Genomics Data 2 (2014) 216–218

Contents lists available at ScienceDirect

Genomics Data

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

Data in Brief

Transcriptome profiling of Set5 and Set1 methyltransferases: Tools for visualization of gene expression

Glòria Mas Martín [1,2], Devin A. King [2], Pablo E. Garcia-Nieto, Ashby J. Morrison *

Department of Biology, Stanford University, Stanford 94305, CA, USA

ARTICLE INFO

Article history:
Received 12 June 2014
Received in revised form 1 July 2014
Accepted 2 July 2014
Available online 11 July 2014

Keywords:
Set5
Set1
Methyltransferase
Gene expression
RNA-Seq

ABSTRACT

Cells regulate transcription by coordinating the activities of multiple histone modifying complexes. We recently identified the yeast histone H4 methyltransferase Set5 and discovered functional overlap with the histone H3 methyltransferase Set1 in gene expression. Specifically, using next-generation RNA sequencing (RNA-Seq), we found that Set5 and Set1 function synergistically to regulate specific transcriptional programs at subtelomeres and transposable elements. Here we provide a comprehensive description of the methodology and analysis tools corresponding to the data deposited in NCBI's Gene Expression Omnibus (GEO) under the accession number GSE52086. This data complements the experimental methods described in Mas Martín G et al. (2014) and provides the means to explore the cooperative functions of histone H3 and H4 methyltransferases in the regulation of transcription. Furthermore, a fully annotated R code is included to enable researchers to use the following computational tools: comparison of significant differential expression (SDE) profiles; gene ontology enrichment of SDE; and enrichment of SDE relative to chromosomal features, such as centromeres, telomeres, and transposable elements. Overall, we present a bioinformatics platform that can be generally implemented for similar analyses with different datasets and in different organisms.

Specifications	
Organism/cell line/tissue	*Saccharomyces cerevisiae*
Sex	N/A
Sequencer or array type	Illumina HiSeq2000
Data format	Raw data: FASTQ Processed data: TXT
Experimental factors	Wildtype BY4741 vs *set1*Δ, *set5*Δ, catalytic inactive $SET5_{Y402A}$ and *set1*Δ *set5*Δ mutant strains
Experimental features	To understand the cooperative function of the methyltransferases Set1 and Set5 in gene expression, total mRNA was obtained from two independent biological replicates each of wildtype (WT), *set1*Δ, *set5*Δ, catalytic inactive $SET5_{Y402A}$ and *set1*Δ *set5*Δ strains. Gene expression profiles of the single and double mutants were generated and analyzed.
Consent	N/A
Sample source location	N/A

Deposited data can be found here: http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE52086.

Experimental design, materials and methods

Yeast strains and media

The *Saccharomyces cerevisiae* haploid strains used to generate gene expression profiles are listed in Table 1. Deletion strains were obtained from the Yeast Knockout Collection (YKO, Open Biosystems) or generated by standard PCR-mediated gene disruption as described [1,2]. The $SET5_{Y402A}$ strain harbors a catalytic inactive Set5 protein and was generated as previously described [1,2].

Two single colonies for each strain were cultured overnight in YPD media containing 2% Dextrose with shaking at 270 rpm at 30 °C. Overnight cultures were diluted in 5 mL of YPD to OD_{600} = 0.1, grown to mid-log phase (OD_{600} = 0.8) shaking at 270 rpm at 30 °C. 1.5 mL of each culture were harvested by centrifugation (3000 rpm 5 min). Pellets were washed in 1 mL of ice-cold water, flash frozen and stored at −80 °C until ready to use.

* Corresponding author. Tel.: +1 650 724 0422.
E-mail addresses: gloria.mas@crg.eu (G. Mas Martín), devking@stanford.edu (D.A. King), paedugar@stanford.edu (P.E. Garcia-Nieto), ashbym@stanford.edu (A.J. Morrison).
[1] Present address: Center for Genomic Regulation, Barcelona 08003, Spain.
[2] These authors contributed equally to this work.

http://dx.doi.org/10.1016/j.gdata.2014.07.001

Table 1
Yeast strains used for RNA-Seq. YKO, Yeast Knockout collection from Open Biosystems.

Strain	Genotype	Background	Reference
Wildtype	*MATa his3Δ1 leu2Δ0 met15Δ0 ura3Δ0*	BY4741	YKO
YGM76	*MATa set5Δ::NATMX*	BY4741	[2]
YGM2	*MATa set1Δ::KANMX*	BY4741	YKO
YGM77	*MATa set5Δ::NATMX set1Δ::KANMX*	BY4741	[2]
YGM168	*MATa SET5::SET5$_{Y402A}$::NATMX*	BY4741	[2]

RNA extraction

Total RNA from 1.5 mL of a mid-log culture pellet was isolated using the MasterPure™ Yeast RNA Purification Kit (Epicentre; cat. no. QER09015) following manufacturer's instructions. RNA samples were treated with DNAseI for 10 min to eliminate contaminating DNA. The Agilent Technologies 2100 Bioanalyzer instrument was used to assess RNA quality and concentration. For all samples, RNA Integrity Number was 6.2 or higher.

RNA-Seq library generation

To enrich for mRNA, a total of 8 μg of purified RNA were used as input material for the Illumina TruSeq™ RNA Sample Preparation v2 Low-Throughput kit (Illumina; cat. no. RS-122-2001). Samples were processed as specified by the manufacturer's protocol, using poly-T oligo-attached magnetic beads and two consecutive rounds of enrichment preceding an mRNA fragmentation step. Next, fragmented mRNA samples were subjected to reverse transcription to generate cDNA using SuperScript II reverse transcriptase and random primers as indicated by the Illumina TruSeq™ protocol. The generated cDNA was then converted to double stranded cDNA and subjected to End repair and 3′Adenylation. Multiple indexing adapters were then ligated to the end of the ds cDNA, followed by a PCR enrichment step. For all of the resulting libraries, the quality and size – with an expected band approximately at 260 bp – were verified on an Agilent Technologies 2100 Bioanalyzer.

RNA-Seq and analysis

Indexed libraries from BY4741 wildtype, *set1*Δ, *set5*Δ, *SET5$_{Y402A}$* and *set1*Δ *set5*Δ mutant strains were subjected to RNA-Seq on an Illumina HiSeq2000 platform according to the manufacturer's protocols (Illumina). The experiment was designed to generate relatively long 101 bp sequences (single-end) to improve specificity of the mapping results. The quality of the raw sequence reads was assessed using FastQC software, with close examination of the "per base sequence quality" results, to ensure accuracy of the base call along the length of the read, and the "overrepresented sequences" results, to ensure the absence of Illumina-specific contaminating oligos. Two replicates of each sample were included, with >10 M mapped reads per replicate. FastQC did not identify any errors in the quality of the sequenced libraries and no additional read pre-processing steps (e.g. trimming) were performed.

Gene expression was quantified using the FPKM (Fragments Per Kilobase of transcript per Million mapped reads) normalization method [3]. This method estimates the transcript level using the number of reads mapped to a given gene (read count), after normalizing the read count by gene length and the total number of mapped reads in the sample. We opted to use the Tuxedo software suite (Bowtie, TopHat, Cufflinks, CummeRbund) for FPKM quantification due to ease of installation and usage, as well as integration with the R statistical computing environment. For gene expression quantification, reads were processed using the "Quantification of reference annotation only" protocol [4]. Specifically, single 101 bp reads were mapped to the *S. cerevisiae* reference Ensembl EF4 genome with TopHat, specifying '-no-novel-juncs'. Gene expression and differential transcription between WT and mutant cells were then determined using Cuffdiff. The Cuffdiff program, included in Cufflinks, was used to assess biases in read distribution across each transcript and to estimate the statistical significance of gene expression changes between samples. The Cuffdiff test provides q-values, which are p-values adjusted using an optimized False Discovery Rate (FDR) approach. FDR provides a powerful means to mitigate statistical artifacts from multiple testing. The Cufflinks results were then accessed in the R statistical computing environment using CummeRbund (v2.0.0), and a table of expression values was generated using the fpkmMatrix() function. The lists of significantly differentially expressed (SDE) genes were obtained using the getSig() function and are included as text files in the supplementary data. FPKM expression data is available through GEO and included in the supplementary files as 'GSE52086_processed_data.txt'. In addition to the 0.05 q-value threshold from Cufflinks, our criteria for defining SDE genes included a fold-change threshold of >1.7, to ensure that biologically relevant gene expression changes were considered for downstream analyses. Using these criteria, we identified a total of 42 SDE genes in *set5*Δ cells, 183 SDE genes in *set1*Δ, and 250 SDE genes in *set5*Δ *set1*Δ cells. A very similar gene expression profile to *set5*Δ was observed for the *SET5$_{Y402A}$* catalytic inactive mutant strain. The complete lists of SDE genes are included in the supplementary files 'set1_sig_genes.txt', 'set5_sig_genes.txt' and 'set5set1_sig_genes.txt'.

After defining SDE genes, we explored the characteristics of the mutant genesets by looking at enrichment in specific biological pathways, expression levels, and locations in the genome relative to annotated genomic features. To assess gene set enrichment near certain genomic regions, locations of chromosomal features were downloaded from the *Saccharomyces* Genome Database (SGD) using YeastMine query builder (included in the supplementary files as 'SGD_genomic_features.tsv'), and distances from gene transcription start sites (TSS) to each chromosomal feature were calculated. Significance of enrichment near specific genomic features was tested using the Wilcoxon rank sum (WRS) statistical method in R. The WRS test is a non-parametric analog to the *t* test, and enables comparisons of distribution location shift in FPKM expression values from two populations of genes. The distribution of gene distances to the nearest feature of the indicated geneset was compared with the genome-wide distribution of distances. Reported *P* values are from the two-sided test, with the alternative hypothesis that the true location shift is not equal to zero. For the reported significant *P* values, the geneset distributions were shifted closer to the indicated feature than would be expected by chance given the genome-wide distribution.

The file 'epi2014_RNAseq_helper_functions.RData' included in the supplementary files provides the complete set of helper functions used in the RNA-Seq analyses described above, and supplementary file 'epi2014_Set5_DAK.R' contains the R script used to generate Figures 1, 2 and 3 as in reference [1].

Discussion

This manuscript provides the full complement of experimental and bioinformatics methods developed in Ref. [1], aimed at examining the cooperative functions of the yeast histone methyltransferases Set5 and Set1 in gene expression. The R code included here comprises the helper functions used to specifically perform the RNA-Seq analysis described above ('epi2014_RNAseq_helper_functions.RData'), and an annotated script ('epi2014_Set5_DAK.R') that outlines in detail the analyses in Ref. [1]. Importantly, the script is intended to provide a template for similar gene expression analyses relating transcriptional patterns to chromosomal features with different datasets and model organisms. Overall, this is a useful bioinformatics toolbox that will bring transparency and adaptability to future transcriptome analysis.

Acknowledgment

Illumina sequencing services were performed by the Stanford Center for Genomics and Personalized Medicine. This work was supported by a National Institutes of Health grant (GM085212).

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.gdata.2014.07.001.

References

[1] G. Mas Martín, D.A. King, E.M. Green, P.E. Garcia-Nieto, R. Alexander, S.R. Collins, N.J. Krogan, O.P. Gozani, A.J. Morrison, Set5 and Set1 cooperate to repress gene expression at telomeres and retrotransposons. Epigenetics 9 (2014) 513–522, http://dx.doi.org/10.4161/epi.2764 (PMID: 24442241).

[2] E.M. Green, G. Mas Martín, N.L. Young, B.A. Garcia, O. Gozani, Methylation of H4 lysines 5, 8 and 12 by yeast Set5 calibrates chromatin stress responses. Nat. Struct. Mol. Biol. 19 (2012) 361–363, http://dx.doi.org/10.1038/nsmb.2252 (PMID: 22343720).

[3] 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, http://dx.doi.org/10.1038/nprot.2012.016 (PMID: 22383036).

[4] 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, http://dx.doi.org/10.1038/nbt.162 (PMID: 20436464).