Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Original Article

Genome Analysis

Dryhic: Increasing Reproducibility of Hi-C Samples with Abnormal Karyotypes

Enrique Vidal [1,2]*, François le Dily [1,2], Javier Quilez [1,2], Ralph Stadhouders [1,2], Yasmina Cuartero [1,2] Guillaume Fillion [1,2] and Miguel Beato [1,2]

[1]Gene Regulation, Stem Cells and Cancer Program, Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology (BIST), Barcelona, Spain
[2]Universitat Pompeu Fabra (UPF), Barcelona, Spain

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Several strategies have been proposed to remove biases from Hi-C experiments. However, samples with abnormal karyotypes (such as cancer cell lines) present challenges not completely addressed by existing methods. We introduce an alternative designed to model the total number of contacts per genomic loci, thus taking into account possible chromosomal abnormalities.

Results: The proposed method increases reproducibility among replicas of Hi-C samples with abnormal karyotype and obtained using different protocols. The implementation is fast and scales well in terms of computing resources for resolutions up to 1 Kbp.

Availability: The strategy proposed is implemented as an R package available at http://www.github.com/qenvio/dryhic.

Contact: enrique.vidal@crg.eu

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

1 Introduction

High throughput chromosomal conformation capture (HiC) experiments allow us to explore the spatial organization of chromatin in cells (Lieberman-Aiden *et al.*, 2009). Contrary to other 3C techniques, HiC is unbiased in the sense that all possible pairwise interactions between restriction fragments might be interrogated. However, this does not guarantee absence of bias. For instance, genome features as CG content, restriction enzyme site distribution and sequence uniqueness (related to mappability) have been described to modify the result in such experiments (Yaffe and Tanay, 2011).

Several strategies have been proposed to remove biases in HiC experiments (Schmitt *et al.*, 2016), either modeling explicitly an a-priori set of known genomic features as potential sources of bias (Yaffe and Tanay, 2011; Hu *et al.*, 2012) or using algorithms to implicitly correct for unknown biases (Imakaev *et al.*, 2012). Notably, the latter has a broader usage due to its fast implementation. Both types of methods are suitable for normal cells but they were not designed for cells with abnormal karyotypes. Although there have been some proposals that consider aneuploidy (Wu and Michor, 2016), to our knowledge, there is a lack of methodology dealing with aberrant karyotypes (common in cancer cell lines) that present abnormalities more complex than entire chromosome duplications.

Here we propose *oned*, an alternative designed to model total number of contacts per genomic loci, thus taking into account possible chromosomal abnormalities. It explicitly models the contribution of known biases via a generalized additive model (GAM). The resulting data present better overall reproducibility between replicas and across different protocols, specially, but not restricted to, in the presence of aberrant karyotypes. Besides, the implementation is fast and scales up to 1 Kbp resolution with reasonable computing resources.

2 Methods

2.1 Model

A common representation for the data yielded in a HiC experiment is a contact matrix, obtained slicing the genome in contiguous bins of fixed length (resolution) and computing the number of contacts between each

© The Author 2017. Published by Oxford University Press. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

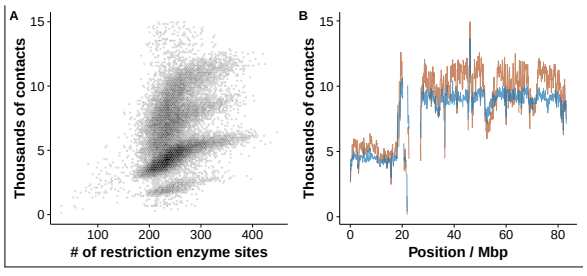

Fig. 1. A Non linear relationship between a genomic feature (number of restriction sites per bin) and the total number of contacts per bin for a T47D sample. B Total number of contacts per bin in chromosome 17 of a T47D sample. Raw (in brown, noisier) and corrected (in blue, more precise) data are depicted.

pair of bins, corresponding to a cell in the contact matrix (x_{ij}), providing a quantification of the interaction between the tow loci.

Our proposal is based on the overall contacts registered for each bin, thus reducing the matrix to one dimension (hence the name). We assumed that the total number of contacts per bin (t_i) could be approximated by a negative binomial distribution with every potentially biasing genomic feature (known a priori) contributing independently to its mean (λ_i). Given that this relationship might not be linear (as can be seen in Figure 1A), we allowed a smooth representation using thin plate penalized regression splines (Wood, 2003) in a generalized additive model (Wood, 2011). The model can be parametrized as follows:

$$t_i = \sum_i x_{ij}$$

$$t_i \sim NB(\lambda_i, \theta)$$

$$log(\lambda_i) \propto \sum_K f_k(z_k)$$

where x_{ij} are the number of contacts between bins i and j, and there is an additive term for each z_k genomic feature. The smooth functions $\{f_k(z_k)\}$ are estimated jointly with the negative binomial dispersion parameter θ using R (R Core Team, 2017) `mgcv` package (Wood, 2011).

Once the model of the total counts has been estimated, we can further rescale $\{\lambda_i\}$ to obtain a correction vector $\{\lambda'_i\}$ that can be used to compute the corrected counts (c_{ij}) or, alternatively, it can be included in downstream analysis:

$$\lambda'_i = \frac{\lambda_i}{\sum_j^n \lambda_j / n}$$

$$c_{ij} = \frac{x_{ij}}{\sqrt{\lambda'_i \lambda'_j}}$$

Following previous works (Yaffe and Tanay, 2011; Hu *et al.*, 2012) we included mappability, CG content and number of restriction sites (comparable to effective length) as genomic features, but the model and implementation allow easy extension with any available genomic feature.

2.2 Data sources

As an environment to test the bias correction, we gather a set of published (Le Dily *et al.*, 2014; Consortium *et al.*, 2012; Rao *et al.*, 2014; Stadhouders *et al.*, 2017; Lin *et al.*, 2012; Dixon *et al.*, 2012) and unpublished data from HiC experiments of different cell types and organisms. The details can be found in Table **??**. Briefly, we used several experiments comprising T47D breast cancer cell lines (7 samples), K562 leukemia cell lines (4 samples), both with aberrant karyotypes, and mouse primary B (6 samples) and ES (7 samples) cells, both with normal diploid karyotypes. The experiments used different protocols (*i.e.* original HiC (Lieberman-Aiden *et al.*, 2009) and in-situ HiC (Rao *et al.*, 2014)), different restriction enzymes (DpnII, HindIII, MboI and NcoI) and they were produced in distinct laboratories.

We also make use of array-based copy-number segmentation of the two cell lines obtained from COSMIC database (Forbes *et al.*, 2010).

2.3 Data processing

All data were processed though an in-house pipeline based on TADBit (Serra *et al.*, 2016). Briefly, after fastq quality control, paired reads were mapped against the corresponding reference genome (hg38 and mm10) taking into account the restriction enzyme motifs. Non-informative contacts were removed applying the following TADBit (Serra *et al.*, 2016) filters: Self-circle, dangling-ends, error, extra dangling-end, duplicated and random-breaks. For more details, see methods of Stadhouders *et al.* (2017). In addition, the pipeline is available, with minor modifications, in Quilez *et al.* (2017) material.

We developed the routines contained in the `dryhic` R package to efficiently create sparse representations of contact matrices and further apply *vanilla*, *ice* and *oned* corrections. `HiTC` (Servant *et al.*, 2012) and `HiCapp` (Wu and Michor, 2016) were used to get *lgf* and *caib* corrections respectively. All results are based on 100 Kbp resolution, although no major differences were found using different ones (data not shown).

2.4 Total number of contacts as copy number proxy

In euploid samples, we expect the coverage of a HiC experiment to be directly related to the number of copies (a duplicated region will show, on average, twice the number of contacts). An ilustrative example can be seen in Figure 2B. Given that our model is based on total number of contacts, which in turn is directly related to the number of copies, we reasoned that a preliminary test would be to check if the corrected number of contacts per bin ($ct_i = \sum_i c_{ij}$) reflects the copy number (as measured by an independent technique such as array-based copy-number segmentation) better than the alternatives.

2.5 Sample comparison

In order to test performance, we relied on pair-wise comparisons of contact matrices. However, there is little consensus on how to compare two HiC matrices. The simplest metric is the Spearman correlation applied to cis (intra-chromosomal) contacts up to a given distance (*i.e.* 5 Mb). Though, the decay in the number of contacts with increasing genomic distance is the main driver of contact abundance and it could mask real differences between samples. This distance decay can be taken into account, for instance, computing the expected number of interactions given the genomic distance between any pair of loci. Thus, we can compare matrices in terms of observed over expected contacts, via the Pearson correlation (again focusing on a given distance range). Alternatively, we can compute the correlation per distance stratum and then obtain a stratum-adjusted correlation coefficient (SCC) as defined in Yang *et al.* (2017). Additionally, Yan *et al.* (2017) proposed a reproducibility score based on the correlation between a set of the Laplacian matrix's eigenvectors (corresponding to the smallest eigenvalues) of the two experiments.

We defined three sets of samples: One with all T47D experiments plus two K562 samples, another with all K562 experiments plus two T47D samples and a third one comprising all mm10 samples (see Table **??** for details). Given a set of experiments and a metric, we first computed all pair-wise combinations between experiments and then classified the comparisons according to the characteristics of each pair (*i.e.* same / different cell types, same / different protocols, etc ...).

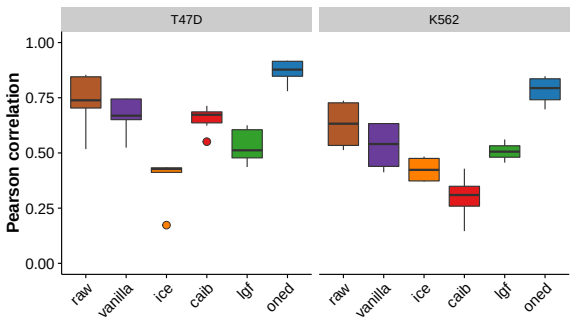

Fig. 2. Pearson correlation between total number of contacts per bin and independent copy number estimation (COSMIC) for each of the methods compared. Left panel T47D breast cancer cell line, right panel K562 leukemia cell line. The new proposal (in blue) outperforms the rest of alternatives.

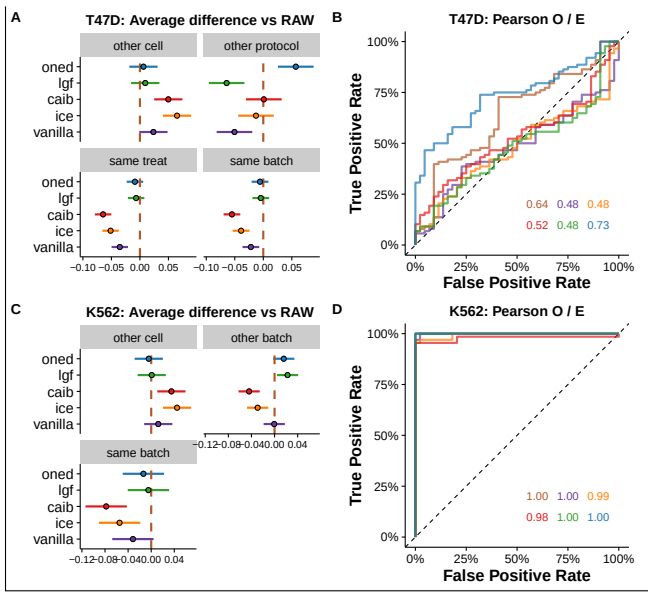

Fig. 3. Results of the comparison between samples with aberrant karyotype. Left panels: Average difference and 95% CI between each correction method (Y axis) and the raw data (brown dashed line) for T47D (A) and K562 (C) sets. Right panels: Corresponding ROC curves and AUCs for T47D (B) and K562 (D) sets. All results in this figure are based on Pearson correlations between the observed over expected counts. Our method increases reproducibility of similar samples while being able to discriminate different ones.

To benchmark our method (***oned***), we also test the performance of other alternatives. Three of them were based on the same implicit iterative correction (Imakaev *et al.*, 2012): A first alternative computing one iteration (***vanilla***), a second one iterating until convergence (***ice***) and a third one using the chromosome adjusted iterative correction (***caib***) (Wu and Michor, 2016). The fourth one was inspired on the explicit bias correction originally presented by Yaffe and Tanay (2011), but based on the local genomic features (***lgf***) approach introduced by Hu *et al.* (2012) as implemented in Servant *et al.* (2012).

Besides, we compared the raw matrices and used these comparisons as a baseline to present the gain / loss of correlation relative to the uncorrected data, thus allowing us to verify the effect of the bias correction procedure. The differences relative to the raw comparisons were estimated using a linear mixed model with the method as fixed effect and the chromosome as a random effect, and using the `lmer` function of the `lme4` R package (Bates *et al.*, 2015).

In an attempt to further summarize the results, we reasoned that the correlation between unrelated samples (*i.e* different cell types) should be lower than the correlation between related samples (*i.e* same cell types), so we could re-interpret each correlation metric as a classifier score. Thus, we computed the receiver operating characteristic (ROC) curves and corresponding area under the curve (AUC) using the R `ROCR` package (Sing *et al.*, 2005).

3 Results

3.1 Copy number

Figure 1B shows the total number of corrected contacts along chromosme 17 of a T47D sample. Our correction approach presents a more stable and precise estimation of the copy number. Quantitatively (Figure 2), Pearson correlation with independent copy number data is high for both T47D and K562 cell lines (averages of 0.87 and 0.78 respectively), not only compared with raw data (averages of 0.74 and 0.63 respectively) but also with any of the alternative methods (averages ranging 0.39-0.68 and 0.30-0.53 respectively). Similar results were obtained using Spearman correlation (Supplementary Figure 2).

3.2 Samples with aberrant karyotypes

Focusing on the Pearson correlation of observed over expected counts, methods based on the iterative correction (mainly *ice* and *caib* but also *vanilla*) present the undesirable feature of higher correlations between unrelated samples than using raw data (top left panels of Figures 3A and 3C, Supplementary Figure 3). Among the comparisons of cells with the same type, *oned* is the only method outperforming uncorrected data when comparing experiments using different protocols (top right panel of Figure 3A) and it is among the top performant when comparing experiments using the same protocol (rest of the panels in Figures 3A and 3C). Consequently, our proposal is the one that better discriminates unrelated samples form similar ones (as it can be seen in the ROC curves and AUCs of Figures 3B and D).

Complementary analysis using other metrics (Supplementary Figures 3, 4 and 5) yielded similar results. Notably, Spearman correlation of contacts presents the worst performance for all scenarios and seems to be a poor choice as a metric to compare HiC matrices.

3.3 Samples with normal karyotypes

Regarding cells with diploid genomes, we confirmed that our bias correction strategy also achieves good results in terms separation between unrelated and similar samples (Figure 4 and Supplementary Figures 6 and 7), increasing the reproducibility across different experimental protocols for primary B and ES mouse cell samples. Again, Spearman correlation of contacts revealed as the worst metric.

3.4 Computing resources

One of the main reasons for the broad usage of the iterative correction (Imakaev *et al.*, 2012) is the easy and fast implementation, moreover compared to existing explicit methods (Servant *et al.*, 2012). Our proposal uses only the total number of contacts per bin instead the full matrix for the model estimation. Thus, it achieves similar computing times that the fastest approaches (Figure 5 A). Yet it scales with increasing resolutions (decreasing bin size) (Figure 5 B).

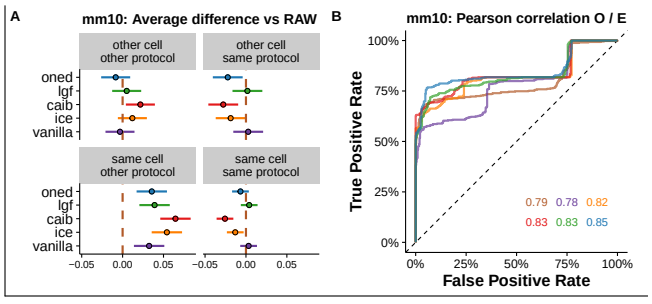

Fig. 4. Results of the comparison between samples with normal karyotype. A Average difference and 95% CI between each correction method (Y axis) and the raw data (brown dashed line) for the mm10 sets. B Corresponding ROC curves and AUCs. The introduced strategy does not under perform compared to the alternatives.

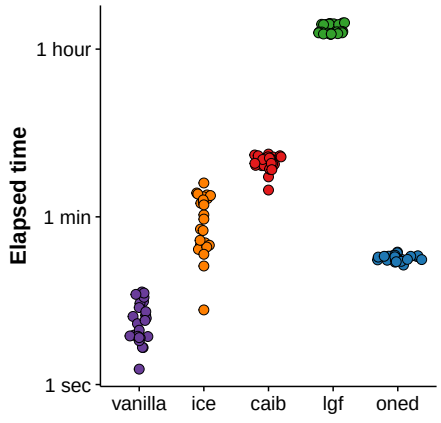

Fig. 5. Computing time (Y axis, log scale) of all the bias correction methods used. Each dot corresponds to one sample. The only method faster than ours under performs in all sample comparisons.

4 Discussion and Conclusions

Acknowledgements

We would like to thank all members of the four labs involved in the 4DGenome Synergy project for the fruitful discussions during project meetings. EV wants to acknowledge specially the members of Miguel Beato's lab for their insights during lab meetings.

Funding

This work has been supported by the... Text Text Text Text.

References

Bates, D. *et al.* (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, **67**(1), 1–48.

Consortium, E. P. *et al.* (2012). An integrated encyclopedia of dna elements in the human genome. *Nature*, **489**(7414), 57–74.

Dixon, J. R. *et al.* (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, **485**(7398), 376–380.

Forbes, S. A. *et al.* (2010). Cosmic: mining complete cancer genomes in the catalogue of somatic mutations in cancer. *Nucleic acids research*, page gkq929.

Hu, M. *et al.* (2012). Hicnorm: removing biases in hi-c data via poisson regression. *Bioinformatics*, **28**(23), 3131–3133.

Imakaev, M. *et al.* (2012). Iterative correction of hi-c data reveals hallmarks of chromosome organization. *Nature methods*, **9**(10), 999–1003.

Le Dily, F. *et al.* (2014). Distinct structural transitions of chromatin topological domains correlate with coordinated hormone-induced gene regulation. *Genes & development*, **28**(19), 2151–2162.

Lieberman-Aiden, E. *et al.* (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *science*, **326**(5950), 289–293.

Lin, Y. C. *et al.* (2012). Global changes in the nuclear positioning of genes and intra-and interdomain genomic interactions that orchestrate b cell fate. *Nature immunology*, **13**(12), 1196–1204.

Quilez, J. *et al.* (2017). Managing the analysis of high-throughput sequencing data. *bioRxiv*, page 136358.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rao, S. S. *et al.* (2014). A 3d map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, **159**(7), 1665–1680.

Schmitt, A. D. *et al.* (2016). Genome-wide mapping and analysis of chromosome architecture. *Nature Reviews Molecular Cell Biology*.

Serra, F. *et al.* (2016). Structural features of the fly chromatin colors revealed by automatic three-dimensional modeling. *bioRxiv*, page 036764.

Servant, N. *et al.* (2012). Hitc: exploration of high-throughput cexperiments. *Bioinformatics*, **28**(21), 2843–2844.

Sing, T. *et al.* (2005). Rocr: visualizing classifier performance in r. *Bioinformatics*, **21**(20), 7881.

Stadhouders, R. *et al.* (2017). Transcription factors orchestrate dynamic interplay between genome topology and gene regulation during cell reprogramming. *bioRxiv*, page 132456.

Wood, S. N. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **65**(1), 95–114.

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **73**(1), 3–36.

Wu, H.-J. and Michor, F. (2016). A computational strategy to adjust for copy number in tumor hi-c data. *Bioinformatics*, page btw540.

Yaffe, E. and Tanay, A. (2011). Probabilistic modeling of hi-c contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nature genetics*, **43**(11), 1059–1065.

Yan, K. K. *et al.* (2017). HiC-Spector: A matrix library for spectral and reproducibility analysis of Hi-C contact maps. *Bioinformatics*.

Yang, T. *et al.* (2017). Hicrep: assessing the reproducibility of hi-c data using a stratum-adjusted correlation coefficient. *bioRxiv*, page 101386.