Genomics Data 7 (2016) 144–147

Contents lists available at ScienceDirect

Genomics Data

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

Transcriptional profiling of the epigenetic regulator Smchd1

Ruijie Liu [a], Kelan Chen [a,b], Natasha Jansz [a,b], Marnie E. Blewitt [a,b,*], Matthew E. Ritchie [a,b,c,*]

[a] Molecular Medicine Division, The Walter and Eliza Hall Institute of Medical Research, Parkville, Victoria 3052, Australia
[b] Department of Medical Biology, The University of Melbourne, Parkville, Victoria 3010, Australia
[c] School of Mathematics and Statistics, The University of Melbourne, Parkville, Victoria 3010, Australia

ARTICLE INFO

Article history:
Received 22 December 2015
Accepted 30 December 2015
Available online 31 December 2015

Keywords:
RNA-sequencing
voom
Sample variability
Epigenetics

ABSTRACT

Smchd1 is an epigenetic repressor with important functions in healthy cellular processes and disease. To elucidate its role in transcriptional regulation, we performed two independent genome-wide RNA-sequencing studies comparing wild-type and *Smchd1* null samples in neural stem cells and lymphoma cell lines. Using an R-based analysis pipeline that accommodates observational and sample-specific weights in the linear modeling, we identify key genes dysregulated by Smchd1 deletion such as clustered protocadherins in the neural stem cells and imprinted genes in both experiments. Here we provide a detailed description of this analysis, from quality control to read mapping and differential expression analysis. These data sets are publicly available from the Gene Expression Omnibus database (accession numbers GSE64099 and GSE65747).

Specifications	
Organism/cell line/tissue	*Mus musculus*, C57BL/6J strain for lymphoma cell lines, FVB/N/C57BL/6J F1 for neural stem cell lines.
Sex	Male.
Sequencer or array type	NSC data: Libraries prepared with the Illumina TruSeq Total Stranded RNA kit and sequenced on an Illumina HiSeq 2000 with Illumina TruSeq SBS Kit v3-HS reagents as 100 bp paired-end reads. Lymphoma data: Libraries prepared with the Illumina TruSeq RNA Sample Preparation Kit v2 and sequenced on an Illumina HiSeq 2000 with Illumina TruSeq SBS Kit v3-HS reagents as 100 bp reads (paired and single-end).
Data format	Raw (fastq) and summarized counts.
Experimental factors	RNA was obtained from Smchd1 null and wild-type samples.
Experimental features	Neural stem cells were derived from E14.5 male embryos. Lymphoma cells were derived from lethally irradiated mice transplanted with fetal liver cells from E14.5 male embryos at the time when animals for sacrificed due to end-stage lymphoma. The lymphoma cells were plated out for growth in vitro and resulting cell lines analyzed here.
Consent	All animal experiments were carried out in accordance with the Walter and Eliza Hall Institute of Medical Research Animal Ethics Committee guidelines (AEC 2011.027).
Sample source location	Melbourne, Australia.

1. Direct link to deposited data

http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE64099
http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE65747

2. Introduction

Smchd1 (structural maintenance of chromosomes hinge domain containing 1) is an important epigenetic modifier that has a critical role in X inactivation [1,2] and genomic imprinting [3,4]. Although initial studies of Smchd1 used these two classic models of epigenetic control, it has become clear that Smchd1 has a broader role in regulating gene expression during normal development [5], in cancer [6] and in the development of facioscapulohumeral muscular dystrophy (FSHD) [7–9].

We were particularly interested to look at the role of Smchd1 in regulating gene expression via RNA sequencing (RNA-seq), as Smchd1 is a repressor protein, and so the very low level of expression of Smchd1 repressed genes best lends itself to RNA-seq over array-based platforms. To this end, we conducted RNA-seq experiments in two model systems, the first was in neural development using neural stem cells and the second was in a cancer model using lymphoma cell lines. In both experiments, samples with wild-type levels of Smchd1 are compared to samples with a null allele of this gene. This article describes our analyses of these two data sets, using a consistent, R-based pipeline that can deal with both observational and sample-level heterogeneity.

* Corresponding authors at: Molecular Medicine Division, The Walter and Eliza Hall Institute of Medical Research, Parkville, Victoria 3052, Australia.
E-mail addresses: blewitt@wehi.edu.au (M.E. Blewitt), mritchie@wehi.edu.au (M.E. Ritchie).

http://dx.doi.org/10.1016/j.gdata.2015.12.027

3. Experimental design, materials and methods

3.1. Mouse strains and sample information

MommeD1 mutant mice were maintained on the FVB/N inbred background, and backcrossed with C57BL/6 mice for more than 15 generations to produce C57BL/6 MommeD1 congenic mice (as previously described in [1]). Neural stem cells were isolated and cultured from the brains of FVB/C57BL/6J F1 E14.5 male embryos, homozygous or wild-type for the *Smchd1*$^{\mathrm{MommeD1}}$ mutation as described in [5]. Lymphoma cell lines were derived from a gene trap allele of Smchd1, described in [6]. This allele was backcrossed onto C57BL/6J, then crossed onto the Eμ-Myc transgenic background to generate *Smchd1*$^{gt/gt}$ *Eμ-MycTg*/+ embryos and their wild-type controls, for transplant and generation of lymphomas. Genotyping was carried out as described in [1,2] and [6]. Experimental animals were treated in accordance with the Australian Government National Health and Medical Research Council guidelines under the approval from the animal ethics committees of the Walter and Eliza Hall Institute (WEHI AEC 2011.027).

3.2. RNA-seq sample preparation and sequencing

Qiagen RNeasy Mini kits were used to extract RNA from *Smchd1*$^{\mathrm{MommeD1/MommeD1}}$ and *Smchd1*$^{+/+}$ wild-type NSCs according to the manufacturer's instructions. RNA was quantified using the NanoDrop 1000 Spectrophotometer (Thermo Scientific) and RNA integrity assessed with the Agilent Bioanalyzer 2100 (Agilent Technologies). Illumina's TruSeq total RNA sample preparation kit was used to prepare libraries for sequencing, which was performed by the Australian Genome Research Facility (Melbourne, Australia) on the Illumina HiSeq 2000 platform to obtain 100 bp paired-end reads.

For the Lymphoma data set, Qiagen RNeasy Mini kits were used to extract RNA from *Smchd1*$^{MommeD1/MommeD1}$;*E*μMycTg/+ and *Smchd1*$^{+/+}$;*E*μMycTg/+ lymphoma cells. Samples were prepared for sequencing at the Australian Genome Research Facility where quality control, library preparation (using Illumina's TruSeq RNA sample preparation kit) and sequencing on the Illumina HiSeq 2000 platform was performed to obtain 100 bp paired-end (for 6 out of 7 samples) or single-end (for 1 sample) reads.

3.3. Quality control and data pre-processing

The FastQC software [10] was used to assess the quality of the raw sequence data. Fig. 1 displays the distribution of sequencing quality (Phred) scores at each base position across reads from a representative RNA-seq sample from each data set. Although variation in base quality is observed across the read, with slightly lower quality at the beginning and end, median quality is above 34 (corresponding to a probability of an incorrect base call below 0.0004) for the entire read. Similar boxplots of base quality scores were observed for other samples (data not shown).

Sequences were then mapped to the mouse reference genome (mm10) using the *Rsubread* program [11] and gene-level counts were obtained by the *featureCounts* procedure [12].

Further analysis was carried out using the *edgeR* [13] and *limma* [14] R/Bioconductor packages. Counts-per-million (CPM) were calculated for each gene to standardize for differences in library-size and filtering was carried out to retain genes with a baseline expression level of at least 0.5 CPM in 3 or more samples. For each data set, TMM normalization [15] was applied and a multidimensional scaling (MDS) plot based on the $\log_2$(CPM) was generated to show relationships between samples (Fig. 2). In both data sets, we observe samples that do not cluster well with their respective replicates of the same genotype. Sample 6 in the NSC data (Fig. 2A) and samples 1 and 7 in the Lymphoma data (Fig. 2B) are more variable than the other replicates of the same type. For NSC sample 6 and Lymphoma sample 7, there was no experimental factor that could be identified to explain this phenomenon. Lymphoma sample 1 on the other hand was the only single-end sample in this experiment that was processed on a different day to the other samples, leading us to conclude that batch processing differences was the likely cause of the additional variation.

3.4. Differential expression analysis

Based on inspection of the MDS plots, which showed variability between replicate samples, linear models [16] with combined observational and sample weights [17,18] were fitted to the $\log_2$(CPM) to summarize over replicate samples. This strategy, implemented in the *voomWithQualityWeights* function, down-weights low abundance observations, which are systematically more variable (Fig. 2C) and

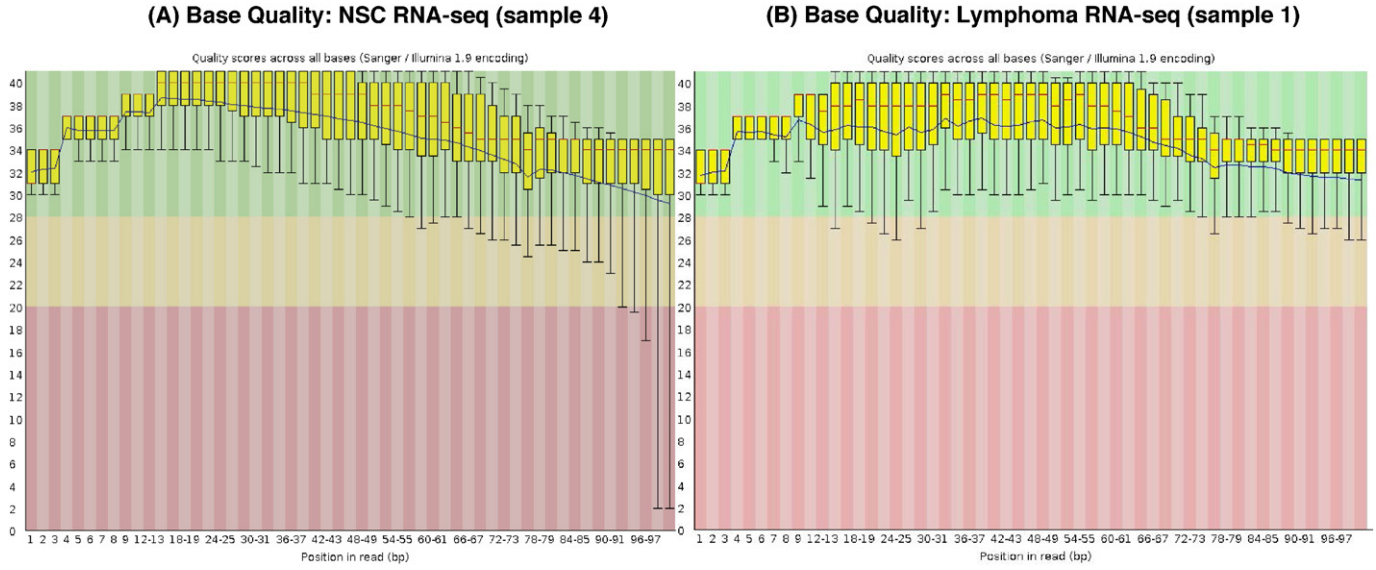

Fig. 1. Quality assessment at the read level. Boxplots of base-calling Phred scores at different base positions across all the reads in representative libraries from NSC RNA-seq (A) and Lymphoma cell line RNA-seq (B) experiments generated by FastQC. The box represents 25% and 75% quantiles of the scores with median score marked by the red line. Whiskers mark the 10% and 90% quantiles and blue lines show the mean quality score.