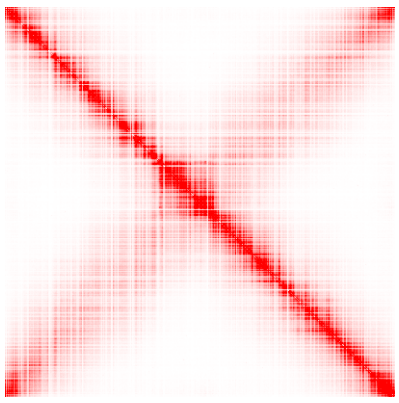
CAUSES FOR BIAS IN 3C-LIKE DATA

Yannick G. Spill[1] and Marc A. Marti-Renom[1,2]

[1]Structural Genomics Group. Centre Nacional d'Anàlisi Genòmica - Centre de Regulaciò Genòmica (CNAG-CRG), Barcelona, Spain.

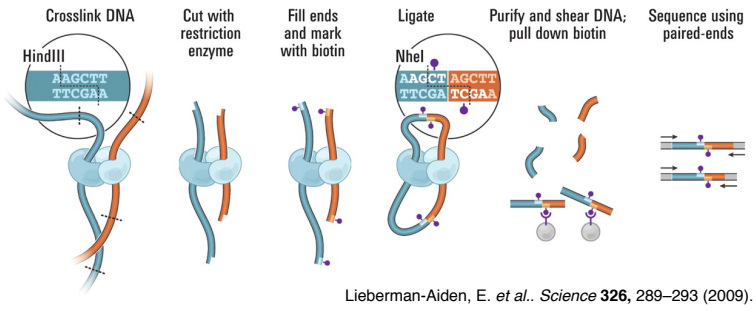
[2]Institució de Recerca i Estudis Avançats (ICREA), Barcelona, Spain



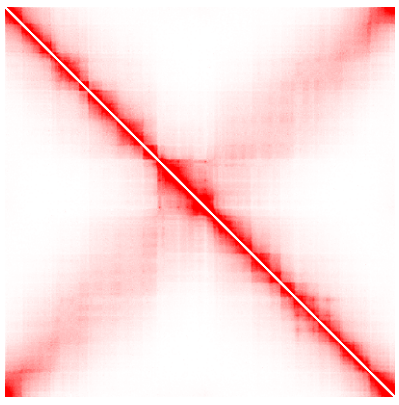
Raw counts

Introduction

Chromatin structure determination is a fast evolving field. It recently emerged with the invention of 3C-like experiments, in particular 3C [1] and Hi-C [2]. These experiments allow to probe for the spatial distance between two genetic loci. Yet, these experiments do not provide the distances themselves, but a contact count, which is prone to be biased by a number of genomic factors. It is therefore crucial to de-bias the data for any practical application.



Lieberman-Aiden, E. *et al.*. *Science* **326,** 289–293 (2009).

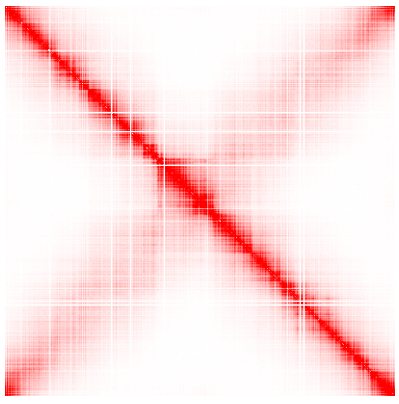


Normalized with ICE

***ICE**: Most popular normalization method*

A number of methods have been proposed to de-bias these datasets, but none so far has reached an overall state of acceptance by the community. ICE [3] rescales the coefficients of the matrix so that all rows and columns have an equal sum, while excluding the diagonal and nearest neighbors. It provides the best treatment, but it has a number of limitations:

- ICE assumes that any locus interacts with someone that is not its closest neighbor.
- ICE does not provide confidence intervals for the contacts
- ICE can only be applied to whole genomes (HiC), or experiments where sequencing is exhaustive for the lines considered (capture-(Hi)C)
- ICE does not provide goodness-of-fit criterions, so we don't know whether ICE has over-cleaned a matrix or not.



Normalized with HiCNorm

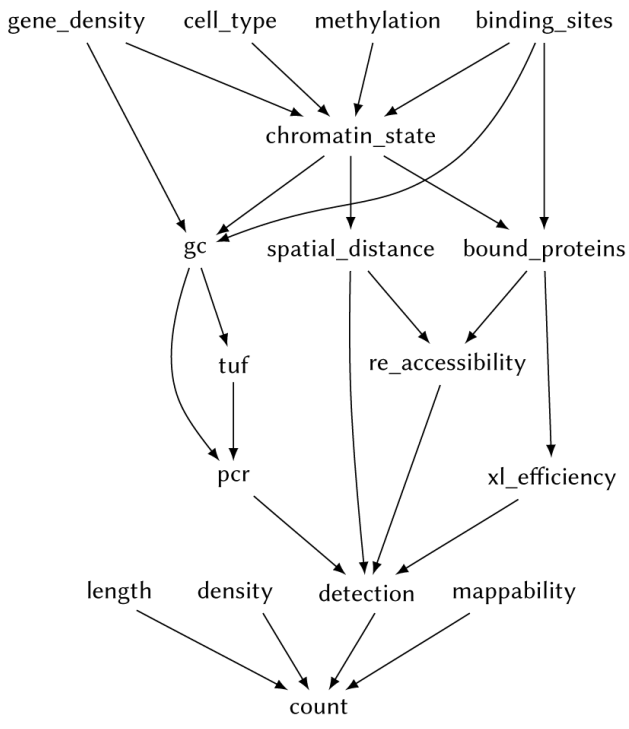
***HiCNorm**: most flexible method*

HiCNorm [4] is based on the method by Yaffe and Tanay [5] and performs a less popular normalization than ICE. It is, however, much more flexible, and can be applied to a large variety of 3C-like datasets.
HiCNorm is based on Poisson regression of the counts using three possible sources of bias: **a)** GC content, **b)** mappability of reads onto the reference genome and **c)** density of restriction fragments per bin.
Compared to ICE, it can give confidence intervals for the contacts, and criteria for goodness of fit (e.g. AIC)

Causal graph

We represent the sources of bias in 3C-like data in a causal graph. As can be seen, a number of factors have been left out in HiCNorm.

- The GC content is too unspecific, because controlling for it would make detection of the chromatin state more difficult.
- The cross-linking efficiency has not been taken into account, and may play a big role if the amount of protein bound to the DNA varies much and non-randomly along the genome.
- The length of the restriction fragments is not entirely taken into account.



Importance

The use of an improved set of factors for normalization would allow for more robust structure determination. In particular, robust estimates of error bars on the normalized counts will translate in a more dynamic representation of consensus structures.

Bibliography

1. Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. *Capturing chromosome conformation*. Science. **295**(5558): 1306–1311 (2002).
2. Belton JM, McCord RP, Gibcus JH, Naumova N, Zhan Y, Dekker J. *Hi–C: A comprehensive technique to capture the conformation of genomes.* Methods. **58**(3): 268-276 (2012).
3. Imakaev M, Fudenberg G, McCord RP, Naumova N, Goloborodko A, Lajoie BR, Dekker J, Mirny LA. *Iterative correction of Hi-C data reveals hallmarks of chromosome organization.* Nature Methods. **9**: 999-1003 (2012).
4. Hu M, Deng K, Selvaraj S, Qin Z, Ren B, Liu JS. *HiCNorm: removing biases in Hi-C data via Poisson regression.* Bioinformatics. **28**(23): 3131-3133 (2012).
5. 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).

cnag
centre nacional d'anàlisi genòmica
centro nacional de análisis genómico

CRG Centre for Genomic Regulation

erc European Research Council
Established by the European Commission

European Commission

>> more information at http://marciuslab.org <<