Supplementary Information

FisHiCal: an R package for iterative FISH-based calibration of Hi-C data

Yoli Shavit*, Fiona Kathryn Hamey and Pietro Lio'

*To whom correspondence should be addressed

Table of Contents

1. Introduction

The growing utility of Hi-C and FISH data will lead to a growing need for FISH-based Hi-C Calibration, for different tissues and for normal and disease conditions. Consequently, FisHical's methods and software will become even more useful and important in the near future. In the following document we provide supplementary information for the manuscript: 'FisHiCal: an R package for iterative FISH-based calibration of Hi-C data' with a comprehensive set of examples and applications.

2. Methods

2.1. Calibration

We suggest here a power law model to relate a set of FISH distances, *D*, and a matching set of contact frequencies, *C*.

$$C \sim \beta D^{\alpha} \tag{1}$$

Taking the log of Eq. (1), gives a linear dependency:

$$\log(C) \sim \log(\beta) + a log(D) \tag{2}$$

and we can estimate a and β with linear regression. Since we assign higher reliability to Hi-C frequencies between segments located in shorter distances, we would like to consider only a subset of matching distances and frequencies for solving Eq. 2. We denote t_r, a reliability threshold that defines this subset, and solve instead:

$$\log(C_{t_r}) \sim \log(\beta) + a log(D_{t_r}) \tag{3}$$

where C_{t_r} and D_{t_r} are the matching subsets of C and D, induced by t_r, respectively. Given the estimated values of a and β we can now apply our calibration model and convert Hi-C frequencies into estimated FISH distances. Since long range frequencies are likely to be noisy, we would like to further discard them as part of the calibration. We define a distance threshold t_n, above which calibrated distances are discarded as non-informative and noisy. Section 7 gives the details for selecting t_r and t_n and provides an analysis of their impact on the calibration and 3D prediction results. Default options are available as part of FisHiCal's calibration model, alongside ad-hoc functionality to refine these thresholds (see Section 3).

2.2. 3D reconstruction and detection of spatial inconsistencies

It is important to note that if we were able to derive the true pairwise 3D Euclidian distances from Hi-C frequencies using our calibration model, then the multi-dimensional scaling (MDS) solution (Torgerson, 1952) would recover the true 3D configuration, up to a rotation. However, Hi-C data are far from providing such a direct measure. Current methods for reconstructing 3D configurations from Hi-C distance approximations typically employ an optimization procedure which aims to minimize the deviation between the input and output distances (Baù *et al.*, 2011; Duan *et al.*, 2010; Zhang *et al*., 2013) or attempt to maximize a likelihood function given the (hypothesized) distribution of the input distances and/or frequencies (Hu *et al*., 2013; Rousseau *et al*., 2011). Here, we aim to minimize the local stress of the predicted configuration through a SMACOF strategy (De Leeuw, 1977), as described in the paper, explicitly addressing the limitations of Hi-C data. The diagonal of the calibrated Hi-C distances matrix is first set to zero and all other zero values (representing discarded information) are replaced with $d_{inf.}$ The value of $d_{inf.}$ can be set according to prior knowledge, for example, based on the radius of the nucleus, or based on available FISH data. The weights can then assigned as described in the paper and used with the latter matrix as input for the SMACOF procedure. In order to further detect spatial inconsistencies, the graph representation of our calibrated Hi-C matrix *G{V, E}* could be generated and searched (testing for inconsistencies for every v∈V). The output of the 3D reconstruction and inconsistency detection could then inform the calibration model, defining an iterative calibration procedure, as explained in the paper.

3. Software implementation (Supp. Table T1)

FisHiCal v1.1 is freely available to install from the R environment and includes comprehensive documentation and examples (?functionName in R, after installing the FisHiCal). Table T1 provides a short description for each of the functions in the package.

Table T1 The functions provided by the FisHiCal package.

Function	Description
findMatchingIndices	Finds matching Hi-C bins for FISH probes, using mid-points.
prepareData	Prepares matching FISH distances and Hi-C frequencies based on matching FISH probe/ Hi-C bin coordinates (that are found with findMatchingIndices).
prepareCalib	Builds the calibration function, by fitting a subset of FISH distances and contact frequencies with a power law model.
updateCalib	Updates the calibration model (useful for updating the noise threshold, see Section 7).
calibrate	Applies a given calibration function on a set of frequencies (typically a Hi-C contact matrix).
getInfoLevelForChr	Computes the information level for a given chromosome after calibration (useful for refining the noise threshold, see Section 7).
lsmacof	Implements the 3D reconstruction algorithm (minimization of a local stress function with SMACOF), described in the paper. The SMACOF component is implemented in c++, using Rcpp (Eddelbuettel and Francois, 2011) and RcppArmadillo (Eddelbuettel and Sanderson, 2014) R packages for R/c++ integration.

Table T1 – Cont.

Function	Description
searchInc	Implements the search for spatial inconsistencies (described in the paper), using the igraph (Csardi and Nepusz, 2006) R package functionality.
summaryInc	Summarizes the details for a list of detected inconsistencies.
plotInc	Plots a neighborhood with inconsistency for a given locus, if such exists.

4. Use case (Supp. Figures S1-S3)

4.1. Data preparation

Hi-C data from 3 human cell lines: IMR90 fibroblasts (Dixon *et al.*, 2012), GM06990 lymphoblasts and K562 erythroleukemia (Lieberman-Aiden *et al,* 2009) were used to generate genome-wide contact matrices of 1 Megabase (Mb) bins, using the chromoR R package and as described by Shavit and Lio' (2014).

4.2. Noise correction

In order to address noise and bias in Hi-C we have applied a method that we have previously developed (implemented as part of the chromoR package), shown to outperform current correction methods (Shavit and Lio', 2014). The resulting corrected matrices were used as input for Hi-C calibration, as described below.

4.3. Calibration

FISH pairwise distances[1] from Human primary fibroblasts between probes in chromosomes 1 and 11 (Mateos-Langerak *et al*., 2009) were matched to Hi-C frequencies with FisHiCal::prepareData based on mid-points of the FISH probes/Hi-C bins. When several FISH probes were mapped to the same 1Mb bin *i*, the FISH distances for this bin with another bin *j* were not unique. In these cases FisHiCal::prepareData takes the minimal non-zero FISH distance between *i* and *j* as representative, in order to generate a unique match and since Hi-C is likely to be biased towards shorter distances. In total, 40 pairwise contact frequencies and FISH distances were used to estimate the calibration for each cell line (39 pairs for IMR90).

We have set the reliability threshold (t_r) to 0.1, taking the subset of 4 (10%) shortest FISH distances and their matching Hi-C counterparts and used FisHiCal::preapreCalib to estimate α, β and t_n for each cell line (see Section 7 for additional details on selecting t_r and t_n). Results (Fig. S1) showed that our model captures FISH/Hi-C dependency across all 3 human cell lines, with the best fit for the fibroblasts cell line and as expected (same cell type for FISH and Hi-C). The estimated slope (α) was -2.59, -2.05 and -2.25 for IMR90, GM06990 and K562 correspondingly, decreasing (absolute value) with sequencing depth and suggesting that resolution increase may not be uniformly distributed but rather provide more information for short range frequencies (see also Section 8 for further analysis of power law behavior). A consistent outlier was detected at 1.49µm between highly transcribed genes in chromosome 1 (Mateos-Langerak *et al*., 2009): probe R21 (chromosome 1: 152,647,445 - 152,789,056), mapped to the genes IL6R, SHE and TDRD10, and probe R33 (chromosome 1: 153,867,613-154,044,636), mapped to the genes YY1AP1, DAP3, MSTO2P and GON4L.

4.4. 3D reconstruction

In order to further explore the spatial basis of the outlier described above we have calibrated the IMR90 Hi-C frequencies with FisHiCal::calibrate and reconstructed the 3D configuration with FisHiCal::lsmacof. We have first verified that the predicted structure presents a spatial segregation as described by Lieberman-Aiden *et al.* (2009), where the researchers showed that chromosomes could be partitioned into 2 compartments, each forming a contact cluster. This pattern emerged from the correlation matrices of the Hi-C contact maps and was further analsyed with Principle Component Analysis (PCA), showing that the first principle component can capture chromosomal decomposition (with positive values corresponding to the first compartment and negative values corresponding to the second compartment). We have repeated this analysis for chromosome 1 and colored loci in the reconstructed 3D configuration

[1] Taking the mean of multiple measurements for each pair.

according to the first principle component (Fig. S2) resulting in a clear partition that is consistent with the Hi-C decomposition. The reconstructed structure provided a possible explanation for the outlier, where the successive 1Mb bins mapped to FISH probes R21 and R33 (encircled in Fig. S2-b) form part of a chromatin loop, that result in relatively high contact frequency values (as Hi-C captures all 1Mb intra-loop contacts) while the FISH probes are at the far ends of the loop due to transcription activity (Mateos-Langerak *et al*., 2009).

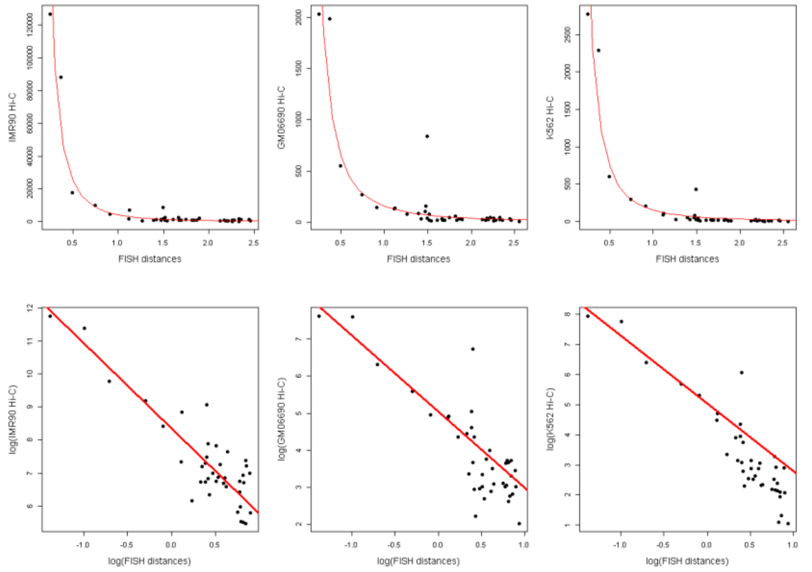


Fig. S1 Hi-C frequencies from 3 human cell lines and matching FISH distances from human fibroblasts are plotted in their original (upper panel) and log-log scale (lower panel), with the estimated calibration curve (in red).

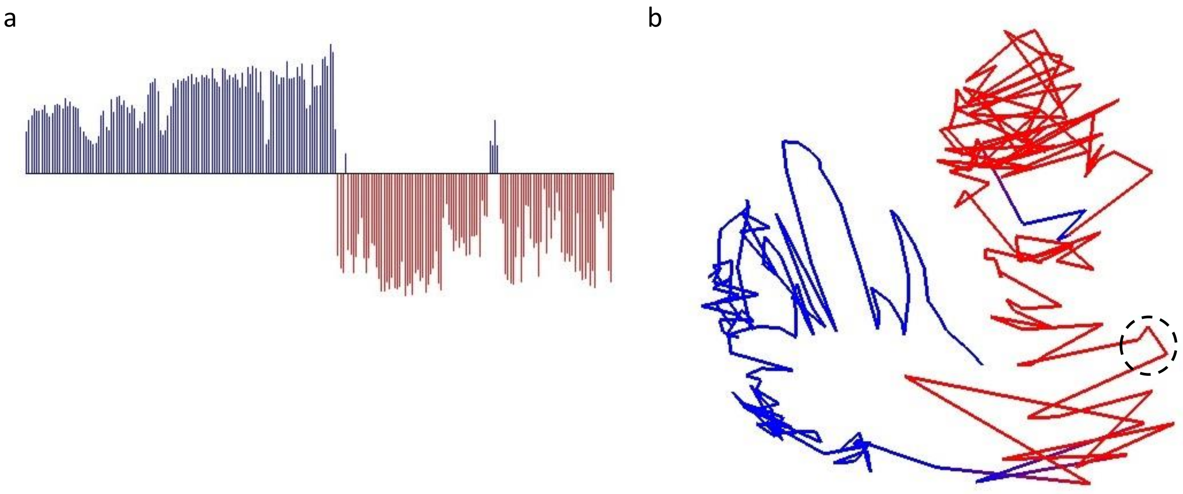


Fig. S2 A predicted structure for chromosome 1 provides a spatial explanation to calibration's outlier. FISH distances from human fibroblasts (Mateos-Langerak *et al*., 2009) were used to calibrate human IMR90 fibroblasts Hi-C (Dixon *et al.*, 2012) at a 1Mb resolution, with FisHiCal::calibrate. The 3D structure of chromosome 1 was then reconstructed with FisHiCal::lsmacof from the calibrated distances. The values of the first principle component of the correlation Hi-C matrix (a) were used to color loci in the reconstructed structure (b), with blue and red for positive and negative values correspondingly, confirming the known partition into 2 chromatin compartments (Lieberman-Aiden *et al*., 2009). The reconstructed structure (b) further provided a possible explanation for the calibration's outlier (Fig. S1). The successive 1Mb bins (encircled in black), mapped to probes R21 (chromosome 1: 152,647,445 - 152,789,056) and R33 (chromosome 1: 153,867,613-154,044,636), formed part of a chromatin loop, resulting in a relatively high contact frequency values (as Hi-C captures all 1Mb intra-loop contacts), while the FISH probes could be located at the far ends of the loop due to high transcription activity (Mateos-Langerak *et al*., 2009).

4.5. Spatial inconsistencies

Further examining the inconsistencies detected for loci in chromosome 1 with FisHiCal::searchInc, highlighted a long genomic domain (184-196Mb), which was in contact with 2 regions: 61-62 Mb in chromosome 7 and 69-70Mb in chromosome 16, that were not connected themselves. Fig. S3 presents the plot of all 12 inconsistencies (1 for each 1Mb loci in the region 184-196Mb), generated with FisHiCal::plotInc, and suggesting that the loci in chromosome 16 and chromosome 7 are in contact with different instances of the genomic domain in chromosome 1, correspondingly.

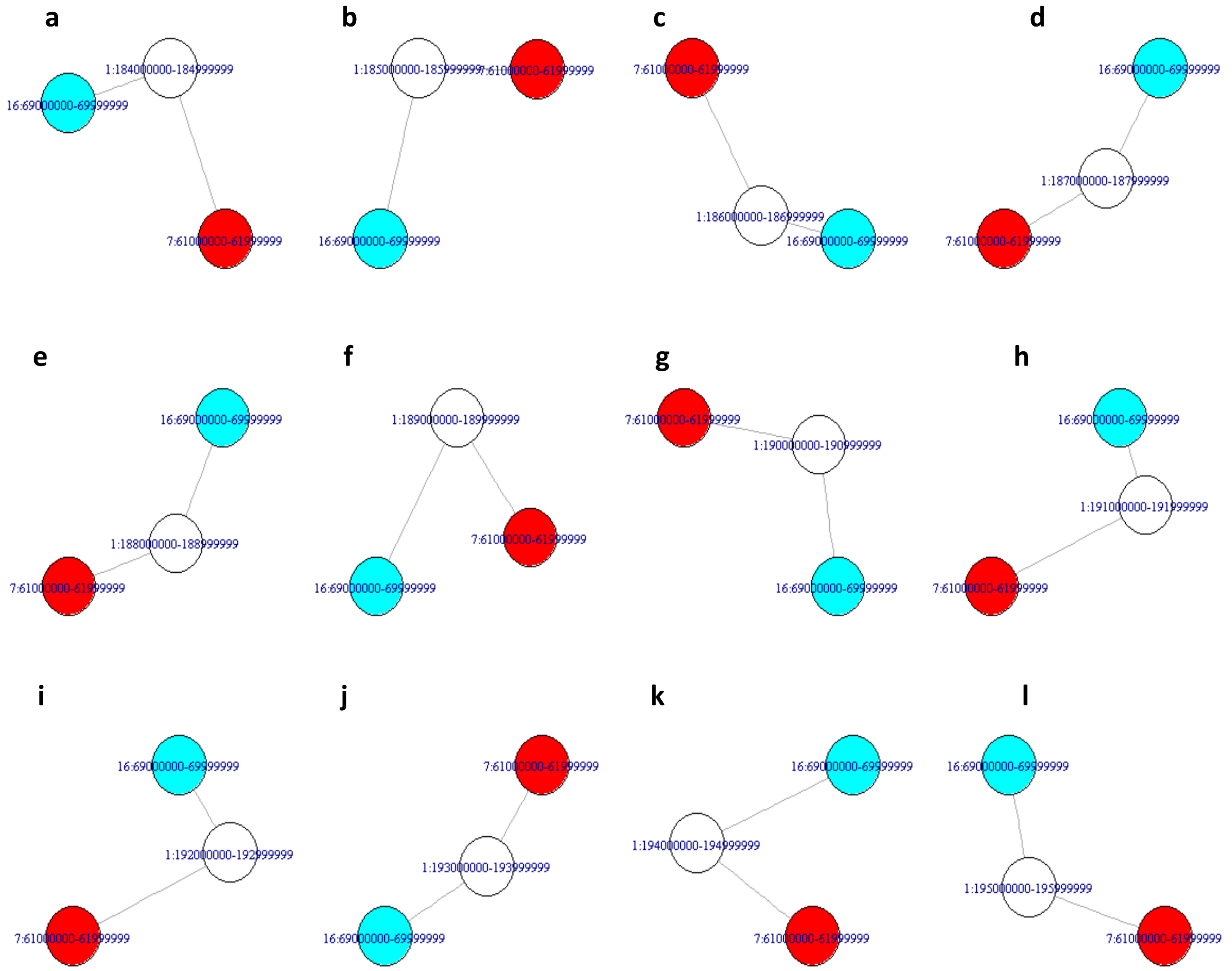


Fig. S3. Spatial inconsistencies detected with FisHiCal::searchInc for 12 consecutive 1 Mb regions in chromosome 1 (a-l) at 184-196Mb. All 12 regions presented a contact with 2 regions in chromosome 7 and 16 correspondingly, that are not themselves connected. Each spatial inconsistency (a-l) was plotted as a graph with FisHiCal::plotInc, where nodes are colored by their connected component (the node corresponding to the region under consideration is colored white) and labeled with their genomic coordinates.

5. Additional applications (Supp. Figures S4-S7)

FisHiCal implements a calibration model for Hi-C data along with 2 key applications: 3D reconstruction and detection of spatial inconsistencies that may occur due to the lack of homology distinction in Hi-C. Here, we discuss several other applications that could be performed with FisHical and prove useful for researchers.

5.1. Quality and reproducibility assessment

Hi-C data include bias and noise, and several methods were developed to address this problem (Cournac *et al*., 2012; Hu *et al*., 2012; Imakaev et *al*., 2012; Shavit and Lio', 2014; Yaffe and Tanay, 2011). Ideally, artefact identification and correction should be carried in light of FISH data. Such verification is now made possible with FisHiCal. Departures of Hi-C frequencies from the expected model can be mapped to specific locations so that researchers can investigate whether these are a result of noise attributed to the Hi-C protocol (for example due to range limitation) or whether they represent meaningful deviation (as explained in Section 4.4 and Section 7). The discrepancy between FISH and Hi-C, captured and measured by the calibration, could further point to functionally important events.

The reproducibility of Hi-C could also be assessed by comparing calibration curves and predicted 3D models of experimental replicates, as shown in Figure S4. Here, we have compared the calibration curves of 2 IMR90 replicates. The curves presented a similar fit, consistent with previous comparison results that detected zero significant changes (Shavit & Lio', 2014). As expected, the fit varied more for long range distances due to noise and range limitation. When comparing the predicted 3D models, we found a strong distance correlation across all chromosomes (Pearson correlation, *r* = 0.9±0.05), compared to a lower correlation between models from different cell types (Pearson correlation, *r* = 0.73±0.08; comparing IMR90 and K562 replicates).

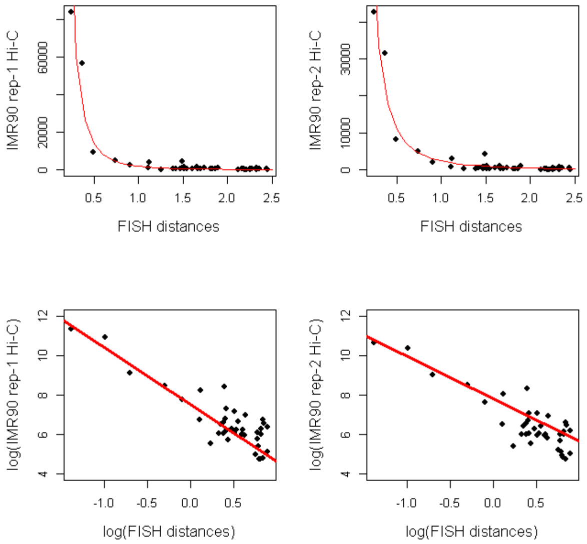
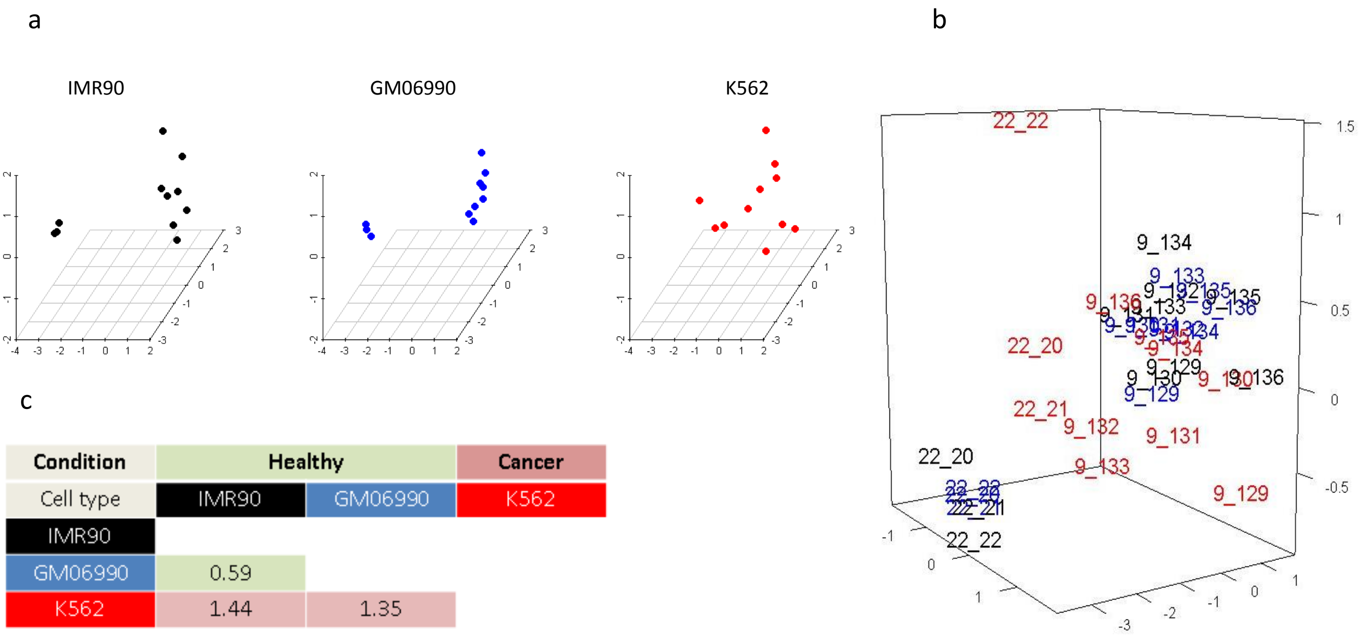


Fig. S4 Hi-C frequencies of 2 IMR90 replicates and matching FISH distances from human fibroblasts (39 matching pairs in total) plotted in their original (upper panel) and log-log scale (lower panel), with the estimated calibration curve (in red), showing a similar behavior.

5.2. 3D Cytogenetics

Using FisHiCal's calibration and 3D prediction functionalities, researchers can carry cytogenetic analysis in 3D. In order to illustrate this idea we have reconstructed the combined 3D structure of chromosome 9 and 22 in IMR90, GM06990 and K562 and examined the ABL/BCR gene fusion (involving the translocation t(9;22)(q34;q11)), typical of the K562 Karyotype (Naumann *et al.*, 2001). The 11 1 Mb regions around this fusion domain (chromosome 9: 129-137Mb, chromosome 22: 20-23Mb) showed a clear chromosome separation in the healthy cell lines (IMR90 and GM066900), and were instead co-located in the K562 cell line (Fig. S5-a). Further comparing the healthy and disease configurations resulted with larger Root Mean Square Deviation (RMSD) values, as expected (Fig. S5-c). Finally, the superimposition of the 3D configurations (Fig. S5-b) pinpointed the translocated regions (chromosome 9: 132-134Mb, chromosome 22: 20-22Mb) in the domain under consideration. We note that a similar analysis could be performed at a genome-wide level (owing to the scalability of Hi-C) in order to identify cytogenetic aberrations and further evaluate the magnitude of co-locations and their link to the clinical phenotype.



Condition	Healthy		Cancer
Cell type	IMR90	GM06990	K562
IMR90			
GM06990	0.59		
K562	1.44	1.35	

Fig. S5 A 3D investigation of 11 1 Mb regions around the ABL/BCR gene fusion domain (chromosome 9: 129-137Mb, chromosome 22: 20-23Mb) in IMR90 (black), GM06990 (blue) and K562 (red). (a) Scatterplots of the 11 1Mb regions under consideration showed a clear separation between regions in chromosome 9 (right-side cluster) and chromosome 22 (left side cluster) in IMR90 and GM06990, that were instead co-located in K562. (b) A 3D plot showing the identity of each region (indicated by “chromosome name_Mb position”) further pinpointed the translocated regions in K562 (red): chromosome 9: 132-134Mb and chromosome 22: 20-22Mb. (c) The comparison of the 3D configurations provided larger RMSD values when comparing healthy and disease conditions, as expected. A similar analysis could be performed at a genome-wide level in order to identify cytogenetic aberrations and further evaluate the magnitude of co-locations and their link to the clinical phenotype.

5.3. Chromosomal maps with scale

A key contribution of a FISH-based Hi-C calibration is in its ability to provide a scale for chromosomal maps. Specifically, the scaling value derived from the calibration can help researchers in reading and accurately interpreting the 3D configurations generated from calibrated Hi-C data (for example, with FisHiCal::lsmacof). Figure S6 presents a 3D map of 6 1Mb regions in chromosome 11, located at short-medium ranges, that were positioned with FisHiCal. The map includes an approximated scale (0.082µm) for 0.1 predicted distance unit, which was computed by taking the mean ratio between the known FISH distances and the predicted distances.

5.4. Spatio-temporal calibration (time series calibration)

Given time-series FISH and Hi-C data, researchers can study spatio-temporal changes. Even in the absence of complete data, calibration curves can still provide valuable information. For example, given FISH data at time t = 0 and matching time-series Hi-C data, the calibration curves can highlight spatial changes over time. Fig S7-a provides the expected calibration curves in a case of expansion. While FISH distances are available only for time t = 0, Hi-C frequencies decrease over time (due to increasing distances), and the expansion is immediately evident from the resulting curves. When instead, both FISH and Hi-C data are given over time, 3D models can be reconstructed (Fig S7-b) and used to further study spatio-temporal behavior, for example, with spatial statistics.

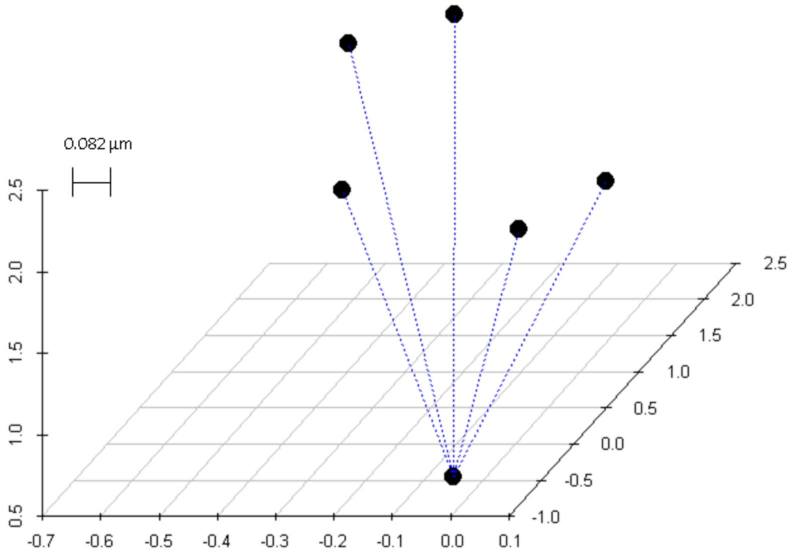


Fig. S6 FISH based Hi-C calibration can provide a scale for chromosomal maps. A 3D map of 6 1Mb regions in chromosome 11 is presented with a scale (for a 0.1 unit distance), derived from the ratio of FISH distances (dashed lines) and predicted (FisHical) distances, allowing for a more accurate interpretation.

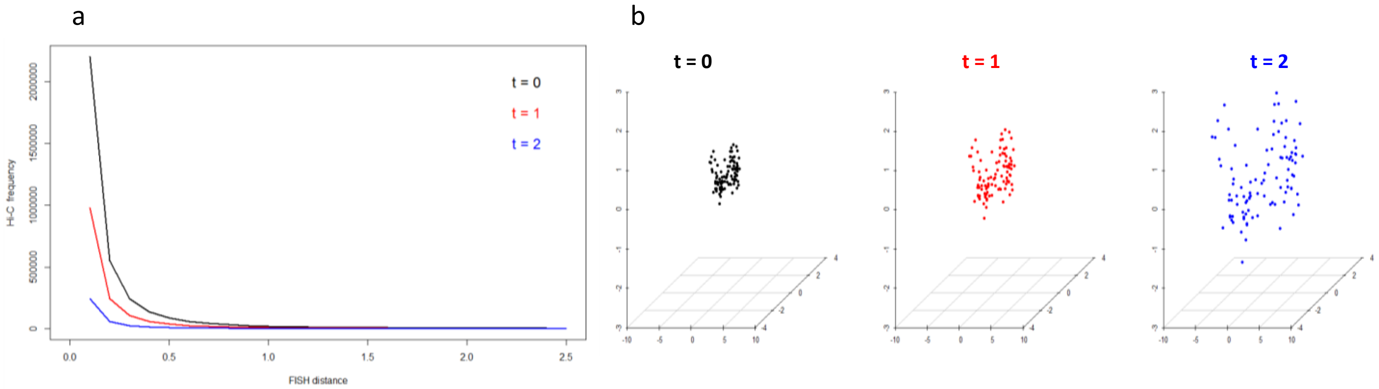


Fig. S7 Spatio-temporal FISH-based Hi-C calibration. In the absence of time series FISH data, calibration curves of Hi-C data at different time points could still provide useful information. Here, the calibration curves in the event of expansion are shown, where Hi-C frequency decrease due to increasing distances. (b) When both FISH and Hi-C data are available over time, researchers can also study the resulting 3D models with spatial statistics.

6. Performance evaluation (Supp. Figures S8-S9, Table T2)

6.1. *in-silico* analysis

In order to test the robustness of our calibration and 3D reconstruction we have evaluated the accuracy of *in-silico* predicted configurations, with increasing levels of noise. We have first generated 100 random configurations, as described by Hu *et al* (2013), in order to create a benchmark of *in-silico* chromosomal configurations. Specifically, following a random walk/giant loop model suggested by Sachs *et al* (1995), we have modelled a chromosomal backbone with a 3D configuration of *n* loci: $\{<x_i, y_i, z_i>, i = 1..n\}$ where the differences between successive coordinates at each axis follow a normal distribution *N(0,1)* and we scale the configuration to ensure that the distance between the first and last loci is 1. We have next generated the empirical RMSD distribution from the *in-silico* benchmark (4950 pairwise RMSD scores), where the RMSD between 2 configurations A, B of size *n* is defined as RMSD(A, B)

$= \sqrt{\frac{\sum_i^n (x_i^A - x_i^B)^2 + (y_i^A - y_i^B)^2 + (z_i^A - z_i^B)^2}{n}}$ where A is first superimposed on B with the Partial Procrustes Superimposition (and we also test for reflection).

Given a randomly generated configuration, contact frequencies can be simulated with a power law model. Random Noise can then be added to all long range contact frequencies up to a given quantile in order to mimic the Hi-C range limitation, where noise is randomly selected from a uniform distribution between 0 and the frequency matching the distance of the noise quantile (for example, 5% noise will result in manipulating all frequencies corresponding to distances at the upper 95% percentile, where noise is sampled between 0 and the 95% distance quantile). Following this procedure, we have generated 100 random configurations and matching contact frequencies with increasing levels of noise (100 configurations and matching contact matrices for each noise value) and then used FisHiCal to calibrate frequencies (taking the 10 shortest distances (10%) for calibration and setting the threshold to be the maximal available distance). FisHiCal::lsmacof was then used to predict the true structure and the RMSD and the Pearson correlation between the true and predicted pairwise distances were calculated. The significance was evaluated using the normalized RMSD, in order to correct for scaling, (as suggested by Hu *et al.*, 2013) - RMSD(A', A)/d99th(A') , where A and A' are the true and predicted configurations respectively, and d99th, is the 99th percentile pairwise distance in A'. The results of this analysis (Fig. S8) showed that FisHiCal can significantly infer the true structure even for high levels of noise and missing information (up to 65%) while achieving a high correlation (r > 0.75) with the true pairwise distances. We further provide here examples of superimposed true and predicted configurations for representative noise levels (Fig. S9) illustrating the impact of noise on calibration and 3D reconstruction.

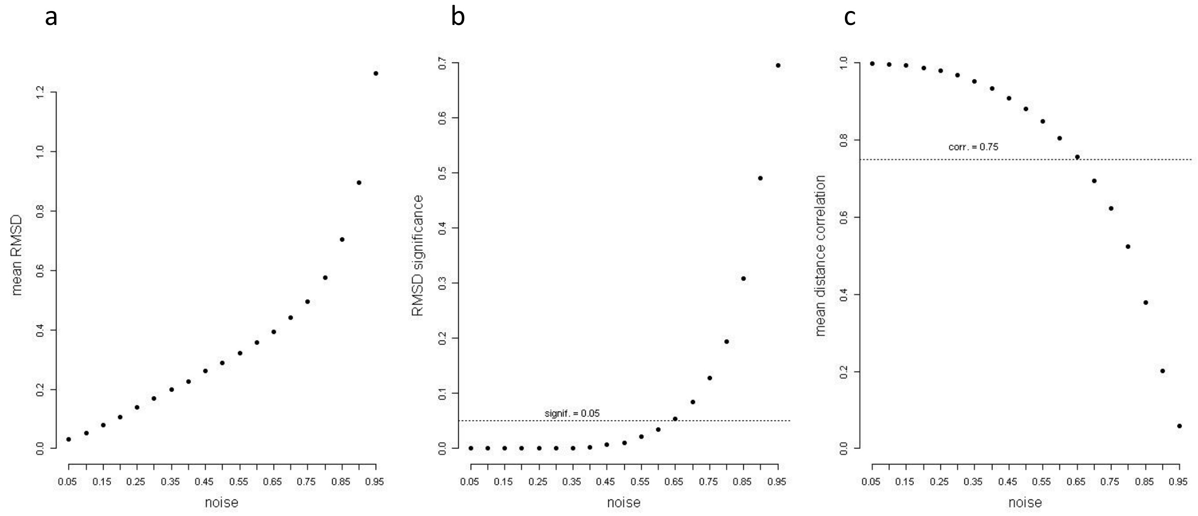


Fig. S8 Mean RMSD (a), significance (b) and correlation (c) for 100 predicted configurations with increasing levels of noise.

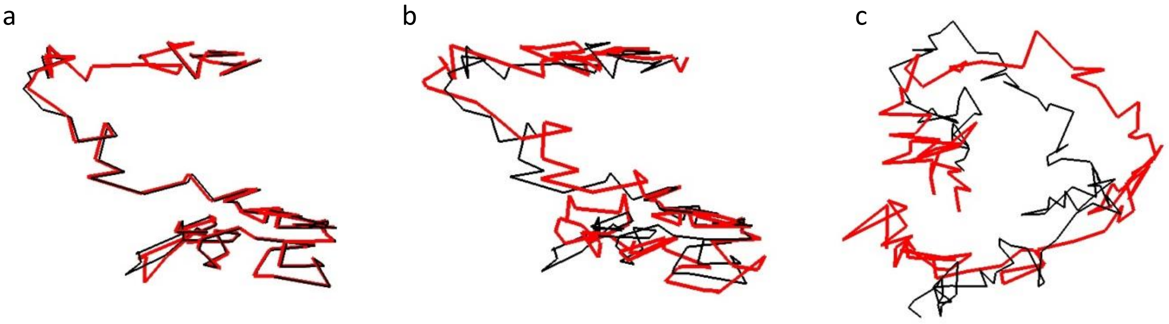


Fig. S9 Examples of superimposed true and predicted *in-silico* configurations (reconstructed with FisHiCal::lsmacof and generated as described above) for representative levels of noise: 5% (a), 25% (b) and 60% (c).

6.2. Running times

We have recorded the running times of calibration (FisHiCal::calibrate), inconsistencies search (FisHical::searchInc) and 3D reconstruction (FisHiCal::lsmacof) on a Windows 7 64 bit OS, with an Intel (R) QUAD CPU Q6600 2.39GHz processor, and 6GB RAM. These tasks were performed for the human fibroblasts data (as described above), for the genome wide Hi-C matrix (all 3 tasks) and for single chromosomes (3D reconstruction only). Table T2 provides the results of this evaluation (also providing the parameters used for each task), showing short running times across tasks. For the genome-wide 3D reconstruction, running times were exponentially longer than for single chromosomes, due to the large amount of missing information (>95%) after calibration, that significantly slowed down the convergence with the given threshold. Nevertheless, we note that the achieved running time is feasible for genome-wide reconstruction (and specifically when compared to current solutions).

Table T2 Summary of running times for calibration, inconsistencies search and 3D reconstruction.

Task	Scale	Number of loci	Parameters	Running time (seconds)
FisHiCal::calibrate	Genome-wide	2880	Calibration threshold: 3.45	2.72
FisHiCal::searchInc	Genome-wide	2880	N\A	3.90
FisHiCal::lsmacof	Genome-wide	2880	d_{Inf} : 4.5,	6055.31
	Chromosome 1	229	convergence threshold: 1e-06	8.31
	Chromosome 2	241		10.34
	Chromosome 3	197		2.92
	Chromosome 4	190		1.36
	Chromosome 5	179		1.22
	Chromosome 6	169		0.91
	Chromosome 7	157		1.75
	Chromosome 8	145		1.5
	Chromosome 9	124		2.42
	Chromosome 10	135		0.54
	Chromosome 11	133		0.38
	Chromosome 12	132		0.37
	Chromosome 13	98		0.69
	Chromosome 14	89		0.35
	Chromosome 15	83		0.59
	Chromosome 16	81		0.3
	Chromosome 17	79		0.15
	Chromosome 18	77		0.33
	Chromosome 19	57		0.07
	Chromosome 20	62		0.14
	Chromosome 21	36		0.01
	Chromosome 22	36		0.05
	Chromosome X	151		3.02

7. Selecting calibration thresholds (Supp. Figures S10-S11)

Hi-C frequencies provide a proxy for FISH distances. However, even in the noise-free case, long-range frequencies are less (or not at all) informative due to the limited range of Hi-C capture. Thus, we would like to consider only the subset of shortest distances (defined by the reliability threshold t_r). Since we assume that reliability decrease with distance, we would like to select the minimal subset that will still provide us with a good estimation for our calibration model. Figure S10 shows the calibration curves for increasing sizes of subsets, presenting a similar behaviour and thus suggesting that a small proportion of distances would suffice to capture Hi-C/FISH dependency (10%), with the advantage of noise reduction. In order to support maximum flexibility, FisHiCal::prepareCalib can take either the size of the subset or a predefined set for cases where some specific distances should not be used for parameter estimation.

Given an estimation of the parameters of our model, a key aspect is to evaluate the noise threshold t_n, (the distance value above which we discard calibrated distances as non-informative/noisy) which corresponds to a trade-off between error-prone data and sparsity (missing information). In order to analyse the impact of t_n we have generated *in-silico* chromosomal conformations with increasing levels of noise, as described in Section 6, and computed the RMSD while decreasing the amount of information (i.e. gradually discarding more distances). The results of this analysis (Fig S11-a) suggested that information levels in the range of [0.6, 0.9] provide a good trade-off, across different noise percentages. We have further compared the information levels achieved with 2 thresholds: a threshold set according to the maximum available FISH (set here to 3.45 μm), suitable for cases where FISH data is limited, and a threshold computed from the fit of the data and the model (2.45 μm):

$$maxD\left(\left|exp^{\frac{\log(C)-\log(\beta)}{\alpha}} - D\right| < \varepsilon\right) \tag{3}$$

Results (Fig. S11-a) showed that together these thresholds achieved information levels at the optimal trade-off range, for all chromosomes, suggesting that they are suitable even for noisy cases. Since the first threshold (black circles in Fig. S11-b) provided a better level of information for chromosome 1, we have chosen it for our use case example (Section 4). While FisHical provides users with the choice of these thresholds (see documentation of FisHiCal::prepareData and FisHiCal::prepareCalib), the analysis described here could be carried in order to guide threshold selection and refinement with FisHiCal:getInforLevelForChr and FisHiCal:updateCalib.

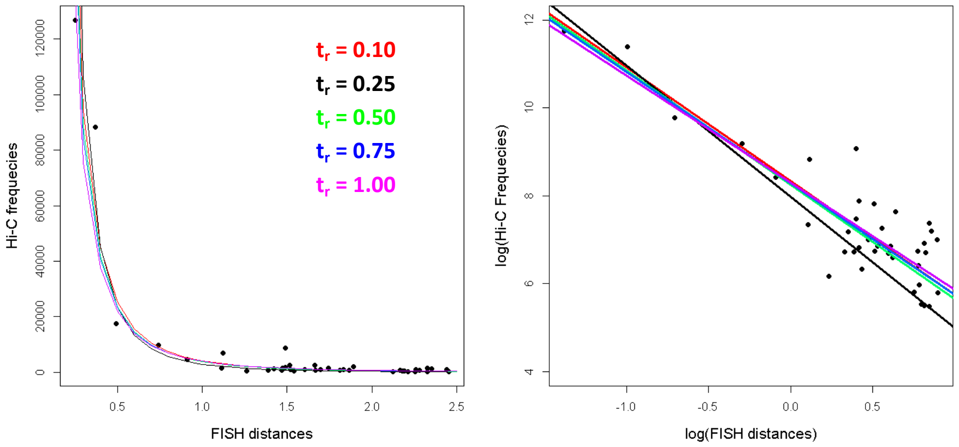


Fig. S10 Calibration curves generated for increasing subsets of distances (increasing t_r values, indicating the proportions of distances considered) presented in their original (left side panel) and log-log (right side panel) scales.

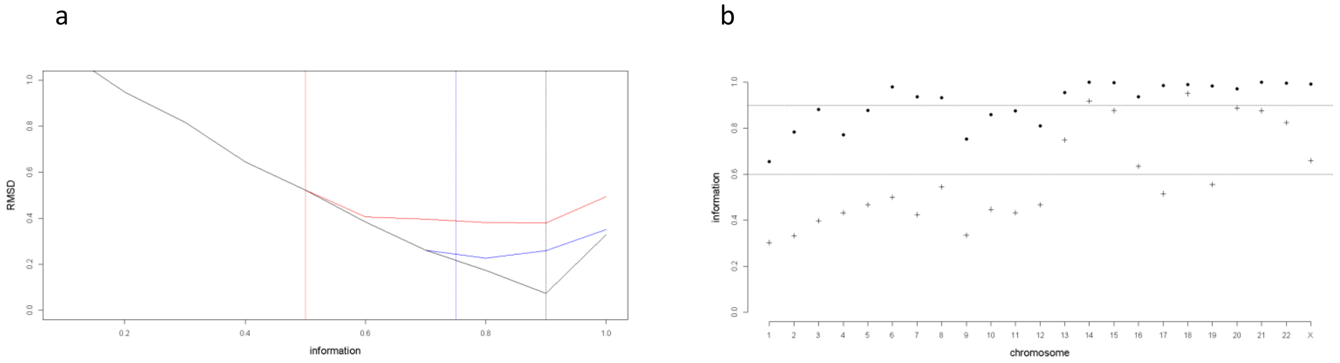


Fig. S11 Assessment of t_n (noise threshold) impact on prediction accuracy. (a) RMSD values decrease as the level of information increases, with a best trade-off in the range [0.6,0.9] across different levels of noise, indicated with a dashed line: 50% (red), 75% (blue) and 90% (black). (b) Information levels across chromosomes in IMR90, given 2 noise thresholds: a threshold set according to the maximum available FISH (circles) and a threshold computed from the fit of the data and the model (Eq. 3; plus symbols).

8. Studying power law behavior (Supp. Figure S12)

Although we assume that short range distances provide more reliable Hi-C frequencies, FisHiCal is designed so that users can explore any subset of distances (see the documentation for FisHiCal::prepareData). This could be important, for example, for studying different regimes that are attributed to open and close chromatin and to looping events (Barbieri *et al.*, 2012). Fig. S12 presents 2 calibration curves that were estimated using different subsets of distances. The first curve (red) was generated as described in section 3 (taking a subset of shortest distances), while the second (blue) was estimated by taking the 2 shortest distances and the outlier investigated in Section 4. The resulting curves show different power law behaviours that may also represent different looping regimes. Since FisHiCal::calibrate can receive any calibration function users may also refine the calibration specification generated by FisHiCal::prepareCalib according to their analysis (see also FisHiCal::udpateCalib).

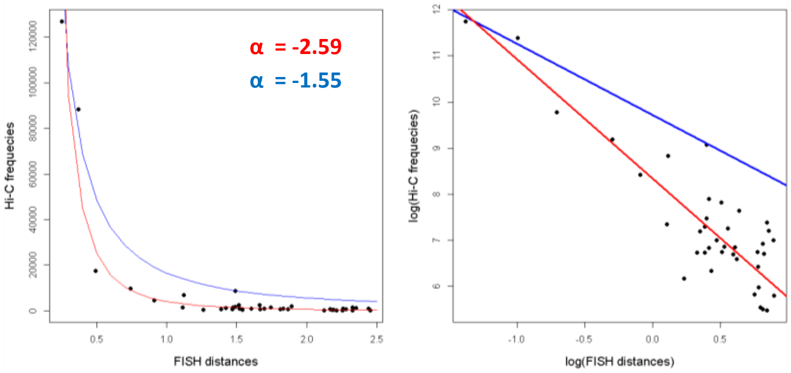


Fig. S12 Calibration curves (presented in their original and log-log scale) show different power law behavior. The red curve represents the default calibration, taking a subset of shortest distances, while the blue curve instead considers the outlier.

9. References

Barbieri, M. *et al.* (2012) Complexity of chromatin folding is captured by the strings and binders switch model. *PNAS*, **109**, 16173–8.

Baù, D., *et al*. (2011) The three-dimensional folding of the α-globin gene domain reveals formation of chromatin globules. *Nat Struct Mol Biol*,. **18**, 107–14.

Cournac, A. *et al*. (2012) Normalization of a chromosomal contact map. *BMC genomics*, **13**, 436.

Csardi, G. and Nepusz, T. (2006) The igraph software package for complex network research. *InterJournal, Complex Systems*, 1695.

De Leeuw, J. (1977). Applications of convex analysis to multidimensional scaling. In Barra, J.R et al (Eds.), Recent developments in statistics, 133–145. Amsterdam, The Netherlands: North-Holland.

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

Duan Z. *et al*. (2010) A three-dimensional model of the yeast genome. *Nature*, **465**, 363–67.

Eddelbuettel, D. and Francois R. (2011). Rcpp: Seamless R and C++ Integration. *Journal of Statistical Software*, **40**, 1-18.

Eddelbuettel, D. and Sanderson, C. (2014). RcppArmadillo: Accelerating R with high-performance C++ linear algebra. *Computational Statistics and Data Analysis*, **71**, 1054-1063.

Hu, M. *et al*. (2012) HiCNorm: removing biases in Hi-C data via Poisson regression. Bioinformatics, **28**, 3–5.

Hu M. *et al*. (2013) Bayesian Inference of Spatial Organizations of Chromosomes. *PLoS Comput Biol*, **9,** e1002893.

Imakaev, M. *et al*. (2012) Iterative correction of hi-c data reveals hallmarks of chromosome organization. *Nat Methods*, **9**, 999-1003.

Lieberman-Aiden E. *et al.* (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, **326**, 289–93.

Mateos-Langerak J. *et al.* (2009) Spatially confined folding of chromatin in the interphase nucleus. *PNAS*,, **106**, 3812-7.

Naumann, S. *et al* (2001) Complete karyotype characterization of the K562 cell line by combined application of G-banding, multiplex-fluorescence in situ hybridization, fluorescence in situ hybridization, and comparative genomic hybridization. *Leukemia research*, **25**, 313–22.

Rousseau M. *et al.* (2011) Three-dimensional modeling of chromatin structure from interaction frequency data using Markov chain Monte Carlo sampling. *BMC bioinformatics*, **12**, 414.

Sachs, R.K *et al.* (1995) A random- walk/giant-loop model for interphase chromosomes. *PNAS*, **92**, 2710–14.

Shavit, Y. and Lio' P. (2014) Combining a wavelet change point and the Bayes factor for analysing chromosomal interaction data. *Mol. BioSyst*, **10**, 1576-85.

Torgerson, W.S. (1952). Multidimensional scaling: I. Theory and method. *Psychometrika*, **17**, 401–19.

Yaffe,E. and Tanay, A. (2011) Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nat. Genet.*, **43**, 1059–65.

Zhang, Z. *et al.* (2013) 3D chromosome modeling with semi-definite programming and Hi-C data. *J Comput Biol*, **20**, 831–46.