Genomics Data 3 (2015) 6–7

Contents lists available at ScienceDirect

Genomics Data

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

Data in Brief

High-throughput whole-genome sequencing of E14 mouse embryonic stem cells

Danny Incarnato, Francesco Neri *

Human Genetics Foundation (HuGeF), via Nizza 52, 10126 Torino, Italy
Next Generation Intelligence (NGI), Torino, Italy

ARTICLE INFO

Article history:
Received 27 October 2014
Received in revised form 30 October 2014
Accepted 30 October 2014
Available online 7 November 2014

Keywords:
ESC
NGS
Whole-genome E14

ABSTRACT

Mouse E14 embryonic stem cells (ESCs) are the most used ESC line, often employed for genome-wide studies involving next generation sequencing analysis [1–5]. More than 2 × 10 E9 sequences made on Illumina platform derived from the genome of E14 embryonic stem cells cultured in our laboratory were used to build a database of about 2.7 × 10 E6 single nucleotide variant [6]. The database was validated using other two sequencing datasets from other laboratory and high overlap was observed. The identified variants are enriched on intergenic regions, but several thousands reside on gene exons and regulatory regions, such as promoters, enhancers, splicing site and untranslated regions of RNA, thus indicating high probability of an important functional impact on the molecular biology of these cells. We created a new E14 genome assembly including the new identified variants and used it to map reads from next generation sequencing data generated in our laboratory or in others on E14 cell line. We observed an increase in the number of mapped reads of about 5%. CpG dinucleotide showed the higher variation frequency, probably because it could be a target of DNA methylation. Data were deposited in GEO datasets under reference GSM1283021 and here: http://epigenetics.hugef-research.org/data.php.

Specifications	
Organism/cell line/tissue	Mouse E14 embryonic stem cells
Sex	Male
Sequencer or array type	Illumina HiScanSQ
Data format	Raw and analyzed
Experimental factors	N/A
Experimental features	Whole genome sequencing of E14 embryonic stem cells
Consent	N/A
Sample source location	Torino, Italy

Direct link to deposited data

http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1283021
http://epigenetics.hugef-research.org/data.php

Experimental design, materials and methods

E14 mouse ES cells were cultured in ESC medium (DMEM high glucose with 15% fetal bovine serum [FBS], NNEA1x, NaPyr1x, 0.1 mM 2-mercaptoethanol, and 1500 U/ml LIF). Genomic DNA was extracted using a DNeasy Blood and Tissue kit (Qiagen).

For sequencing of E14 genome, DNA was sonicated for 17′ pulse 30″ ON/30″OFF high with Bioruptor Twin (Diagenode). Libraries were generated with DNA Sample Prep Kit (Illumina) and sequenced on Illumina HiScanSQ Platform. Basecalls performed using CASAVA version 1.8 following default parameters. Reads quality was estimated using FastQC tool v0.10.1 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Nucleotide positions with a quality score under 30 (Phred33 scale) were trimmed using the *fastx_trimmer* tool from the FASTX Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/).

After low-quality positions trimming, reads in which sequencing continued through the 3′ adapter sequence were clipped using the *fastx_clipper* tool from the FASTX Toolkit. Then, reads were aligned to the mouse genome assembly mm9 using Bowtie [7] v0.12.7 with the following parameters: -q –max /dev/null -v 1 -S –sam-nohead -m 1. Reads with the same mapping positions were collapsed into one using the rmdup tool from SAMtools. Variants calling was performed using the mpileup tool from SAMtools [8]. Next, we used VCFtools [9]

* Corresponding author at: Via Nizza 52, HuGeF, Torino, Italy. Tel.: +39 011 6709531.
E-mail addresses: danny.incarnato@hugef-torino.org (D. Incarnato), francesco.neri@hugef-torino.org, fneri@nextgenintelligence.com (F. Neri).

http://dx.doi.org/10.1016/j.gdata.2014.10.023
2213-5960/

v0.1.11 (http://vcftools.sourceforge.net/) to select only SNVs with coverage of ≥10 and a frequency of ≥0.5. Moreover, using custom Perl scripts we discarded sites with more than one variant call at the same place. Finally, using the GATK v2.7-4 (http://www.broadinstitute.org/gatk/) *FastaAlternateReferenceMaker* function we created the new reference E14 assembly from the mm9 genome assembly.

These data can be found at: http://epigenetics.hugef-research.org/data.php.

References

[1] A. Krepelova, F. Neri, M. Maldotti, S. Rapelli, S. Oliviero, Myc and max genome-wide binding sites analysis links the Myc regulatory network with the polycomb and the core pluripotency networks in mouse embryonic stem cells. PLoS ONE 9 (2014) e88933, http://dx.doi.org/10.1371/journal.pone.0088933.

[2] F. Neri, D. Incarnato, A. Krepelova, S. Rapelli, A. Pagnani, R. Zecchina, et al., Genome-wide analysis identifies a functional association of Tet1 and Polycomb PRC2 in mouse embryonic stem cells. Genome Biol. 14 (2013) R91, http://dx.doi.org/10.1186/gb-2013-14-8-r91.

[3] X. Chen, H. Xu, P. Yuan, F. Fang, M. Huss, V.B. Vega, et al., Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. Cell 133 (2008) 1106–1117, http://dx.doi.org/10.1016/j.cell.2008.04.043.

[4] F. Neri, A. Krepelova, D. Incarnato, M. Maldotti, C. Parlato, F. Galvagni, et al., Dnmt3L antagonizes DNA methylation at bivalent promoters and favors DNA methylation at gene bodies in ESCs. Cell 155 (2013) 121–134, http://dx.doi.org/10.1016/j.cell.2013.08.056.

[5] K. Williams, J. Christensen, M.T. Pedersen, J.V. Johansen, P.A.C. Cloos, J. Rappsilber, et al., TET1 and hydroxymethylcytosine in transcription and DNA methylation fidelity. Nature 473 (2011) 343–348, http://dx.doi.org/10.1038/nature10066.

[6] D. Incarnato, A. Krepelova, F. Neri, High-throughput single nucleotide variant discovery in E14 mouse embryonic stem cells provides a new reference genome assembly. Genomics 104 (2014) 121–127, http://dx.doi.org/10.1016/j.ygeno.2014.06.007.

[7] B. Langmead, C. Trapnell, M. Pop, S.L. Salzberg, Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 10 (2009) R25, http://dx.doi.org/10.1186/gb-2009-10-3-r25.

[8] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, et al., The sequence alignment/map format and SAMtools. Bioinformatics 25 (2009) 2078–2079, http://dx.doi.org/10.1093/bioinformatics/btp352.

[9] P. Danecek, A. Auton, G. Abecasis, C.A. Albers, E. Banks, M.A. DePristo, et al., The variant call format and VCFtools. Bioinformatics 27 (2011) 2156–2158, http://dx.doi.org/10.1093/bioinformatics/btr330.