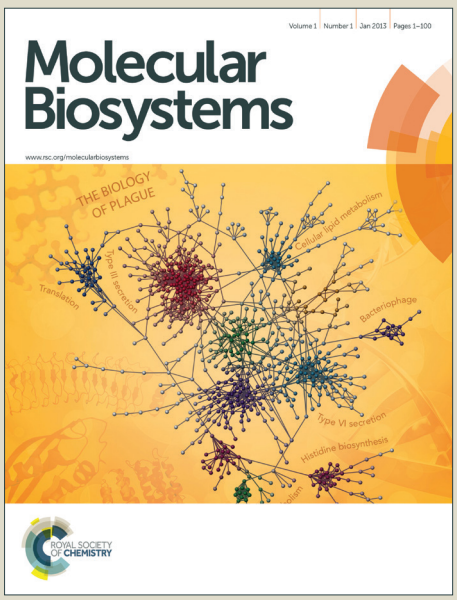
Molecular BioSystems

Accepted Manuscript



ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/molecularbiosystems

Cite this: DOI: 10.1039/c0xx00000x

www.rsc.org/xxxxxx

ARTICLE TYPE

Combining wavelet change point and Bayes Factor for analysing chromosomal interactions data

Yoli Shavit,*[a] Pietro Lio'[b]

Received (in XXX, XXX) Xth XXXXXXXXX 20XX, Accepted Xth XXXXXXXXX 20XX

DOI: 10.1039/b000000x

Over the past decades we have witnessed great efforts to understand the cellular function at the cytoplasm level. Nowadays there is a growing interest in understanding the relationship between function and structure at the nuclear, chromosomal and sub-chromosomal levels. Data on chromosomal interactions that are now becoming available in unprecedented resolution and scale are opening the way to address this challenge. Consequently, there is a growing need for new methods and tools that will transform these data into knowledge and insights. Here, we have developed all the steps required for the analysis of chromosomal interactions data (Hi-C data). The result is a methodology which combines wavelet change point with Bayes Factor, for useful correction, segmentation and comparison of Hi-C data. We have further developed chromoR, an R package that implements the methods presented here. chromoR provides researchers with a means to analyse chromosomal interactions data using statistical bioinformatics, offering a new and comprehensive solution for this task.

Introduction

The Hi-C procedure detects and quantifies long range chromosomal interactions in a population of cells, without constraining itself to a specific set of primers[1]. In this way, interaction frequencies between all genomic loci are reported, where a high interaction frequency is a proxy for spatial proximity (on average). With Hi-C data, efforts, so far focused at the cytoplasm level, are now extending to the nucleus compartment in order to understand the relationship between function and structure at the nuclear, chromosomal and sub-chromosomal levels.

However, Hi-C data are also noisy and biased[2], where coverage of different regions vary, resulting with an over- and under- representation of interaction frequencies. Recently, several methods were developed to correct contact Hi-C maps, explicitly or implicitly stating the sources of bias[2-5]. However, neither of these methods considered the multiple scales at which noise and bias may appear. For example, Guanine-Cytosine (GC) genome composition, that was shown to be a source of bias in Hi-C data[2], presents different scales of heterogeneity, such as genes and isochores[6-8] (See Supp. Text T1 for a complementary analysis we have performed to evaluate GC based bias in Hi-C data at different scales). Addressing variance of coverage at different scales could provide an improved way for correcting bias in Hi-C data.

The next step, after correcting for bias and noise, is to detect changes in interaction frequencies that reflect changes in structure. 1-dimensional (1D) contact profiles, generated for a given chromosome by taking the row sums of its contact map with another chromosome (or with itself), can be used to detect such changes. This could be important, for example, for identifying chromosomal aberrations such as deletions, duplications and translocations, and shed light on their predisposition mechanism, previously shown to correlate with spatial proximity[9-10].

Finally, contact maps can be compared in order to detect regions that are interacting significantly different under different conditions or over time. Such a comparison can pinpoint genomic loci and changes in interaction that are important for function.

To the best of our knowledge, there is currently no available methodology for performing change point analysis and comparison, as presented here, or for following the entire pipeline of correction, segmentation and comparison of Hi-C data. Here, we address this need and provide the methods and software (chromoR, see Software Implementation section) required for analysing chromosomal interactions data (Fig. 1):

1 We explain how wavelet variance stabilization can address the multi-scaled variance in Hi-C coverage and apply this method to correct Hi-C contact maps (Fig. 1c, Methods: section 1).
2 We show that wavelet change point analysis is a useful tool for segmentation of Hi-C 1D contact profiles and present its application for detecting chromosomal aberrations (Fig. 1f, Methods: section 2).
3 We describe how the Bayes Factor can be adapted for comparison of Hi-C contact maps and provide a comparison methodology for detecting significant changes (Fig. 1h, Methods: section 3).

These methods provide together a comprehensive solution while independently addressing one problem at a time. In addition, their output can be further coupled with existing tools that address related tasks, such as visual comparison[11] and integration[12] of Hi-C data (Fig. 1j)

In the next sections we provide the technical details of our methods followed with key examples and tests used to verify their performance. For the sake of space we provide additional information in the supplementary materials (referred in the text).

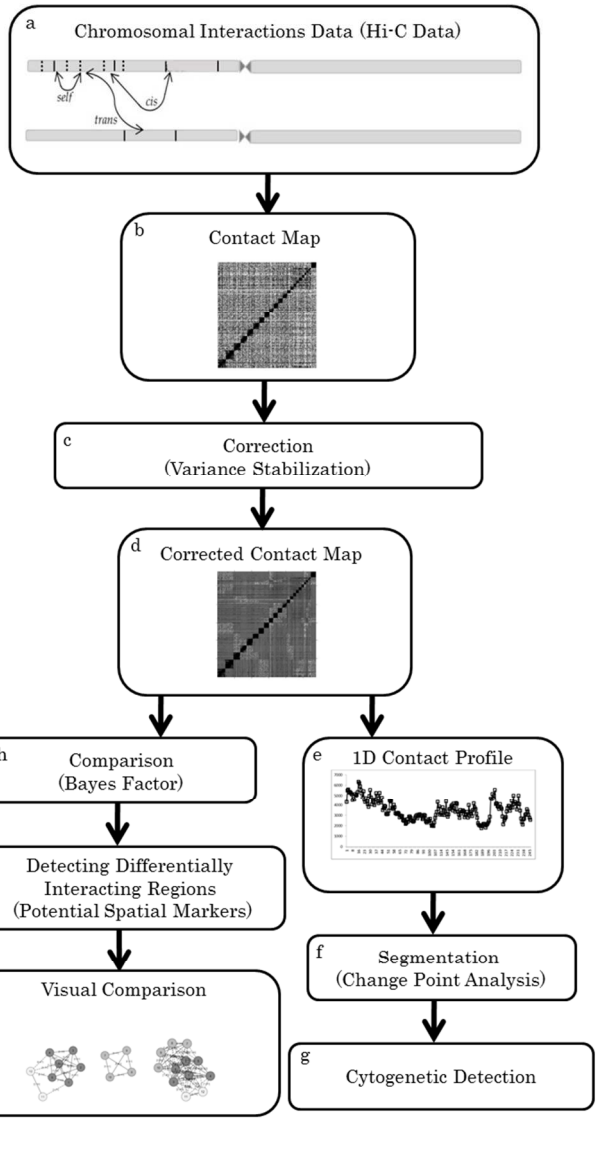


Fig. 1 Main steps of the analysis of chromosomal interactions data. We offer methods and software for analysing chromosomal interactions data. First, contact maps are generated from Hi-C data (a-b). Correction of noise and bias is then carried out with variance stabilization (c). Next, 1D contact profiles are generated from corrected contact maps (d-e) and segmented with change point analysis (f). The resulting segmentation could be used for detecting chromosomal aberrations (g, cytogenetic detection) as shown in this paper. Corrected contact maps are further compared, using Bayes Factor analysis, in order to detect regions that are differentially interacting (h-i). The output of this comparison could provide insights on functional spatial markers and further visualized and integrated with tools developed for this purpose[11-12] (j).

Methods

In this section we formalize the problems of correcting bias in Hi-C contact maps, comparing them and segmenting 1D contact profiles. We first provide a brief description of variance stabilization and change point analysis and explain how they can be applied to solve bias correction and segmentation, respectively. Specifically, we propose here wavelet statistics to address these tasks, as they (and in particular wavelet-based change point detection and variance stabilization) were previously successfully applied to various real-life and biological problems such as genome sequence analysis, protein structure and microarrays data analysis[13] (some notable examples are[14-17]). We next describe how the Bayes Factor[18] can be used for comparing Hi-C contact maps, so that regions that interact significantly different can be detected. Finally, we provide details about the software implementation of our methods, made available as an R package (chromoR).

Variance stabilization of Hi-C coverage with Haar-Fitz Transform

Hi-C coverage varies depending on different factors[2] that may act at different scales. As a result, the interaction frequency between 2 regions may be an over or under representation of the true interaction frequency. We would like to remove noise that is a result of variance in coverage while taking into consideration the multiple scales at which variance may appear.

The Haar-Fitz Transform (HFT)[19-20] decomposes Poisson distributed observations into coefficients in multiple scales. These coefficients are Gaussian distributed variables[19-20], so that the variance no longer depends on λ, the parameter of the distribution[20]. Moreover, wavelet shrinkage methods for Gaussian noise can now be applied for de-noising[21], by filtering coefficients (at different scales) that are most likely to consist only of noise. As a result, variance is stabilized and noise is reduced in the reconstructed sequence of observations. This method, previously shown to perform well for de-noising and intensity estimation of Poisson data[20] could be applied to correct Hi-C contact maps, as explained below.

A Hi-C contact map is a matrix M of size $N*N$, where N is the number of loci in the genome under consideration and we assume that $M[i,j]$, the interaction frequency between a locus i and a locus j, is a Poisson distributed variable. M could be converted into a set R of $m(1+m)/2 + 1$ vectors, where m is the number of chromosomes and we have $m(1+m)/2$ vectors, one for each pair of chromosomes, and another vector for the entire set of self-interactions (the diagonal of M). R represents pairwise interactions within (*cis*) and between chromosomes (*trans*), separately for each chromosome pair, and separately from all self-interactions (considered here together).

We then apply HFT for R, separately for *cis*, *trans* and self-vectors, followed by a de-noising procedure and then reconstruct the sequence with the inverse HFT. Here, de-noising is achieved with cross-validation wavelet shrinkage[20] based on exploratory analysis with Poisson data sets and with Hi-C data. After de-noising and reconstruction, M is reconstructed from R, resulting in a new contact map M', where $M'[i,j]$ is the corrected interaction frequency between the loci i and j. We note here that our correction could be applied without separating *cis* and *trans* interactions, or could be applied independently and separately for each vector in R.

Segmentation of 1D contact profiles with change point detection

We define the 1D contact profile, x, of chromosome a, with respect to chromosome b (a and b may be the same chromosome), as an n-tuple which corresponds to the row sums of the contact map of chromosome a and chromosome b, where n is the number of loci in chromosome a. The i-th element in x is thus the total number of interactions observed between the i-th locus in chromosome a with all loci in chromosome b.

Given a contact profile x, we are interested in finding $1 < t_1 < t_2 < .. t_m < n$, where m is the number of change points and $t_1, t_2, ..., t_m$ are their respective positions (loci). We assume here that the elements in x are realizations of random variables that are Poisson distributed, where λ_i is the expected value of the i-th element. Detecting the position(s) of change, $1 < t_1 < t_2 < .. t_m < n$, is then to test the following null hypothesis:

$$H_0: \lambda_1 = \lambda_2 = ... = \lambda_n \quad (1)$$

versus the alternative:

$$H_1: \lambda_1 = ... = \lambda_{k1} \neq \lambda_{k1+1} = ... = \lambda_{k2} \neq \lambda_{k2+1} \quad ... = \lambda_{km} \neq \lambda_{km+1} = ... = \lambda_n \quad (2)$$

In order to solve this problem we follow a wavelet Poisson change point detection algorithm[15], originally developed to detect multi-scale changes in Poisson data and applied to photon intensities for single molecule detection. While various parametric and non-parametric approaches were suggested for solving the change point problem, this approach has the advantage of addressing the multi-scaled and fluctuated nature of the data, while assuming a Poisson distribution. In addition, it allows to control the resolution of the search by specifying the coarsest level at which the search for change points will be performed (for segmentation purposes we have set the coarsest level to be 1 in order to perform a full multi-scaled search).

Given a 1D contact profile, this algorithm[15] applies a discretized version of the Continuous HFT and decomposes the profile into multi-scaled coefficients. The coefficients at each scale correspond to the difference between aggregated neighbours at the previous scale, where the top scale consists of the sequence of observations. The algorithm then searches for local maxima coefficients within each scale and continues by linking them across scales. Using the resulting maxima lines, the algorithm selects candidate change points according to the location and scale that gives the maximum value (along each maxima line). The final set of change points is then the subset of candidate change points that maximizes the fit with the data, evaluated with the Akaike's Information Criterion (a detailed description and analysis of this algorithm with Poisson data appears in the original paper[15]).

The detected change point locations define a segmentation of the input contact profile that can be further analysed and compared. Here (Results section), we explain how such analysis could be applied for detecting chromosomal aberrations and also provide an example for its ability to detect topological boundaries (Supp. Text T6). We note here that the above change point segmentation could also be applied to the distribution of restriction sites along each chromosome for evaluating Hi-C coverage (where the number of change points and their spread are an indicator for coverage heterogeneity and variance). Supp. Text T2 provides the results of this additional application for the interested reader.

Comparison of contact maps with the Bayes Factor

Given 2 contact maps M_1 and M_2 of the same dimensions $N*N$, denote $x^r_{i,j}$ to be $M_r[i,j]$, the interaction frequency between loci i and j in the contact map M_r, $r = 1,2$. We are interested in finding pairs of loci i,j that present a significantly different interaction frequency between M_1 and M_2. Taking a Bayesian approach, this could be achieved by comparing the probability of observing $x^1_{i,j}$ and $x^2_{i,j}$ given the null hypothesis (that a spatial disposition has not occurred), versus the alternative. The ratio between the probability of the data given the alternative versus their probability given the null is defined by the Bayes Factor[18] (BF). The BF of a loci pair i and j, denoted as BFi,j, is then:

$$BF_{ij} = \frac{P(x^1_{i,j}, x^2_{i,j} | H_1)}{P(x^2_{i,j}, x^2_{i,j} | H_0)} = \frac{P(x^1_{i,j})\, P(x^2_{i,j})}{P(x^1_{i,j}, x^2_{i,j})} \quad (3)$$

where $P(x^1_{i,j}, ... x^r_{i,j})$ is the marginal likelihood of the data. This ratio was previously shown to provide a better measure for statistical significance, compared to the *p. value* computed with frequentist statistics[22].

We assume here that corrected and standardized $x^r_{i,j}$ come from a Gaussian distribution with a Gamma prior, where both mean (μ) and variance (σ2) vary. This assumption is based on the mean-variance stabilization (achieved with our correction method) and on the standardization of the data (removing the mean and dividing by the standard deviation). Standardization is performed here in order to support comparison even in the presence of overall large differences in read count between contact maps. Under this assumption the marginal likelihood is then given by:

$$P(x^1_{i,j}, ... x^n_{i,j}) = \int \prod_{r=1}^{n} [\, p(x^r_{i,j} | \mu, \sigma^2) \,]\, p(\mu | \mu_0, \sigma_0^2)\, d\mu \quad (4)$$

Eq. 4 can be re-written with a close form[23-24] so that the BF value for each pair of loci can be computed analytically. (Supp. Text T3 provides the full details of this derivation). Here, we have calculated separated prior mean (μ_0) and variance (σ_0^2) for *cis, trans* and *self-* interactions (Supp. Text T3), so that the prior used for a given pair of loci depends on the type of interaction between them.

Given corrected and standardized contact maps M_1, M_2, the BF values for all loci pairs can be calculated (note that BFi,j equals to BFj,i as the interaction between loci i and j is symmetric). Pairs assigned with a BF value larger than a threshold t are reported as differentially interacting regions in M_1 with respect M_2 and can be further investigated. In the Results section we explain how to choose t (also relating it to the recommended thresholds of BF[18]) and provide an evaluation of the False Discovery Rate (FDR) of our method.

Software Implementation

We have developed an R package (chromoR) that implements all the stages of chromosomal data analysis, as described in this paper, from pre-processing (generating contact maps from mapped positions) to correction, segmentation and comparison.

chromoR is available on CRAN (http://cran.r-project.org/) and includes examples and data from this paper (see also http://www.cl.cam.ac.uk/~ys388/chromoR/) as well as additional guidelines for generating (or retrieving) Hi-C mapped positions from raw Hi-C data (fragment pairs). Although chromoR is accompanied with a complete documentation we provide here a brief description of the implementation of the methods we have developed (correction, segmentation and comparison).

Our correction procedure uses the haarfisz[25] and wavethresh[26] R packages for performing HFT and cross-validation de-noising. The user can apply the *correctCIM* function for correcting while separating *cis*, *trans*, and *self*-interactions (as explained above). This function requires a contact map and a genome segmentation defining the coordinates of each locus. In order to correct a single pairwise contact map or to apply the correction on the entire contact map, the function *correctPairCIM* can be used instead. Both functions will return the corrected version of the input contact map as output.

The function *compareCIM* implements our BF comparison procedure. Here, the user is required to provide 2 contact maps of the same dimension and a genome segmentation file defining the coordinates of each locus. The output is the coordinates of the loci pairs that were found to be differentially interacting (achieving a BF value larger than a given threshold) along with their BF values and their corresponding (standardized) interaction frequencies in the compared contact maps. Additional optional input is the BF threshold (set to be the permissive threshold by default, see Results section) and any property of interest provided by the user for each locus as part of the genome segmentation file (e.g. cytogenetic band).

For performing change point detection, we have downloaded the code of the described algorithm[18] from http://homepages.ulb.ac.be/~majansen/software/index.html and written additional MATLAB code to perform the segmentation analysis described here. For the sake of completeness, we have further implemented this algorithm in R and made it available (the *segmentCIM* function) as part of the chromoR package.

Results

Correcting Hi-C contact maps with variance stabilization

The objective of our correction method is to control over- and under- representation of genomic regions, while preserving true trends of interaction. Ideally, this correction will be evaluated by comparing results to a benchmark or with a mechanistic model. In the absence of such data, we use here statistical estimators. Specifically, we assume that a good correction will improve the correlation between replicates generated with different enzymes (reproducibility). The reproducibility measure was previously used by Lieberman *et al.*[1], and Hu *et al.*[3], for evaluating Hi-C data and their correction.

For this evaluation we have used publicly available 1 megabase (Mb) HindIII and NcoI contact maps of GM06990[1] (a lymphoblastoids cell line), generated after fragment filtration[2] (Supp. Text T4 provides the details of data preparation). We have then calculated the Spearman correlation coefficients for all pairwise contact maps of the 2 replicates (in *trans* and *cis)*, before and after applying our correction (variance stabilization, VS). Here, we have applied our correction with (VS1) and without separating *cis* and *trans* interactions (VS2). In addition, we have considered 3 other *state-of-the-art* correction methods: HiCNorm[3], iterative correction and eigenvector decomposition (ICE)[4] and a method developed by Yaffe and Tanay[2], referred from here on as YT. For HiCNorm and YT we have downloaded the corresponding corrected contact maps of the HindIII and NcoI replicates (YT: http://compgenomics.weizmann.ac.il/tanay/?page_id=283; HiCNorm: http://www.people.fas.harvard.edu/~junliu/HiCNorm/Lieberman-Aiden.rar). Since corrected contact maps were not available for ICE and as its current software implementation required complicated pre-processing, we have implemented ICE and applied it to the same contact maps used as input for our correction.

The results of this assessment (Table 1) showed that our method significantly improves reproducibility, in *cis* (one sided paired t-test, $p < 0.0005$) and in *trans* (one sided paired t-test, $p < 2.2e-16$), with VS2 (VS1) increasing the average of (absolute) Spearman correlation from 0.750 to 0.819 (0.799) and from 0.062 to 0.127 (0.115), in *cis* and *trans* correspondingly. Comparison with other methods (Table 1) further showed that our method provides a better means of correction. Specifically, for *trans* contact maps, our method (VS1 and VS2) achieved significantly better reproducibility than all 3 other methods (one sided paired t-test, $p < 1.6e-07$). For *cis* contact maps, VS2 outperformed all 3 other methods while VS1 provided better reproducibility than ICE and YT and a lower, although not significantly (one sided paired t-test $p > 0.9$), correlation than HiCNorm. We note that in addition to providing better performance, both VS1 and VS2 also presented higher consistency, while other methods varied in their performance. For example, while VS2 achieved the best results in *cis* and *trans,* HiCNorm was the second best for *cis* correction, but only fourth for *trans*.

Table 1 Average Spearman correlation, calculated across pairwise *cis* and *trans* contact maps, between Hi-C contact maps of GM06990 NcoI and HindIII replicates. In addition to the observed contact maps (OBS), the following correction methods were evaluated: variance stabilization (VS1, and VS2, 2 variants of our correction), YT[2], HiCNorm[3] and ICE[4]. The highest correlation value achieved for each interaction type is highlighted in bold.

Type[a]	OBS	VS1	VS2	YT	HiCNorm	ICE
cis	0.750	0.799	**0.819**	0.779	0.817	0.795
trans	0.062	0.117	**0.127**	0.086	0.066	0.064

[a] Spearman correlation was calculated for pairwise contact maps in *cis* and in *trans*. The table provides the mean (of absolute) correlation coefficients across chromosome pairs.

In order to further evaluate our method, we have next tested whether known features emerge after correction. Here, we have compared distance maps generated from observed (generated as described in Supp. Text T1) and corrected (as described above for VS1) contact maps of GM06990, when summing the HindIII and NcoI contact maps (Fig 2). As pairwise interaction frequencies were found to be proportionally inversed to spatial proximity[27], we have used here the inverse function for transforming an interaction frequency to a distance proxy. We note that this distance function could be refined and improved using findings

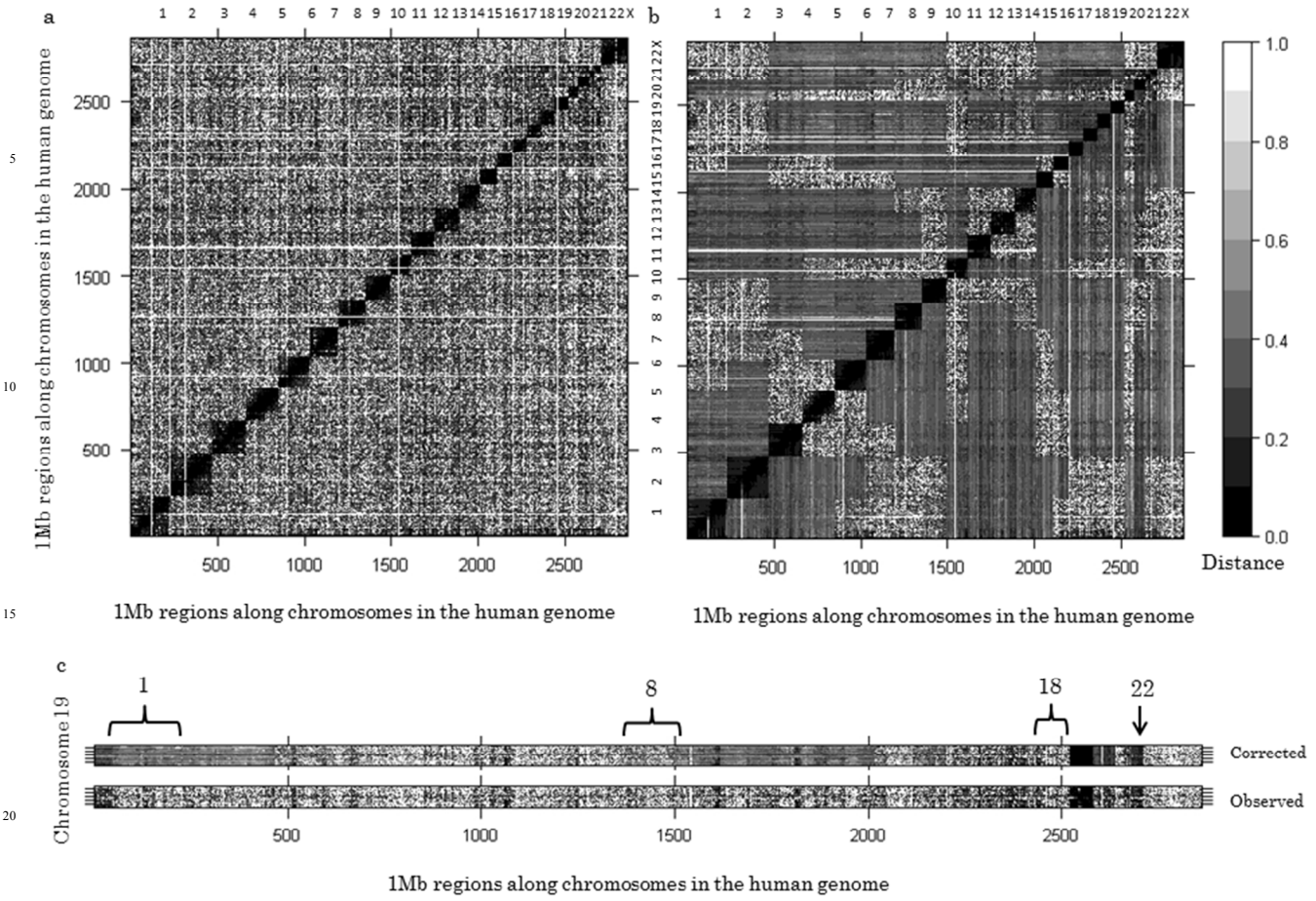


Fig. 2 Heatmaps of observed and corrected distance maps. Contact maps of GM06990 (summation of the NcoI and HindIII contact maps) were transformed to distance maps by taking their inverse. The corrected distance map (b) provides a better distinction between closer and distant regions, compared to the observed distance map (a). In the corrected map of chromosome 19 (c), chromosomes 1 and 22 appear closer (darkened regions) and chromosome 8 and 18 appear more distant (light regions). This distinction (that is missing from the observed distance heatmap) is consistent with previous findings on the nuclear organization of lymphoblastoids[28] where chromosomes 1,19,22 were shown to reside in the nuclear interior while chromosomes 8 and 18 were located at the nuclear periphery.

coming from different experiments, for example, distances measured with fluorescent in situ hybridization (FISH).

Our corrected distance map (Fig. 2b) provided a better distinction between regions that are less or more distant. Figure 2c presents the corrected and observed distance maps for chromosome 19 with all other chromosomes. In the corrected map chromosomes 1 and 22 appear closer (darkened regions) and chromosome 8 and 18 appear more distant (lighter regions). This distinction (that is missing from the observed distance map) is consistent with previous findings on the nuclear organization of lymphoblastoids[28] where chromosomes 1,19 and 22 were shown to reside in the nuclear interior while chromosomes 8 and 18 were located at the nuclear periphery. We note that even with the simplified distance function used here, results from exploratory analysis we have performed showed that our corrected distances are in agreement with distances obtained from previous FISH experiments, and provide better estimation when compared to other correction methods (Supp. Text T5 provides comparison of FISH distances[1] and distances calculated from corrected contact maps, using our correction, and compared to YT, HiCNorm and ICE). These results suggest that our correction method provides an improved discrimination between noise and true trends and that corrected distance maps could be further used to provide a better input for spatial models of chromosomal organization.

Detecting chromosomal aberrations with change point analysis of 1D contact profiles

A 1D contact profile provides for each region, a measure for its overall contact frequency, or 'interactivity', with respect to another chromosome (the same chromosome in the *cis* case). When a chromosomal aberration occurs in *cis* (i.e. when a deletion or a duplication takes place), we expect that the interactivity of the affected region/s will change, and significantly increase (duplication) or decrease (deletion), depending on the aberration type and size. For example, if a deletion of size a, occurs in a region with a *cis* interactivity value p (the value of that region in the *cis* contact profile), then its expected interactivity value, after deletion, is $p - x$, where x is proportional to a. Similarly, when a translocation takes place (chromosomal aberration in *trans)*, the corresponding *trans* contact profiles are expected to present a *cis* like interactivity for the translocated regions. Here, the change from *trans* to *cis* interactivity reflects the translocation event, where regions that were previously located in *trans (*separate chromosomes) are now in *cis* (same

chromosome). Change point analysis could thus be applied to 1D Hi-C contact profiles, where the resulting segmentation is compared across different healthy and disease conditions in order to detect chromosomal aberrations and related organization patterns.

We note that our change point approach could be applied for other detection purposes. For example, Dixon *et al.*[29] have recently shown that our genome could be divided into topological domains, so that each domain presents a significantly higher self-interaction with respect to interactions with other domains. In this seminal work, the quantity under investigation was the tendency of each region to form upstream or downstream interactions, in a 2Mb range (termed 'directionality index'). While Dixon *et al.* employed a Hidden Markov Model (HMM) in order to identify domain boundaries, based on the 'directionality index', we suggest here that change point analysis could potentially achieve a similar segmentation, when applied to 1D contact profiles of high resolution. Supp. Text T6 provides such an example for a 10 Kilobases (Kb) contact profile, showing that our method can recover topological domains, consistent with the previous findings of Dixon *et al*[29] and with key genomic features.

Here, instead, our goal is to identify changes in interactivity and then compare them across conditions, with the rationale of identifying aberrations and distortions in chromosomal organization (as described above). For this purpose we have used Hi-C data with simulated and real chromosomal aberrations. We have first generated 1Mb 1D contact profiles from publicly available IMR90[29] (a lung fibroblasts cell line) and K562[1] (a myelogenous leukemia cell line) Hi-C data, as described above (Methods section), after applying our correction (Supp. Text T7 provides the details of data preparation). For change point analysis of simulated deletions and duplications we have used the IMR90 *cis* 1D contact profile of chromosome 3. In order to simulate a deletion (duplication) in a region *i*, we have removed (added) contacts proportional to the size of the aberration size, *a*, and the average *cis* contacts per base pair, *c*:

$$Pi,a = Pi + j \cdot c \cdot a \qquad (5)$$

Where *j* is -1 for deletions and 1 otherwise, and *Pi,a* is the value of the contact profile at region *i* after an aberration of size *a* has occurred. For sizes of 1Mb or more (2Mb or 3Mb), the contact value of corresponding regions was set to 1 for deletions and for twice the average *cis* contact for duplications. For simulating translocations, we have considered the IMR90 *trans* 1D contact profile of chromosome 9 with respect to chromosome 12. Here, we have simulated a translocation by adding contacts proportional to the size of the translocation and to the average *cis* contact per base pair (see Eq. 5). For sizes of 1Mb or more we have replaced the *trans* contact with the average *cis* contact in chromosome 9. At each iteration of our simulation, a single region was chosen and change points were detected given a simulated aberration (for aberration sizes ranging between 0.01 – 3 Mb). For each of the considered aberration sizes we have then calculated the proportion of segmentations that detected the modified region as a change point within a range of 5 Mb.

Results (Figure 3) showed that for large aberration sizes (2-3Mb) we get a high detection proportion (> 0.9) for duplications, deletions and translocations. The detection rate decreased with aberration size with a better performance for deletions and translocations (compared to duplications).

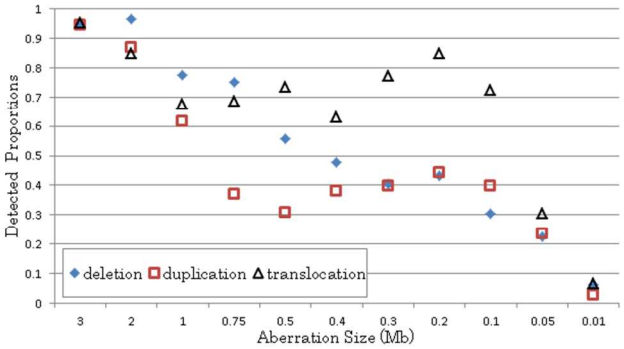


Fig. 3 Proportions of detected chromosomal aberrations (deletions, duplication sand translocations) within a 5Mb range, with change point analysis. Corrected 1D profiles of IMR90 were used for simulation. Deletions and duplications were simulated using the *cis* 1D profile of chromosome 3. The *trans* 1D profile of chromosomes 9, with respect to chromosome 12, was used for simulating translocations. At each iteration (in a given simulation), a single region was chosen and change points were detected given a simulated aberration (for each of the given sizes). The proportion of simulated aberrations that were detected as change points (within a 5Mb range), for different aberrations sizes were reported (Y axis).

Aberrations of size 0.01 Mb (10 Kb) were practically undetectable, suggesting that the segmentation size (1 Mb) acts as a limiting factor. Detection of translocations was better for mid-sizes (0.1-1Mb) when compared to other aberrations, with a high detection proportion for sizes of 0.1-0.3 Mb.

The average detection offset was 1.03 Mb for large and mid-sized aberrations (≥0.5Mb) and increased as aberration size decreased (Supp. Fig. S1). We note that the introduction of aberrations also had an impact on the overall change point segmentation, where up to 2 change points, on average, were added, with a larger impact for deletions and duplications (Supp. Table ST1). These additional change points are a result of introducing an aberration, which affects all scales. These points can be filtered out based on their relative contribution, but may also be used as an indication for a possible distortion.

To further evaluate change point analysis for detecting chromosomal aberrations we have next compared change points detected in 1D contact profiles of IMR90 and K562. Specifically, we have compared change point segmentation of 1D profiles of chromosome 9, with respect to chromosomes 12 and 22, as these chromosome pairs are known to translocate in chronic myeloid leukemia[30]. We found that for the K562 contact profiles, additional change points were detected that correspond to the known translocation positions in chromosome 9 (Figure 4): region 21-22Mb, (cytogenetic band p21.3, a known translocation with chromosome 12[30]), and region 133-134Mb (cytogenetic band q34.11- q34.12, a known translocation with chromosome 22 and chromosome 13[30]). We note that these change points were identified in both of the K562 contact profiles, suggesting that each *trans* contact profile provides a partial view of the spatial segmentation, while major changes emerge consistently. Putting together the results described above suggest the usefulness of employing change point analysis to Hi-C contact profiles for cytogenetic purposes.

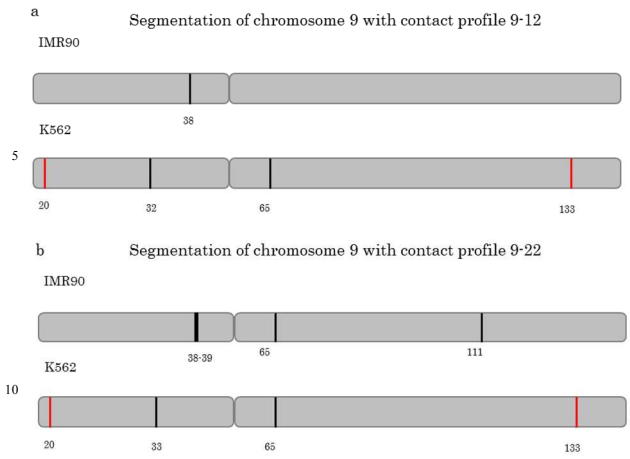


Fig. 4 Comparison of segmentation of contact profiles. 2 contact profiles of chromosome 9 were segmented with change point analysis: the *trans* contact profile of chromosome 9 with chromosome 12 (a) and the *trans* contact profile of chromosome 9 with chromosome 22 (b). The change points detected with the contact profile of IMR90 and K562 are marked with a line and with the associated 1Mb position on chromosome 9. Known chromosomal aberrations were detected as change points for the K562 profile (highlighted in red) and were missing from the IMR90 segmentation.

Detecting differentially interacting regions with Bayes Factor (Comparing Hi-C contact maps)

Due to the probabilistic nature of Hi-C data, we expect variations to occur even between replicates and after appropriate correction. We attribute these variations, or changes, to noise and distinguish them from significant changes (differentially interacting regions) that occur, for example, in disease conditions or between different types of cells. Here, we first set out to evaluate the ability of our method to discriminate between noise and significant changes (and in particular, its FDR) and to determine the BF threshold that is associated with such discrimination.

For this evaluation we have first generated a corrected 1Mb map for a Hi-C replicate of IMR90[29], denoted here as M^{orig}. In order to simulate a case where all changes are attributed to noise, we have then generated replicates of M^{orig}, while gradually increasing the level of noise. This was done by adding random gaussian noise to M^{orig}, where the mean is fixed (and set to zero) and the variance takes the value of the ratio σ^2/SNR, where σ^2 is the variance of M^{orig} and the SNR is decreased in order to increase the level of noise.

Significant changes were further introduced by turning *cis*-like values into *trans*-like values and vice versa, mimiking a structural change. Specifically we have assigned randomly selected *cis* pairs in a given noisy replicate with the interaction frequency of a *trans* pair from M^{orig} (and similarly for *trans* pairs in this replicate). Here, pairs were selected for each chromosome, where we first select a random region *i1* in the given chromosome and then randomly select 2 other regions: *i2* and *i3*, from regions that are located in *cis* and *trans* with respect to region *i1*, correspondingly. The 3 selected regions then define the pairs to be switched: <*i1,i2*> and <*i1, i3*>.

We have next calucalted the BF values for the comparison of M^{orig} with noisy replicates and noisy replicates with significant changes, generated with a range of SNR values ([2,2.5,3..,9.5,10]). Figure 5 presents the maximal BF value obtained for each comparison, for the noisy (blue) and significant (red) cases, indicating that significant changes are associated with significantly higher BF values (one sided t-test, $p < 1.2e-10$), independently of noise. We note here that the maximal BF value (8.962) of all comparisons with noisy replicates, was in close agreement with a general BF threshold previsuly suggested to provide strong evidence against the null[18], providing a restrictive threshold for comparison.

In order to choose an appropriate threshold while controlling the FDR we have further compared the values of upper percentiles of BF values across different comparisons (range of SNR values). As expected (since only a small number of significant changes was introduced), the BF percentiles values computed for comparison with noisy replicates and with noisy replicates with significant changes, were not significantly different (one sided t-test, $p > 0.05$) up to the 99.9995th percentile (exclusive), after which significant differences have emerged (one sided t-test $p < 0.05$). Following this observation we have further defined a permissive threshold, as the average value of the 99.9995th percentile (6.10), across comparisons with noisy replicates We note that this threshold was stable across different levels of SNR (6.10±0.025), thus providing a low FDR even in the presence of a noisy setting.

We have then assessed the sensitivity and specificity of our method for comparisons of M^{orig} with noisy replicates with significant changes (setting the SNR to 5). When using the restrictive threshold, our method achieved a high specificity (99.99%) and a low sensitivity (10.87%). The sensitivity was further improved (19.56%) while maintaining the high specificity (99.99%), with the permissive threshold. We note that using a lower threshold (5.83, the 99.99th percentile), did not improve sensitivity, suggesting that the highest values indeed correspond to the most significant changes.

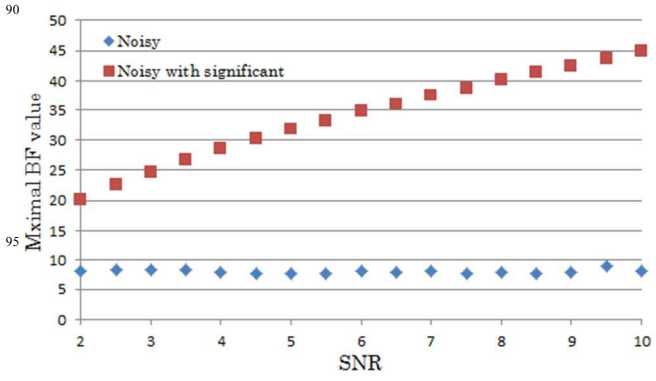


Fig. 5 Maximal BF values computed for comparisons of Morig with noisy replicates (blue) and with noisy replicates with significant changes (red), across different levels of SNR (lower SNR corresponds to a higher level of noise). The maximal BF value for comparison with noisy replicates was relatively constant (8.148+/-0.285), suggesting that our method is robust to noise. When, in addition to noise, significant changes were introduced (red), the maximal value was significantly higher (one sided t-test, $p < 1.2e-10$).

After establishing the robustness of our method with simulated data and determining an appropriate BF threshold (permissive

threshold, 6.10), we have next compared corrected 1Mb contact maps, generated for replicates of IMR90[29] and embryonic Human Stem Cells (ESC)[29] (Supp. Text T8 provides the details of data preparation). When comparing replicates from the same cell type, no changes were detected for the IMR90 replicates, while 0.0002% (10 changes) were detected as significant for the ESC replicates (here, comparison with the restrictive threshold resulted with 1 change detected as significant).

Conversely, when comparing replicates from different cell types, more changes were detected overall (Table 2). We note however that some comparisons yielded more changes than others. The differences in comparison results could be attributed to sequencing depth, fluctuations and technical factors. In order to evaluate the comparability of 2 contact maps, one can first use visualization aids (e.g. heatmaps). Furthermore, our correction method could be used to evaluate the consistency between replicates, using different significant threshold. In the comparison presented here, despite the different in their number, changes were consistent across comparisons. For example, out of the 4 changes detected when comparing replicate ESC-2 and replicate IMR90-1, 2 changes were detected in the comparison of replicates ESC-1 and IMR90-1 within an offset of several 1Mb in one of the dimensions (1 change is shared and another is shared within a close range). Table ST2 lists all the changes that were detected as significant in the cross comparisons of IMR90 and ESC replicates, showing that the most significant changes (highest BF values) were found in 2 or more cross comparisons, in most of the cases. Interestingly, the 2 region pairs that achieved the highest BF value (highlighted in Supplementary Table ST2) were mapped to the gene FEN1P1 (chromosome 1: 91,793,795 - 91,795,813), associated with lung cancer[31] and the gene MYO1D (chromosome 17: 30,819,540 - 31,204,195), associated with intracellular movements and mostly expressed in brain, followed by lungs[32].

Table 2 The number of changes detected as significant in cross comparisons of IMR90 and ESC replicates (using the permissive threshold).

Replicate[a]	ESC-1	ESC-2
IMR90-1	12	4
IMR90-2	68	32

[a] 2 replicates were considered for each cell type: IMR90-1,-2 and ESC-1, -2

We note here that a further comparison of the IMR90 contact map (summing the contact maps of the 2 replicates) with a contact map generated from K562 Hi-C data[1] identified the regions pair chromosome 9: 132-33Mb and chromosome 22: 21-22Mb as the most significant change. This pair is mapped to a known translocation in K562[30] and was successfully recovered despite the differences in sequencing depth of the corresponding Hi-C data sets[1,29].

The results of our assessment suggest that our comparison method could be applied to Hi-C contact maps in order to perform a 1Mb genome wide scan for interesting spatial changes, even in the presence of noise and fluctuations.

Discussion

In this paper we provide a step-by-step statistical bioinformatics solution to the processing of Hi-C data from fragment abundance to chromosome signal differences between control and disease and across replicates. This solution includes 3 main steps: bias correction of contact maps with variance stabilization (Fig. 1c), change point segmentation of 1D contact profiles (Fig. 1f) and comparison of contact maps using the Bayes Factor (Fig. 1h).

We have chosen a wavelet approach for variance stabilization as it addresses the multiple scales at which bias may appear (see Supp. Text T1). We have shown that this approach provides better reproducibility (also when compared to existing methods) and can better capture chromosomal (Figure 2c) and intra-chromosomal distances (Supp. Text T5). By simulating chromosomal aberrations we have next demonstrated how wavelet change point analysis can be successfully applied to 1D contact profiles for detecting structural changes. We have further shown that this method can detect known translocations in a cancer cell line. We have also provided an example showing that our approach is useful for detecting topological domains (Supp. Text T6). Finally, we have shown that the Bayes Factor can be adapted to act as a useful estimator of significant changes between contact maps, as part of a comparison method we have developed. Using *in-silico* and real Hi-C data, we have shown that we can detect significant changes in interaction frequency, in a robust manner, successfully addressing both replicates and different cell comparison.

The methods described here and the accompanying R package (chromoR), which implements them, result in a complete pipeline for analysing chromosomal interactions data (Fig. 1). This pipeline offers a novel nuclear telescope, allowing researchers to pinpoint spatial changes, at a resolution of 1Mb (and up to ~10Kb). A future continuation of this pipeline would be to develop methods that can investigate regions of interest at a higher resolution and at the level of fragments and that performs integration with epigenetic data. Based on the output of this pipeline, researcher could identify important spatial markers and develop new models for the nuclear architecture.

Conclusions

Data on chromosomal interactions are now becoming available in unprecedented resolution and their quality is under ongoing improvement. There is now a growing need for reliable and user friendly computational tools that will support a better interpretation of these data. The methods and software, as described here, provide a comprehensive solution for addressing this task, with the hope to help researchers in gaining a better understanding of the nuclear architecture.

Acknowledgments

We thank Dr Viet Anh Nguyen for a productive discussion on the Bayes Factor in the context of Hi-C comparison. We further thank our 2 anonymous reviewers for their invaluable comments.

Funding: we thank FP7-Health-F5-2012, under grant agreement n° 305280 (MIMOmics).
Conflict of Interest: none declared.

Notes and references

[a] *Computer Laboratory, University of Cambridge, Cambridge, CB3 0FD, UK; E-mail: ys388@cam.ac.uk*
[b] *Computer Laboratory, University of Cambridge, Cambridge, CB3 0FD, UK; E-mail: pl219@cam.ac.uk*

† Electronic Supplementary Information (ESI) available: 10 supplementary pdf files: (1) Supp. Text T1: a complementary analysis of GC content and Hi-C coverage at different scales, (2) Supp. Text T2: the details of applying change point analysis for evaluating Hi-C coverage, (3) Supp. Text T3: a detailed derivation of the Bayes Factor, used for our comparison method, (4) Supp. Text T4: additional details for the evaluation of our correction method and its comparison to HiCNorm[3] and YT[2], (5) Supp. Text T5: exploratory comparison of FISH distances and distances calculated from corrected contact maps, using our correction and compared to HiCNorm[3] and YT[2], (6) Supp. Text T6: example of applying our change point approach for detecting topological domains, (7) Supp. Text T7: details of generating 1D contact profiles used for change point analysis of simulated and real chromosomal aberrations, (8) Supp. Text 8: details of generating contact maps (from publicly available Hi-C data) for testing our comparison procedur, (9) Supp. Table ST1: a table providing the average number of additional change points (impact) after introducing a chromosomal aberration, (10) Supp. Table ST2: a table providing the coordinates of region pairs that presented a significantly different interaction frequency between IMR90 and ESC, (11) Supp. Figure S1: a graph of the average offset for detecting simulated chromosomal aberrations. See DOI: 10.1039/b000000x/

1 E. Lieberman-Aiden, N. L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R Lajoie, P.J. Sabo, M.O. Dorschner, R. Sandstrom, B. Bernstein, M. A. Bender, M. Groudine, A.Gnirke, J. Stamatoyannopoulos, L. A. Mirny, E. S. Lander and J. Dekker, *Science*, 2009, **326**, 289.
2 E. Yaffe and A., *Nat. Genet*., 2011, **43**, 1059.
3 M. Hu, K. Deng, K. Selvaraj, Z. Qin, B. Ren, and S. J. Liu, *Bioinformatics,* 2012, **28**, 3.
4 M. Imakaev, G. Fudenberg, R. P. Mccord, N. Naumova, A. Goloborodko, B. R. Lajoie, J. Dekker and L. A. Mirny, *Nat Methods*, 2012, **9**, 999.
5 A. Cournac, H. Marie-Nellyn, M. Marbouty, R. Koszul and J. Mozziconacci, *BMC genomics*, 2012, **13**, 436.
6 C. Burge, A.M. Campbel and S. Karlin, *Proc Natl Acad Sci USA* ,1992, **89**, 1358.
7 D. Mouchiroud, G. D'Onofrio, B. Aïssani, G. Macaya, C. Gautier and G. Bernardi, *Gene*, 1991, **100**, 181.
8 G. Bernardi, *Gene* 2000, **241**, 3.
9 Y. Zhang, R. P. McCord, Y. J. Ho, B. R. Lajoie, D. J. Hildebrand, A. C. Simon, M. S. Becker, W. A. Frederick and J. Dekker, *Cell*, 2012, **148**, 908.
10 J. M. Engreitz, V. Agarwala and L. A. Mirny, *PloS one*, 2012, **7**, e44196.
11 Y. Shavit and P. Lio', *Bioinformatics*, 2013, **29**, 1206.
12 I. Merelli and P. Lio', *PLoS One*, 2013, **8**, e75146.
13 P. Liò, *Bioinformatics*, 2003, **19**, 2.
14 E. S. Motakis, G. P. Nason, P. Fryzlewicz and G. A. Rutter, *Bioinformatics*, 2006, **22**, 2547.
15 M. Jansen, *Chemometrics and intelligent laboratory systems*, 2007, **85**, 159.
16 P. Fryzlewicz, V. Delouille and G. P. Nason, *Appl. Statist.*, 2007, **56** ,99.
17 B. Audit, A. Baker, C. L. Chen, A. Rappailles, G. Guilbaud, H. Julienne, A. Goldar, Y. d'Aubenton-Carafa, O. Hyrien, C. Thermes and A. Arneodo, *Nat Protoc*, 2013, **8**, 98.
18 R. E. Kass and A. E. Raftery, *Journal of the American Statistical Association*, 1995, **90**, 791.
19 M. Fisz, *Colloquium Mathematicum*, 1955, **3**, 138.
20 P. Fryzlewicz and G. P. Nason, *J. Comp. Graph. Stat*, 2004, **13**, 621.
21 G. P. Nason, *J R Statist Soc B*, 1996, **58**, 463.
22 J. Wakefield, *Genetic epidemiology*, 2009, **33**, 79.
23 M. H. DeGroot, *New York: McGraw-Hill*, 1970.
24 K. P. Murphy, *Technical report. University of British Columbia*, 2007.
25 P. Fryzlewicz, 2010, http://CRAN.R-project.org/package=haarfisz.
26 G. Nason, 2013, http://CRAN.R-project.org/package=wavethresh.
27 J. Fraser, M. Rousseau, S. Shenker, M. A. Ferraiuolo, Y. Hayashizaki, M. Blanchette and J. Dostie, *Genome Biol.*, 2009, **10**, R37.
28 S. Boyle, S. Gilchrist, J. M. Bridger, N. L. Mahy, J. A. Ellis and W. A. Bickmore, *Hum. Mol Genet*, 2001, **10**, 211.
29 J. R. Dixon, S. Selvaraj, F. Yue, A. Kim, Y. Li, Y. Shen, M. Hu, J. S. Liu and B. Ren, *Nature*, 2012, **485**, 376.
30 N. Sabine, R. Dirk, S. Michael and D. Hans-Jochen, *Leukemia Research*, 2001, **25**, 313.
31 M. Sato, L. Girard, I. Sekine, N. Sunaga, R. D. Ramirez, C. Kamibayashi, J. D. Minna, *Oncogene*, 2003, **22**, 7243.
32 T. Nagase, K. Ishikawa, M. Suyama, R. Kikuno, N. Miyajima, A. Tanaka, H. Kotani, N. Nomura, O. Ohara, *DNA Res*, 1998, **5**, 277.