BIOINFORMATICS APPLICATIONS NOTE
Vol. 28 no. 23 2012, pages 3131–3133
doi:10.1093/bioinformatics/bts570

Genome analysis
Advance Access publication September 27, 2012

HiCNorm: removing biases in Hi-C data via Poisson regression

Ming Hu[1], Ke Deng[1], Siddarth Selvaraj[2,3], Zhaohui Qin[4], Bing Ren[2] and Jun S. Liu[1,*]

[1]Department of Statistics, Harvard University, Cambridge, MA 02138, USA, [2]Department of Cellular and Molecular Medicine, UCSD School of Medicine, La Jolla, CA 92093, USA, [3]Bioinformatics and Systems Biology Graduate Program, University of California, San Diego, La Jolla, CA 92093, USA and [4]Department of Biostatistics and Bioinformatics, Emory University, Atlanta, GA 30322 USA

Associate Editor: Alex Bateman

ABSTRACT

Summary: We propose a parametric model, HiCNorm, to remove systematic biases in the raw Hi-C contact maps, resulting in a simple, fast, yet accurate normalization procedure. Compared with the existing Hi-C normalization method developed by Yaffe and Tanay, HiCNorm has fewer parameters, runs >1000 times faster and achieves higher reproducibility.

Availability: Freely available on the web at: http://www.people.fas.harvard.edu/~junliu/HiCNorm/.

Contact: jliu@stat.harvard.edu

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

Received on May 24, 2012; revised on August 9, 2012; accepted on September 17, 2012

1 INTRODUCTION

Understanding how genome is organized in three-dimensional (3D) space inside the nucleus is of great importance to the study of genome function (Naumova and Dekker, 2010). Harnessing the power of the massively parallel sequencing technologies, Dekker and his colleagues developed a genome-wide approach to characterize the long-range chromatin interactions and 3D organization of chromosomes (Lieberman-Aiden *et al.*, 2009). This technique, known as Hi-C, uses proximity ligation and high throughput sequencing to interrogate the spatial distances between pairs of genomic loci, resulting in a matrix of pairwise contact frequencies along the genome, which could be used to infer 3D folding of chromosomes inside the nucleus (Lieberman-Aiden *et al.*, 2009).

In Hi-C experiments, the frequency of chromatin interaction between two genomic loci is represented by the number of paired-end reads linking the two genomic sequences. In principle, higher read counts indicate increased frequency of chromatin interaction and closer spatial distances, but many additional factors may also contribute to the actual read counts observed. Recently, Yaffe and Tanay reported systematic biases inherent in these Hi-C read counts, including the distance between restriction enzyme cut sites, the GC content of fragment ends and uniqueness (or mappability) of short sequence reads (Yaffe and Tanay, 2011). These factors dramatically affect the read counts in Hi-C contact maps. To remove these systematic biases, a non-parametric probabilistic model (referred to hereafter as the YT approach) was proposed that explicitly models the probability of observing a paired-end read spanning two fragment ends. This approach can remove the majority of systematic biases and substantially improve the reproducibility in Hi-C contact maps.

One limitation of the YT approach is its extremely intensive computation cost. Because the likelihood function contains probabilities of observing a paired-end read spanning all possible fragment end pairs (in the order of 10^{12}), the computation cost for any statistical analysis based on such a model is overwhelming. Furthermore, the YT approach uses step functions with 420 unknown parameters to approximate the joint effect of multiple systematic biases. It is very challenging to interpret such step functions, if not impossible.

Here, we propose HiCNorm, a greatly simplified normalization procedure that is much less computationally intensive. HiCNorm uses a generalized linear model (McCullagh and Nelder, 1989) to correct the systematic biases in Hi-C contact maps. We show that, in addition to being many orders of magnitude faster, HiCNorm achieves a higher reproducibility than the YT approach on published Hi-C datasets.

2 RESULTS

We analyzed the same Hi-C dataset used in Yaffe and Tanay's work, which was originally published in Lieberman-Aiden *et al.*, 2009. In Lieberman-Aiden's study, two restriction enzymes (HindIII and NcoI) were used to prepare two Hi-C libraries. We referred to these two biological replicates as the HindIII dataset and the NcoI dataset, respectively. The goal is to normalize the read count data in the contact map at certain resolution level and achieve high reproducibility between these two biological replicates. With the current sequencing depth, the resolution is much coarser than the length of a fragment (~4 KB). Instead of modeling Hi-C data at the fragment end level as in the YT method, we directly model the read count data at the desired resolution level (1 MB). We first used Yaffe and Tanay's protocol (Yaffe and Tanay, 2011) to pre-process the Hi-C raw reads (Supplementary Material) and then explored sources of Hi-C biases at 1 MB resolution level. First, as only reads that mapped within 500 bp of the nearest restriction site were kept for downstream analysis, we truncated each fragment end up to 500 bp, and calculated the total length of truncated fragment end for each genomic locus (effective length feature).

*To whom correspondence should be addressed.

We further calculated the GC content within a 200 bp region upstream of each fragment end (GC content feature) and the mappability score within a 500 bp region next to each fragment end (mappability feature). The GC content feature and the mappability feature of each genomic locus are defined as the average of the corresponding features among all fragment ends falling into this locus. Supplementary Figure S1 illustrated these three systematic biases in Hi-C contact maps at 1 MB resolution level, which share similar patterns to the biases at the fragment end level (Yaffe and Tanay, 2011).

To remove aforementioned systematic biases, we propose HiCNorm, a Poisson regression model for Hi-C contact maps normalization (Supplementary Material). In HiCNorm, we estimated the bias effects of the effective length feature and the GC content feature while fixing the mappability feature as a Poisson offset. We applied HiCNorm to the Hi-C *cis* (*trans*) contact map of each chromosome (chromosome pair) separately, and reported the estimated biases in Supplementary Figure S2. We observed a positive correlation between the effective length in each locus pair and the number of contacts in both HindIII dataset and NcoI dataset. The GC content feature is negatively associated with the number of contacts in the HindIII dataset and positively associated with that in the NcoI dataset. Noticeably, the magnitude of the estimated effective length biases is significantly larger than that of the estimated GC content biases (paired *t*-test *P*-value < 2.8e-11). These results suggest that the effective length is the dominant bias factor at 1 MB resolution level, and the bias owing to the GC content is secondary. We further conducted the residual analysis (Supplementary Material) and found that HiCNorm fits data well in spite of over-dispersion. To account for the over-dispersion, we also implemented a negative binomial (NB) regression model for Hi-C contact maps normalization (Supplementary Material). The NB regression model achieved a slightly better fit to the data than the Poisson regression model in terms of the Akaike information criterion (Akaike, 1974), but these two models provided very similar bias reductions. We therefore focused on the Poisson regression model henceforth, which is mathematically simpler. Our R package provides a choice for the user to use either the NB or the Poisson regression models.

We used the residuals of the Poisson regression as the normalized Hi-C contact maps (Supplementary Material). For comparison, the normalization results of the YT approach were downloaded from Yaffe's webpage: http://compgenomics.weizmann.ac.il/tanay/?page_id=283. The better normalization method achieves higher reproducibility (measured by Spearman's rank correlation coefficient) between the HindIII dataset and the NcoI dataset. We also reported the reproducibility of raw Hi-C contact maps.

We first calculated the Spearman correlation coefficients of the two raw Hi-C *cis* contact maps obtained from the HindIII dataset and the NcoI dataset for each of the 23 chromosomes (chromosome 1∼22 and X). We found that the YT approach improves the average Spearman correlation coefficients across the 23 chromosomes from 0.7241 to 0.8041, whereas HiCNorm further improves the correlation coefficient to 0.8111 (Supplementary Fig. S3a). HiCNorm achieved significantly higher reproducibility than raw Hi-C reads (paired *t*-test *P*-value = 1.3e-14) and the YT approach (paired *t*-test *P*-value = 0.0004) for the population of 23 chromosomes.

We further evaluated the reproducibility of Hi-C *trans* contact maps obtained from the HindIII dataset and the NcoI dataset for each chromosome pair (253 chromosome pairs in total). As the 1 MB Hi-C *trans* contact maps are sparse, instead of measuring the reproducibility of Hi-C *trans* contact maps directly, we studied the reproducibility of two one-dimensional (1D) coverage vectors, which are defined as vectors of row and column summations, respectively, of each Hi-C *trans* contact map and used in Yaffy and Tanay's study (Yaffe and Tanay, 2011). Across the total 506 1D coverage vectors, the average Spearman correlation coefficient of raw Hi-C 1D coverage vectors obtained from the HindIII dataset and the NcoI dataset was − 0.1106. After normalization, the correlation coefficient was increased to 0.5818 (highly significant) using the YT approach, and further increased to 0.5832 using HiCNorm (Supplementary Fig. S3b) (statistically insignificant).

While providing comparable reproducibility of Hi-C *trans* contact maps with respect to the YT approach, HiCNorm is more robust. The standard deviations of the Spearman correlation coefficients of 1D coverage were 0.1199 and 0.1679 for HiCNorm and the YT approach, respectively (Supplementary Fig. S3b). As an illustrative example, the Spearman correlation coefficient of the raw 1D coverage of chromosome 22 (using the *trans* contact between chromosome 4 and chromosome 22) was − 0.2058 before normalization and − 0.0689 after normalization using the YT approach, but it was improved dramatically to 0.6518 using HiCNorm (Supplementary Fig. S4).

Due to the intensive computation requirement, it was not feasible to normalize the whole genome contact maps (Lieberman-Aiden *et al.*, 2009) with the YT approach on a computer with one single CPU, but with HiCNorm, it took ∼8 min. To make a feasible and fair comparison of the computation times of the YT approach and HiCNorm, we used the testing dataset published with the source code of the YT approach (downloaded from http://compgenomics.weizmann.ac.il/tanay/?page_id=283), which is a subset of the Hi-C dataset (Lieberman-Aiden *et al.*, 2009) used in this study (*cis* and *trans* interactions of chromosome 19, 20, 21 and 22 in the NcoI dataset), as the benchmark to evaluate the computation time. For this testing dataset, the YT approach took ∼4 h (14 041 s), whereas HiCNorm only took ∼2 s, i.e. HiCNorm is >7000 times faster than the YT approach.

In this work, we propose HiCNorm, a novel normalization procedure based on Poisson regression to remove systematic biases in Hi-C contact maps. Compared with the YT approach, HiCNorm improves the reproducibility between the Hi-C datasets obtained from libraries prepared using different restriction enzymes. Additionally, because HiCNorm involves a much smaller number of parameters (only three parameters in HiCNorm compared with 420 parameters in the YT approach), we are able to significantly reduce the computation cost. The significant improvement is made possible due to two reasons: first, we directly model the frequency of chromatin interaction in the desired resolution level Hi-C contact maps so as to enable sufficient data reduction at an early stage; second, the hyperbola-shaped contour plot of the effective length bias and the GC content bias (Supplementary Fig. S1) motivated us to adopt a parametric method to account for multiple systematic biases, which makes the inference much simpler than the non-parametric method used in the YT approach. We also found that increasing the model complexity (adding more interaction terms among the features in HiCNorm) did not lead to further improvement in terms of the reproducibility using real data. Our study confirms the critical importance of proper normalization of the Hi-C read map data. With the increased popularity of the Hi-C technology in the biomedical field, HiCNorm has the potential to become a convenient and powerful normalization tool for Hi-C data analysis.

Funding: This work was supported by US National Institutes of Health grants R01HG005119 (to Z.Q.), 5R01GM080625 (to J.S.L.) and the Ludwig Institute for Cancer Research (to B.R.).

Conflict of Interest: none declared.

REFERENCES

Akaike,H. (1974) A new look at the statistical model identification. *IEEE Trans. Automat. Control*, **19**, 716–723.

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

McCullagh,P. and Nelder,J.A. (1989) *Generalized Linear Models*. Chapman & Hall/CRC, London.

Naumov,N. and Dekker,J. (2010) Integrating one-dimensional and three-dimensional maps of genomes. *J. Cell Sci.*, **123**, 1979–1988.

Yaffe,E. and Tanay,A. (2011) Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nat. Genet.*, **43**, 1059–1065.