Simulations of knotting of DNA during genome mapping

Aashish Jain and Kevin D. Dorfman^{a)}

Department of Chemical Engineering and Materials Science, University of Minnesota—Twin Cities, 421 Washington Ave. SE, Minneapolis, Minnesota 55455, USA

(Received 3 February 2017; accepted 21 March 2017; published online 11 April 2017)

Genome mapping involves the confinement of long DNA molecules, in excess of 150 kilobase pairs, in nanochannels near the circa 50 nm persistence length of DNA. The fidelity of the map relies on the assumption that the DNA is linearized by channel confinement, which assumes the absence of knots. We have computed the probability of forming different knot types and the size of these knots for long chains (approximately 164 kilobase pairs) via pruned-enriched Rosenbluth method simulations of a discrete wormlike chain model of DNA in channel sizes ranging from 35 nm to 60 nm. Compared to prior simulations of short DNA in similar confinement, these long molecules exhibit both complex knots, with up to seven crossings, and multiple knots per chain. The knotting probability is a very strong function of channel size, ranging from 0.3% to 60%, and rationalized in the context of Odijk's theory for confined semiflexible chains. Overall, the knotting probability and knot size obtained from these equilibrium measurements are not consistent with experimental measurements of the properties of anomalously bright regions along the DNA backbone during genome mapping experiments. This result suggests that these events in experiments are either knots formed during the processing of the DNA prior to injection into the nanochannel or regions of locally high DNA concentration without a topological constraint. If so, knots during genome mapping are not an intrinsic problem for genome mapping technology. *Published by AIP Publishing.* [<http://dx.doi.org/10.1063/1.4979605>]

I. INTRODUCTION

Genome mapping in nanochannels¹ is an emerging genomic method² that is ideally suited for creating scaffolds for *de novo* genome assembly,^{3–5} as well as searching for structural variations in genomes^{6–8} that have profound impacts on phenotype and disease states.⁹ As illustrated in Fig. 1(a), this technology relies on labeling long pieces of DNA with a sequence-specific label.¹ The labeled DNA is then inserted into a square nanochannel whose width (45 nm) is close to the persistence length of double-stranded DNA in a high ionic strength buffer.¹⁰ The DNA stretch within the channel, and the location of the labels are obtained by fluorescence microscopy. By also staining the DNA backbone with YOYO-1, each label can be associated with a particular molecule of DNA. The resulting data are digitized to create the genomic map; the genomic distance between labels on each molecule comprising the consensus map is computed from their physical distance through the average stretching of the DNA molecule in the channel.

The presence of knots within the DNA is a potential problem for the accuracy of genomic mapping in nanochannels.¹¹ Figure 1(b) illustrates one possible effect of a knot. Here, the knot pulls two contiguous labels closer together along the channel axis, leading to a reduction in the distance between labels “2” and “3.” In a genomics experiment, this reduction might be called as a deletion, even though the total amount of DNA between the two labels has remained the same. Figure 1(c) provides an even more dramatic example, where the formation of the knot

^{a)}Electronic mail: dorfman@umn.edu

FIG. 1. Illustration of the effect of knot formation on genome mapping in nanochannels. The upper schematics correspond to what would be obtained in two-color fluorescence microscopy images of the DNA backbone (green) and the sequence-specific labels (red). The image data are reduced to a one-dimensional genomic map indicating the relative positions of the labels within the channel, illustrated in this cartoon by the lower schematics. (a) In the ideal circumstance, the DNA is in a linear form without any knots, producing a genomic map where the physical distance between labels in the channel is proportional to the genomic distance between them on the DNA. The consensus of many such measurements is the reference genomic map. (b) A trefoil knot can produce an apparent deletion in the genomic map by pulling together the two interior labels within the channel. (c) A trefoil knot can scramble the labels on the DNA backbone. The corresponding genomic map suggests a complicated set of structural variations with respect to the reference genomic map.

causes labels “2” and “3” to appear out of order. Such an event might be interpreted as a complicated structural variation, but is in fact a result of topological rearrangement of the DNA in the channel rather than some biological outcome.

It is thus important to understand the frequency and size of knots in DNA in a nanochannel. In a recent experimental study, Reifengerger *et al.* examined the frequency and size of anomalously large YOYO intensity signals during genome mapping, which presumably correspond to regions of the nanochannel that have multiple strands of overlapping DNA.¹¹ Of the 189 153 molecules in that study, 7.3% exhibited a high YOYO fluorescence somewhere along their backbone. Over 50% of these molecules were folded at the leading edge. The remaining anomalous intensity events, which were distributed throughout the molecule backbone, have relative intensities of up to $10\times$ the typical YOYO fluorescence, with a peak at $3\times$. Such intensities are consistent with the presence of knots; a trefoil knot would produce approximately $3\times$ intensity and more complex knots could produce even higher intensities. However, these experiments do not provide a definitive result on the topology of these regions. For example, an S-loop would produce the same threefold intensity increase as a trefoil knot.¹¹

In the course of their discussion, Reifengerger *et al.* attempted to compare their experimental data to the extant simulation literature on knotting of DNA in nanochannels.¹¹ While quite a lot is known^{12–14} about knotting of polymers confined in three dimensions (e.g., in a bacterial capsid) or in the plane, there are considerably less data for the thermodynamics of knotting of linear DNA in a nanochannel. The most relevant work falls into two categories. On the one hand, Micheletti and co-workers have developed a substantial body of work on the knotting of DNA in a wide range of channel sizes,^{15–19} but these simulations generally involve contour lengths that are almost an order of magnitude smaller than those used in genome mapping. For example, Suma *et al.*¹⁹ studied the dynamics of DNA knotting and unknotting in a 56 nm channel, a channel size that is highly relevant for genomic mapping. However, the maximum length of the DNA in their study was $4.8\ \mu\text{m}$, whereas the minimum length of DNA used for genomic mapping is around $50\ \mu\text{m}$. Moreover, the typical size of a potentially knotted region of DNA in the experiments by Reifengerger *et al.*¹¹ is around 5 kilobases (approximately $1.7\ \mu\text{m}$), which would represent a substantial fraction of such a small molecule. On the other hand, Doyle and coworkers have studied knotting of longer chains,^{20,21} up to $20\ \mu\text{m}$. However, these chains were confined to channels that are over 100 nm, which is too large for genomic mapping.²²

Motivated by this gap in the literature and its importance for addressing unanswered questions raised by Reifengerger *et al.*,¹¹ we have performed simulations of DNA knotting in nanochannels ranging from 35 nm to 60 nm at a contour length of $56\ \mu\text{m}$ (approximately 164 kilobase pairs), which is the relevant range of channel sizes and near the lower bound for

the molecular weight for genomic mapping. Our work takes advantage of a pruned-enriched Rosenbluth method (PERM) for generating configurations of long chains in confinement and their statistical weights,^{23–25} which are then analyzed in a post-processing step using standard Alexander polynomials.²⁶ Our results suggest that the knotted events observed by Reifengerger *et al.*¹¹ are not representative of the equilibrium ensemble. This speculation has implications for the practice of genome mapping in nanochannels.

II. SIMULATION METHOD

We model the DNA as a discrete wormlike chain of N_b touching beads of diameter w , following our previous work.²⁷ The bending penalty between trios of continuous beads is accounted by the potential

$$U_{\text{bend}} = k_B T \kappa (1 - \cos \theta), \quad (1)$$

where k_B is Boltzmann's constant, T is the absolute temperature, κ is the bending penalty parameter, and θ is the angle between the beads. The bending penalty κ is related to the persistence length, l_p , through the relation²⁸

$$\frac{l_p}{w} = \frac{\kappa}{\kappa - \kappa \coth(\kappa) + 1}. \quad (2)$$

Excluded volume interaction is incorporated by enforcing hard-core repulsion between all beads and between the beads and the walls, treating the beads as hard spheres of diameter w . Due to the bead-wall interaction, the effective channel size available to the beads is $H_{\text{eff}} = H - w$, where H is the actual channel size. The parameters $w = 5.6$ nm and $l_p = 52.2$ nm are the same as our recent work comparing simulations and experiments during genome mapping.²⁹

This confined DNA model is then simulated by PERM,^{23–25} following the methodology outlined by Muralidhar *et al.*³⁰ For each channel size $H = 35, 40, \dots, 60$ nm, we run PERM simulations to grow chains up to $N_b^{\text{max}} = 10\,000$ beads, corresponding to a contour length of $L = 56$ μm . A single tour in PERM involves growing the chain bead-by-bead, where growing configurations with high statistical weights are enriched (copied) to improve sampling while those with low weights are pruned (growth stopped) to avoid spending time resolving configurations that have negligible contributions to the ensemble average. When a configuration j within a tour reaches $N_b = N_b^{\text{max}}$, we output its cumulative Rosenbluth weight W_j and the position of each bead i within that chain, $\{\mathbf{r}_i\}_j$. Since a given tour can produce multiple configurations that reach $N_b = N_b^{\text{max}}$ (most often when the channel size is large) and some tours fail to reach $N_b = N_b^{\text{max}}$ (most often when the channel size is small), each set of simulations produces a different number of configurations for analysis. The total number of configurations, N_c , produced by 1.5×10^5 tours for each channel size are reported in Table I.

For a given configuration j , we then need to determine the number of knots within that chain from $\{\mathbf{r}_i\}_j$. To minimize the computational overhead, this operation is done in the two steps illustrated by a typical example in Fig. 2. In the first step, we screen the chain for potential knots in a manner similar to what was done in experiments.¹¹ We divide the channel into bins of width w_b , and then count the number of beads in each bin. The axial segment density profile, $\rho_s(x)$, along the channel axis, x , is then simply the number of beads in a bin divided by w_b . As this is only a screening step, there is some flexibility in the choice of w_b ; a small value of w_b will count too many small fluctuations and flag many bins as possible knots, while a large value of w_b will reduce the dynamic range of $\rho_s(x)$ and make it challenging to determine an appropriate cutoff value for the segment density. We found that $w_b = 20w$ (equivalent to 112 nm) and a cutoff value $\rho_s = 2$ provided robust results, illustrated in the example in Fig. 2. While there are a number of regions in Fig. 2 with $\rho_s > 2$ that are not knotted, it is important to be conservative at this stage to avoid missing knotting events.

TABLE I. Statistics for the number of configurations, N_c , number of configurations with knots, N_c^k , and the total number of knots, N_k for each channel size, H .

H [nm]	N_c	N_c^k	N_k
35	1 548 716	3 827	3 827
40	1 445 324	34 415	34 420
45	1 205 133	115 053	121 180
50	1 438 909	240 512	269 265
55	1 737 945	537 268	666 708
60	1 409 708	831 929	1 195 735

We then analyze the knotting only for those chains having at least one bin with $\rho_s > 2$. Typically, multiple contiguous segments with $\rho_s > 2$ are observed, which we group together using a clustering algorithm to identify a region with a potential knot. It is possible to have multiple clustered structures scattered throughout a chain, and hence, a possibility of multiple knots in a single chain. For example, Fig. 2 has 10 clustered regions that could be knots. We then test for knotting by extracting the bead coordinates $\mathbf{r}_i$ that lie within a given clustered region with $\rho_s > 2$ and analyze the resulting subchain with the R package `RKnots`.^{31,32} Since these subchains are linear and mathematically any knot is defined in a closed loop, a centroid closure algorithm³³ is applied first to obtain closed loops, and then, the Alexander polynomials²⁶ are calculated. The output of `RKnots` is a symbolic expression for the Alexander polynomial, which can be used to identify the knot type. For some subchains, the polynomial is 1, which corresponds to the unknotted objects shown by purely green segments in Fig. 2. Such subchains have high segment density, but no knots. Subchains whose polynomial is not 1 contain a knot somewhere therein, as depicted by the orange segments in Fig. 2. Note that there are some limitations in using the Alexander polynomial to detect very complex knots.¹⁴ Since our work focuses on a reasonably tight confining geometry, we hardly observe such cases but denote these rare events as “unknown” or “UN.”

The output of the knot analysis provides the number of knots on chain j , N_j^k , the type of each of these knots, and the number of beads, N_b^k , involved in the knot. It proves convenient in what follows to convert the latter value into a knot contour length via

$$L^k = N_b^k w. \quad (3)$$

FIG. 2. An example of segment density profile, $\rho_s(x)$, along the channel axis, x , for a single chain containing three knots (two 3_1 and one 4_1) in a 55 nm channel. The horizontal dashed line indicates the cutoff value on ρ_s , above which all segments are analyzed for knot detection. Snapshots of the DNA segments with $\rho_s > 2$ are displayed, pointing to their corresponding peaks. The orange color is used to highlight the knots. Segments without an orange section have high ρ_s but no knots are detected in them.

PERM is a biased chain growth method, so we need to remove the bias to construct statistics from the data. As is the case in typical PERM simulations,²³ the total Rosenbluth weight of the simulation is the sum of the weights of each chain in the ensemble

$$W = \sum_{j=1}^{N_c} W_j. \quad (4)$$

The probability of observing a knot is then given by

$$p = W^{-1} \sum_{j=1}^{N_c^k} W_j. \quad (5)$$

The probability of observing k knots within a chain is computed in an analogous manner, where the summation in Eq. (5) now runs over all chains containing k knots.

Since one chain with weight W_j can contain multiple knots (e.g., the chain in Fig. 2), we need to also compute a total Rosenbluth weight for the knots

$$W_k = \sum_{j=1}^{N_c^k} N_j^k W_j. \quad (6)$$

The values of L^k obtained from knots within configuration j are then re-weighted by W_j/W_k to remove the bias for constructing histograms for the distribution of L^k .

III. RESULTS

The experiments by Reifengerger *et al.*¹¹ provide observations on the probability of suspected knot formation and the size of those knots, assuming that the anomalously high YOYO intensity events located on the interior of the chain are indeed knotted DNA. We thus analyze our data with respect to these two observables, beginning with the probability of observing a knot. Since our simulations have higher spatial resolution than the experiments, we also provide data on different knot types and their frequencies.

A. Knotting probability

Table I reports both the number of configurations obtained for each channel size and the number of such configurations that contain at least one knot. Note that the ratio of these values is not the probability of obtaining a knot, since PERM is a biased simulation method—different configurations contribute different weights to the ensemble average. After removing the bias, we obtain the knotting probability in Fig. 3. The knotting probability drops precipitously as the channel size decreases, from a maximum value of 59.3% at $H = 60$ nm to 0.3% at $H = 35$ nm.

The strong decrease in knotting probability as the channel size decreases is related to the concomitant increase in the global persistence length. As explained by Odijk,³⁴ the global persistence length g is the renormalized persistence length for an effective one-dimensional random walk along the channel axis for an ideal chain. In other words, the global persistence length is the typical distance between hairpin turns in the absence of excluded volume interactions. Subsequently, Muralidhar *et al.*³⁰ have proposed a correction to Odijk's global persistence length theory for a square nanochannel³⁴ that provides good correspondence between simulations and the theory down to channel sizes $H \approx 0.5l_p$. Figure 3(a) provides the value of the global persistence length for the channel sizes used here, which fall within the range where the corrected theory is valid.³⁰ Inasmuch as a knot requires making multiple hairpins, we would expect the probability of knotting to drop precipitously when $L \gg g$ is not satisfied. This is indeed the case for the smallest channel sizes.

FIG. 3. Probability of knotting as a function of channel size (black symbols) and (a) global persistence length (red line) or (b) Odijk parameter ξ (blue line).

Satisfying the criterion $L \gg g$ is a necessary but not sufficient condition for frequent hairpin formation, since the global persistence length is a property of an ideal chain.³⁰ It is also important that the excluded volume created by a hairpin turn be small. Odijk³⁵ proposed that the relevant quantity is the ratio of the excluded volume created by a hairpin turn to the channel volume available to the polymer involved in the hairpin. For a square channel, this ratio is

$$\xi = \frac{gw}{H^{5/3}l_p^{1/3}}. \quad (7)$$

Figure 3(b) provides the companion data for ξ as a function of channel size, plotted again with the knotting probability to facilitate comparison with Fig. 3(a). When $\xi > 1$, the excluded volume created by a hairpin is strong and hairpin formation is infrequent.^{30,35} Thus, even if we simulate extremely long chains at $H = 35$ nm, we would continue to expect to have low knotting probabilities due to the rarity of hairpin formation owing to excluded volume. Conversely, the excluded volume penalty at $H = 60$ nm is not especially strong, so hairpin formation is frequently provided $L \gg g$, which is indeed the case in our simulations.

When knotting is rare, there is the possibility for correlations in the PERM data *within* a given tour, bearing in mind that the 1.5×10^5 tours themselves are independent. In the context of studying knotting, we might expect either of the two correlations illustrated in Fig. 4. In one case, there are configurations that, if they are enriched, have a tendency to promote knot formation nearby during the chain growth. In this case, there is a tendency for the knots to be located in the same region within the channel but the knots themselves are different. In the second case, the chain growth leads to a knot and subsequent enrichment leads to that same knot appearing many times in the data set.

For the smaller channel sizes ($H = 35$ nm and 40 nm), the knots tend to be localized at particular points along the channel. To aid in this analysis, let us define a normalized knot position

FIG. 4. Illustration of potential correlation errors in PERM when hairpin formation is infrequent. (a) The chain growth leads to the formation of a particular configuration that is favorable for subsequent knot formation. During enrichment, this configuration generates many knots but the knots themselves have different configurations. (b) The chain growth leads to a knot within the chain. If this chain is enriched, all of the copies will have exactly the same knot.

$$d^k = \frac{x^k - x_{\min}}{X}, \quad (8)$$

where x^k is the axial position of the center of mass of knot k , $x_{\min}$ is the axial position of the bead located at the smallest value of x , and X is the span of the chain in the channel

$$X = x_{\max} - x_{\min}. \quad (9)$$

Figure 5 shows the distributions for d^k for the different channels studied here. It is clear that the 35 nm channel (which exhibits very infrequent knotting) had all of its knots approximately 60% and 80% of the way along the chain. In addition to correlations of the type in Fig. 4, this localization results from the relatively large value of g and small value of ζ , implying that substantial contour length from the start of chain growth is required to form a knot. We analyzed these seemingly localized knots and found, fortunately, that their statistics were what we would expect from Fig. 4(a) rather than the identical configurations of Fig. 4(b). As such, the correlation errors in our simulation are essentially the same as we would expect from a bead displacement Monte Carlo method where the knots are infrequent and have low mobility; we sample many configurations of knots within a particular location of the chain. Note that, in dynamic simulations, knot diffusion along the chain is not equivalent to the formation and destruction of many knots as the diffusing knot has memory of its previous configuration. As such, even if a dynamic simulation allows the knot to move, it may still exhibit the same correlation errors as a chain growth method.

The distribution of knot locations in Fig. 5 is noisier than previous results obtained for knot formation in free solution.³⁶ The distributions are also flatter than those reported previously.³⁶ This flattening may arise from the difference in methodology, as the chain growth method used here does not involve diffusion of contour along the chain, or could be due to the shorter segment-segment correlation length in confinement.

In what follows, we will continue to include data for the smallest channels for completeness, keeping in mind the possibility for correlation errors when knotting is rare. It is worth noting that correlation errors are well controlled in our prior PERM work on chain extension and its variance^{25,30,37–40} by simply running more tours. Since each tour is independent, the

FIG. 5. Probability density for observing a knot at a fractional distance d^k away from the bead with the minimum position for different channel sizes.

TABLE II. Percentage of knotted chains that have k knots in a single chain as a function of channel size H . Entries listed as ... indicate no chains with that number of knots were identified. The columns in the table are normalized to sum to 100%.

Number of knots k	Channel size H [nm]					
	35	40	45	50	55	60
1	100	99.993	95.770	86.441	78.629	63.974
2	...	0.007	4.079	12.578	18.938	29.091
3	...	...	0.151	0.980	2.304	6.522
4	...	...	...	...	0.127	0.402
>4	...	...	...	...	...	0.011

statistics eventually become independent of the number of tours. Typically, this requires around 10^6 tours. Unfortunately, given the amount of data produced when we store the configurations and the post-processing required to compute the knots, it is infeasible to run another order-of-magnitude more tours here. In prior work,^{25,30,37–40} all thermodynamic data were computed on-the-fly without storing configuration data to disk or performing any post-processing.

In addition to the overall knotting probability, we can also determine the probability that a given chain contains multiple knots. As seen in Table I, the total number of knots that we observed in our simulations exceeds the number of knotted chains for all channel sizes with the exception of $H = 35$ nm. Table II provides the probability of observing multiple knot formation in each channel size up to $k = 4$ knots, obtained after removing the PERM bias. The frequency of observing multiple knots on a given chain increases as we increase the channel size, consistent with the arguments about hairpin formation and excluded volume made in the context of Fig. 3.

Previously, Michelletti and Orlandini¹⁷ proposed a combinatorial model for determining the number m of identical prime knots of the same type within a chain. It is interesting to see whether this model also applies in general to the overall probability of finding multiple knots within a chain. The combinatorial model predicts that the probability p_k of observing k knots within a chain is given by¹⁷

$$p_k = (-\ln \tilde{p})^k \frac{\tilde{p}}{k!}, \quad (10)$$

where $\tilde{p} = 1 - p$ is the probability that a chain contains no knots. As we see in Fig. 6, Eq. (10) is a very good approximation in terms of the absolute value of the probability and the overall

FIG. 6. Probability of observing k knots within a chain. The solid lines are the combinatorial model in Eq. (10) and the solid points are the data obtained for different channel sizes.

trend in the data. The relative error can become large for rare knots, since the probability of observing such knots is small. However, we cannot determine whether the disagreement there is due to limitations of the model or sampling errors in the simulations.

B. Knot type and size

In addition to identifying the prevalence of knotting and the frequency of multiple knots on a given chain, the Alexander polynomials²⁶ provide information on the type of knot. Table III provides the percentage of each knot type observed in our simulations. For the smallest channels, we only obtain trefoil knots, presumably due to the strong excluded volume created by forming a hairpin. As the channel size increases, the frequency of trefoil knots decreases in favor of the formation of increasingly more complex knots. We have even observed knots that may have up to 8 crossings. Given the limitations of the Alexander polynomial, these apparent 8-fold knots may be compound knots with the same polynomial. As such, we can only state with confidence that we observe knots with up to seven crossings.

The statistics for the 5-fold knots in Table III are of particular interest with respect to previous work on knotting in the absence of confinement.⁴¹ In previous simulations of unconfined circular DNA, Rybenkov *et al.* observed that the 5_2 twist knot occurs at a higher frequency than the 5_1 torus knot.⁴¹ We observe similar behavior in Table III for the smaller channel sizes, but then the trend reverses. We have tested the result as a function of the number of tours using 5×10^4 , 1×10^5 , and 1.5×10^5 tours, and find that the ordering in Table III is robust to the number of tours, albeit with differences in the values of the probabilities due to the sampling errors. Establishing firm bounds on the probabilities in Table III would require at least one order of magnitude in tours, based on our experience in previous work,^{25,30,37–40} where expectations were computed on-the-fly during the PERM simulations. Given that the knotting problem requires storing and post-processing the configurations, such an analysis is infeasible. Nonetheless, it is intriguing that the reversal in the trend for 5-fold knots takes place at the point where the global persistence length begins to approach the DNA persistence length. Fully

TABLE III. Percentage of different knot types as a function of channel size H appearing in Fig. 7. Entries listed as ... indicate that no knots were identified of that type. The row UN corresponds to knots of unknown type that cannot be identified using an Alexander polynomial. The 8-fold knots are identified by Alexander polynomials and may be composite knots. The columns in the table are normalized to sum to 100%.

Knot type	Channel size H [nm]					
	35	40	45	50	55	60
3_1	100	99.99	89.071	86.079	87.934	84.618
4_1	...	0.01	9.750	5.788	9.440	10.153
5_1	...	...	...	1.094	1.163	2.538
5_2	...	...	0.968	7.031	0.610	1.671
6_1	...	...	...	0.004	0.212	0.331
6_2	...	...	...	0.004	0.163	0.513
6_3	...	...	0.001	...	...	0.075
7_1	...	...	...	...	...	^a
7_2	...	...	...	...	...	0.001
7_3	...	...	...	...	...	^a
7_5	...	...	...	...	0.014	...
7_6	...	...	...	...	^a	^a
7_7	...	...	...	...	...	^a
8_{20}	...	...	0.210	...	0.445	0.090
8_{21}	...	...	...	...	0.018	0.003
UN	...	...	...	...	^a	0.006

^aKnot type was observed but the percentage is less than 0.001%.

determining how 5-fold knots evolve as a function of confinement represents an interesting but computationally challenging avenue for future work.

Inasmuch as we have obtained a reasonable amount of data for knots with complexities exceeding the 3_1 knot, it is worthwhile to see whether these more complex knots contain substantially more DNA. To this extent, Fig. 7 provides box plots for the distributions of contour length within a knot, L^k , for each knot type reported in Table III. There is a modest increase in the contour length involved in the knot as its complexity increases. However, simply having more DNA involved in the knot does not imply that the knot is more complex. The “outliers” in the box plots for the trefoil knots can contain more DNA than any of the more complex knots, which is likely the result of better sampling of the trefoil knots. Nevertheless, Fig. 7 illustrates that knots of a particular type can have a very wide range of compactness.

Figure 7 breaks out the contour length inside the knot in terms of the knot type. While convenient for understanding how knot complexity is related to knot size, such granularity in the data exceeds what is possible from genomic mapping experiments.¹¹ In the experiments, the amount of DNA contained within a purported knot was measured by aligning the DNA on the two sides of the knot to the reference genome, and then computing how much the knot contracted the DNA between the two linear regions, analogous to Fig. 1(b).

To provide the optimal correspondence between our simulations and experiments, Fig. 8 provides the composite information on the distribution of contour length within the knots as a function of channel size, independent of the complexity of the knot. Overall, we find that the contour lengths L^k corresponding to the bulk of the distributions remain essentially unchanged as the

FIG. 7. Box plot of the knot contour length statistics as a function of knot complexity for each channel size. The label “UN” on the x axis denotes “unknown” knots which Alexander polynomials fail to identify. The 8-fold knots are identified by Alexander polynomials and may be composite knots. The proportions of the data corresponding to each box are listed in Table III.

FIG. 8. Statistics for knot contour lengths as a function of channel size. The blue box spans from the 25th percentile to 75th, the red line indicates the median of the distribution, and the black line spans from the minimum to the maximum in the distribution.

channel size increases, although there is a slight uptick for $H = 55$ nm and $H = 60$ nm. Rather, the range of possible knot sizes increases substantially with channel size. The minimum contour length within a knot remains relatively unchanged with channel size at approximately 280 nm, reflecting the inability to tighten the knot any further due to excluded volume within the knot. As the channel size increases, the longest knot observed in our simulations increases as the knot can swell to fill the larger channel volume. If the knot is a compact object, we might expect that the maximum in the distribution would scale like H^3 . The data in Fig. 8 suggest that this power law might be a reasonable description of the data, but the limited range of channel sizes and using the maximum in the distribution limits the probative value of such a power law.

IV. DISCUSSION

The objective of our simulations was to fill in the gap in the simulation literature for long DNA chains in small channels, thereby addressing directly the situation studied by Reifengerger *et al.*¹¹ in their genomic mapping experiments. To do so, we took an alternate simulation approach (PERM) that is well suited to generate very long molecules in the Odijk regimes.^{25,30,37–40} Overall, the result obtained for long chains in our simulations are in agreement with those obtained by Orlandini and coworkers for short chains.^{16,17,19} For example, the distributions for $H = 55$ and 60 nm channels in Fig. 8 approximately agree with previous simulations¹⁹ carried out for shorter DNA (1.2–4.8 μm) in a 56 nm channel and for circular DNA of length 4.8 μm confined in a slit.¹⁶ Furthermore, the typical value of the knot size in our simulation is about 0.5 μm , which is 3 times smaller than that observed in experiments,¹¹ but close to that found in previous simulations for short DNA in similar confinement.^{17,19} This disagreement between the experiment and simulation was also discussed in the article on experimental data,¹¹ and was left as an open question: Would the chain size affect the agreement? Our results clearly indicate that increasing chain length beyond that used previously^{17,19} does not affect the knot size. This statement is also consistent with previous simulation work in one-dimensional confinement,^{19,42,43} where the chain sizes were small and similar conclusions about the knot localization were made.

Our simulations uncovered a number of very complex knots (Table III) that have not been seen previously in simulations at such strong confinement, and we also have instances where there are multiple knots within a single chain (Table II) that are not seen for shorter molecules. Both of these new results arise from our change in methodology, but they do not change the

key conclusion reached by Reifenger *et al.*¹¹—the experimental data are not consistent with predictions from equilibrium simulations of confined DNA.

What are the reasons for the discrepancy between simulations and experiments? While it is always possible that the model itself is not sufficient, this seems unlikely to be the dominant factor given the success of this model in capturing other properties of confined DNA in experiments.^{27,44–47} The equilibrium knots tend to be much smaller than those observed in experiments. The distribution of contour lengths in experiments¹¹ is centered around 5 kilobase pairs (approximately $1.7\ \mu\text{m}$) and the maximum extends out to almost 40 kilobase pairs (approximately $13.6\ \mu\text{m}$). These results disagree with the data reported in Fig. 8; the simulated knot formation is always substantially smaller than what is observed in experiments. We would expect the knot size in the simulations to be smaller than the experiments, since we can trim the subchain down to the minimal size required to form a knot, whereas the experimental measurement has much less precision. However, it seems highly unlikely that this difference in measurement precision could result in a several-fold change in the results.

Addressing the knotting frequency is a more subtle question than the knot size. The knotting probabilities in Fig. 3 vary from 10% at $H=45\ \text{nm}$ to 16% at $H=50\ \text{nm}$, which is the relevant range of channel sizes in genomic mapping experiments.⁴⁸ In this first approximation, the knotting probabilities in simulations and experiments seem to be in qualitative agreement.

However, the comparison between experiments and simulations is not as simple as this first approximation suggests, with competing factors that would lead to the simulations overestimating or underestimating the knotting probability. On one hand, the simulations include only the number of knotted events, while the experiments include all anomalously high intensity events that are not $2\times$ intensity relative to the baseline and are not located at the leading edge of the molecule. As indicated in the example of Fig. 2, an anomalously high segmental density $\rho_s(x)$ is not necessarily coincident with the existence of a knot. Indeed, while two of the knots in that particular molecule have the two highest segmental densities, the third knot is the 8th highest density spike. Thus, by only considering the knotting events, the simulations should underestimate the frequency of anomalously bright events (which may or may not be knots). On the other hand, the simulations have an infinite signal-to-noise ratio, while the experimental analysis required a smoothing algorithm to deal with the relatively low signal-to-noise ratio of the backbone intensity under the 38:1 dye-to-base-pair ratio used in genome mapping.¹¹ Similarly, the spatial resolution of the simulations is set by the bead size ($5.6\ \text{nm}$), while the experiments are limited by the pixel size of the image ($108.3\ \text{nm}$).¹¹ The higher spatial resolution and signal-to-noise ratio in the simulations should overestimate the frequency of anomalous intensity events in experiments. Given that these two effects point in opposite directions, any agreement between the predicted knotting probability in simulations and the frequency of anomalously bright events in experiments could arise from a fortuitous cancellation of errors.

In an attempt to provide a better correspondence between the simulations and experiments, we also analyzed the $45\ \text{nm}$ channel simulation data in a manner that is closer to the experimental analysis.¹¹ Explicitly, we computed the segmental density profile $\rho(x)$ and determined the probability that a chain will exceed a threshold value ρ^* in at least one bin position x along the chain. Using the cutoffs provided in experiments, we found that the probability of our simulated chains having a $2\times$ intensity spike is 88.8%, whereas a $3\times$ intensity spike had a probability of 28.1%. These results suggest a 28.1% probability of a knot-like intensity event and a 60.7% probability of a fold (i.e., the remaining events that are at least $2\times$). Both values are substantially higher than the experimental results. One potential problem in this comparison is the aforementioned difference in the signal-to-noise ratio. If we increase the threshold to a $4\times$ intensity spike, we observe a much lower value of 0.5%. The sensitivity of the analysis method to the particular choice of cut-off, which itself is connected to the signal-to-noise, makes it challenging to directly compare the probability of observing anomalously bright events in experiments to simulations.

Taken together, our analysis of both the knotting frequency and probability of knotting calls into question whether the presumed knots observed by Reifengerger *et al.*¹¹ are indeed knots at all, a point that was also raised in that publication. A definitive answer to this question requires further experiments on the dynamics of these regions of anomalously high segmental density. Given the low mobility of DNA within such a small channel^{28,49} and the very large DNA molecules, it seems unlikely that there is sufficient time for the knots to diffuse to the ends of the chain without photobleaching the dye. In fact, recent dynamic simulations^{18,19} for short chains confined in a nanochannel showed that the knot lifetime is up to two orders of magnitude shorter than the time span from untying a knot to the formation of a new knot when the channel size is close to the persistence length. In particular, Suma *et al.*¹⁹ found that, for a 4.8 μm long chain in 56 nm channel, the mean knot lifetime was around 20 ms, whereas the mean time between formation of knots was about 150 ms. They also varied the chain length from 1.2 to 4.8 μm and showed that the knot lifetime increases by increasing the chain length because the knot needs to diffuse much farther to reach the adsorbing boundary condition at the chain end. However, the waiting time between two knotting events was almost independent of the chain size because the rate of formation of a new knot depends on the rate at which back-foldings occur at the chain end, which is a local nucleation event. Overall, we might expect an increased value of the knot duration for much longer chains similar to what we have in this work. However, if the anomalously bright regions are not topologically constrained, we might expect their size to diminish on experimental time scales due to excluded volume interactions. Thus, dynamic experiments are likely to be a key step towards understanding anomalous bright regions during genomic mapping.

V. CONCLUDING REMARKS

In the present contribution, we have used PERM simulations of a discrete wormlike chain model to investigate the knotting of DNA in the strong confinement used for genome mapping in nanochannels. Overall, we found that increasing the contour length of the chain did not qualitatively change the conclusions reached by simulations of shorter chains in similar degrees of confinement.^{16,17,19} We did observe more complicated knots and multiple knots per chain, both of which are likely due to the ability of PERM to reach high molecular weights in the Odijk regime. These results are encouraging for future work on knotting in strong confinement, in particular, for dynamic simulations, since shorter chains are considerably less expensive to simulate than their longer counterparts. They also suggest that PERM could be an efficient method to look at rare knotting events.

Our study of longer chains was motivated by the experiments by Reifengerger *et al.*,¹¹ who reported the presence of anomalously bright regions along the DNA backbone during genome mapping in nanochannels. Our analysis of both the knotting probability and, more importantly, the contour length of DNA involved in a knot suggest that the anomalously bright regions observed during genome mapping experiments¹¹ are not consistent with the predictions of the sizes of knots in the equilibrium ensemble of confined DNA chains in channels near the persistence length. This is a promising result with respect to the genomic technology. Even when the DNA is knot-free, it still undergoes thermal fluctuations that lead to small changes in the distance between labels on the DNA during genome mapping. Since the equilibrium knots are typically small, the presence of knots within the DNA is unlikely to have a significant effect on the quality of the genomic maps obtained by stretching DNA in nanochannels.

While there are several positive aspects of our study with respect to both simulations of knotted DNA in nanochannels and their effect on genome mapping, the simulations reported here do not explain the anomalously bright regions observed during genome mapping experiments.¹¹ However, if they do not arise from thermal fluctuations at equilibrium, then they are not an intrinsic limitation of the technology. Rather, these anomalously bright regions must arise from some other part of the experimental protocol. Possible reasons include knots on the DNA that form prior to injection into the nanochannel, defects within the nanochannel that lead

to transient adsorption to the surface, or failure to completely remove DNA binding proteins during extraction from the cells. It is also possible that the large brightness is due to knots that are passing between one another.⁵⁰ We are optimistic that further experimental work, in particular, visualizing the dynamics of these bunched regions of DNA, will ultimately reveal their origin and an approach to minimize their effect on genome mapping.

ACKNOWLEDGMENTS

This work was supported by the National Institutes of Health (R01-HG006851). Computational resources were provided in part by the Minnesota Supercomputing Institute at the University of Minnesota.

- ¹E. T. Lam, A. Hastie, C. Lin, D. Ehrlich, S. K. Das, M. D. Austin, P. Deshpande, H. Cao, N. Nagarajan, M. Xiao, and P.-Y. Kwok, *Nat. Biotechnol.* **30**, 771 (2012).
- ²M. Levy-Sakin and Y. Ebenstein, *Curr. Opin. Biotechnol.* **24**, 690 (2013).
- ³A. R. Hastie, L. Dong, A. Smith, J. Finklestein, E. T. Lam, N. Huo, H. Cao, P.-Y. Kwok, K. R. Deal, J. Dvorak, M.-C. Luo, Y. Gu, and M. Xiao, *PLoS One* **8**, e55864 (2013).
- ⁴M. Pendleton, R. Sebra, A. W. C. Pang, A. Ummat, O. Franzen, T. Rausch, A. M. Stutz, W. Stedman, T. Anantharaman, A. Hastie, H. Dai, M. H. Y. Fritz, H. Cao, A. Cohain, G. Deikus, R. E. Durrett, S. C. Blanchard, R. Altman, C.-S. Chin, Y. Guo, E. E. Paxinos, J. O. Korbel, R. B. Darnell, W. R. McCombie, P.-Y. Kowk, C. E. Mason, E. E. Schadt, and A. Bashir, *Nat. Methods* **12**, 780 (2015).
- ⁵J. Seo, A. Rhie, J. Kim, S. Lee, M. Sohn, C. Kim, A. Hastie, H. Cao, J. Yun, J. Kim, J. Kuk, G. H. Park, J. Kim, H. Ryu, J. Kim, M. Roh, J. Baek, M. W. Hunkapiller, J. Korlach, J. Shin, and C. Kim, *Nature* **538**, 243 (2016).
- ⁶H. Cao, A. R. Hastie, D. Coa, E. T. Lam, Y. Sun, H. Huang, X. Liu, L. Lin, W. Andrews, S. Chan, S. Huang, X. Tong, M. Requa, T. Anantharaman, A. Krogh, H. Yang, H. Cao, and X. Xu, *GigaScience* **3**, 34 (2014).
- ⁷C. L. Usher, R. E. Handsaker, T. Esko, M. A. Tuke, M. N. Weedon, A. R. Hastie, H. Cao, J. E. Moon, S. Kashin, C. Fuchsberger, A. Metspalu, C. N. Pato, M. T. Pato, M. I. McCarthy, M. Boehnke, D. M. Altshuler, T. M. Frayling, J. N. Hirschhorn, and S. A. McCarrroll, *Nat. Genet.* **47**, 921 (2015).
- ⁸A. C. English, W. J. Salerno, O. A. Hampton, C. Gonzaga-Jauregui, S. Ambreth, D. I. Ritter, C. R. Beck, C. F. Davis, M. Dahdouli, S. Ma, A. Carroll, N. Veeraraghavan, J. Bruestle, B. Drees, A. Hastie, E. T. Lam, S. White, P. Mishra, M. Wang, Y. Han, F. Zhang, P. Stankiewicz, D. A. Wheeler, J. G. Reid, D. M. Muzny, J. Rogers, A. Sabo, K. C. Worley, J. R. Lupski, E. Boerwinkle, and R. A. Gibbs, *BMC Genomics* **16**, 286 (2015).
- ⁹L. Feuk, A. Carson, and S. Scherer, *Nat. Rev. Genet.* **7**, 85 (2006).
- ¹⁰C. Bustamante, J. F. Marko, E. D. Siggia, and S. Smith, *Science* **265**, 1599 (1994).
- ¹¹J. G. Reifenger, K. D. Dorfman, and H. Cao, *Analyst* **140**, 4887 (2015).
- ¹²E. Orlandini and S. G. Whittington, *Rev. Mod. Phys.* **79**, 611 (2007).
- ¹³D. Marenduzzo, C. Micheletti, and E. Orlandini, *J. Phys.: Condens. Matter* **22**, 283102 (2010).
- ¹⁴C. Micheletti, D. Marenduzzo, and E. Orlandini, *Phys. Rep.* **504**, 1 (2011).
- ¹⁵C. Micheletti, D. Marenduzzo, E. Orlandini, and D. Sumners, *Biophys. J.* **95**, 3591 (2008).
- ¹⁶C. Micheletti and E. Orlandini, *Macromolecules* **45**, 2113 (2012).
- ¹⁷C. Micheletti and E. Orlandini, *Soft Matter* **8**, 10959 (2012).
- ¹⁸C. Micheletti and E. Orlandini, *ACS Macro Lett.* **3**, 876 (2014).
- ¹⁹A. Suma, E. Orlandini, and C. Micheletti, *J. Phys.: Condens. Matter* **27**, 354102 (2015).
- ²⁰L. Dai, J. R. C. van der Maarel, and P. S. Doyle, *ACS Macro Lett.* **1**, 732 (2012).
- ²¹L. Dai, C. B. Renner, and P. S. Doyle, *Macromolecules* **48**, 2812 (2015).
- ²²S. K. Das, M. D. Austin, M. C. Akana, P. Deshpande, H. Cao, and M. Xiao, *Nucleic Acids Res.* **38**, e177 (2010).
- ²³P. Grassberger, *Phys. Rev. E* **56**, 3682 (1997).
- ²⁴T. Prellberg and J. Krawczyk, *Phys. Rev. Lett.* **92**, 120602 (2004).
- ²⁵D. R. Tree, Y. Wang, and K. D. Dorfman, *Phys. Rev. Lett.* **110**, 208103 (2013).
- ²⁶M. D. Frank-Kamenetskii and A. V. Vologodskii, *Sov. Phys. Usp.* **24**, 679 (1981).
- ²⁷Y. Wang, D. R. Tree, and K. D. Dorfman, *Macromolecules* **44**, 6594 (2011).
- ²⁸A. Muralidhar and K. D. Dorfman, *Macromolecules* **48**, 2829 (2015).
- ²⁹A. Jain, J. Sheats, J. G. Reifenger, H. Cao, and K. D. Dorfman, *Biomicrofluidics* **10**, 054117 (2016).
- ³⁰A. Muralidhar, D. R. Tree, and K. D. Dorfman, *Macromolecules* **47**, 8446 (2014).
- ³¹F. Comoglio and M. Rinaldi, *PLoS One* **6**, e18693 (2011).
- ³²F. Comoglio and M. Rinaldi, *Bioinformatics* **28**, 1400 (2012).
- ³³R. C. Lua and A. Y. Grosberg, *PLoS Comput. Biol.* **2**, e45 (2006).
- ³⁴T. Odijk, *J. Chem. Phys.* **125**, 204904 (2006).
- ³⁵T. Odijk, *Phys. Rev. E* **77**, 060901 (2008).
- ³⁶L. Tubiana, A. Rosa, F. Fragiaco, and C. Micheletti, *Macromolecules* **46**, 3669 (2013).
- ³⁷A. Muralidhar, D. R. Tree, Y. Wang, and K. D. Dorfman, *J. Chem. Phys.* **140**, 084905 (2014).
- ³⁸D. R. Tree, W. F. Reinhart, and K. D. Dorfman, *Macromolecules* **47**, 3672 (2014).
- ³⁹A. Muralidhar and K. D. Dorfman, *Macromolecules* **49**, 1120 (2016).
- ⁴⁰A. Muralidhar, M. Quevillon, and K. D. Dorfman, *Polymers* **8**, 79 (2016).
- ⁴¹V. V. Rybenkov, N. R. Cozzarelli, and A. V. Vologodskii, *Proc. Natl. Acad. Sci. U.S.A.* **90**, 5307 (1993).
- ⁴²C. H. Nakajima and T. Sakaue, *Prog. Theor. Phys. Suppl.* **191**, 205 (2011).
- ⁴³C. H. Nakajima and T. Sakaue, *Soft Matter* **9**, 3140 (2013).

- ⁴⁴D. R. Tree, Y. Wang, and K. D. Dorfman, [Biomicrofluidics](#) **7**, 054118 (2013).
- ⁴⁵D. Gupta, J. J. Miller, A. Muralidhar, S. Mahshid, W. Reisner, and K. D. Dorfman, [ACS Macro Lett.](#) **4**, 759 (2015).
- ⁴⁶D. Gupta, J. Sheats, A. Muralidhar, J. J. Miller, D. E. Huang, S. Mahshid, K. D. Dorfman, and W. Reisner, [J. Chem. Phys.](#) **140**, 214901 (2014).
- ⁴⁷L. Dai, D. R. Tree, J. R. C. Van Der Maarel, K. D. Dorfman, and P. S. Doyle, [Phys. Rev. Lett.](#) **110**, 168105 (2013).
- ⁴⁸W. F. Reinhart, J. G. Reifenberger, D. Gupta, A. Muralidhar, J. Sheats, H. Cao, and K. D. Dorfman, [J. Chem. Phys.](#) **142**, 064902 (2015).
- ⁴⁹D. R. Tree, Y. Wang, and K. D. Dorfman, [Phys. Rev. Lett.](#) **108**, 228105 (2012).
- ⁵⁰B. Trefz, J. Siebert, and P. Virnau, [Proc. Natl. Acad. Sci. U.S.A.](#) **111**, 7948 (2014).