BIOINFORMATICS APPLICATIONS NOTE

Vol. 30 no. 21 2014, pages 3120–3122
doi:10.1093/bioinformatics/btu491

Genome analysis

Advance Access publication July 23, 2014

FisHiCal: an R package for iterative FISH-based calibration of Hi-C data

Yoli Shavit[1,*], Fiona Kathryn Hamey[2] and Pietro Lio[1]

[1]Computer Laboratory, University of Cambridge, Cambridge CB3 0FD and [2]Cambridge Systems Biology Centre, University of Cambridge, Cambridge CB2 1GA, UK

Associate Editor: John Hancock

ABSTRACT

Summary: The fluorescence *in situ* hybridization (FISH) method has been providing valuable information on physical distances between loci (via image analysis) for several decades. Recently, high-throughput data on nearby chemical contacts between and within chromosomes became available with the Hi-C method. Here, we present FisHiCal, an R package for an iterative FISH-based Hi-C calibration that exploits in full the information coming from these methods. We describe here our calibration model and present 3D inference methods that we have developed for increasing its usability, namely, 3D reconstruction through local stress minimization and detection of spatial inconsistencies. We next confirm our calibration across three human cell lines and explain how the output of our methods could inform our model, defining an iterative calibration pipeline, with applications for quality assessment and meta-analysis.

Availability and implementation: FisHiCal v1.1 is available from http://cran.r-project.org/.

Contact: ys388@cam.ac.uk

Supplementary information: Supplementary Data is available at *Bioinformatics* online.

Received on April 16, 2014; revised on June 26, 2014; accepted on July 16, 2014

1 INTRODUCTION

With molecular biologists getting a closer look at the spatial organization of the nucleus, important methodologies currently used enlist Hi-C, FISH and electron microscopy (EM). In many ways, these methods offer complementary information. Hi-C data provide a genome-wide capture of chromatin contacts, with the advantages of high-throughput and scale; however, their spatial interpretation is complicated. FISH data, on the other hand, are usually limited in scale but offer a direct measure for physical distance (through image analysis). They can be obtained regardless of range or physical obstacles that affect the chemical capture of contacts (in Hi-C) between far away or inaccessible loci, and with a distinction between homologs, which is absent in Hi-C data.

Tanizawa *et al.* (2010) previously suggested an exponential model to convert yeast Hi-C frequencies to FISH distance approximations and used them to reconstruct a 3D model for the yeast genome. However, it was not clear whether the same model could be used for higher-order organisms with larger genomes, where Hi-C limitations are likely to be more prominent, and given that previous studies have suggested a power law model to relate physical distances and contact frequencies (Duan *et al.*, 2010; Fraser *et al.*, 2009). Whether the same parameters are appropriate across short, medium and long ranges, or for different chromosomes and cell types, also remained an open question. More recently, Trieu and Cheng (2014) used FISH data for parameterizing an objective function that will reconstruct the 3D configuration from human Hi-C data. Although innovative, this approach did not exploit in full the relationship between Hi-C and FISH data and lacked a model to study their discrepancies. Here we present FisHiCal, an R package for integrating Hi-C and FISH data, which offers a modular and easy-to-use tool for chromosomal spatial analysis. With FisHiCal, researchers can prepare and apply FISH-based Hi-C calibration, which converts contact frequencies into distances while taking into consideration range limitations in Hi-C data. To make our calibration especially useful, FisHiCal also includes 3D inference methods that we have developed, that can in turn iteratively refine the calibration model. We confirm our calibration across three human cell lines and show that our methods can provide valuable information for Hi-C/FISH calibration refinement, meta-analysis and quality assessment. To the best of our knowledge, FisHiCal is the only tool available for performing these tasks.

2 MATERIALS AND METHODS

Given a set D of pairwise FISH distances and a matching set C of Hi-C contact frequencies, we assume a power law model such that $C \sim \beta D^{\alpha}$. Taking the log of this model gives a linear dependency that could be solved with linear regression to estimate α and β. As long range frequencies are typically noisy and less reliable, we would like to consider only a subset of (shortest) matching distances and frequencies. We denote t_r, a reliability threshold that defines this subset, and solve instead for C_{t_r} and D_{t_r} the matching subsets of C and D, induced by t_r, respectively. After we have estimated the values of α and β, Hi-C frequencies could be converted into calibrated Hi-C distances, approximating FISH distances. Because of Hi-C limitations, the suggested model may be appropriate only up to a certain distance threshold t_n, above which corresponding contact frequencies should be discarded as non-informative or noisy (the Supplementary information gives the details of selecting t_r and t_n).

Calibrated Hi-C data can then provide input for 3D reconstruction. We denote $\delta_{i,j}$ as the calibrated distance between loci i and j, and $d_{i,j}(Y)$ as their Euclidian distance, in the true underlying 3D configuration Y. Here, a zero $\delta_{i,j,}$ for different loci i and j, represents missing information that was discarded in the calibration step. Our goal is then to minimize

[*] To whom correspondence should be addressed.

the following function, usually termed *stress* in a multidimensional scaling setting (Kruskal, 1964): $\sum_{i<j} w_{i,j}(\delta_{i,j} - d_{i,j}(Y)^2$, where $w_{i,j}$ are the weights we assign according to the reliability of $\delta_{i,j}$. As we mostly rely on local information, we can use here a local stress function (Chen and Buja, 2009), where missing $\delta_{i,j}$ are replaced with a constant d_{inf} (d_{inf} >> known $\delta_{i,j}$) and $w_{i,j}$ take the value of $1/d_{inf}$ for missing distances and 1 otherwise (for $d_{inf} \leq 1$, weights of missing distances should be set to a small constant <<1). As $w_{i,j}$ define an irreducible matrix, the stress minimization could be performed through Scaling by Majorizing a Complicated Function (SMACOF) (De Leeuw, 1977), a well-established strategy for this task, which guarantees convergence.

The calibrated Hi-C distances $\delta_{i,j}$ further define a weighted undirected graph $G\{V, E\}$, where V is the set of loci and E is the set of edges: $\{(v_i,v_j)| \delta_{i,j}>0,\ i!=\ j\}$ with weights $\delta_{i,j.}$ Here we distinguish between immediate neighbors from the same chromosome (*cis)* and from different chromosomes (*trans)* and detect a spatial inconsistency for a node *v* in *G*, if the subgraph *G'* of all (immediate) *trans* neighbors of *v* is not connected. Further identifying the connected components in *G'* can highlight the cause of inconsistency and the underlying spatial division. An inconsistency represents an event where several loci that are in an accessible range from another locus (in *trans*), are not in an accessible range themselves, which may occur, for example, owing to homology or noise.

The iterative calibration process starts with measuring the discrepancies between Hi-C and FISH data, which could aid in reproducibility assessment and point to functionally meaningful events (Supplementary information, Section 5). The output of the calibration is then used for 3D inference applications, as described above (and for other applications, Section 4). Finally, studying the resulting spatial models, for example, across chromosomes or different tissues, provides the input for the next iteration to further refine the calibration model.

3 SOFTWARE IMPLEMENTATION

FisHiCal v1.1 implements the methods described in Section 2. Users can first prepare and apply their calibration (Supplementary Table S1: prepareData, prepareCalib, calibrate) and use 3D inference methods (Supplementary Table S1: lsmacof, and –Inc functions) to spatially explore their data and further refine their calibration (Supplementary Table S1: updateCalib, getInfoLevelForChr). To make it user-friendly and accessible, FisHiCal further includes examples and comprehensive documentation. Finally, scalability and feasible running times are achieved through C++/R integration (Supplementary information, Section 6).

4 USE CASE AND APPLICATIONS

Hi-C data from human IMR90 fibroblasts (Dixon *et al.*, 2012), GM06990 lymphoblasts and K562 erythroleukemia (Lieberman-Aiden *et al.*, 2009) cell lines were pre-processed and corrected for noise and bias, as described by Shavit and Lio' (2014). The data were then matched with FisHiCal::prepareData, at a resolution of 1 Mb, to FISH distances from human primary fibroblasts (Mateos-Langerak *et al.*, 2009). Supplementary Figure S1 presents these data and the calibration curves that were estimated with FisHiCal::prepareCalib, confirming our model across the three cell lines. The best fit was achieved for FISH and Hi-C data from the same cell type (fibroblasts) with a consistent outlier at 1.49 µm, mapped to a *cis* distance in chromosome 1. To further explore the spatial basis of this outlier, we have reconstructed the 3D structure of chromosome 1 from the calibrated (fibroblasts) Hi-C matrix with FisHiCal::lsmacof. The predicted structure (Supplementary Figure S2) confirmed the known partition into two chromatin compartments (Lieberman-Aiden *et al.*, 2009) and suggested a possible explanation for the outlier. The corresponding two successive 1-Mb bins (encircled in Supplementary Figure S2b) formed part of a chromatin loop in our model. While Hi-C captured all intra-loop frequencies (leading to a relatively high value), FISH distances were likely to be measured between the far ends of the loop, as the corresponding probes were mapped to highly transcribed genes (Mateos-Langerak *et al.*, 2009). Further exploring the calibration curves of silenced and activated regions could confirm whether this example illustrates a new measure for looping events (see also Supplementary information, Section 8). Additional examination of the inconsistencies detected for loci in chromosome 1 with FisHiCal::searchInc, highlighted a long genomic domain (184–196 Mb), which was in contact with two loci in chromosomes 7 and 16 that were not themselves connected (Supplementary Figure S3). As previous findings position chromosome 1 homologs at different locations in the nucleus (Bolzer *et al.*, 2005), this inconsistency suggests that the loci in chromosome 16 and chromosome 7 are in contact with different instances of the genomic domain in chromosome 1, correspondingly. Additional experimental knowledge (e.g. from EM) could be used to further explore this finding. The use case described above illustrates an iterative FISH-based procedure for Hi-C calibration. Further applications of calibration include quality assessment and genome wide cytogenetic studies (Supplementary Figures S4 and S5). With more FISH and Hi-C data becoming available, building chromosomes maps with accurate scale and carrying time-series calibration analysis will be made possible with FisHiCal (Supplementary Figures S6 and S7).

5 CONCLUSION

Being the first tool to integrate FISH and Hi-C data, FisHiCal shows the importance of calibration for studying the nuclear architecture with different techniques, with applications spanning cytogenetics, differentiation and spatio-temporal studies.

ACKNOWLEDGEMENTS

We thank Dr Sandra Goetze for kindly providing us with FISH measurements. We thank Dr Sabrina Tosi and Dr Chris Town for helpful discussion on technical and clinical aspects of FISH analysis.

Funding: We thank FP7-Health-F5-2012, under grant agreement no 305280 (MIMOmics).

Conflict of Interest: none declared.

REFERENCES

Bolzer,A. *et al.* (2005) Three-dimensional maps of all chromosomes in human male fibroblast nuclei and prometaphase rosettes. *PLoS Biol.*, **3**, e157.

Chen,L. and Buja,A. (2009) Local multidimensional scaling for nonlinear dimension reduction, graph drawing, and proximity analysis. *J. Am. Stat. Assoc.*, **104**, 209–219.

De Leeuw,J. (1977) Applications of convex analysis to multidimensional scaling. In: Barra,J.R. *et al.* (ed.) *Recent Developments in Statistics*. North-Holland, Amsterdam, The Netherlands, pp. 133–145.

Dixon,J.R. *et al.* (2012) Topologoical domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, **485**, 376–380.

Duan,Z. *et al.* (2010) A three-dimensional model of the yeast genome. *Nature*, **465**, 363–367.

Fraser,J. *et al.* (2009) Chromatin conformation signatures of cellular differentiation. *Genome Biol.*, **10**, R37.

Kruskal,J.B. (1964) Nonmetric multidimensional scaling: a numerical method. *Psychometrika*, **29**, 115–129.

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

Mateos-Langerak,J. *et al.* (2009) Spatially confined folding of chromatin in the interphase nucleus. *Proc. Natl Acad. Sci. USA*, **106**, 3812–3817.

Shavit,Y. and Lio',P. (2014) Combining a wavelet change point and the Bayes factor for analysing chromosomal interaction data. *Mol. BioSyst.*, **10**, 1576–1585.

Tanizawa,H. *et al.* (2010) Mapping of long-range associations throughout the fission yeast genome reveals global genome organization linked to transcriptional regulation. *Nucleic Acids Res.*, **38**, 8164–8177.

Trieu,T. and Cheng,J. (2014) Large-scale reconstruction of 3D structures of human chromosomes from chromosomal contact data. *Nucleic Acids Res.*, **42**, e52.