Supplementary Data: Quantification of DNA cleavage specificity in Hi-C experiments

Dario Meluzzi and Gaurav Arya

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

1 SUPPLEMENTARY METHODS

1.1 Cleavage fractions due to enzyme activity and random DNA breakage

The OLTD in the \( N \)-element vector \( b \) can be expressed as a linear combination of the CLTDs in the columns of the \( N \times N \) matrix \( S \), i.e., \( b = Sr + \epsilon \), where \( \epsilon \) is a vector of random errors. Let \( E = \{i_1, i_2, \ldots, i_E\} \) be the set of \( E \ll N \) targets corresponding to the cognate recognition sequence of the restriction enzyme and to all possible single-base mutants of that sequence. If we assume that the enzyme can cleave DNA only at those targets, then the cleavage fractions \( x_i = (1 - p_b) r_i \) due to enzyme activity must be zero for \( i \not\in E \). Thus, the linear combination \( Sr \) of \( N \) CLTDs can be expressed as a new linear combination \( Ax \) involving only \( E \) CLTDs and a weighted average of all \( N \) CLTDs. The weights of this average are the proportions \( q_i \) of occurrences of each target \( i \) in the reference genomic sequence. Specifically,

\[
Sr = \sum_{i=1}^{N} r_i s_i \\
= \sum_{i=1}^{N} \left[ (1 - p_b) r_{i|e} + p_b r_{i|b} \right] s_i \\
= \sum_{i=1}^{N} (x_i + p_b q_i) s_i \\
= \sum_{i=1}^{N} x_i s_i + \sum_{i=1}^{N} p_b q_i s_i \\
= \sum_{i \in E} x_i s_i + p_b \sum_{i=1}^{N} q_i s_i \\
= [s_{i_1} s_{i_2} \cdots s_{i_E} \sum_{i=1}^{N} q_i s_i]^{T} \begin{bmatrix} x_{i_1} & x_{i_2} & \cdots & x_{i_E} & p_b \end{bmatrix} \\
= Ax
\]

Hence, by solving \( b = Ax + \epsilon \), we can estimate the fractions \( x_i \) of cleavages due to enzyme activity at target sites \( i \in E \), as well as the fraction \( p_b \) of cleavages due to random DNA breakage.

1.2 Simulations of Hi-C experiments

To validate our computational method, we generated artificial Hi-C products using a computer program that simulates Hi-C experiments on a given genome. Each generated product contains a single ligation junction that occurs randomly anywhere along the length of the product, thus mimicking the randomness of the shearing process (Iyengar, 1980). Each ligation junction results from joining two blunt ends, each derived from a randomly drawn and randomly cleaved target site. Specifically, for each cleavage that eventually yields a blunt end, a chromosome is randomly picked with probability proportional to the size of that chromosome. Next, the cleavage is randomly attributed either to random breakage, with probability \( p_b \), or to enzyme activity, with probability \( 1 - p_b \). If cleavage is due to random breakage, a site is randomly picked with uniform probability among all sites where target \( i \in E \) occurs on the chosen chromosome. If cleavage is due to enzyme activity, a site is randomly picked with uniform probability among all sites where target \( i \in E \) occurs on the chosen chromosome. Then, the picked site is cleaved with probability \( p_{e|i} \). If cleavage fails for this site, other sites are repeatedly picked on the chosen chromosome until successful cleavage ensues. Because we disregard reads that contain ligation junctions or that align to multiple genomic locations, we discard a read pair if either read matches such

* to whom correspondence should be addressed
criteria. We also discard read pairs that are not “samestrand” (Jin et al., 2013). In this way, all of the read pairs generated by the simulations are directly usable for computing OLTDs without any additional filtering.

We simulated Hi-C experiments on two possible reference genomes (Table S1). The first one contained one chromosome of 5,141,828 random bases, each drawn with uniform probability from \{A, C, G, T\}. The second reference genome consisted of a single chromosome, the 61,431,566 base-long chr19 from the reference mouse genome (Genome Reference Consortium GRCh38, UCSC version mm10). Depending on the simulation, the length \( L \) of the Hi-C products was either constant and equal to 500 bp, or varied according to a modified normal distribution with mean 370 bp, standard deviation 50 bp, and a lower tail truncated at the length of two reads, i.e., 100 bp. The mean and standard deviation were estimated by trial and error to reproduce approximately the location and spread of the peaks of the apparent product length distributions obtained from experimental Hi-C data sets (Fig. S5 and Fig. S6), as described below.

To assess the accuracy of the cleavage fractions \( \hat{r}_i \), estimated by solving Eq. (2) and using Eq. (1) in the main text, we compared these fractions to values \( \tilde{r}_i = n_i/M \) measured from the simulations, where \( n_i \) is the number of product ends connected to a blunt end that was derived from cleavage at target \( i \), and \( M = \sum_{i=1}^{N} n_i \) is the total number of simulated blunt ends. Then, to obtain an upper bound for the error in \( \hat{r}_i \), we computed the residual \( R_{CP} = \sqrt{\sum_{i=1}^{N} (\hat{r}_i - \tilde{r}_i)^2} \).

### 1.3 Alignment of reads from Hi-C experiments

Each sequence read from each Hi-C read pair was submitted to the Bowtie software (Langmead et al., 2009) for alignment against the mm10 reference mouse genome. To obtain only genomic locations of uniquely aligned reads, i.e., those that align without mismatches to a unique location in the reference sequence, the Bowtie option “-v 0” was used. Among the read pairs uniquely aligned to the reference sequence, “inward” read pairs were discarded, as not necessarily resulting from DNA cleavage (Jin et al., 2013). Pairs containing reads that uniquely aligned to the same genomic locations as reads in some other pair, were also discarded as likely artifacts of PCR amplification (Imakaev et al., 2012). To estimate cleavage fractions due to enzyme activity and random breakage in Hi-C experiments, we selected only read pairs aligned to one chromosome of the mm10 mouse genome, either chr1 or chr19, and used those reads to construct the OLTDs. Counts of read pairs at various stages of the alignment procedure are reported in Table S2 for each experiment considered.

### 1.4 Histograms of apparent product lengths

To assess the quality of our simulations, we determined histograms of apparent lengths of both simulated and experimental Hi-C products, as described in Yaffe and Tanay (2011), assuming cleavage only at the cognate target of the HindIII restriction enzyme. This assumption is what makes the resulting product lengths “apparent,” rather than actual, in the presence of non-specific cleavage. For each read in each experimental read pair uniquely aligned to chr19, the genomic location of the first target site was found downstream of the alignment location of the read on the reference sequence, but outside the span of the read itself. The “downstream” direction was from the 5′-end to the 3′-end of the strand to which the read was aligned. The sum of the distances from the location of each read in the pair to the location of the corresponding target site was taken as the apparent length \( L \) of the product represented by the examined read pair. This length included the length of each read and the length of the single-stranded portion of the HindIII sticky ends, which are filled to generate blunt ends in Hi-C experiments (Fig. 1A in main text). A similar procedure for computing product lengths was followed with Hi-C products obtained from simulations.

### 1.5 Areas of peaks in histograms of apparent product lengths

The location and width of the peak in each histogram of apparent product lengths were found to vary across the different experimental data sets examined (Fig. S5). To approximate the area under the peak in a consistent way despite changes in peak location and width, we calculated the area between upper and lower bounds that were obtained by following the same procedure for each histogram. First, the location and value of the maximum height in the histogram were determined. Next, the intersections of the horizontal line at half height with the rising and falling edges of the peak were found. Then, the distance of each intersection from the location of the maximum height was doubled to obtain the location of the bound on the same side of the peak as the intersection in question. Finally, the areas of the histogram bins between the resulting lower and upper bounds were summed to obtain the area under the peak. To calculate error bars for this area, the half-areas of the bins corresponding to the upper and lower bounds of the sum were added in quadrature.

### 2 SUPPLEMENTARY RESULTS

#### 2.1 Histograms of apparent product lengths

We initially probed the first two experimental data sets by computing histograms of apparent product lengths. We found such histograms to be peaked over the range of expected product lengths. The histogram for the second data set also displayed a long upper tail ranging from about 500 bp to more than 10,000 bp (SRX128473 in Fig. S5). Similar histograms were reported in a previous study (Yaffe and Tanay, 2011) that analyzed Hi-C products from the experiments of Lieberman-Aiden et al. (2009) on GM06990 human lymphoblast cells. The long upper tails were attributed to non-specific cleavage of the chromatin (Yaffe and Tanay, 2011). Our histograms show that approximately 67% of the products derived from pro-B cells in data set SRX128473 had apparent lengths greater than the upper bound for cleavages only at the enzyme’s cognate target (Fig. S5), whereas the fraction of such products from the other data sets analyzed was no more than about...
Table S1. Enzymatic cleavage patterns and corresponding cleavage fractions in simulations of Hi-C experiments with 6-base or 4-base targets.

| target | single | 4 targets | 10 targets / 7 targets | 10 evenly / 7 evenly |
|--------|--------|----------|-----------------------|---------------------|
|        | | $f_{i,k}$ | | $f_i$ |
|        | | $p_{i|k}$ | | $p_i$ |
| $i^b$ sequence$^c$ | group 1 | 2, 3 | group 1 | 2, 3 | group 1 | 2, 3 | group 1 | 2, 3 |
| 825 AAGCTT | 100 | 100.00 | 100.00 | 100 | 22.51 | 19.88 | 100 | 53.57 | 52.98 | 100 | 5.45 | 5.55 |
| 818 AAACTT | 80 | 34.01 | 41.55 | 9 | 9.08 | 12.45 | 100 | 10.27 | 14.53 |
| 822 AACCTT | 60 | 26.17 | 20.83 | 8 | 8.30 | 7.41 | 100 | 10.55 | 9.71 |
| 824 AAGATT | 40 | 17.30 | 17.74 | 7 | 7.20 | 8.28 | 100 | 10.46 | 12.39 |
| 778 AAGCAT | 5 | 5.17 | 5.47 | 100 | 10.53 | 11.47 |
| 794 AAGCCT | 4 | 4.24 | 4.56 | 100 | 10.73 | 12.39 |
| 810 AAGGTT | 3 | 3.91 | 2.38 | 100 | 10.20 | 9.37 |
| 58 AAGCTA | 2 | 2.06 | 1.75 | 100 | 10.45 | 9.17 |
| 314 AAGCTC | 1 | 1.04 | 1.33 | 100 | 10.61 | 13.91 |
| 570 AAGCTG | 100 | 100.00 | 100.00 | 100 | 21.93 | 17.17 | 100 | 65.18 | 57.44 |
| 120 GATC | 80 | 34.67 | 38.42 | 9 | 11.59 | 14.46 | 100 | 15.34 | 15.83 |
| 117 GAAAC | 60 | 26.01 | 29.85 | 7 | 9.01 | 11.65 | 100 | 15.33 | 16.40 |
| 119 GAGC | 40 | 17.38 | 14.56 | 5 | 6.46 | 6.09 | 100 | 15.39 | 12.01 |
| 118 GACC | 3 | 3.87 | 5.04 | 100 | 15.36 | 16.56 |
| 20 AATC | 2 | 2.60 | 3.81 | 100 | 15.44 | 18.79 |
| 78 CATC | 1 | 1.29 | 1.50 | 100 | 15.39 | 14.74 |

$^a$ Four sets of enzymatic cleavage probabilities $p_{i|k}$ were chosen to produce different patterns of cleavage fractions in three groups of computer simulations used to generate Hi-C products, either from an artificial reference genome consisting of uniformly random bases (group 1) or from chr19 of the mm10 reference mouse genome (groups 2 and 3). The three groups of simulations were carried out with 6-base targets (top portion of the table) and then again with 4-base targets (bottom portion). $^b$ Index $i$ identifies uniquely the sequence and its reverse complement for a particular 6-base or 4-base target. Sequences for all possible targets are listed, in order of ascending index, in supplementary files targets6.txt and targets4.txt. $^c$ Forward sequence of the target cleaved in computer simulations of Hi-C experiments. Bases mutated relative to the sequence of the enzyme’s cognate target are shown.

13%. Similar low fractions were also reported by Yaffe and Tanay (2011). These results suggest that different cell types and experimental conditions may yield notably different proportions of non-specific DNA cleavage.

Possible mechanisms of non-specific cleavage in Hi-C experiments are believed to be the star activity of the restriction enzyme and random DNA breakage (Imakaev et al., 2012). To investigate these mechanisms, we determined the histograms of apparent lengths for products generated by our third group of simulations, were product length $L$ was varied randomly. As expected, in the absence of star activity and random breakage, the histogram lacks a upper tail and the largest fraction of products is under the peak of the histogram (Fig. S6A). A long tail and a shrunken peak, however, arise when enzymatic cleavage involves 3 or 9 additional targets (Fig. S6B–D), and these changes increase with increasing simulated star activity (Fig. S6C versus S6D). Similar changes are also seen in the absence of star activity when the fraction $p_0$ of cleavages due to random breakage increases (Fig. S6E–G).

The qualitative similarity in peak shape and tail extent of the histograms obtained from experimental Hi-C data sets, e.g., SRX178473 and SRX118420 in Fig. S5, and some of the histograms obtained from simulated data sets, e.g., panels C and E in Fig. S6, respectively, suggests that both hypothesized mechanisms of non-specific cleavage, star activity and random breakage, may be active in real Hi-C experiments. Our proposed method for estimating cleavage fractions enables quantifying the contribution of each mechanism to the final Hi-C products.
Fig. S1: Estimated cleavage fractions $\hat{r}_i$ (gray bars) converge to corresponding cleavage fractions $\tilde{r}_i$ (blue arrows) measured from Hi-C simulations as the number $M$ of windows used to compute the $\hat{r}_i$’s increases. Each bar is the average of three estimates obtained from independent simulations, with the specified number of windows per estimate. Error bars are standard deviations over the three estimates. Simulations were performed with $p_b = 0$ and using the patterns of enzymatic cleavage probabilities reported in Table S1: (A) “single”, (B) “4 targets”, (C) “10 targets”, (D) “10 evenly”.
Fig. S2: Errors in cleavage fractions estimated from results of simulations with 6-base targets. The residual $R_{CF} = \sqrt{\sum_{i=1}^{N}(\hat{r}_i - \tilde{r}_i)^2}$ between estimated and measured cleavage fractions is plotted as a function of number of windows $M$ used to compute the OLTD from products generated by computer simulation of Hi-C experiments. Each row of plots corresponds to a different fraction $p_b$ of cleavages due to random DNA breakage, as indicated in the first column. Each column of plots corresponds to a different group of simulations. In each group, a particular combination of reference sequence and choice of Hi-C fragment length $L$, i.e., constant or variable, was used as explained in the footnotes of Table S1. See the caption of Fig. 4 in the main text for more details.
**Fig. S3:** Errors in cleavage fractions estimated from results of simulations with 4-base targets. For more details, see the captions of Fig. S2 and Fig. 4 in the main text.
Fig. S4: Errors in estimated fractions of cleavages due to random DNA breakage. The residual $R_b = |\hat{p}_b - \tilde{p}_b|$ between estimated and measured fraction of cleavages due to random breakage is plotted as a function of number $M$ of windows used to compute the OLTD from simulated Hi-C products. The arrangement and details of the plots are the same as those described in Fig. S2. The horizontal dashed line in each plot, where present, represents the value of $p_b$ used in the simulations. Missing points are zero.
Fig. S5: Normalized histograms of apparent Hi-C product lengths determined from Hi-C data sets of Lin et al. (2012) (SRX178471 and SRX178473) and Zhang et al. (2012) (all others) using read pairs uniquely aligned to chr19. All histograms were constructed as described in Yaffe and Tanay (2011), by searching for the first occurrence of the cognate target downstream of each read. Bin size was 10 bases for all histograms. The approximate fraction of products attributable to cleavage only at the cognate target is shown in each plot and equals the area under the peak between the lower and upper bounds indicated by dashed vertical lines. These bounds were determined as described in Section 1.5.
**Fig. S6:** Histograms of apparent Hi-C product lengths obtained through computer simulations generating variable-length products from chr19 of the mm10 reference mouse genome. (A-D) Enzymatic cleavage occurred with the probabilities $p_{i|l}$ shown in Table S1 and there were no cleavages due to random DNA breakage, i.e., $p_b = 0$. (E-G) Enzymatic cleavage occurred only at the cognate target (CT) and cleavages due to random DNA breakage occurred with the indicated values of $p_b$. Percentages and vertical dashed lines have the same meanings as in Fig. S5. Because in this case the distribution of product lengths was known a priori to have mean $\mu_L$ and standard deviation $\sigma_L$, the lower and upper bounds used to calculate the area under the peak were chosen to be $\mu_L - 2\sigma_L$ and $\mu_L + 2\sigma_L$, respectively, which approximate the values that would result from the procedure described in Section 1.5. A bin size of 10 bases was used to construct all histograms.
Table S2. Results of alignments to the mm10 reference mouse genome performed on sequence reads from experimental Hi-C data sets on murine pre-pro-B and pro-B cells, and on mESCs.

|               | SRX178471 | SRX178471‡ | SRX178473 | SRX178473‡ | SRX118420 | SRX118420‡ | SRX118421 | SRX118421‡ | SRX118422 | SRX118422‡ | SRX118423 | SRX118423‡ | SRX118424 | SRX118424‡ | SRX118425 | SRX118425‡ | SRX118426 | SRX118426‡ | SRX116341 | SRX116341‡ | SRX116342 | SRX116342‡ |
|---------------|-----------|------------|-----------|------------|-----------|------------|-----------|------------|-----------|------------|-----------|------------|-----------|------------|------------|-----------|------------|-----------|------------|-----------|------------|-----------|------------|
| input         | 578 183 785 | 578 183 785 | 445 079 258 | 445 079 258 | 99 822 785 | 106 962 722 | 113 929 401 | 45 551 008 | 110 147 984 | 37 947 323 | 248 464 013 | 465 473 330 | 340 651 343 |
| uniquely aligned | 305 787 824 | 305 787 824 | 205 871 336 | 205 871 336 | 38 463 157 | 43 115 522 | 47 476 553 | 14 456 494 | 46 481 393 | 14 138 464 | 94 178 616 | 213 423 650 | 144 493 582 |
| duplicated     | 128 410 198 | 128 410 198 | 29 370 636 | 29 370 636 | 597 728 | 1 131 805 | 1 522 569 | 265 822 | 1 064 177 | 294 173 | 1 772 859 | 3 761 056 | 2 661 865 |
| usable         | 396 405 291 | 396 405 291 | 2 813 099 | 2 813 099 | 777 109 | 864 233 | 1 044 471 | 265 822 | 1 064 177 | 294 173 | 1 772 859 | 3 761 056 | 2 661 865 |
| chr1 or chr19  | 687 674 | 687 674 | 2 813 099 | 2 813 099 | 394 309 | 455 734 | 272 226 | 113 914 | 303 090 | 100 732 | 493 145 | 1 592 661 | 803 591 |

‡ Each column contains counts of read pairs associated with a particular SRA “experiment”. Values in these columns are for reads aligned to chr1. Values in all other columns are for reads aligned to chr19.

a Total number of read pairs contained in the SRA files examined.
b Number of pairs of reads that could both be uniquely aligned to mm10 using Bowtie (Langmead et al., 2009).
c Uniquely aligned read pairs that were discarded as possible artifacts of PCR amplification, because both reads in each such pair were aligned to the same genomic locations as the reads in some other pair.
d Uniquely aligned read pairs that remained after discarding the duplicates.
e Usable read pairs with reads aligned both either to chr1 or to chr19 of mm10.
f Usable read pairs aligned to chr21 but excluding “inward” pairs. These are the numbers of experimental read pairs that were used to estimate cleavage fractions (Table S3) and to compute histograms of apparent product lengths (Fig. S5).

Table S3. Enzymatic cleavage fractions estimated from experimental Hi-C data sets on murine pre-pro-B and pro-B cells, and on mESCs.

| target       | SRX178471 | SRX178471‡ | SRX178473 | SRX178473‡ | SRX118420 | SRX118420‡ | SRX118421 | SRX118421‡ | SRX118422 | SRX118422‡ | SRX118423 | SRX118423‡ | SRX118424 | SRX118424‡ | SRX118425 | SRX118425‡ | SRX118426 | SRX118426‡ | SRX116341 | SRX116341‡ | SRX116342 | SRX116342‡ |
|--------------|-----------|------------|-----------|------------|-----------|------------|-----------|------------|-----------|------------|-----------|------------|-----------|------------|------------|-----------|------------|-----------|------------|-----------|------------|-----------|
| 825 AAGCTT   | 67.04 (.18) | 58.60 (.09) | 49.30 (.11) | 49.00 (.00) | 89.26 (.09) | 89.77 (.08) | 84.88 (.08) | 86.77 (.54) | 86.45 (.06) | 83.03 (.49) | 83.91 (.06) | 83.80 (.09) | 2.18 (.02) | 5.27 (.19) | 2.38 (.09) | 1.17 (.53) | 2.72 (.15) | 3.80 (.03) | 1.71 (.15) | 4.25 (.04) | 2.51 (.05) |
| 821 AA       | 1.71 (.15) | 1.20 (.08) | 3.10 (.07) | 2.38 (.09) | 1.17 (.53) | 2.72 (.15) | 3.80 (.03) | 1.71 (.15) | 1.01 (.02) | 1.80 (.04) | 1.01 (.02) | 2.25 (.04) | 2.51 (.05) | 1.85 (.06) | 2.08 (.12) | 1.01 (.02) | 1.80 (.04) | 1.01 (.02) | 2.25 (.04) | 2.51 (.05) | 1.85 (.06) | 2.08 (.12) |
| 822 AACCTT   | 4.29 (.14) | 4.57 (.13) | 3.35 (.15) | 1.11 (.39) | 2.16 (.17) | 1.38 (.09) | 1.80 (.04) | 1.01 (.02) | 2.25 (.04) | 2.51 (.05) | 1.85 (.06) | 2.08 (.12) |
| 794 AAGGCTG  | 5.83 (.09) | 9.53 (.08) | 7.96 (.02) | 11.10 (.01) | 1.06 (.11) | 1.73 (.07) | 1.01 (.02) | 2.25 (.04) | 2.51 (.05) | 1.85 (.06) | 2.08 (.12) |
| 810 AAGCCTG  | 4.64 (.06) | 3.09 (.02) | 12.69 (.07) | 12.94 (.03) | 1.11 (.39) | 2.16 (.17) | 1.38 (.09) | 1.80 (.04) | 1.01 (.02) | 2.25 (.04) | 2.51 (.05) | 1.85 (.06) | 2.08 (.12) |
| 314 AAGCTA   | 12.74 (.05) | 14.17 (.02) | 11.65 (.07) | 12.25 (.03) | 1.61 (.17) | 2.22 (.17) | 1.14 (.18) | 1.01 (.02) | 2.25 (.04) | 2.51 (.05) | 1.85 (.06) | 2.08 (.12) |

‡ Each column contains values of the enzymatic cleavage fraction $\hat{x} = \hat{p}_i e_{ri}$ obtained by solving Eq. (2) in the main text for the targets assumed to be cleavable by the enzyme. Values are shown as percentages and reported as mean (standard deviation) over three pseudo-samples. Values estimated to be less than 1% are omitted for clarity.

b Index $i$ identifies uniquely each target, with $i = 825$ for the cognate target of the HindIII enzyme.
c Forward sequence of cleaved target. Bases mutated relative to cognate sequence are shown with smaller letters.
d SRA experiments SRX178471 and SRX178473 are from Lin et al. (2012).
e Experiments SRX118420 through SRX118426 are from Zhang et al. (2012).
f Experiments SRX116341 and SRX116342 are from Dixon et al. (2012).
g Target sites known to be cleaved by HindIII star activity (Nasri and Thomas, 1988). h Also obtained by solving Eq. (2) in the main text, estimated fraction of cleavages due to random DNA breakage.
REFERENCES

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.

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.

Iyengar, S. (1980). A computer model for hydrodynamic shearing of DNA — Further investigation on distribution of break lengths: Part III. Computer Programs in Biomedicine, 12, 183–190.

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.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology, 10, R25.

Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B. R., Sabo, P. J., Dorschner, M. O., Sandstrom, R., Bernstein, B., Bender, M. A., Grondin, M., Gurke, A., Stamatoyannopoulos, J., Mirny, L. A., Lander, E. S., and Dekker, J. (2009). Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. Science, 326, 289–293.

Lin, Y. C., Benner, C., Mannson, 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.

Nasri, M. and Thomas, D. (1988). Increase of the Potentialities of Restriction Endonucleases by Specificity Relaxation in the Presence of Organic Solvents. Annals of the New York Academy of Sciences, 542, 255–265.

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.

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.