Genomics Data 7 (2016) 247–249

Contents lists available at ScienceDirect

Genomics Data

journal homepage: www.elsevier.com/locate/gdata

Data in Brief

Genome-wide RNA-seq and ChIP-seq reveal Linc-YY1 function in regulating YY1/PRC2 activity during skeletal myogenesis

Kun Sun [a,b], Liang Zhou [b], Yu Zhao [b], Huating Wang [b,c,*], Hao Sun [a,b,*]

[a] Department of Chemical Pathology, The Chinese University of Hong Kong, Hong Kong, China
[b] Li Ka Shing Institute of Health Sciences, The Chinese University of Hong Kong, Hong Kong, China
[c] Department of Orthopaedics and Traumatology, The Chinese University of Hong Kong, Hong Kong, China

ARTICLE INFO

Article history:
Received 19 January 2016
Received in revised form 30 January 2016
Accepted 31 January 2016
Available online 2 February 2016

Keywords:
Linc-YY1
Myogenesis
RNA-seq
ChIP-seq
C2C12 cells

ABSTRACT

Little is known how lincRNAs are involved in skeletal myogenesis. Here we describe the discovery and functional annotation of Linc-YY1, a novel lincRNA originating from the promoter of the transcription factor (TF) Yin Yang 1 (YY1). Starting from whole transcriptome shotgun sequencing (a.k.a. RNA-seq) data from muscle C2C12 cells, a series of bioinformatics analysis was applied towards the identification of hundreds of high-confidence novel lincRNAs. Genome-wide approaches were then employed to demonstrate that Linc-YY1 functions to promote myogenesis through associating with YY1 and regulating YY1/PRC2 transcriptional activity *in trans*. Here we describe the details of the ChIP-seq, RNA-seq experiments, and data analysis procedures associated with the study published by Zhou and colleagues in the Nature Communications Journal in 2015 Zhou et al. (2015) [1]. The data was deposited on NCBI's Gene Expression Omnibus (GEO, http://www.ncbi.nlm.nih.gov/geo/) with accession number GSE74049.

Specifications	
Organism/cell line/tissue	*Mus musculus*/C2C12
Sex	NA
Sequencer or array type	Illumina HiSeq 1500, Illumina Hiseq 2000, Illumina GA IIx
Data format	Raw data: FASTQ files Processed data: BEDGRAPH, TXT
Experimental factors	Myoblast vs myotube
Experimental features	Using RNA-seq, we identified hundreds of high confidence novel lincRNAs in skeletal muscle. Among these lincRNAs, one near the YY1 transcription factor caused our attention, which we named Linc-YY1. We then used ChIP-seq and RNA-seq to investigate the function of this novel lincRNA during myogenesis.
Consent	NA
Sample source location	Manassas, VA, USA

1. Direct link to deposited data

http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE74049.

* Corresponding authors at: Li Ka Shing Institute of Health Sciences, The Chinese University of Hong Kong, Hong Kong, China.
E-mail addresses: huating.wang@cuhk.edu.hk (H. Wang), haosun@cuhk.edu.hk (H. Sun).

2. Experimental design, materials, and methods

2.1. Cell culture

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

2.2. 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 (Santa Cruz Biotechnology), Ezh2 (Active Motif), Eed (Millipore), trimethyl-histone H3-K27 (Millipore), or normal mouse IgG (Santa Cruz Biotechnology) 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, Eed and H3K27me3) (pair-end with 36 bp).

http://dx.doi.org/10.1016/j.gdata.2016.01.022

Table 1
Sequencing and mapping information for the experiments.

Type	Sample	Treatment	Raw reads	Pre-processed*	Mapped reads	Mappability (%)
ChIP-seq	Ezh2	Vector transfected	33,033,076	NA	22,480,371	68.1
ChIP-seq	EED	Vector transfected	42,772,454	NA	32,563,476	76.1
ChIP-seq	H3K27me3	Vector transfected	42,897,641	NA	33,224,324	77.5
ChIP-seq	IgG	Vector transfected	46,757,140	NA	35,189,696	75.3
ChIP-seq	Yy1	Linc-YY1 transfected	120,778,675	28,875,468	22,368,411	77.5
ChIP-seq	Ezh2	Linc-YY1 transfected	45,143,782	NA	35,378,560	78.4
ChIP-seq	EED	Linc-YY1 transfected	45,229,883	NA	34,536,361	76.4
ChIP-seq	H3K27me3	Linc-YY1 transfected	46,306,903	NA	35,814,707	77.3
ChIP-seq	IgG	Linc-YY1 transfected	38,713,272	NA	27,863,812	72.0
RNA-seq	siNC	Negative control oligo transfected	248,983,238	112,409,997	94,952,575	84.5
RNA-seq	siLinc-YY1	siLinc-YY1 oligo transfected	275,819,571	64,514,698	52,448,804	81.3

* The preprocessing procedure is only performed on samples with high duplication rate designed to remove the duplicated reads while not for others with low duplication rate (shown as 'NA').

Technical replicates were prepared by sequencing the same library twice. A data analysis pipeline CASAVA 1.8 (Illumina) was employed to perform the initial bioinformatics 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).

2.3. Read alignment and peak calling

The sequenced reads were mapped to the mouse reference genome (UCSC mm9, non-repeat-masked) using SOAP2 [5] (v2.20) allowing a maximum of two mismatches and only the uniquely aligned reads were kept. The sequencing and mapping information from each dataset were shown in Table 1. The protein-DNA binding peaks were identified using Model-based Analysis for ChIP-seq (MACS [6], v1.4.2) with IgG control sample as background. The p-value cutoff was set as 10^{-5} to call high-confidence binding sites.

2.4. Whole transcriptome sequencing experiments

Preparation of RNA-seq libraries for sequencing on the Illumina platforms was carried out using the RNA-Seq Sample Preparation Kit (catalog number RS-930-1001) according to the manufacturer's standard protocol. Briefly, purified RNA was fragmented *via* incubation for 5 min at 94'C with the Illumina supplied fragmentation buffer. The first strand of cDNA was next synthesized by reverse transcription using random oligo primers. Second-strand synthesis was conducted by incubation with RNase H and DNA polymerase I. The resulting double-stranded DNA fragments were subsequently end-repaired and A-nucleotide overhangs were added by incubation with Taq Klenow lacking exonuclease activity. After the attachment of anchor sequences, fragments were PCR amplified using Illumina-supplied primers and loaded onto the Hiseq 2000 or GAIIx flow cell. DNA clusters were generated with an Illumina cluster station with Paired-End Cluster Generation Kit v2 (Illumina), followed by 50 (or 26 bp)*2 cycles of sequencing on sequencer with Sequencing Kit v3 (Illumina). Genome Analyzer Sequencing Control Software (SCS) v2.5, which could perform real-time image analysis and base calling, was used to carry out the image processing and base calling during the chemistry and imaging cycles of a sequencing run. The default parameters within the data analysis software (SCS v2.5) from Illumina were used to filter poor-quality reads. In the default setting, a read would be removed if a chastity of <0.6 is observed on two or more bases among the first 25 bases.

2.5. Transcriptome assembly and novel lincRNA identification

To generate a comprehensive catalog of lincRNAs in muscle cells, we applied an integrated analysis on RNA-seq data generated by Trapnell et al. [7] and our own RNA-seq data from proliferating and differentiating C2C12 cells. The raw sequencing reads were aligned to the mouse reference genome (UCSC mm9) using Tophat (v2.0.4) [7], during which procedure the UCSC gene annotation file downloaded from Cufflinks website (http://cole-trapnelllab.github.io/cufflinks/igenome_table/index.html) was used (the '-G' option). The transcriptome assembly was then performed using Cufflinks (v2.0.4) [7], and a total of 46,627 transcripts were obtained. Then sebnif (v1.2.2) [8] was employed to identify the high-confidence novel lincRNAs. In this procedure, the annotated genes, transcript size, expression level and coding potential were all considered. As a result, a total of 2413 novel lincRNAs were identified. After further annotating each of them with features including K4-K36 domain, EST tag and MyoD binding, a stringent set of 158 lincRNAs were obtained.

2.6. Differentially expressed genes analysis

To detect the differentially expressed genes between siNC- and siLinc-YY1-transfected C2C12 cells, the raw RNA-seq data were first preprocessed (adapter trimming and duplicate removing using in-house programs) and then aligned to the reference genome (UCSC mm9) using Tophat (version 2.0.4), during which procedure the UCSC gene annotation file downloaded from Cufflinks website (http://cole-trapnelllab.github.io/cufflinks/igenome_table/index.html) was used (the '-G' option). The sequencing and mapping information were shown in Table 1. Cuffdiff (version 2.0.4) [7] was then applied on the aligned data set against the RefSeq gene annotation, to determine differentially expressed genes with a 'significant' status. The GO analysis of the differentially expressed genes was performed using DAVID (http://david.abcc.ncifcrf.gov/).

References

[1] L. Zhou, K. Sun, Y. Zhao, S. Zhang, X. Wang, Y. Li, L. Lu, X. Chen, F. Chen, X. Bao, X. Zhu, L. Wang, L.Y. Tang, M.A. Esteban, C.C. Wang, R. Jauch, H. Sun, H. Wang, Linc-YY1 promotes myogenic differentiation and muscle regeneration through an interaction with the transcription factor YY1. Nat. Commun. 6 (2015) 10026.
[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] 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.
[8] K. Sun, Y. Zhao, H. Wang, H. Sun, Sebnif: an integrated bioinformatics pipeline for the identification of novel large intergenic noncoding RNAs (lincRNAs)–application in human skeletal muscle cells. PLoS One 9 (2014) e84500.