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 conformation capture (3C) experiments is to reconstruct spatial distances and three-dimensional genome structures from observed contacts between genomic loci. We propose a two-step algorithm, ShRec3D, and assess its accuracy using both *in silico* data and human genome-wide 3C (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 and 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, presumably reflecting the average spatial organization of their chromosomes.

Several methods have been developed to derive three-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 owing to algorithm trapping in local optima, and computation time is often prohibitive for large data sets. Therefore, they resort to data binning at the cost of lowering genomic resolution. We propose a two-step alternative: a method adapted from network analysis for translating contact maps into distances, followed by a 3D reconstruction.

It is straightforward to get all the distances between a set of N points given their 3D coordinates (**Fig. 1a**, step 1). From this matrix, one can also infer easily the binary contact matrix given a contact threshold ε (**Fig. 1a**, step 2). Our goal was to infer the optimal 3D coordinates knowing only the contact matrix. The mathematical problem of reconstructing a spatial structure from the distances between its elements is solved by distance geometry[12] or classical multidimensional scaling[13] (MDS; Online Methods). These methods involve the computation of the first three eigenvectors in intermediary matrix[14], the Gram matrix (**Fig. 1a**, steps 4 and 5) and have been used previously in the context of 3C experiments[1,10,11].

An important step in MDS-based methods of chromosome reconstruction is therefore the derivation of a complete set of distances from a (possibly sparse) contact map (**Fig. 1a**, step 6). We introduce a weighted 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[15]. We then take for the distance between any two nodes the length of the shortest path relating them on the graph, computed using the Floyd-Warshall algorithm. Our method accommodates binary contact maps (for example, single-cell Hi-C data)[7] by taking link lengths equal to 1 between contacting points, or else infinite (no link). Although it is approximate and gives distances in a dimensionless unit, this shortest-path metric assigns a sound distance (symmetric and satisfying the triangular inequality) to all pairs of points, as required for the application of MDS results (Online Methods). It offers a 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.

This algorithm, which we call 'shortest-path reconstruction in 3D' (ShRec3D), combines this shortest-path distance with MDS to achieve chromosome reconstruction (**Fig. 1a**, steps 4–6).

We tested the efficiency of ShRec3D in a controlled *in silico* case. We generated a yeast genome 3D structure[16] represented as $N = 26{,}538$ beads (each corresponding to approximately three nucleosomes) linked by springs accounting for intrachromosomal DNA connectivity. The 16 yeast chromosomes were confined into a nucleus of radius 1.6 μm (**Fig. 1b**). From the bead coordinates we computed the associated Gram and distance matrices (**Fig. 1c,d**) and a binary contact map (**Fig.1a**, steps 1 and 2 and **Fig. 1e**) and the distance matrix was then obtained by applying step 6 to this contact map (**Fig. 1f**). The consecutive application of steps 5 and 4 (**Fig. 1a**) reconstructs the coordinates up to a global transformation (some rotation, dilation and possibly mirror symmetry). To quantitatively assess the original structure recovery, we compared in a scatter plot the actual (**Fig. 1d**) and reconstructed (**Fig. 1g**) distances and computed their Spearman rank correlation[11] (**Fig. 1h**). A spectral analysis supported the dimensional reduction of step 4 (**Supplementary Fig. 1a**).

We compared, for data sets of various sizes, both the reconstruction accuracy and the speed of ShRec3D and two other methods, BACH[6] and ChromSDE[11]. All gave satisfactory results in terms of accuracy (**Fig. 2a**); however, on a personal computer, the run time for our script ranged from tens of seconds for small data sets (~1,000 points) to 50 h for the largest one (26,538 points),

[1]Laboratoire de Physique Théorique de la Matière Condensée, CNRS UMR 7600, Université Pierre et Marie Curie, Sorbonne Universités, Paris, France. [2]Institut de Génétique Moléculaire de Montpellier, CNRS UMR 5535, Université de Montpellier, Montpellier, France. [3]Institut Pasteur, Group Spatial Regulation of Genomes, Department of Genomes and Genetics, Paris, France. Correspondence should be addressed to J.M. (mozziconacci@lptmc.jussieu.fr).

RECEIVED 12 MARCH; ACCEPTED 6 AUGUST; PUBLISHED ONLINE 21 SEPTEMBER 2014; DOI:10.1038/NMETH.3104

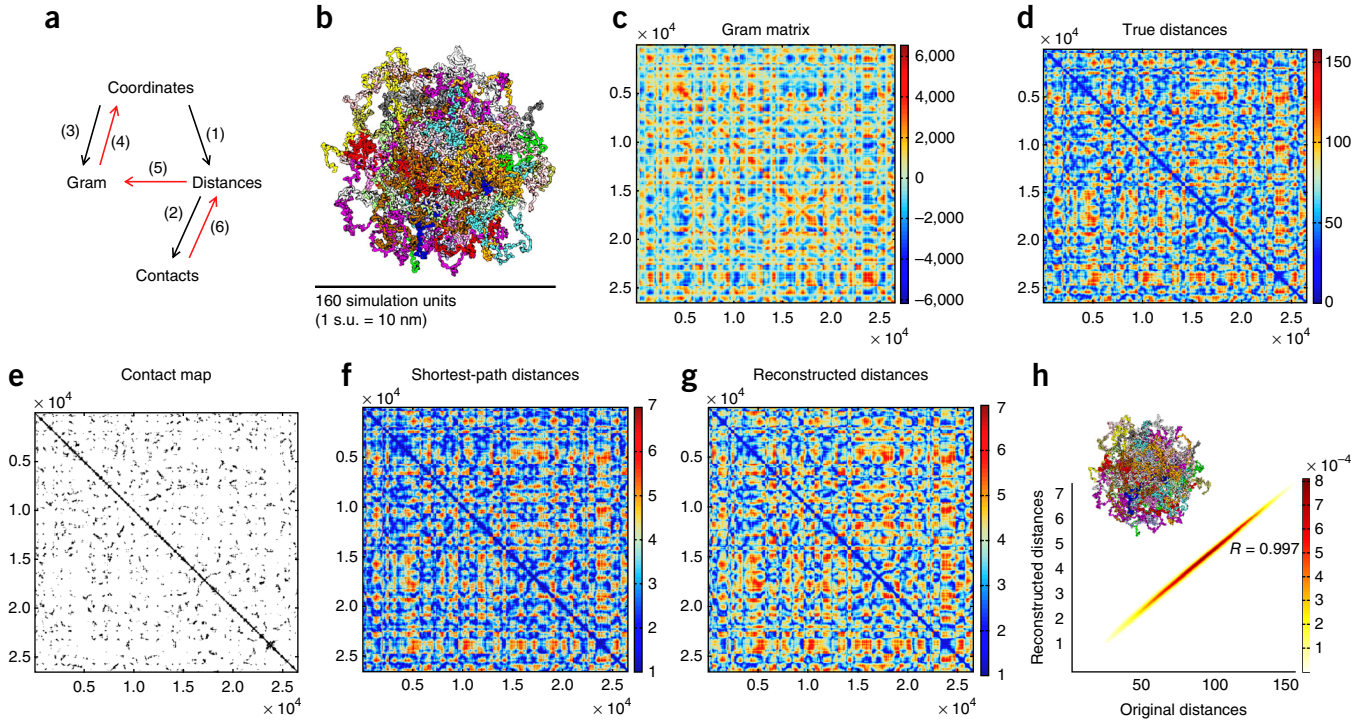

Figure 1 | Application of ShRec3D to a simulated data set. (**a**) Algorithm flowchart; steps 1–6 are detailed in the text and Online Methods. (**b**) 3D structure of *in silico* yeast genome generated using polymer dynamics for a chain of $N = 26{,}538$ beads (each chromosome is shown in a different color). (**c**) Gram matrix. (**d**) Distance matrix (s.u.). (**e**) Contact map (binary map, threshold $\varepsilon = 60$ nm). (**f**) Distance matrix derived from contacts. (**g**) Distance matrix of the reconstructed structure (dimensionless). (**h**) Scatter plot of original and reconstructed distances; heat map colors indicate the local density of points; Spearman rank correlation coefficient R is indicated. Inset, reconstructed 3D structure. s.u., simulation unit.

several orders of magnitude faster than other methods (**Fig. 2b**). The limiting step for ShRec3D computation time is the Floyd-Warshall algorithm computing shortest paths on the contact map, whose worst-case performance scales as $O(N^3)$. We also found that the accuracy of the MDS reconstruction applied directly to inverse-frequency distances was poor (**Fig. 2** and **Supplementary Fig. 1b–d**), demonstrating the importance of using our shortest-path metric before MDS reconstruction.

We tested and compared ShRec3D to the above-mentioned alternative methods in conditions closer to those of real Hi-C experiments. The robustness of the ShRec3D reconstruction with respect to experimental noise (mimicked by misplaced contacts, Online Methods) was insured for noise levels lower than 1% (maximal level in typical Hi-C experiments (**Fig. 2c**), 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 and is thus limited to small-size structures (**Fig. 2b**). We then reproduced the superposition of single-cell contact maps reflecting the genome fold variations over a cell population that is characteristic of Hi-C maps; accordingly, one could reach only an average 3D structure. From a Langevin dynamic simulation[16] of our *in silico* genome, we extracted a variable number k of independent structures and computed the average of their contact maps (Online Methods). The distances reconstructed with ShRec3D from this simulated average Hi-C contact map quantitatively matched the average distances in the superposition of structures (**Fig. 2d**). This was 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 largest size manageable in

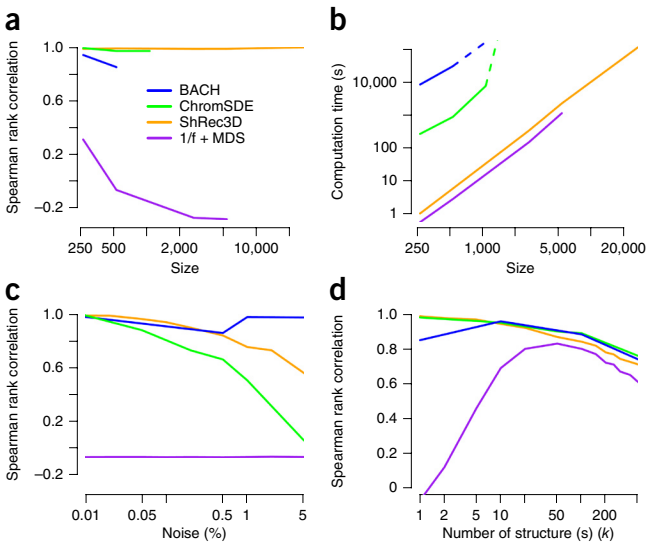

Figure 2 | Quantitative assessment of ShRec3D performance and reliability. (**a**,**b**) Comparison of ShRec3D with BACH[6], ChromSDE[11] and MDS applied to inverse-frequency distances for simulated data of increasing size N (number of beads) in terms of reconstruction accuracy (Spearman rank correlation between original and reconstructed distances) (**a**) and computation time (**b**). (**c**) Robustness to a controlled amount of randomly misplaced contacts mimicking experimental noise (semilog 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 of structure in **Fig. 1b** coarse-grained to $N = 480$ points) and distances reconstructed from the corresponding average contact map.

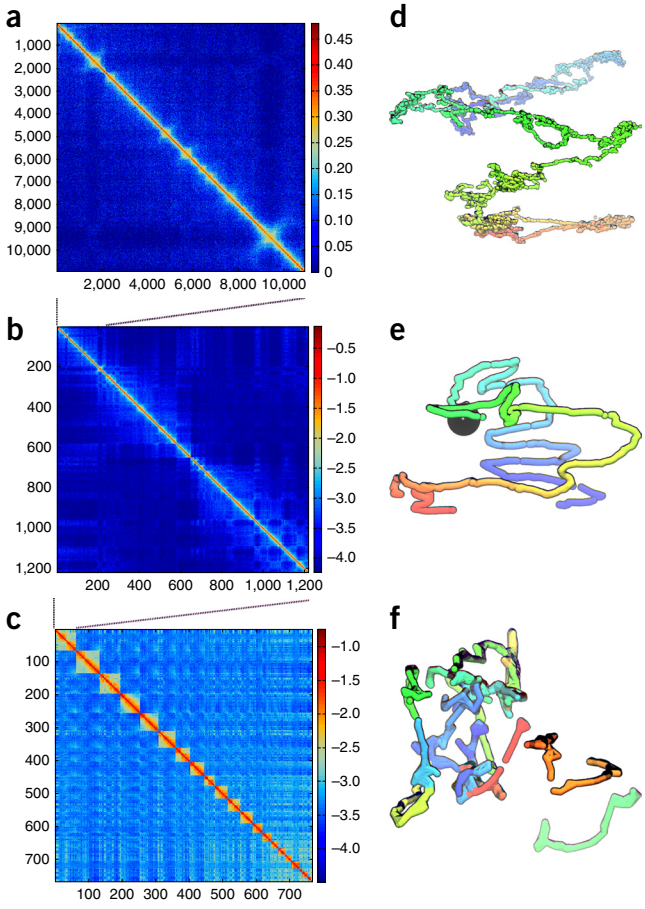

Figure 3 | 3D multiscale visualization of human autosomal chromosomes from Hi-C data. (**a**–**f**) Experimental contact maps (**a**–**c**) and corresponding 3D reconstructions (**d**–**f**) at genomic resolutions, from the scale of restriction fragments in chromosome 1 (embryonic stem cells (hESCs))[17] (SRX128221) (**a**) to that of bins containing 50 (**b**) or 1,000 (**c**) restriction fragments covering the whole chromosome set[18] (SRX030110) (**Supplementary Fig. 3**). Color gradients in **d**–**f** indicate the position along the genome.

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 was expected, as 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 data sets (where loci correspond to many restriction fragments). ShRec3D's ability to visualize average structures at different scales is illustrated by reconstructing a 30-Mbp region of chromosome 1 at 3-kb resolution (**Fig. 3a**), the chromosome average structure at 150-kb resolution (**Fig. 3b**) and the average arrangement of autosomal chromosomes within nuclear space at 3-Mbp resolution (**Fig. 3c**). The genome connectivity and chromosome partitioning achieved by ShRec3D (**Fig. 3d–f** and **Supplementary Figs. 2** and **3**) would make it an efficient tool for genome scaffolding from Hi-C data[19,20]. Alternative methods (BACH[6], MDS applied to inverse frequency distances and ChromSDE[11]) did not manage to properly reconstruct fine-resolution structures in reasonable amounts of time. The potential of ShRec3D to devise 3D genome browsers is illustrated with the coloring of a 3D structure of chromosome 1 at resolution 30-kb according to the chromatin partition in two compartments[2] (**Supplementary Fig. 4a**). Any chemical, structural or functional annotation available on linear genomes can be similarly overlaid on chromosome 3D structures (for example, two histone H3 modifications, H3K9Ac and H3K9me3 (GEO GSM409308)) (**Supplementary Fig. 4b**).

ShRec3D involves no *ad hoc* constraints or tunable parameters and is free from convergence issues and misleading transient outcomes. Its speed makes it applicable to both 3C or carbon-copy 3C (5C) data sets, which typically involve tens of loci, and high-resolution Hi-C data sets, comprising sparse contacts between hundreds of thousands of points. Its accurate reconstruction of average distances between genomic loci and visualization of a consensus structure enable a meaningful use of cell-population Hi-C data, especially when extended into 3D genome browsers.

URLs. The ShRec3D algorithm is available at https://sites.google.com/site/julienmozziconacci/#TOC-Downloads.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

The authors thank D. Arendt for the online-available implementation of the Floyd-Warshall algorithm. They acknowledge funding from UPMC (Université Pierre et Marie Curie, Sorbonne-Universités), grant CONVERGENCE2011, project CVG1110 (J.M.), from the French National Cancer Institute, grant INCa_5960 (A.L.), from the French National Research Agency (ANR), grant ANR-13-BSV5-0010-03 (A.L.) and from the French National Research Agency (ANR), grant ANR-09-PIRI-0024 (A.C.).

AUTHOR CONTRIBUTIONS

J.M., A.L. and J.R. designed the algorithm. J.M. implemented it. J.R., P.R., A.C. and J.M. tested its validity. A.C. analyzed experimental data sets. J.M. and A.L. wrote the paper.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

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, Z. *et al. Nature* **465**, 363–367 (2010).
6. Hu, M. *et al. PLOS Comput. Biol.* **9**, e1002893 (2013).
7. Nagano, T. *et al. Nature* **502**, 59–64 (2013).
8. Rousseau, M. *et al. BMC Bioinformatics* **12**, 414 (2011).
9. Trieu, T. & Cheng, J. *Nucleic Acids Res.* **42**, e52 (2014).
10. Varoquaux, N., Ay, F., Noble, W.S. & Vert, J.P. *Bioinformatics* **30**, i26–i33 (2014).
11. Zhang, Z. *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 (2012).
19. Burton, J.N. *et al. Nat. Biotechnol.* **31**, 1119–1125 (2013).
20. Kaplan, N. & Dekker, J. *Nat. Biotechnol.* **31**, 1143–1147 (2013).

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 an 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 an $N \times N$ positive semidefinite 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 an $N \times N$ 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, otherwise 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[21].

Explicit and reciprocal relationships (**Fig. 1a**) can be established between the above matrix representations of an N-point structure. As detailed below, steps 1–3 (numbering as in **Fig. 1a**) 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 reconstruction in 3D') involves first the translation of a contact map into a distance matrix using a graph-theoretic method (**Fig. 1a**, step 6; see below) then the reconstruction of a 3D structure using standard results from distance geometry and classical multidimensional scaling (MDS) (**Fig. 1a**, step 5 followed by step 4; see below).

*Step 1 from coordinates to distances (*V *to* D*).* The Euclidean distance between the points P_i and P_j straightforwardly is expressed as a function of the coordinates, i.e.,

$$D_{ij} = \sqrt{\sum_{\alpha=1}^{n} |V_{\alpha i} - V_{\alpha j}|^2}$$

*Step 2 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 3 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. Distance geometry has been developed to solve the issue of recovering coordinates from the sole knowledge of distances[12,14,22]. 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[13,23]; as explained below, it is based on algebra and explicit (analytical) formulas. It should be noted 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 ref. 24 for an overview and ref. 10 for an application to chromosome reconstruction).

*Step 4 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 semidefinite, the coordinates of the N points P_i ($i = 1,\ldots,N$) in an 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

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

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). This result is presented clearly and demonstrated in an accessible way in ref. 14 (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 should be noted that there is 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)[1]. However, the Cholesky decomposition does not exist if G has vanishing diagonal elements and is unstable if these elements are small[12]. The constructive result used in our step 4, although basically similar, is conceptually and practically simpler, as 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[14].

*Step 5 from distances to Gram matrix (*D *to* G*).* The mathematical derivation of a Gram matrix from the knowledge of distances is presented in ref. 14, theorems 3.1 and 3.3. 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

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

doi:10.1038/nmeth.3104

The argument introduces an auxiliary matrix, the metric matrix M, whose elements are

$$M_{ij} = \frac{1}{2}[d_{0i}^2 + d_{0j}^2 - D_{ij}^2] \quad (3)$$

It is shown in ref. 22 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 positive semidefinite of rank n. Then M coincides with the Gram matrix G of the N points, and step 4 applies.

In practice, small errors in the distances may cause some eigenvalues of M to be negative (though small). This means that an exact 3D Euclidean embedding does not exist. This problem has been mentioned in the context of 3C data[1] 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 3D 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[14]. Importantly, this reconstruction, step 5 above, requires that the full set of distances, i.e., D_{ij} for any pair (i,j), is known. This requires a preprocessing step of the contact map, converting contact frequencies into distances. Filling this gap is an important advance provided by step 6 (see below).

MDS reconstruction truncates the metric matrix M, equation (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 more valid when the neglected eigenvalues are small and more separated by a large spectral gap from the three retained eigenvalues (**Supplementary Fig. 1a**).

Shortest-path method: from contact map to distance matrix. We introduce a 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 or 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, as 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 $(i_0, i_1,., i_k)$ 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 $A_{ij}^k > 0$. In mathematical terms, this 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. This path is not necessarily unique.

In the case of Hi-C experiments, performed over a cell population and hence over an accumulation of structures, each pair (i,j) of nodes is associated with a normalized contact frequency f_{ij}, although it may be vanishing or very small. A current method, henceforth termed the inverse-frequency method, is to assign the value $1/f_{ij}$ to the distance between points P_i and P_j (ref. 15). 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, however, gives unreliable or even meaningless distance values for small f_{ij}. Moreover, it does not define a true distance, as 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 $(i_0, i_1,., i_k)$ with $i = i_0$ and $j = i_k$, such that the path length is minimal over all the paths relating P_i and P_j. Although 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 the shortest path relating them. Other choices are possible for relating link length and contact frequency, for example, $1/f^{\alpha}$. Basically, the exponent modifies the relative weight of rare contacts (decreasing for $\alpha > 1$, increasing for $\alpha < 1$). The value of this exponent α appeared to have little effect on the reconstruction quality, owing 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 Floyd-Warshall algorithm for computing shortest paths and their lengths. Interestingly, by construction, weak or vanishing contact frequencies do not contribute to the distances, as the shortest paths will bypass the corresponding links (of large or infinite lengths). This method thus makes it 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. First, it is obviously symmetrical and vanishes only if the points are identical. Second, 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 described previously[15]. The shortest-path method improves the distance assigned to pairs of loci with low or vanishing contact frequencies, for which the inverse-frequency expression is unsatisfactory. It is also beneficial in case of a binary contact map (for example, for single-cell Hi-C data)[7], where the inverse-frequency method yields distance values either equal to 1 or infinite. Finally, the distance between neighboring loci along the genome will be satisfactorily small, as 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 (**Supplementary Fig. 2**).

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 three nucleosomes. The chain is confined in a spherical nucleus of radius 1.6 μm. The simulation spatial unit

corresponds to 10 nm 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 1 and 2 above. Considering a regularly spaced sampling of the N points allows one to investigate different data sizes. In particular, such coarse graining is mandatory in the comparison with alternative methods (see below) that cannot handle sizes as large as $N = 26{,}538$. Finally, we used this Langevin simulation to generate an ensemble of structures.

Implementation of the algorithm ShRec3D. By supplementing our shortest-path distance derivation, step 6, with MDS reconstruction, steps 5 and 4, we obtained a constructive algorithm, ShRec3D. It allows us to visualize a 3D structure from the knowledge of any contact map (either binary, in terms of contact presence or 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 was written with MATLAB (http://www.mathworks.com/products/matlab/).

As mentioned above, the coordinate reconstruction involves only the first three eigenvectors of the metric matrix M (equation (3)), as if the other eigenvalues were vanishing (an approximation previously proposed[1] and quantitatively assessed)[14]. The validity of this eigenvalue truncation, approximating M by a positive semidefinite matrix G of rank 3, is assessed by investigating the spectrum of M and ensuring 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 (**Fig. 1**). **Supplementary Figure 1a,b** shows 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 negative eigenvalues originates from the inaccuracies in the distance matrix derived from the data and propagated to the metric matrix. Indeed, although 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 or absence). Because of this loss of information in the generation of our synthetic data, the reconstructed metric matrix differs slightly from that of the original structure.

Implementation of available alternative methods. We compared the performance of our algorithm ShRec3D with alternative reconstruction methods in terms of both reconstruction accuracy and speed (see below). To make a fair comparison, we limited ourselves to methods for which the original codes, optimized by their authors, were available, namely BACH software[6] and ChromSDE[11]. They all involve an optimization procedure and yield a single consensus structure from Hi-C contact maps. We ran all the software programs on the same Linux machine. We ran BACH for 5,000 iteration steps (default value) and used ChromSDE in linear mode. Comparison on real data is limited at present, as 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 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 sites is set equal to their inverse contact frequency $1/f$ (ref. 15), both completed by the MDS 3D reconstruction described above. Although 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 genomic loci considered). A first comparison (**Supplementary Fig. 1a,b**) shows that the dimensional reduction step in MDS reconstruction is not legitimate when the preprocessing step 6 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 (**Supplementary Fig. 1c,d** shows a scatter plot of the reconstructed and original matrix distances and the Spearman correlation coefficient R for both methods, applied to the same simulated benchmark as in **Fig. 1** and **Supplementary Fig. 1a,b**). The poor performance of the simple method originates, presumably, from the fact that it 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 (**Supplementary Fig. 1d**) could be even stronger when larger structure sizes are considered. An alternative improvement to 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 data set and each description scale[11]. Although this method is satisfactory in terms of reconstruction quality, the inherent optimization procedure makes it very costly in computation time (**Fig. 2b**).

Performance of the ShRec3D algorithm in terms of computation time. We plotted the computation time (in seconds) using our method and alternative methods (**Fig. 2b**) as a function of N (number of points in the structure to be reconstructed). This time scales roughly as $O(N^3)$ using both the BACH software and our method, whereas it does not scale as any power of N for ChromSDE. For a size of 1,000 points, our algorithm runs in ~20 s, whereas it takes more than 1 d to converge for BACH. For a size >1,000 points, the ChromSDE software did not reach convergence. MDS reconstruction applied to inverse-frequency distances is obviously faster than ShRec3D, as it skips the Floyd-Warshall computation of shortest path.

Quantitative assessment of reconstruction quality. 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 because 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, for example, in ref. 11 for the same purpose), which avoids both

doi:10.1038/nmeth.3104

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 confirmed that our reconstruction ShRec3D preserves polymer connectivity, that is, that neighboring loci along the genome are also close neighbors in the 3D space (**Supplementary Fig. 2a**).

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 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. This largely preserves the total number of contacts (the moves displacing a positive entry toward an already positive entry are very scarce for realistic contact densities). 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 the concentrations of both the ligase and the cross-linking factor. One way to quantify the noise in Hi-C experiments is to estimate the proportion of random ligation events in the bank. One can, for instance, use the fact that organisms such as *S. cerevisiae* have mitochondria 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 that this proportion was typically smaller than 1%. Such an estimate can also be derived from a recently developed analysis of metagenomic samples[25].

We investigated whether the reconstruction accuracy was affected by the presence of a fraction of misplaced contacts ranging from 0.01% (no noise) to 5% (above the upper boundary on the experimental noise strength). As in the noiseless case, the quantitative assessment was done by plotting the Spearman rank correlation coefficient R between the original distance matrix and the distances in the structure reconstructed from the noisy contact map as a function of noise strength (**Fig. 2c**). For comparison, we also implemented alternative methods (BACH[6] and ChromSDE[11]) in the same noisy conditions (**Fig. 2c**). 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[26]. MDS applied to inverse-frequency distances yields a uniformly poor result (R close to 0) (**Supplementary Fig. 1d**).

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 a 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 thoroughly explore the conformation space. Up to 500 snapshots, separated by a sufficiently long run of the dynamical simulation, were extracted. Taken together, they yield a realistic Hi-C contact map (steps 1–2). 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 (**Fig. 2d**). Owing to the prohibitive run time of BACH and ChromSDE for large structures, we considered coarse-grained structures with only $N = 480$ points.

Normalization and representation of real Hi-C data sets. The procedure we used to normalize the data was presented previously[21]. 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 1. 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. 3a**) the contact map is very sparse, and the color bar is graded with respect to the contact frequency to the power of 0.3 to increase the contrast. At lower resolutions the number of contacts is computed between groups ('bins') of 50 and 1,000 restriction fragments. For these two resolutions (**Fig. 3b,c**), the color bar is established in log scale. We confirmed (**Supplementary Fig. 2b**) that polymer connectivity was preserved by our reconstruction (i.e., that neighboring loci along the genome were 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. It is possible to label the chromosomes and distinguish which chromosome each genomic locus belongs to (**Supplementary Fig. 3**). We normalized intra- and interchromosomal contacts independently and then added the two normalized matrices 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 the partition of the human chromatin structure in two compartments inferred from the spectral analysis of the correlation matrix between the lines of the contact map[1] (**Supplementary Fig. 4a**) and the linear profiles of acetylation and trimethylation of lysine 9 of histone H3 (GSM409308) (**Supplementary Fig. 4b**).

21. Cournac, A. *et al. BMC Genomics* **13**, 436 (2012).
22. Schoenberg, I.J. *Ann. Math.* **36**, 724–732 (1935).
23. Young, G. & Householder, A.S. *Psychometrika* **3**, 19–22 (1938).
24. Kruskal, J.B. & Wish, M. Sage University papers series on quantitative applications in the social sciences (no. 07-011) (SAGE Publications, Newbury Park, 1978).
25. Burton, J.N. *et al. G3*, **4**, 1339–1346 (2014).
26. Franzke, B. & Kosko, B. *Phys. Rev. E* **84**, 041112 (2011).