BIOINFORMATICS APPLICATIONS NOTE

Vol. 28 no. 5 2012, pages 717–718
doi:10.1093/bioinformatics/bts007

Genome analysis

Advance Access publication January 11, 2012

CHROMATRA: a Galaxy tool for visualizing genome-wide chromatin signatures

Thomas Hentrich[1,2], Julia M. Schulze[2], Eldon Emberly[3,*] and Michael S. Kobor[2,*]

[1]Department of Computing Science, University of British Columbia, Vancouver, BC V6T 1Z4, [2]Department of Medical Genetics, Centre for Molecular Medicine and Therapeutics, Child and Family Research Institute, Vancouver, BC V5Z 4H4 and [3]Department of Physics, Simon Fraser University, Burnaby, BC V5A 1S6, Canada

Associate Editor: Alfonso Valencia

ABSTRACT

Summary: CHROMATRA (CHROmatin Mapping Across TRAnscripts) is a visualization tool available as plug-in for the Galaxy platform. It allows detailed yet concise presentations of data derived from ChIP-chip or ChIP-seq experiments by visualizing enrichment scores across genes or other genomic features while accounting for their length and additional characteristics such as gene expression. It integrates into typical analysis workflows and enables rapid graphical assessment and comparison of genome-wide data at a glance.

Availability: https://github.com/cmmt/chromatra

Contact: msk@cmmt.ubc.ca; eemberly@sfu.ca

Received on September 22, 2011; revised on November 30, 2011; accepted on January 3, 2012

1 INTRODUCTION

The pace of advancements in DNA microarray (Stoughton, 2005) and sequencing technology (Shendure and Ji, 2008) continues to transform biological research. Both techniques complement each other and provide powerful platforms to interrogate entire genomes at high resolution, ranging from sequence decoding, to gene expression profiling, to chromatin modification mapping (Collas, 2010; Kapranov *et al.*, 2007; Metzker, 2010).

It is in large part due to these technologies that data are becoming available to scientists and allow questions to be asked of unprecedented depth. At the same time, however, the increasingly large volume and complexity of datasets demands adequate computational methods to analyze and present the data effectively.

While raw numbers may contain all relevant information of an analysis, they are not always necessarily suitable to convey biological meaning efficiently, especially when presenting data in scientific publications. Graphic visualizations of data can aid in representing complex content and reveal salient information at a glance (Gehlenborg *et al.*, 2010). Yet, satisfying this claim is particularly challenging for genome-wide data. For instance, when mapping chromatin modifications or transcription factors using ChIP-chip or ChIP-seq, it is a major goal to assess their localization and distribution across all genes of an organism while preserving structural and functional features of their respective genome.

Current approaches for visualizing binding or modification profiles can be broadly distinguished into two categories: first, tools of single-gene resolution, like the Genome Browser (Fujita *et al.*, 2011), offer the most detailed level of data presentation, but make it difficult to evaluate profiles across genes and derive broader principles as they disperse the information over multiple plots to cover the genome. Second, averaging approaches try to circumvent this problem by grouping genes into classes, e.g. based on length, Gene Ontology terms (Ashburner *et al.*, 2000) or transcription rate (Mayer *et al.*, 2010; Pokholok *et al.*, 2005) and calculating average profiles for each class. While they succeed in condensing the mosaic representation of single-gene visualizations, they inevitably sacrifice detail dependent on the applied grouping given the heterogeneity of genes with respect to their length and other features.

Here, we present CHROMATRA (CHROmatin Mapping Across TRAnscripts), a visualization tool for genome-wide DNA-protein binding and chromatin modification maps that balances detail and compactness of the existing visualization approaches. Developed as a plug-in for the Galaxy analysis environment (Goecks *et al.*, 2010), CHROMATRA portrays enrichment profiles in a comprehensive yet condensed form by accounting for feature length and additional characteristics such as gene expression. CHROMATRA plots reveal the spatial distribution of chromatin marks or binding events across all genomic features of interest, and may aid in discovering unexpected correlations and patterns in enrichment profiles.

2 FUNCTIONALITY AND WORKFLOW

CHROMATRA is a plug-in for the Galaxy bioinformatics platform (Goecks *et al.*, 2010). As one of the most comprehensive analysis solutions for genomics data, Galaxy integrates different software tools under a unified user interface with advanced workflow management capabilities (Blankenberg *et al.*, 2007). Since CHROMATRA was written in Python it easily integrates with Galaxy, itself implemented in Python, and has no dependencies on external libraries not already used by Galaxy.

CHROMATRA consists of two visualization modules, which can be integrated into the analysis workflow for ChIP-chip or ChIP-seq experiments. Typically, the raw data of ChIP experiments are normalized and enrichment scores calculated and stored in standardized formats such as WIG, BED and GFF for subsequent analysis. After these steps, CHROMATRA can be linked to the workflow and used to visualize ChIP profiles by reading GFF files that contain enrichment scores for the whole genome.

The first module, CHROMATRA-L, accounts for differences in lengths of genomic features, such as transcripts, exons or telomeric

*To whom correspondence should be addressed.

© The Author 2012. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

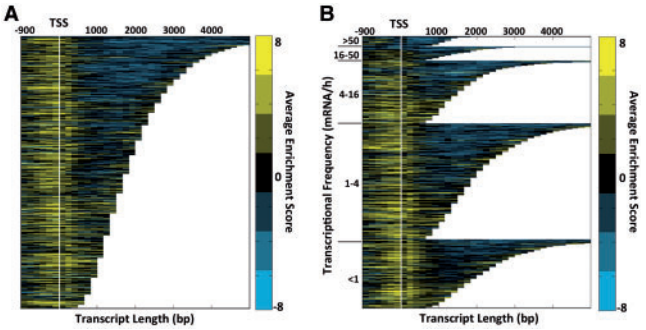

Fig. 1. CHROMATRA plot based on ChIP-chip profile of histone variant H2A.Z in *S.cerevisiae* (data from Wang *et al.*, 2009). (**A**) CHROMATRA-L plot for H2A.Z across transcripts in yeast, sorted by length and aligned by TSS, showing that H2A.Z is enriched primarily in promoters but absent in transcript bodies. Normalized enrichment scores were binned into segments of 150 bp. (**B**) CHROMATRA-T plot for same dataset as in (A), showing that H2A.Z preferentially marks infrequently transcribed genes and is absent in transcripts with high transcription rate. Transcripts sorted by length and transcription rate range (Holstege *et al.*, 1998).

elements, and eliminates potential biases in assessing their ChIP profiles that usually arise when using non-absolute length scales. It does so by sorting the features by length and calculating mean enrichment values for bins of absolute size according to the user-specified settings, for example at nucleosome resolution. Therefore, the module requires coordinates of genomic features aligned to a common position, for example coordinates of genes aligned to their transcription start site (TSS). These coordinates can be uploaded by the user or readily derived from other Galaxy modules in a tab-delimited format. The resulting enrichment profiles are color-coded according to user-defined settings and visualized in a heatmap-like plot, which is available for download in different image formats, such as PNG, PDF or SVG. Figure 1 shows a representative example, mapping histone variant H2A.Z profiles derived from ChIP-chip experiments across ∼4500 transcripts of *Saccharomyces cerevisiae* with known TSS and transcription rate.

The second module, CHROMATRA-T, extends the first one by allowing an additional metric, such as transcription rate (Fig. 1B) or a pre-computed numerical classification scheme, e.g. based on their Gene Ontology term, to be accounted for when visualizing enrichment profiles of genomic features. According to user-defined intervals/classes for the second metric, CHROMATRA-T first partitions the set of features accordingly and then sorts each group by length. Enrichment values are subsequently calculated and displayed as described above. CHROMATRA-T hence allows assessing ChIP enrichment profiles at a glance according to various characteristics of genomic features and avoids length-dependent biases.

3 CONCLUSION AND OUTLOOK

CHROMATRA implements two visualization approaches for exploring genome-wide chromatin signatures and facilitating the discovery of unexpected patterns and correlations. The comprehensive yet compact visualization allows assessing enrichment profiles for data from any organism. The CHROMATRA modules have proven their effectiveness in a recent study, addressing the localization and dependencies of chromatin modifications in the context of transcription (Schulze *et al.*, 2011). Besides assessing genes or transcripts, any other genomic feature that can be aligned according to a common position can be visualized, including introns and alternative transcripts. Furthermore, data derived through related techniques, such as MeDIP-seq (Jacinto *et al.*, 2008; Weber *et al.*, 2005) or MBD-seq (Serre *et al.*, 2010), can be used as input as well. As a plug-in for Galaxy, CHROMATRA easily integrates into existing workflows and enables direct graphical exploration of genome-wide data.

ACKNOWLEDGEMENTS

We thank Dr Arvind Gupta (UBC) for support and helpful discussions during the development of this project. T.H. is a fellow of the CIHR/MSFHR Bioinformatics Training Program.

Funding: Canadian Institute of Health Research (MOP-79442).

Conflict of Interest: none declared.

REFERENCES

Ashburner,M. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.

Blankenberg,D. *et al.* (2007) A framework for collaborative analysis of ENCODE data: making large-scale analyses biologist-friendly. *Genome Res.*, **17**, 960–964.

Collas,P. (2010) The current state of chromatin immunoprecipitation. *Mol. Biotechnol.*, **45**, 87–100.

Fujita,P.A. *et al.* (2011) The UCSC Genome Browser database: up-date 2011. *Nucleic Acids Res.*, **39**, D876–D882.

Gehlenborg,N. *et al.* (2010) Visualization of omics data for systems biology. *Nat. Methods*, **7**, S56–S68.

Goecks,J. *et al.* (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.*, **11**, R86.

Holstege,F. *et al.* (1998) Dissecting the regulatory circuitry of a eukaryotic genome. *Cell*, **95**, 717–728.

Jacinto,F.V. *et al.* (2008) Methyl-DNA immunoprecipitation (MeDIP): hunting down the DNA methylome. *BioTechniques*, **44**, 35, 37, 39 passim.

Kapranov,P. *et al.* (2007) Genome-wide transcription and the implications for genomic organization. *Nat. Rev. Genet.*, **8**, 413–423.

Mayer,A. *et al.* (2010) Uniform transitions of the general RNA polymerase II transcription complex. *Nat. Struct. Mol. Biol.*, **17**, 1272–1278.

Metzker,M.L. (2010) Sequencing technologies - the next generation. *Nat. Rev. Genet.*, **11**, 31–46.

Pokholok,D. *et al.* (2005) Genome-wide map of nucleosome acetylation and methylation in yeast. *Cell*, **122**, 517–527.

Serre,D. *et al.* (2010) MBD-isolated Genome Sequencing provides a high-throughput and comprehensive survey of DNA methylation in the human genome. *Nucleic Acids Res.*, **38**, 391–399.

Schulze,J.M. *et al.* (2011) Splitting the task: Ubp8 and Ubp10 deubiquitinate different cellular pools of H2BK123. *Genes Dev.*, **25**, 2242–2247.

Shendure,J. and Ji,H. (2008) Next-generation DNA sequencing. *Nat. Biotechnol.*, **26**, 1135–1145.

Stoughton,R.B. (2005) Applications of DNA microarrays in biology. *Annu. Rev. Biochem.*, **74**, 53–82.

Wang,A.Y. *et al.* (2009) Asf1-like structure of the conserved Yaf9 YEATS domain and role in H2A.Z deposition and acetylation. *Proc. Natl Acad. Sci.*, **106**, 21573–21578.

Weber,M. *et al.* (2005) Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. *Nat. Genet.*, **37**, 853–862.

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016