BIOINFORMATICS APPLICATIONS NOTE Vol. 30 no. 11 2014, pages 1620–1622
doi:10.1093/bioinformatics/btu082

Genome analysis

Advance Access publication February 7, 2014

HiBrowse: multi-purpose statistical analysis of genome-wide chromatin 3D organization

Jonas Paulsen[1,*], Geir Kjetil Sandve[2], Sveinung Gundersen[3], Tonje G. Lien[4], Kai Trengereid[5] and Eivind Hovig[1,2,3,*]

[1]Institute for Cancer Genetics and Informatics, Oslo University Hospital, PO Box 4950, Nydalen, 0424 Oslo, [2]Department of Informatics, University of Oslo, Problemveien 7, 0313 Oslo, [3]Department of Tumor Biology, Institute for Cancer Research, Oslo University Hospital, PO Box 4950, Nydalen, 0424 Oslo, [4]Department of Mathematics, University of Oslo, Problemveien 7, 0313 Oslo and [5]ELIXIR project, Department of Informatics, University of Oslo, Problemveien 7, 0313 Oslo, Norway

Associate Editor: Michael Brudno

ABSTRACT

Summary: Recently developed methods that couple next-generation sequencing with chromosome conformation capture-based techniques, such as Hi-C and ChIA-PET, allow for characterization of genome-wide chromatin 3D structure. Understanding the organization of chromatin in three dimensions is a crucial next step in the unraveling of global gene regulation, and methods for analyzing such data are needed. We have developed HiBrowse, a user-friendly web-tool consisting of a range of hypothesis-based and descriptive statistics, using realistic assumptions in null-models.

Availability and implementation: HiBrowse is supported by all major browsers, and is freely available at http://hyperbrowser.uio.no/3d. Software is implemented in Python, and source code is available for download by following instructions on the main site.

Contact: jonaspau@ifi.uio.no

Supplementary Information: Supplementary data are available at *Bioinformatics* online.

Received on October 18, 2013; revised on January 17, 2014; accepted on February 3, 2014

1 INTRODUCTION

Methods for detection of genome-wide chromatin 3D conformation, such as Hi-C (Lieberman-Aiden *et al.*, 2009) and ChIA-PET (Fullwood *et al.*, 2009), are drastically expanding our understanding of genome biology. However, statistical and computational methods to analyze chromatin conformation capture-based data are needed. Many of the available methods focus on data visualization, or are not suited for genome-wide statistical investigations (Baù *et al.*, 2010; Servant *et al.*, 2012; Thongjuea *et al.*, 2013; Zhou *et al.*, 2013). The structure of chromatin makes statistical analysis complicated, due to correlations between the interaction frequencies caused by both sequence-dependent and topological constraints (Paulsen *et al.*, 2013). A few statistical tests have been proposed, with varying possibilities to account for structural dependencies (Botta *et al.*, 2010; Kruse *et al.*, 2013; Paulsen *et al.*, 2013; Wang *et al.*, 2013; Witten and Noble, 2012). Two useful command-line tools are the hiclib-package (Imakaev *et al.*, 2012), and the HOMER software suit (Heinz *et al.*, 2010), which both allow for noise-removal, outlier detection and compartment identification. The HOMER software additionally allows for identification of significant interactions in a given dataset, assuming a binomial distribution and a background model taking into account sequence-based and compartmental biases.

The global nature of these data allow for other types of statistical investigations beyond detecting significance of individual interactions. A common type of analysis is to analyze a set of genomic elements (genes, regulatory elements, transcription factors, etc.), and ask how this subset, or 'query track', is spatially arranged in 3D space as represented by a Hi-C dataset, for example. Here we present HiBrowse, a web-based analysis server for performing statistical analysis of 3D genomes in a range of different settings. The available statistics provide a flexible and expandable catalog of tools based on state-of-the-art statistical methods utilizing Monte Carlo (MC) and analytic methods as suited, in addition to a range of tools for visualization and hypothesis-generating investigations.

2 FEATURES AND METHODS

2.1 Data representation and analysis framework

We build on general software components of the Genomic HyperBrowser (Sandve *et al.*, 2010, 2013), a web-based analysis server for genome-scale data. The graphical user interface (GUI) is based on Galaxy (Goecks *et al.*, 2010), a user-friendly point-and-click environment familiar to many researchers. All tracks are based on a representation of elements as mathematical objects, consisting of points, segments, functions and variants of these [see Gundersen *et al.* (2011) for an in-depth discussion]. Any given analysis can be performed on all chromosomes, specific chromosomes or selected sub-parts of chromosomes, depending on the needs.

In practice, an analysis is initiated by selecting one or more tracks either from the HyperBrowser repository, or from the user history. At least one of the selected tracks must be a Hi-C (3D) track, and the accompanying selected tracks (called 'query tracks') determine the types of statistical analyses that are possible, and therefore selectable in the system.

A range of publicly available 3D-datasets have been installed in the repository. Since it has been shown that Hi-C and similar data can contain systematic biases, all the available Hi-C datasets have been corrected

*To whom correspondence should be addressed.

for such biases using the method of Imakaev *et al.* (2012). Furthermore, a specialized tool has been developed to allow users to upload their own Hi-C data (or similar) into the history, even if the dataset itself does not conform to well-known formats. See Supplementary Table S1 for a list of already installed and pre-processed Hi-C datasets.

2.2 Overview of statistical methods

Statistical tools are divided into two broad categories: hypothesis tests and descriptive statistics. Hypothesis tests are both MC based and analytical. Due to the complex structure of chromatin conformation capture data, finding suited explicit null distributions is generally not possible (Paulsen *et al.*, 2013; Witten and Noble, 2012), and even randomization of the data through MC is difficult. Therefore, we consistently perform permutations on the query track only. The hypothesis tests can be divided into three types, defined by the query track type, as illustrated in Figure 1A. For example, Points (P) are used to analyze general (all-versus-all) 3D co-localization by specifying a set of genomic elements using the BED format, while Linked Points (LP) are used to analyze 3D co-localization between selected pairs of elements by providing additional information about which genomic elements that should be linked together.

In the most basic case, if the user selects a set of points (genomic elements) in BED-format in addition to a Hi-C data track, one may ask whether all the genomic elements in the BED-file are more/less co-localized in 3D, in an all-versus-all fashion, than what would be expected by chance. In this case, the mean of the observed standardized interaction frequencies is compared to the expected value estimated from the permuted positions in representative regions of the rest of the Hi-C (3D) track. This analysis was introduced in Paulsen *et al.* (2013), and in this article we expand the methodologies by allowing a much wider variety of query tracks. For example, by specifying two point-tracks (two BED files), in addition to a Hi-C (or similar) track, the user can ask whether the points in track 1 are more/less co-localized with track 2, than expected by chance. In this type of statistical question, the permutations can be performed on both of the point-tracks, or by preserving one of the point-tracks completely.

It is also possible to specify particular interactions between a set of genomic elements, and compare these interactions with randomly permuted interactions within the same set of elements. In HiBrowse, interactions between genomic elements are defined using LP, a format described in detail elsewhere (Gundersen *et al.*, 2011). Such linked track types can easily be created by using a dedicated tool that converts from a simple BED file format containing information about which elements that should be linked together (see Supplementary Fig. S1, for an example). Since this type of analysis only permutes interactions intrinsically with regards to the query track, the positions of all elements will be completely preserved. This type of analysis should be used whenever specific interactions between genomic elements are considered, and it would be natural to compare with random links between the same elements. Since regions of the genome can have varying properties (active/inactive genes, open/closed chromatin, etc.), global shuffling of links between all selected elements is not always preferable. To take such properties into account during the permutation, each of the points can be marked by a value, such that the link-permutations will be performed by preserving the value-combinations on both sides of the links.

If the user wants full control over exactly what pairs of interactions that are allowed to take part in the link-permutations, it is possible to specify a case/control value on each of the links via a dedicated tool which accepts two BED files ('case' and 'control') of the same format as described above (see Supplementary Fig. S2, for an example). The case/control-linked elements can then be selected together with a Hi-C (3D) track, allowing the user to compare the interaction frequency of all the links marked as 'case' with the expected interaction frequency given by permuting the case/control labels. This type of statistic is optimal for data that is only sampled from a pre-defined set of elements of the genome, and where the user wants to find out whether a subset of these elements are co-localized in 3D.

Finally, it is possible to find statistically significant differences between two Hi-C datasets, for example comparing treatments [as e.g. in

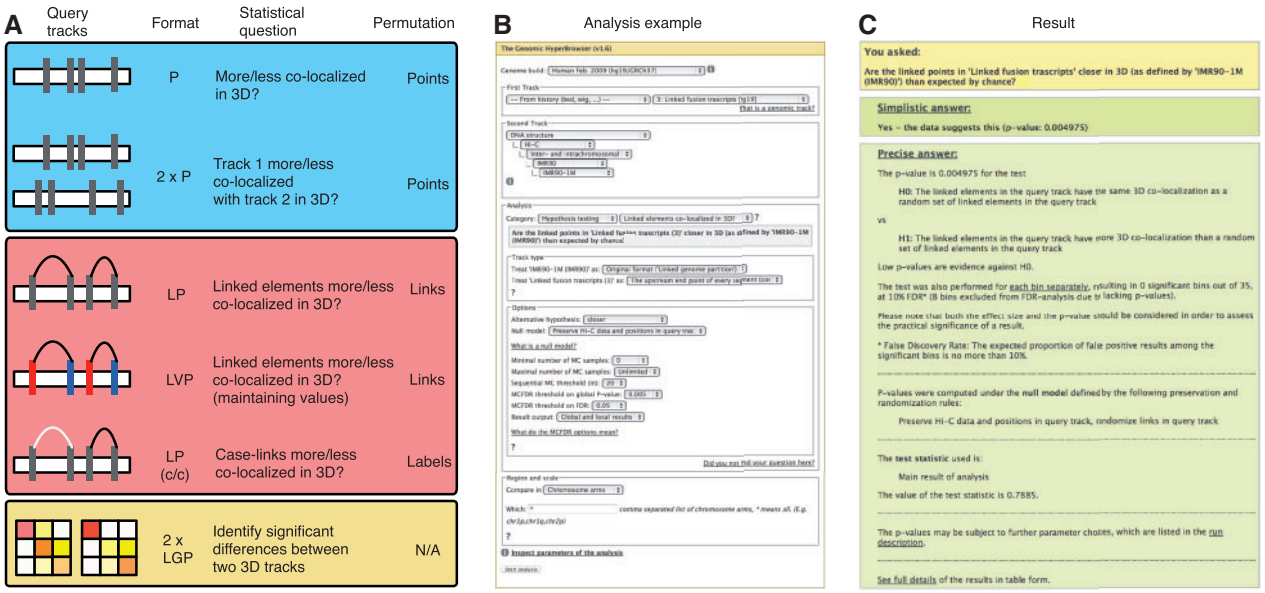

Fig. 1. (**A**) Overview of statistical hypothesis tests implemented in HiBrowse. See Gundersen *et al.* (2011) for an in-depth explanation of track types, and the Supplementary Material for details about each statistic. (**B**) Example of a HiBrowse analysis using the 'Linked elements more/less co-localized in 3D?' statistic, investigating whether fusion transcripts are co-localized in 3D. (**C**) Result page from the analysis, presenting the question asked by the user together with both a simplistic and a more detailed answer giving the *P*-value and model assumption details. Links are provided to full details of the results at individual chromosome regions

Rickman *et al*. (2012)]. The statistical test implemented for this type of analysis is based on the edgeR-tool (Robinson *et al*., 2010). Details about the mathematical formulation of the different types of statistics and their corresponding null-hypotheses are found in the Supplementary Material.

In addition to hypothesis tests, a range of descriptive statistics have been implemented. For example, each hypothesis test is accompanied by an enrichment score, giving the degree of over/under-representation of 3D co-localization, compared to the expected 3D co-localization (see Supplementary Material for details). Other types of available descriptive statistics are visualization of clustered Hi-C matrices as heatmaps or graphs, principal component analysis on Hi-C matrices and other summary statistics (see Supplementary Table S2 for a comprehensive list). All available analyses are described thoroughly on the help pages linked from the main site, where example histories are provided such that users can explore each statistic in detail. Demo-buttons are provided for all tools, giving small example runs. See Figure 1B and C for an analysis example.

Funding: This work was supported by the Norwegian Cancer Society [PR-2006-0433].

Conflict of Interest: none declared.

REFERENCES

Baù,D. *et al*. (2010) The three-dimensional folding of the α-globin gene domain reveals formation of chromatin globules. *Nat. Struct. Mol. Biol.*, **18**, 107–114.

Botta,M. *et al*. (2010) Intra- and inter-chromosomal interactions correlate with CTCF binding genome wide. *Mol. Syst. Biol*, **6**.

Fullwood,M.J. *et al*. (2009) An oestrogen-receptor-α-bound human chromatin interactome. *Nature*, **462**, 58–64.

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.

Gundersen,S. *et al*. (2011) Identifying elemental genomic track types and representing them uniformly. *BMC Bioinformatics*, **12**, 494.

Heinz,S. *et al*. (2010) Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell*, **38**, 576–589.

Imakaev,M. *et al*. (2012) Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nat. Methods*, **9**, 999–1003.

Kruse,K. *et al*. (2013) A complex network framework for unbiased statistical analyses of DNA–DNA contact maps. *Nucleic Acids Res.*, **41**, 701–710.

Lieberman-Aiden,E. *et al*. (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, **326**, 289–293.

Paulsen,J. *et al*. (2013) Handling realistic assumptions in hypothesis testing of 3D co-localization of genomic elements. *Nucleic Acids Res.*, **41**, 5164–5174.

Rickman,D.S. *et al*. (2012) Oncogene-mediated alterations in chromatin conformation. *Proc. Natl Acad. Sci. USA*, **109**, 9083–9088.

Robinson,M.D. *et al*. (2010) edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.

Sandve,G.K. *et al*. (2010) The Genomic HyperBrowser: inferential genomics at the sequence level. *Genome Biol.*, **11**, R121.

Sandve,G.K. *et al*. (2013) The Genomic HyperBrowser: an analysis web server for genome-scale data. *Nucleic Acids Res.*, **41**, W133–W141.

Servant,N. *et al*. (2012) HiTC: exploration of high-throughput 'C' experiments. *Bioinformatics*, **28**, 2843–2844.

Thongjuea,S. *et al*. (2013) r3Cseq: an R/Bioconductor package for the discovery of long-range genomic interactions from chromosome conformation capture and next-generation sequencing data. *Nucleic Acids Res.*, **41**, e132.

Wang,H. *et al*. (2013) Topological properties of chromosome conformation graphs reflect spatial proximities within chromatin. In: *Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics*. ACM, Washington, DC, USA, p. 306.

Witten,D.M. and Noble,W.S. (2012) On the assessment of statistical significance of three-dimensional colocalization of sets of genomic elements. *Nucleic Acids Res.*, **40**, 3849–3855.

Zhou,X. *et al*. (2013) Exploring long-range genome interactions using the WashU Epigenome Browser. *Nat. Methods*, **10**, 375–376.