Bioinformatics, 32(8), 2016, 1121–1129
doi: 10.1093/bioinformatics/btv736
Advance Access Publication Date: 17 December 2015
Original Paper

Genome analysis

Hierarchical block matrices as efficient representations of chromosome topologies and their application for 3C data integration

Yoli Shavit[1],*, Barnabas James Walker[2,3] and Pietro Lio'[1],*

[1]Computer Laboratory, University of Cambridge, Cambridge CB3 0FD, UK, [2]University of Cambridge, Cambridge CB3 0FD, UK and [3]Department of Life Sciences, Imperial College London, London SW7 2AZ, UK

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on 6 February 2015; revised on 21 November 2015; accepted on 12 December 2015

Abstract

Motivation: Recent advancements in molecular methods have made it possible to capture physical contacts between multiple chromatin fragments. The resulting association matrices provide a noisy estimate for average spatial proximity that can be used to gain insights into the genome organization inside the nucleus. However, extracting topological information from these data is challenging and their integration across resolutions is still poorly addressed. Recent findings suggest that a hierarchical approach could be advantageous for addressing these challenges.

Results: We present an algorithmic framework, which is based on hierarchical block matrices (HBMs), for topological analysis and integration of chromosome conformation capture (3C) data. We first describe chromoHBM, an algorithm that compresses high-throughput 3C (HiT-3C) data into topological features that are efficiently summarized with an HBM representation. We suggest that instead of directly combining HiT-3C datasets across resolutions, which is a difficult task, we can integrate their HBM representations, and describe chromoHBM-3C, an algorithm which merges HBMs. Since three-dimensional (3D) reconstruction can also benefit from topological information, we further present chromoHBM-3D, an algorithm which exploits the HBM representation in order to gradually introduce topological constraints to the reconstruction process. We evaluate our approach in light of previous image microscopy findings and epigenetic data, and show that it can relate multiple spatial scales and provide a more complete view of the 3D genome architecture.

Availability and implementation: The presented algorithms are available from: https://github.com/yolish/hbm.

Contact: ys388@cam.ac.uk or pl219@cam.ac.uk

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

HiT-3C techniques provide a means for studying the genome architecture at a range of resolutions. As oppose to light microscopy, 3C gives a population-based measure that relies on spatial proximity but does not directly convey a spatial context (Belmont, 2014). All 3C derivatives consist of cross-linking nuclei, followed by chromatin digestion and re-ligation. Quantifying the number of ligation events between two chromatin fragments gives their 'contact frequency' in the examined population of nuclei, providing a pairwise estimator for their distance. HiT-3C methods (most notably, the Hi-C method) detect contacts between groups of loci, per chromosome or genome-wide, which are typically summarized in a non-negative matrix called a 'contact map'.

Analyzing contact maps can reveal different scales of topological organization. At a 1 Megabase (Mb) resolution, chromatin compartments and sub-compartments were identified with principal component analysis (PCA) (Lieberman-Aiden *et al.*, 2009) and clustering (Rao *et al.*, 2014). (Dixon *et al.* 2012) further suggested that at shorter length scales (40–100 kilobases (kb)), a genomic bin can be assigned with a state which represents its preference to interact with other bins along the sequence (directionality index): upstream, downstream or none. Using hidden Markov models (HMM) the researchers have shown that the genome can be segmented into regions of bins with the same state, leading to the definition of Topologically Associated Domains (TADs, 'upstream' or 'downstream' regions) and their boundaries ('none' regions). Dynamic programming (Levy-Leduc *et al.*, 2014) and change point detection (Shavit and Lio', 2014) were also employed for identifying transitions in contact frequency enrichment which induce a topological genome segmentation. Owing to the definition of TADs, their position and boundaries can change considerably depending on the sequencing depth and the bin size used (Filippova *et al.*, 2014). For example, studying HiT-3C contact maps at different resolutions revealed that previously identified large TADs can be divided into sub-TADs (Filippova *et al.*, 2014; Phillips-Cremins *et al.*, 2013). This hierarchical organization suggests the need for methods that could go beyond the TAD-based segmentation and capture topologies at different length scales. Hierarchical topologies derived from contact maps can further provide a spatial context for studying point-wise and relational properties (e.g., histone modifications and gene co-expression, respectively). This can in turn uncover associations between spatial organization (structure) and genomic features (function) that could not have been discovered when considering only the proximity of the sequence. Such evidence synthesis requires an approach which can relate datasets at different resolutions and link genetic and epigenetic features to the spatial scale at which they work.

Physics and statistical mechanics have provided valuable lessons about the genome organization. By modelling the chromatin fibre as a polymeric chain of beads, researchers have applied polymer physics to study the expected genome topology and dynamics. In particular, the distance and probability of contact of between beads (loci), given their distance on the sequence, were a subject for much research (Mirny, 2011). Measurements obtained with fluorescence in-situ hybridization (FISH) and HiT-3C experiments were used to motivate and validate hypotheses concerning these properties, respectively, and, together with general characteristics of polymer chain behaviour were used to propose different models of chromatin organization (Barbieri *et al.*, 2012; Bohn *et al.*, 2007; Lieberman-Aiden *et al.*, 2009; Mateos-Langerak *et al.*, 2009; Munkel *et al.*, 1999; Sachs *et al.*, 1995). The increasing volume of HiT-3C data calls for the development of algorithms and data structures that can summarize, compress and integrate datasets while considering issues of time and space complexity. Such a computational approach can complement physics simulations and provide the necessary framework for large scale studies of the 3D genome architecture and its function.

In this paper, we present a HBM-based algorithmic framework for topological analysis and integration of HiT-3C data. We first introduce the HBM representation for contact maps and give the necessary definitions (Section 2.1). Next, we describe chromoHBM, an algorithm which compresses a contact map into a HBM by iteratively detecting dense modules (communities) of interacting chromatin segments (Section 2.2). We then focus on the application of HBMs for 3C data integration. We explain the difficulties involved in directly combining HiT-3C datasets and propose to merge their HBM representations instead, using the chromoHBM-3C algorithm (Section 2.3). We also note that spatial inference can benefit from incorporating topological knowledge and describe chromoHBM-3D, an algorithm which takes a 3D reconstruction method and iteratively guides it by means of HBM traversal (Section 2.4). In order to evaluate our approach, we study Hi-C datasets and show that HBMs highlight key topologies and that the merged HBM representation can capture multiple scales of spatial organization that could not have been detected by separately analyzing each dataset. For the sake of space, robustness and running time evaluation is given in the supplementary (suppl.) information (Table 1). Additional results of related or exploratory analysis are also available on request.

Table 1. Evaluation results and their availability.

Evaluation	Algorithm	Availability
Consistency with FISH	chromoHBM	Section 3
Consistency with FISH	chromoHBM-3C	Section 3
Consistency with epigenetic data	chromoHBM-3C	Section 3
Consistency with FISH	chromoHBM-3D	Suppl.
Consistency with topological domains	chromoHBM	Suppl.
Robustness (community detection method)	chromoHBM	Suppl.
Robustness (filtering)	chromoHBM	Suppl.
Runtime analysis	chromoHBM	Suppl.
Runtime analysis	chromoHBM-3C	Suppl.
Runtime analysis	chromoHBM-3D	Suppl.
Consistency of contact maps with FISH	Pre-processing	AOR
Reproducibility between replicates	Pre-processing	AOR

AOR, available on request; Suppl., supplementary information.

2 Methods

2.1 From contact maps to HBMs: an introduction for bioinformaticians

3C experiments measure the frequency of ligation events between chromatin fragments in a population of nuclei. For a given pair of genomic bins, each spanning one or more chromatin fragments, we define their contact frequency as follows:

Definition 1 (**contact frequency**): Let a genomic bin be a vector of consecutive chromatin fragments. Let $a = (a_1, a_2, ..., a_k)$ be a genomic bin with k fragments and $b = (b_1, b_2, ..., b_l)$ a genomic bin with l fragments. The contact frequency between a and b is then: $f_{a,b} = \sum_{i=1}^{k} \sum_{j=1}^{l} e_{a_i,b_j}$ where e_{a_i,b_j} is the number of ligation events between fragments a_i and b_j, for $1 \leq i \leq k$, $1 \leq j \leq l$.

We can further summarize these data with a matrix whose entries give the pairwise contact frequencies between two vectors of genomic bins. This leads to the following definition:

Definition 2 (**contact map**): Let $p = (p_1, p_2, ..., p_m)$ be a vector of m consecutive genomic bins and $q = (q_1, q_2, ..., q_n)$ a vector of n consecutive genomic bins. The contact map of p and q is a $m \times n$ matrix A, with $A_{i,j} = f_{p_i,q_j}$, where f_{p_i,q_j} is the contact frequency between bins p_i and q_j, for $1 \leq i \leq m$, $1 \leq j \leq n$.

When all the genomic bins in p and q are from the same chromosome, we call A a *cis* contact map. Otherwise, A is a *trans* contact map. From Definition 1 we also get that A is a non-negative matrix and that if p and q are identical then it is also symmetric.

In practice, contact frequencies are biased measures of ligation events (Yaffe and Tanay, 2011). Consequently, contact maps need to be corrected before any subsequent analysis. In addition, even in the bias-free case, their exact spatial interpretation is not straightforward. For example, we have recently found that contact frequencies and distances measured with FISH are correlated only for short range distances (Shavit *et al.*, 2014). Thus, de-noised contact maps should also be filtered in order to discard artefacts and data that are not spatially meaningful. Here, we concentrate on the topological analysis and data integration of *cis* chromosomal contact maps and assume that they are de-noised and filtered (an evaluation of the effect of filtering is described in the suppl. information).

Since we are interested in recovering the topology of chromosomes from their contact maps, the concept of networks immediately comes to mind. In fact, *cis* contact maps were previously modelled as adjacency matrices of weighted undirected graphs (Boulos *et al.*, 2013). This leads to the definition of a contact network:

Definition 3 (**contact network**): Let $p = (p_1, p_2, \ldots, p_m)$ be a vector of m genomic bins, from the same chromosome. The contact network of p is a weighted undirected graph $N(V, E, w)$ where:

- $V = \{p_1, p_2, \ldots, p_m\}$ is a set of nodes,
- $E = \{\{p_i, p_j\} | f_{p_i,p_j} > 0\}$ is a set edges, where f_{p_i,p_j} is the contact frequency between the nodes (bins) p_i and p_j, for $1 \le i, j \le m$, and
- $w : E \rightarrow \mathbb{R}^+$ is a weight function which assigns each edge with the contact frequency between its elements: $w(\{p_i, p_j\}) = f_{p_i,p_j}$ and $\mathbb{R}^+$ is the set of strictly positive real numbers.

It follows that A, the contact map of p (a symmetric *cis* contact map), is the $m \times m$ adjacency matrix of N:

$$A_{i,j} = \begin{cases} w(\{p_i, p_j\}) & f_{p_i,p_j} > 0, \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

Based on Definition 3 and Equation 1 we can derive a contact map from a contact network and vice versa. The network representation of a contact map (i.e., its contact network) provides an intuitive starting point for analyzing its topology.

Recently, several studies have demonstrated that both the fine structure and long range scaling behaviour of Hi-C data is consistent with a fractal globule model of genome folding (Grosberg *et al.*, 1988; Lieberman-Aiden *et al.*, 2009; Mirny, 2011; Nazarov *et al.*, 2015). From a modelling point of view, the packing of the crumpled globule is formed by iteratively folding the polymer (chromosome) chain such that folds at one scale are grouped together in the next scale, ultimately forming a ‘fold of folds’ (Grosberg *et al.*, 1988). The resulting structure can be described with a hierarchical block Parisi matrix that is consistent with characteristics observed in Hi-C data (Nazarov *et al.*, 2015). Parisi matrices consist of a growing block, placed along the diagonal, which is itself a Parisi matrix of a smaller size. This class of matrices is a sub-class of hierarchical matrices which can be used to represent large-scale and dense matrices with logarithmic-linear complexity (Hackbusch, 1999). Similarly to Parisi matrices, HBMs can capture self-similar hierarchical structures but in these data structures an entry in the matrix takes the lowest level in the hierarchy. More formally, we define an HBM as follows:

Definition 4 (**hierarchical block matrix (HBM)**): Let N be an undirected graph $N(V, E, w)$, with V a set of m nodes, E a set of edges between them and w a function which assigns each edge with a weight. If w is defined as $w : E \rightarrow \{1\}$ then we say that N is unweighted, otherwise N is a weighted graph. Let C be the set of clusters in N, $C = \{c_l\}_{l=1}^{k}$ where $c_l \subseteq V$ and $k \ge 1$. We denote $B(1)$ to be a $m \times m$ matrix, with:

$$B(1)_{i,j} = \begin{cases} 1 & i, j \in c, \quad c \in C, \quad c \subseteq V, \\ 0 & \text{otherwise} \end{cases}$$

Let $N_1(V_1, E_1, w_1)$ be an undirected graph whose nodes are the clusters in N and $C_1 = \{c_{1,l}\}_{l=1}^{k_1}$ is the set of the clusters in N_1, with $c_{1,l} \subseteq V$ and $k_1 \ge 1$. Note that each cluster in C_1 is a union of sets (clusters in C) that contain nodes in V. Using a recursive definition, we denote N_s to be an undirected graph whose nodes are the clusters in N_{s-1} and B(s) to be a m × m matrix, with:

$$B(s)_{i,j} = \begin{cases} 1 & i, j \in c, \quad c \in C_{s-1}, \quad c \subseteq V, \\ 0 & \text{otherwise} \end{cases}$$

where $C_{s-1} = \{c_{s-1,l}\}_{l=1}^{k_{s-1}}$ is the set of clusters in N_{s-1} with $c_{s-1,l} \subseteq V$ and $k_{s-1} \ge 1$, for $s \ge 2$. Note that if $B(s)_{i,j} = 1$ than for all $s' > s$, $B(s')_{i,j} = 1$ as well.

The hierarchical block matrix (HBM) of N is a non-negative symmetric m × m matrix, H, with: $H_{i,j} = \min_s\{s | B(s)_{i,j} = 1\}$, for $s \ge 1$.

If the underlying topology of a chromosome is hierarchical then we expect its contact network to consist of communities (clusters) that will form mega-communities and so forth. This is in keeping with the general definition of communities as dense sub-graphs of a sparser graph (Radicchi *et al.*, 2004). The resulting hierarchy can then be described with an HBM; given a contact map A and its contact network N, we can compute H from N using Definition 4

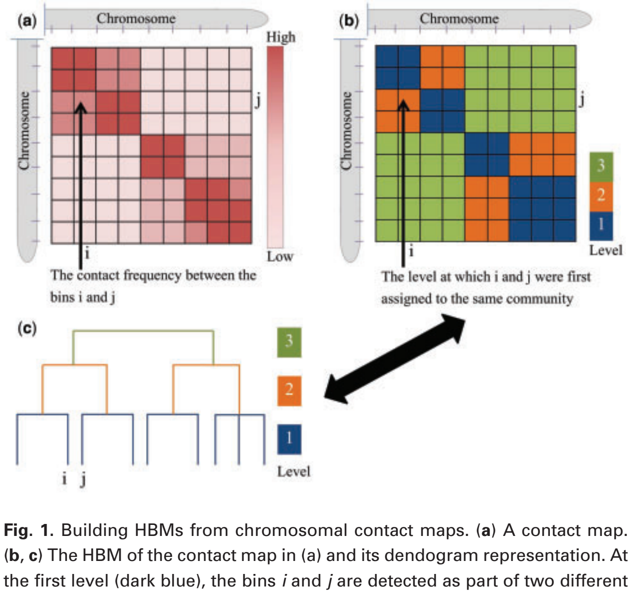

(a)
Chromosome
High
Chromosome
j
i
Low
The contact frequency between the bins i and j
(b)
Chromosome
Chromosome
j
3
2
1
i
Level
The level at which i and j were first assigned to the same community
(c)
3
2
1
i j
Level

Fig. 1. Building HBMs from chromosomal contact maps. (**a**) A contact map. (**b**, **c**) The HBM of the contact map in (a) and its dendogram representation. At the first level (dark blue), the bins *i* and *j* are detected as part of two different communities in the contact network of the contact map in (a). They appear as leaves in the dendogram and belong to two separate clusters. At level 2 (orange), *i* and *j* are first assigned to the same community and the clusters in the dendogram are merged into one. The *i*, *j*th entry of the HBM (pointed to by a black arrow) is set accordingly to 2. At the third and last level (green) all the bins are merged into one single cluster. The HBM in (b) can be derived from the dendogram in (c) (and vice versa) using a traversal from the lowest to the highest level

by iteratively detecting and merging communities in N until all communities are merged into one or cannot be merged any further. Figure 1 illustrates this idea. H can be further summarized with a hierarchical tree (or a dendogram), which describes the merges of communities at each level in a succinct way (Figure 1c). Taking a data-driven perspective, H also allows us to first incorporate local information and gradually include more global constraints, which is desirable when considering contact frequencies (Shavit *et al.*, 2014).

In the next sections we describe three HBM-based algorithms: chromoHBM, chromoHBM-3C and chromoHBM-3D. The implementation of these algorithms is available at: http://www.cl.cam.ac.uk/ys388/hbm/. Details of data preparation (datasets used and subsequent pre-processing) and performance evaluation are given in the suppl. information (see also Table 1).

2.2 chromoHBM: compressing contact maps into topological features

The chromoHBM algorithm (Algorithm 1) derives the HBM representation of a (de-noised and filtered) *cis* contact map. It takes a $m \times m$ symmetric *cis*-contact map A and returns its HBM H. chromoHBM starts by initializing H to be a $m \times m$ matrix (Algorithm 1, line 3) and then removes self-interactions from A (by setting its diagonal to zero) and creates Adj, a copy of A (Algorithm 1, lines 4–5). Next, H is iteratively populated. At each iteration, s, chromoHBM calls *detectCommunities* which returns the communities in N, the network whose adjacency matrix is defined by Adj. It then updates Adj to be a $m_s \times m_s$ matrix, with m_s the number of identified communities. $Adj_{i,j}$ is set to be a normalized sum of the contact frequencies between the bins in A that are elements in the i^{th} and j^{th} communities in N, respectively:

$$Adj_{i,j} = \frac{\sum_{k \in c_i} \sum_{l \in c_j} A_{k,l}}{|c_i||c_j|} \qquad (2)$$

where c_r is the r^{th} community in N and $|c_r|$ is its size, for $1 \leq r \leq m_s$ and $i \leq j$. Finally, chromoHBM updates H to record the communities detected at the current iteration (level) s by setting $H_{k,l}$ to s if the bins k and l in A are assigned to the same community for the first time, for $1 \leq k, l \leq m$ (Algorithm 1, lines 26–30). The iterative process halts when a single community has been identified (all communities were merged) or when it is not possible to merge any of the communities detected at the previous level (Algorithm 1, while-loop condition, line 10). At this point, s is $\max(H) + 1$ and H may include entries that have not been set (if some communities could not have been merged). chromoHBM updates such entries to take the value of s in order to represent the global community formed by all the communities in the network (Algorithm 1, lines 33–37).

The time complexity of the s_{th} iteration in chromoHBM is given by:

$$T(s, m_s) = T(detectCommunities, Adj_s) + \sum_{i=1}^{m_s} \sum_{j=i+1, i<m_s}^{m_s} |c_i||c_j| \qquad (3)$$

where $T(g, m)$ is the time complexity of g with an input of size m. The first term in Equation 3 gives the time complexity of *detectCommunities* which depends on the size of Adj at iteration s, Adj_s. This method is intentionally left unspecified since we expect that the underlying hierarchy will consistently emerge regardless of the implementation used (an evaluation of the robustness of HBMs generated with chromoHBM, when using different detection methods, is described in the suppl. information). The second term describes the number of operations required to update Adj and H, which depends on m_s, the number of communities detected at iteration s and on $|c_r|$, the size of community c_r, for $1 \leq r \leq m_s$. Since the final update of H takes another $O(m^2)$ operations (Algorithm 1, lines 33–37), the total time complexity of the algorithm is given by:

$$T(chromoHBM) = O(m^2) + \sum_s T(s, m_s) \qquad (4)$$

Algorithm 1. chromoHBM

```
 1: procedure chromoHBM($A$)
 2:   $m \leftarrow Nrow(A)$
 3:   $H \leftarrow$ matrix($m$, $m$)
 4:   $diag(A) \leftarrow 0$                    ▷ remove self-interactions
 5:   $Adj \leftarrow A$          ▷ the adjacency matrix of a network $N$
 6:   $m_0 \leftarrow m$
 7:   $m_s \leftarrow m_0$
 8:   $s \leftarrow 1$
 9:
10:    while $m_s > 1$ and ($m_s < m_0$ or $s = 1$) do
11:        $c \leftarrow detectCommunities(Adj)$
12:        $m_0 \leftarrow m_s$
13:        $m_s \leftarrow length(c)$      ▷ number of communities in $N$
14:        $Adj \leftarrow matrix(m_s, m_s)$
15:
16:        for $i \leftarrow 1$ to $m_s$ do
17:          for $j \leftarrow i$ to $m_s$ do
18:            if $i \neq j$ then
19:              $f \leftarrow 0$
20:              for all $k \in c_i$ do
21:                for all $l \in c_j$ do
22:                   $f \leftarrow f + A_{k,l}$
23:              $Adj_{i,j} \leftarrow \frac{f}{|c_i||c_j|}$
24:              $Adj_{i,j} \leftarrow Adj_{j,i}$
25:
26:          for all $k \in c_i$ do
27:            for all $l \in c_i$ do
28:              if $H_{k,l}$ is NULL then              ▷ first time we
                 find $k$ and $l$ together
29:                 $H_{k,l} \leftarrow s$
30:                 $H_{l,k} \leftarrow H_{k,l}$
31:    $s \leftarrow s + 1$
32:
33:  for $i \leftarrow 1$ to $m - 1$ do
34:     for $j \leftarrow i + 1$ to $m$ do
35:        if $H_{i,j}$ is NULL then                     ▷ first update
36:           $H_{i,j} \leftarrow s$
37:           $H_{j,i} \leftarrow H_{i,j}$
38:  return $H$
```

The space complexity of chromoHBM is $O(m^2)$ since we store Adj and H in memory. This requirement could be relaxed by implementing H as a hierarchical tree.

2.3 chromoHBM-3C: merging HBMs with different bin sizes

High-throughput 3C techniques generate contact maps at various resolutions and bin sizes. To date, however, there are no methods for putting together contact maps that differ in their binning, resolution or protocol. We focus on the integration of *cis* symmetric contact maps and define the problem of HiT-3C data integration as follows:

Definition 5 (**3C data integration**): Let $x = [x_b, x_e]$ be a genomic range starting at position x_b and ending at position x_e on a given genomic sequence. A y-binning of x divides it into $\frac{x_e - x_b + 1}{y}$ genomic bins of length y, where y is a natural number. Let $u = (u_1, u_2, ..., u_m)$ and $v = (v_1, v_2, ..., v_k)$ be two vectors of genomic bins, generated by a r and a r/n binning of x, respectively. For simplicity, we will assume that n and $\frac{x_e - x_b + 1}{r}$ are natural numbers. Since $\frac{r}{n}$ is also a natural number (by the definition of binning), we get that $k = nm$. Let A and B be the contact maps of u and v, respectively; the problem of integrating A and B is to create a new k × k contact map which faithfully represents the information in A and in B.

Several factors make the 3C data integration problem particularly challenging. First, differences in resolution and in the molecular protocol can lead to a considerable variation in the contact frequency range, which in turn makes it difficult to combine (and compare) contact maps. Other variations in the experimental setting can also contribute to this effect. In addition, merging contact maps with different bin sizes require data processing, such as expansion or normalization of contact frequencies, which is not straightforward to perform and may lead to data distortion. Since we are ultimately interested in putting together the underlying topologies that are captured by contact maps, the HBM representation offers an attractive surrogate. Thus, instead of directly combining two contact maps, which is a difficult task, we can merge their HBMs and generate a unified representation of their topologies.

We follow the definition of the 3C data integration problem (Definition 5) and denote H^A to be the $m \times m$ HBM of A and H^B to be the $k \times k$ HBM of B. Since $k = nm$, we can expand H^A to be a $k \times k$ matrix, $\hat{H}^A$, with:

$$\hat{H}^A_{i,j} = H^A_{p,q}, \qquad p = \lceil \frac{i}{n} \rceil, \qquad q = \lceil \frac{j}{n} \rceil \tag{5}$$

for $1 \le i, j \le k$ and $1 \le p, q \le m$.

We define a 'merging matrix', M, given by:

$$M = \frac{\hat{H}^A + H^B}{2} \tag{6}$$

where $M_{i,j}$ takes the 'average level' between the levels at which i and j were first assigned to the same community. Given M, we can compute the 'merged HBM', $H^{A,B}$ which integrates A and B by merging the topologies of H^A and H^B. Let $s = \{s_i\}_{i=1}^{l_M}$ be the set of levels (unique values) in M, sorted in an increasing order, where l_M is the total number of levels in M. $H^{A,B}_{i,j}$ is given by:

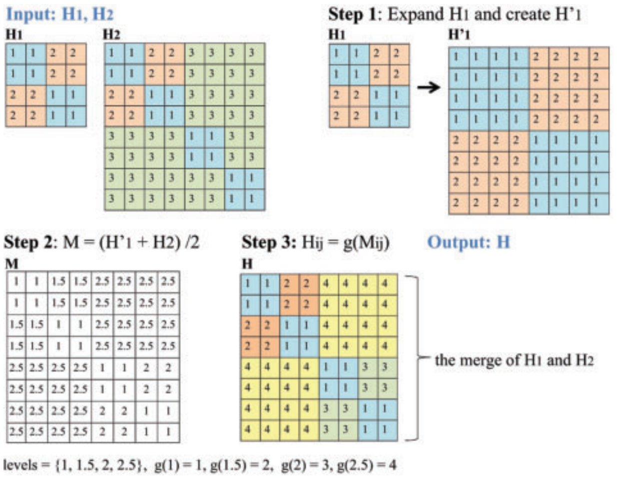
$$H^{A,B}_{i,j} = g(M_{i,j}) \tag{7}$$

where g is a function which takes the value of a level in M and returns its index in s:

$$g(s_i) = i, \qquad s_i \in s \tag{8}$$

The computation of $H^{A,B}$ mimics a bottom-up traversal of M, where new levels are added in order to accommodate for the, possibly different, topologies of H^A and H^B (note that when $H^A = H^B$ then also $H^{A,B} = H^B$). Figure 2 illustrates the steps involved in merging two HBMs, using a toy example.

The chromoHBM-3C algorithm (Algorithm 2) takes a $m \times m$ HBM H^A and a $k \times k$ HBM H^B where $k = nm$ and n is a natural number. It follows the steps described above and returns $H^{A,B}$, the merged HBM of H^A and H^B. First, it generates $\hat{H}^A$, the expanded version of H^A, according to Equation 5 (Algorithm 2, lines 6–11). Second, it creates the merging matrix, M, by taking the 'average' of $\hat{H}^A$ and H^B. Lastly, chromoHBM-3C traverses the levels of M, from the lowest to the highest, and updates $H^{A,B}$ according to Equation 7 (Algorithm 2, lines 16–21). For each level s_l in M, chromoHBM-3C iterates through the entries of M and updates the entry $H^{A,B}_{i,j}$ to be l if $M_{i,j}$ equals s_l. At the end of the traversal H contains the number of levels required to merge the topologies in H^A and H^B. The expansion and merge steps take $O(k^2)$ operations each and the bottom-up traversal requires another $O(l_M k^2)$ operations (where l_M is the number of levels in M). We also create and update matrices of size k^2. Hence, the total time and space complexities of the algorithm are $O(l_M k^2)$ and $O(k^2)$, respectively.


Input: H1, H2
H1
H2
Step 1: Expand H1 and create H'1
H1
H'1
Step 2: M = (H'1 + H2) /2
M
Step 3: Hij = g(Mij)
H
Output: H
the merge of H1 and H2
levels = {1, 1.5, 2, 2.5}, g(1) = 1, g(1.5) = 2, g(2) = 3, g(2.5) = 4

H1

1	1	2	2
1	1	2	2
2	2	1	1
2	2	1	1

H2

1	1	2	2	3	3	3	3
1	1	2	2	3	3	3	3
2	2	1	1	3	3	3	3
2	2	1	1	3	3	3	3
3	3	3	3	1	1	3	3
3	3	3	3	1	1	3	3
3	3	3	3	3	3	1	1
3	3	3	3	3	3	1	1

H'1

1	1	1	1	2	2	2	2
1	1	1	1	2	2	2	2
1	1	1	1	2	2	2	2
1	1	1	1	2	2	2	2
2	2	2	2	1	1	1	1
2	2	2	2	1	1	1	1
2	2	2	2	1	1	1	1
2	2	2	2	1	1	1	1

M

1	1	1.5	1.5	2.5	2.5	2.5	2.5
1	1	1.5	1.5	2.5	2.5	2.5	2.5
1.5	1.5	1	1	2.5	2.5	2.5	2.5
1.5	1.5	1	1	2.5	2.5	2.5	2.5
2.5	2.5	2.5	2.5	1	1	2	2
2.5	2.5	2.5	2.5	1	1	2	2
2.5	2.5	2.5	2.5	2	2	1	1
2.5	2.5	2.5	2.5	2	2	1	1

H

1	1	2	2	4	4	4	4
1	1	2	2	4	4	4	4
2	2	1	1	4	4	4	4
2	2	1	1	4	4	4	4
4	4	4	4	1	1	3	3
4	4	4	4	1	1	3	3
4	4	4	4	3	3	1	1
4	4	4	4	3	3	1	1

Fig. 2. Merging HBMs with different bin sizes. Two HBMs, H_1 and H_2 are generated from contact maps with a different binning, where a bin in H_1 corresponds to two bins in H_2. In Step 1, we expand H_1 so that the binning of the two HBMs will match (Equation 5). The average between the expanded HBM, H_1', and H_2 gives the merging matrix M (Step 2, Equation 6). Finally, at Step 3, we compute the merged HBM, H, using M. Each entry in H takes the index of the level value in the matching entry in M, which can be computed with the function g (Equations 7–8)

2.4 chromoHBM-3D: guiding 3D positioning with HBMs

Current methods for 3D positioning typically use available constraints (contact frequencies) all together. If chromosomes are hierarchically organized, then reconstructing their configuration could benefit from a bottom-up approach, which starts by positioning the smaller, more local domains, and gradually proceeds by placing them relative to each other (instead of inferring all the positions at once). Since HBMs capture and relate multiple levels of topological features, they provide a natural means for implementing such an approach.

The chromoHBM-3D algorithm (Algorithm 3) takes five arguments as input:

- A, a $m \times m$ contact map,
- H, the HBM of A,
- f, a 3D reconstruction method which takes a contact map and returns its 3D positioning,
- t, a function which converts distances into contact frequencies (for example, using the inverse function), and
- arg, a list of any additional arguments for f.

Algorithm 2. chromoHBM-3C

```
procedure chromoHBM-3C($H^A$ , $H^B$)
    $m \leftarrow Nrow(H^A)$
    $k \leftarrow Nrow(H^B)$
    $n \leftarrow \frac{k}{m}$    ▷ we assume that $n$ is a natural number
    $\hat{H}^A \leftarrow matrix(k,k)$    ▷ expand $H^A$
    for $i \leftarrow 1$ to k do
        $p \leftarrow \lceil \frac{i}{n} \rceil$
        for $j \leftarrow i$ to k do
            $q \leftarrow \lceil \frac{j}{n} \rceil$
            $\hat{H}^A_{i,j} \leftarrow H^A_{p,q}$
            $\hat{H}^A_{j,i} \leftarrow \hat{H}^A_{i,j}$
    $M \leftarrow \frac{\hat{H}^A + H^B}{2}$
    $H^{A,B} \leftarrow matrix(0,k,k)$
    $s \leftarrow getLevels(M)$    ▷ level values in increasing order
    $l_M \leftarrow length(s)$    ▷ number of levels
    for $l \leftarrow 1$ to lM do
        for $i \leftarrow 1$ to k do
            for $j \leftarrow i$ to k do
                if $M_{i,j} = s_l$ then
                    $H^{A,B}_{i,j} \leftarrow l$
                    $H^{A,B}_{j,i} \leftarrow H^{A,B}_{i,j}$
    return $H^{A,B}$
```

Given this input, chromoHBM-3D performs a bottom-up traversal of H and returns the 3D positioning of A. It first creates a $m \times m$ distance matrix, D, that will be updated during the traversal. For each level l, for each community $c_k \in c$ (for c, the set of communities detected at level l), chromoHBM-3D calls f with the contact map of c_k, A^{c_k}, given by $A^{ck}_{i,j} = A_{i,j}$ for all $i,j \in c_k$. Note that small communities are skipped since they do not provide enough spatial constraints (Algorithm 3, line 9, where $|c_k|$ is the size of c_k). Given Y, the 3D configuration reconstructed with f, chromoHBM-3D updates $D_{i,j}$ to take the Euclidean distance between i and j in Y, for $i,j \in c_k$. This update is carried out only for entries that have not been updated at previous iterations (based on 'more local' constraints). At the end of the iterative procedure, chromoHBM-3D transforms D into a 'contact map' A', using t, and calls f with A'. The resulting 3D configuration is a 'refined' 3D positioning of A based on the information in H.

The running time of chromoHBM-3D depends on the number of levels and communities in H and on the time complexity of f. Let l_H be the number of levels in H, m_l, the number of communities at level l, and $T(f, n)$, the time complexity of f with an input of size n. The time complexity of chromoHBM-3D is given by:

$$T(chromoHBM - 3D) = O(m^2 + T(f,m) + \sum_{l=1}^{l_h} \sum_{k=1}^{m_l} T(f,|c_k|) + |c_k|^2) \quad (9)$$

The first and second terms in Equation 9 give the number of operations required to derive A' from D and the time complexity of reconstructing Y from A', respectively (Algorithm 3, lines 18–22). The time complexity of the bottom-up traversal is given by the third term (nested sums). For each level l, we call f with the contact map of the community c_k which takes $T(f,|c_k|)$ operations. Updating D requires additional $|c_k|^2$ operations.

Algorithm 3. chromoHBM-3D

```
 1: procedure chromoHBM-3D($A$, $H$, $f$, $t$, $arg$)
 2:     $m \leftarrow Nrow(A)$
 3:     $D \leftarrow matrix(m,m)$
 4:     $l_H \leftarrow Nlevels(H)$    ▷ number of levels in $H$
 5:     for $l \leftarrow 1$ to $l_H$ do
 6:         $c \leftarrow communitiesAtLevel(l)$    ▷ communities detected at level $l$
 7:         for all $c_k \in c$ do
 8:             $m_l \leftarrow |c_k|$
 9:             if $m_l > 4$ then
10:                 $A^{c_k} \leftarrow matrix(m_l, m_l)$
11:                 for all $i,j \in c_k$ do
12:                     $A^{c_k}_{i,j} \leftarrow A_{i,j}$
13:                 $Y \leftarrow f(A^{c_k}, args)$    ▷ the 3D positioning of $A^{c_k}$
14:                 for all $i,j \in c_k$ do
15:                     if $D_{i,j}$ is NULL then    ▷ first update
16:                         $D_{i,j} \leftarrow \sqrt{\sum_{a=1}^{3} (Y_{i,a} - Y_{j,a})^2}$
17:     $A' \leftarrow matrix(m,m)$
18:     for $i \leftarrow 1$ to $m - 1$ do
19:         for $j \leftarrow i + 1$ to $m$ do
20:             $A'_{i,j} \leftarrow t(D_{i,j})$    ▷ convert distances into 'contact frequencies'
21:             $A'_{j,i} \leftarrow A'_{i,j}$
22:     $Y \leftarrow f(A', args)$
23:     return $Y$
```

The space complexity of chromoHBM-3D depends on the memory requirements of f but has a lower-bound of $O(m^2)$ (the size of the distance matrix D).

3 Results

Let A and B be two contact maps, H^A and H^B their HBMs, respectively, and $H^{A,B}$ the merged HBM of H^A and H^B, we would like to test:

- **Is $H^{A,B}$ 'consistent'?** If $H^{A,B}$ is consistent then its topology should be in agreement with the separate topologies of H^A and H^B as well as with previous findings.
- **Is $H^{A,B}$ 'powerful'?** If $H^{A,B}$ is powerful then it should add new information, which is not available when independently analyzing H^A and H^B. Also, there should be a clear motivation for not directly comparing A and B (for example, large variation in contact frequency distribution and range).

Further evaluation of the robustness, consistency and performance (running time) of chromoHBM, chromoHBM-3C and chromoHBM-3D is described in the suppl. information (see Table 1).

In order to evaluate how 'consistent' and 'powerful' separate and merged HBMs are, we study two contact maps and their HBMs, which considerably differ in their binning: a 500 and a 25 kb contact maps generated from human fibroblasts nuclei for chromosome 1 (Rao *et al.*, 2014). We first compare the corrected (de-noised) contact maps. We find that the 25 kb contact map is sparser and that the distributions of the two contact maps are significantly different (one-sided *t*-test (significant differences are observed also when considering non-parametric tests such as the Kolmogorov-Smirnov test), *P*-value

$< 2.2 \times 10^{-16}$). The 500 kb contact frequencies range between 0 and 101 500 with a mean of 427.10 and a median of 31.57, while the distribution of the 25 kb contact frequencies is more (rightly) skewed; the 25 kb contact frequencies range between 0 and 6221 and have a mean of 1.074 and a zero median. In order to allow for a more direct comparison, we standardize the two contact maps and compute the expanded version of the 500 kb contact map, using Equation 5. Testing for agreement between the two standardized contact maps (which now match in their dimensions) yields a low Pearson correlation value ($r = 0.15$). Put together, the notable variations between the 500 and 25 kb contact maps, make it difficult to integrate them and motivate the use of their HBM representation for this purpose.

We next apply chromoHBM in order to generate the HBM for each of the contact maps, using the Infomap algorithm (Rosvall and Bergstrom, 2008) for detecting communities. This algorithm, which is recognized as a leading approach for community detection in undirected weighted graphs (Lancichinetti and Fortunato, 2009), aims at finding an optimal compression of the information flow in a graph, using random walks as a proxy for the amount of information transferred between nodes. Its key idea, inspired by Information Theory, is that finding a compression that can closely recover the information flow in a graph is equivalent to detecting its topological modules or communities (which dictate the dynamics of the system). In the suppl. information we further show that chromoHBM recovers similar HBMs, independently of the detection method used (robustness evaluation, see Table 1).

Given their HBMs, we apply chromoHBM-3C to compute the merged HBM representation of the two contact maps. For description purposes, we use H_{500}, H_{25} and H_{merged} to denote the 500 kb, 25 kb and the merged HBMs, respectively. Figure 3 shows H_{500} (enlarged for visual clarity), H_{25} and H_{merged}. We notice that some topological patterns that appear in H_{500} are missing from H_{25} and vice versa. Small communities that are detected at the first level of H_{25}, are missing or merged in H_{500}. At the second level, the two HBMs detect mega communities that roughly correspond or overlap. These communities are further combined into two large clusters (in blue) at the third level of H_{500}. This level of hierarchical organization is completely absent from H_{25}. At the last level (level 5 and level 4 for H_{500} and H_{25}, respectively), bins which correspond to regions with a poor mappability (and thus poorly covered) are merged with the others to form the global community.

From a signal processing point of view (according to the Heisenberg uncertainty principle), the observed trade-off, between H_{500} and H_{25}, is expected. Smaller bins can better detect small scale topologies but miss large scale features (good knowledge of frequency but poor knowledge of position). In contrast, larger bins capture large scale organization on the expense of finer details (good knowledge of position but poor knowledge of frequency). H_{merged} (Figure 3c) accommodates for the distinct scales that are captured by H_{25} and H_{500}, and provides a multi-scale view of the organization of chromosome 1. Its first level matches the first level of H_{25} (a small scale topology) while its fifth level roughly corresponds to the third level of H_{500} (a large scale topology). At intermediate levels, overlapping features detected by both HBMs are combined together, forming a unified view of the separate hierarchies.

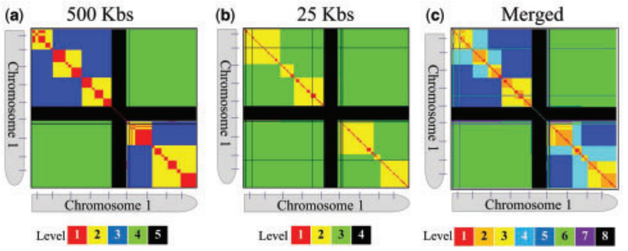


Fig. 3. HBMs of chromosome 1: before and after merge. (**a**) 500 kb HBM of chromosome 1 (enlarged for visual clarity). (**b**) 25 kb HBM of chromosome 1. (**c**) The merged HBM of the HBMs in (a) and (b), generated with chromoHBM-3C

In order to further evaluate the consistency of H_{merged}, we study active and repressed domains in chromosome 1. These domains belong to a group of regions which appear throughout the genome, termed 'ridges' and 'anti-ridges', which show distinct functional and structural characteristics. Ridge domains are rich with genes and active regulatory elements, and are highly transcribed and expressed ('active') (Gilbert *et al.*, 2004; Versteeg *et al.*, 2003). They possess an 'open' chromatin topology which further correlates with their high expression level. In contrast, anti-ridge domains are gene poor, lowly transcribed and expressed and present a 'close' and packed topology ('repressed'). Figure 4 shows the HBMs of chromosome 1, at positions 1.50×10^8 to 1.54×10^8 (green frame, upper panel) and 1.73×10^8 to 1.76×10^8 (red frame, lower panel), which correspond to a ridge and an anti-ridge domain, respectively. For these specific regions, differences in density were also confirmed with FISH (distances within the ridge domain shown to be larger than distances within the anti-ridge domain) (Mateos-Langerak *et al.*, 2009).

In line with the trend observed when analyzing the entire chromosome, H_{25} captures finer topological details compared to H_{500}. At the first level, the ridge domain of H_{25} consists of four communities, which roughly correspond to two communities in H_{500}. In a similar way, the two communities detected in the anti-ridge domain in H_{25} form a single community in H_{500}. When considering H_{merged}, we find that the anti-ridge domain presents a flatter hierarchy compared to the ridge domain (two levels versus three levels, respectively) with a smaller number of larger communities. This is in agreement with the characteristics of anti-ridge domains, which are more dense than ridge

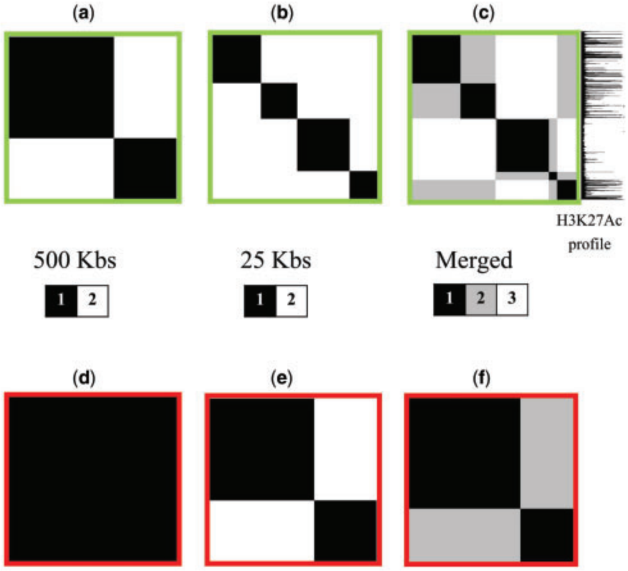


Fig. 4. HBMs of chromosome 1 at ridge and anti-ridge domains. The upper panel (green frame) shows the 500 kb (**a**), 25 kb (**b**) and merged (**c**) HBMs of chromosome 1 at positions 1.50×10^8 to 1.54×10^8, which correspond to a ridge ('active') domain. The respective HBMs at the lower panel correspond to an anti-ridge ('repressed') domain at positions 1.73×10^8 to 1.76×10^8. The vertical track next to the merged HBM of the ridge domain (c) shows the picks of the H3K27Ac histone modification mark at the corresponding positions in the genome (chromosome 1: 1.50×10^8 to 1.54×10^8). This track was retrieved from the UCSC Genome Browser website (Kent *et al.*, 2002)

domains. For the ridge domain, the first and third level of H_{merged} closely correspond to the first and second level of H_{25}, respectively. The second level of H_{merged}, reveals, however, an additional scale that is missing from H_{25} and H_{500}; the last and first two communities (along the diagonal), which are detected at the first level, form one mega community while the other two communities are merged into a second mega community. This topology recapitulates the picks of a regulatory mark (shown as a vertical track next to H_{merged} in Figure 4c), which is associated with a high transcription activity (the H3K27Ac histone mark). The first mega community corresponds to H3K27Ac picks while the second matches a flat profile. This partition, which is missing from the separate HBMs, further illustrates the advantage of a merged view of multiple scales.

4 Outlook

Since the introduction of the 3C technique in 2002 (Dekker *et al.*, 2002), various high-throughput derivatives have been developed. Given the latest enhancements in resolution (kb and sub-kb) (Hsieh *et al.*, 2015; Rao *et al.*, 2014; Sexton *et al.*, 2012) and signal-to-noise ratio (Kalhor *et al.*, 2012a,b), as well as the on-going reduction in sequencing costs, HiT-3C techniques are expected to become even more widely used for studying the 3D genome organization. Recent single-cell protocols (Nagano *et al.*, 2013, 2015) also hold the promise to extend the conventional population-based analysis to the single cell level. Integrating these noisy, big and high-dimensional data across resolutions, sub-populations and with other Omics, pose challenges that are not yet addressed.

The HBM representation summarizes the topological features of chromosomal contact maps. This summary is useful not only for gaining biological insights into genome topology, but importantly as a data structure that can facilitate data integration. Putting together HiT-3C HBMs can be performed across resolutions (as demonstrated above) but also across sub-populations. If we consider *n* contact maps of *n* sub-populations, their HBMs give a means to compare the topologies which characterize each sub-population and allow to create an ensemble which captures the variance in the global population. This can be achieved using a simple voting algorithm, where each HBM 'votes' for the communities which comprise it. In this way, one can identify the communities which dominate the population. In addition, simple HBM statistics (e.g., mean and variance) can give insights about the importance of the relative positioning of genes and domains. Beyond the challenge of integrating different 3C datasets, a key problem is how to relate Omics and 3C information (i.e., at what resolution?). Merged HBMs allow to identify the spatial scales and topologies which are relevant for a given Omic of interest and to better link structure and function (see Section 3).

The algorithmic framework presented in this paper can be used as a foundation for analyzing multi-omic data, while considering both their sequence and 3D contexts. Since analysis can be carried independently for each scale, HBMs also lend themselves to parallel computations which is advantageous for big data. Finally, HBMs could complement genome Google-like browsers, which are now starting to emerge (He *et al.*, 2013), since they link information at different resolutions.

Acknowledgements

We thank FP7-Health-F5-2012, under grant agreement 105 no 305280 (MIMOmics).

Conflict of Interest: none declared.

References

Barbieri,M. *et al.* (2012) Complexity of chromatin folding is captured by the strings and binders switch model. *Proc. Natl. Acad. Sci. USA*, **109**, 16173–16178.

Belmont,A.S. (2014) Large-scale chromatin organization: the good, the surprising, and the still perplexing. *Curr. Opin. Cell Biol.*, **26**, 69–78.

Bohn,M. *et al.* (2007) Random loop model for long polymers. *Phys. Rev. E Stat. Nonlinear Soft Matter Phys.*, **76**, 051805.

Boulos,R.E. *et al.* (2013) Revealing long-range interconnected hubs in human chromatin interaction data using graph theory. *Phys. Rev. Lett.*, **111**, 118102.

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

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

Filippova,D. *et al.* (2014) Identification of alternative topological domains in chromatin. *Algorithms Mol. Biol.*, **9**, 14

Gilbert,N. *et al.* (2004) Chromatin architecture of the human genome: gene-rich domains are enriched in open chromatin fibers. *Cell*, **118**, 555–566.

Grosberg,A.Y. *et al.* (1988) The role of topological constraints in the kinetics of collapse of macromolecules. *J. Phys.*, **49**, 2095–2100.

Hackbusch,W. (1999) A sparse matrix arithmetic based on H-matrices. Part I: Introduction to H-matrices. *Computing*, **62**, 89–108.

He,C. *et al.* (2013). GMol: A Tool for 3D Genome Structure Visualization. In: *The Great Lake Bioinformatics Conference, Pittsburgh, PA.*

Hsieh,T.H. *et al.* (2015) Mapping nucleosome resolution chromosome folding in yeast by micro-C. *Cell*, **162**, 108–119.

Kalhor,R. *et al.* (2012a) Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. *Nat. Biotechnol.*, **30**, 90–98.

Kalhor,R. *et al.* (2012b) Solid-phase chromosome conformation capture for structural characterization of genome architectures. *Nat. Biotechnol.*, **30**, 90–98.

Kent,W.J. *et al.* (2002) The human genome browser at UCSC. *Genome Res.*, **12**, 996–1006.

Lancichinetti,A. and Fortunato,S. (2009) Community detection algorithms: a comparative analysis. *Phys. Rev. E Stat. Nonlinear Soft Matter Phys.*, **80**, 056117

Levy-Leduc,C. *et al.* (2014) Two-dimensional segmentation for analyzing Hi-C data. *Bioinformatics*, **30**, i386–i392.

Lieberman-Aiden,E. *et al.* (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, **326**, 289–293.

Mateos-Langerak,J. *et al.* (2009) Spatially confined folding of chromatin in the interphase nucleus. *Proc. Natl. Acad. Sci. USA*, **106**, 3812–3817.

Mirny,L.A. (2011) The fractal globule as a model of chromatin architecture in the cell. *Chromosome Res.*, **19**, 37–51.

Munkel,C. *et al.* (1999) Compartmentalization of interphase chromosomes observed in simulation and experiment. *J. Mol. Biol.*, **285**, 1053–1065.

Nagano,T. *et al.* (2013) Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature*, **502**, 59–64.

Nagano,T. *et al.* (2015) Comparison of Hi-C results using in-solution versus in-nucleus ligation. *Genome Biol.*, **16**, 175.

Nazarov,L.I. *et al.* (2015) A statistical model of intra-chromosome contact maps. *Soft Matter*, **11**, 1019–1025.

Phillips-Cremins,J.E. *et al.* (2013) Architectural protein subclasses shape 3D organization of genomes during lineage commitment. *Cell*, **153**, 1281–1295.

Radicchi,F. *et al.* (2004) Defining and identifying communities in networks. *Proc. Natl. Acad. Sci. USA*, **101**, 2658–2663.

Rao,S.S. *et al.* (2014) A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, **159**, 1665–1680.

Rosvall,M. and Bergstrom,C.T. (2008) Maps of random walks on complex networks reveal community structure. *Proc. Natl. Acad. Sci. USA*, **105**, 1118–1123.

Sachs,R.K. *et al.* (1995) A random-walk/giant-loop model for interphase chromosomes. *Proc. Natl. Acad. Sci. USA*, **92**, 2710–2714.

Sexton,T. *et al*. (2012) Three-dimensional folding and functional organization principles of the Drosophila genome. *Cell*, **148**, 458–472.

Shavit,Y. and Lio',P. (2014) Combining a wavelet change point and the Bayes factor for analysing chromosomal interaction data. *Mol. Biosyst.*, **10**, 1576–1585.

Shavit,Y. *et al*. (2014) FisHiCal: an R package for iterative FISH-based calibration of Hi-C data. *Bioinformatics*, **30**, 3120–3122.

Versteeg,R. *et al*. (2003) The human transcriptome map reveals extremes in gene density, intron length, GC content, and repeat pattern for domains of highly and weakly expressed genes. *Genome Res.*, **13**, 1998–2004.

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.