frontiers in GENETICS

OPINION ARTICLE
published: 31 October 2012
doi: 10.3389/fgene.2012.00230

Current bioinformatic approaches to identify DNase I hypersensitive sites and genomic footprints from DNase-seq data

Pedro Madrigal* and Paweł Krajewski

Laboratory of Biometry, Institute of Plant Genetics, Polish Academy of Sciences, Poznań, Poland
**Correspondence: pmad@igr.poznan.pl*

Edited by:
Alexandre V. Morozov, Rutgers University, USA

Reviewed by:
Alexandre V. Morozov, Rutgers University, USA
Zhi Wei, New Jersey Institute of Technology, USA

CHROMATIN ACCESSIBILITY

The formation of regions of open chromatin or nucleosome loss in eukaryotic genomes is an important factor elucidating potential regulatory activity. Nucleosome packaging, which organizes the DNA structure, acts as a regulator of transcription by enabling or restricting protein binding, and therefore facilitating the replication and coordination of gene activity (Cockerill, 2011). In addition, chromatin accessibility, which has been determined traditionally by regions of "open" or "closed" conformation, is subjected to dynamically changing events at accessible *cis*-regulatory elements (Bell et al., 2011).

Chromatin accessibility can be examined by DNase I digestion, and then uncovered by the DNase I cleavage pattern (Wu et al., 1979). The combination of DNase I digestion and high-throughput sequencing (DNase-seq) has been used to map chromatin accessibility *in vivo* in a given tissue or cell-type on a genome-wide scale (Song and Crawford, 2010). This technique allows for an unprecedented increase both in resolution and the range spanned, compared to the pre-next generation sequencing era (Kodama et al., 2007). The current DNase-seq protocol has been adapted from the methodology described by Boyle et al. (2008a), achieving higher resolution than DNase-chip, and can be applied to any species with a sequenced genome.

Although, the analysis of data coming from sequencing technologies such as chromatin immunoprecipitation followed by sequencing (ChIP-seq), or whole transcriptome shotgun sequencing (RNA-seq) have concentrated a huge level of research effort, methodologies for the analysis of DNase-seq data are relatively immature (Song and Crawford, 2010). This data presents its own peculiarities and should not be merely treated as ChIP-seq data, but instead linked to it to provide biological insights of chromatin domains and transcriptional regulation. The general view conceives regions of open chromatin spanning nucleosome-free or nucleosome-depleted regions often in the vicinity of transcription factor binding events.

DNase I HYPERSENSITIVE SITES

DNase I hypersensitive sites (DHSs) indicate regions of an open chromatin state obtained as DNase-seq highly reproducible tag-enriched sites. The coverage formed by reads mapped uniquely, after artifact filtering (Baek et al., 2012), can be obtained as a standard format file, and visualized in a genome browser. The obtained profiles resemble to some extent the ones usually obtained by ChIP-seq, but there are several important differences: (1) Whereas ChIP data relate to a two-state situation of "bound" or "unbound" regions, DNase I acts as a generic indicator of chromatin state, and allows the handling of multiple states of chromatin accessibility (Shu et al., 2011); (2) ChIP-seq analyzers can employ two strand-specific approaches for peak detection: tag shifting or tag extension. Both strategies will hide the actual location of protein-DNA binding within a DHS; and (3) ChIP-seq peaks for a transcription factor are usually well-defined and can be identified by visual inspection, whereas DHSs are less evident due to tag enrichment over wide stretches of genomic sequence.

It is important to stress that there are two influencing factors that can change the DNase pattern: (1) How accessible is the region, determined by the fold-enrichment of the DHS and (2) How protected is the sequence where a transcription factor is binding (depth of the footprint). Therefore, the utilization of a ChIP-seq peak finder does not completely fit the patterns formed in a DNase-seq assay. However, due to the lack of well-established algorithms to handle DNase-seq data, popular ChIP-seq peak finders are used instead to pinpoint DHSs (Zhang et al., 2008; He et al., 2012). Among those peak callers, only F-seq (Boyle et al., 2008b) considers an algorithm adjustment specially dedicated to identify DHSs in its kernel density estimation approach, concerning the average fragment size of the experiment. The DHSs reported by this program have helped to integrate and interrelate data among several platforms (Shu et al., 2011; Song et al., 2011), for instance aiding the correlation *in vivo* of footprints with ChIP-seq enrichment (Boyle et al., 2011). F-seq has been also used to identify enriched sites in formaldehyde-assisted isolation of regulatory elements followed by sequencing (FAIRE-seq), but without any available statistical assessment concerning false discovery rate (FDR) or *p*-value calculation, with the DHSs obtained under different qualitative cutoffs, depending on a user-defined standard deviation threshold over the average signal with respect to a local background (Gaulton et al., 2010). As a consequence, the F-seq users need to employ time and effort on designing a proper statistical test for their experiment (Zhang et al., 2012).

Opposed to the F-seq approach, in Baek et al. (2012), the read extension to the average fragment size of the experiment is recommended. In this method, mappability-adjusted z-scores enable one to obtain statistical significance for the list of DHSs reported, to our knowledge being the first DNase-seq algorithm reporting FDR values for a list of DHSs. However, these two methodologies do not allow the inclusion of control samples in the analysis which, as has been demonstrated for ChIP-seq, can potentially reduce the FDR. Therefore, new statistical algorithms should be developed to exploit the potential of DNase-seq data more efficiently than, as is the current approach, peak callers developed originally for the analysis of ChIP-seq datasets (some of such tools are reviewed in Wilbanks and Facciotti, 2010).

FOOTPRINT DETECTION

At a very high sequencing depth it is possible to identify depleted narrow regions in the DHSs core, corresponding to protein footprints, ranging typically from 8 to 30 bp. Both the kernel density estimation approach (Boyle et al., 2008b) and the hotspot detection algorithm (Baek et al., 2012) will smooth the tag density profile and report the location of DHS peaks, making difficult the visualization and detection of confined depleted regions protected against DNase I cleavage. This problem can be solved by using DNase I cuts (read-start sites) for coverage determination instead of full-size or extended aligned reads. Thus, additional software to identify protein-DNA footprints is much needed. With this in mind, Hesselberth et al. (2009), presented a computational algorithm able to detect substantial DNase I cleavage reduction in the tag density compared to its adjacent flanking regions at nucleotide resolution. After computing depletion scores, non-overlapping footprints within intergenic regions have been reported. However, this method does not scale well for large genomes (Baek et al., 2012). Its modified version was introduced by Chen et al. (2010) who used a generalization of hidden Markov models and Bayesian networks, and considered non-uniquely mappable regions as missing data. This improved the precision of their previous approach in terms of FDR.

With sufficiently deep sequencing, the so-called "digital genomic footprinting" technique can reveal single protein-binding events (Hesselberth et al., 2009). Unlike ChIP-seq, which is specific for the protein under study, footprints identify narrow DNA regions that can be bound by any factor (Hager, 2009), showing significant enrichment for known motifs upstream of the transcription start sites (TSSs).

DIFFERENTIAL DNase I HYPERSENSITIVITY

With the popularization and drastic cost decrease of sequencing leading to the generation of multiple sequenced samples, quantitative analysis of differential ChIP-seq binding across conditions, time stages or different tissues has been the subject of a great amount of research in the last 2 years (Bardet et al., 2011; Liang and Keles, 2012). However, adapting DNase-seq data singularities for differential analysis has just begun to be approached (He et al., 2012). The only proposed methodology computes scores of stimulus-dependent DHS changes, proving the utility of quantitative measures of chromatin accessibility differences between conditions to predict transcription factor binding. Coupling information of known motifs found within the DHS can improve the prediction, and using instead the changes in DHS (ΔDHS), produces the best prediction. These results are coherent with the theory that the interaction between a specific sequence and a transcription factor may be guided by different types of chromatin configuration (van Steensel, 2011).

DATA INTEGRATION

Apart from the usual structural annotation and downstream analysis (including enrichment of known motifs or *de novo* motif discovery, with the canonical motif placed typically in the peak of a DHS) of the regions of interest, both for footprints or DHSs, the combination with other genomic data sources can unravel a plethora of novel biological insights. DHSs have positive correlation with active histone marks, whereas the correlation is negative for repressive histone marks, and DHSs score is higher for active genes than for silent ones (Shu et al., 2011). Furthermore, it has been shown recently that DNase-seq data, aided by regulatory genome sequences, can predict gene expression in a cell-type specific fashion (Natarajan et al., 2012). The utilization of prior knowledge can group the footprints or DHSs into more biologically meaningful target clusters, allowing a better understanding of how chromatin accessibility affects TF-DNA interaction. Although, the spatial distribution of DHSs/footprints is highly informative about binding, no one data source is fully enlightening when taken alone. For example, Centipede (Pique-Regi et al., 2011) improves TF-binding prediction by scanning the genome in search for known motifs or positional weight matrices, and integrating evolutionary sequence conservation, proximity to the nearest TSS, DNase I cuts, and histone modifications data into a Bayesian mixture model. However, not all factors influence the model in the same way: histone marks do not significantly improve the predictive power of DNase I accessibility. Centipede also shows the potential to extract quantitative measures of TF-binding from DNase-seq data. The disadvantage of Centipede is the compulsory requirement to know *a priori* the consensus sequence (motif) for each TF, which makes DNase-seq, if we consider the current state-of-the-art, a complementary tool of ChIP-seq rather than an independent assay to determine TF-binding sites genome-wide.

The correlation between gene expression and active and repressive histone marks have revealed four distinct modes of chromatin structure in humans, further invalidating the simplistic assumption that chromatin can only be in an "open" or "closed" conformation (Shu et al., 2011). Additionally, a cross-validated set of DNase-seq and FAIRE-seq sites allowed the creation of high-confidence open chromatin maps (Song et al., 2011).

From this year, a manually curated web-server storing DNase-seq and ChIP-Seq data from human and mouse studies is publicly available (Qin et al., 2012).

New open questions should redirect the efforts to adapt each methodology to fruitfully map chromatin accessibility by DNase-seq, from the former stages of getting significant broad DNase I hypersensitive regions or narrow footprints, to the latter steps that include the differential assessment of chromatin accessibility changes and the correlation with other available genomic data. The question whether DNase-seq will eventually serve as a substitute for ChIP-seq, and to what extent, will be unraveled in the upcoming years.

ACKNOWLEDGMENTS

Pedro Madrigal is supported by the EU Marie Curie Initial Training Network SYSFLO (agreement number 237909).

REFERENCES

Baek, S., Sung, M. H., and Hager, G. L. (2012). "Quantitative analysis of genome-wide chromatin remodeling," in *Chromatin Remodeling: Methods and Protocols*, ed R. H. Morse (New York, NY: Springer Science+Business Media), 433–441.

Bardet, A. F., He, Q., Zeitlinger, J., and Stark, A. (2011). A computational pipeline for comparative ChIP-seq analyses. *Nat. Protoc.* 7, 45–61.

Bell, O., Tiwari, V. K., Thomä, N. H., and Schübeler, D. (2011). Determinants and dynamics of genome accessibility. *Nat. Rev. Genet.* 12, 554–564.

Boyle, A. P., Davis, S., Shulha, H. P., Meltzer, P., Margulies, E. H., Weng, Z., et al. (2008a). High-resolution mapping and characterization of open chromatin across the genome. *Cell* 132, 311–322.

Boyle, A. P., Guinney, J., Crawford, G. E., and Furey, T. S. (2008b). F-Seq: a feature density estimator for high-throughput sequence tags. *Bioinformatics* 24, 2537–2538.

Boyle, A. P., Song, L., Lee, B. K., London, D., Keefe, D., Birney, E., et al. (2011). High-resolution genome-wide *in vivo* footprinting of diverse transcription factors in human cells. *Genome Res.* 21, 456–464.

Chen, X., Hoffman, M. M., Bilmes, J. A., Hesselberth, J. R., and Noble, W. S. (2010). A dynamic Bayesian network for identifying protein-binding footprints from single molecule-based sequencing data. *Bioinformatics* 26, i334–i342.

Cockerill, P. N. (2011). Structure and function of active chromatin and DNase I hypersensitive sites. *FEBS J.* 278, 2182–2210.

Gaulton, K. J., Nammo, T., Pasquali, L., Simon, J. M., Giresi, P. G., Fogarty, M. P., et al. (2010). A map of open chromatin in human pancreatic islets. *Nat. Genet.* 42, 255–259.

Hager, G. (2009). Footprints by deep sequencing. *Nat. Methods* 6, 254–255.

He, H. H., Meyer, C. A., Chen, M. W., Jordan, V. C., Brown, M., and Liu, X. S. (2012). Differential DNase I hypersensitivity reveals factor-dependent chromatin dynamics. *Genome Res.* 22, 1015–1025.

Hesselberth, J. R., Chen, X., Zhang, Z., Sabo, P. J., Sandstrom, R., Reynolds, A. P., et al. (2009). Global mapping of protein-DNA interactions *in vivo* by digital genomic footprinting. *Nat. Methods* 6, 283–289.

Kodama, Y., Nagaya, S., Shinmyo, A., and Kato, K. (2007). Mapping and characterization of DNase I hypersensitive sites in *Arabidopsis* chromatin. *Plant Cell Physiol.* 48, 459–470.

Liang, K., and Keles, S. (2012). Detecting differential binding of transcription factors with ChIP-seq. *Bioinformatics* 28, 121–122.

Natarajan, A., Yardimci, G. G., Sheffield, N. C., Crawford, G. E., and Ohler, U. (2012). Predicting cell-type-specific gene expression from regions of open chromatin. *Genome Res.* 22, 1711–1722.

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 Res.* 21, 447–455.

Qin, B., Zhou, M., Ge, Y., Taing, L., Liu, T., Wang, Q., et al. (2012). CistromeMap: a knowledgebase and web server for ChIP-Seq and DNase-Seq studies in mouse and human. *Bioinformatics* 28, 1411–1412.

Shu, W., Chen, H., Bo, X., and Wang, S. (2011). Genome-wide analysis of the relationships between DNaseI HS, histone modifications and gene expression reveals distinct modes of chromatin domains. *Nucleic Acids Res.* 39, 7428–7443.

Song, L., and Crawford, G. E. (2010). DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harb. Protoc.* 2010, pdb.prot5384.

Song, L., Zhang, Z., Grasfeder, L. L., Boyle, A. P., Giresi, P. G., Lee, B. K., et al. (2011). Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity. *Genome Res.* 21, 1757–1767.

van Steensel, B. (2011). Chromatin: constructing the big picture. *EMBO J.* 30, 1885–1895.

Wilbanks, E. G., and Facciotti, M. T. (2010). Evaluation of algorithm performance in ChIP-seq peak detection. *PLoS ONE* 5:e11471. doi: 10.1371/journal.pone.0011471

Wu, C., Bingham, P. M., Livak, K. J., Holmgren, R., and Elgin, S. C. (1979). The chromatin structure of specific genes: I. Evidence for higher order domains of defined DNA sequence. *Cell* 16, 797–806.

Zhang, W., Zhang, T., Wu, W., and Jiang, J. (2012). Genome-wide identification of regulatory DNA elements and protein-binding footprints using signatures of open chromatin in Arabidopsis. *Plant Cell* 24, 2719–2731.

Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Berstein, B. E., et al. (2008). Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* 9, R137.

Received: 28 July 2012; accepted: 13 October 2012; published online: 31 October 2012.

Citation: Madrigal P and Krajewski P (2012) Current bioinformatic approaches to identify DNase I hypersensitive sites and genomic footprints from DNase-seq data. Front. Gene. ***3****:230. doi: 10.3389/fgene.2012.00230*

This article was submitted to Frontiers in Bioinformatics and Computational Biology, a specialty of Frontiers in Genetics.