Genomics Data 2 (2014) 89–91

Contents lists available at ScienceDirect

Genomics Data

journal homepage: http://www.journals.elsevier.com/genomics-data/

Data in Brief

Genome-wide profiling of YY1 binding sites during skeletal myogenesis

Kun Sun [a], Leina Lu [b], Huating Wang [b], Hao Sun [a,*]

[a] Department of Chemical Pathology, Li Ka Shing Institute of Health Sciences, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong SAR, China
[b] Department of Obstetrics and Gynaecology, Li Ka Shing Institute of Health Sciences, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong SAR, China

ARTICLE INFO

Article history:
Received 28 April 2014
Received in revised form 14 May 2014
Accepted 15 May 2014
Available online 23 May 2014

Keywords:
YY1
ChIP-seq
Myogenesis
LincRNA

ABSTRACT

Skeletal muscle differentiation is regulated by a network of transcription factors, epigenetic regulators and non-coding RNAs. We have recently performed ChIP-seq experiments to explore the genome-wide binding of transcription factor YY1 in skeletal muscle cells. Our results identified thousands of YY1 binding peaks, underscoring its multifaceted functions in muscle cells. In particular, we identified a very high proportion of YY1 binding peaks residing in the intergenic regions, which led to the discovery of some novel lincRNAs under YY1 regulation. Here we describe the details of the ChIP-seq experiments and data analysis procedures associated with the study published by Lu et al. in the EMBO Journal in 2013 [1].

© 2014 The Authors Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/3.0/).

Specifications	
Organism/cell line/tissue	*Mus musculus/C2C12*
Sex	*NA*
Sequencer or array type	*Illumina Hiseq 2000, Illumina GA IIx*
Data format	*Raw data: FASTQ files* *Processed data: BEDGRAPH, TXT*
Experimental factors	*Myoblast vs myotube*
Experimental features	*Using ChIP-seq, we generated genome-wide maps of YY1 in skeletal myoblasts and myotubes with biological replicates. We found that a large proportion of the binding sites reside in the intergenic regions; therefore, many lincRNAs are regulated by YY1.*
Consent	*NA*
Sample source location	*Manassas, VA, USA*

Direct link to deposited data

http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE45875

Experimental Design, Materials and Methods

Cell culture

Mouse C2C12 myoblast cell line was purchased from American Type Culture Collection (ATCC). The myoblasts were maintained in a growth medium (DMEM, 10%FBS and 1% penicillin/streptomycin), and induced to myotubes by culturing in a differentiation medium (DMEM, 2% horse serum and 1% penicillin/streptomycin).

ChIP assays and sequencing experiments

ChIP assays were performed as previously described [2,3]. About 2×10^7 C2C12 cells and 5 μg of antibodies were used in one immunoprecipitation. The antibodies include YY1 #1 (Santa Cruz Biotechnology, Cat# SC-1703, rabbit polyclonal), YY1 #2 (Abcam, Cat# AB58066, mouse monoclonal), Ezh2 (Cell Signaling, MA, USA, Cat# AC22), trimethyl-histone H3-K27 (Millipore, Cat# 07-449), trimethyl-histone H3-K4 (Millipore, Cat# 07-473), or normal mouse IgG (Santa Cruz Biotechnology, Cat# SC-2025) as a negative control.

For library construction, we used a protocol as described before [4]. Briefly, the immunoprecipitated DNA (~10 ng) were end-repaired, and A-nucleotide overhangs were then added, followed by adapter ligation, PCR enrichment, size selection and purification. The purified DNA library products were evaluated using Bioanalyzer (Agilent) and SYBR qPCR and diluted to 10 nM for sequencing on Illumina Hi-seq 2000 sequencer (YY1) (pair-end with 50 bp) or Illumina Genome Analyzer II sequencer (Ezh2, H3K27me3 and H3K4me3) (pair-end with 36 bp). Technical replicates were prepared by sequencing the same library twice. A data analysis pipeline CASAVA 1.8 (Illumina) was employed to perform the initial bioinformatic analysis (base calling). Table 1 lists all the experiments that we had performed. For MB YY1, we performed two biological replicates with the antibody SC-1703 and a third biological replicate with a second antibody AB58066. We also performed two technical replicates for each antibody (run 1 and run 2).

* Corresponding author at: Department of Chemical Pathology, The Chinese University of Hong Kong, Room 503A, Li Ka Shing Institute of Health Sciences, Prince of Wales Hospital, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong SAR, China. Tel.: +852 3763 6048; fax: +852 3763 6033.
E-mail address: haosun@cuhk.edu.hk (H. Sun).

http://dx.doi.org/10.1016/j.gdata.2014.05.008
2213-5960/© 2014 The Authors Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/3.0/).

Reads alignment, peak defining and motif analysis

The sequenced reads were mapped to the mouse reference genome (UCSC mm9, non-repeat-masked) using SOAP2 [5] (version 2.20, with the following parameters: "-v 2 -r 0 -m 0 -p 20") allowing a maximum of two mismatches and only the uniquely aligned reads were kept. The protein–DNA binding peaks were identified using Model-based Analysis for ChIP-seq (MACS [6], version 2.0.9; for YY1 ChIP-seq (MB rep1); the parameters are "-g mm -m 8,30 -p 0.001" and then the peaks were filtered by *q*-values; for others, the parameters were "-g mm -m 8,30 -q 0.01") with the IgG control sample as background. During the peak calling, a *q*-value (adjusted *P*-value calculated using the Benjamini–Hochberg procedure) was set under 10^{-5} for YY1; it corresponds to an empirical FDR (False Discovery Rate) of 3.4%; 10^{-2} was used for Ezh2 and H3K27me3, where the FDRs were estimated to be around 1%. This difference in data processing for ChIP-seq experiments was because the performance of MACS on a large dataset (e. g., YY1 ChIP-seq (MB rep1) sequenced on Hi-seq 2000) is not as good as on a small dataset (e.g., others on GA IIx). The mapping information and number of peaks from each dataset were shown in Table 1.

The raw ChIP-seq sequencing reads for transcription factor MyoD were downloaded from NCBI's Sequence Read Archive (SRA; http://www.ncbi.nlm.nih.gov/sra) with accession number SRX016191 and SRX016040 for MBs and MTs, respectively. The reads were aligned using the above method and the MyoD-binding peaks were identified using a *q*-value cutoff of 10^{-2}. The processed ChIP-seq data (binding sites) for Pol II and H3K4me3 were obtained from NCBI's GEO under accession number GSE25308. When comparing peaks from different experiments, two peaks were considered as "overlapped peaks" if the distance between them was less than 1 kb.

In order to search for highly occurring motifs in the DNA sequences underlying the putative binding peaks, Discriminative Regular Expression Motif Elicitation (DREME [7], version 4.8.0) was applied on the 100 bp (± 50 bp) sequences flanking the peak summit. The analysis was run on both strands to search for motifs that are no more than 8 bp in length with E-values <0.01.

Quality control

In peak defining, we used the IgG as a negative control and also carefully selected the *q*-values for a reasonable FDR. According to the ENCODE ChIP-seq guidelines [8], we calculated the Fragments in Peaks (FRiP) value using in-house programs (See Supplementary Material). Moreover, for the YY1 biological replicates, we performed Irreproducible Discovery Rate (IDR) analysis using the package developed by Li et al. [9].

Table 1
List of ChIP-seq experiments.

IP	Read length	Total reads[d]	Mapped reads[e]	No. of peaks
YY1 (MB, rep1)[a]	50 bp	106.4	80	1820
YY1 (MB, rep2run1)[b]	36 bp	34.0	25.6	996
YY1 (MB, rep2run2)[b]	36 bp	34.5	26.0	1061
YY1 (MB, rep3run1)[c]	36 bp	28.9	22.4	1504
YY1 (MB, rep3run2)[c]	36 bp	30.9	23.9	1655
YY1 (MT)[d]	50 bp	86.8	62.5	626
Ezh2	50 bp	37.0	25.7	1801
H3K4me3	36 bp	10.2	6.9	21,051
H3K27me3	36 bp	26.5	20.8	10,674

[a,c,d] Using SC1703 antibody.

[b] Using AB58066 antibody.

[e] Total reads and mapped reads are reported as millions of reads. The number of uniquely mapped reads, the number of reads aligning and the number of pairs concordantly are all the same as the number of the mapped reads in this experiment based on the alignment protocol used.

Functional annotation

To identify putative YY1 target gene, each identified peak was associated with the closest RefSeq gene when it falls into the 4 kb (±2 kb) flanking region of the gene's TSS, and these genes were considered as potentially regulated by YY1. For analysis of differentially expressed genes, we used Cufflinks [10] (version 1.3) to evaluate the expression profile (using Fragments Per Kilobase of exon model per Million mapped reads, FPKM) of all the RefSeq transcripts using the publically available RNA-Seq data obtained from −24 h (myoblasts, MBs) and 60 h (myotubes, MTs) C2C12 [11]. Differentially expressed genes were defined as those up- or down-regulated in MTs as compared to MB. If a gene is differentially expressed and bound by YY1, we reason that it could be potentially regulated by YY1 since the YY1 level decreases during C2C12 differentiation. Up-regulated YY1 bound genes were defined if their expression in MTs is >1.2 fold higher compared with MBs and these genes are possibly repressed by YY1 in MBs. Down-regulated YY1 bound genes were defined if their expression in MTs is less than 0.8 fold compared with MBs and they are likely activated by YY1 in MBs. Then Gene Ontology (GO) analysis was performed on both up- and down-regulated genes using Database for Annotation, Visualization and Integrated Discovery (DAVID, http://david.abcc.ncifcrf.gov/) [12,13] for functional annotations.

Identification of YY1 bound novel lincRNAs

Since we observed that more than 1/4 of the YY1 peaks were in the intergenic regions, we suspected that YY1 may regulate unannotated lincRNAs. To validate this hypothesis, we used the list of novel lincRNA identified by Guttmanet et al. from four mouse cell types [14]. YY1-binding sites were searched in the flanking regions (±100 kb on both sides) of these lincRNAs. The resultant list of lincRNAs was considered as YY1-associated muscle lincRNAs or Yams.

Conflict of interest statement

The authors declare no conflict of interest.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.gdata.2014.05.008.

References

[1] L. Lu, K. Sun, X. Chen, Y. Zhao, L. Wang, L. Zhou, H. Sun, H. Wang, Genome-wide survey by ChIP-seq reveals YY1 regulation of lincRNAs in skeletal myogenesis. EMBO J. 32 (2013) 2575–2588.

[2] L. Lu, L. Zhou, E.Z. Chen, K. Sun, P. Jiang, L. Wang, X. Su, H. Sun, H. Wang, A novel YY1-miR-1 regulatory circuit in skeletal myogenesis revealed by genome-wide prediction of YY1-miRNA network. PLoS One 7 (2012) e27596.

[3] L. Zhou, L. Wang, L. Lu, P. Jiang, H. Sun, H. Wang, A novel target of microRNA-29, Ring1 and YY1-binding protein (Rybp), negatively regulates skeletal myogenesis. J. Biol. Chem. 287 (2012) 25255–25265.

[4] Y. Diao, X. Guo, Y. Li, K. Sun, L. Lu, L. Jiang, X. Fu, H. Zhu, H. Sun, H. Wang, Z. Wu, Pax3/7BP is a Pax7- and Pax3-binding protein that regulates the proliferation of muscle precursor cells by an epigenetic mechanism. Cell Stem Cell 11 (2012) 231–241.

[5] R. Li, C. Yu, Y. Li, T.W. Lam, S.M. Yiu, K. Kristiansen, J. Wang, SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics 25 (2009) 1966–1967.

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

[7] T.L. Bailey, DREME: motif discovery in transcription factor ChIP-seq data. Bioinformatics 27 (2011) 1653–1659.

[8] S.G. Landt, G.K. Marinov, A. Kundaje, P. Kheradpour, F. Pauli, S. Batzoglou, B.E. Bernstein, P. Bickel, J.B. Brown, P. Cayting, Y. Chen, G. DeSalvo, C. Epstein, K.I. Fisher-Aylor, G. Euskirchen, M. Gerstein, J. Gertz, A.J. Hartemink, M.M. Hoffman, V. R. Iyer, Y.L. Jung, S. Karmakar, M. Kellis, P.V. Kharchenko, Q. Li, T. Liu, X.S. Liu, L. Ma, A. Milosavljevic, R.M. Myers, P.J. Park, M.J. Pazin, M.D. Perry, D. Raha, T.E. Reddy, J. Rozowsky, N. Shoresh, A. Sidow, M. Slattery, J.A. Stamatoyannopoulos, M. Y. Tolstorukov, K.P. White, S. Xi, P.J. Farnham, J.D. Lieb, B.J. Wold, M. Snyder, ChIP-

seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Res. 22 (2012) 1813–1831.

[9] Q. Li, J.B. Brown, H. Huang, P.J. Bickel, Measuring reproducibility of high-throughput experiments. Ann. Appl. Stat. 5 (2011) 1699–2264.

[10] C. Trapnell, B.A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M.J. van Baren, S.L. Salzberg, B.J. Wold, L. Pachter, Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol. 28 (2010) 511–515.

[11] C. Trapnell, A. Roberts, L. Goff, G. Pertea, D. Kim, D.R. Kelley, H. Pimentel, S.L. Salzberg, J.L. Rinn, L. Pachter, Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat. Protoc. 7 (2012) 562–578.

[12] W. Huang da, B.T. Sherman, R.A. Lempicki, Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat. Protoc. 4 (2009) 44–57.

[13] W. Huang da, B.T. Sherman, R.A. Lempicki, Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 37 (2009) 1–13.

[14] M. Guttman, I. Amit, M. Garber, C. French, M.F. Lin, D. Feldser, M. Huarte, O. Zuk, B.W. Carey, J.P. Cassady, M.N. Cabili, R. Jaenisch, T.S. Mikkelsen, T. Jacks, N. Hacohen, B.E. Bernstein, M. Kellis, A. Regev, J.L. Rinn, E.S. Lander, Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature 458 (2009) 223–227.