CENTIPEDE: application to ATAC-seq footprinting of DNA:protein interactions

Roger Pique-Regi [1]*, Molly Estill [1], Donovan Watza [1], Elisabeth Doman [1] and Francesca Luca [1]*

[1]Center for Molecular Medicine and Genetics, Department of Obstetrics and Gynecology, Wayne State University, 540 E Canfield, Scott Hall, Detroit, MI 48201, USA

ABSTRACT

Motivation: Deciphering the regulatory sequences which control gene transcription is a critical step in understanding both cellular and condition-specific regulatory programs encoded in the human genome. Transcriptional response is typically regulated by transcription factors (TFs) which are known to bind specific regulatory sequence motifs. Profiling the binding activity of these factors can be quickly accomplished at a genome-wide scale with the recently developed technique ATAC-seq, which utilizes the Tn5 transposase to fragment and tag accessible DNA. When coupled with an advanced computational method such as CENTIPEDE, ATAC-seq data can be used to generate binding models for TFs with known motifs across the genome. To date, there are no methods that efficiently incorporate the information provided by paired-end sequencing, which allows both the identification of the library fragment length as well as the two cleavage locations that generated the fragment.

Results: We have extended CENTIPEDE to utilize fragment length information in a simple yet efficient way. Our results indicate that paired-end sequencing provides a more informative footprint model for ATAC-seq libraries, which leads to greater accuracy in predicting TF binding. These results were validated with ChIP-seq data (ENCODE Project) for multiple factors including CTCF, NRSF, NRF-1, and NFkB.

Availability: `http://github.com/piquelab/CENTIPEDE` (under development)

Contact: `fluca@wayne.edu;rpique@wayne.edu`

Recent technological advances in molecular biology combined with high-throughput sequencing have made possible the analysis of many different types of regulatory function across the entire genome and across a large number of cell-types. At the molecular level the regulatory function is usually accomplished by transcription factor (TF) proteins that recognize specific DNA sequence motifs. Functional genomics data collected by ENCODE (The ENCODE Project Consortium, 2012), Roadmap Epigenomics (Roadmap Epigenomics Consortium, 2015), and other groups (e.g., Visel *et al.* 2009) have provided a great deal of information about regulatory regions, however we are still far away of obtaining complete maps of active regulatory regions for many TFs and cellular conditions.

Although the association of transcription factors to DNA can be empirically determined through assays such as Chromatin ImmunoPrecipitation-Sequencing(ChIP-Seq), ChIP-Seq has been performed for only a subset of all DNA-associated proteins and only in selected cell types. However, alternative experimental procedures have been devised to indirectly recover actively used binding sites for a wide panel of TF by measuring chromatin accessibility, chromatin modifications, P300 ChIP-seq, enhancer RNAs and others. Many of these experimental data types are not TF specific, thus circumventing the need for TF-specific antibodies. Several computational methods have been developed that can exploit the correlation structure of these types of data to identify tissue-specific regulatory regions. For example, different types of histone modifications have been integrated using hidden Markov models (HMMs) (Ernst and Kellis, 2010), and dynamic Bayesian networks (DBNs) (Hoffman *et al.*, 2013). Chromatin accessibility measured by DNase-seq and sequence motifs have been integrated in different types of mixture models (CENTIPEDE (Pique-Regi *et al.*, 2011), PIQ (Sherwood *et al.*, 2014) and others (Boyle *et al.*, 2012; Neph *et al.*, 2012)). The advantage of DNase-seq over other indirect methods is that the DNase I cleavage pattern at single base-pair resolution around the binding site (i.e., the footprint) is very informative. A systematic comparison of DNase-seq derived CENTIPEDE predictions and ChIP-seq data from ENCODE on LCLs and K562 cells demonstrated a remarkable agreement in classifying motif instances as bound or unbound, and CENTIPEDE was used to create one of the most extensive map of transcription factor (TF) binding in LCLs.

Here we focus on adapting and applying the CENTIPEDE method to ATAC-seq data. ATAC-seq is a new experimental protocol (Buenrostro *et al.*, 2013) that similarly to DNase-seq seeks to map chromatin accessible regions of the genome. The basic difference between ATAC-seq and DNase-seq is the use of the synthetic Tn5 transposase enzyme instead of DNase I to cleave the DNA at accessible chromatin locations. A major advantage of ATAC-seq is that Tn5 can be pre-loaded with the sequencing primers as in the case of the Nextera DNA library preparation kit (Illumina), which greatly simplifies and accelerates library preparation. Similar to DNase-seq, the small sequence region actively bound by a TF is less accessible to Tn5, thus factor-specific footprint can be detected. Tn5 is less efficient in cleaving linker DNA between tightly compacted nucleosomes, but can more easily cleave between less compacted nucleosomes such as those surrounding enhancer/promoter regions. CENTIPEDE was originally designed to model at base-pair resolution the DNase I cleavage sensitivity spatial profile (i.e., footprint) characteristic of

*to whom correspondence should be addressed active TF-DNA binding and ignores paired-end information. Here we modify the CENTIPEDE model to analyze ATAC-seq data and exploit the paired-end information.

Experimental data. The lymphoblastoid cell line (LCL) GM18508 was purchased from Coriell Cell Repository. LCLs were cultured and starved according to (Maranville *et al.*, 2011). We then followed the protocol by (Buenrostro *et al.*, 2013) to lyse the cells and prepare ATAC-seq libraries, with the exception that we used the Illumina Nextera Index Kit (Cat#15055290) in the PCR enrichment step. Individual libraries fragment distribution was assessed on the Agilent Bioanalyzer and pooling proportions were determined using the qPCR Kapa library quantification kits (KAPA Biosystems). Library pools were run on the Illumina NextSeq 500 Desktop sequencer in the Luca/Pique-Regi lab and on the Illumina HiSeq 2500 at the Michigan State University Genomics Core. Libraries from three replicates were pooled and sequenced on multiple sequencing runs for a total of 753M 50bp PE reads. Additional DNase-seq data and ChIP-seq data was retrieved from the ENCODE project for evaluation purposes.

Pre-processing. Reads were aligned to the reference human genome hg19 using `bwa mem` (Li and Durbin, 2009 `http://bio-bwa.sourceforge.net`). Reads with quality <10 and without proper pairs were removed using `samtools` (`http://github.com/samtools/`) while putative duplicated reads were kept as is. Reads with different fragment length were partitioned into four bins: 1) [39-99], 2) [100-139], 3) [140-179], 4) [180-250]. For each fragment the two Tn5 insertion sites were calculated as the position 4bp after the 5'-end in the 5' to 3' direction. Then for each candidate motif a matrix $\boldsymbol{X}$ was constructed to count Tn5 insertion events: each row represents a sequence match to motif in the genome (motif instance), and each column a specific cleavage site at a relative bp and orientation with respect to the motif instance. We built a matrix $\{\boldsymbol{X}_l\}_{l=1}^4$ for each fragment length bin, each using a window half-size S=150bp resulting in $(2 \times S + W) \times 2$ columns, where W is the length of the motif in bp. The motif instances were scanned in the the human genome hg19 using position weight (PWM) models from TRANSFAC and JASPAR as previously described (Pique-Regi *et al.*, 2011).

Results. We used CENTIPEDE with four modeling alternatives. Each alternative model we tested corresponds to a different parametrization of the Negative-Multinomial distribution modeling the total number of reads in each row of $\boldsymbol{X}$ as well as the spatial distribution across columns. Namely, these different alternatives are:

- ATAC-seq reads from different fragment lengths are separated and modeled as independent Negative-Multinomial (NM) components $\{\boldsymbol{X}_l\}_{l=1}^4$.
- ATAC-seq reads from different fragment lengths are all added together (one NM component, fragment length information is ignored) $\boldsymbol{X} = \boldsymbol{X}_1 + \boldsymbol{X}_2 + \boldsymbol{X}_3 + \boldsymbol{X}_4$
- ATAC-seq reads from different fragment lengths are concatenated together (one NM component) $\boldsymbol{X} = [\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3, \boldsymbol{X}_4]$
- ATAC-seq reads from only short fragments used (one NM component) $\boldsymbol{X} = \boldsymbol{X}_1$

We ran the CENTIPEDE model for a selection of TFs with DNase-seq and ChIP-seq data available from the ENCODE project. The CENTIPEDE results on TF binding were then compared using an ROC curve (Figure 1) and the area under the curve (AUC) for each model is summarized in Table 1.

Area under the curve (AUC) values					
		ATAC-seq, Fragments with different fragment lengths are modeled:			
ChIP-seq	DNase-seq	Separated	Added	Concatenated	Short Only
CTCF	0.93	0.94	0.94	**0.95**	0.93
NRSF	0.79	**0.78**	0.75	0.71	0.74
NRF-1	0.88	0.8	0.77	**0.94**	0.74
NFkB	0.96	0.76	0.72	**0.9**	0.72
NFYA	0.97	0.89	0.86	**0.95**	0.83
SP-1	0.97	0.77	0.67	**0.93**	0.75
USF-1	0.92	0.64	0.52	**0.91**	0.62
EGR-1	0.93	0.62	0.6	**0.87**	0.59

Table 1. AUC values for ROC analysis of ATAC-seq and DNase-seq performance. Several CENTIPEDE models were tested to optimize performance on ATAC-seq data (in bold we indicate the best model for each TF).

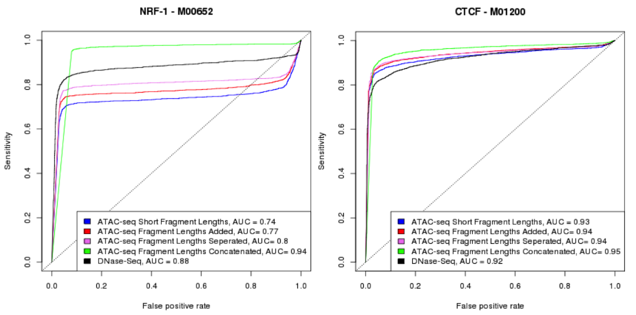

Fig. 1. ROC analysis of ATAC-seq and DNase-seq performance with CENTIPEDE. Receiver operating characteristic curve of ATAC-seq and DNase-seq performance for ENCODE ChIP-seq sites (left, CTCF; right, NRF-1). Four different CENTIPEDE models were tested to optimize performance on ATAC-seq data.

For many TFs, ATAC-seq performs similarly to DNase-seq but may require higher sequencing depth. Additional data across many cell-types and conditions may be necessary to further confirm this result. ATAC-seq has several advantages over DNase-seq: the experimental assay protocol is simpler and much faster, it uses a very small number of cells (50,000) and is therefore easier to perform and replicate on several cell types (including primary cells). Our implementation shows that CENTIPEDE with appropriate modifications can take advantage of the paired-end information provided by the ATAC-seq data and improve binding site detection as determined by comparison to ChIP-seq data, thus yielding a resolution similar to DNase-seq with a streamlined experimental workflow.

ACKNOWLEDGEMENT

We would like to thank Wayne State University HPC Grid for computational resources, the Michigan State University Genomics Core for sequencing services, and Jacob Degner and members of the Luca/Pique group for helpful comments and discussions.

Funding: NIH 1R01GM109215-01 (RPR, FL)
AHA 14SDG20450118 (FL)

REFERENCES

Boyle, A. P., Hong, E. L., Hariharan, M., Cheng, Y., Schaub, M. A., Kasowski, M., Karczewski, K. J., Park, J., Hitz, B. C., Weng, S., Cherry, J. M., and Snyder, M. (2012). Annotation of functional variation in personal genomes using RegulomeDB. *Genome Research*, **22**(9), 1790–1797.

Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y., and Greenleaf, W. J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature methods*, **10**, 1213–8.

Ernst, J. and Kellis, M. (2010). Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nature Biotechnology*, **28**(8), 817–825.

Hoffman, M. M., Ernst, J., Wilder, S. P., Kundaje, A., Harris, R. S., Libbrecht, M., Giardine, B., Ellenbogen, P. M., Bilmes, J. a., Birney, E., Hardison, R. C., Dunham, I., Kellis, M., and Noble, W. S. (2013). Integrative annotation of chromatin elements from ENCODE data. *Nucleic acids research*, **41**(2), 827–41.

Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**(14), 1754–1760.

Maranville, J. C., Luca, F., Richards, A. L., Wen, X., Witonsky, D. B., Baxter, S., Stephens, M., and Di Rienzo, A. (2011). Interactions between glucocorticoid treatment and cis-regulatory polymorphisms contribute to cellular response phenotypes. *PLoS Genet.*, **7**, e1002162.

Neph, S., Stergachis, A. B., Reynolds, A., Sandstrom, R., and Borenstein, E. (2012). Resource Circuitry and Dynamics of Human Transcription Factor Regulatory Networks. *Cell*, pages 1–13.

Pique-Regi, R., Degner, J. F., Pai, A. a., Gaffney, D. J., Gilad, Y., and Pritchard, J. K. (2011). Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Research*, **21**(3), 447–55.

Roadmap Epigenomics Consortium (2015). Integrative analysis of 111 reference human epigenomes. *Nature*, **518**(7539), 317–330.

Sherwood, R. I., Hashimoto, T., O'Donnell, C. W., Lewis, S., Barkal, A. A., van Hoff, J. P., Karun, V., Jaakkola, T., and Gifford, D. K. (2014). Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape. *Nature Biotechnology*, **32**, 171–178.

The ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.

Visel, A., Rubin, E. M., and Pennacchio, L. A. (2009). Genomic views of distant-acting enhancers. *Nature*, **461**(7261), 199–205.