Review of existing methods for Hi–C data analysis & some ideas …

H.L

April 11, 2014

After a literature review, I collected the following reports describing Hi–C data analysis methods:

1. Peng, C. et al. The sequencing bias relaxed characteristics of Hi-C derived data and implications for chromatin 3D modeling. Nucleic Acids Research 41, e183e183 (2013).

   Method: It uses a parameter ("sequencing-bias-relaxed") to deal with biases due to a) differences in sequencing depth, b) different chromatin regions within the same experiment.
   Pipeline used: AutoChrom3D Source code & 3D models: `http://ibi.hzau.edu.cn/3dmodel`
   Result: Method for automatic generation of chromatin 3D structure models. Takes into account the aforementioned biases. The authors admit that there is space for improvement.

2. Zhang, Z., Li, G., Toh, K.-C. & Sung, W.-K. 3D Chromosome Modeling with Semi-Definite Programming and Hi-C Data. Journal of Computational Biology 20, 831-846 (2013).

   Method:
   A deterministic method which uses "semi-definite programming techniques" to fit the observed data.
   Pipeline used: ChromSDE. It is based on RMSD ("root mean square deviation").
   Source code and instructions: `http://biogpu.ddns.comp.nus.edu.sg/~zzz/ChromSDE/`.
   Result:
   They used simulated and real Hi-C data and they claim that their method is more accurate than the existing ones (meaning ChromSDE

and BACH). hey used Spearman correlation to demonstrate the superiority of the method.

3. Ay, F., Bailey, T. L. & Noble, W. S. Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts. Genome Research (2014). doi:10.1101/gr.160374.113

   Method: Fit-Hi-C: Statistical confidence estimation that takes into account technical Hi–C biases and polymer looping.
   Pipeline used: A set of Python scripts that can be found on: `http://noble.gs.washington.edu/proj/fit-hi-c`. It accepts as input list of locus pairs and the corresponding counts and output the list with P-values and Q-values. Corrected maps (after iterative correction for example) can also be used as input.
   Result: It is interesting that while they confirm previously described promoter-enhancer interactions (77% of the promoter-enhancer interactions mediated by RNApolII for example), they claim that most contacts happen in insulators and heterochromatin regions and not in enhancers and euchromatin regions (differences with Ren paper here?). They also worked on NANOG which is very interesting and they found many contacts mediated by this master regulator in embryonic stems cells.

4. Lu, Y., Zhou, Y. & Tian, W. Combining Hi-C data with phylogenetic correlation to predict the target genes of distal regulatory elements in human genome. Nucleic Acids Research 41, 10391-10402 (2013).

   Method: They combine phylogenetic information with available Hi-C data to predict interactions of promoters with distal regulatory elements (DREs).
   Pipeline used: No code available.
   Result: No comparisons available with other methods.

5. Hu, M., Deng, K., Qin, Z. & Liu, J. S. Understanding spatial organizations of chromosomes via statistical analysis of Hi-C data. Quant Biol 1, 156-174 (2013).

   Method: Not a new method but a very good review of the existing ones. It also provides a very good description of the method itself, the biases and the statistical challenges. Table 1A gives a comprehensive overview of the current methods for Hi–C data analysis. It mentions pros and cons and the corresponding references as well.

Pipeline used: -
Result: -

6. Hu, M. et al. Bayesian Inference of Spatial Organizations of Chromosomes. PLoS Comput. Biol. 9, (2013).

   Method: A Bayesian probabilistic approach named “Bayesian 3D constructor for Hi-C data”.
   Pipeline used: BACH and BACH-MIX algorithms created in their study. Account for systematic biases and problems due to differences in sequencing depth.
   Result: Successful detection of euchromatic–heterochromatic regions. Demonstrated superiority of these algorithms when compared with MCMC5C algorithm, as the results from BACH and BACH-MIX agree more with FISH data.

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

   Method: Iterative correction and eigenvector decomposition (ICE). An iterative normalization method is used. It is not biologically relevant though as it assumes “equal visibility” of all loci, which cannot be the case.
   Pipeline used: Software (Python code) is available on `https://bitbucket.org/mirnylab/hiclib`.
   Result: -

8. Yaffe, E. & Tanay, A. Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. Nature Genetics 43, 1059-1065 (2011).

   Method: An integrated probabilistic background model. Takes into account GC bias, mappability problems and the distance between restriction sites.
   Pipeline used: The corresponding Hi-C pipeline (hicpipe) is available on: `http://compgenomics.weizmann.ac.il/tanay/?page_id=283`
   Result: This method removes the aforementioned biases but it has a major disadvantage: it is very slow.

9. Hu, M. et al. HiCNorm: removing biases in Hi-C data via Poisson regression. Bioinformatics 28, 3131-3133 (2012).

   Method:
   Much simplified normalization method when compared with that of Yaffe et al. (see above). It uses a generalized linear model and it corrects the same biases mentioned above. **Pipeline used:** Available on: http://www.people.fas.harvard.edu/ junliu/HiCNorm/.
   Result: > 1000 times faster than Yaffe *et al.*

10. Jin, F. et al. A high-resolution map of the three-dimensional chromatin interactome in human cells. Nature 503, 290-294 (2013).

    Method: -
    Pipeline used: -
    Result: -

Interesting Questions

Idea 1 Gene content per chromosome (in terms of active genes) and gene interactions. (You have more *cis* than *trans* interactions).

Idea 2 To challenge the so-called "transcription factories". If "transcription factories" exist, given that the vast majority of 3D chromosomal interactions are *cis* interactions, the co-regulated genes should tend to be on the same chromosome. While this could be true in certain cases, I am pretty sure that there are genes on different chromosomes that are regulated by the same factor. Also mobility of chromatin in the nucleus is rather limited [2, 6]. The "transcription factory" model could potentially apply in certain cases of master regulators (KLF-1 for example) [5] but it is has clearly not been established that it is the rule.

Idea 3 Mentioned by Misteli [3]. Hi-C data refer to populations. This may obscure the relationship of 3D chromatin structure and function. Even in the same population of cells, there are cells expressing a gene and others that do not. What is the difference in the chromatin structure though when certain genes are expressed vs. when they are not? The recent development of single-cell Hi-C (if it really works as advertised) [4] may help answer this question.

Idea 4 The "kill many birds with one stone" hypothesis. If we suppose that co-regulated genes (by master transcription factors) are truly in close proximity in 3D space, a favorite and very economical –in terms of changes required– strategy in cancer would be to create master regulator behavior where it does not exist and mess up many genes all together. This could be done by fusing the activation-domain of a master regulator with the binding domain of an abundant transcription factor (Aris).

Idea 5 Mentioned by Iannis during the lab meeting on April 11, 2014. To check if there are enhancers that control the expression of genes that reside on different chromosomes. Panos mentioned that Thanos has shown it with NF-kappaB. You may want to ask Panos about the exact publication. Bryan also mentioned that a nice experiment would be to check chromatin structure changes upon cell differentiation. Would monoallelic expression be a nice system for that? Based on what Spector published earlier this year [1], embryonic stem cells tend to express both alleles while mature cells (at least neurons) tend to express one of the two alleles. If this is the case with cells of the immune system as well, the chromatin structure of the precursors and the mature cells could be checked in order to check if differences in chromatin structure explain the difference in expression. Advantage: y showing differences between alleles we are sure that we are talking about the same cells (same environment). Disadvantage: Hi–C experiments have to be performed in the lab (so cells of the immune system have to be used as model).

References

[1] Melanie A Eckersley-Maslin, David Thybert, Jan H Bergmann, John C Marioni, Paul Flicek, and David L Spector. Random Monoallelic Gene Expression Increases upon Embryonic Stem Cell Differentiation. *Developmental Cell*, 28(4):351–365, 2014.

[2] R Ileng Kumaran and David L Spector. A genetic locus targeted to the nuclear periphery in living cells maintains its transcriptional competence. *The Journal of Cell Biology*, 180(1):51–65, January 2008.

[3] Tom Misteli. Parallel genome universes. *Nature Reviews Genetics*, 30(1):55–56, January 2012.

[4] Takashi Nagano, Yaniv Lubling, Tim J Stevens, Stefan Schoenfelder, Eitan Yaffe, Wendy Dean, Ernest D Laue, Amos Tanay, and Peter Fraser. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature*, 502(7469):59–64, 2013.

[5] Stefan Schoenfelder, Tom Sexton, Lyubomira Chakalova, Nathan F Cope, Alice Horton, Simon Andrews, Sreenivasulu Kurukuti, Jennifer A Mitchell, David Umlauf, Daniela S Dimitrova, Christopher H Eskiw, Yanquan Luo, Chia-Lin Wei, Yijun Ruan, James J Bieker, and Peter Fraser. Preferential associations between co-regulated genes reveal a transcriptional interactome in erythroid cells. *Nature Genetics*, 42(1):53–61, January 2010.

[6] Evi Soutoglou, Jonas F Dorn, Kundan Sengupta, Maria Jasin, Andre Nussenzweig, Thomas Ried, Gaudenz Danuser, and Tom Misteli. Positional stability of single double-strand breaks in mammalian cells. *Nature Cell Biology*, 9(6):675–682, June 2007.