) adjusted p-value) and expressed if it had a non-zero estimated level (baseMean in DESeq) in some condition.

2 NH response elements in RAR regions

2.1 Evaluating NHRE motif enrichment in RAR regions

We expected a large proportion of our RAR regions to contain the known binding motif of the RAR-RXR complex, at least significantly more than in random regions. In order to quantify this expected enrichment, we scanned each binding region with the motif defined by Balmer *et al.* [5] and reported its best alignment score. We did the same in a control set consisting of random DNA sequences with the same length and GC distribution. Now for a given score threshold, we could calculate the proportion of regions with at least one match (called positive regions) in both sets (RAR regions and control). By varying the score threshold from $-\infty$ to $+\infty$, we obtained a ROC curve, which summarizes the trade-off between the sensitivity of the detection and the false positive rate. This was repeated for all direct, everted and inverted repeats; we also considered a DR{0,1,2,5} (resp. DR{1,2,5}) compound motif that has an occurrence as soon as one among DR0, DR1, DR2, DR5 (resp. DR1, DR2, DR5) has one. Fig. 1 shows the obtained results for the DR0, DR1, DR2, DR5 and DR{0,1,2,5} motifs.

The area under the ROC curve (AUC) can be used to quantify the enrichment of the motif in our dataset, as it is equal to the probability that a RAR region has a better score than a random region. In other words, it measures the ability of the alignment score to distinguish between experimentally bound regions and random sequences. We reported the AUC for all direct, everted and inverted repeats in Fig. 3C of the paper.

2.2 Determination of an alignment score threshold for site prediction

We sought to establish a threshold for alignment scores, in order to predict true NHRE with an acceptable trade-off between sensitivity and false positive rate. This threshold can be calculated theoretically for random sequences generated by a Markovian process [14]; however we preferred to estimate these threshold empirically, in order to use a control set consisting of sequences from an actual genome.

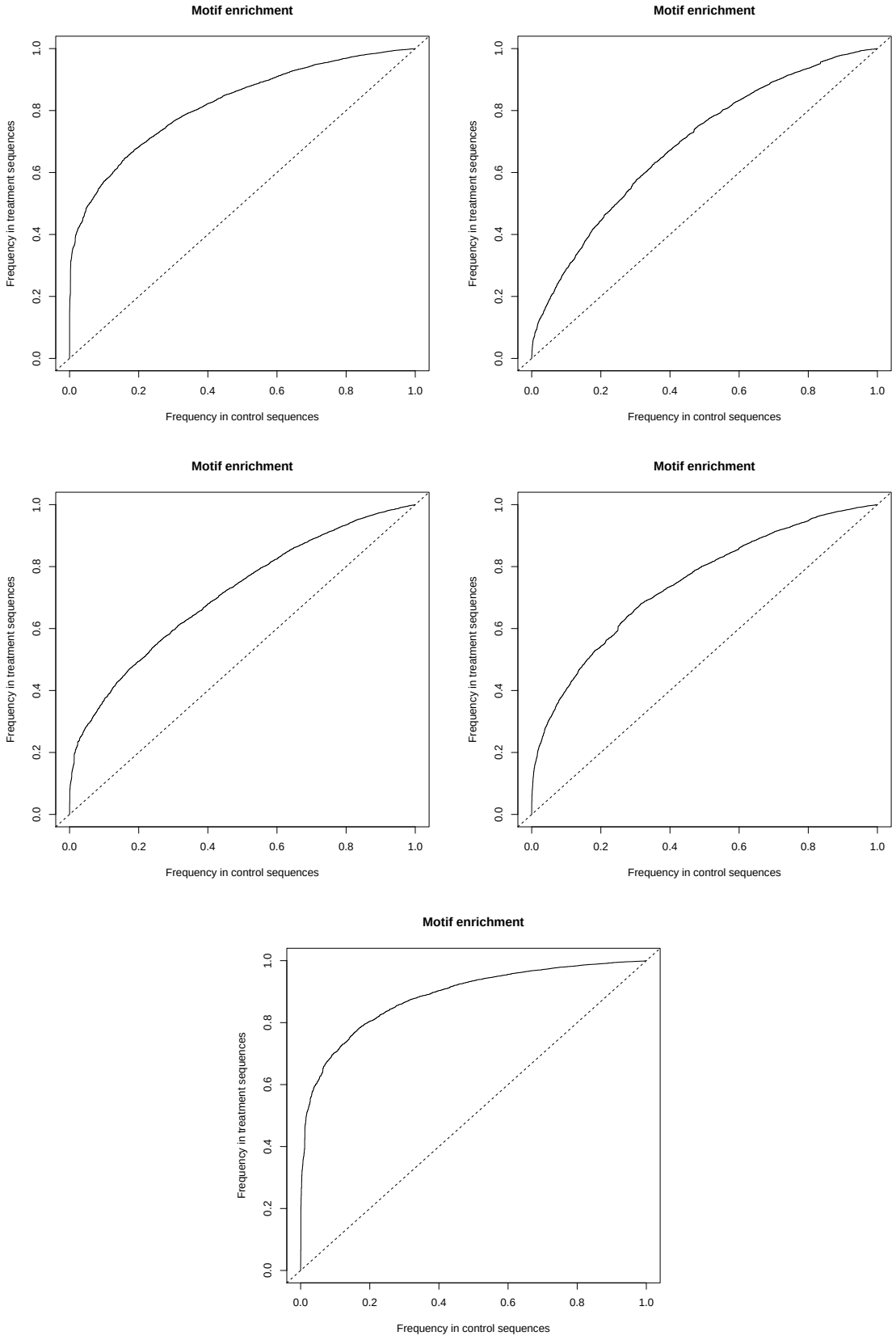

Figure 1: ROC curves for DR0, DR1, DR2, DR5 and DR{0,1,2,5} motifs.

We built a control set of random sequences having the same length and G+C composition distribution than in the panRAR binding regions. Each nucleotide was generated independently. This control set was scanned for all NHR motifs, and reported the score threshold yielding respectively 1, 5, 10, 20 and 30% positive (that is, bearing a motif occurrence) sequences in the control set. Said differently, we thus estimated the thresholds for false-positive rates of 0.01, 0.05, 0.1, 0.2 and 0.3. We also reported the fraction of positive sequences in the panRAR binding regions, that is the corresponding sensitivity (assuming, as an approximation, that all binding regions have an actual RAR motif).

We finally performed all motif predictions by taking the score threshold corresponding to 10% of matching sequences in the control set.

3 Temporal profiles and association with TF binding

The primary analysis of our ChIP-seq and RNA-seq datasets (see Section 1) yielded time series measurements for both RAR/RXR binding at every RAR binding site and transcription level for every gene. We then performed a clustering analysis of these data, as means to detect significant associations between an element (gene or binding site) response and its chromatinic environment.

3.1 ChIP-seq time series

ChIP-seq time series clustering RAR/RXR binding regions were filtered by keeping the only regions that experience a significant change during the time course. More precisely, we tested for each binding region if there were two time points between which the ChIP-seq signal exhibited a significant difference, as detected by the Poisson Margin Test [10] at a 10^{-6} significance level. These so-called "dynamic binding regions" were then clustered using a standard k-means algorithm; each region was represented as a vector with one coordinate per time-point; each coordinate was computed as the number of reads falling into the region in the associated time-point divided by the number of millions of aligned reads in the corresponding sample (recall that all binding regions were set to the same length). We used the program `Cluster 3`[5] with the following options:

- normalization on (each profile is scaled to (Euclidean) norm 1)
- k-means algorithm
- 100 runs
- use uncentered Pearson correlation as a distance.

The number of clusters was chosen empirically, by visual inspection.

[5] `http://bonsai.hgc.jp/~mdehoon/software/cluster/software.htm#ctv`

Transcription factor occupancy datasets We used the occupancy maps provided in publications on mouse ES [7, 12, 9, 8, 16, 17] and F9 cells [2]. All the datasets were collected through the GEO database [1]. We chose to reuse the called peaks distributed with each dataset.

Association with RAR/RXR binding clusters Each of our RAR binding regions was annotated as having or not a detected peak for other TFs less than 1kb away. Formally, the association score between some cluster membership and some other TF binding was quantified with a log odds ratio as computed for the Fisher Exact test.

3.2 RNA-seq time series

We applied an analoguous methodology to detect association between expression profiles and TF binding. First we distinguished three groups of genes in the genome:

- **modulated**, that is genes that exhibit a significant difference between two time points, as tested with `DEseq` at an (adjusted) p-value level of 0.05,
- **expressed**, that is genes that have a non-zero expression level at some time point,
- non-expressed, that is genes which have not been detected at any time point.

Only modulated genes were included for the cluster analysis, which was performed by `Cluster 3.0`. Each gene was represented as a vector containing the log fold change with respect to the initial (t=0h) time point for the three other time points. The number of classes in the clustering was also chosen by visual inspection of the clusters.

Abbreviations

- AUC : Area Under Curve
- NH : Nuclear Hormone
- NHRE : Nuclear Hormone Receptor Element
- RA : Retinoic Acid
- ROC : Receiver OPerator Curve

References

[1] A. Acland et al. "Database resources of the National Center for Biotechnology Information". In: *Nucleic Acids Res.* (2013).

[2] I. Aksoy et al. "Oct4 switches partnering from Sox2 to Sox17 to reinterpret the enhancer code and specify endoderm". In: *EMBO J.* 32.7 (2013), pp. 938–953.

[3] S. Anders and W. Huber. "Differential expression analysis for sequence count data". In: *Genome Biol.* 11.10 (2010), R106.

[4] Simon Anders, Paul Theodor Pyl, and Wolfgang Huber. "HTSeq – A Python framework to work with high-throughput sequencing data". In: *bioRxiv* (2014). DOI: 10.1101/002824. eprint: http://biorxiv.org/content/early/2014/02/20/002824.full.pdf. URL: http://biorxiv.org/content/early/2014/02/20/002824.

[5] J. E. Balmer and R. Blomhoff. "A robust characterization of retinoic acid response elements based on a comparison of sites in three species". In: *J. Steroid Biochem. Mol. Biol.* 96.5 (2005), pp. 347–354.

[6] B. E. Bernstein et al. "An integrated encyclopedia of DNA elements in the human genome". In: *Nature* 489.7414 (2012), pp. 57–74.

[7] X. Chen et al. "Integration of external signaling pathways with the core transcriptional network in embryonic stem cells". In: *Cell* 133.6 (2008), pp. 1106–1117.

[8] J. Han et al. "Tbx3 improves the germ-line competency of induced pluripotent stem cells". In: *Nature* 463.7284 (2010), pp. 1096–1100.

[9] J. C. Heng et al. "The nuclear receptor Nr5a2 can replace Oct4 in the reprogramming of murine somatic cells to pluripotent cells". In: *Cell Stem Cell* 6.2 (2010), pp. 167–174.

[10] A. Kowalczyk et al. "The poisson margin test for normalization-free significance analysis of NGS data". In: *J. Comput. Biol.* 18.3 (2011), pp. 391–400.

[11] Ben Langmead et al. "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome". In: *Genome Biology* 10.3 (2009), R25. ISSN: 1465-6906. DOI: 10.1186/gb-2009-10-3-r25. URL: http://genomebiology.com/2009/10/3/R25.

[12] A. Marson et al. "Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells". In: *Cell* 134.3 (2008), pp. 521–533.

[13] Guillaume Rey et al. "Genome-Wide and Phase-Specific DNA-Binding Rhythms of BMAL1 Control Circadian Output Functions in Mouse Liver". In: *PLoS Biol* 9.2 (Feb. 2011), e1000595. DOI: 10.1371/journal.pbio.1000595. URL: http://dx.doi.org/10.1371%2Fjournal.pbio.1000595.

[14] E. Roquain and S. Schbath. "Improved compound Poisson approximation for the number of occurrences of any rare word family in a stationary Markov chain". In: *Advances in applied probability* 39.1 (2007), pp. 128–140.

[15] C. Trapnell, L. Pachter, and S. L. Salzberg. "TopHat: discovering splice junctions with RNA-Seq". In: *Bioinformatics* 25.9 (2009), pp. 1105–1111.

[16] C. Xu et al. "Genome-wide roles of Foxa2 in directing liver specification". In: *J Mol Cell Biol* 4.6 (2012), pp. 420–422.

[17] M. Yamaji et al. “PRDM14 ensures naive pluripotency through dual regulation of signaling and epigenetic pathways in mouse embryonic stem cells”. In: *Cell Stem Cell* 12.3 (2013), pp. 368–382.

[18] Y. Zhang et al. “Model-based analysis of ChIP-Seq (MACS)”. In: *Genome Biol.* 9.9 (2008), R137.