OneD: increasing reproducibility of Hi-C samples with abnormal karyotypes

Working with cancer cell lines Hi-C? Stop here 3 min ...

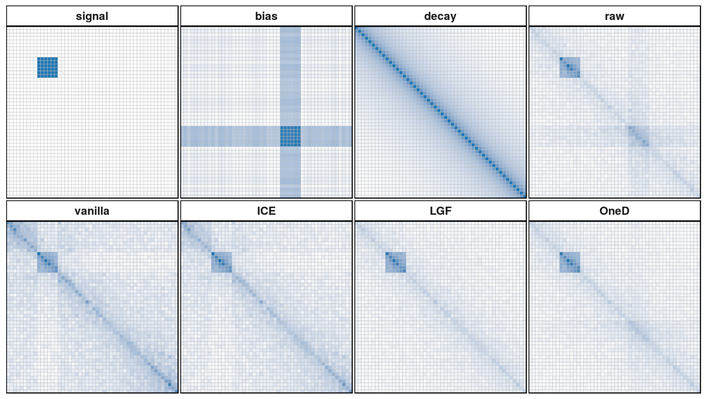
Enrique Vidal[1,2], François le Dily[1,2], Javier Quilez[1,2], Ralph Stadhouders[1,2], Yasmina Cuartero[1,2], Thomas Graf[1,2], Marc A. Marti-Renom[1,2,3,4], Miguel Beato[1,2] and Guillaume J. Filion[1,2]

[1] Gene Regulation, Stem Cells and Cancer Program, Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology (BIST), Dr. Aiguader 88, 08003, Barcelona, Spain [2] Universitat Pompeu Fabra (UPF), Barcelona, Spain [3] CNAG-CRG, Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), Baldiri i Reixac 4, 08028 Barcelona, Spain [4] ICREA, Pg. Llus Companys 23, 08010 Barcelona, Spain

enrique.vidal@crg.eu

Motivation

Hi-C provides information about the spatial organization of the chromatin. It interrogates all possible pairwise interactions between restriction fragments [3]. Several genomic features (E.g. mappability, number of restriction sites, GC content) impact the result [6]. Although there are correction methods that try remove those biases [4], some of them also generate artifacts and/or need large computational resources. Here we present OneD, a bias correction method that improves reproducibility between samples. OneD is fast and scalable both in resolution and sequencing depth.



Schematic example. Top panels depict a true signal (TAD), a genomic bias, the distance decay and the composite result (with Poisson noise) of these three components as a simulation of *raw* data. Bottom panels show the results of applying different correction methods.

Data

We gathered Hi-C samples from two cell lines with abnormal karyotypes (6 *T47D* and 8 *K562*). We created two sets of samples, one per cell line. On each set, we spiked in two samples from the other cell line. Thus, we end up with a pool of samples with different cell types and technical protocols. Here, protocol includes application type (dilution Hi-C vs in-situ Hi-C), restriction enzyme (both four and six base cutters), laboratory and sequencer. This design allows us to define comparisons between samples with different degrees of similarity. All results presented correspond to 100 *Kbp* resolution.

CELL	APPLICATION	RE	LAB	N
T47D	dilution HIC	HindIII	CRG	1
T47D	dilution HIC	HindIII	UMass	2
T47D	dilution HIC	NcoI	CRG	1
T47D	INHIC	DpnII	CRG	2
K562	INHIC	HindIII	4DGU	1
K562	INHIC	MboI	4DGU	1
K562	INHIC	MboI	Baylor	6

Set of samples. RE stands for restriction enzyme and N for number of samples.

Method

$$t_i = \sum_j^n x_{i,j} \sim NB(\lambda_i, \theta)$$

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

$x_{i,j}$: counts of matrix entry i, j

f_k: thin plate regression spline (I.g. `mgcv::s` default)

z_k: genomic feature (E.g. mappability)

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

$$\hat{x}_{ij} = \frac{x_{ij}}{\sqrt{\lambda'_i \lambda'_j}}$$

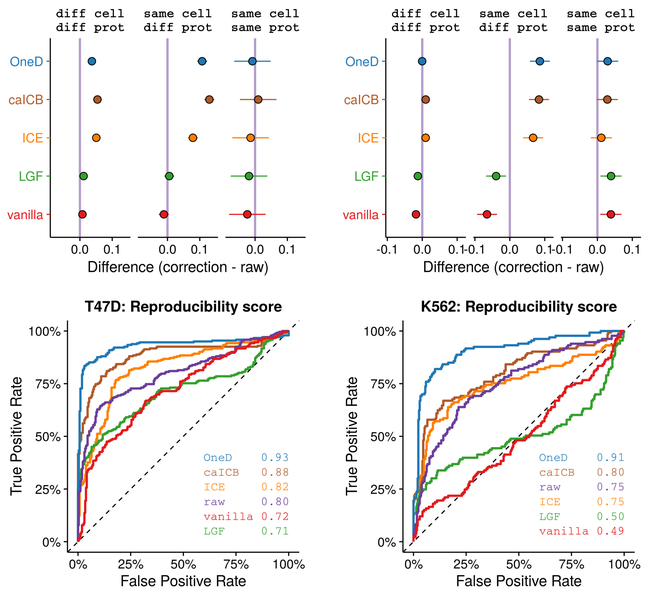
λ'_i: correction associated to bin i

$\hat{x}_{ij}$: corrected counts of matrix entry i, j

Comparisons

We benchmarked OneD, against vanilla, ICE [2], caICB [5] and the Local Genomic Features method (LGF) [1]. We also considered the uncorrected matrices (raw). We used the reproducibility score [7] to estimate the similarity between pairs of samples. Cell type and technical protocol define informative categories to summarize the results. Also, a pair of samples from the same type are expected to be more similar than a pair from different types. Thus, we create ROC curves (and compute AUCs) to test the performance of the methods.

Results



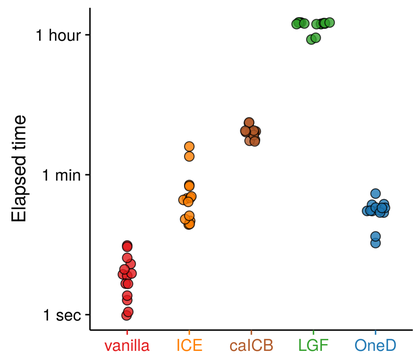
Comparison between samples with aberrant karyotypes. Top panels: Average changes compared to the raw data and corresponding 95 $\%C.I.$. Bottom panels: ROC curves for cell type discrimination. The area under the curve is indicated in the bottom right corner. Left panels correspond to T47D set and right panels to K562 set.

Conclusions

OneD:

- improves reproducibility in Hi-C samples with aberrant karyotype
- is competitive in terms of computational speed
- performs as well as best method in normal karyotypes
- execution time scales well with sequencing depth and resolution
- is suitable for targeted designs (E.g. capture Hi-C)
- is stable at low sequencing depths / high resolutions
- **BONUS!** can be used to estimate copy number

Results supporting grey bullets can be found in the manuscript.



Computing time. One point per sample (Y axis on log scale).

Availability

github code

`github.com/qenvio/dryhic`

NAR paper

`doi.org/10.1093/nar/gky064`

References

[1] M. Hu, K. Deng, S. Selvaraj, Z. Qin, B. Ren, and J. S. Liu. HiCNorm: removing biases in Hi-C data via Poisson regression. *Bioinformatics*, 28(23):3131–3133, 2012.

[2] M. Imakaev, G. Fudenberg, R. P. McCord, N. Naumova, A. Goloborodko, B. R. Lajoie, J. Dekker, and L. A. Mirny. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nature methods*, 9(10):999–1003, 2012.

[3] E. Lieberman-Aiden, N. L. Van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326(5950):289–293, 2009.

[4] A. D. Schmitt, M. Hu, and B. Ren. Genome-wide mapping and analysis of chromosome architecture. *Nature Reviews Molecular Cell Biology*, 2016.

[5] H.-J. Wu and F. Michor. A computational strategy to adjust for copy number in tumor Hi-C data. *Bioinformatics*, page btw540, 2016.

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

[7] K. K. Yan, G. Gurkan Yardimci, C. Yan, W. S. Noble, and M. Gerstein. HiC-Spector: A matrix library for spectral and reproducibility analysis of Hi-C contact maps. *Bioinformatics*, Mar 2017.

Funding

This work was partially supported by the Spanish Ministry of Economy and Competitiveness 'Centro de Excelencia Severo Ochoa 2013-2017' (SEV-2012-0208) and ACER to CRG. R.S. was supported by an EMBO Long-term Fellowship (ALTF 1201-2014) and a Marie Curie Individual Fellowship (H2020-MSCA-IF-2014). We received funding from the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013)/ERC Synergy grant agreement 609989 (4DGenome). The content of this poster reflects only the author's views and the Union is not liable for any use that may be made of the information contained therein.