3D genome reconstruction from chromosomal contacts

Annick Lesne[1,2], Julien Riposo[1], Paul Roger[1], Axel Cournac[3] & Julien Mozziconacci[1,*]

A computational challenge raised by Chromosome Conformational Capture experiments is to reconstruct spatial distances and 3D genome structures from observed contacts between genomic loci. We propose a two-step algorithm, ShRec3D, based on a graph-theoretic metric and distance geometry. We assessed its accuracy using both *in-silico* data and human Hi-C data. This algorithm avoids convergence issues, accommodates sparse and noisy contact maps, and is orders of magnitude faster than existing methods.

Chromosomal Conformation Capture (3C) has been developed for identifying DNA segments in close proximity within a cell nucleus[1]. It involves *in-vivo* formaldehyde cross-linking of protein-mediated DNA-DNA contacts, sonication and re-ligation of cross-linked fragments, followed by sequencing. Next-generation sequencing techniques brought this protocol to the whole-genome scale (Hi-C) in cell populations[2]. Hi-C experiments provide genome-wide maps of contact frequencies between genomic loci for various cell types in human, mouse, fly, yeast and bacteria, presumably reflecting the average spatial organization of their chromosomes.

Several methods have been developed to derive 3-dimensional (3D) chromosomal structures from Hi-C contact maps[3]. Most involve optimization of loci coordinates[4-11] until experimentally measured contacts are satisfactorily reproduced. These methods perform well but convergence issues may arise due to algorithm trapping in local optima, and computation time is often prohibitive for large datasets. Therefore they resort to data binning, at the cost of lowering genomic resolution. We propose a two-step alternative: a novel method adapted from graph analysis for translating contact maps into distances, followed by 3D reconstruction.

Reconstructing a spatial structure from the distances between its elements is an old mathematical problem solved by distance geometry[12] and classical multidimensional scaling[13] (MDS). Four matrices can be associated to a set of N points in 3D space (Fig.1a-e): the 3xN coordinate matrix V whose columns are the barycentric coordinates (Fig.1b); the NxN Gram matrix G comprising the scalar products of coordinate vectors (Fig.1c); the NxN distance matrix D whose elements are the Euclidean distances between the points (Fig.1d); the NxN binary contact map A, where an element equals 1 if there is a contact between the corresponding points, that is, if their distance is smaller than a given threshold ϵ (Fig.1e). Explicit relationships can be established (Fig.1a). Going from **(i)** V to D, **(ii)** D to A and **(iii)** V to G, is straightforward. Two theorems[14] relate **(iv)** G to V and **(v)** D to G, in a way minimizing the sum of squared errors: 3D-coordinates are retrieved by taking the first three eigenvectors of G, and elements of G express as a linear combination of the squared distances. A similar algebraic passage from D to G to V has been proposed in the context of 3C experiments[1]; the above theorems offer a simpler and explicit coordinate reconstruction involving only the diagonalization of G (see Online Methods).

An unsettled intermediary step in MDS-based chromosome reconstruction methods is the derivation of a complete set of distances from a (possibly sparse) contact map, step **(vi)** in Fig.1a. A simple choice is to set the distance between loci i and j equal or proportional to some power α of the inverse[15] contact frequency. However, triangular inequality is not satisfied, and unreliably large distances are assigned to pairs of loci having a low or vanishing contact frequency. An option is to iteratively tune the exponent[11] α and/or a pseudo-count associated to non-contacting pairs. We avoid these time-

consuming optimizations by introducing a graph whose nodes are the *N* loci detected in the experiment. The length of a link is determined as the inverse contact frequency between its end-nodes. We take for the distance between any two nodes the length of the shortest path relating them on the graph, computed using Floyd-Warshall algorithm. Contrary to the inverse-frequency derivation, our method accommodates binary contact maps (e.g. single-cell Hi-C data[7]) by taking link lengths equal to 1 between contacting points, else infinite (no link). Although approximate and giving values up to an overall scaling factor, this shortest-path metric assigns a sound distance (symmetric and satisfying the triangular inequality) to all pairs of points, as required to apply MDS results (Online Methods). It offers a novel way to achieve the preprocessing step common to all 3C-based techniques of converting observed contact frequencies into a complete set of distances, independently of the downstream reconstruction method.

ShRec3D algorithm (Shortest-path Reconstruction in 3D) combines this shortest-path distance with MDS, to achieve chromosome structure reconstruction through consecutive application of procedures (vi,v,iv), Fig.1a.

We tested ShRec3D efficiency in a controlled *in-silico* case. We generated a yeast genome 3D-structure[16] represented as N=26 538 beads (each corresponding to approximately 3 nucleosomes) linked by springs accounting for intra-chromosomal DNA connectivity. The 16 yeast chromosomes were confined into a nucleus of radius 1.6 mm (Fig.1b). From the bead coordinates, we computed a binary contact map (steps (i-ii), Fig.1e). Fig.1f shows the distance matrix obtained by applying step (vi) to this contact map. Procedures (v-iv) reconstruct the coordinates, up to a global transformation (some rotation, dilation and possibly mirror symmetry). Fig.1g displays the distances in this reconstructed structure (inset, Fig. 1h). To quantitatively assess the original structure recovery, we compared in a scatter plot the actual and reconstructed distances, and computed their Spearman rank correlation[11],Fig.1h. A spectral analysis supports the dimensional reduction, step (v) (Fig.S1a).

We compared for various data sizes both the reconstruction accuracy and the speed performance of ShRec3D and two other methods, BACH[6] and ChromSDE[11]. All gave satisfactory results in terms of reconstruction accuracy (Fig.2a). However our script runs on a personal computer from 1 minute for small datasets (~1000 points) to 50 hours for the largest one (26 538 points) overstepping the performances of other methods by several orders of magnitude (Fig.2b). ShRec3D running-time limiting-step is the Floyd-Warshall algorithm computing shortest paths on the contact map, whose worst-case performance scales as $O(N^3)$. We tested MDS reconstruction applied directly to inverse-frequency distances, and its poor accuracy is shown in Figs.2 and S1b-d, demonstrating the importance of using our novel shortest-path metric prior to MDS reconstruction.

We tested and compared ShRec3D to the above-mentioned alternative methods in conditions closer to real Hi-C experiments,. Robustness with respect to experimental noise (mimicked by misplaced contacts, Online Methods) is shown on Fig.2c, demonstrating the good performance of ShRec3D for noise levels lower than 1% (maximal level in typical Hi-C experiments, see Online Methods). The probabilistic nature of BACH[6] makes it efficient in the presence of high levels of noise; however, it remains slower than ShRec3D-reconstruction by several orders of magnitude hence limited to small-size structures (Fig. 2a). We reproduced the Hi-C map feature of being a superposition of single-cell contact maps reflecting the genome fold variations over a cell population; accordingly, only an average 3D-structure can be reached. From a Langevin dynamic simulation[16] of our *in-silico* genome, we extracted a variable number *k* of independent structures and compute the average of their contact maps (Online Methods). The distances reconstructed with ShRec3D from this simulated average Hi-C contact map quantitatively matches the average distances in the superposition of structures, Fig.2d. This is also achieved by the alternative methods for a large number of structures, however the

comparison had to be limited to coarse-grained structures with 480 points, the maximal size manageable in a reasonable time using BACH[11] and ChromSDE[11]. The increase in quality of MDS applied to inverse-frequency distances with the number of structures is expected since the inverse-frequency expression becomes closer to the shortest-path distance when the average contact map becomes denser.

We implemented ShRec3D on experimental Hi-C data obtained in human embryonic stem cells[17] and lymphoblastoids[18] exploiting both the very sparse Hi-C data obtained at the best available genomic resolution (restriction fragments) and coarse-grained datasets (where loci correspond to many restriction fragments). ShRec3D ability to visualize average structures at different scales is illustrated by reconstructing a 30Mbp region of chromosome 1 at 3kb resolution (Fig.3a), the chromosome average structure at 150kb resolution (Fig.3b), and the average arrangement of autosomal chromosomes within nuclear space at 3Mbp resolution (Fig.3c). Genome connectivity and chromosome partitioning achieved by ShRec3D (Figs.3d-f, S2b and S3) would make it an efficient tool for genome scaffolding from Hi-C data[19,20]. Alternative methods (BACH[6], ChromSDE[11]) do not manage fine-resolution data in reasonable time. MDS applied to inverse-frequency distances does not properly reconstruct the fine-resolution structure. The potential of ShRec3D to devise 3D genome browsers is illustrated with the coloring of a 3D-structure of chromosome 1 at resolution 30kb according to the chromatin partition in two compartments[2] (Fig.3g). Two histone-H3 modifications (H3K9Ac and H3K9me3, GSM469974) are added in Fig.3h. Any chemical, structural or functional annotation available on linear genomes can be similarly overlaid on chromosome 3D-structures.

ShRec3D allows a fast and accurate visualization of average chromosome 3D structures from Hi-C datasets. The shortest-path metric used in the preliminary step offers a reliable way for translating contact frequencies into distances, potentially able to improve any analysis involving a 3C-based distance matrix. ShRec3D involves no *ad hoc* constraints and tunable parameters, and is free from convergence issues and misleading transient outcomes. Its speed makes it applicable to both 3C or 5C datasets, which typically involve tens of loci, and high-resolution Hi-C datasets, comprising sparse contacts between hundreds of thousands of points. Its accurate reconstruction of average distances between genomic loci and visualization of a consensus structure provide a meaningful use of cell-population Hi-C data, especially when extended into 3D genome browsers.

METHODS

Methods and associated references are available in the online version of the paper. Programs are freely available at publisher site and https://sites.google.com/site/julienmozziconacci

ACKNOWLEDGMENTS

The authors thank Dustin Arendt for the online-available implementation of the Floyd-Warshall algorithm. They acknowledge funding from UPMC, grant CONVERGENCE2011-projet CVG1110 and from the Institut National du Cancer, grant INCa_5960. UPMC belongs to Sorbonne-Universités,

AUTHOR CONTRIBUTIONS

JM, AL and JR designed the algorithm. JM implemented it. JR, PR, AC and JM tested its validity. AC analyzed experimental datasets. JM and AL wrote the paper.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Affiliations

[1]Laboratoire de Physique Théorique de la Matière Condensée, CNRS UMR 7600, UPMC Univ. Paris 06, F-75005, Paris, France. [2]Institut de Génétique Moléculaire de Montpellier, CNRS UMR 5535, Univ. Montpellier, F-34293, Montpellier, France. [3]Institut Pasteur, F-75015, Paris, France.
*Correspondence should be addressed to J.M. (mozziconacci@lptmc.jussieu.fr)

Fig.1 | Application of our algorithm ShRec3D to a simulated dataset. (a) Algorithm flowchart. **(b)** 3D structure of *in-silico* yeast genome generated using polymer dynamics for a chain of *N =26 538* beads (simulation units, s.u.=10nm, each chromosome in a different color) **(c)** Gram matrix (s.u.) **(d)** Distance matrix (s.u.). **(e)** Contact map (binary map, threshold ϵ=60 nm). (f) Distance matrix derived from contacts (in number of steps). **(g)** Distance matrix of the reconstructed structure (dimensionless) **(h)** Scatter plot of original and reconstructed distances; heat-map colors encode the local density of points; Spearman rank correlation coefficient *R* is indicated; **(inset)** reconstructed 3D structure.

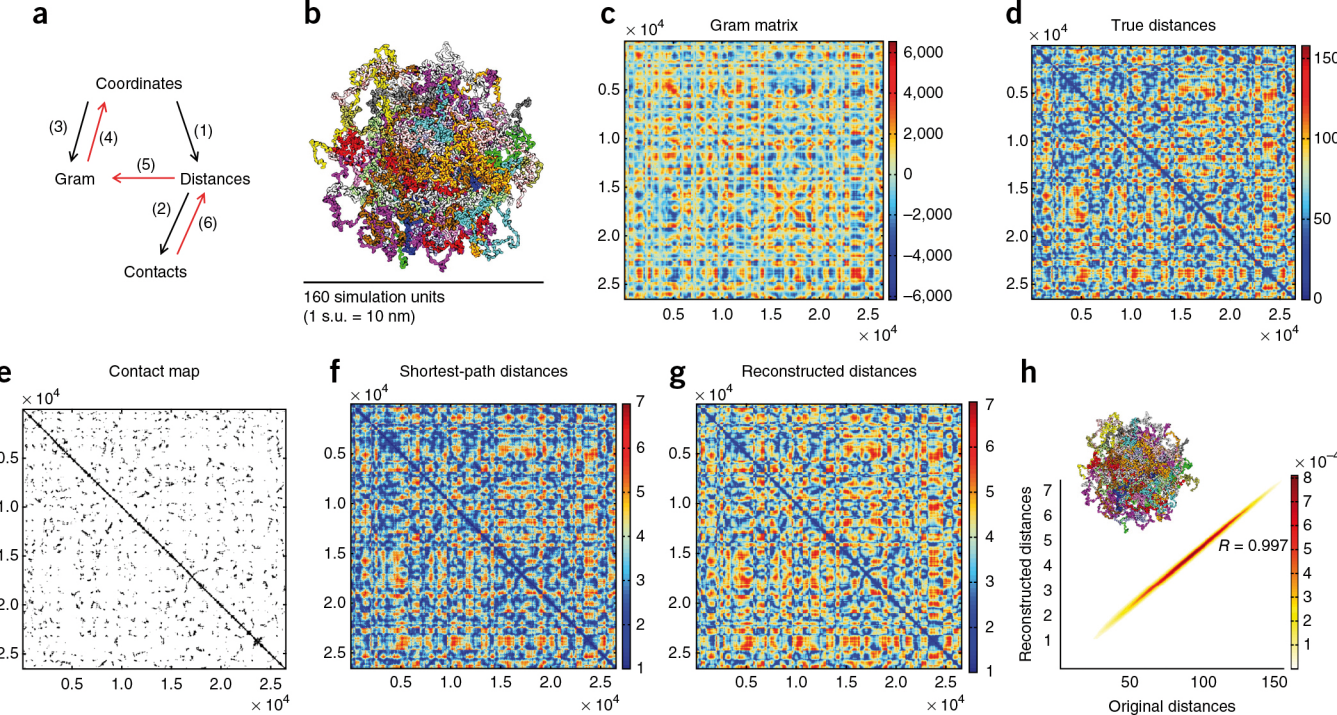
Fig.2 | Quantitative assessment of our algorithm ShRec3D performance and reliability. Comparison of ShRec3D with three alternative methods (BACH[6], ChromSDE[11] and MDS applied to inverse-frequency distances) for simulated data of increasing size *N* in terms of **(a)** reconstruction accuracy (Spearman rank correlation coefficient between original and reconstructed distances) and **(b)** computation time. **(c)** Robustness to a controlled amount of randomly misplaced contacts mimicking experimental noise (semi-log plot) **(d)** Comparison of average distances in a population of an increasing number *k* of simulated structures (up to *k*=500 independent snapshots of a Langevin dynamics, Fig.1, coarse-grained to *N*=480 points) and distances reconstructed from the corresponding average contact map.

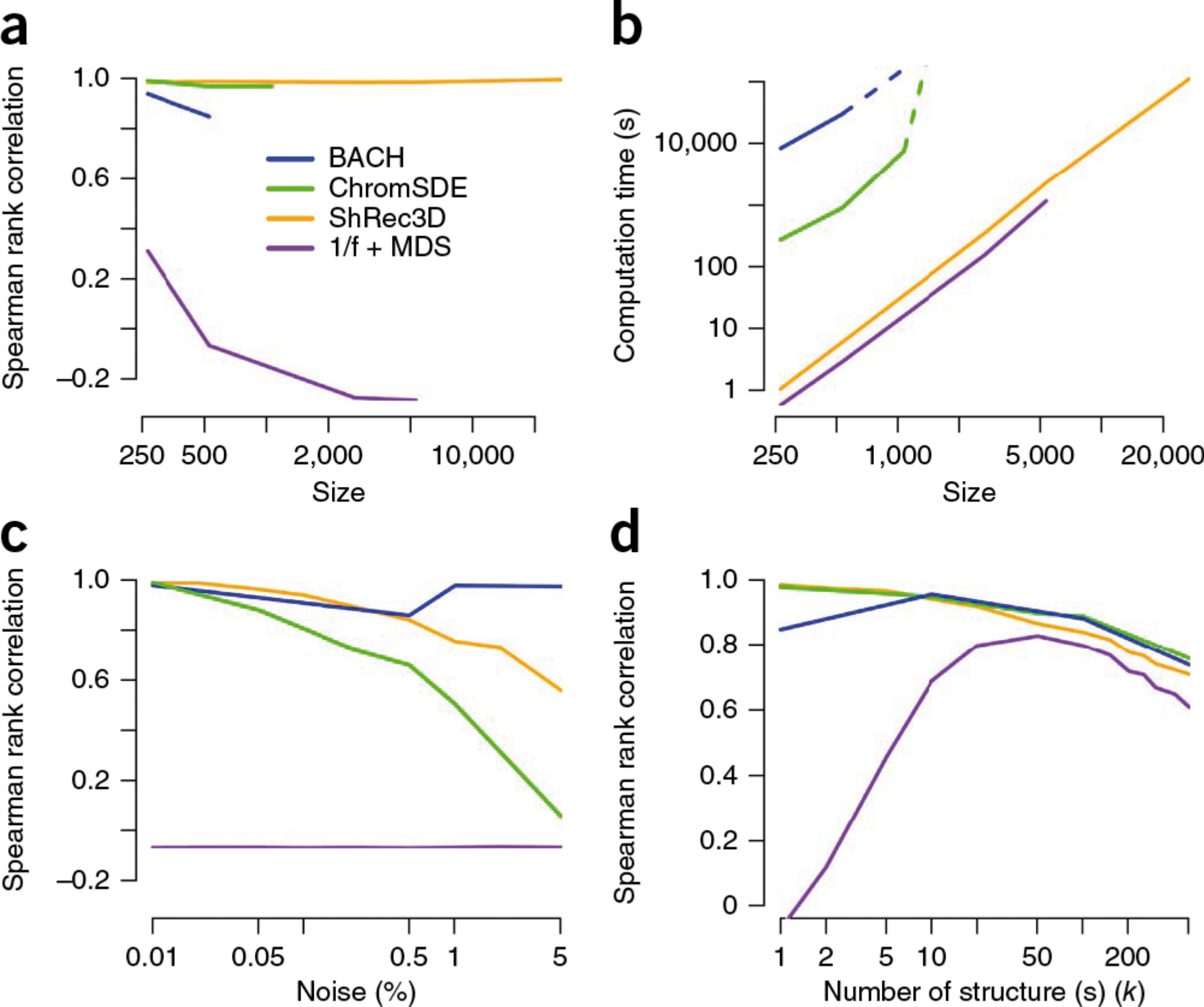
Fig. 3 | **3D multi-scale visualization and annotation of human autosomal chromosomes from Hi-C data.** Experimental contact maps at various genomic resolutions, from **(a)** the scale of restriction fragments in chromosome 1 (embryonic stem cells, hESC[17]) to **(b)** that of bins each containing 50 (hESC,centromere as a black ball) to **(c)** 1000 restriction fragments covering the whole chromosome set (data from [18], see also Fig S3). **(d-f)** Corresponding reconstructed 3D structures. Colors indicate the position along the genome. **(g)** Overlay in hESC[17] of human chromatin partition in two compartments[2] (highlighted by color code on the structure and boxes) **(h)** Additional overlay of acetylation and tri-methylation of lysine 9 in histone H3.

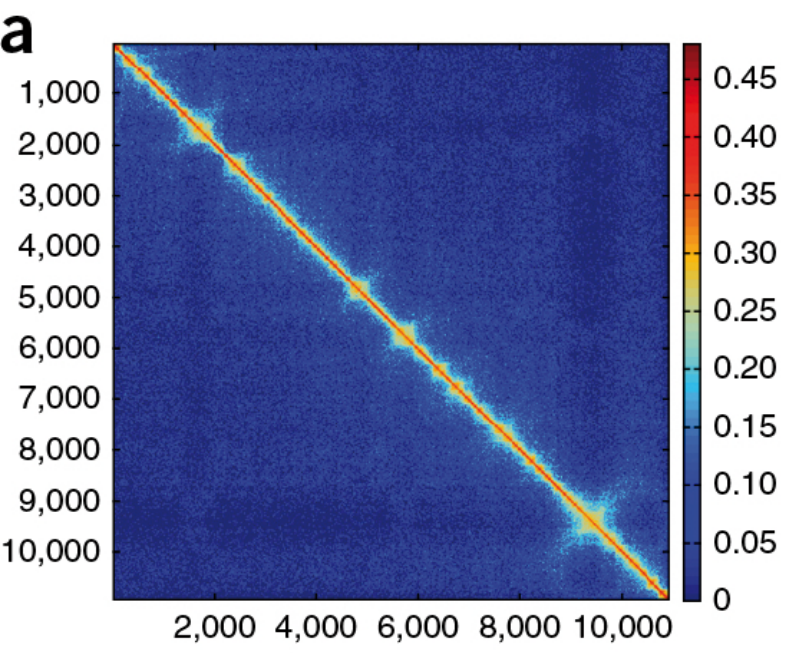
References

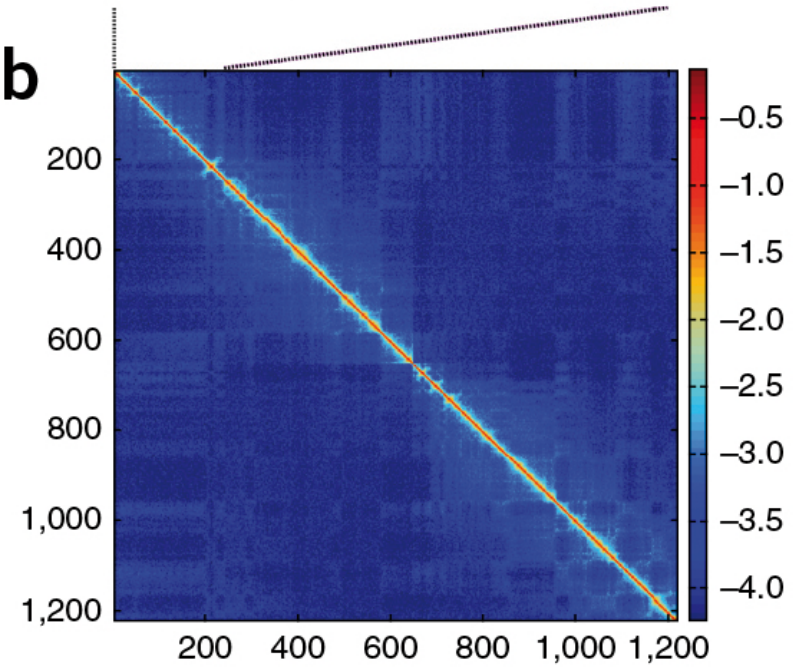
1. Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. *Science* **295**, 1306–1311 (2002).
2. Lieberman-Aiden, E. *et al*. *Science* **326**, 289–293 (2009).
3. Marti-Renom, M.A. & Mirny, L.A. *PLoS Comput. Biol*. **7**, e1002125 (2011).
4. Baù, D. & Marti-Renom, M.A. *Methods* **58**, 300–306 (2012)
5. Duan *et al*. *Nature* **465**,363–367 (2010).
6. Hu, M. *et al*. *PLoS Comput Biol*. **9**, e1002893 (2013).
7. Nagano *et al*. *Nature*, **502**, 59–64 (2013).
8. Rousseau, M. *et al*. *BMC Bioinformatics* **12**, 414 (2011).
9. Trieu, T. & Cheng, J. *Nucl. Acids Res*. **42**, e52 (2014).
10. Varoquaux, N., Ay, F., Noble , W.S. & Vert, J.P. *Bioinformatics* **30**, i26–i33 (2014)
11. Zhang *et al, J. Comput. Biol*. **20**, 831–846 (2013).
12. Sippl, M.J. & Scheraga, H.A. *Proc. Natl. Acad. Sci. USA* **82**, 2197–2201 (1985)

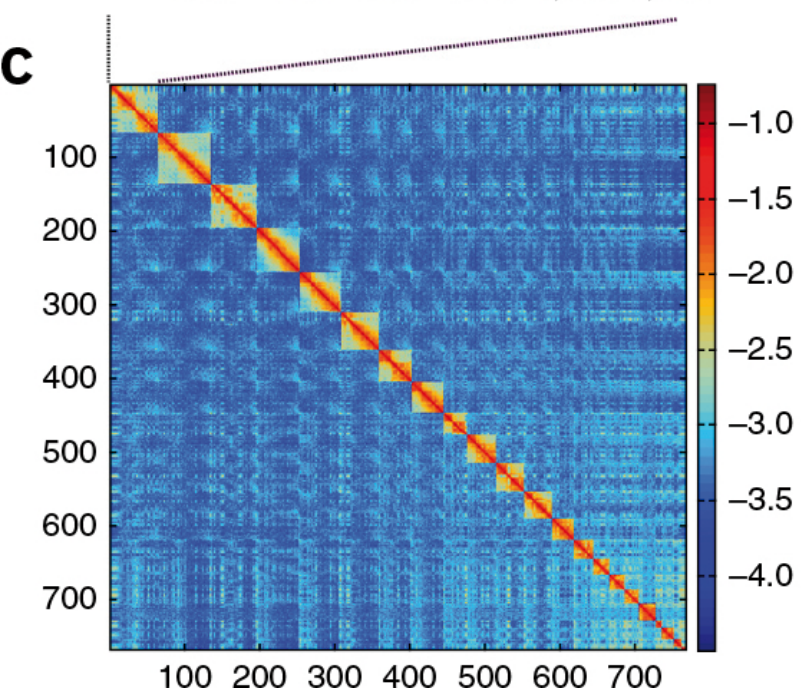
13.Torgerson, W.S. *Psychometrika* **17**, 401–419 (1952)
14. Havel, T.F., Kuntz, I. & Crippen, G.M. *Bull. Math. Biol.*, **45**, 665–720 (1983).
15. Fraser, J. *et al. Genome. Biol.* **10**, R37 (2009).
16. Hajjoul, H. *et al. Genome Res.* **23**, 1829–1838 (2013).
17. Dixon, J.R. *et al. Nature* **485**, 376–380 (2012).
18. Kalhor, R. *et al. Nat. Biotechnol.* **30**, 90–98 (2011).
19. Burton, J.N. *et al. Nat Biotechnol.* **31**, 1119–1125 (2013).
20. Kaplan, N. & Dekker, J. *Nat Biotechnol.* **31**, 1143–1147 (2013).

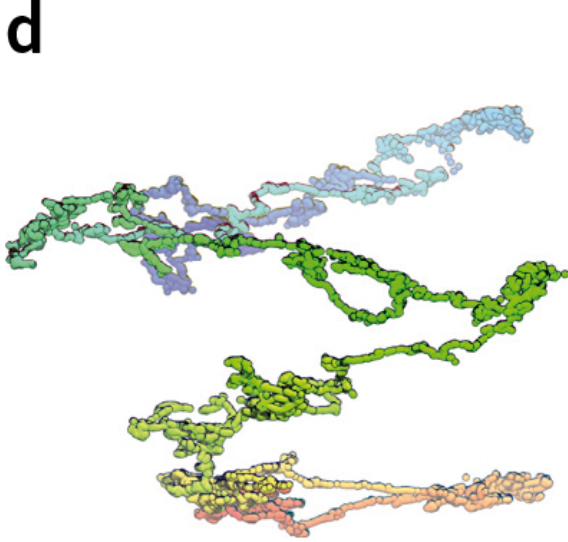

a
Coordinates
(1)
(2)
(3)
(4)
(5)
(6)
Gram
Distances
Contacts
b
160 simulation units
(1 s.u. = 10 nm)
c
Gram matrix
× 10^4
6,000
4,000
2,000
0
−2,000
−4,000
−6,000
d
True distances
150
100
50
0
e
Contact map
f
Shortest-path distances
g
Reconstructed distances
h
R = 0.997
Reconstructed distances
Original distances
× 10^−4

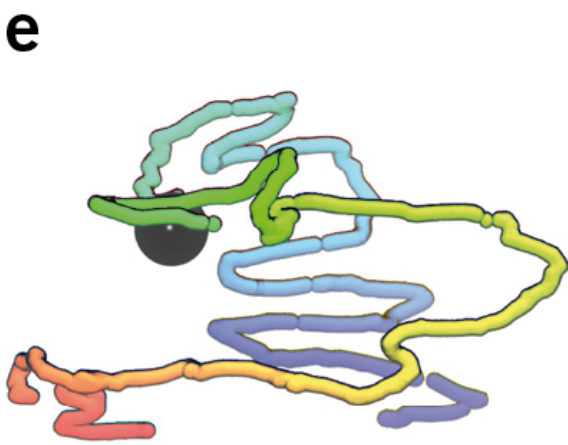

a
Spearman rank correlation
1.0
0.6
0.2
–0.2
BACH
ChromSDE
ShRec3D
1/f + MDS
250
500
2,000
10,000
Size
b
Computation time (s)
10,000
100
10
1
250
1,000
5,000
20,000
Size
c
Spearman rank correlation
1.0
0.6
0.2
–0.2
0.01
0.05
0.5
1
5
Noise (%)
d
Spearman rank correlation
1.0
0.8
0.6
0.4
0.2
0
1
2
5
10
50
200
Number of structure (s) (k)


a
1,000
2,000
3,000
4,000
5,000
6,000
7,000
8,000
9,000
10,000
2,000 4,000 6,000 8,000 10,000
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0


b
200
400
600
800
1,000
1,200
200 400 600 800 1,000 1,200
–0.5
–1.0
–1.5
–2.0
–2.5
–3.0
–3.5
–4.0


c
100
200
300
400
500
600
700
100 200 300 400 500 600 700
–1.0
–1.5
–2.0
–2.5
–3.0
–3.5
–4.0


d


e


f

SUPPLEMENTARY INFORMATION

about "A fast algorithm to reconstruct 3D genome structures from chromosomal contact maps"

by Annick Lesne[1,2], Julien Riposo[1], Paul Roger[1], Axel Cournac[3] & Julien Mozziconacci[1]

(1) Laboratoire de Physique Théorique de la Matière Condensée, CNRS UMR 7600, UPMC Univ. Paris 06, F-75005, Paris, France.

(2) Institut de Génétique Moléculaire de Montpellier, CNRS UMR 5535, Univ. Montpellier, F-34293, Montpellier, France.

(3) Institut Pasteur, F-75015, Paris, France.

CONTENTS

SUPPLEMENTARY FIGURES

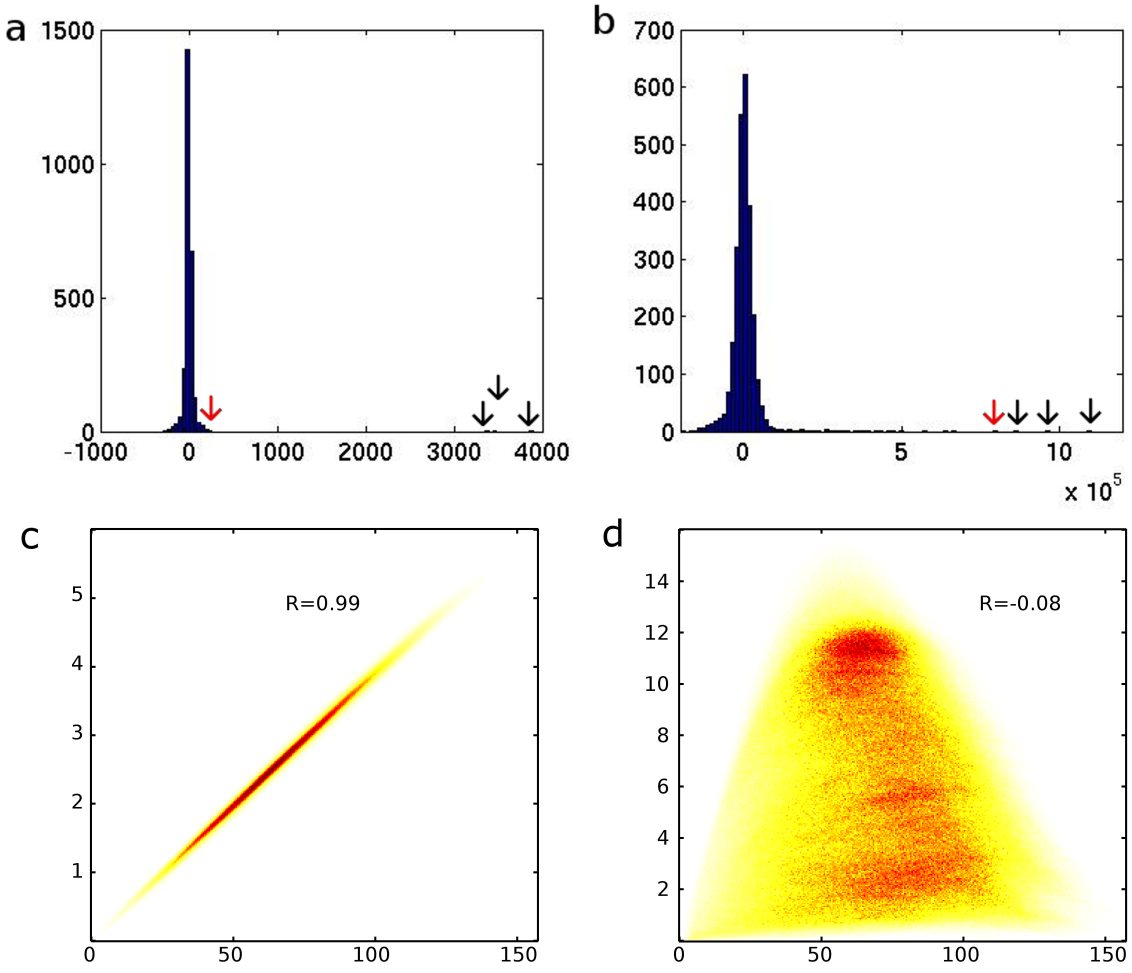


Fig. S1 | Comparison of ShRec3D with the simple MDS-inverse-frequency approach. The top panels displays the spectrum (eigenvalue histogram) of the metric matrix, Eq. 3, derived from our simulated benchmark (with here $N = 2600$) using (a) our shortest-path distance and (b) the simple distance equal to the inverse contact frequency (both distances are dimensionless: the units on the abscissa axis depend on the chosen normalization for the contact frequencies). The three rightmost black arrows underline the first three eigenvalues and the leftmost red arrow underlines the fourth one, demonstrating the presence of a significant spectral gap in (a) and the absence of spectral gap in (b). The bottom panels present a scatter plot (high densities in red) of the original distances in our simulated benchmark (horizontal axis, simulation unit equal to 10 nm, $N = 2600$ points) and the distances reconstructed from the corresponding contact map, using as a preliminary step either (c) our shortest-path distance or (d) distances obtained as the inverse contact frequencies $1/f$, followed in both cases by the MDS procedure described in Fig. 1(a), steps v-vi (vertical axis, dimensionless distances). Spearman rank correlation coefficient R is indicated in inset.

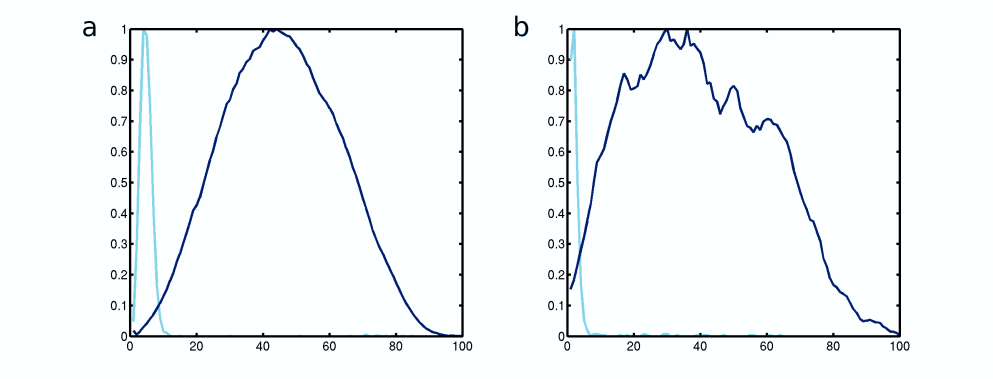


Fig. S2 | Polymer connectivity in ShRec3D reconstruction. Normalized histogram of the reconstructed distances $D_{i,i+1}$ between neighbors along the genome (narrowly peaked light blue curves), for (a) our simulated benchmark (here $N = 2600$ points) and (b) genome-wide real Hi-C data (Kalhor et al. 2011) compared to the normalized histogram of all distances taken as a reference (widespread dark blue curves).

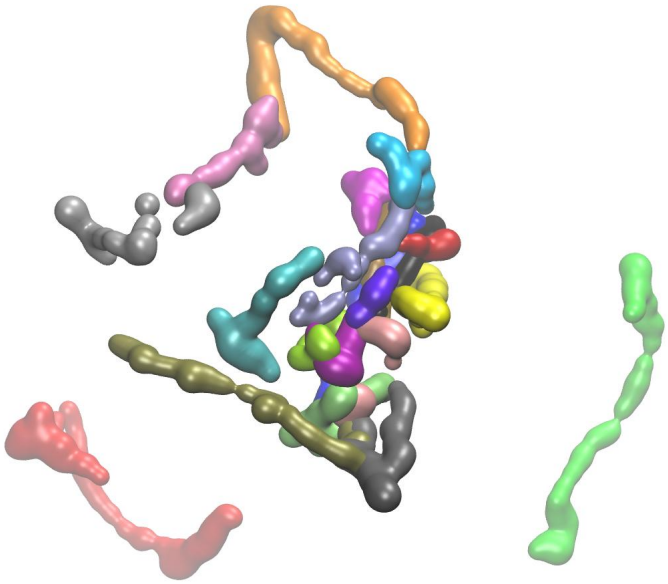


Fig. S3 | Visualization of human autosomal chromosomes using ShRec3D (Hi-C data in lymphoblastoid cells, Kalhor *et al.* 2011). Color labeling of the different chromosomes: 1: blue, 2: red, 3: grey, 4: orange, 5: yellow, 6: gold, 7: silver, 8: green, 9: pink, 10: cyan, 11: purple, 12: lime, 13: mauve, 14: ochre, 15: iceblue, 16: black, 17: lightgreen, 18: light cyan, 19: violet ,20: magenta, 21: dark red, 22: light orange.

ONLINE METHODS

Matrix definitions: coordinate, Gram, distance and contact matrices

Various matrices can be associated with a structure comprising N points P_i ($i = 1, \ldots, N$) in a n-dimensional space, having here in mind application to experimental structures with $n = 3$. The origin O of the coordinate system (point with null coordinates) is taken to be the barycenter of the set of points; barycentric coordinates are indeed geometrically more suitable for structure visualization. The coordinate matrix V is a $n \times N$ matrix comprising the Euclidean coordinates of the points, namely the element $V_{\alpha i}$ is the coordinate of the point P_i along the α-axis ($\alpha = 1, \ldots, n$-. The Gram matrix G is a $N \times N$ positive semi-definite matrix whose element G_{ij} is the scalar product of the coordinate vectors associated with points P_i and P_j. The distance matrix D is a $N \times N$ matrix whose element D_{ij} is the Euclidean distance between the points P_i and P_j. A binary contact matrix A can be defined for a given threshold ε: its element A_{ij} equals 1 if the distance D_{ij} between the points P_i and P_j is smaller than ε, else it equals 0.

In practice, a contact is said to occur between two genomic loci P_i and P_j if their distance within the cell nucleus is smaller than a given threshold ε, prescribed by the experimental technique (cross-linking step in chromosome conformation capture experiments) and its sensitivity. The experimental results are expressed either in a binary way (presence or absence of a contact) which yields a binary contact map (typically the case for single-cell Hi-C experiments), or in terms of contact counts c_{ij} (typically the case for Hi-C experiments performed in a cell population) then normalized into contact frequencies f_{ij} during the data processing, see Cournac *et al.* (2012).

Explicit and reciprocal relationships, Fig. 1(a), can be established between the above matrix representations of a N-point structure. As detailed below, steps (i), (ii) and (iii) are straightforward. They will be used in the generation of benchmark *in-silico* Hi-C data, starting from simulated chromosome structures. Our algorithm ShRec3D (for *Shortest-path 3D Reconstruction* involves first the translation of a contact map into a distance matrix using a graph-theoretic method (step (vi) in Fig. 1(a), see below section "Shortest-path method"), then the reconstruction of a 3D-structure using standard results from distance geometry and classical multidimensional scaling (MDS, steps (iv-v) in Fig. 1(a), see below section "Multidimensional scaling").

• *Step (i) from coordinates to distances (V to D)* — The Euclidean distance between the points P_i and P_j straightforwardly writes as a function of the coordinates, i.e. $D_{ij} = \sqrt{\sum_{\alpha=1}^{n} |V_{\alpha i} - V_{\alpha j}|^2}$.

• *Step (ii) from distances to contacts (D to A)* — Given a threshold ε, a binary contact matrix is obtained by setting to 1 the elements A_{ij} such that D_{ij} is smaller than ε and the others to 0.

• *Step (iii) from coordinates to Gram matrix (V to G)* — The Gram matrix of the set of points is obtained by computing the scalar product of their coordinate vectors (columns of V): $G = V^T V$ where V^T is the transpose of V, i.e. $G_{ij} = \sum_{\alpha=1}^{n} V_{\alpha i} V_{\alpha j}$

Multidimensional scaling: from distance matrix to 3D structure (steps iv-v, Fig. 1(a))

Distance geometry has been developed to solve the issue of recovering coordinates from the sole knowledge of distances (see e.g. Schoenberg (1935), Havel, Kuntz & Crippen (1983), Sippl & Scheraga (1985) and references therein). Multidimensional scaling brings in another notion, here central, of dimensional reduction: given a dimension n (currently small, $n = 3$ in our case), find the n-dimensional structure optimally approximating a given distance matrix. This goal is often achieved by minimizing a cost function involving the (possibly weighted) differences between the given features and the reconstructed coordinates. To avoid a time consuming optimization, we instead followed the original line of multidimensional scaling, today called "classical MDS" (see e.g.

the seminal papers by Youth and Householder (1938) and Torgerson (1952)); as explained below, it is based on algebra and explicit (analytical) formulas. (It is to note that the scope of MDS rapidly extended beyond distance matrices, to the treatment of ordinal information ("nonmetric MDS"), where the involvement of a cost function is then mandatory, see e.g. the handbook by Kruskal & Wish (1978) for an overview and Varoquaux *et al.* 2014 for a recent application to chromosome reconstruction.)

• *Step (iv) from Gram matrix to coordinates (G to V)* — One of the most powerful theorems of distance geometry states that, provided the Gram matrix is positive semi-definite, the coordinates of the N points P_i ($i = 1, \ldots, N$) in a n-dimensional space can be recovered from the first n eigenvectors E_α ($\alpha = 1, \ldots, n$) of the Gram matrix, normalized to 1 then rescaled by the square root of their associated eigenvalue λ_α, namely:

$$V_{\alpha i} = E_\alpha(i) \times \sqrt{\lambda_\alpha} \quad \text{with} \quad \sum_{i=1}^{N} E_\alpha(i)^2 = 1, \tag{1}$$

where $E_\alpha(i)$ is the i-th component of the eigenvector E_α and $V_{\alpha i}$ is defined above (α-coordinate of the point P_i). We refer to Havel, Kuntz & Crippen (1983), where this result is clearly presented and demonstrated in an accessible way (part of Theorem 3.1); however this result is older and progressively established during the parallel development of distance geometry and multidimensional scaling. The rank of the Gram matrix G determines the minimum embedding dimension n. Geometrically, the first n eigenvectors of G are the principal axes of the N-point structure and the eigenvalues are the corresponding moments. It is to note that an alternative way for passing from a Gram matrix G to the coordinates V, relying on Cholesky decomposition of G (namely, writing $G = LL^T$ where L is a lower triangular matrix with real positive or vanishing diagonal elements) is mentioned in Dekker *et al.* (2002), footnote 40. However, the Cholesky decomposition does not exist if G has vanishing diagonal elements, or is unstable if these elements are small (Sippl and Scheraga, 1985). The above constructive result, although basically similar, is conceptually and practically simpler since it involves only the diagonalization of G.

Importantly, if the rank of the Gram matrix G is larger than the desired dimension n, the above formula give the coordinates of the n-dimensional structure best approximating the underlying one (see e.g. Havel, Kuntz & Crippen (1983)).

• *Step (v) from distances to Gram matrix (D to G)* — The mathematical derivation of a Gram matrix from the knowledge of distances is presented in theorems 3.1 and 3.3 in Havel, Kuntz & Crippen (1983). We reformulate these theorems in a form more tractable for our algorithmic purposes. The first step is to express for any $i = 1, \ldots, N$ the distance d_{0i} between the barycenter O and the point P_i as a function of the elements of D:

$$d_{0i}^2 = \frac{1}{N} \sum_{j=1}^{N} D_{ij}^2 - \frac{1}{N^2} \sum_{j=1}^{N} \sum_{k>j}^{N} D_{jk}^2. \tag{2}$$

The argument introduces an auxiliary matrix, the metric matrix M, with elements:

$$M_{ij} = \frac{1}{2} \left[d_{0i}^2 + d_{0j}^2 - D_{ij}^2 \right]. \tag{3}$$

It is shown in Schoenberg (1935) that the condition that D is a distance matrix associated with a N-point Euclidean structure in a space of dimension n, is that M be semi-definite positive of rank n. Then M coincides with the Gram matrix G of the N points and (iv) applies.

In practice, small errors in the distances may cause some eigenvalues of M to be negative (though small). This means that an exact 3-dimensional Euclidean embedding does not exist.

This problem has been mentioned in the context of 3C data by Dekker *et al.* (2002), footnote 40, and solved by replacing all but the largest three eigenvalues of M by 0. It is a central result of MDS that this procedure yields the best 3-dimensional Euclidean approximation of the original matrix. Choosing the barycenter (center of mass) of the N points to be the reference point O in the matrix M minimizes the ensuing approximation (Havel, Kuntz & Crippen 1983). Importantly, this reconstruction, step (v), requires to known the full set of distances, i.e. to know D_{ij} for any pair (i, j). This requires a preprocessing step of the contact map, converting contact frequencies into distances. Filling this gap is an important advance provided by the next step (vi), section "Shortest-path method" below.

In conclusion, MDS reconstruction truncates the metric matrix M, Eq. 3, into the rank-3 Gram matrix G of what will be the reconstructed 3D-structure by considering only the dominant three eigenvalues and associated eigenvectors, which yields the optimal 3D approximation of the original structure (i.e. the approximation minimizing the sum of squared errors). This dimensional reduction (from M to G) and subsequent step of coordinate reconstruction (from G to V) are all the more valid when the neglected eigenvalues are small, and the more separated by a large spectral gap from the three retained eigenvalues. Such a check is performed in a typical instance in Fig. S1(a).

Shortest-path method: from contact map to distance matrix (step (vi), Fig. 1(a))

We introduce a novel way to derive the full set of distances from the knowledge of a (possibly sparse) contact map, using the concept of shortest path in graph theory. In the case of a single underlying structure, the graph is defined by the binary contact map A (presence/absence of a contact) seen as its adjacency matrix. This graph, whose nodes are the points P_i $(i = 1, \ldots, N)$ has to be connected, since it would not be possible to assign a distance between points belonging to two distinct components. Connectedness means that for any pair of points P_i and P_j, one can find a path $(i_0, i_1, \ldots, i_k)$ with $i = i_0$ and $j = i_k$, such that $A_{i_q i_{q+1}} = 1$ or in practice, find a strictly positive integer k such that $A^k_{ij} > 0$. In mathematical terms, it means that A has to be irreducible (which is the case in the experimental situations considered here). The path with the minimal number of steps is termed the shortest path between the points P_i and P_j.

In the case of Hi-C experiments, performed over a cell population hence an accumulation of structures, each pair (i, j) of nodes is associated with a normalized contact frequency f_{ij}, however possibly vanishing or very small. A current method, henceforth termed inverse-frequency method, is to assign the value $1/f_{ij}$ to the distance between points P_i and P_j (see e.g. Fraser *et al.* (2009). Vanishing contact frequencies are replaced by pseudo-counts to avoid ill-defined (infinite) distance values, which introduces the arbitrary choice of a pseudo-count value in the distance matrix derivation. This simple method anyhow gives unreliable or even meaningless distance values for small f_{ij}. Moreover, it does not define a true distance since it does not satisfy triangular inequality. To circumvent these shortcomings, we considered a weighted adjacency matrix, where the link (i, j) between nodes i and j is endowed with a length equal to the value $1/f_{ij}$. The shortest path between P_i and P_j is now a path $(i_0, i_1, \ldots, i_k)$ with $i = i_0$ and $j = i_k$, such that the length $\sum_{q=0}^{k} 1/f_{i_q i_{q+1}}$ is minimal over the paths relating P_i and P_j. While the shortest path is not necessarily unique, its length takes a unique value. We propose to define the distance between two points by the length of shortest path relating them. Note that other choices are possible for relating link length and contact frequency, e.g. $1/f^\alpha$. Basically, the exponent modify the relative weight of rare contacts (decreasing for $\alpha > 1$, increasing for $\alpha < 1$). The value of α appeared to have little effect on the reconstruction quality, due to the fact that low contact frequencies do not contribute to our shortest-path distance, hence we kept the original choice $\alpha = 1$.

We use the acknowledged Floyd-Warshall algorithm for computing shortest paths and their lengths. Interestingly, by construction, weak or vanishing contact frequencies do not contribute to the distances since the shortest paths will bypass the corresponding links (of large or infinite

lengths). This method thus makes possible to both reconstruct the whole set of distances and filter some of the experimental noise (low contact frequencies that may correspond to noise are not used). Importantly, this method defines a true distance. Firstly, it is obviously symmetric and vanishes only if the points are identical. Secondly, by construction, the minimal path length to go from node i to node j is always smaller or equal to the sum of the minimal path length to go from node i to some node k and the minimal path length to go from node k to node j. Accordingly, the shortest-path distance satisfies the triangular inequality (with equality when a shortest path from i to j passes through k).

For pairs of loci with high contact frequency, our shortest-path distance recovers the simple inverse-frequency expression of Fraser *et al.* (2009). On the contrary, the shortest-path method improves the distance assigned to pairs of loci with low or vanishing contact frequencies. It is also beneficial in case of a binary contact map (e.g. for single-cell Hi-C data, Nagano et al. 2013), where the inverse-frequency method yields distance values either equal to 1 or infinite. Finally, the distance between neighboring loci along the genome will satisfactorily be small, since the closer the loci are along the genome, the more they establish contacts; this distance is thus consistent with the polymer-like connectedness of each chromosome (see also Fig. S2 below).

Preparation of the *in-silico* yeast genome structure

The *in-silico* yeast genome structure used to test our reconstruction algorithm has been generated using a simple polymer model (with excluded volume) for a chain of $N = 26\,538$ beads, each corresponding approximately to 3 nucleosomes. The chain is confined in a spherical nucleus of radius 1.6 μm. The simulation spatial unit corresponds to 10nm in a real nucleus. The barycentric coordinates are taken from a snapshot of a simulated Langevin dynamics simulation after it has reached thermal equilibrium. The binary contact map mimicking single-cell Hi-C data has been obtained using steps (i, ii) above. Considering a regularly spaced sampling of the N points allows to investigate different data sizes. In particular, such a coarse-graining is mandatory in the comparison with alternative methods (see below) which cannot handle sizes as large as $N = 26\,538$. Finally, we used also this Langevin simulation to generate an ensemble of structures.

Implementation of the algorithm ShRec3D

By supplementing our shortest-path distance derivation, step (vi), with MDS reconstruction, steps (v) and (iv), we obtained a constructive algorithm, that we named ShRec3D (for Shortest-path 3D Reconstruction). It allows to visualize a 3D structure from the knowledge of any contact map (either binary, in terms of contact presence/absence, or quantitative, in terms of contact frequencies). Overall, the 3D coordinates are reconstructed up to an arbitrary rotation, dilation and possibly mirror symmetry. The algorithm presented above has been written with MATLAB (http://www.mathworks.fr/products/matlab/).

As mentioned above, the coordinate reconstruction involves only the first three eigenvectors of the metric matrix M, Eq. 3, as if the other eigenvalues were vanishing (an approximation also proposed in Dekker *et al.* (2002) and quantitatively assessed –among others– in Havel, Kuntz & Crippen (1983)). The validity of this eigenvalue truncation, approximating M by a positive semi-definite matrix G of rank 3, is assessed by investigating the spectrum of M and checking that the largest three eigenvalues are separated by a large spectral gap from the remaining spectrum concentrated near 0.

We performed this spectral check for the above-described simulated benchmark data (presented in Fig. 1). Fig. S1(a,b) displays the comparison between the spectrum of the metric matrix obtained when using our shortest-path metric and that obtained using the simple inverse-frequency distance, both followed by the same MDS reconstruction. The presence of small and partly neg-

ative eigenvalues originates from the inaccuracies in the distance matrix derived from the data, propagated to the metric matrix. Indeed, while the shortest-path distance is a true metric (satisfying the triangular inequality), the data from which the distance matrix is derived have been discretized in the form of binary contact map (presence/absence). Due to this loss of information in the generation of our synthetic data, the reconstructed metric matrix slightly differs from that of the original structure.

Implementation of available alternative methods

We compared the performance of our algorithm ShRec3D with alternative reconstruction methods as regards both reconstruction accuracy and speed (see below). We limited ourselves to methods for which the original codes, optimized by their authors, were available, so as to make a fair comparison, namely BACH software (Hu *et al.* 2013) and ChromSDE (Zhang*et al.* 2013). They all involve an optimization procedure, and yield a single consensus structure from Hi-C contact maps. We ran the softwares on the same Linux machine (or a similar machine when the code has to be run on Windows software). We run BACH for 5000 iteration steps (default value). We implemented ChromSDE in its linear mode. Comparison on real data is presently limited since these alternative methods are unable to deal with full-size data sets; we have thus favored the use of simulated benchmark data, for which the underlying structure is known.

Comparison of our shortest-path distance with the inverse-frequency distance

We compared the respective performances of our shortest-path distance and the simple derivation where the distance between two genomic site is set equal to their inverse contact frequency $1/f$ (Fraser *et al.* 2009), both completed by the MDS 3D-reconstruction described above. While the inverse-frequency method corresponds to a faster procedure, it yields a poor reconstruction (whatever the choice of the pseudo-count value, here chosen equal to a given fraction of the average contact frequency, that is, the total number of contacts divided by N^2 where N is the number of considered genomic loci). A first comparison has been presented above, Fig. S1(a,b). It shows that the dimensional reduction step in MDS reconstruction is not legitimate when the preprocessing step (vi) converting the contact map into distances uses the simple inverse-frequency method. We performed additional tests to assess the improvement brought by our shortest-path distance. Fig. S1(c,d) displays a scatter plot of the reconstructed *vs* original matrix distances and their Spearman correlation coefficient R for both methods, applied to the same simulated benchmark as in Figs. 1 and S1(a,b). The poor performance of the simple method presumably originates from the fact that is is not a true distance (it does not satisfy the triangular inequality), and gives an important weight to less reliable low-frequency contacts. Accordingly, some points are placed spuriously far from the core of the structure by the simple method, and the anti-correlation observed in Fig. S1(d) could be even stronger when larger structure sizes are considered. Note that an alternative improvement of the simple inverse-frequency method has been to consider that distances are proportional to $1/f^{\alpha}$, and iteratively optimize the value of the exponent α for each dataset and each description scale (Zhang *et al.* 2013). While this method is satisfactory in terms of reconstruction quality, the inherent optimization procedure makes it very costly in computation time, as seen on Fig. 2(b).

Performance of our ShRec3D algorithm in terms of computation time

The computation time (in seconds) using our method and alternative methods has been plotted, Fig. 2(b), as a function of the data size N (number of points in the structure to be reconstructed). This time scales roughly as $O(N^3)$ using both the BACH software (Hu *et al.* 2013) and our method whereas it does not scales as any power of N for ChromSDE (Zhang *et al.* 2013). For a size of 5000 points, our algorithm runs in about 15 minutes, whereas it takes 5 days to converge for BACH.

For such a size, the ChromSDE software did not converge at all. MDS reconstruction applied to inverse-frequency distances is obviously faster than ShRec3D, since it skips the Floyd-Warshall computation of shortest path.

Quantitative quality assessment of the reconstruction

The quality assessment of our reconstruction algorithm has been done on synthetic data, obtained from simulated 3D structures as described above (real data cannot be used as a benchmark since the underlying 3D structures are unknown). Direct comparison of coordinates would require a preliminary alignment of the coordinates, including the possible mirror symmetry between the original and reconstructed structures, and a global rescaling of the dimensionless reconstructed distances. We thus favored a comparison of the original and the reconstructed distances in terms of their Spearman rank correlation coefficient R (used e.g. in Zhang *et al.* (2013) for the same purpose) which avoids both alignment and rescaling issues. It is satisfactorily equal to 1 for identical distance matrices and to 0 between a matrix and a random shuffle of its elements. We also checked, Fig. S2(a), that our reconstruction ShRec3D preserves polymer connectivity, that is, neighboring loci along the genome are also close neighbors in the 3D space.

Robustness of the reconstruction with respect to experimental noise

Hi-C experiments are intrinsically flawed by spurious re-ligations (i.e. re-ligations occurring between different cross-linked complexes, instead of occurring only within cross-linked complexes), falsely interpreted as contacts. To mimic the effect of this noise in our *in-silico* benchmark, we modified the original binary contact map and generated a controlled amount of disorder by moving randomly a given fraction of positive entries. So doing, the total number of contacts is largely preserved (the moves displacing a positive entry towards a yet positive entry are very scarce for realistic contact densities). In the experiment, the noise strength is controlled by the dilution of the cross-linked complexes, which is limited by the required concentration of the enzyme used for DNA re-ligation; the total number of detected contacts depends on both the ligase concentration and the concentration of cross-linking factor. One way to quantify the noise in Hi-C experiment is to estimate the proportion of random ligation events in the bank. One can for instance use the fact that organisms like yeast *S. Cerevisiae* has mitochondria which lay outside the nucleus hence cannot make contacts as detected in the cross-linking step of 3C techniques. We calculated the proportion of ligations between loci from the main genome and loci from mitochondria as a minimum estimation of the random ligation content. From our experiments, we found this proportion typically smaller than 1%. Such an estimate can also be derived from the newly developed analysis on metagenomic samples, Burton *et al.* 2014).

We investigated whether the reconstruction accuracy is affected by the presence of a fraction of misplaced contacts ranging from 0, 01% (no noise) to 5% (above the upper bound on the experimental noise strength). As in the noiseless case, the quantitative assessment has been done by plotting the Spearman rank correlation coefficient) between the original distance matrix and the distances in the structure reconstructed from the noisy contact map as a function of noise strength, Fig. 2(c). For comparison, we also implemented alternative methods (BACH software by Hu *et al.* (2013) and ChromSDE by Zhang *et al.* (2013)) in the same noisy conditions, and presented the results in the same panel, Fig. 2(c). The improvement by structural disorder of the convergence of BACH software in the considered instance can be explained by the fact that noise prevents trapping in local optima (see e.g. Franzke & Kosko 2011). MDS applied to inverse-frequency distances yields a uniformly poor result (R close to 0, see Fig. S1(d)).

Accuracy of the average structure reconstructed from a superposition of contact maps (Hi-C experiments)

Hi-C experiments are performed over a cell population, hence the experimentally obtained contact maps are in fact an accumulation of individual contact maps (or an average after normalization) each corresponding to an individual cell. We mimicked an Hi-C experiment by considering the superposition of a variable number of simulated structures, reproducing the different chromosome structures present in the cell population. The structures were simulated as above, with a proper tuning of the parameter dynamics to throughly explore the conformation space. Up to 500 snapshots, separated by a sufficiently long run of the dynamical simulation, were extracted. They altogether yield a realistic Hi-C contact map (using procedures (i-ii) above). We evaluated the accuracy of our treatment of Hi-C data by comparing the distance matrix reconstructed from the average contact map and the mean (over the different structures) of the actual distances. Results are shown on Fig. 2(d), where we also present the performance of alternative methods. Due to the prohibitive running time of BACH and ChromSDE for large structures, we there considered coarse-grained structures with only $N = 480$ points.

Normalization and representation of real Hi-C datasets

The procedure we used to normalize the data has been presented in (Cournac *et al.* 2012). The resulting contact maps display relative contact frequencies between genomic loci normalized in such a way that the sum of the contact frequencies for each fragment is equal to one. The color code used in the maps, Fig. 3, depends on the genomic resolution and associated contact density. At the finest resolution (restriction fragments, Fig. 3(a), the contact map is very sparse and the color bar is graded with respect to the contact frequency to the power 0.3, in order to increase the contrast. At lower resolutions the number of contacts are computed between groups ("bins") of respectively 50 and 1000 restriction fragments for the middle and bottom representation. For these two resolutions (Figs. 3(b,c)), the color bar is established in log scale. We checked, Fig. S2(b), that polymer connectivity is preserved by our reconstruction (namely neighboring loci along the genome are also spatial neighbors in the 3D-space) by computing the normalized histogram of the reconstructed distances $D_{i,i+1}$ between neighbors along the genome, compared to the normalized histogram of all reconstructed distances. As shown on Fig. S3, it is possible to label the chromosomes, i.e. to distinguish to which chromosome each genomic locus belongs to. We normalized independently intra- and inter-chromosomal contacts and then added the two normalized matrices in order to reconstruct at the same time the chromosome folding patterns and their relative orientation within the cell nucleus.

3D genome browsers can be obtained by superimposing existing chemical, structural or chemical annotations onto our reconstructed chromosome structures. We illustrated this approach using for instance (Fig. 3(g)) the partition of human chromatin structure in two compartments inferred from the spectral analysis of the correlation matrix between the lines of the contact map (Lieberman-Aiden *et al.* 2009) and the linear profiles of acetylation and trimethylation of lysine 9 of histone H3 (GSM469974).

References

[1] Burton, J.N., Liachko, I., Dunham, M.J. & Shendure, J. Species-level deconvolution of metagenome assemblies with Hi-C-based contact probability maps, *G3 (Bethesda)*, published online May 2014, doi: 10.1534/g3.114.011825.

[2] Cournac, A., Marie-Nelly, H., Marbouty, M., Koszul, R. & Mozziconacci, J. Normalization of a chromosomal contact map. *BMC Genomics* **13**, 436 (2012).

[3] Dekker, J., Rippe K., Dekker, M. & Kleckner, N. Capturing chromosome conformation, *Science* **295**, 1306-1311 (2002).

[4] Franzke, B. & Kosko, B. Noise can speed convergence in Markov chains, *Phys. Rev. E* **84**, 041112 (2011).

[5] Fraser, J., Rousseau, M., Shenker, S., Ferraiuolo, M.A., Hayashizaki, Y., Blanchette, M. & Dostie, J. Chromatin conformation signatures of cellular differentiation, *Genome. Biol.* **10**, R37 (2009).

[6] Havel, T., Kuntz, I. & Crippen, G. The theory and practice of distance geometry. *Bull. Math. Biol.* **45**, 665-720 (1983).

[7] Hu, M., Deng, K., Qin, Z., Dixon, J., Selvaraj, S., Fang, J., Ren, B. & Liu, J.S. Bayesian inference of spatial organizations of chromosomes, *PLoS Comput Biol.* **9**, e1002893 (2013).

[8] Kalhor, R., Tjong, H., Jayahilaka, N., Alber, F. & Chen, L. Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. *Nat. Biotechnol.* **30**, 90-98 (2011).

[9] Kruskal, J.B. & Wish, M. *Multidimensional scaling. Quantitative applications in the social sciences* (SAGE Publications Newbury Park, 1978).

[10] Lieberman-Aiden, E., van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., Sandstrom, R., Bernstein, B., Bender, M.A., Groudine, M., Gnirke, A., Stamatoyannopoulos, J., Mirny, L.A., Lander, E.S. & Dekker, J. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289-293 (2009).

[11] Nagano, T., Lubling, Y., Stevens, T.J., Schoenfelder, S., Yaffe, E., Dean, W., Laue, E.D., Tanay, A. & Fraser, P. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature* **502**, 59-64 (2013).

[12] Schoenberg, I.J. Remarks to Maurice Fréchet's article"Sur la définition axiomatique d'une classe d'Espace distanciés vectoriellement applicable sur l'espace de Hilbert". *Ann. Math.* **36**, 724-732 (1935).

[13] Sippl, M.J. & Scheraga, H.A. Solution of the embedding problem and decomposition of symmetric matrices. *Proc. Natl. Acad. Sci. USA* **82**, 2197-2201 (1985).

[14] Varoquaux, N., Ay, F., Noble, W.S. & Vert, J.P. A statistical approach for inferring the three-dimensional structure of the genome. *Bioinformatics* **30**, i26-i33 (2014).

[15] Torgerson, W.S. Multidimensional scaling. I. Theory and method. *Psychometrika* **17**, 401-419 (1952).

[16] Young, G., & Householder, A.S. Discussion of a set of points in terms of their mutual distances, *Psychometrika* **3**, 19-22 (1938).

[17] Zhang, Z., Li, G., Toh, K.C. & Sung, W.K. 3D chromosome modeling with semi-definite programming and Hi-C data. *J. Comput. Biol.* **20**, 831-846 (2013).