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

OXFORD

Genome Analysis

OneD: 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], Miguel Beato [1,2] and Guillaume Fillion [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: The three-dimensional conformation of genomes is an essential component of their biological activity. The advent of the Hi-C technology enabled an unprecedented advance in our understanding of genome structures. However, Hi-C is subject to systematic biases that can compromise downstream analyses. Several strategies have been proposed to remove those biases, but none of them addresses the common issue of abnormal karyotypes. Many experiments are performed in cancer cell lines, which typically harbor large-scale copy number variations that create visible defects on the raw Hi-C maps. The consequences of these widespread artifacts on the normalized maps are mostly unexplored.
Results: We observed that current normalization methods perform badly in the presence of large-scale copy number variations, obscuring biological variations and enhancing batch effects. To address this issue, we developed an alternative approach designed to take into account chromosomal abnormalities. The method, called OneD, increases reproducibility among replicates of Hi-C samples with abnormal karyotype, thus significantly outperforming previous methods. In the absence of chromosomal aberrations, OneD fared equally well as state-of-the-art methods, making it a safe choice for Hi-C normalization. OneD is fast and scales well in terms of computing resources for resolutions up to 1 kbp.
Availability: OneD 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

One of the crown achievements of modern biology was to realize that genomes have an underlying three-dimensional structure contributing to their activity (Rowley and Corces, 2016). In mammals, this organization plays a key role in guiding enhancer-promoter contacts (CITE), in V(D)J recombination (CITE) and in X chromosome inactivation (CITE). One of the most significant breakthroughs towards this insight was the development of the high throughput chromosomal conformation capture technology (Hi-C), assaying chromosomal contacts at a genome-wide scale (Lieberman-Aiden *et al.*, 2009). Nowadays, exploring the spatial organization of chromatin has become a priority in many fields and Hi-C has become part of the standard molecular biology toolbox (CITE).

Contrary to the precursor technologies 3C, 4C and 5C (de Wit and de Laat, 2012), Hi-C interrogates all possible pairwise interactions between restriction fragments. However, this does not guarantee that the method has no bias. On the contrary, local genome features such as the G+C content, the availability of restriction enzyme sites and the mappability of the sequencing reads have been shown to impact the results (Yaffe and Tanay, 2011). In addition, the general experimental biases such as batch effects and protocol variation also apply. It is thus important to normalize Hi-C data in order to remove biases and artifacts, so that they are not confused with biological signal.

Several methods have been proposed to remove biases in Hi-C experiments (Schmitt *et al.*, 2016). The first strategy is to model biases

explicitly from a defined set of local genomic features, such as the G+C content and the availability of restriction sites. This approach is used in the method of Yaffe and Tanay (2011) and in Hicnorm by Hu *et al.* (2012). The second strategy is to implicitly correct unknown biases by enforcing some regularity on the data. This approach is used in the ICE method of Imakaev *et al.* (2012), where the total amount of contacts of every bin is imposed to be the same. ICE is currently the most popular method, due in part to its speed.

Neither of these strategies were designed with regard for cell types with karyotypic aberrations, most common in cancer cells in culture. Hi-C is very sensitive to aneuploidy, copy number variations and translocations. Actually, these aberrations have so much influence on the outcome that the artifacts can be used to re-assemble the target genome (Korbel and Lee, 2013). So far the only attempt to address the issue was the correction method caICB by Wu and Michor (2016). However, caICB applies a chromosome-wide correction, effectively excluding the numerous cases of partial aneuploidy and regional copy number variations.

Here we propose *OneD*, a method to correct local chromosomal abnormalities. OneD explicitly models the contribution of known biases via a generalized additive model. The normalized data is more reproducible between replicates and across different protocols. Importantly, OneD is also applicable when cells have a normal karyotype, where it performs as well as the best normalization methods. Finally, the implementation is faster than ICE and scales up to 1 kbp resolution with reasonable computing resources.

2 Methods

2.1 Model

The most common representation of Hi-C data is a contact matrix, obtained by slicing the genome in n consecutive bins of fixed size (the resolution) and computing the number of contacts between each pair of bins. The values are stored in the cells of the contact matrix (x_{ij}), quantifying of the interaction between the tow loci at positions i and j.

Our approach is to model the tally of contacts for each bin, thus reducing the matrix to a one dimension score (hence the name OneD). We assume that the total number of contacts per bin (t_i) can be approximated by a negative binomial distribution. This choice is sensible because the amount of contacts is a discrete variable and because the negative binomial distribution allows for overdispersion. We further assume that the explicit sources of bias have independent contributions to the mean of the distribution for a given bin (λ_i).

Given that this relationship might not be linear (see for instance 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

$$t_i = \sum_j^n x_{ij} \sim NB(\lambda_i, \theta)$$
$$\log(\lambda_i) \propto \sum_k f_k(z_k)$$

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

Once the parameters of the model are estimated, we rescale the estimated means $\{\lambda_i\}$ to obtain a correction vector $\{\lambda'_i\}$ that can be used to compute the corrected counts ($\hat{x}_{ij}$).

$$\lambda'_i = \frac{\lambda_i}{\sum_j^n \lambda_j / n}$$
$$\hat{x}_{ij} = \frac{x_{ij}}{\sqrt{\lambda'_i \lambda'_j}} \quad (1)$$

In line with previous methods (Yaffe and Tanay, 2011; Hu *et al.*, 2012), the features used to fit the model are the local G+C content, read mappability and the content and number of restriction sites. The model and the implementation can be easily extended with any user-provided genomic feature.

2.2 Copy number correction

Briefly, a hidden Markov model with emissions distributed as a Student's t variable is fitted on the corrected total amount of contacts per bin. The model consists of 7 states that correspond to 1, 2, 3, 4, 5, 6 and 8 copies of the target, for a total of 3 emission parameters (a single scaling parameter, a single standard deviation and a single degree of freedom for all the states) and 21 transition parameters.

The model is fitted with the Baum-Welch algorithm (Baum and Petrie, 1966) until convergence, following the detail given by Filion *et al.* (2010). The Viterbi path is then computed and corresponds to the inferred copy number of each bin (c_i).

A correction equal to the square root of the copy number is then applied to the whole matrix. More specifically, each cell is updated to

$$\hat{x}^*_{ij} = \frac{\hat{x}_{ij}}{\sqrt{c_i c_j}}$$

2.3 Data sources

To test the correction of biases, we gathered 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 Hi-C data of different cell types and organisms. The details can be found in Table 1.

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 were carried out in different laboratories, following either the original Hi-C protocol (Lieberman-Aiden *et al.*, 2009) or the newer *in situ* version (Rao *et al.*, 2014), and using different restriction enzymes (DpnII, HindIII, MboI and NcoI).

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

2.4 Data processing

All data were processed though a pipeline based on TADBit (Serra *et al.*, 2016). Briefly, after controlling the quality of fastq files, paired-end reads were mapped in the corresponding reference genome (hg38 and mm10) taking into account the restriction enzyme site. Non-informative contacts were removed applying the following TADBit filters: self-circle, dangling-ends, error, extra dangling-end, duplicated and random-breaks. For more details, see the methods section of Stadhouders *et al.* (2017). In addition, the pipeline is available from the supplementary material published by Quilez *et al.* (2017).

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 carry out the *lgf* and *caib*

Table 1. Hi-C data sets used in this study.

Sample ID	Cell type	Application	RE	Sequencing core	Set(s)	Source
dc3a1e069_51720e9cf	T47D	*in situ* Hi-C	DpnII	CRG	T47D, K562	NA
b1913e6c1_51720e9cf	T47D	*in situ* Hi-C	DpnII	CRG	T47D	NA
dc3a1e069_ec92aa0bb	T47D	*in situ* Hi-C	DpnII	CRG	T47D, K562	NA
HindIII_T0	T47D	dilution Hi-C	HindIII	CRG	T47D	SRR1054341
NcoI_T0	T47D	dilution Hi-C	NcoI	CRG	T47D	SRR1054343
ENCLB758KFU	T47D	dilution Hi-C	HindIII	UMass	T47D	ENCLB758KFU
ENCLB183QHG	T47D	dilution Hi-C	HindIII	UMass	T47D	ENCLB183QHG
HIC069	K562	*in situ* Hi-C	MboI	Baylor	T47D, K562	SRR1658693
HIC070	K562	*in situ* Hi-C	MboI	Baylor	T47D, K562	SRR1658694
HIC071	K562	*in situ* Hi-C	MboI	Baylor	K562	SRR1658695,SRR1658696
HIC074	K562	*in situ* Hi-C	MboI	Baylor	K562	SRR1658701,SRR1658702
b7fa2d8db_bfac48760	B-cell	*in situ* Hi-C	DpnII	CRG	mm10	GSE96611
fc3e8b36a_7bf1bf374	ES-cell	*in situ* Hi-C	DpnII	CRG	mm10	GSE96611
b7fa2d8db_7284b867a	B-cell	*in situ* Hi-C	DpnII	CRG	mm10	GSE96611
fc3e8b36a_38bfd1b33	ES-cell	*in situ* Hi-C	DpnII	CRG	mm10	GSE96611
b7fa2d8db_58e812fc2	B-cell	*in situ* Hi-C	DpnII	CRG	mm10	GSE96611
fc3e8b36a_c990a254e	ES-cell	*in situ* Hi-C	DpnII	CRG	mm10	GSE96611
b7fa2d8db_73f11d923	B-cell	*in situ* Hi-C	DpnII	CRG	mm10	GSE96611
fc3e8b36a_4bf044f18	ES-cell	*in situ* Hi-C	DpnII	CRG	mm10	GSE96611
GSM987817	B-cell	dilution Hi-C	HindIII	UCSC	mm10	SRR543428-SRR543431
GSM987818	B-cell	dilution Hi-C	HindIII	UCSC	mm10	SRR543432-SRR543442
GSM862720	ES-cell	dilution Hi-C	HindIII	UCSC	mm10	SRR443883-SRR443885
GSM862721	ES-cell	dilution Hi-C	HindIII	UCSC	mm10	SRR400251-SRR400256
GSM862722	ES-cell	dilution Hi-C	NcoI	UCSC	mm10	SRR443886-SRR443888

corrections respectively. All the results are based on a resolution of 100 kbp, but we found no major differences for different values (not shown).

2.5 Comparison of Hi-C matrices

There is no universally accepted standard to compare Hi-C matrices. The simplest metric is the Spearman correlation applied to intra-chromosomal contacts up to a given distance (5 Mb in what follows). The second option is to measure the similarity of observed over expected contacts via the Pearson correlation up to a given distance range. Compared to the first, this metric gives more weight to changes occurring away from the diagonal. The third option is to compute a correlation per distance stratum and then obtain a stratum-adjusted correlation coefficient (SCC) as defined in Yang *et al.* (2017). Finally, the last option, proposed by Yan *et al.* (2017) is to measure the Pearson correlation between the last eigenvectors of the Laplacian of the Hi-C matrix. This approach borrows the concepts of spectral clustering (Von Luxburg, 2007) and amounts to comparing high level features of the matrix.

We defined three data sets to measure experimental reproducibility after normalization: The first contained the samples from T47D plus two samples from K562, the second contained the samples from K562 plus two samples from T47D, the third contained all the mouse samples (see Table 1 for details). Given a set of experiments and a metric, we first computed all pairwise combinations between experiments and then classified the comparisons according to the characteristics of each pair (cell type, protocol, batch and treatment).

We benchmarked ***oned***, against ***vanilla***, ***ice*** (Imakaev *et al.*, 2012), ***caib*** (Wu and Michor, 2016) and ***lgf*** (Hu *et al.*, 2012; Servant *et al.*, 2012). The first three methods correct biases implicitly, whereas the fourth method does it explicitly.

To measure the gain or loss of similarity upon normalization, we compared raw matrices to obtain a baseline. The differences with this baseline were estimated using a linear mixed model fitted with the `lmer` function of the `lme4` R package (Bates *et al.*, 2015), where the fixed effect

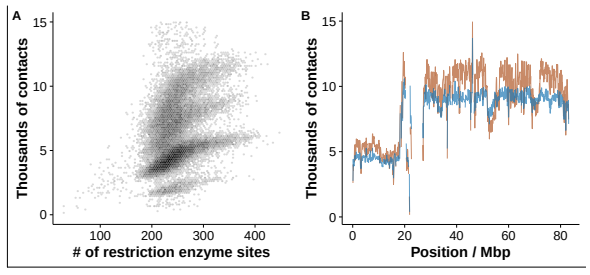

Fig. 1. A. Non linear relationship between the number of restriction sites per bin and the total number of contacts per bin in T47D. B. Total number of contacts per bin in chromosome 17 of T47D. Raw (brown) and corrected (blue) signals are shown. The region that corresponds to the first 20 Mbp of the chromosome is present in single copy in this clone of T47D, explaining that the signal is approximately half compared to the rest of the chromosome. Also notice that the raw signal is noisier than the corrected one.

was the normalization method and the random effect was the chromosome. Receiver operating characteristic (ROC) curves were then computed using `ROCR` package (Sing *et al.*, 2005).

3 Results

3.1 Correction method

The principle of OneD is to explicitly model Hi-C biases on a single variable: the total amount of contacts for each bin of the matrix (see 2.1). The reason for this choice is that the total amount of contacts is approximately proportional to the local copy number. For instance, a duplicated region in a diploid genome will show on average a 50% increase in the number of contacts. Discontinuities of the amount of contacts thus correspond to changes of the copy number.

Other biases affect the total amount of contacts in a continuous but not necessarily linear way. Figure 1A shows the relationship between the

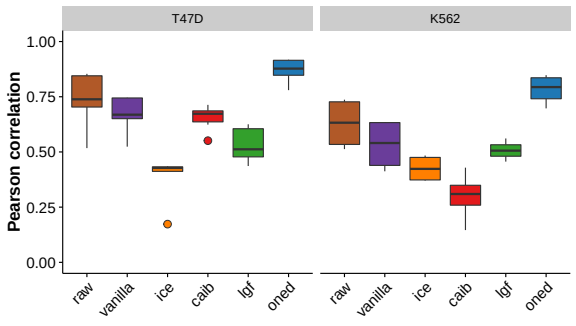

Fig. 2. Pearson correlation between total number of contacts per bin and independent copy number estimation (COSMIC) for each of the methods compared. Left panel T47D breast cancer cell line, right panel K562 leukemia cell line. The new proposal (in blue) outperforms the rest of alternatives.

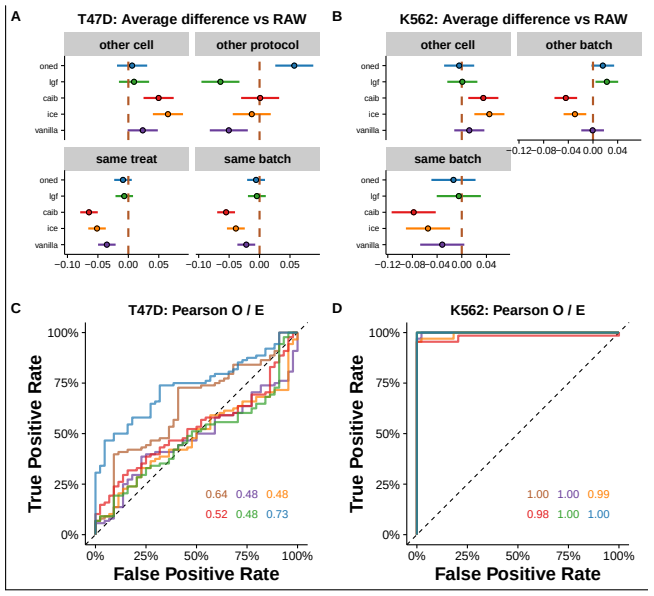

Fig. 3. Results of the comparison between samples with aberrant karyotype. A and B. Average changes compared to raw on the T47D and K562 sets. The bars represent 95% confidence intervals centered on the mean difference of the correlation score between a given correction method and the raw data. The brown dashed line indicates the value of the average score on raw matrices (set to 0). C and D. ROC curves on the T47D and K562 sets. The areas under the curve are indicated in the bottom right corner. The color code is the same as panels A and B. The brown lines correspond to raw matrices. All results in this figure are based on Pearson correlations between the observed over expected counts.

amount of contacts and the number of restriction enzyme sites in T47D, a breast cancer cell line with aberrant and unstable karyotype. Four clouds are visible. Each corresponds to a copy number between one and four. In all of them, the relationship flattens as the number of restriction sites increases. For this reason, OneD fits a non-linear relationship between the total amount of contacts and the known sources of bias (by default the G+C content, the number of restriction sites and the mappability of the reads).

The biases are estimated genome-wide and each cell of the matrix is then corrected using equation (1). Note that the corrected amount of contacts is still proportional to the copy number. The correction thus represents a more stable and precise estimation of the copy number. Figure 1B shows the corrected number of contacts along chromosome 17 of a T47D. OneD greatly reduces the wiggling of the total amount of contacts.

We tested the validity of this approach against the Catalogue Of Somatic Mutations In Cancer (COSMIC, Forbes *et al.*, 2010). Figure 2 shows the Pearson correlation between the corrected number of contacts and the copy number estimation for both T47D and K562 cell lines. Similar results were obtained using Spearman correlation (Supplementary Figure 2). All the methods except OneD decrease the agreement with the copy number because they partially correct this bias. In contrast, OneD enhances the conformity of the signal with the copy number. Not correcting for variable copy number at that stage may seem to defeat the purpose of a normalization method, but we will see below how this can lead to better performance.

3.2 Aberrant karyotypes

We first benchmarked the performance of OneD on biological samples with an aberrant karyotype. A good normalization method should increase the similarity between biological replicates by reducing irrelevant experimental variation, such as batch effects, laboratory of origin and protocol variations. Similarly, a good normalization should decrease the similarity between samples that received a different treatment in order to enhance the biological differences.

We assembled two Hi-C data sets obtained from T47D and K562, two cancer cell lines with aberrant karyotypes. In each set we spiked two samples from the other cell line (see Table 1) in order to introduce biological variability. We measured the Pearson correlation of observed over expected counts to compare matrices before and after normalization by different methods (see 2.5). This gave an indication of the impact of a given normalization method. The results are summarized in Figure 3.

The *caib* and *ice* methods increased the similarity between the different cell lines (Figures 3A and 3B and Supplementary Figure 3). This is an undesirable effect, as it obscures the biological variability. In the same vein, these methods decreased the similarity between samples that received the same treatment (Figure 3A), suggesting that the normalization process is detrimental to the biological signal in these two cases. The method *vanilla* followed the same trend but to a lesser extent, consistent with the fact that it consists of a single *ice* iteration.

OneD was the only method to increase the similarity between experiments carried out on the same material but with a different protocol (Figure 3A). In these conditions, *lgf* was decreasing the similarity. Taken together, the results show that OneD enhanced biological variation while decreasing experimental variation.

An important application of normalization methods in experimental setups is to identify outliers. We thus investigated the capacity of the different methods to help identify the samples from the other cell type spiked in the data set. We interpreted the pairwise comparison scores as classifier scores and summarized the results with a ROC curve (Figures 3C and D).

All the methods, including the absence of normalization, succeeded in recognizing the T47D outliers among the K562 samples, but recognizing the K562 outliers among the T47D samples proved more challenging. OneD increased the discrimination power compared to the raw matrices, but all the other methods decreased it. Actually, they performed little better than random on this task.

Using the other metrics described in 2.5 yielded similar results (Supplementary Figures 3, 4 and 5). Note that Spearman correlation of contacts presented the worst performance for all scenarios, and it thus seems to be a poor choice as a metric to compare HiC matrices.

Taken together, these results show that OneD enhances biological variation and reduces experimental noise on samples with an aberrant karyotype.

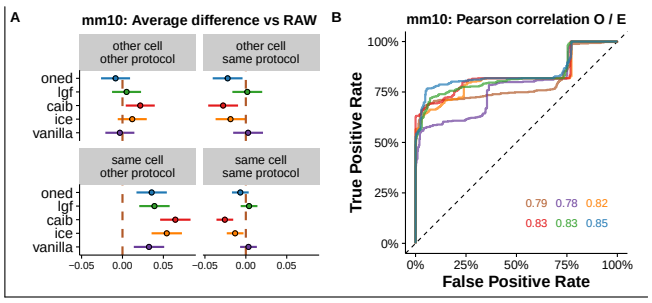

Fig. 4. Results of the comparison between samples with normal karyotype. A. Average changes compared to raw on the mouse data set. The bars are as in B. ROC curves on the mouse data set. The legend is as in Figure 3.

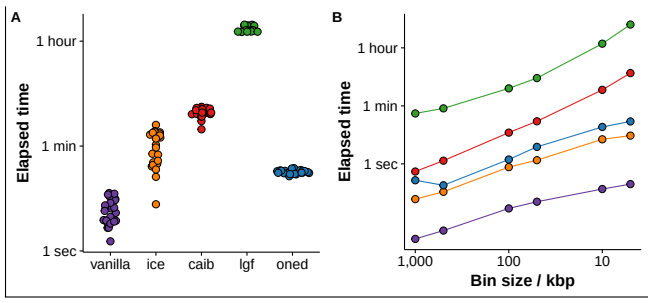

Fig. 5. Computing time of the bias correction methods. A. Total time for the entire genome at 100 kbp resolution. Each dot corresponds to one sample. The only method faster than ours under performs in all sample comparisons. B. Time to correct a reduced genome (chr19-22) of one sample at different resolutions. Note the logarithmic scale on the y-axis on both panels.

3.3 Normal karyotypes

Does the performance of OneD with aberrant karyotypes comes at the cost of decreased performance with normal karyotypes? To address this question, we assembled another data set comprised of normal mouse B cells and embryonic stem (ES) cells. The cell types were pooled in almost equal proportions (see Table 1) and the same tests as above were performed.

In these conditions, variations of the protocol had a strong effect on the impact of the different normalization methods (Figure 4A). For instance, *caib* and *ice* increased the similarity when the protocols were different, but decreased it when the protocols were the same. The effect was stronger when comparing identical cell types, but the same trend appeared when comparing different cell types, indicating that these methods may enhance or reduce biological variation, depending on the context. Once more, *vanilla* followed the same trend as *ice* to a lesser extent. The *lgf* method increased the similarity when comparing the same cells with different experimental protocols, and had little to no effect in the other cases. This indicates that *lgf* is a safe choice in this case.

OneD decreased the similarity between different cell types when using the same protocol and increased it between identical cell types when using different protocols. In this other two cases, it had little effect. In summary, OneD never enhanced the experimental noise and it reduced it in one more case than *lgf*.

When interpreting the similarity scores as classification scores, we observed that all the methods could identify approximately 50% of the biological pairs, after which their performance diverged (Figure 4B). OneD achieved the highest area under the curve on this problem, even though it was not always above the others methods at every position of the graph.

Using other metrics to compare matrices gave similar results (Supplementary Figures 6 and 7). OneD was always among the top scoring methods. In these conditions, Spearman correlation of contacts again appeared as as the worst comparison metric as it showed a lower performance for all the normalization methods.

Taken together, these results indicate that OneD performs as well as the best normalization methods even with normal karyotypes.

3.4 Speed

Finally, we compared the computational speed of the different normalization methods. One of the main reasons for the broad acceptance of *vanilla* and *ICE* as standards is the speed of the available implementation (Imakaev *et al.*, 2012). This is even more significant as current explicit methods (Servant *et al.*, 2012) are much slower in comparison.

Unlike the other methods, OneD corrects a single variable instead of the whole matrix. It uses only the total number of contacts per bin and thus estimates the model parameters much faster than previous explicit methods. We measured the running time of the different tools on a 3.3 GHz machine with 62 GB RAM, always using the default parameters. Figure 5A shows the running times of the different methods on the samples described in Table 1 at 100 kbp resolution. The fastest method is *vanilla* and the slowest is *lgf*, with an over 100-fold span between the two. OneD also scales well in terms of bin size, as it can be seen in Figure 5B, where the different corrections were applied to a reduced part of the genome at increasing resolutions.

OneD is the second fastest method and it always ran in less than 1 min in the conditions of the benchmark. In all cases, the memory footprint was less than 300 MB. Interestingly, the running time of OneD is much more homogeneous than that of the other methods. The reason is that the size of the regression problem to be solved by the `mgcv` package is always the same at fixed resolution

Taken together, these results show that OneD is very competitive in terms of computational speed compared to existing methods.

4 Discussion and Conclusions

Here we introduced OneD, a fast computational method to normalize Hi-C matrices. OneD was developed ground up to address the need to normalize data collected on samples with aberrant karyotypes, but it applies seamlessly to the case of normal karyotypes. We showed that OneD performs significantly better than other methods when the cells present karyotypic aberrations (Figure 3), and that it performs equally well as the best methods otherwise (Figure 4). We also showed that OneD is on average faster than *ice*, which makes it very competitive from the point of view of computational speed.

The originality of OneD lies in that it projects all the biases onto a single variable: the total amount of contact per bin. This allows greater running speed, while preserving a good performance on samples with karyotypic aberrations. One of the reasons that OneD is better able to highlight the biological distinctions between samples is that it does not correct the copy number but enhances it (Figures 1 and 2).

This raises the question whether variations of the copy number constitute a biological signal or an artifact. Note that in the presence of translocations, deletions and large scale rearrangements, correcting the copy number is equally faulty as not correcting it. Indeed, in this case the reference genome on which the matrix is drawn is incorrect and changing the values of the pixels cannot account for this information.

Depending on the intention of the user, the effect of the copy number should either be kept or removed. This is why OneD does not perform the correction by default, but allows the user to obtain a diploid-equivalent Hi-C map computed from a hidden Markov model. The resulting matrices have a near constant amount of contacts per bin, but the artifacts caused

by the misfit between the genome of the sample and the reference genome are still present (for instance, the artifacts caused by large scale inversions are not changed in any way).

Overall, OneD constitutes a safe choice to normalize Hi-C matrices. If the karyotype of the sample is aberrant, it enhances the biological variation. If not, it performs at least equally well as other methods with an advantage in terms of computational speed.

Acknowledgements

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

Funding

This work was supported by the Spanish Ministry of Economy and Competitiveness, Centro de Excelencia Severo Ochoa 2013-2017, SEV-2012-0208, ERC Synergy Grant 609989 and...

References

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

Baum, L. E. and Petrie, T. (1966). Statistical inference for probabilistic functions of finite state markov chains. *The Annals of Mathematical Statistics*, **37**(6), 1554–1563.

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

de Wit, E. and de Laat, W. (2012). A decade of 3c technologies: insights into nuclear organization. *Genes & development*, **26**(1), 11–24.

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

Filion, G. J. *et al.* (2010). Systematic protein location mapping reveals five principal chromatin types in drosophila cells. *Cell*, **143**(2), 212–224.

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.

Korbel, J. O. and Lee, C. (2013). Genome assembly and haplotyping with hi-c. *Nature biotechnology*, **31**(12), 1099.

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.

Rowley, M. J. and Corces, V. G. (2016). The three-dimensional genome: principles and roles of long-distance interactions. *Current opinion in cell biology*, **40**, 8–14.

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.

Von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and computing*, **17**(4), 395–416.

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.