



METHOD ARTICLE

**REVISED** Improved inference of chromosome conformation from images of labeled loci [version 3; peer review: 2 approved]

Brian C. Ross , James C. Costello

Computational Bioscience Program, Department of Pharmacology, University of Colorado, Anschutz Medical Campus, Aurora, CO, 80045, USA

**v3** **First published:** 21 Sep 2018, 7(ISCB Comm J):1521 (<https://doi.org/10.12688/f1000research.16252.1>)  
**Second version:** 11 Mar 2019, 7(ISCB Comm J):1521 (<https://doi.org/10.12688/f1000research.16252.2>)  
**Latest published:** 28 Mar 2019, 7(ISCB Comm J):1521 (<https://doi.org/10.12688/f1000research.16252.3>)

**Abstract**

We previously published a method that infers chromosome conformation from images of fluorescently-tagged genomic loci, for the case when there are many loci labeled with each distinguishable color. Here we build on our previous work and improve the reconstruction algorithm to address previous limitations. We show that these improvements 1) increase the reconstruction accuracy and 2) allow the method to be used on large-scale problems involving several hundred labeled loci. Simulations indicate that full-chromosome reconstructions at 1/2 Mb resolution are possible using existing labeling and imaging technologies. The updated reconstruction code and the script files used for this paper are available at: <https://github.com/heltilda/align3d>.

**Keywords**





chromosome, conformation, reconstruction, fluorescence, genetic, loci






This article is included in the **International Society for Computational Biology Community Journal gateway**.

**Open Peer Review**

**Reviewer Status**  

	Invited Reviewers	
	1	2
<b>REVISED</b> version 3 published 28 Mar 2019		 report
<b>REVISED</b> version 2 published 11 Mar 2019	 report	↑
version 1 published 21 Sep 2018	↑  report	↑  report

- 1 **Ewan Birney** , European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Cambridge, UK  
**Carl Barton**, European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Cambridge, UK
- 2 **Marco Cosentino Lagomarsino** , University of Milan, Milan, Italy  
 IFOM Foundation—FIRC Institute of Molecular Oncology, Milan, Italy  
**Vittore Scolari** , Pasteur Institute, Paris, France

Any reports and responses or comments on the article can be found at the end of the article.

**Corresponding authors:** Brian C. Ross ([brian.ross@ucdenver.edu](mailto:brian.ross@ucdenver.edu)), James C. Costello ([JAMES.COSTELLO@ucdenver.edu](mailto:JAMES.COSTELLO@ucdenver.edu))

**Author roles:** **Ross BC:** Conceptualization, Formal Analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Costello JC:** Funding Acquisition, Project Administration, Resources, Supervision, Writing – Review & Editing

**Competing interests:** No competing interests were disclosed.

**Grant information:** Funding was provided by the Boettcher Foundation (J.C.C.), National Institutes of Health [2T15LM009451 to B.C.R.], and a Cancer League of Colorado grant (J.C.C. and B.C.R.).

*The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.*

**Copyright:** © 2019 Ross BC and Costello JC. This is an open access article distributed under the terms of the [Creative Commons Attribution Licence](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

**How to cite this article:** Ross BC and Costello JC. **Improved inference of chromosome conformation from images of labeled loci [version 3; peer review: 2 approved]** F1000Research 2019, 7(ISCB Comm J):1521 (<https://doi.org/10.12688/f1000research.16252.3>)

**First published:** 21 Sep 2018, 7(ISCB Comm J):1521 (<https://doi.org/10.12688/f1000research.16252.1>)

**REVISED** Amendments from Version 2

We improved a sentence comparing our method to ChromoTrace, in the Introduction section.

See referee reports

## Introduction

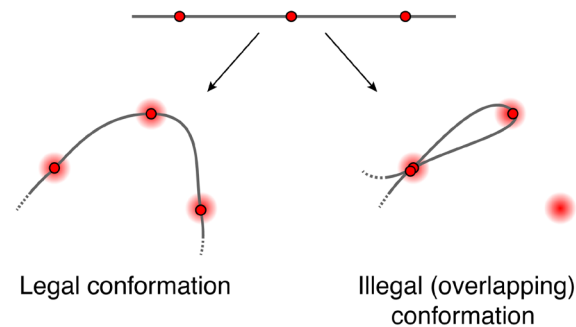
Measurement of *in vivo* chromosome conformation is a major unsolved problem in structural biology despite its known biological importance<sup>1</sup>. The present state-of-art is to either obtain indirect information about conformations using 3C-derived methods which measure DNA-DNA contacts (typically in a cell-averaged population)<sup>2</sup>, or else directly measure the cellular locations of individual chromosomal loci in single cells by microscopy<sup>3</sup>. The major limitation of direct localization is one of throughput: only ~ 3–5 labeled loci can be uniquely identified ‘by color’ in a standard microscope image, whereas a whole-chromosome reconstruction would involve labeling and identifying hundreds or thousands of loci.

Several research efforts aim to remove the color limitation either by experimental improvements or computational inferences. The experimental approaches aim to allow an increased number of labels that can be distinguished in an image<sup>4–6</sup>. Alternatively, attempts have been made to infer the identity of labels that cannot be uniquely identified in an image, by comparing the image to the known label positions along the DNA contour. The first attempt to do this was ‘by eye’<sup>7</sup>, but subsequently two computational algorithms have been developed to automate this inference: *align3d*<sup>8</sup> and *ChromoTrace*<sup>9</sup>. There are two important differences between these algorithms. First, *align3d* has less stringent experimental requirements than *ChromoTrace*, as it allows for missing labels in the image and does not require a uniform label spacing along the chromosome. Second, *ChromoTrace* outputs explicit conformations, whereas *align3d* outputs likelihoods of the various possible identities for each labeled locus. Both approaches have their advantages: *ChromoTrace* output is straightforward to interpret, whereas *align3d* output gives information on the range of possible conformational solutions along with their likelihoods.

This paper presents improvements to *align3d*<sup>8</sup> that allow it to generate high-quality, chromosome-scale conformational reconstructions. First, we briefly describe the algorithm. Using a) the genomic locations and colors of labeled loci and b) the spatial locations and colors of spots in a microscope image, together with a relation tying the genomic distance between two loci to their average spatial displacement, this method constructs a table of ‘mapping probabilities’  $p(L \rightarrow s)$  for a given labeled genomic locus  $L$  having produced spot  $s$  in the microscope image. Each mapping probability  $p(L \rightarrow s)$  is calculated by dividing the summed statistical weights of conformations where locus  $L$  maps to spot  $s$ , which we term a mapping partition function and denote  $Z_{L \rightarrow s}$ , by the full partition function  $Z$  that is the summed weight of all conformations. A proper calculation of  $Z_{L \rightarrow s}$  and  $Z$  would consider all conformations having no

more than one locus at any given spot in the image<sup>1</sup>, similar to a traveling salesman tour<sup>10</sup>, but this exact calculation is intractable for large problems. Instead, *align3d* counts all conformations for which *adjacent* loci do not overlap at the same spot (see **Figure 1**), using a variant of the forward-backward algorithm<sup>11</sup> that can propagate between non-adjacent layers. This is a major source of error as the vast majority of conformations contributing to the partition function overlap at non-adjacent loci, and one consequence is that the normalization of mapping probabilities makes no sense for a non-overlapping conformation, as  $\sum_L p(L \rightarrow s)$  can exceed 100% for certain spots. To recover from this error, *align3d* assigns a penalty to each spot and iteratively adjusts these penalties until the spot normalization is sensible. Although somewhat ad hoc, use of spot penalties recovers significant information about medium-sized conformations (~ 30 labeled loci), although larger simulated experiments (~ 300 loci) have convergence problems due to the cost function plateauing at very small or large values of the spot penalties.

The final step is to use the mapping probabilities to construct the range of likely conformations compatible with the microscope image. Uncertainty in the conformation results from inaccuracy or uncertainty in the mapping probabilities due to three factors: inaccuracy in the DNA model (the relation between genomic and spatial distance), error in estimating the partition functions, and the inherent uncertainty in the data even with a perfect reconstruction algorithm. The DNA model can be calibrated by a control experiment, and we argue that the remaining model error can reduce our method’s confidence in its results but it generally does *not* cause our method to reconstruct mistaken conformations. The main focus of this paper is on improving the partition function estimate, using two different strategies. First, we give an efficient method for optimizing the spot penalties when there



**Figure 1. Legal versus illegal (overlapping) conformations.** Schematic showing one legal and one illegal conformation passing through spots  $A$ ,  $B$  and  $C$ . *align3d* counts both legal and overlapping conformations in estimating the partition  $Z$  (although it is able to prevent *adjacent* loci from overlapping).

<sup>1</sup>Depending on how the experiment is done, two spots of the same color sufficiently close in the image may appear as a single spot where the conformation self-overlaps. We prefer to treat this scenario as a missing-spot measurement error rather than relax the one-spot-per-locus rule. If the spots have been properly localized, then the underlying conformation visits any given spot once at most.

are hundreds of spots in the image. Next, we provide formulas for the partition functions which allow them to be estimated to arbitrarily high accuracy (given enough computation time), without using spot penalties or any optimization. As we show using simulations, these two methods used individually or in tandem permit confident, chromosome-scale conformational reconstructions using existing experimental technologies.

**Methods**

First we provide a method for efficiently optimizing the spot penalties regardless of the number of labeled loci. This rule guarantees that a) the rate of missing spots is as expected, and b) the mapping probabilities are properly normalized. Let  $q_s$  denote the penalty attached to spot  $s$ ; then the update rule for that spot penalty is:

$$q'_s = \frac{\frac{1}{P(s)/N} - 1}{\frac{1}{1 - p_{jn}(c)} - 1} \cdot \frac{\frac{1}{P(s)/N} - 1}{\frac{1}{\min(l, P(s))N} - 1} \cdot q_s \quad (1)$$

where  $N$  is the number of loci,  $P(s) = \sum_L p(L \rightarrow s)$  is the total probability of mapping any locus to spot  $s$ , and  $p_{jn}(c)$  is the estimated rate of missing spots having color  $c$ . The justification for this rule is given in [Appendix 1 \(Supplementary File 1\)](#).

We can also update a penalty  $\bar{q}_c$  that is associated with *missing* spots of color  $c$ . This gives a faster way to enforce a desired missing spot rate because there are fewer  $\bar{q}$  penalties than  $q$  penalties. An update to  $\bar{q}_c$  is equivalent to a reverse update to all  $q_s$  for spots  $s$  of color  $c$ , so the update rule is:

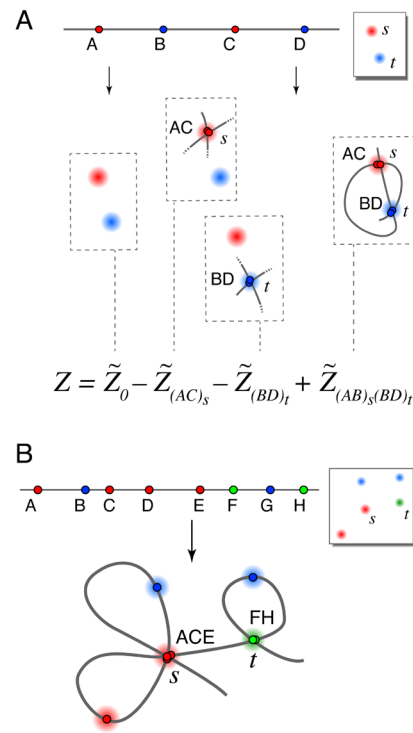
$$\bar{q}'_c = \frac{\frac{1}{1 - p_{jn}(c)} - 1}{\frac{1}{P(s)/N} - 1} \cdot \bar{q}_c \quad (2)$$

Typically, we first optimize the  $\bar{q}$  parameters to achieve a target missing spot rate, then optimize the  $q$  parameters to enforce  $P(s) \leq 1$  while maintaining the missing spot rate. In either case, we apply [Equation 1](#) or [Equation 2](#) to bring the  $q$  or  $\bar{q}$  parameters close to their final values. When the cost function stops improving, we switch to the steepest-descent algorithms used in [Ross and Wiggins, 2012<sup>8</sup>](#) to polish  $q$  or  $\bar{q}$ .

Next, we give two exact formulas for the partition functions  $Z_{L \rightarrow s}$  and the full partition function  $Z$  that determine our locus-to-spot mapping probabilities. We focus on the full partition function  $Z$  since the formulas for  $Z_{L \rightarrow s}$  are identical. The largest term in each formula, which we denote  $\tilde{Z}_0$  (or  $\tilde{Z}_0^{opt}$  when spot penalty optimization is used), is the original estimate from [Ross and Wiggins, 2012<sup>8</sup>](#) calculated using a variant of the forward-backward algorithm<sup>11</sup>. Additional terms are computed in the same way, except that certain loci are constrained to map to certain spots. All of the constraints we will apply are *illegal constraints*, in that they force multiple loci to overlap at some spot in the image; therefore these terms only count illegal conformations that we would like to remove from

the baseline calculation. By computing these terms and subtracting them from  $\tilde{Z}_0$  we eliminate the overlapping conformations and improve the calculation. It turns out that this process erroneously subtracts conformations with multiple overlaps more than once and thus we have to add back in higher-order corrections (i.e. partition functions having multiple constrained spots). Repeating this logic leads to exact formulas for  $Z$  taking the form of series expansions, which are dominated by the lowest-order terms as those have the fewest restrictions on conformational overlaps. [Figure 2A](#) illustrates an example of such a series expansion, where each parenthetical subscript  $(XY \dots)_s$  on a term label denotes an illegal constraint forcing loci  $X, Y, \dots$  to overlap at spot  $s$  when that term is calculated. We use this notation throughout.

There are two ways we might remove conformations containing overlapping loci, leading us to two different series expansions for the true partition function  $Z$ . Suppose that we are calculating the term  $\tilde{Z}_{(AC \dots)_s}$  whose single illegal constraint forces loci  $A, C, \dots$  to overlap at spot  $s$ . One option is to forbid any of the other unconstrained loci from also mapping to spot  $s$ , since spot  $s$  is already overused. This leads to series expansion 1. Alternatively, allowing



**Figure 2. Series expansions. A.** Schematic showing terms in a series expansion, in a case where series 1 and series 2 have the same terms. The full series gives the exact partition function for the 4-locus experiment shown where only 2 spots appeared in the image (due to a high rate of missing spots). Cartoons show only the constrained loci for each term (so for example each term includes the illegal conformation visiting spots  $s \rightarrow t \rightarrow s \rightarrow t$ ). **B.** An illegal conformation for which loci  $A, C$  and  $E$  overlap at spots  $s$ , and loci  $F$  and  $H$  overlap at spot  $t$ . Series expansion 1 includes this conformation in terms  $\tilde{Z}_0$ ,  $\tilde{Z}_{(ACE)_s}$ ,  $\tilde{Z}_{(FH)_t}$  and  $\tilde{Z}_{(ACE)s(FH)_t}$ . Series expansion 2 includes this conformation in the same terms with the addition of  $\tilde{Z}_{(AC)_s}$ ,  $\tilde{Z}_{(AE)_s}$ ,  $\tilde{Z}_{(CE)_s}$ ,  $\tilde{Z}_{(AC)s(FH)_t}$ ,  $\tilde{Z}_{(AE)s(FH)_t}$  and  $\tilde{Z}_{(CE)s(FH)_t}$ .

further overlaps with spot  $s$  from the unconstrained loci gives us series expansion 2. [Figure 2B](#) illustrates the differences between the two series.

Each of the two series expansions is a weighted sum over *all possible illegally-constrained terms* having two properties: 1) each locus and each spot appear at most once in the indices, and 2) two or more loci map to each constrained spot. To be formal, we use  $\Omega$  to represent the set of all possible illegal constraints: each element of  $\Omega$  consists of a set of two or more non-adjacent loci and a single spot where they are forced to overlap. Each expansion thus takes the form

$$Z = \sum_{\phi \subseteq \Omega} w_{\phi} \tilde{Z}_{\phi} \quad (3)$$

where  $\tilde{Z}_{\phi}$  is zero if any two constraints share a locus or spot. We will choose the integer weights  $w_{\phi}$  so as to cancel out the overlapping conformations. By symmetry arguments, the weighting factor should not depend on the identities of the loci or spots, but rather only by the number of constrained spots  $n_{\phi}$  and the number of loci  $n_k^{\phi}$  involved in each  $k^{\text{th}}$  constraint. For example,  $w_{(ACE)_s(BD)_t}$  is determined by  $n_{\phi} = 2$ ,  $n_1^{\phi} = 3$  and  $n_2^{\phi} = 2$ .

Here we specify each series expansion by giving a formula for the weights  $w_{\phi}$  in terms of  $n_{\phi}$  and the various  $n_k^{\phi}$ . We also explain how to select an appropriate set of terms  $\psi$  when there are too many terms to evaluate. Our selection prohibits any legal or overlapping conformation from contributing a negative weight to the partition function estimate, thereby guaranteeing positive mapping probabilities and allowing use of the reconstruction-quality metrics given in Ross and Wiggins, 2012<sup>8</sup>. Derivations of the coefficient formulas and the term-selection criteria for each series expansion appear in [Appendix 2 \(Supplementary File 1\)](#).

**Series expansion 1** For series expansion 1, we do not allow the unconstrained loci to map to spots that were used in constraints. Then the weights  $w_{\phi}$  in the series formula given by [Equation 3](#) are:

$$w_{\phi} = (-1)^{n_{\phi}} \quad (4)$$

To select terms for a series approximation, we first choose a set of illegal constraints  $\psi$  to disallow, then include all series terms  $\tilde{Z}_{\phi}$  containing only those constraints: i.e.  $\phi \subseteq \psi$ . This guarantees non-negative mapping probabilities. In order to efficiently evaluate the largest terms, we recommend selecting the  $N_{\psi}$  constraints having the highest product of mapping probabilities in the baseline calculation  $\tilde{Z}_0$  (or  $\tilde{Z}_0^{\text{opt}}$  if spot penalties will be used). For example, we would include  $(AC)_s$  if  $p(A \rightarrow s) \cdot p(C \rightarrow s)$  is sufficiently large.

**Series expansion 2** For series expansion 2, the unconstrained loci are allowed to map to spots that were used in constraints. Then the weights  $w_{\phi}$  in [Equation 3](#) are:

$$w_{\phi} = \prod_{k=1}^{n_{\phi}} (-1)^{n_k^{\phi}-1} (n_k^{\phi} - 1). \quad (5)$$

To select terms for a series approximation, we first choose a set of  $N_{\psi}$  single-locus-to-spot mappings  $\Psi$ , then include all terms  $Z_{\phi}$  whose illegal constraints use only mappings in  $\Psi$ . For example, the constraint  $(AC)_s$  would be included if  $\Psi \subseteq \{A \rightarrow s, C \rightarrow s\}$ . In order to select the largest terms, we recommend building  $\Psi$  from the  $N_{\psi}$  largest mapping probabilities calculated from  $\tilde{Z}_0$  or  $\tilde{Z}_0^{\text{opt}}$ .

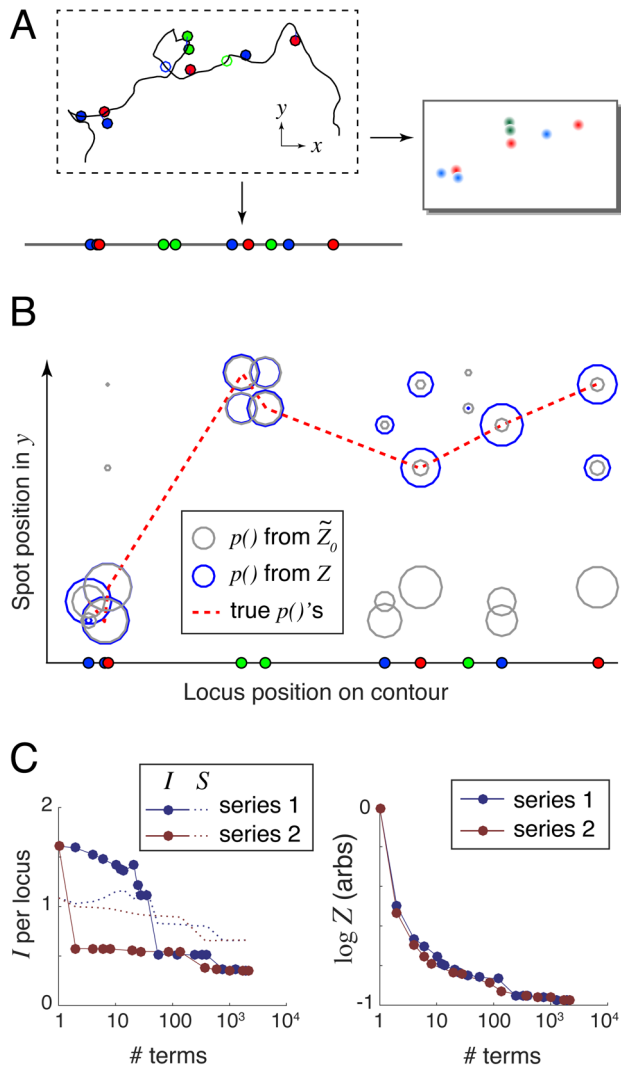
## Results

We tested the improved `align3d` method by generating random chromosome conformations using our software tool `wormulator` (version 1.1), and simulating the process of error-prone labeling, imaging and finally producing the locus-to-spot mapping probabilities. We considered three scenarios for our simulations. 1) The ‘Toy’ scenario involves 10 genomic loci, where each locus is labeled using one of 3 colors. For these simple problems the partition function can be calculated exactly. 2) Our simulated Experiment 1 uses standard DNA labeling methods and traditional 3-color microscopy to label 30 loci with 3 colors, thus interrogating a significant fraction of a chromosome contour. 3) Our simulated Experiment 2 labels 300 loci across a chromosome-length contour. The reconstruction of Experiment 2 is made possible by using the Oligopaints labeling technique<sup>4</sup> to label in 20 different colors.

For each scenario, we randomly generated 100 conformations using a wormlike chain model (packing density  $n_l = 0.3$  kb/nm, persistence length  $l_p = 300$  kb, as suggested by the measurements of Trask, Pinkel and van den Engh, 1989<sup>12</sup>); applied a random labeling at a mean density of 1 locus per megabase; and simulated experimental error: 100/200-nm Gaussian localization error in xy/z, a 10% rate of missing labels, and a 10% rate of nonspecific-bound labels. A typical simulated experiment from the Toy scenario is shown in [Figure 3A](#).

Next, we specified a DNA model relating the genomic distance between two loci  $L$  to their expected RMS spatial distance  $R$ , which is used by `align3d` to estimate the probability density of spatial displacement  $\mathbf{r}$  using a Gaussian chain model:  $\sigma(\mathbf{r}) \propto \exp[-3|\mathbf{r}|^2/2R^2]$ . Our current implementation requires a power relation between  $R$  and  $L$ , where the exponent may depend on  $L$ . Since any realistic polymer model predicts straight DNA on very short scales, we chose the model  $R = n_l L$  for  $L < l_p$  and  $R = A_{\rho} L^{\rho}$  for  $L > l_p$ , where  $A_{\rho} = l_p^{-(\rho-1)} \cdot n_l$  for continuity. In a real experiment the three free parameters  $n_l$ ,  $l_p$  and  $\rho$  would be fit to pairwise distance distributions between different pairs of loci in a separate calibration experiment. For our purposes  $n_l$  and  $l_p$  were set to the same values used to generate the wormlike chain conformations, and since these conformations were random walks we set  $\rho = 1/2$ .

For each simulated conformation, we fed the label positions and colors together with the simulated 3D images and our DNA model into the `align3d` algorithm to produce locus-to-spot mapping probabilities. For example, the simulated experiment shown in [Figure 3A](#) produced the mapping probabilities shown graphically in [Figure 3B](#) using circles, where the size of each circle indicates probability magnitude. Here grey circles show the mapping probabilities computed from  $\tilde{Z}_0$  with no use of spot



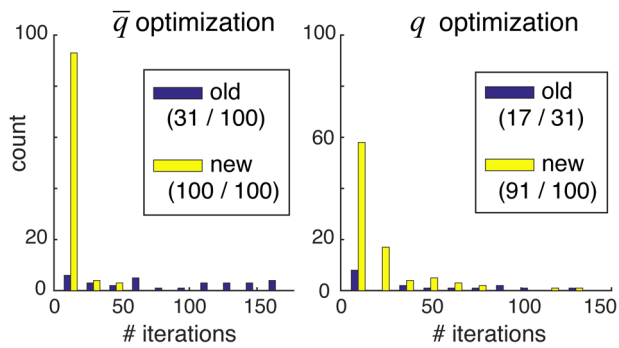
**Figure 3. Example reconstruction.** **A.** Randomly generated and labeled chromosome contour with simulated experimental error: localization error (lines offsetting spots from the labeled genomic loci) and missing labels (open circles). This example lacks nonspecifically-bound labels (floating spots). **B.** Spot mapping probabilities calculated using both the largest series term  $\tilde{Z}_0$  (grey circles), and the exact  $Z$  that can be computed using 2210 series terms (blue circles). The dotted red line connects the true locus-to-spot mappings, which are used to calculate the unrecovered information. In this example  $I(\tilde{Z}_0) = 1.54$  bits/locus and  $I(Z) = 0.32$  bits/locus. **C.** Unrecovered information  $I$  and entropy  $S$  (left panel) and  $\log Z$  (right panel) versus the number of terms used in the series expansions.

penalties, and blue circles show those same probabilities computed using the exact  $Z$ . This example shows how excluding high-weight and heavily-overlapping conformations reduces and improves the partition function estimate (see Figure 3C) and concentrates the probability mass into the ‘true’ locus-to-spot mappings (shown connected by the dotted red line in Figure 3B).

Our reconstruction quality metric is the amount of *unrecovered information* from the mapping probabilities, defined as  $I = -\langle \log p(L_i \rightarrow s_j) \rangle_i$  where the average  $\langle \cdot \rangle$  is taken over the set of true locus-to-spot mappings  $(L_i, s_j)$ . The ideal case of  $I \rightarrow 0$  implies a perfect reconstruction with no mistakes and zero uncertainty, but in practice  $I$  is always positive. In a real experiment where the true mappings are not known, we use a proxy for unrecovered information that we term entropy, defined as  $S = -\langle p(L_i \rightarrow s_j) \log p(L_i \rightarrow s_j) \rangle_{ij}$  whose average is taken over all locus-to-spot mappings, not just the correct mappings. The goal is to have  $S \approx I$  so that a real experiment will have an accurate estimate of the reconstruction performance. The left-hand panel of Figure 3C shows how  $I$  and  $S$  depend on the accuracy of the calculation for the simple example shown, using either of the two series expansions and varying the number of terms from 1 (simply  $\tilde{Z}_0$ ) to 2210 which is the full set of terms for either series and thus computes  $Z$  exactly. Entropy generally overestimates the amount of unrecovered information (see Supplementary Figure S1 and Supplementary Figure S2, Supplementary File 1), because the large mapping probabilities should be even larger, and the small ones even smaller, than their assigned values (see Supplementary Figure S3, Supplementary File 1). Appendix 3 (Supplementary File 1) argues that this miscalibration is caused by the mismatch between the wormlike chain DNA model used to generate the simulated conformations and the Gaussian chain model used by align3d in the reconstruction.

**Validation of Equation 1–Equation 5.** We first validated each of the two series expansions by comparing them against exact partition function calculations for the simulated Toy experiments. In all cases, both series expansions, when taken to their maximum number of terms, exactly reproduced the partition function calculations obtained by direct enumeration over all possible non-overlapping conformations. This test validates Equation 4 and Equation 5. We also verified that both series expansions could be used in conjunction with spot penalty optimization (Equation 1 and Equation 2), both by numerically validating the cost function gradient calculation and by testing for convergence on these small problems.

**Improved optimization allows large-scale reconstructions.** Next, we tested whether the iterative spot-penalty optimization rules given by Equation 1 and Equation 2 could work on large-scale problems such as those of Experiment 2, where the old gradient descent optimizer in align3d had difficulty<sup>8</sup>. The results are shown in Figure 4, which compares the number of iterative steps required to converge the  $\bar{q}$  (missing-spot penalty) and  $q$  (spot penalty) parameters without/with use of our improved optimization rules (labeled ‘old’/‘new’ respectively in the legend). Since the spot penalties  $q$  are optimized for probability normalization only after  $\bar{q}$  parameters have been optimized to achieve a desired missing spot frequency, we only attempted to optimize the  $q$  parameters for simulations where  $\bar{q}$  converged. There were two results from this experiment. First, more attempts to optimize the  $\bar{q}$  and  $q$  parameters successfully converged when using the new optimization rules in conjunction with gradient descent, as indicated by the greater volume of the ‘new’



**Figure 4. Comparison of old and new optimization methods.** Each panel compares the number of iterations required to achieve convergence using the old (purple) versus new (yellow) optimization methods. Only trials that successfully converged are counted, so the histograms are not normalized relative to each other. The first number in parentheses of each legend entry shows the number of converged trials, and the second number shows the total number of trials. Note that the second numbers in the right-hand panel equal the first numbers in the left-hand panel, since we required convergence in  $\bar{q}$  in order to attempt optimization of the  $q$  parameters.

histogram and the correspondingly larger numbers shown in the legends. Secondly, of the trials that did converge, our new method required significantly fewer iterations and thus less computation time than the old method, as indicated by the relative skews of the distributions.

#### Use of more colors dramatically improves reconstructions.

Our most striking result is that simulations of the ambitious Experiment 2 produce far better results than even the Toy scenario, despite the fact that these simulations have more loci per color than either the Toy scenario or Experiment 1. This can be seen in the amount of unrecovered information  $I$  shown in the simulation-averaged plots of Figure 5A. High-quality reconstructions using  $\sim 20$  colors were also observed by the ChromoTrace reconstruction method<sup>9</sup> even for large numbers of labeled loci. Our explanation is that the reconstruction quality has more to do with the average spatial density of loci per color than the total number of loci per color, because each ‘propagator’ evolving one potential locus-to-spot mapping to the next sees only the spots within some reasonable radius, as determined by the genomic distance to the next locus. These arguments really pertain to the information recovery of the baseline calculation of  $\tilde{Z}_0$ ; the story is more complicated when better approximating the true  $Z$  which forbids spot reuse between loci, but a simple heuristic is that some average fraction of the competing spots were used earlier along the contour and should thus be removed from consideration. If our reasoning is correct, then reconstructions based on huge numbers of labeled loci (for example whole-genome reconstructions) should be possible as long as the spot density does not get too high.

At the end of this section we revisit Experiment 2, in order to assess the reconstruction quality when analyzing more realistic DNA contours having tighter confinement and thus more closely-packed spots.

**Series expansion 2 outperforms series expansion 1.** Next, we compared the convergence properties of our two expansions on the three scenarios of simulated experiments. Figure 5A gives a sense of how the amount of unrecovered information varies with the number of terms taken in each series, without (solid lines) and with (dotted lines) the use of spot penalties. Each of the 3 panels summarizes all 100 simulated experiments of that scenario, and each experiment in that scenario shows a unique relationship between information recovery and number of series terms computed. Representative curves of individual experiments in each scenario are shown in Supplementary Figure S1 (Supplementary File 1). In order to summarize these very dissimilar curves, Figure 5A shows a median average of all 100 individual experimental curves taken at each data point. Note that this averaging process does not necessarily preserve the shape of the curves from typical individual simulations.

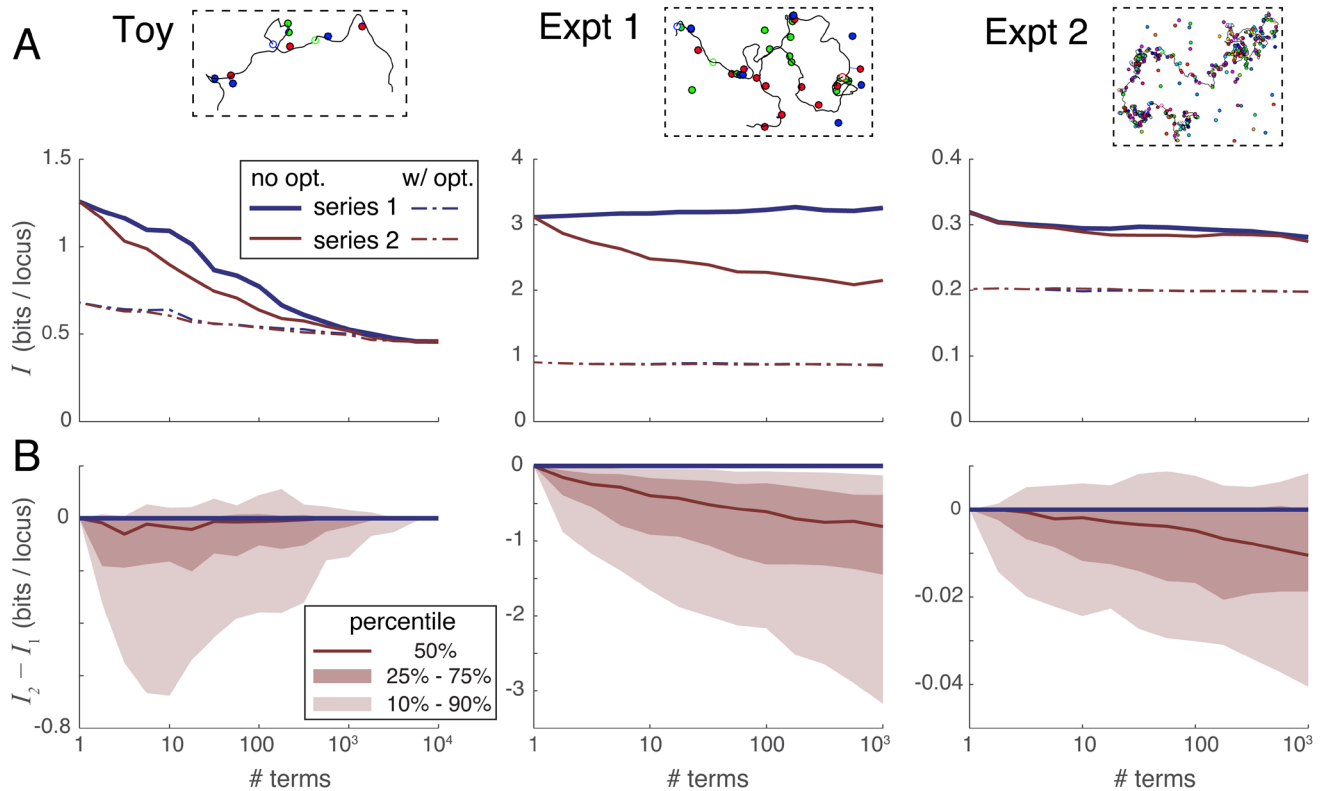
In order to directly compare the two series expansions, we plotted their difference in unrecovered information  $I_2 - I_1$  versus the number of series terms in Figure 5B. In this case, we plotted the full distribution showing the median (50th percentile) as well as the 10th, 25th, 75th and 90th percentile curves. These plots show directly that series 2 almost always outperforms series 1 when only a few terms can be evaluated. The reason is that the terms in series 2 are larger in magnitude owing to their looser constraints, and thus remove the extraneous part of the partition function more quickly than the terms of series 1 (see Supplementary Figure S1 and Supplementary Figure S4, Supplementary File 1). Based on these results, we recommend using series expansion 2 in all situations where the partition function cannot be evaluated exactly.

#### Spot penalty optimization is the most efficient way to recover information.

Spot penalty optimization is an iterative process where each iterative step requires the evaluation of some number of series terms. An optimization requiring  $t$  iterations thus multiplies computation time by a factor of  $t$  relative to the simple evaluation of the series. Alternatively, one could spend the extra computation time on taking the series to a higher order without spot penalty optimization. Figure 6A plots the unrecovered information when a) taking series 2 to a certain order without optimization, versus b) using spot penalty optimization on only the first term yielding  $\tilde{Z}_0^{opt}$ . The dotted line in each panel shows the median number of terms requiring the same computation time as  $\tilde{Z}_0^{opt}$ . The Toy scenario shows that, if the series expansion is carried deep enough, it becomes more accurate than  $\tilde{Z}_0^{opt}$ : in other words the difference  $I - I_0^{opt}$  becomes negative. However, for the practical scenarios of Experiments 1 and 2 this crossover point requires taking more terms than would be needed to match the computational cost of calculating  $\tilde{Z}_0^{opt}$  (the dotted line). Based on these results, we recommend always performing spot penalty optimization, especially for larger reconstructions.

#### Series expansions can improve optimization information recovery.

Although spot penalty optimization is the most efficient way to recover information, that process alone can only extract a certain fraction of the recoverable information: once the cost function is zero, optimization can proceed no further



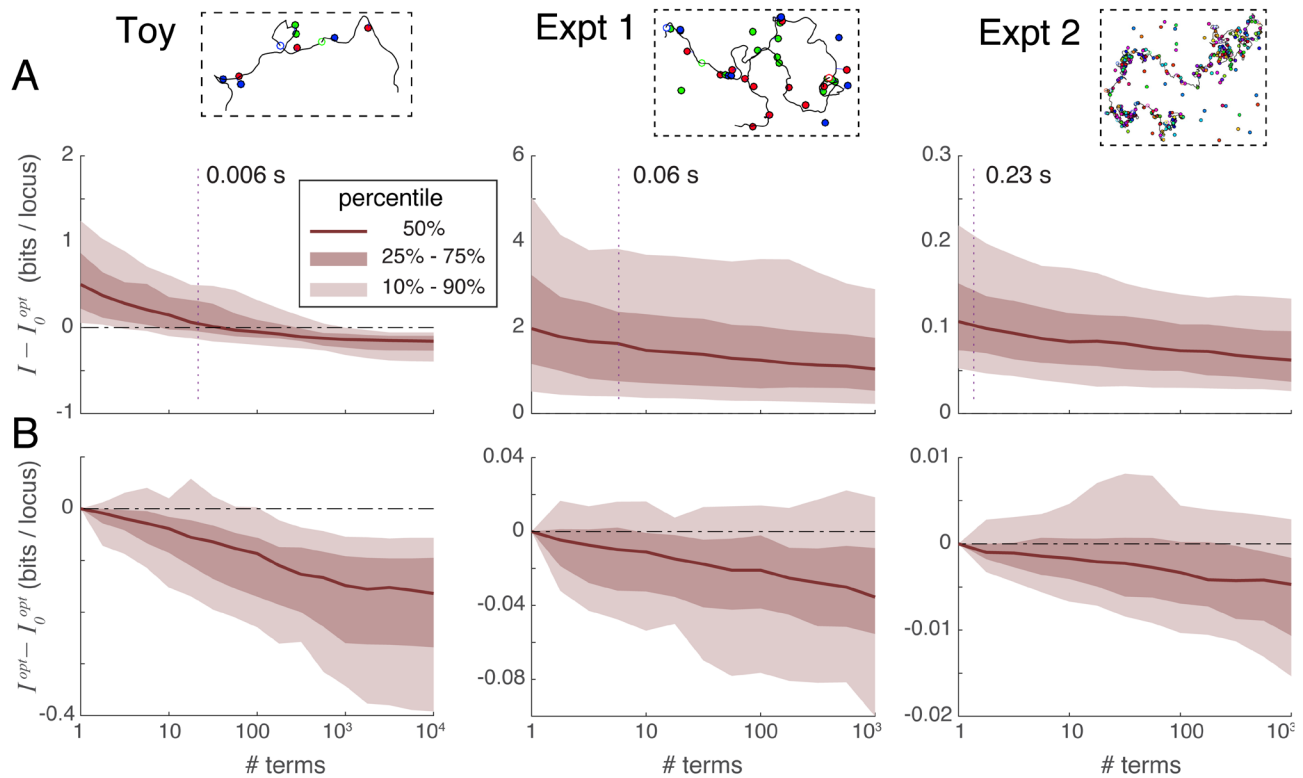
**Figure 5. Comparison of the convergence rates of series expansion 1 and series expansion 2.** **A.** Median unrecovered information  $I$  as a function of the number of terms used in each series expansion, without using spot penalty optimization (solid lines) versus with optimization (dotted lines), and over the three simulation scenarios (panels left-to-right). Each curve was derived from the 100 individual curves corresponding to the 100 simulations in each scenario using a simple point-by-point median average. **B.** Percentile distribution of the difference between the unrecovered information using series 2 minus the unrecovered information using series 1; the fact that this difference quickly drops below zero in nearly all individual simulations shows that series 2 recovers more information in the first few terms than does series 1.

despite the problem not having been solved exactly. At this point, the only way forward is to go higher in the order of series terms used; we can still apply spot penalties to this sum of terms and iteratively optimize them as before using Equation 1 and Equation 2. Figure 6B plots the difference in unrecovered information when applying spot penalty optimization between a) a variable number of terms in series expansion 2, and b) only  $\tilde{Z}_0$  (the first series term). This figure shows that including additional series terms in the optimization improves the information recovery, albeit at a slow rate (especially for large problems).

**20-color labeling leads to near-perfect reconstructions.** As shown in Figure 5A, the unrecovered information for the whole-chromosome Experiment 2 averages around 0.2 bits per locus, implying near perfect mapping probabilities. However, because these results were based on randomly-generated unconfined conformations, they do not establish whether such good information recovery is possible with real chromosomes which are likely to be more compact. To test Experiment 2 on realistic chromosome conformations, we generated four plausible conformations of human chromosome 4 by running the GEM software package<sup>13</sup>

on the smoothed human Hi-C data set provided by Yaffe and Tanay, 2011<sup>14</sup> and using a 3D spline interpolation to increase the resolution from 1 Mb to 50 kb. These conformations were then virtually labeled at 300 randomly-selected loci and simulated experimental error was added in as before. One set of experiments assumed diffraction-limited 100/200 nm localization error in xy/z, and a second set of experiments assumed superresolution 30/50 nm localization error in xy/z; in both sets the missing- and extra-spot rates were 10%. For this experiment we determined the DNA model parameters  $n_i$  and  $l_p$  by fitting pairwise locus distributions, as one would do in an experiment, and for  $L > l_p$  we set  $\rho = 1/3$  as that has been reported in the literature for locus separations under 7 Mb<sup>4</sup>. Mapping probabilities were reconstructed by taking series expansion 2 to the lowest order that included at least 1000 terms, then applying and optimizing spot penalties. Compared with the random-walk conformations used to test the Experiment 2 scenario, the diffraction-limited reconstructions did somewhat worse ( $\sim 0.4$  versus  $\sim 0.2$  bits of unrecovered information per locus) owing to fact that physical confinement of chromosomes increases the density of competing spots in the image. The superresolution reconstruction quality was unchanged at  $\sim 0.2$  bits of unrecovered information.





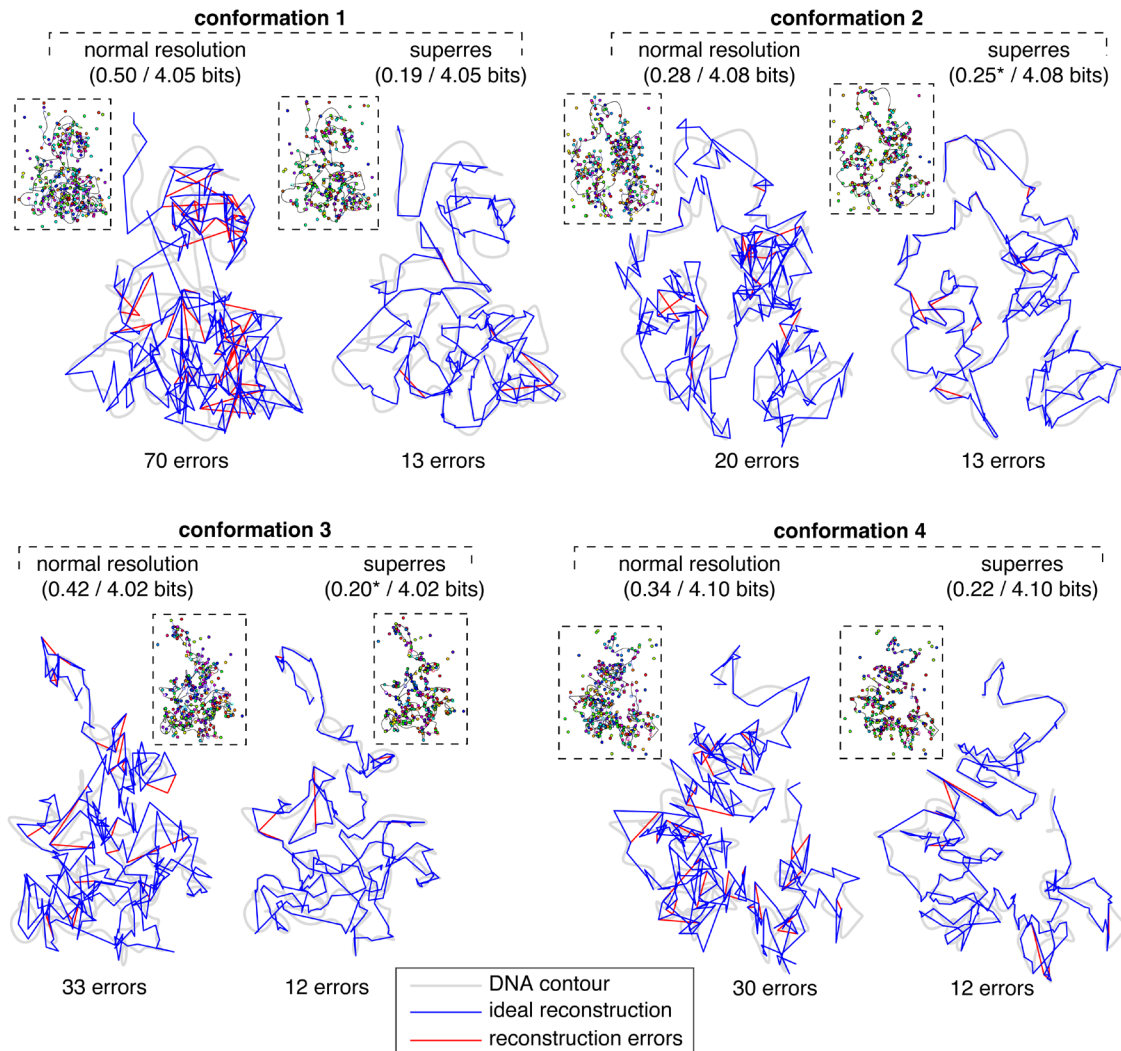
**Figure 6. Optimization in conjunction with series expansions. A.** Comparison of unrecovered information using series expansions without iteration, denoted  $I$ , to the unrecovered information obtained by optimizing spot penalties using only the first series term, denoted  $I_0^{opt}$ , over three experimental situations. Vertical dotted lines indicate the median number of series terms computable with the same computational time as was required to obtain  $I_0^{opt}$ . For Experiments 1 and 2 the difference  $I - I_0^{opt}$  is typically positive at the intersection of the dotted line, indicating that spot penalty optimization method is the more efficient way of recovering information. **B.** Comparison of unrecovered information using spot-penalty optimization in conjunction with multiple series terms versus optimization of  $\tilde{Z}_0$  alone, showing the added benefit of including more terms in the series.

Despite the drop in performance when localizing spots at the diffraction limit, 0.4 bits of unrecovered information per locus is still an extremely strong reconstruction, implying that the correct locus-to-spot mappings are assigned  $p$ -values averaging around  $2^{-0.4} \approx 76\%$ . Starting from such accurate and confident mapping probabilities, one can infer a reasonable conformation simply by assigning each locus to the unassigned spot to which it maps with the highest probability (or calling a missing spot if  $1 - \sum_{s'} p_{L \rightarrow s'} >$  any  $p_{L \rightarrow s'}$ ), repeating the process for overlapping loci, and drawing a line in the image that connects these spots in genomic order. The conformations produced by this simple rule are shown in **Figure 7**: the correct conformation is shown with a blue line and errors in the inferred conformation are shown in red. The reconstructed conformations are  $\sim 90\%$  accurate at diffraction-limited resolution and  $\sim 96\%$  accurate at superresolution, as determined by an alignment between the true and inferred spot sequences traveling along the DNA contour. Most mistakes are of a sort that does not change the large-scale structure. For example, one common error is to erroneously skip one or more spots in the image, thus ‘looping out’ a small part of the conformation and effectively lowering the resolution.

**Figure 7** shows that the benefit of superresolution is two-fold: 1) the locus-to-spot mapping quality improves relative to diffraction-limited resolution (i.e. fewer red lines), and 2) the small-scale structure of an ideal mapping (blue line) more faithfully traces the underlying contour (grey line). This shows the importance of measuring spot locations to sub-pixel resolution, even in experiments where normal-resolution microscopes using standard fluorophores are used to localize spots separated by two pixels or more. In our GEM conformations 23 spots were closer than 200 nm to another spot of the same color, which would indicate problems localizing these spots, but this is inconsistent with the data shown in Wang *et al.*, 2016<sup>4</sup> which indicates that virtually all spots in our experimental scenarios should be well-separated in at least in some cell lines.

### Discussion

We have developed and evaluated two improvements to the align3d method for reconstructing chromosome structure. Both of these improve the partition function estimates that determine the locus-to-spot mapping probabilities, which can provide the basis for (probabilistic) reconstructed conformations.



**Figure 7. Simulated reconstructions of 4 plausible conformations of human chromosome 4.** The left-hand reconstruction of each conformation was obtained using a simulated image from diffraction-limited microscopy (shown in inset; localization error is shown as lines connecting spots to DNA), and the right-hand reconstruction used a simulated superresolution image. Grey shaded lines indicate the underlying DNA contours; blue lines trace the ideal reconstructed contours given the measured spot positions; red lines show our reconstructed contours where they deviate from the ideal contours. Captions above each reconstruction indicate the amount of unrecovered information  $I$  per locus after/before the reconstruction process; captions below indicate the number of alignment errors between the spot ID sequences read along the true versus inferred conformations. For both superresolution reconstructions 2 and 3 we calculated  $I$  excluding a single locus whose true spot mapping was given 0 probability; including that locus sends  $I \rightarrow \infty$ .

The first improvement is a more robust spot-penalty optimizer that allows for large-scale reconstructions involving hundreds of labeled loci, such as will be needed to uncover whole-chromosome conformations. The second improvement is two series expansion formulas for the partition functions, which in principle allow the mapping probabilities to be solved to arbitrary accuracy within the limitations of the experiment and the underlying DNA model. In practice, the series approach is difficult for two reasons: 1) there are a huge number of terms in each series expansion, and 2) the lowest-order approximation  $\tilde{Z}_0$  overestimates  $Z$  by many orders of magnitude, unlike other series expansions where the initial approximation is close to the final answer. Despite the difficulties, the series formulas that we give

offer some way forward to improve on the original estimate  $\tilde{Z}_0^{opt}$ . Of the two formulas, we recommend using series expansion 2, which has the larger-magnitude terms and thus recovers the most information when only a few terms can be evaluated.

Our problem of finding likely (i.e. low-free-energy) DNA conformations passing through a set of imaged spots is similar to the well-known traveling salesman problem (TSP), in which a salesman must find the shortest route connecting a set of cities. Somewhat more closely related is a generalization of the TSP called the time-dependent traveling salesman problem (TDTSP)<sup>10</sup>, where the intercity distances change every step on the tour; this is analogous to our situation where the free energy

needed to thread DNA between two spots depends not only on their separation but also on the length of DNA used to connect them. In our case, the presence of missing and extra spots generalizes our problem still further: in the TDTSP analogy the salesman would be allowed to skip stops and cities for a penalty. Our main finding is that the partition function of this generalized TDTSP (which encompasses traditional TSP and TDTSP problems) can be expressed as a sum of terms computable using a (modified) forward-backward algorithm, a result which should also apply to other route-finding applications where research has historically focused on route optimization rather than route inference.

Both our mapping  $p$ -values and our entropy proxy for information recovery show a systematic bias, which comes from the use of a different DNA model for reconstruction than was used to create the simulated DNA contours. The fact that our reconstructions were nonetheless quite strong shows that the reconstruction method itself is quite robust to model error. This is very fortunate given the uncertainty in the true *in vivo* biological model describing the cells in a real experiment. For our results to be accurate, we had to calibrate our model so as to reproduce the peak in the distance distribution of pairs of distinguishable loci. An experimenter would perform this calibration by imaging distinguishable pairs of loci in a parallel experiment. Due to align3d's use of a very permissive Gaussian chain DNA model, both systematic biases work in the direction of causing the method to underestimate its performance: high  $p$ -values should be higher (and low  $p$ -values lower) than reported, and the unrecovered information tends to be less than the entropy estimate. Thus the results are at least as good as they appear to be.

From a genomic standpoint, our most exciting result is that the combination of our computational improvements together with 20-color labeling technology gives almost perfect reconstructions at the whole-chromosome scale. Out of  $\sim 4$  bits per locus of uncertainty inherent in the reconstruction problem, our method recovers  $\sim 3.6$ – $3.8$  bits. Such confident mapping probabilities allow for the direct construction of individual conformations that are  $\geq 90\%$  accurate. High-quality piecewise reconstructions are likewise possible with two overlapping copies of the same chromosome (data not shown), although sometimes the fragments cannot be assembled. We want to emphasize that our reconstructions require only a few parameters that would be known experimentally with proper controls: the 3 DNA model parameters which in a real scenario would be calibrated using a control experiment, and the correct average rates of missing and extra spots averaged over all experiments, used by align3d to estimate the actual number of missing spots per color in each

experiment. The robustness of the analysis to experimental unknowns gives evidence that reconstructions using real-world experimental data will be of similar quality to those in our simulations, and if so then direct measurement of chromosome conformations is possible today with current technology.

### Data availability

The simulated conformations and labelings used to generate the plots in this paper, together with the output of the align3d analysis, can be found at: <https://github.com/heltilda/align3d/blob/master/seriesExpansions/a3dRawData.zip>

### Software availability

Results in this paper were generated using version 1.1 of align3d, built using version 1.1 of Cicada scripting language. Simulated conformations and labelings were generated using version 1.1 of wormulator.

All source files used in preparing this paper are available from the GitHub page for this paper: <https://github.com/heltilda/align3d/tree/master/seriesExpansions>.

License: GPL 3.0

Archived code at time of publication:

align3d: <https://doi.org/10.5281/zenodo.2580342><sup>15</sup>

License: GPL 3.0

wormulator: <https://doi.org/10.5281/zenodo.1411503><sup>16</sup>

License: GPL 3.0

Cicada scripting language: <https://doi.org/10.5281/zenodo.1411505><sup>17</sup>

License: MIT License

---

### Grant information

Funding was provided by the Boettcher Foundation (J.C.C.), National Institutes of Health2 [T15LM009451 to B.C.R.], and a Cancer League of Colorado grant (J.C.C. and B.C.R.).

*The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.*

### Acknowledgements

The authors want to thank Rani Powers and Jenny Mae Samson for helping review the manuscript.

### Supplementary material

Supplementary File 1: align3dSupplement.pdf. File containing three appendices giving the derivations of the equations used in this text (Appendix 1 and Appendix 2), a discussion of model error (Appendix 3), and the supplemental figures (Appendix 4).

[Click here to access the data](#)

## References

1. Dekker J, Belmont AS, Guttman M, *et al.*: **The 4D nucleome project.** *Nature.* 2017; **549**(7671): 219–226.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
2. Imakaev MV, Fudenberg G, Mirny LA: **Modeling chromosomes: Beyond pretty pictures.** *FEBS Lett.* 2015; **589**(20 Pt A): 3031–3036.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
3. Kocanova S, Goiffon I, Bystricky K: **3D FISH to analyse gene domain-specific chromatin re-modeling in human cancer cell lines.** *Methods.* 2018; **142**: 3–15.  
[PubMed Abstract](#) | [Publisher Full Text](#)
4. Wang S, Su JH, Beliveau BJ, *et al.*: **Spatial organization of chromatin domains and compartments in single chromosomes.** *Science.* 2016; **353**(6299): 598–602.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
5. Takei T, Shah S, Harvey S, *et al.*: **Multiplexed Dynamic Imaging of Genomic Loci by Combined CRISPR Imaging and DNA Sequential FISH.** *Biophys J.* 2017; **112**(9): 1773–1776.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
6. Ma H, Tu LC, Naseri A, *et al.*: **Multiplexed labeling of genomic loci with dCas9 and engineered sgRNAs using CRISPRainbow.** *Nat Biotechnol.* 2016; **34**(5): 528–30.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
7. Lowenstein MG, Goddard TD, Sedat JW: **Long-range interphase chromosome organization in *Drosophila*: a study using color barcoded fluorescence *in situ* hybridization and structural clustering analysis.** *Mol Biol Cell.* 2004; **15**(12): 5678–5692.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
8. Ross BC, Wiggins PA: **Measuring chromosome conformation with degenerate labels.** *Phys Rev E Stat Nonlin Soft Matter Phys.* 2012; **86**(1 Pt 1): 011918.  
[PubMed Abstract](#) | [Publisher Full Text](#)
9. Barton C, Morganello S, Oedegaard O, *et al.*: **Chromotrace: Reconstruction of 3D chromosome configurations by super-resolution microscopy.** *bioRxiv.* 2017; 115436.  
[Publisher Full Text](#)
10. Gouveia L, Voß S: **A classification of formulations for the (time-dependent) traveling salesman problem.** *Eur J Oper Res.* 1995; **83**(1): 69–82.  
[Publisher Full Text](#)
11. Baum L: **An inequality and associated maximization technique in statistical estimation of probabilistic functions of a markov process.** *Inequalities.* 1972; **3**: 1–8.  
[Reference Source](#)
12. Trask B, Pinkel D, van den Engh G: **The proximity of DNA sequences in interphase cell nuclei is correlated to genomic distance and permits ordering of cosmids spanning 250 kilobase pairs.** *Genomics.* 1989; **5**(4): 710–717.  
[PubMed Abstract](#) | [Publisher Full Text](#)
13. Zhu G, Deng W, Hu H, *et al.*: **Reconstructing spatial organizations of chromosomes through manifold learning.** *Nucleic Acids Res.* 2018; **46**(8): e50.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
14. Yaffe E, Tanay A: **Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture.** *Nat Genet.* 2011; **43**(11): 1059–65.  
[PubMed Abstract](#) | [Publisher Full Text](#)
15. heltilda: **heltilda/align3d: Final version of align3d incorporating series formulas (Version 1.1.1).** *Zenodo.* 2019.  
<http://www.doi.org/10.5281/zenodo.2580342>
16. heltilda: **heltilda/wormulator: wormulator version for paper (Version 1.1).** *Zenodo.* 2018.  
<http://www.doi.org/10.5281/zenodo.1411503>
17. heltilda: **heltilda/cicada: Cicada version used in F1000Research paper (Version 1.1).** *Zenodo.* 2018.  
<http://www.doi.org/10.5281/zenodo.1411505>

# Open Peer Review

Current Peer Review Status:  

---

## Version 3

Reviewer Report 19 July 2019


<https://doi.org/10.5256/f1000research.20533.r46353>

© 2019 Lagomarsino M et al. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution Licence](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



**Marco Cosentino Lagomarsino** 

Génophysique/Genomic Physics Group, UMR 7238, CNRS (French National Centre for Scientific Research), Genomics of Microorganisms, Pierre and Marie Curie University (UPMC), Paris, France

**Vittore Scolari** 

Pasteur Institute, Paris, France

We have now both read the replies and the new manuscript and we are satisfied.

**Competing Interests:** No competing interests were disclosed.

**We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.**

---

## Version 2

Reviewer Report 27 March 2019

<https://doi.org/10.5256/f1000research.20223.r45547>

© 2019 Birney E et al. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution Licence](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



**Ewan Birney** 

European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Cambridge, UK

**Carl Barton**

European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Cambridge, UK

In my initial review of the paper "Improved inference of chromosome conformation from images of labeled loci" a few concerns were raised. My comments mainly related to putting the current work of the authors in better context with respect to the previous work as well as explaining a number of the details mentioned in the paper in more detail.

I believe the authors have satisfactorily addressed my comments. They have added an appendix giving more information on the convergence properties of the used equations. They have addressed differences between align3d and previous work explaining how they are quite different approaches.

Some concerns were also raised relating to the effect of the number of colors on reconstruction quality and they have added another paragraph explaining in more detail why this effect is seen. In addition to this it seems they have addressed or responded to all of the other reviewer comments.

We are happy to approve indexing.

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** genomics

**We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.**

---

### Version 1

Reviewer Report 10 December 2018

<https://doi.org/10.5256/f1000research.17750.r40510>

© 2018 Lagomarsino M et al. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution Licence](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



**Marco Cosentino Lagomarsino**

Génophysique/Genomic Physics Group, UMR 7238, CNRS (French National Centre for Scientific Research), Genomics of Microorganisms, Pierre and Marie Curie University (UPMC), Paris, France

**Vittore Scolari**

Pasteur Institute, Paris, France

In the manuscript "Improved inference of chromosome conformation from images of labeled loci", Ross and Costello present a computational inference method to reconstruct genome conformation from measurements of the positions of  $m$  labelled loci of known coordinates with  $n \ll m$  colored fluorescent foci. The presented tool is a new version of their computational tool "align3d" with multiple improvements. The tool has the aim of inferring the polymer conformation of chromosomes in-vivo, starting from images of fluorescently-tagged genomic loci, where each color tags different loci at the same time.

The authors provide a test of the algorithm with data that are generated computationally, in a simple (short polymers) or more complex (longer polymers, more colors) setting. Finally they provide a simplified (and

limited - see point 5 below) test using data that are derived from empirical data.

The question appears interesting due to its experimental motivation, although probably the method is not yet close to something with concrete applicability to experimental data. We think that this could develop into a useful tool for the global effort of understanding the chromosome conformation of organisms in-vivo. As physicists, we are concerned with some aspects related to the representation (modelling) of the polymer and the experimental situation. Our observations might be useful for the authors or for other scientists that intend to analyse this kind of experimental data.

1. The authors state that the inaccuracy of the DNA (conformation) model, i.e. how the physical distance of two loci scales with arclength distance along the genomic coordinate is a major factor of error (more precisely, this is a conditional distribution of distances given distance along the chain). They further state that nothing is known about this. However, this is not really the case, as both Hi-C and FiSH experiments with labelled loci give information about these quantities (Lagomarsino *et al.*, 2015<sup>1</sup> and Fudenberg and Mirny, 2012<sup>2</sup>).

In particular, the assumption that the polymer is a Gaussian chain seems very restrictive. A much less restrictive (though still limited) assumption would be that this scaling relation is a tuneable power-law. This assumption is particularly interesting because in this case the scaling law relating physical distance to distance along the genome is related to the contact probability measured in Hi-C data (Fudenberg and Mirny, 2012<sup>2</sup>). Indeed, in this scenario the contact probability (sometimes called “P(s)”, where s is the arclength distance) and the connection between genomic distance and typical spatial distance R(s) are related by a scaling (Polovnikov *et al.*, 2018<sup>3</sup>). Thus Hi-C data could be used to directly constrain the inference, or to compare with the results.

In this last scenario one could use the inference to learn the scaling from data. It seems quite reasonable to us that this scaling should be one of the main observables to infer from the data. Imposing this scaling appears like imposing a specific behaviour on the configurations that we are attempting to infer. In this regard, one big question is whether the observable “scaling of physical distance with arclength distance” can be inferred from the data without making the problem under-determined. We would like to stimulate the authors to spend some words to address this question.

As we suggest above, there are multiple possible approaches to this practical issue, such as the use of the observable quantity “P(s)”, the contact probability measured with Hi-C, or the use of an ansatz, such as a power law (Marie-Nelly *et al.*, 2014<sup>4</sup>), accompanied by a procedure to optimize the parameters.

2. The authors’ main hypothesis is that only one locus can map to each identified spot in the image, and, for this reason, the solution proposed is a heuristic method to solve the traveling salesman problem for the polymer on those loci. We observe that this might be a good practical assumption but it is not necessarily a good one for the chromosome, and for polymers in general. Polymers can have loops, even randomly. The definition of those loops depends on the resolution of observation (which experimentally will be limited by diffraction). The frequency of loops in chromosomes depends on important physical and biological parameters such as active looping (Fudenberg *et al.*, 2016<sup>5</sup>), the presence of different solvent phases and the balance between steric and other kinds of interactions (Scolari and Lagomarsino, 2015<sup>6</sup>) as well as from steps of the experimental protocols

(Scolari *et al.*, 2018<sup>7</sup>). Hi-C experiments, measure loops and quantify their specific and generic properties. In terms of the genomic distance, it has been shown that at small distances the chromosomes are very compact, and the amount of this compaction varies widely across conditions (Lazar-Stefanita *et al.*, 2017<sup>8</sup> and Muller *et al.*, 2018<sup>9</sup>) even for the same organism. For increasingly longer distances, generally, the probability of making a loop normally decreases monotonically with genomic distance. Thus, we think that the authors' approach should be applicable to an increasing number of cases by increasing the scale of observation and modelling, under the condition that the relation that ties the genomic distance to the three-dimensional distance is chosen correctly.

3. The algorithm is focused on a single chain conformation and does not exploit ensembles. Typically in such experiments one expects to have fairly low precision of localisation, but almost arbitrarily large amount of realisations (different cells). Each will be different but will also have common properties, and relaxing the question could make the inference process much easier. After all, inferring precisely a single configuration is not so relevant, because it will change in time due to natural fluctuations of the system. It is more useful (and well defined) to infer some ensemble properties (at fixed conditions for the cells such as time and phase into the cell cycle), and then quantify the cell-to-cell diversity with respect to such average behavior.
4. These images will come from microscopy and they will likely be 2D projections, or have lower resolution in the z direction. The authors do not address this issue (and in general the issue of resolution seems underestimated), but we expect it to be quite important in any concrete situation.
5. In regard to the final example, we notice that the data is binned at 1mb and then interpolated at 100kb with a spline, we wonder if this resolution improvement introduces any alterations in the reconstructed conformations of the polymer. For this reason, it seems reasonable to perform a more thoughtful statistical analysis with different levels of interpolation to support this choice.

## References

1. Lagomarsino MC, Espéli O, Junier I: From structure to function of bacterial chromosomes: Evolutionary perspectives and ideas for new experiments. *FEBS Lett.* 2015; **589** (20 Pt A): 2996-3004 [PubMed Abstract](#) | [Publisher Full Text](#)
2. Fudenberg G, Mirny LA: Higher-order chromatin structure: bridging physics and biology. *Curr Opin Genet Dev.* 2012; **22** (2): 115-24 [PubMed Abstract](#) | [Publisher Full Text](#)
3. Polovnikov KE, Gherardi M, Cosentino-Lagomarsino M, Tamm MV: Fractal Folding and Medium Viscoelasticity Contribute Jointly to Chromosome Dynamics. *Phys Rev Lett.* 2018; **120** (8): 088101 [PubMed Abstract](#) | [Publisher Full Text](#)
4. Marie-Nelly H, Marbouty M, Cournac A, Flot JF, Liti G, Parodi DP, Syan S, Guillén N, Margeot A, Zimmer C, Koszul R: High-quality genome (re)assembly using chromosomal contact data. *Nat Commun.* 2014; **5**: 5695 [PubMed Abstract](#) | [Publisher Full Text](#)
5. Fudenberg G, Imakaev M, Lu C, Goloborodko A, Abdennur N, Mirny L: Formation of Chromosomal Domains by Loop Extrusion. *Cell Reports.* 2016; **15** (9): 2038-2049 [Publisher Full Text](#)
6. Scolari VF, Cosentino Lagomarsino M: Combined collapse by bridging and self-adhesion in a prototypical polymer model inspired by the bacterial nucleoid. *Soft Matter.* 2015; **11** (9): 1677-87 [PubMed Abstract](#) | [Publisher Full Text](#)



7. Scolari VF, Mercy G, Koszul R, Lesne A, Mozziconacci J: Kinetic Signature of Cooperativity in the Irreversible Collapse of a Polymer. *Phys Rev Lett*. 2018; **121** (5): 057801 [PubMed Abstract](#) | [Publisher Full Text](#)
8. LazarStefanita L, Scolari V, Mercy G, Muller H, Guérin T, Thierry A, Mozziconacci J, Koszul R: Cohesins and condensins orchestrate the 4D dynamics of yeast chromosomes during the cell cycle. *The EMBO Journal*. 2017; **36** (18): 2684-2697 [Publisher Full Text](#)
9. Muller H, Scolari VF, Agier N, Piazza A, Thierry A, Mercy G, Descorps-Declere S, Lazar-Stefanita L, Espeli O, Llorente B, Fischer G, Mozziconacci J, Koszul R: Characterizing meiotic chromosomes' structure and pairing using a designer sequence optimized for Hi-C. *Mol Syst Biol*. 2018; **14** (7): e8293 [PubMed Abstract](#) | [Publisher Full Text](#)

**Is the rationale for developing the new method (or application) clearly explained?**

Partly

**Is the description of the method technically sound?**

Yes

**Are sufficient details provided to allow replication of the method development and its use by others?**

Partly

**If any results are presented, are all the source data underlying the results available to ensure full reproducibility?**

Partly

**Are the conclusions about the method and its performance adequately supported by the findings presented in the article?**

Partly

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** Statistical Physics, Quantitative Biology, Chromosome Organization

**We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however we have significant reservations, as outlined above.**

Author Response 11 Mar 2019

**Brian Ross**, University of Colorado, Anschutz Medical Campus, Aurora, USA

We wish to thank Reviewers 2 for their many helpful comments and insights. To address their comments as well as several concerns of our own, we have made a number of changes to our analysis, our results and the content of the main paper, and added a new appendix. The changes made to the code required us to regenerate all 9 figures that show results, although only Figure 7 changed significantly. We have also updated our GitHub repository containing all our code and example data.

Reviewers 2 pointed out that we were too strong in our language stating that 'nothing is known'

about the DNA model. We agree and we have removed this wording from the last paragraph in the Introduction. They also suggest the use of contact probabilities as a basis for inferring the distance function used in the model. This would be possible, but there is actually a direct measurement of the distance function (Wang *et al.*, 2016) which we have decided to use instead. Wang *et al.* found a 1/3 scaling exponent between L and R at the relevant length range, which we incorporated into our model for analyzing the Hi-C reconstruction (but not the random chains, whose exponent is 1/2) and then reran the results shown in Figure 7. The results did not change very much because most adjacent loci are close enough that  $R = k \cdot L$ , i.e. the exponent is 1. The reviewers point out this could be due to our spline interpolation used to increase the resolution of the conformation. Unfortunately we were not able to directly infer higher-resolution conformations because the Hi-C data set recommended by Zhu *et al.*, 2018 (the Hi-C inference method called 'GEM') is binned at 1 Mb resolution, and this sets the resolution of the GEM conformations. We do not believe this is a problem for 2 reasons: 1) our average label spacing is  $\sim 2/3$  Mb, not far below the 1 Mb GEM conformation subunit length, and 2) the inferred conformations seem to have a persistence length somewhat above the length of a subunit, although we cannot rule out that this might change with a different bin density. We agree that it would be ideal to have a higher-resolution structure (although bin sampling would become an issue using this Hi-C data set), but we suspect that the errors in Hi-C inferences probably overwhelm the resolution issue.

We want to point out that our use of a Gaussian chain for reconstruction (but not for producing the DNA chains) is not incompatible with the scaling relations mentioned by Reviewer 2, because these scaling relations determine average spatial separation R of two loci based on their genomic separation L, but not the form of the distribution  $p(r|R)$ . We have chosen to model  $p(r|R)$  with a Gaussian (partly for convenience since it is easy to factor in localization error, but partly for other reasons; see below) having  $\sigma = R^2$ , but R is in turn calculated as  $R = L^{\text{power}}$ . In our earlier draft this power was fixed at 1/2 for distances above a persistence length, but as Reviewer 2 pointed out recent experimental data show exponents of 1/3 - 1/5. To address this issue, we generalized our program to accept more general DNA models consisting of different power laws at different inter-locus distance regimes, and our new results use exponents of either 1/2 and 1/3 for long DNA segments, depending on the simulated experiment. To make this clearer, we have added a new paragraph to the Results section (2nd paragraph) explaining the model selection in our simulations, as well as how a model would be chosen in a real experiment.

In our initial submission we claimed that a systematic error seen in the mapping probabilities was due to overestimation of the missing-spot rate. Since then we have both fixed the missing-spot rate estimation and made major progress in figuring out the real cause of the error, which we explain in detail in a new Appendix 3 and refer to in several places in the text. The error comes from the fact that the Gaussian chain model used for reconstruction differs significantly from the wormlike chain model used to generate our simulated contours. While Reviewer 2 was concerned that the use of a 'wrong' model would skew the results, we believe that the opposite interpretation is more accurate: the fact that we obtain high-quality results even when the reconstruction model differs from the model used to produce the conformations shows that our approach is robust to model error. Appendix 3 justifies this intuition, by showing that model error causes our results to appear less certain than they are, but does not cause reconstruction errors if the reconstruction model is less sharply-peaked than the true model. This is the other justification for using the Gaussian chain model, which is quite permissive of unexpected behavior that we may find given that the true *in-vivo* DNA model may behave unexpectedly sometimes which may be very difficult to measure exactly in calibration experiments. We have also added a new 3rd paragraph to the Discussion explaining this.

We agree with Reviewers 2 that loops certainly can happen and, to the extent that they can be distinguished by microscopy, our algorithm is certainly capable of finding looped conformations (even if the loop is over 2 adjacent loci -- our Gaussian model peaks at  $r = 0$ ). If two loci of the same color happen to overlap in a microscope image, one may be missed -- this is considered a missing-spot or false-negative error, as mentioned in the footnote in the Introduction. Since our algorithm is capable of handling both false negatives and false positives (extra unbound spots), we do not anticipate loops to be a problem. If there are many points of overlap coming from an identical color sequence (e.g. if two copies of the same chromosome overlap) then the reconstruction fragments can become fragmented, with ambiguity as to which piece goes with which other piece -- we have added a brief note about this to the Discussion section.

Reviewers 2 point out that align3d is a single-cell method, not an ensemble method, and we completely agree. We believe that aggregating single-cell conformations will give many interesting insights that one could not get by aggregating, for example, pairwise distances. Our method should be seen as one possible means of obtaining these cell conformations.

Finally, Reviewers 2 raise the issue of resolution in the z dimension: we certainly do consider localization error in z, both in generating the spot localizations (which have z error as well as x/y error) and in the reconstructions (where the errors in x/y/z are required inputs). In all simulations we set the z localization error higher than x/y error (200 vs 100 nm in normal resolution, 50 vs 30 nm in superresolution), reflecting the fact that axial resolution is worse in most setups. We have updated the main text to more explicitly give the localization error in the various simulations.

#### References:

Wang, S., Su, J. H., Beliveau, B. J., Bintu, B., Moffitt, J. R., Wu, C. T., & Zhuang, X. (2016). Spatial organization of chromatin domains and compartments in single chromosomes. *Science*, 353 (6299), 598-602.

Zhu, G., Deng, W., Hu, H., Ma, R., Zhang, S., Yang, J., ... & Zeng, J. (2018). Reconstructing spatial organizations of chromosomes through manifold learning. *Nucleic acids research*, 46(8), e50-e50.

**Competing Interests:** No competing interests were disclosed.

Reviewer Report 07 November 2018

<https://doi.org/10.5256/f1000research.17750.r38643>

© 2018 Birney E et al. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution Licence](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



**Ewan Birney**

European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Cambridge, UK

**Carl Barton**

European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Cambridge, UK

In this paper the authors update and build on a method they have previously published known as 'align3d'. This method attempts to infer the chromosome conformation based on images of fluorescently tagged genomic loci. The authors claim that this updated method increases the accuracy of the inferred conformation as well as allowing the method to run on larger instances of the problem. They then go on to demonstrate where the method allows for the near perfect reconstruction of larger scale, simulated, labelled images. We believe that the article is worthy of indexing on the condition that some minor issues, outlined below, are addressed.

In the introduction the authors mention a couple of other methods attempting to resolve similar problems. I think that this section should be expanded as there is no critical comparison of how this method compares to each of those mentioned. In particular the computational methods should be compared and contrasted so it is clear to the reader how this method differs from others.

Whilst reviewing this paper we were unable to access the supplementary data. This must be made available before the paper can be indexed. Some things the authors could include in the supplementary section that would be useful from the computational perspective would be the type of series expansion being used and any information on how quickly the series expansion converge to the original formula. The authors have experimentally checked on the convergence properties but sometimes it's quite simple to determine theoretically how quickly some approximation converges. This information could be useful in determining better expansions and would explain more concretely why they get some of the results they see.

In the experimental section of the paper the authors generate three different types of simulation that they denote 'Toy', 'Experiment 1' and 'Experiment 2'. In the discussion of the results the authors make the following comment:

'Use of more colors dramatically improves reconstructions. Our most striking result is that simulations of the ambitious Experiment 2 produce far better results than even the Toy scenario, despite the fact that these simulations have more loci per color than either the Toy scenario or Experiment 1. This can be seen in the amount of unrecovered information shown in the simulation-averaged plots of Figure 5A. Thus a push to 20-color labeling could prove critical for genomic reconstruction at the chromosome scale and beyond. At the end of this section we revisit Experiment 2, in order to assess the reconstruction quality when analyzing more realistic DNA contours having tighter confinement.'

The authors should make some attempt to explain this situation. Actually if you increase the number of colours and also increase the number of loci with the same colour then you would not obviously assume that the problem should be harder. It very much depends on how each is increased within proportion to each other.

An increase in the number of unique colours available should lead to the problem being exponentially easier as you are effectively exponentially decreasing the ambiguity in the data set. Should you also increase the number of loci labelled with the same colour then you wouldn't expect the problem to become harder unless that increase was large enough to outweigh the effects of the increase in the number of available colours. In this sense it could be argued that many instances of the 'Toy' example are fundamentally more challenging than the (on the face of it) more complicated 'Experiment 2'. This should somehow be addressed by the authors.

Also in the discussion of the experimental results the authors note that '20-color labelling leads to near-perfect reconstructions.' This result is consistent with our results reported by Barton *et al.* (2017<sup>1</sup>). It

would be good to mention this as although the computational methods are different, the simulations are generated in different ways and the resolution simulated is different, both methods suggest that if ~20 colours are available then near perfect reconstruction is possible. The authors should also point out the similarity of the number of colours needed in both their method and ours.

The differences in computational methodology yet similarity in the numbers of colours needed for near perfect reconstruction perhaps suggests to me that both methods are in some sense 'naive'. There must exist a minimum number of colours required for a certain average reconstruction performance (with the appropriate caveats) but we would be surprised if it was as high as 20. It could be interesting to see the authors add some discussion about this connection and any insight they might have into it.

Finally there have been a number of different attempts to simulate super resolved images of the type used in this and other computational methods. If the authors can use this data as input or the data can easily be coerced into an appropriate format for this method then the paper would be much stronger with the addition of results of using the method against these datasets. In this way the authors can clearly demonstrate that the method they propose is not simply good on their own simulated data, but also performs robustly on other independently generated simulations.

### References

1. Barton C, Morganella S, Ødegård-Fougner Ø, Alexander S, Ries J, Fitzgerald T, Ellenberg J, Birney E: ChromoTrace: Computational reconstruction of 3D chromosome configurations for super-resolution microscopy. *PLoS Comput Biol.* 14 (3): e1006002 [PubMed Abstract](#) | [Publisher Full Text](#)

### Is the rationale for developing the new method (or application) clearly explained?

Yes

### Is the description of the method technically sound?

Partly

### Are sufficient details provided to allow replication of the method development and its use by others?

Partly

### If any results are presented, are all the source data underlying the results available to ensure full reproducibility?

Partly

### Are the conclusions about the method and its performance adequately supported by the findings presented in the article?

Partly

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** genomics

**We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however we have significant reservations, as outlined above.**

Author Response 11 Mar 2019

**Brian Ross**, University of Colorado, Anschutz Medical Campus, Aurora, USA

We wish to thank Reviewers 1 for their many helpful comments and insights. Based on these comments and those of Reviewers 2, we have made some changes to our analysis, updated our results (particularly those shown in Figure 7) and the content of the main paper, and added a new appendix.

We apologize for the problems Reviewers 1 had in accessing the Supplemental Material. The material was uploaded and available (to us), but the 'Appendix x' links lead nowhere in the published version. These seemingly dead links have been removed.

Reviewers 1 suggested that in the Introduction we compare our align3d method to the other published method that we are aware of (ChromoTrace) to highlight their differences. We agree that this is indeed a useful addition, and so we have added several sentences to the Introduction (2nd paragraph) contrasting the two algorithms. We are not experts on ChromoTrace, and if we have mischaracterized it in some way we apologize and hope the reviewers will correct us.

Reviewers 1 also inquired about the exact series expansion formulas we used. The expansion formulas are in the Methods section of the main text, not the Supplemental material. To make this clearer we have added an equation number to the series definition preceding the coefficient formulas (this is the new Equation 3), and referenced that equation explicitly in the two coefficient formulas (which are now Equations 4-5). Thus the series definitions are fully in the main body of the paper, and only their derivations are in Appendix 2.

One technical detail is that our original code could not use our series expansions in conjunction with the preexisting capability to 'fix' certain loci to map to certain spots in the image, in order to obtain mapping probabilities that are conditional on the fixed loci. This has been addressed in the new version of the code. This oversight did not affect the results shown in the paper, but it did require us to add an explanatory paragraph to the end of Appendix 2.

Reviewers 1 asked about the finding that our simulated Experiment 2 reconstructions came out much better than the Experiment 1 reconstructions, despite having more labeled loci of a given color. We have added several sentences to the Results section ("Use of more colors dramatically improves reconstructions" section) explaining that we believe that it is the spatial density of labeled loci rather than the absolute number that determines the reconstruction quality. Reviewers 1 noticed the same finding in Barton *et al.* (2018); we have added this citation. We have not systematically tested performance as a function of the number of colors; we chose 20 simply based on the fact that 10 sequential hybridizations is reasonable for our planned experiment based on conversations with our collaborator (Wang *et al.*, 2016, demonstrate 17 rounds). Since we haven't noticed a plateau in reconstruction performance versus number of colors, as evidenced by the fact that the 20-color reconstructions still have some uncertainty, we do not see a reason to go towards fewer colors.

A final question raised by Reviewers 1 concerned the issue of superresolution in the simulated images. Since our spots are presumed well-separated (based on the data of Wang *et al.*, 2016) we believe we can get super-resolved spot localization without having to use special microscopes or fluorophores, and without having to resolve individual fluorophores. Thus the superresolution comes for free on normal images at the scale we consider here. We have added text explaining

this (new final paragraph of Results), and also a second set of conformational reconstructions to Figure 7 showing explicitly the benefit of superresolving the spot locations. If we were to push to higher-genomic-resolution labeling (say, 10s-100s kb locus separation; current simulations are at ~600 kb) then we would indeed need superresolution microscopes, but since those are not the experiments simulated here we did not try to simulate those images. In fact this is why we chose to label these simulations at the 600 kb resolution.

Although we were not able to increase the Hi-C inferred resolution, we did discover that we had misinterpreted the scale of the Hi-C-derived conformations of Figure 7, thus underestimating the relative magnitude of microscope error in these simulations. Our new plots have corrected this error. Owing to the larger microscope error our new reconstruction quality is somewhat worse as measured by our information metric. To compensate we improved our script that estimates a likely conformation from our output (mapping p-values), and as a result these likely conformations are roughly of the same quality as before. We also added a parallel set of superresolution reconstructions to this figure, in order to show explicitly the benefit of reducing microscope error.

#### References:

Barton, C., Morganella, S., Oedegaard, O., Alexander, S., Ries, J., Fitzgerald, T., ... & Birney, E. (2018). Chromotrace: reconstruction of 3D chromosome configurations by super-resolution microscopy. *bioRxiv*, 115436.

Wang, S., Su, J. H., Beliveau, B. J., Bintu, B., Moffitt, J. R., Wu, C. T., & Zhuang, X. (2016). Spatial organization of chromatin domains and compartments in single chromosomes. *Science*, 353 (6299), 598-602.

Zhu, G., Deng, W., Hu, H., Ma, R., Zhang, S., Yang, J., ... & Zeng, J. (2018). Reconstructing spatial organizations of chromosomes through manifold learning. *Nucleic acids research*, 46(8), e50-e50.

**Competing Interests:** No competing interests were disclosed.

---

## Comments on this article

### Version 1

Reader Comment 11 Nov 2018

**Peter Rogan**, University of Western Ontario, Canada

Having had the opportunity to recently see Oligopaint data and speak with a number of practitioners of this technique, reproducibility remains a significant issue. The simulations described in this paper cannot be practically applied for chromosome conformation analysis until convincing data are obtained:

"The robustness of the analysis to experimental unknowns gives evidence that reconstructions using real-world experimental data will be of similar quality to those in our simulations, and if so then direct measurement of chromosome conformations is possible today with current technology."

FISH using short single copy DNA probes is the underlying basis of this technique ([PMID 11381034](#)). We successfully labeled single copy oligonucleotides by ligation or hybridization of fluorescent detection

reagents, but it was technically challenging (PMID 16460913). Considerable effort is required to ensure low fluorescence background for clinical applications of single copy probes (PMID 12923866), because Cot-1 repeat blocking DNA can produce artifacts in short probe hybridization (PMID 23376933). Expertise in human cytogenetic identification of hybridized chromosomes (e.g. reverse DAPI banding patterns) would also to convincingly demonstrate accuracy of Oligopaint results. .

**Competing Interests:** Founder of CytoGnomix. The company holds patents and markets single copy probes.

---

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact [research@f1000.com](mailto:research@f1000.com)

F1000Research