*Published online ** ***** ****

*Nucleic Acids Research, 2017, Vol. **, No. ** **1–8*
doi:10.1093/nar/gkn000

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 [1,2,3], 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 and [2]Universitat Pompeu Fabra (UPF), Barcelona, Spain and [3]CNAG-CRG, Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), Baldiri i Reixac 4, 08028 Barcelona, Spain and [4]ICREA, Pg. Lluís Companys 23, 08010 Barcelona, Spain

Received January 1, ****; Revised February 1, ****; Accepted March 1, ****

ABSTRACT

The three-dimensional conformation of genomes is an essential component of their biological activity. The advent of the Hi-C technology enabled an unprecedented progress 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 the issue of abnormal karyotypes received little attention. 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. We observed that current normalization methods are not robust to the presence of large-scale copy number variations, potentially obscuring biological differences 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, outperforming previous methods significantly. On normal karyotypes, 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 5 Kbp.

INTRODUCTION

One of the crown achievements of modern biology was to realize that genomes have an underlying three-dimensional structure contributing to their activity (Dekker and Mirny, 2016, Pezic *et al.*, 2017, Rowley and Corces, 2016). In mammals, this organization plays a key role in guiding enhancer-promoter contacts (De Laat and Duboule, 2013), in V(D)J recombination (Choi and Feeney, 2014) and in X chromosome inactivation (Galupa and Heard, 2015). A significant breakthrough 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 (Dekker *et al.*, 2013).

Contrary to the precursor technologies 3C, 4C and 5C (de Wit and de Laat, 2012, Dekker *et al.*, 2002, Dostie *et al.*, 2006, Simonis *et al.*, 2006), 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 to general experimental biases such as batch effects. 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. 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 condition on the data. This approach is used in the Iterative Correction and Eigenvector decomposition method (*ICE*) of (Imakaev *et al.*, 2012), whereby 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 for cell types with karyotypic aberrations, most common in cancer. Yet, Hi-C is very sensitive to aneuploidy, copy number variations and translocations. Actually, these aberrations have so much influence on the outcome that they can be used as signature to re-assemble the target genome (Korbel and Lee, 2013). An additional complication is that karyotypic aberrations are not

*To whom correspondence should be addressed. Tel: +34 93 316 01 15; Fax: +34 93 316 00 99; Email: enrique.vidal@crg.eu

experimental biases, so it is unclear whether they should be corrected at all or be considered part of the biological signal.

So far, the only attempt to address the issue was the chromosome-adjusted Iterative Correction Bias method (*caICB*) of (Wu and Michor, 2016). However, *caICB* applies a uniform chromosome-wide copy number 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 in Hi-C experiments. *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 efficient when cells have a normal karyotype, where it performs as well as the best normalization methods. Finally, the implementation is as fast as *ICE* and it scales up to 5 Kbp resolution with reasonable computing resources.

MATERIALS AND METHODS

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 the interaction between the two 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 the default arguments of the `mgcv::s` and `mgcv::nb` functions (Wood, 2011) of the R software (R Core Team, 2017).

Once the parameters of the model are determined, the estimated means $\{\lambda_i\}$ are rescaled 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 (Hu *et al.*, 2012, Yaffe and Tanay, 2011), the default features used to fit the model are the local G+C content, the read mappability and the number of restriction sites. The model and the implementation can be modified or extended with any user-provided genomic features.

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 ($\hat{t}_i = \sum_j^n \hat{x}_{ij}$). The model consists of 8 states that correspond to 1, 2, 3, 4, 5, 6, 7, and '8 or more' copies of the target, for a total of 4 emission parameters (a single position parameter for states 1-7, a position parameter for state '8 or more', a single standard deviation for all the states and a single degree of freedom for all the states) and 56 transition parameters.

The model is fitted with the Baum-Welch algorithm (Baum and Petrie, 1966) until convergence, following a previously described implementation (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

$$\hat{x}^*_{ij} = \frac{\hat{x}_{ij}}{\sqrt{c_i c_j}}. \quad (2)$$

Data sources

To test the correction of biases, we gathered a set of published (Dixon *et al.*, 2012, ENCODE Project Consortium, 2012, Le Dily *et al.*, 2014, Lin *et al.*, 2012, Rao *et al.*, 2014, Stadhouders *et al.*, 2017) and unpublished Hi-C data of different cell types and organisms (Supplementary Table 1).

We used several experiments comprising T47D breast cancer cell lines (6 samples), K562 leukemia cell lines (8 samples), both with aberrant karyotypes; mouse primary B cells (6 samples) and ES cells (7 samples), both with normal diploid karyotypes; GM12878 B-lymphoblastoid cells (58 samples), with diploid karyotype; BT474, MCF10, MCF-7 and SKBR3 breast cell lines (1 sample each). 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, MspI, MboI and NcoI). We also used array-based copy-number segmentation of the two cell lines obtained from the COSMIC database (Forbes *et al.*, 2010).

All data were processed through a pipeline based on TADbit (Serra *et al.*, 2017). Briefly, after controlling the quality of FASTQ files, paired-end reads were mapped to 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-ends`, `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 (available at http://www.github.com/qenvio/dryhic) to efficiently create sparse representations of contact matrices and further apply *vanilla*, *ICE*, *KR* and *oneD* corrections. `HiTC` (Servant *et al.*, 2012) and `HiCapp` (Wu and Michor, 2016) were used to carry out the *LGF* and *caICB* corrections respectively. All the results are based on a resolution of 100 Kbp, but we found no major differences at different resolutions (not shown).

Simulations

Simulations were based on altering four experimental Hi-C samples obtained from diploid mouse cell lines (b7fa2d8db_bfac48760, fc3e8b36a_7bf1bf374, GSM987817 and GSM862720), either B cell or ES cell, and from either CRG or UCSC. The simulation strategy was the following: First we selected uniformly at random the amount of copy number break points (from 3 to 10). We then placed the break points along chromosomes 18 and 19 uniformly at random. Next, we assigned a copy number (2, 3, 4 or 10 with equal probability) to each segment delimited by the break points. We computed the outer product of this simulated copy number profile and multiplied it element-wise with the original contact matrices of chromosomes 18 and 19. The resulting matrices were used as input for correction methods. For each simulation (100 in total), we measured the pairwise reproducibility score defined by Yan *et al.* (2017) and computed the average for pairs from the same cell type minus the average for pairs from different cell types.

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 Mbp 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 by (Yang *et al.*, 2017). Finally, the "reproducibility score" proposed by Yan *et al.* (2017) sums the distances between the leading twenty 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 five 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, the fourth contained all GM12878 samples, the fifth contained one sample of each breast cells (see Supplementary 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).

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 was the normalization method and the random effect was the chromosome. To test the ability of the different methods to separate samples from different cell origin we generated receiver operating characteristic (ROC) curves using the similarity metric (e.g. reproducibility score) as the classifier score and the relative cell type (i.e. same or different cell type) as the binary classification. ROC curves were computed using the `ROCR` package (Sing *et al.*, 2005) for the three first sets (those including samples of different cell types).

RESULTS

Bias correction in Hi-C experiments

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, also referred to as the *contact profile*. 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.

Experimental biases affect the contact profile in a continuous but not necessarily linear way. Figure 1A shows the relationship between the amount of contacts and the number of restriction enzyme sites in T47D, a breast cancer cell line with an aberrant karyotype. Four clouds are visible. Each corresponds to sequences present in one to four copies. In all of them, the relationship flattens as the number of restriction sites increases. To capture this relationship, *OneD* fits a non-linear model 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 experimental biases are estimated genome-wide and each cell of the Hi-C matrix is then corrected using equation (**1**). Note that the corrected contact profile is still proportional to the copy number. In cancer cell lines, the contact

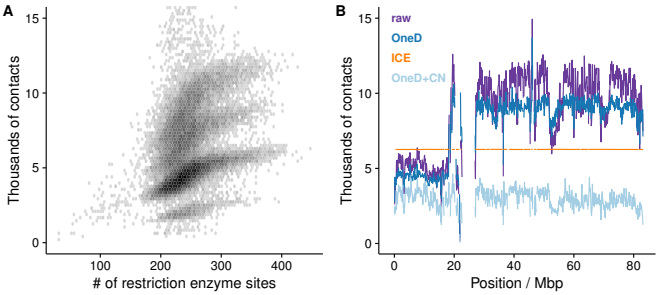

Figure 1. Modeling biases from the contact profile. A. Non linear relationship between the number of restriction sites and the total number of contacts per bin in T47D. B. Total number of contacts per bin on chromosome 17 of T47D. The purple line represents the raw signal, the dark blue line represents the signal after OneD bias correction, the light blue line represents the signal after OneD bias and copy number (CN) correction and the orange line represents the signal after ICE bias correction. The long arm of chromosome 17 (the region corresponding to 20-80 Mbp) is present in four copies, explaining that the signal is about twice higher than for the short arm.

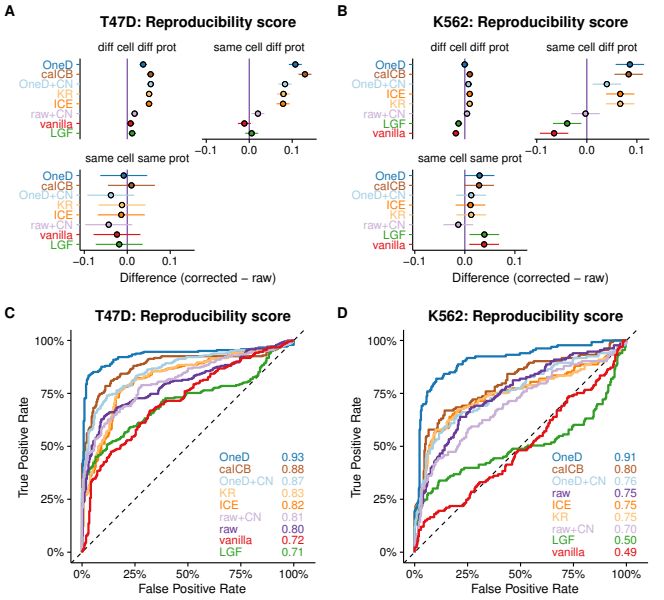

Figure 2. Removing biases from Hi-C on aberrant karyotypes. A and B. Average changes compared to the raw data on the T47D and K562 data sets. The bars represent 95% confidence intervals centered on the mean difference of the reproducibility score between a given correction method and the raw data. Upper left panels: Different cell type samples processed using different protocols. Upper right panels: Same cell type samples processed using different protocols. Lower left panels: Same cell type samples processed using the same protocol. 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 in panels A and B.

profile corrected by *OneD* is highly correlated to the copy number estimated from an external source (Supplementary Figure 1). Figure 1B shows the corrected contact profile along chromosome 17 of T47D. *OneD* greatly reduces the wiggling of the contact profile (dark blue line).

Optionally, *OneD* allows the user to also correct the Hi-C signal for the copy number. In this case, a hidden Markov model is fitted on the corrected contact profile in order to infer the local copy number. Each cell of the Hi-C matrix is then further corrected using equation (**2**), *i.e.* it is divided by the square root of the inferred copy number. In essence, this approach flattens the contact profile so that it is no longer proportional to the local copy number. The light blue line in figure 1B shows the result of this process. After this correction, the contact profile fluctuates around a genome-wide constant value.

In what follows, we benchmarked *OneD* and *OneD* with copy number correction (*OneD+CN*) against *vanilla*, *ICE* (Imakaev *et al.*, 2012), the *KR* algorithm (Knight and Ruiz, 2013), *caICB* (Wu and Michor, 2016) and the Local Genomic Features method (*LGF*, Hu *et al.*, 2012, Servant *et al.*, 2012). The first four methods correct biases implicitly, whereas the fifth method does it explicitly.

We used four metrics to compare Hi-C matrices (see Materials and Methods). For consistency and concision, the results based on the reproducibility score of Yan *et al.* (2017) are shown in the main figures, and the results based on the other metrics are shown in the Supplementary Material.

Aberrant karyotypes

We first benchmarked the performance of our approach on biological samples with an aberrant karyotype. A good normalization method should increase the similarity between biological replicates by reducing irrelevant experimental variance, such as batch effects, laboratory of origin and protocol variations. Likewise, a good normalization should increase the contrast between different samples to enhance the biological differences.

We collected multiple Hi-C data sets obtained from T47D and K562 cells, both with aberrant karyotypes. The first pool, referred to as the T47D data set contained six T47D samples and two K562 samples, the second, referred to as the K562 data set contained eight K562 samples and two T47D samples. We compared matrices before and after normalization by different methods (see Materials and Methods, distributions shown in Supplementary Figures 2 and 3). This gave an indication of the impact of a given normalization method on the biological variation (differences between samples from T47D and K562) and on the technical variation (differences among samples from the same cell type). The results are summarized in Figure 2 and Supplementary Table 3.

On the T47D data set, the methods that increased the reproducibility between identical cell types also increased it between different cell types (Figures 2A), meaning that none of them enhanced exclusively the biological signal. However, the difference was most marked for *OneD* and *caICB*. On the K562 data set (figure 2B), both methods increased the reproducibility between identical cell types but not between different cell types , which is the desired behavior of a correction method. On both data sets, *OneD+CN*, *ICE* and *KR* gave inferior results, while *raw+CN* (copy number correction alone), *LGF* and *vanilla* were not competitive. The conclusions were unchanged when using other metrics to compare Hi-C matrices (Supplementary Figures 4, 5 and 6).

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 2C and D). *OneD* had the largest area under the curve in both tests, and the corresponding ROC curve was above the others throughout. Using other metrics to compare Hi-C matrices yielded similar results (Supplementary Figures 4, 5 and 6).

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

Normal karyotypes

Does the performance of *OneD* on aberrant karyotypes come at the cost of decreased performance on normal karyotypes? To address this question, we collected data from mouse B cells and embryonic stem (ES) cells, both with a normal karyotype. The cell types were pooled in almost equal proportions (see Supplementary Table 1) and we performed the same tests as above using the same metrics (distributions shown in Supplementary Figure 7)

In this test, the reproducibility between different cell types was barely affected (Figure 3A). All the methods increased the reproducibility between identical cell types (except one, see below), but only when the experimental conditions were different. This means that all the methods were able to remove some technical biases. *ICE* and *KR* showed the strongest increase in reproducibility, but *OneD*, *OneD+CN*, and *caICB* had competitive performance. *vanilla* and *LGF* were less competitive, and *raw+CN* (copy number correction alone) performed poorly, because it is equivalent to no normalization on diploid cells. Other metrics gave similar results (Supplementary Figures 8, 9 and 10, Supplementary Table 3).

As above, we used the reproducibility score for classification and compared the tools with a ROC curve (Figure 3B). The performance of all the tools were high and less variable than for aneuploid cell lines. *OneD* achieved the highest area under the curve on this problem, but with a small margin over *KR* and *ICE*. The performance of *OneD+CN* was the same as that of *OneD* because copy number correction has no effect on diploid cell lines. Using other metrics to compare matrices gave similar results (Supplementary Figures 8, 9 and 10).

We took advantage of the wide range of protocols used in the GM12878 set to further investigate the performance of each method. The differences between methods were minor and *OneD* showed the greatest improvement in reproducibility (Supplementary Figure 11).

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

Copy number correction

The results so far suggest that the copy number correction lowers the performance of *OneD*. A possible limitation in Hi-C is the absence of ground truth. We thus used simulations to generate matrices with defined karyotypic aberrations (see Materials and Methods) and tested the capacity of the correction methods to distinguish different cell types (Supplementary Figure 12). On this test, *LGF* and *OneD* performed best. In comparison, *OneD+CN* performed less

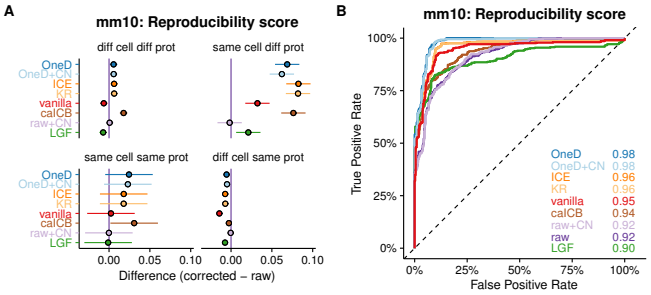

Figure 3. Removing biases from Hi-C on normal karyotypes. A. Average changes compared to the raw data on the mouse data set. Upper left panel: Different cell type samples processed using different protocols. Upper right panel: Same cell type samples processed using different protocols. Lower left panel: Same cell type samples processed using the same protocol. Lower right panel: Different cell type samples processed using the same protocol. B. ROC curves on the mouse data set. The legend and color code are as in Figure 2.

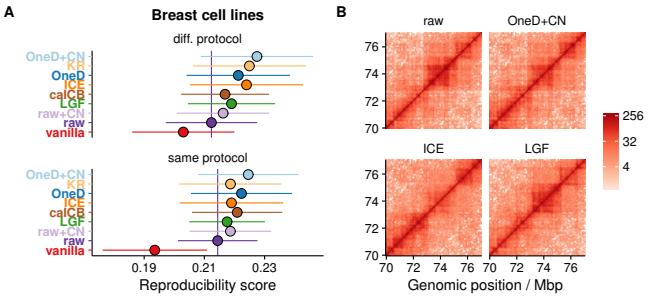

Figure 4. Copy number correction. A. Removing biases from Hi-C on breast cancer cells. The reproducibility score between samples from the same breast cancer cell type was computed after normalization. The vertical bar shows the reproducibility score of the raw data. Shown are comparisons for different (top) or identical (bottom) experimental conditions. B. Detail of a Hi-C matrix normalized with different methods. The central portion has an increased copy number, which affects normalization. *ICE* fades it away, *LGF* enhances it and *OneD* reduces the signal by about half.

well, suggesting that correcting the copy number attenuates the biological signal. This is corroborated by the fact that the copy number correction alone has a lower performance than no correction at all. Thus, copy number correction can indeed have side effects that lower the biological signal.

However, those conclusions are based on data sets where the karyotype is uniform. One sometimes needs to compare samples from cells with different karyotype, for instance in cancer samples where the karyotypes may change through time. In such conditions, copy number correction can remove a feature that is considered irrelevant in order to highlight other biological differences.

To test this idea, we collected a Hi-C data set from breast cancer cell lines and measured the reproducibility score between identical cell types after normalization with the different tools. In this case, we found that *OneD+CN* outperformed *OneD* and gave the highest overall reproducibility (Figure 4A). Even though the variability is high in this case, this suggests that the karyotypes of these cells have diverged, so that the copy number correction performed by *OneD* increased the reproducibility between samples from the same cell type.

Copy number correction is also useful to give euploid-equivalent representations of samples from aberrant karyotypes. Figure 4B shows an example where a region at the center of the matrix is duplicated. *ICE* overcompensated the original bias and faded the signal almost entirely. Concomitantly, the signal at the bottom left of the matrix was enhanced and showed a structure that was not visible in the raw data. On the contrary, *LGF* strengthened the central region and the diagonal. *OneD* reduced the level of the central portion by a factor 2 approximately, but did not otherwise distort the main features of the region.

Besides matrix correction, copy number estimation could have a potential stand-alone utility. For instance, we applied *OneD+CN* to a T47D sample (dc3a1e069_51720e9cf) at 10 Kbp and compared the correlation with the array-based copy number. The correlation increased from .80 for the raw totals, to .88 for the *OneD* corrected totals to .92 for the estimated CN. We provide the table of the estimated start and end points as Supplementary Table 2.

In summary, *OneD+CN* can improve the reproducibility of samples with unstable karyotypes, and it can be used to obtain an euploid-equivalent normalized matrix.

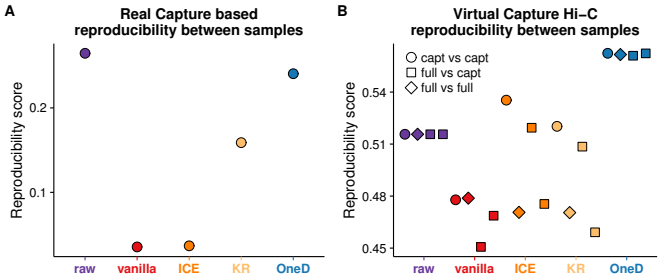

Figure 5. Bias removal on Capture Hi-C data A. Performance on the Capture Hi-C data. Two replicates centered on a 6 Mbp domain on chromosome 6 in T47D cells were normalized by the indicated tools and their reproducibility score was computed. B. Performance on virtual Capture Hi-C data. Two Hi-C experiments were post-processed to produce an output similar to the Capture Hi-C experiment depicted in panel A. The signal outside the domain on chromosome 6 was removed and the data were normalized (capt), or the data were normalized before removing the signal outside the domain on chromosome 6 (full). Four matrices per method were generated and the reproducibility score between the pairs from different replicates were computed. Note that the scores are identical for the raw data because in this case the matrices of the same replicate are identical.

Capture Hi-C

Matrix-balancing methods (here *ICE*, *vanilla* and *KR*) are specifically tailored for Hi-C because it is the dominant 3C-derived technology. As a consequence, their performance is typically lower when they are used for variants such as Capture Hi-C (Jäger *et al.*, 2015). In this case, the contacts with one or more regions of interest are enriched through direct purification, essentially erasing all the other contacts from the Hi-C matrix. The lack of homogeneity of the signal in such technologies makes normalization particularly challenging.

We reasoned that *OneD* should handle such cases gracefully, because the signal is formally equivalent to an aberrant karyotype where most of the genome is present in zero copy. We thus evaluated the performance of *OneD* on a Capture Hi-C data set at 5 Kbp resolution centered on a 6 Mbp domain of chromosome 6 in T47D cells. We found that the reproducibility score between replicates was higher after normalization with *OneD* than with matrix-balancing methods (figure 5A). The performance of *OneD* was only slightly lower than the raw data, suggesting that it barely disturbs the biological signal. Other tools were not included beause they had prohibitive run time.

To gain additional insight, we also set up a "virtual Capture Hi-C" data set by removing the signal outside the same region of chromosome 6 from Hi-C experiments performed in the same cell type. In this setup, we could restrict the data and then normalize, or normalize and then restrict the data. This allowed us to probe how sensitive the methods are to the shape of the input data. Here we found that *OneD* significantly outperformed the other methods: it featured the highest reproducibility score between replicates with practically no influence from the shape of the data (figure 5B). Also, in this case, *OneD* enhanced the reproducibility between replicates compared to the raw data. This confirms the view that capture methods are a natural framework for the statistical model underlying *OneD*. In comparison, matrix-balancing methods showed lower reproducibility between replicates, together with more variability due to the shape of the data.

In conclusion, *OneD* is suitable for removing experimental biases from data acquired with Capture Hi-C.

Speed

Finally, we compared the computational speed of the different normalization methods. *vanilla* and *ICE* are broadly used because of their conceptual simplicity, ease of use and speed (Imakaev *et al.*, 2012). This is even more significant as current explicit methods (Servant *et al.*, 2012) are much slower in comparison.

OneD corrects a single variable instead of the whole matrix, 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 6A shows the running times of the different methods on all the samples shown in Supplementary Figure 1 at 100 Kbp resolution. The fastest method was *vanilla* and the slowest *LGF*, with an over 100-fold span between the two.

On average, *OneD* was the second fastest method and it always ran in less than 1 min in the conditions of the benchmark. The copy number correction increased the running time, but it remained under 1 min in general. Throughout this benchmark, the memory footprint of *OneD* was less than 300 MB. Interestingly, the running time of *OneD* was 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.

To better understand these results, we also investigated the influence of bin size and sequencing depth on the running time. We performed a benchmark on the sample with highest sequencing depth in the data set (HIC070, see Supplementary Table 1) at different bin sizes while downsampling the contact matrices to mimic lower resolutions. On this test, we

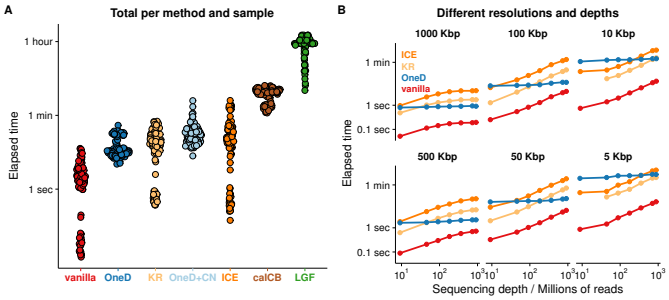

Figure 6. Computing time of the bias correction methods. A. Total time to process samples listed in Supplementary Table 1 at 100 Kbp resolution. Each dot corresponds to a sample. B. Total time to process sample HIC070 at the indicated bin size and sequencing depths. Note the logarithmic scale on the y-axis on both panels.

included only the methods competing with *OneD* in terms of speed, *i.e.* the matrix-balancing methods. Figure 6B shows that the running time of *OneD* does not depend on the sequencing depth. Strikingly, *OneD* also gives nearly identical results at very different sequencing depths (Supplementary Figure 13). *vanilla* was the overall fastest method and *OneD* was usually faster than *ICE* and *KR* at high but not at low sequencing depth. At low resolution (100 Kbp) the speed advantage of *OneD* appeared from low sequencing depth, but at high resolution resolution (5 Kbp) it appeared only at high sequencing depth.

Taken together, these results show that *OneD* is competitive in terms of speed. It is particularly adapted to projects where the sequencing depth is high, as the running time is essentially not affected.

DISCUSSION

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 from biological 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 2), and that it performs equally well as the best methods on euploid genomes (Figure 3). We also showed that *OneD* is approximately as fast as *ICE* (Figure 6), which makes it competitive from the point of view of computational speed.

The originality of *OneD* lies in the estimation from a single variable of the explicit biases with a Negative Binomial model and of the copy number. This allows greater running speed, while preserving a good performance on samples with karyotypic aberrations. One of the reasons why *OneD* is able to better highlight the biological distinctions between samples is that it only corrects the copy number if specifically requested by the user. The impact of copy number variations on normalization is rather opaque, especially if they are treated as implicit biases (Figure 4B). Treating them as explicit biases with optional removal seems to be an overall safer strategy. In homogeneous data sets, correcting for the copy number can blur the biological signal, but when karyotypes are variable or unstable, it may increase the reproducibility (Figure 4A).

This raises the question whether variations of the copy number constitute a biological signal or an artifact. If the biological sample contains karytoypic aberrations, then its genome is grossly different from the reference genome, which makes signal correction very challenging. The proper approach would be to use the actual genome of the biological sample as a reference to construct the contact map. However, this strategy is presently unfeasible because assembling mammalian genomes is still a hard problem.

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 euploid-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 mismatch 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).

CONCLUSION

Overall, *OneD* constitutes a novel computational approach 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 in terms of quality and of computational speed.

ACKNOWLEDGEMENTS

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

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 manuscript reflects only the author's views and the Union is not liable for any use that may be made of the information contained therein.

Conflict of interest statement. None declared.

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.

Choi, N. M. and Feeney, A. J. (2014). CTCF and ncRNA regulate the three-dimensional structure of antigen receptor loci to facilitate V (D) J recombination. *Frontiers in immunology*, **5**, 49.

De Laat, W. and Duboule, D. (2013). Topology of mammalian developmental enhancers and their regulatory landscapes. *Nature*, **502**(7472), 499–506.

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

Dekker, J. and Mirny, L. (2016). The 3D genome as moderator of chromosomal communication. *Cell*, **164**(6), 1110–1121.

Dekker, J. *et al.* (2002). Capturing chromosome conformation. *science*, **295**(5558), 1306–1311.

Dekker, J. *et al.* (2013). Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nature Reviews Genetics*, **14**(6), 390–403.

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

Dostie, J. *et al.* (2006). Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome research*, **16**(10), 1299–1309.

ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**(7414), 57–74.

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.

Galupa, R. and Heard, E. (2015). X-chromosome inactivation: new insights into cis and trans regulation. *Current opinion in genetics & development*, **31**, 57–66.

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.

Jäger, R. *et al.* (2015). Capture hi-c identifies the chromatin interactome of colorectal cancer risk loci. *Nature communications*, **6**.

Knight, P. A. and Ruiz, D. (2013). A fast algorithm for matrix balancing. *IMA Journal of Numerical Analysis*, **33**(3), 1029–1047.

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.

Pezic, D. *et al.* (2017). More to cohesin than meets the eye: complex diversity for fine-tuning of function. *Current Opinion in Genetics & Development*, **43**, 93–100.

Quilez, J. *et al.* (2017). Parallel sequencing lives, or what makes large sequencing projects successful. *GigaScience*.

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.* (2017). Automatic analysis and 3d-modelling of hi-c data using tadbit reveals structural features of the fly chromatin colors. *PLoS computational biology*, **13**(7), e1005665.

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

Simonis, M. *et al.* (2006). Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture–on-chip (4C). *Nature genetics*, **38**(11), 1348–1354.

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.