Exploring Lineage-Specific Enhancers by Integrating Enhancer Transcription, Epigenomic Features, Sequence Motifs, and Transcription Factor Expression

This manuscript was automatically generated from vsmalladi/tfsee-manuscript@d64ea19 on November 6, 2017.

Authors

- **Venkat Malladi**
  0000-0002-0144-0564 · vsmalladi · katatonikkat

Abstract

The identification of transcription factors (TF) driving the formation of active enhancers that regulate the expression of target genes remains an open problem. We have developed a computational framework that identifies cell type-specific enhancers and their cognate TFs by integrating multiple genomic assays that probe the transcriptomes (GRO-seq and RNA-seq) and epigenomes (ChIP-seq) of various samples. Our method, called Total Functional Score of Enhancer Elements (TFSEE), integrates the magnitude of enhancer transcription (GRO-seq), enrichment of marks associated with enhancers (H3K4me1 and H3K27ac ChIP-seq), TF mRNA expression levels (RNA-seq), and TF motif p-values (MEME). This method has allowed us to explore the enhancer landscape in different cell types that share common origins or are biologically related, including distinct molecular subtypes of breast cancer, and embryonic stem cells (ESCs) and their derived lineages. Using TFSEE, we have identified key breast cancer subtype-specific transcription factors that are bound at active enhancers and dictate gene expression patterns determining growth outcomes. To demonstrate the broader utility of our approach, we have used this algorithm to identify transcription factors during the differentiation of embryonic stem cells into pancreatic cells. Taken together our results show that TFSEE can be used to perform multilayer genomic data integration to uncover novel cell type-specific transcription factors that control lineage-specific enhancers.

Introduction

Results

Discussion

Acknowledgments

Material and Methods

Genomic Data Curation

We used previously published GRO-seq, ChIP-seq and RNA-seq data from [1,2] of time course differentiation of human embryonic stem cells (hESC) to pancreatic endoderm (PE). All data sets are available from NCBI's Gene Expression Omnibus [3] or EMBL-EBI's ArrayExpress [4] repositories using the accession numbers listed in Table 1.

Table 1: **Description and accession numbers of GRO-seq, ChIP-seq and RNA-seq datasets.**

Assay	Accessions
GRO-seq	GSM1316306, GSM1316313, GSM1316320, GSM1316327, GSM1316334
H3K4me3 ChIP-seq	ERR208008, ERR208014, ERR207998, ERR20798, ERR207999
H3K4me1 ChIP-seq	GSM1316302, GSM1316303, GSM1316309, GSM1316316, GSM1316317, GSM1316310, GSM1316323, GSM1316324, GSM1316330, GSM1316331
H3K27ac ChIP-seq	GSM1316300, GSM1316301, GSM1316307, GSM1316308, GSM1316314, GSM1316315, GSM1316321, GSM1316322, GSM1316328, GSM1316329

Assay	Accessions
Input ChIP-seq	ERR208001, ERR208012, ERR207984, ERR208011, ERR207986, GSM1316304, GSM1316305, GSM1316311, GSM1316312, GSM1316318, GSM1316319, GSM1316325, GSM1316326, GSM1316332, GSM1316333
RNA-seq	ERR266333, ERR266335, ERR266337, ERR266338, ERR266341, ERR266342, ERR266344, ERR266346, ERR266349, ERR266351

Analysis of ChIP-seq Data Sets

The raw reads were aligned to the human reference genome (GRCh37/hg19) using default parameters in Bowtie (ver. 1.0.0) [5]. The aligned reads are subsequently filtered for quality and uniquely mappable reads using Samtools (ver. 0.1.19) [6] and Picard (ver. 1.127) [7]. Library complexity is measured using BEDTools (v 2.17.0) [8] and meet ENCODE data quality standards [9]. Relaxed peaks were called using MACS (v2.1.0) [10] with a p-value of 1×10^{-2} for each replicate, pooled replicates' reads and pseudoreplicates. Peak calls that are replicated from the pooled replicated that are either observed in both replicates, or in both pseudoreplicates are used for subsequent analysis.

Analysis of RNA-seq Data Sets

The raw reads were aligned to the human reference genome (GRCh37/hg19) using default parameters in STAR (v2.4.2a) [11]. Quantification of genes against Gencode (v.19) [12] annotations was done using default parameters in RSEM (v 1.2.31) [13].

Analysis of GRO-seq Data

The GRO-seq reads were trimmed to the first 36 bases, to trim adapter and low quality sequence, using default parameters of fastx_trimer in fastx-toolkit (v.0.0.13.2) [14]. The trimmed reads were aligned to the human reference genome (GRCh37/hg19) using default parameters in BWA (v0.7.12) [15].

Kernel Density

Kernel density plot representations were used to express the univariate distribution of ChIP-seq reads under peaks, RNA-seq reads for protein-coding genes and GRO-seq reads for short paired and short unpaired eRNAs. The kernel density plots were calculated in Python (ver. 2.7.11) using the kdeplot function from seaborn libary [16] with default parameters.

Defining Transcription Start Sites

We made distinct transcription start sites (TSS) for protein-coding genes from Gencode (v.19) [12] annotations using MakeGencodeTSS [17].

References

1. **Dynamic Chromatin Remodeling Mediated by Polycomb Proteins Orchestrates Pancreatic Differentiation of Human Embryonic Stem Cells**
Ruiyu Xie, Logan J. Everett, Hee-Woong Lim, Nisha A. Patel, Jonathan Schug, Evert Kroon, Olivia G. Kelly, Allen Wang, Kevin A. D'Amour, Allan J. Robins, … Maike Sander
Cell Stem Cell (2013-02) https://doi.org/10.1016/j.stem.2012.11.023

2. **Epigenetic Priming of Enhancers Predicts Developmental Competence of hESC-Derived Endodermal Lineage Intermediates**
Allen Wang, Feng Yue, Yan Li, Ruiyu Xie, Thomas Harper, Nisha A. Patel, Kayla Muth, Jeffrey Palmer, Yunjiang Qiu, Jinzhao Wang, … Maike Sander
Cell Stem Cell (2015-04) https://doi.org/10.1016/j.stem.2015.02.013

3. **GEO**
Gene Expression Omnibus
https://www.ncbi.nlm.nih.gov/geo/

4. **ArrayExpress**
ArrayExpress – functional genomics data
http://www.ebi.ac.uk/arrayexpress/

5. **Ultrafast and memory-efficient alignment of short DNA sequences to the human genome**
Ben Langmead, Cole Trapnell, Mihai Pop, Steven L Salzberg
Genome Biology (2009) https://doi.org/10.1186/gb-2009-10-3-r25

6. **The Sequence Alignment/Map format and SAMtools**
H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin,
Bioinformatics (2009-06-08) https://doi.org/10.1093/bioinformatics/btp352

7. **Picard**
Broad Institute
GitHub http://broadinstitute.github.io/picard/

8. **BEDTools: a flexible suite of utilities for comparing genomic features**
Aaron R. Quinlan, Ira M. Hall
Bioinformatics (2010-01-28) https://doi.org/10.1093/bioinformatics/btq033

9. **ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia**
S. G. Landt, G. K. Marinov, A. Kundaje, P. Kheradpour, F. Pauli, S. Batzoglou, B. E. Bernstein, P. Bickel, J. B. Brown, P. Cayting, … M. Snyder
Genome Research (2012-09-01) https://doi.org/10.1101/gr.136184.111

10. **Identifying ChIP-seq enrichment using MACS**
Jianxing Feng, Tao Liu, Bo Qin, Yong Zhang, Xiaole Shirley Liu
Nature Protocols (2012-08-30) https://doi.org/10.1038/nprot.2012.101

11. **STAR: ultrafast universal RNA-seq aligner**
Alexander Dobin, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, Thomas R. Gingeras
Bioinformatics (2012-10-25) https://doi.org/10.1093/bioinformatics/bts635

12. **GENCODE: The reference human genome annotation for The ENCODE Project**
J. Harrow, A. Frankish, J. M. Gonzalez, E. Tapanari, M. Diekhans, F. Kokocinski, B. L. Aken, D. Barrell, A. Zadissa, S. Searle, … T. J. Hubbard
Genome Research (2012-09-01) https://doi.org/10.1101/gr.135350.111

13. **RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome**
Bo Li, Colin N Dewey
BMC Bioinformatics (2011) https://doi.org/10.1186/1471-2105-12-323

14. **FASTX-Toolkit**
Hannon Lab
http://hannonlab.cshl.edu/fastx_toolkit/

15. **Fast and accurate short read alignment with Burrows-Wheeler transform**
H. Li, R. Durbin
Bioinformatics (2009-05-18) https://doi.org/10.1093/bioinformatics/btp324

16. **seaborn: statistical data visualization — seaborn 0.8.1 documentation**(2017-09-03) http://seaborn.pydata.org/

17. **MakeGenecodeTSS**
Sarah Djebali
GitHub https://github.com/sdjebali/MakeGencodeTSS