Published online 11 August 2015

Nucleic Acids Research, 2016, Vol. 44, No. 1 ***e4***
doi: 10.1093/nar/gkv820

Quantification of DNA cleavage specificity in Hi-C experiments

Dario Meluzzi and Gaurav Arya*

Department of NanoEngineering, University of California San Diego, 9500 Gilman Dr., La Jolla, CA 92093, USA

Received March 23, 2014; Revised July 9, 2015; Accepted August 2, 2015

ABSTRACT

Hi-C experiments produce large numbers of DNA sequence read pairs that are typically analyzed to deduce genomewide interactions between arbitrary loci. A key step in these experiments is the cleavage of cross-linked chromatin with a restriction endonuclease. Although this cleavage should happen specifically at the enzyme's recognition sequence, an unknown proportion of cleavage events may involve other sequences, owing to the enzyme's star activity or to random DNA breakage. A quantitative estimation of these non-specific cleavages may enable simulating realistic Hi-C read pairs for validation of downstream analyses, monitoring the reproducibility of experimental conditions and investigating biophysical properties that correlate with DNA cleavage patterns. Here we describe a computational method for analyzing Hi-C read pairs to estimate the fractions of cleavages at different possible targets. The method relies on expressing an observed local target distribution downstream of aligned reads as a linear combination of known conditional local target distributions. We validated this method using Hi-C read pairs obtained by computer simulation. Application of the method to experimental Hi-C datasets from murine cells revealed interesting similarities and differences in patterns of cleavage across the various experiments considered.

INTRODUCTION

The recently developed technique of chromosome conformation capture (3C) and its derivatives known as 4C, 5C and Hi-C (1–5) provide valuable information about interactions between different genomic loci (6). Such interactions can in turn be analyzed to infer the spatial organization of chromatin (7–13). Among the more advanced of these methods, Hi-C experiments can probe chromatin interactions across an entire genome (14).

Each Hi-C experiment involves an elaborate protocol that eventually yields a large number of pairs of short sequence reads. First, the genomic DNA inside intact nuclei is covalently cross-linked by treatment with formaldehyde. The cross-linked DNA is then digested with a restriction enzyme, and the resulting sticky ends are labeled with biotin and filled to generate blunt ends. These blunt ends are ligated and the ligation products are sheared, size-selected and enriched by biotin-streptavidin pulldown. Owing to size-selection, the final Hi-C library consists of DNA molecules whose lengths vary over a narrow range, e.g. from 300 to 500 bp. The ends of these final products are sequenced using next-generation DNA sequencing technology (15). Thus each product molecule yields a pair of short sequence reads, and all reads have the same length, e.g. 50 bases. The read pairs can in turn be analyzed to deduce interactions, or contacts, between different parts of the original chromatin fiber (16).

A key step in Hi-C experiments is the digestion of cross-linked chromatin by a Type II restriction endonuclease that cleaves DNA specifically at locations containing the enzyme's recognition sequence (5), which is generally 4 or 6 bases long and is palindromic (17). For example, the HindIII restriction enzyme recognizes AAGCTT and hydrolyzes the phosphodiester bond between the two adenosines on each strand. In reality, however, DNA cleavage in Hi-C experiments may be less specific than expected (18,19), because restriction enzymes can also cleave alternate sequences that differ in one base from the cognate recognition sequence, a phenomenon known as star activity (20). Moreover, additional cleavages may result from random DNA breakage (19). These alternative mechanisms then give rise to different relative fractions of DNA cleavages. A quantitative estimation of such cleavage fractions would enable accurate computer simulations for generating known Hi-C products, which could then be used to test downstream computational pipelines for data analysis. Cleavage fractions could also be used to compare different Hi-C datasets or monitor experimental conditions, e.g. through the fraction of cleavages due to random DNA breakage and to investigate biophysical properties, such as chromatin compaction, that may correlate with DNA cleavage patterns in Hi-C experiments.

*To whom correspondence should be addressed. Tel: +1 858 822 5542; Fax: +1 858 534 9553; Email: garya@ucsd.edu

Here we describe a computational method for analyzing Hi-C datasets to estimate the fractions of cleavages occurring at both cognate and non-cognate recognition sequences of a given restriction enzyme, as well as the fraction of cleavages due to random DNA breakage. This method does not require additional experimental data besides the read pairs provided by Hi-C experiments, and can therefore be used with available datasets to investigate past experiments. We validate the proposed method using known Hi-C fragments generated by computer simulation and illustrate the application of the method by obtaining cleavage fractions from published Hi-C datasets on murine pre-pro-B, pro-B and ES cells.

MATERIALS AND METHODS

Model of the cleavage process in Hi-C experiments

We model the cleavage process in Hi-C experiments by assuming two possible mechanisms: enzyme activity and random DNA breakage. The former results in DNA cleavage only at genomic locations containing the enzyme's cognate recognition sequence or its single-base mutants, whereas the latter mechanism results in cleavage at any genomic location, independently of DNA sequence. We assume that cleavage may occur at any n-base sequence, where n is the length of the enzyme's recognition sequence, e.g. $n = 6$ for HindIII and $n = 4$ for MboI. We also assume that cleavage efficiency does not depend on DNA duplex orientation. Thus, we define a cleavage 'target' to be any n-base sequence or its reverse complement and assign a unique integer index i to each target. The number of different possible n-base targets is $N =$ (palindromic targets) + (all other targets) $= 4^{n/2} + (4^n - 4^{n/2})/2$. The $N = 136$ possible 4-base targets and the $N = 2080$ possible 6-base targets are listed in supplementary files targets4.txt and targets6.txt, respectively. The 'cognate target' (CT) is the recognition sequence of the restriction enzyme. For simplicity, cleavage of any target is assumed to produce the same staggered profile of sticky DNA ends as enzymatic cleavage at the CT (Figure 1A).

We next define the 'cleavage fraction' r_i, $i = 1, \ldots, N$, with $\sum_{i=1}^{N} r_i = 1$, as the fraction of cleavages occurring at sites with target i prior to the ligation of biotin-labeled blunt ends in the Hi-C protocol. Suppose that target i occurs at S_i sites in the reference sequence. Then, among the cleavages due to random breaks, the fraction occurring at target i is $r_{i|\mathrm{b}} = q_i$, where $q_i = S_i/\sum_i S_i$ is the proportion of sites with target i in the reference sequence. Among the cleavages due to enzyme activity, the fraction occurring at target i is $r_{i|\mathrm{e}} = p_{\mathrm{e}|i} q_i / \sum_k p_{\mathrm{e}|k} q_k$, where $p_{\mathrm{e}|i}$ is the enzymatic cleavage probability for target i, i.e. the probability of successful cleavage by the enzyme when it binds to a site with sequence i. Therefore, the total fraction of cleavages at target i is:

$$r_i = p_{\mathrm{b}} r_{i|\mathrm{b}} + (1 - p_{\mathrm{b}}) r_{i|\mathrm{e}} = p_{\mathrm{b}} q_i + (1 - p_{\mathrm{b}}) \frac{p_{\mathrm{e}|i} q_i}{\sum_k p_{\mathrm{e}|k} q_k}, \quad (1)$$

where p_{b} is, among all cleavages, the fraction due to random DNA breakage and $1 - p_{\mathrm{b}}$ is the fraction due to enzyme activity. Hence, in our model, the problem of quantifying cleavage specificity in Hi-C experiments boils down to estimating r_i for all possible targets.

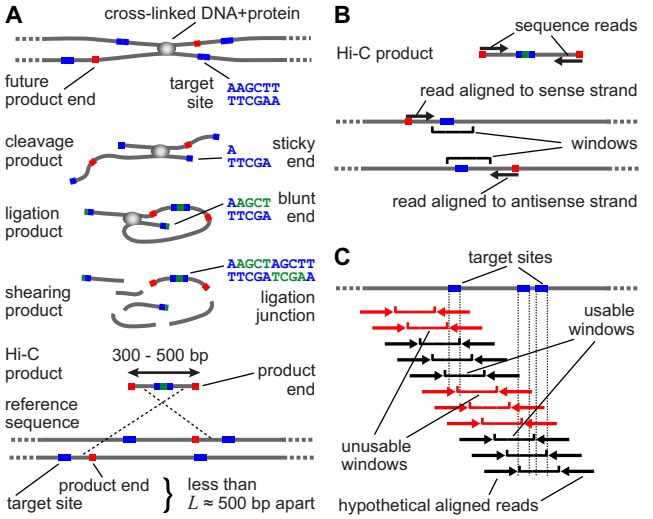

Figure 1. (**A**) Essential experimental steps yielding a Hi-C product each of whose ends (red) maps to the reference sequence at a location near the corresponding cleaved target site (blue). Cleavage at the HindIII recognition sequence is shown to illustrate the sequence of a ligation junction. Several experimental details, such as biotin labeling of blunt ends, are omitted for clarity. (**B**) Windows used to determine the OLTD begin at the downstream end of each aligned read. Arrows indicate extent and orientation of single reads before and after alignment to the reference sequence. (**C**) Windows usable for constructing a CLTD are those in which the chosen target occurs at least once. Target sites, products, reads and windows are not drawn to scale. In the present study, reads are 50-bp long and windows are 400-bp long.

Estimation of cleavage fractions

The final Hi-C products result from pulldown and size-selection of sheared, biotin-labeled ligation products (5). Consequently, each Hi-C product should contain at least one ligation junction that resulted from joining two blunt ends, each derived from cleavage of some target (Figure 1A). Because the Hi-C product length L has a limited range, say 300 to 500 bp (5), each of the two ends of a product molecule is connected to a corresponding blunt end by a DNA stretch of up to ~500 bp. Then, the genomic location of the target from which a particular blunt end was derived should be within 500 bp from the genomic location of the aligned read associated with that blunt end (see bottom of Figure 1A). Thus, by examining the targets present in the downstream vicinity of each aligned read, we should observe the cleaved targets more often than the uncleaved ones.

To quantify this observation, we construct a 'local target distribution', which specifies, for each possible target i, the frequency of that target within a particular set of windows of length W over the given reference sequence. By choosing each window to start immediately downstream of each aligned read from a given Hi-C dataset, we obtain an 'observed local target distribution' (OLTD). If cleavage occurred only at target i, the resulting OLTD would show a higher frequency at the cleaved target than at others. We can predict this OLTD by constructing a 'conditional local target distribution' (CLTD) for target i. A CLTD can be obtained from the reference sequence by counting targets in windows immediately downstream of each possible aligned read consistent with cleavage at target i (Figure 2A and B).

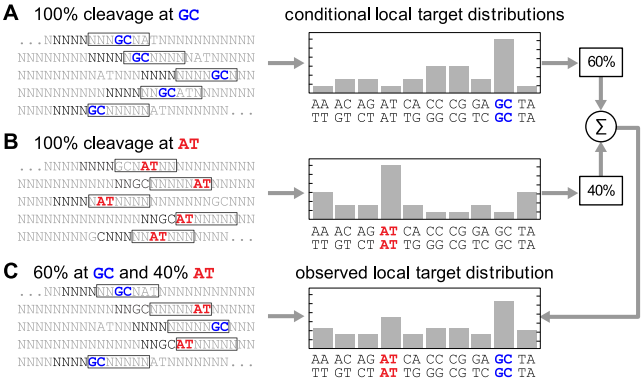

Figure 2. Schematic representation of the hypothesis underlying the proposed method for estimating cleavage fractions from genomic locations of aligned reads. In this trivial example, the lengths of targets (blue and red), aligned reads (black) and windows (rectangular outlines) on the reference sequence are assumed to be 2, 4 and 7 bp, respectively. In actual Hi-C experiments, such lengths may be 6, 50 and 400 bp, respectively (see Figure 3). There are 10 different possible targets of length 2 bp. Averaging counts of target sites over windows consistent with cleavage at only one target, GC in (**A**) or AT in (**B**), yields a conditional local target distribution (CLTD) that emphasizes the cleaved target. The observed local target distribution (OLTD), i.e. the distribution observed when cleavage occurs at more than one target (**C**), is assumed to be a linear combination of CLTDs corresponding to the cleaved targets, with weights equal to the unknown cleavage fractions.

There are N different CLTDs for a given reference sequence, one for each possible target.

By construction, the CLTD for target i approximates the OLTD that would be obtained from a large number of product ends if cleavage occurred randomly but uniformly only at target i. On the other hand, if cleavage occurred at various targets, we would expect the resulting OLTD to be a mixture of the CLTDs corresponding to the cleaved targets, and the contribution of each CLTD should reflect the proportion of cleavages at the corresponding target. We therefore propose to express an OLTD as a linear combination of CLTDs (Figure 2C), each weighted by the cleavage fraction for the corresponding target, i.e., $\mathbf{b} = \mathbf{S}\mathbf{r} + \boldsymbol{\epsilon}$, where $\mathbf{b}$ is an N-element vector containing the OLTD, $\mathbf{S}$ is a $N \times N$ matrix each of whose columns $\mathbf{s}_i$ contains the CLTD for a particular target i, $\mathbf{r}$ is a N-element vector containing the cleavage fractions r_i for $i = 1, \ldots, N$, and $\boldsymbol{\epsilon}$ is an N-element vector of random errors due to finite averaging in the construction of the OLTD.

In practice, we want to estimate the cleavage fraction p_b due to random DNA breakage and the cleavage fractions $x_i = (1 - p_\mathrm{b})r_{i\in\mathrm{e}}$ due to enzyme activity assumed to involve only $E \ll N$ possible targets, namely the CT of the restriction enzyme and the $3n/2$ targets corresponding to all possible single-base mutations in the palindromic sequence of the CT. For 6-base targets, $E = 10$. Let the subset $\mathcal{E}$ contain such targets. Then, as shown in Supplementary Data, we can rearrange the above linear combination to:

$$\mathbf{b} = \mathbf{A}\mathbf{x} + \boldsymbol{\epsilon}, \qquad (2)$$

where $\mathbf{A}$ is an $N \times (E+1)$ matrix and $\mathbf{x}$ is an $(E+1)$-element vector. The first E columns of $\mathbf{A}$ are the columns $\mathbf{s}_i$ of $\mathbf{S}$ for $i \in \mathcal{E}$, whereas the last column is equal to $\sum_{i=1}^{N} q_i \mathbf{s}_i$ and represents the local target distribution due to random breakage, i.e. the OLTD that would be observed if cleavages occurred uniformly at all possible targets. The first E elements of $\mathbf{x}$ are $x_i = (1 - p_\mathrm{b})r_{i\in\mathrm{e}}$, and thus represent the contribution of enzyme activity to the cleavage fractions for targets $i \in \mathcal{E}$ in Equation (1). The last element of $\mathbf{x}$ is equal to p_b. To find an optimal $\mathbf{x}$ for given $\mathbf{b}$ and $\mathbf{A}$ in Equation (2), we solve a non-negative least squares problem (21) using the function scipy.optimize.nnls provided by SciPy tools (22). Equation (1) then gives r_i from x_i and p_b.

Determination of OLTD. In Hi-C experiments, the ends of each final product molecule are sequenced to yield a pair of short reads. Each read can be aligned to zero, one, or multiple locations on the reference sequence. We disregard pairs containing reads that align with mismatches or align to multiple locations. Also rejected are 'inward' read pairs, i.e. pairs of reads that point toward each other after alignment to the reference sequence, because such read pairs are not always the result of cleaving a target site downstream of each read (23). Each read from each accepted pair defines the start of a downstream window that should contain the cleaved target. Because a read containing a ligation junction cannot be aligned to the reference sequence, neither read in an accepted pair may contain the cleaved target. Thus, the length of the downstream window is $W = L - 2m$, where L is the assumed product length and m is the length of a read (Figure 1B). Over M such windows we compute the average number b_i of occurrences of target i, for $i = 1, \ldots, N$. Then, the vector $\mathbf{b} = [b_1\ b_2\ \ldots\ b_N]^\mathrm{T}$ contains the desired OLTD.

Determination of CLTDs. A CLTD is the OLTD predicted when cleavage occurs only at a particular target i. Thus, we construct a CLTD just like an OLTD by averaging the number of target occurrences within a set of windows over the same reference sequence used to construct the OLTD. In this case, however, we consider all windows consistent with cleavage at target i, i.e. target i must occur at least once in each window of length W (Figure 1C). Because reads that align to multiple locations in the reference sequence are ignored when constructing an OLTD, the windows that start at those locations must be omitted in the construction of all CLTDs. The CLTD for target i is placed in column $\mathbf{s}_i = [s_{1i}\ s_{2i}\ \ldots\ s_{Ni}]^\mathrm{T}$ of the $N \times N$ matrix $\mathbf{S}$, where s_{ki} is the average number of occurrences of target k in windows consistent with cleavage at target i.

Simulations of Hi-C experiments

To validate our computational method, we generated artificial read pairs using a computer program that simulates Hi-C experiments on a given genome, as described in Supplementary Data.

Analysis of experimental Hi-C datasets

To analyze fragments from real Hi-C experiments, we retrieved datasets from the NCBI Sequence Read Archive (24) for experiments SRX178471, SRX178473, SRX118420 through SRX118426, SRX116341 and SRX116342. These datasets contain pairs of 50- or 36-bp sequence reads from

experiments on murine pre-pro-B, pro-B and ES cells (25–27). The read pairs from these datasets were aligned and selected as described in Supplementary Data. Then, for each experiment, the selected pairs were sorted by genomic location of the read with smallest location in each pair, and split into three interleaved samples of approximately equal size. Each sample was then used to estimate the desired cleavage fractions with Equation 2, and the resulting values were used to calculate sample means and standard deviations.

RESULTS

Validation with simulated Hi-C products

Our method for estimating cleavage fractions from Hi-C datasets involves three computational steps: determination of N CLTDs from the reference sequence, determination of an OLTD from a Hi-C dataset and solution of a non-negative least squares problem formulated in terms of those distributions, Equation (2).

CLTDs identify cleaved targets. Each CLTD is constructed from the reference sequence by assuming that cleavage involves only a particular target. Intuitively, the cleaved target should be more strongly represented than others in the associated CLTD, because at least one such target must occur in the windows used to construct the associated CLTD, while other targets are not necessarily present in each window. Thus, each CLTD should specifically identify the corresponding cleaved target.

These expectations are confirmed by the CLTDs obtained for the HindIII CT and for three of its single-base mutants (Figure 3A–D). Because these CLTDs were constructed from an artificial chromosome consisting of uniformly random bases, most targets can be seen to occur with frequencies close to $2 \times (400 + 1)/4^6 \approx 0.2$, i.e. twice the frequency expected for any 6-base sequence in a 400-bp window. There are also targets with frequencies close to 0.1, the frequency expected for targets with palindromic 6-base sequences. As anticipated, in each CLTD the cleaved target occurs with the highest frequency. However, frequencies enhanced above the background are also seen for targets whose sequences partially overlap the sequence of the cleaved target, because these targets are more likely than non-overlapping ones to occur in windows where the cleaved target is present. The existence of enhanced frequencies for a small number of targets that depend on the sequence of the cleaved target supports the notion that each CLTD identifies a particular cleaved target. It should therefore be possible to compute the unique set of weights needed to express a given OLTD as a linear combination of CLTDs.

Hi-C products from computer simulations. To test the proposed method with read pairs at known genomic locations, we performed three groups of computer simulations of Hi-C experiments. The first group generated constant-length Hi-C products from a reference sequence containing uniformly random bases. The second group generated products from chr19 of the mm10 mouse genome. The third group generated variable-length products from the same chromosome. In each group of simulations, we induced four different patterns of enzymatic cleavage (Supplementary Table S1). Cleavage at a 'single' target involved only the CT AAGCTT of the HindIII restriction enzyme. Cleavage at '4 targets' used a decreasing probability of enzymatic cleavage p_{eli} at the CT and three of its 1-base mutants, namely AAACTT, AACCTT, AAGATT, where smaller letters indicate mutated bases. Cleavage at '10 targets' again used decreasing p_{eli}, but at the CT and all of its 1-base mutants. Cleavage at '10 evenly' used $p_{eli} = 100\%$ for all '10 targets.' Additionally, to investigate various extents of random DNA breakage, each pattern of cleavage was simulated with four different values of p_b, namely 0, 10, 20 and 50%. To obtain error bars, each simulation was carried out three times with different seeds of the random number generator. Finally, we repeated all simulations using 4-base targets. Thus we performed a total of $3 \times 4 \times 4 \times 3 \times 2 = 288$ simulations, which generated a total of 1.44B read pairs aligned to known genomic locations.

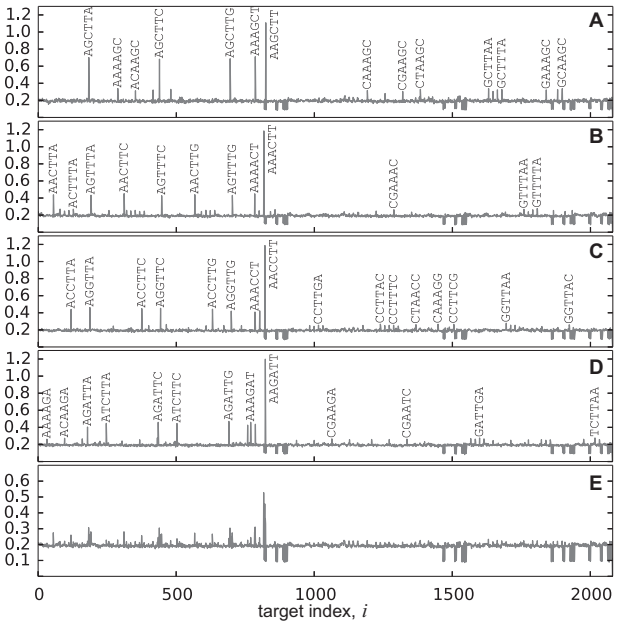

Figure 3. Examples of (**A**–**D**) conditional and (**E**) measured local target distributions derived from an artificial chromosome consisting of 5 141 828 uniformly random bases. Counts of sites for each possible target i were averaged over all possible windows of length $W = 400$ bp on the reference sequence, such that each window contained at least one site with the target sequence (A) AAGCTT, (B) AAACTT, (C) AACCTT, (D) AAGATT, as explained in Figure 1C, or such that (E) each window started immediately downstream of an aligned read from one or the other end of a Hi-C product, as explained in Figure 1B. Products for (E) were obtained from simulations with decreasing enzymatic cleavage probabilities p_{eli} at the above four target sequences, and with zero fraction p_b of cleavages due to random DNA breakage (column '4 targets' in Supplementary Table S1). The file targets6.txt lists 6-base targets in order of index i.

Cleavage fractions from simulations on random sequence. To begin, we analyzed the constant-length Hi-C products obtained from an artificial reference sequence of uniformly random bases. We verified that the OLTD (Figure 3E) closely approximates a linear combination of the CLTDs for the cleaved targets (Figure 3A–D) with weights equal to the cleavage fractions $\tilde{r}_i$ *measured* from the simulations (data not shown).

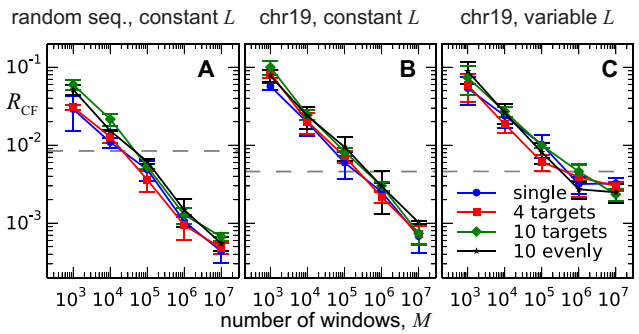

Figure 4. Variation of the residual R_{CF} between estimated and measured cleavage fractions with increasing number M of windows used to compute the OLTD from simulated Hi-C read pairs. The simulations involved (**A**) a random reference sequence and products with constant length L, (**B**) chr19 of mm10 and constant-length products or (**C**) chr19 of mm10 and variable-length products. Each point is the mean of three residuals R_{CF}, each calculated using a set of estimates $\hat{r}_i$ and measurements $\tilde{r}_i$ obtained from an independent simulation. Error bars are standard deviations of the three R_{CF} values. Each series of connected points corresponds to a different pattern of enzymatic cleavage probabilities p_{eli}, all with $p_b = 20\%$. The horizontal dashed line indicates the smallest measured cleavage fraction $\tilde{r}_i$ among all simulations (see Supplementary Table S1 for the case with $p_b = 0$). Supplementary Figure S2 shows similar plots for other values of p_b.

To obtain cleavage fractions $\hat{r}_i$ *estimated* from the simulated read pairs, we solved the non-negative least squares problem of Equation (2) using the appropriate OLTD and CLTDs. Then, to assess how the accuracy of $\hat{r}_i$ varies with the number M of windows used to compute the OLTD, we varied M and compared the resulting estimates $\hat{r}_i$, for $i \in \mathcal{E}$, with corresponding measurements $\tilde{r}_i$ obtained from the simulations. We found that the sample mean of $\hat{r}_i$, over three replicates of the simulation, approaches $\tilde{r}_i$ for each cleaved target as M increases, while the sample standard deviation of $\hat{r}_i$ becomes negligible (Supplementary Figure S1), suggesting a lack of systematic error in the estimates for this simple test case.

Plotting R_{CF}, the residual between the estimated and measured cleavage fractions (see Supplementary Data), against M (Figure 4A and first column in Figures S2 and S3) confirmed that the $\hat{r}_i$'s are free from systematic error for values of p_b up to 50%. The residual $R_b = |\hat{p}_b - \tilde{p}_b|$ between estimated and measured values of p_b was also found to decrease with increasing M (Supplementary Figure S4, first column).

Cleavage fractions from simulations on chr19. To test our method with reads from a more realistic reference sequence, we used the constant-length products obtained from Hi-C simulations on chr19 of the mm10 reference mouse genome. The cleavage fractions measured in this case differed notably from the corresponding cleavage fractions measured in the previous simulations (Supplementary Table S1), a result consistent with the non-random character of real genomic sequences. Despite this non-random character, we found that the residuals R_{CF} and R_b again approach zero as M increases (Figure 4B and second columns in Supplementary Figures S2, S3 and S4), indicating that the estimated cleavage fractions approach the measured cleavage fractions and are therefore unbiased even for a real chromosome.

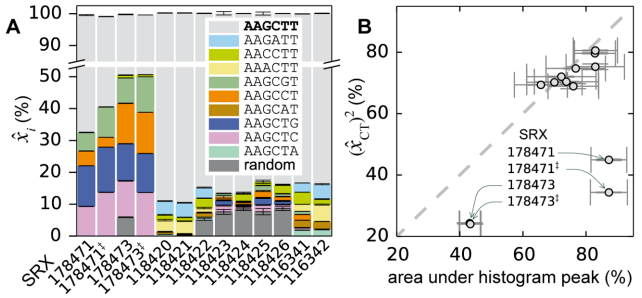

Figure 5. (**A**) Enzymatic cleavage fractions $\hat{x}_i$ and proportion p_b of cleavages due to random DNA breakage estimated from Hi-C datasets of (25), (26) and (27). Error bars represent standard deviations over three pseudo-samples. (**B**) The square $(\hat{x}_{CT})^2$ of the enzymatic cleavage fraction estimated for the cognate target (CT) (y-axis) and the area under the peak in the histograms (Supplementary Figure S5) of experimental product lengths (x-axis) are alternative approximations to the proportion of Hi-C products resulting from cleavage only at the CT of the enzyme. The two approximations agree in most of the cases analyzed. Each point corresponds to an experiment whose cleavage fraction estimates are reported in (A) and in Supplementary Table S3. Vertical error bars are the same as in (A). Horizontal error bars account for the width of the bins bounding the histogram peak. The line $y = x$ (dashed) is shown as a guide. ‡ Cleavage fractions estimated for chr1, rather than chr19.

We next analyzed the products generated by the third group of Hi-C simulations, which again involved chr19 of mm10 but also introduced a spread in the length L of the simulated products to better approximate the products from real experiments (Supplementary Figures S5 and S6). In this case, increasing M caused R_{CF} to level off at around 0.3% (Figure 4C), suggesting the presence of a small systematic error. This error is fairly insensitive to p_b (Supplementary Figures S2 and S3, third column) and is present also in $\hat{p}_b$ (Supplementary Figure S4, third column). As such error was absent in estimates from constant-length products, it likely ensued from analyzing variable-length products with OLTDs and CLTDs constructed over windows of constant length W, effectively assuming a constant length L for all products. Although real Hi-C products do vary in length, using constant-length windows to estimate cleavage fractions from such products may provide adequate estimates $\hat{r}_i$ for applications that can tolerate errors on the order of 1%.

Application to experimental Hi-C products

To perform our calculations on real Hi-C products from different cells and different sources, we obtained two datasets from murine E2A-deficient hematopoietic progenitor (pre-pro-B) cells and RAG-1-deficient pro-B cells (25), seven datasets from murine Ataxia Talangiectasia mutated kinase deficient (ATM-/-) and wild-type pro-B cells with an I-SceI site in chr2, chr7, chr15 or chr18 (26), and two datasets from murine embryonic stem cells (mESCs) (27). We then applied our method to estimate cleavage fractions due to enzyme activity and random breakage from each experimental dataset. To limit computational effort, we obtained estimates only for cleavages on chr1 and chr19 (Figure 5A and Supplementary Table S3).

The fraction $\hat{p}_b$ of cleavages attributed to random breakage was negligible in 7 of the 13 cases analyzed (Supple-

mentary Table S3). In all cases, the largest enzymatic cleavage fraction was estimated at the CT of the HindIII enzyme and ranged from 49 to 90%. Such fraction, however, was smaller for experiments SRX178471 and SRX178473 than for all other experiments. Also, SRX178471 and SRX178473 gave the same composition of enzymatically cleaved targets, on both chr1 and chr19, but such composition differed from that of the other experiments, possibly owing to procedural differences. Interestingly, data for SRX118420 and SRX118421, which were biological replicates on ATM-/- I-SceI-chr2 pro-B cells, gave similar cleavage patterns, in agreement with the expectation that different biological replicates produce similar results. Similarity of cleavage patterns also ensued from mESC biological replicates SRX116341 and SRX116342. However, data for SRX118423 and SRX118424, biological replicates on ATM-/- I-SceI-chr18 pro-B cells, gave similar values of $\hat{p}_b$ and $\hat{x}_{CT}$ but dissimilar cleavage fractions at other enzymatically cleaved targets, suggesting the presence of small estimation errors, as seen in Figure 4C, or variation in experimental conditions.

Previous studies have characterized the star activity of the HindIII restriction endonuclease under non-standard conditions involving high pH, high ionic strength, high enzyme concentration or the addition of the organic solvent DMSO. Under these conditions, the HindIII endonuclease was found to cleave at targets AAGCCT, AAGATT, AAGCGT, AAGCTC, AAACTT and AAGCAT, which differ in one base from the CT AAGCTT (28,29). For each of these non-CTs, some or all of the experiments analyzed gave cleavage fraction estimates greater than 1% (Supplementary Table S3). Significant fractions were also estimated for non-CTs AACCTT, AAGCTG and AAGCTA, which were not reported in Refs. (28,29).

To probe the validity of our estimates, we collected histograms of apparent product lengths from each dataset (Supplementary Figure S5). As the tails of these histograms correspond to Hi-C products resulting from non-specific DNA cleavage (18), the area under the peak in the normalized histograms should approximate the fraction of products resulting from cleavage only at the enzyme's CT. Another quantity approximating such fraction is the square $(\hat{x}_{CT})^2$ of the cleavage fraction estimated for the CT. Comparing the two approximations revealed a qualitative agreement for all experiments considered except SRX178471, which gave a value of $\hat{x}_{CT}$ noticeably lower than expected (Figure 5B). These results suggest the presence of systematic errors, which can be large in some cases, but appear to be small in general.

DISCUSSION

Large Hi-C datasets are typically analyzed to obtain contact maps that contain information about the 3D organization of chromatin (9). The present study shows that the same datasets can be analyzed to obtain information about DNA cleavage specificity in Hi-C experiments. In particular, we have presented a computational method for estimating cleavage fractions, which quantify the cleavages resulting from restriction enzyme activity and from random DNA breakage during the experimental steps carried out to digest cross-linked chromatin in preparation for blunt ends.

We validated our method using artificial Hi-C products generated by computer simulation. Our validation revealed a small systematic error in the cleavage fractions estimated from products with variable length (Figure 4C). Although perhaps acceptable in some applications, the observed error could be reduced by constructing OLTDs and CLTDs that account for the distribution of actual product lengths. Such distribution could in turn be inferred from histograms of apparent product lengths (Supplementary Figure S5), or from fluorescence intensity profiles of agarose gels used for size-selection of Hi-C products.

Additional validation may also be possible by using appropriate experimental data. For example, estimates of cleavage fractions could be compared to more direct measurements from careful analysis of ligation junctions found in full sequences of Hi-C products. Full product sequences, though shorter than 200 bp, were obtained in previous work assessing a modified Hi-C protocol (30). Larger numbers of full sequences, necessary for accurate measurements of cleavage fractions, may become accessible with further advances in next-generation sequencing methods (15).

Our proposed method relies on a model that includes several simplifying assumptions about the process of transforming genomic DNA into a Hi-C library. Among the assumptions are the independence of p_b on genomic location, the absence of product length variation, the absence of products with zero or more than one ligation junction, the inclusion of products with ligation junctions lacking a biotinylated cytosine, the absence of cleavage events closely spaced on the same DNA molecule, the absence of biases due to GC content and restriction fragment lengths (18) and generally the absence of any read coverage bias. We did, however, account for mappability bias (18) by omitting windows associated with unmappable reads from the computation of CLTDs. Some of the above assumptions may be responsible for the unexpectedly low value of $\hat{x}_{CT}$ obtained for experiment SRX178471 (Figure 5B). By addressing the above assumptions, future refinements to the described computational method should yield more accurate results.

Although simulations of Hi-C experiments were used to validate our computational method, they may also be valuable for other purposes. For example, simulations could provide an inexpensive means to test experimental protocol details (31), such as the choice of restriction enzyme for a given genome, or to assess the accuracy of existing and future computational methods that infer contact maps from Hi-C data, a task not easily accomplished with experimental datasets alone. Because the accuracy of contact maps inferred from Hi-C data may depend on cleavage specificity, a validation of such maps could benefit from artificial Hi-C products that resemble experimental ones in terms of cleavage fractions. Our estimation of such fractions would provide the parameters necessary to perform simulations that yield products with the desired characteristics.

Besides enabling accurate simulations of Hi-C experiments, the ability to estimate cleavage fractions from Hi-C read pairs may be useful to compare datasets or to monitor experimental conditions. For example, the observed vari-

ation of cleavage patterns across experiments (Figure 5A) suggests that cleavage fractions could be used as a signature to confirm the origin of different datasets or to assess the reproducibility of the same procedures performed at different times or by different investigators. Also, datasets obtained from biological replicates may be expected to produce similar patterns of cleavage, as we observed for experiments SRX118420 and SRX118421. Thus comparing cleavage fractions could provide a means to gauge the quality of different experimental runs.

The observed variation across experiments also suggests that cleavage fractions may correlate with biologically relevant aspects of the cells under study. For instance, the cleavage fractions we estimated from the experimental Hi-C datasets of Ref. (25) on pre-pro-B and pro-B cells varied significantly with cell type. In particular, the proportion of cleavages due to enzyme activity at the CT was found to decrease in pro-B cells relative to pre-pro-B cells (Figure 5A and Supplementary Table S3), suggesting that chromatin reorganization or condensation in pro-B cells may have altered the accessibility of CT sites relative to non-cognate ones. Future studies may consider estimating cleavage fractions for specific genomic domains, chromosomes or entire genomes, and correlating those fractions with biophysical properties that affect the possible mechanisms of DNA cleavage in Hi-C experiments.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors thank Dr Amy Davenport Migliori and Chaitanya Murthy for insightful discussions. The authors are also grateful to the anonymous reviewers, whose thorough comments and suggestions resulted in significant improvements to this manuscript.

FUNDING

American Cancer Society Institutional Research Grant [70-002 to Moores Cancer Center, in part]; National Science Foundation Research Grant [1200850]. Funding for open access charge: UCSD.
Conflict of interest statement. None declared.

REFERENCES

1. Dekker,J., Rippe,K., Dekker,M. and Kleckner,N. (2002) Capturing chromosome conformation. *Science*, **295**, 1306–1311.
2. Simonis,M., Klous,P., Splinter,E., Moshkin,Y., Willemsen,R., deWit,E., vanSteensel,B. and deLaat,W. (2006) Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nat. Genet.*, **38**, 1348–1354.
3. Zhao,Z., Tavoosidana,G., Sjolinder,M., Gondor,A., Mariano,P., Wang,S., Kanduri,C., Lezcano,M., Singh Sandhu,K., Singh,U. *et al.* (2006) Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. *Nat. Genet.*, **38**, 1341–1347.
4. Dostie,J., Richmond,T.A., Arnaout,R.A., Selzer,R.R., Lee,W.L., Honan,T.A., Rubio,E.D., Krumm,A., Lamb,J., Nusbaum,C. *et al.* (2006) Chromosome conformation capture carbon copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome Res.*, **16**, 1299–1309.
5. Lieberman-Aiden,E., vanBerkum,N.L., Williams,L., Imakaev,M., Ragoczy,T., Telling,A., Amit,I., Lajoie,B.R., Sabo,P.J., Dorschner,M.O. *et al.* (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, **326**, 289–293.
6. deWit,E. and deLaat,W. (2012) A decade of 3C technologies: insights into nuclear organization. *Genes Dev.*, **26**, 11–24.
7. Bau,D., Sanyal,A., Lajoie,B.R., Capriotti,E., Byron,M., Lawrence,J.B., Dekker,J. and Marti-Renom,M.A. (2011) The three-dimensional folding of the beta-globin gene domain reveals formation of chromatin globules. *Nat. Struct. Mol. Biol.*, **18**, 107–114.
8. Meluzzi,D. and Arya,G. (2013) Recovering ensembles of chromatin conformations from contact probabilities. *Nucleic Acids Res.*, **41**, 63–75.
9. Dekker,J., Marti-Renom,M.A. and Mirny,L.A. (2013) Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nat. Rev. Genet.*, **14**, 390–403.
10. Paulsen,J., Lien,T.G., Sandve,G.K., Holden,L., Borgan,Ø., Glad,I.K. and Hovig,E. (2013) Handling realistic assumptions in hypothesis testing of 3D co-localization of genomic elements. *Nucleic Acids Res.*, **41**, 5164–5174.
11. Kruse,K., Sewitz,S. and Babu,M.M. (2013) A complex network framework for unbiased statistical analyses of DNA–DNA contact maps. *Nucleic Acids Res.*, **41**, 701–710.
12. Hu,M., Deng,K., Qin,Z., Dixon,J., Selvaraj,S., Fang,J., Ren,B. and Liu,J.S. (2013) Bayesian inference of spatial organizations of chromosomes. *PLoS Comput. Biol.*, **9**, e1002893.
13. Serra,F., Di Stefano,M., Spill,Y.G., Cuartero,Y., Goodstadt,M., Baù,D. and Marti-Renom,M. A. (2015) Restraint-based three-dimensional modeling of genomes and genomic domains. *FEBS Letters*, doi:10.1016/j.febslet.2015.05.012.
14. Belton,J.-M., McCord,R.P., Gibcus,J.H., Naumova,N., Zhan,Y. and Dekker,J. (2012) Hi–C: a comprehensive technique to capture the conformation of genomes. *Methods*, **58**, 268–276.
15. Mardis,E.R. (2013) Next-generation sequencing platforms. *Annu. Rev. Anal. Chem.*, **6**, 287–303.
16. vanBerkum,N.L., Lieberman-Aiden,E., Williams,L., Imakaev,M., Gnirke,A., Mirny,L.A., Dekker,J. and Lander,E.S. (2010) Hi-C: a method to study the three-dimensional architecture of genomes. *J. Vis. Exp.*, **39**, e1869.
17. Pingoud,A., Fuxreiter,M., Pingoud,V. and Wende,W. (2005) Type II restriction endonucleases: structure and mechanism. *Cell. Mol. Life Sci.*, **62**, 685–707.
18. Yaffe,E. and Tanay,A. (2011) Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nat. Genet.*, **43**, 1059–1065.
19. Imakaev,M., Fudenberg,G., McCord,R.P., Naumova,N., Goloborodko,A., Lajoie,B.R., Dekker,J. and Mirny,L.A. (2012) Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nat. Meth.*, **9**, 999–1003.
20. Wei,H., Therrien,C., Blanchard,A., Guan,S. and Zhu,Z. (2008) The Fidelity Index provides a systematic quantitation of star activity of DNA restriction endonucleases. *Nucleic Acids Res.*, **36**, e50.
21. Lawson,C.L. and Hanson,R.J. (1995) *Solving Least Squares Problems*, Society for Industrial and Applied Mathematics, Philadelphia, PA.
22. Oliphant,T.E. (2007) Python for scientific computing. *Comput. Sci. Eng.*, **9**, 10–20.
23. Jin,F., Li,Y., Dixon,J.R., Selvaraj,S., Ye,Z., Lee,A.Y., Yen,C.-A., Schmitt,A.D., Espinoza,C.A. and Ren,B. (2013) A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature*, **503**, 290–294.
24. Kodama,Y., Shumway,M. and Leinonen,R. (2012) The sequence read archive: explosive growth of sequencing data. *Nucleic Acids Res.*, **40**, D54–D56.
25. Lin,Y.C., Benner,C., Mansson,R., Heinz,S., Miyazaki,K., Miyazaki,M., Chandra,V., Bossen,C., Glass,C.K. and Murre,C. (2012) Global changes in the nuclear positioning of genes and intra- and interdomain genomic interactions that orchestrate B cell fate. *Nat. Immunol.*, **13**, 1196–1204.
26. Zhang,Y., McCord,R., Ho,Y.-J., Lajoie,B., Hildebrand,D., Simon,A., Becker,M., Alt,F. and Dekker,J. (2012) Spatial organization of the mouse genome and its role in recurrent chromosomal translocations. *Cell*, **148**, 908–921.

27. Dixon,J.R., Selvaraj,S., Yue,F., Kim,A., Li,Y., Shen,Y., Hu,M., Liu,J.S. and Ren,B. (2012) Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, **485**, 376–380.
28. Nasri,M. and Thomas,D. (1986) Relaxation of recognition sequence of specific endonuclease HlndIII. *Nucleic Acids Res.*, **14**, 811–821.
29. Nasri,M. and Thomas,D. (1988) Increase of the potentialities of restriction endonucleases by specificity relaxation in the presence of organic solvents. *Ann. N. Y. Acad. Sci.*, **542**, 255–265.
30. Sexton,T., Yaffe,E., Kenigsberg,E., Bantignies,F., Leblanc,B., Hoichman,M., Parrinello,H., Tanay,A. and Cavalli,G. (2012) Three-dimensional folding and functional organization principles of the drosophila genome. *Cell*, **148**, 458–472.
31. Rodley,C., Bertels,F., Jones,B. and O'Sullivan,J. (2009) Global identification of yeast chromosome interactions using Genome conformation capture. *Fungal Genet. Biol.*, **46**, 879–886.