Sequence-Dependent Persistence Length of Long DNA

Hui-Min Chuang,[1] Jeffrey G. Reifenberger,[2] Han Cao,[2] and Kevin D. Dorfman[1,*]
[1]*Department of Chemical Engineering and Materials Science, University of Minnesota–Twin Cities,*
421 Washington Avenue SE, Minneapolis, Minnesota 55455, USA
[2]*BioNano Genomics, 9640 Towne Centre Drive, Suite 100, San Diego, California 92121, USA*

(Received 25 July 2017; published 29 November 2017)

Using a high-throughput genome-mapping approach, we obtained circa 50 million measurements of the extension of internal human DNA segments in a 41 nm × 41 nm nanochannel. The underlying DNA sequences, obtained by mapping to the reference human genome, are 2.5–393 kilobase pairs long and contain percent GC contents between 32.5% and 60%. Using Odijk's theory for a channel-confined wormlike chain, these data reveal that the DNA persistence length increases by almost 20% as the percent GC content increases. The increased persistence length is rationalized by a model, containing no adjustable parameters, that treats the DNA as a statistical terpolymer with a sequence-dependent intrinsic persistence length and a sequence-independent electrostatic persistence length.

DOI: 10.1103/PhysRevLett.119.227802

Over the past two decades, long molecules of double-stranded DNA have emerged as an important model system in polymer physics, with applications in rheology [1], confined polymers [2–4], and transport in model porous media [4,5]. A particularly salient advantage of DNA is the ability to visualize the polymer by fluorescence microscopy, thereby directly interrogating the underlying physical models at the single-molecule level. The proper interpretation of these experiments requires an accurate measurement of the DNA persistence length. Often, the persistence length is obtained from force-extension experiments [6] or polyelectrolyte theory [7]. These approaches often assume that the persistence length of DNA is, at most, a weak function of the sequence. In this Letter, we present data obtained from a high-throughput genomic-mapping method [8,9] that call into question this widespread assumption. Using circa 5×10^7 measurements of DNA extension in nanochannels, we show that the 2% increase in fractional extension as percent GC content increases (which does not affect the genome-mapping strategy employed here) translates into a persistence length that varies by almost 20% due to the relatively weak dependence of the fractional extension on persistence length in the Odijk regime [10]. Building on existing concepts [7,11], we rationalize our result by modeling long DNA as a statistical terpolymer with a sequence-dependent intrinsic persistence length.

The neglect of the DNA sequence in many polymer physics experiments stands in stark contrast to that in biophysics. The so-called "intrinsic curvature" of DNA, which emerges over circa 100 base pairs, depends strongly on the DNA sequence [11–13] and is purported to play a role in biological processes such as nucleosome positioning [14–17]. Likewise, certain sequences such as poly(A) tracts introduce local bends in DNA [18–21], again at very short length scales. These local properties are modeled by a sequence-dependent bending energy that depends on the dinucleotide pair being bent [11,12]. The dependence of intrinsic curvature on sequence implies, *inter alia*, that the dinucleotide bending energies differ substantially. As such, they should manifest at long length scales in the DNA persistence length in the same way that hindered rotation around carbon-carbon bonds leads to a 13% increase in statistical segment length for polystyrene when compared to polyethylene [22].

Measuring how the DNA persistence length depends on the sequence, while simultaneously ensuring the sequence is long enough to average over the intrinsic curvature, is an onerous task. Standard methods, such as light or neutron scattering (see references in Ref. [23]), magnetic tweezers [6], and atomic force microscopy [18], are inherently low-throughput. We thus adopted the genome-mapping approach described in Fig. 1. A detailed explanation of the experimental method is included in Supplemental Material [24]. Briefly, DNA were extracted from a human cell line (Hapmap NA 12878, female, Caucasian). The DNA were nick labeled using Nt·BspQI (New England Biolabs) to insert cy-3-like fluorescent nucleotides at the nick site GCTCTTC [8], and the backbone was stained with YOYO-1 (Invitrogen) at a ratio of one dye molecule to 37 base pairs [Fig. 1(a)] [33]. The DNA were then stretched by electrokinetic injection into an array of 41-nm-wide, square nanochannels on an Irys® v2 chip (BioNano Genomics) and imaged on a research-grade version of the Irys® system [Fig. 1(b)] using the Bionano Prep™ flow buffer (BioNano Genomics, ionic strength = 48 mM). We obtained data on 452 219 DNA molecules at least 150 kilobase pairs (kbp) in size. These molecules aligned to the human reference genome (hg19) at a hit rate of 85.2%, yielding a final data set with 36× coverage of the human

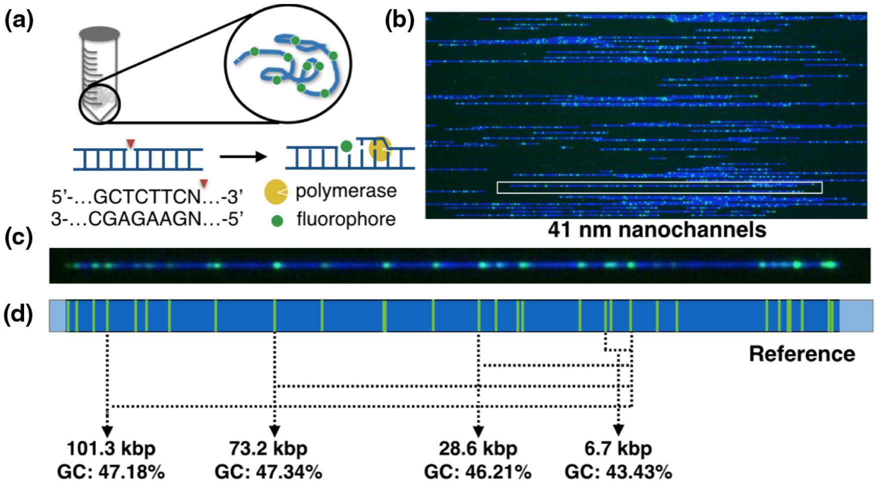

FIG. 1. Experimental approach to measure the persistence length of DNA over a wide range of sequences. (a) Human DNA are fluorescently labeled at the GCTCTTC sequence by nick labeling, and the backbone is stained with YOYO-1. (b) The labeled DNA are stretched in a 41 nm × 41 nm nanochannel using the high-throughput Irys® genome-mapping system. (c) Individual molecules are mapped to the (d) reference human genome, which reveals the underlying sequence—and percent GC content—between nick sites. This particular molecule (154.6 kbp) has 30 nick sites; the percent GC content values for 4 of the 435 possible pairs of nicking sites on this molecule are indicated.

reference genome. We considered only the extension between pairs of nick sites in a given chromosome that are (i) separated by at least 2.5 (for adequate resolution) and 393 kbp (for adequate sampling) and (ii) do not contain any N-base (unknown) regions in the human genome. Removing N-base regions is essential, as these unknown sequences in the reference genome introduce systematic errors [24,34]. Figure 1(c) shows a representative molecule with 30 nick sites; the percent GC content for the sequences between 4 of the 435 possible pairs of nicking sites on this molecule is indicated in Fig. 1(d).

Human DNA and the high-throughput afforded by genome mapping in nanochannels are essential to the robustness of our experiments. In contrast to microorganisms and viruses, whose DNA are commonly exploited for polymer physics [35], human DNA possesses a wide range of percent GC content. As an extreme example, we identified pairs of nick sites with very similar separations on chromosome 6 (2555 bp separation) and chromosome 15 (2504 bp separation) with percent GC contents of 16.4% and 74.7%, respectively. To ensure adequate sampling, we restricted our attention to percent GC contents from 32.5% to 60%; each pair of nick sites in this range is sampled at least 10 times in our experiment.

Figure 2 summarizes the resulting data set, which contains 50 493 547 measurements obtained from single molecules of DNA. The trend in percent GC content at fixed N_{kbp} reflects the sequence of the human genome, which is AT rich. The trend in N_{kbp} at fixed percent GC content arises because each DNA molecule [e.g., Fig. 1(d)] will contribute many measurements with short distances between nick sites but only a few measurements at long distances.

Figure 3 shows how the fractional extension between each pair of nick sites depends on the percent GC content and the genomic distance N_{kbp} between those nick sites. We report our results here in terms of the fractional

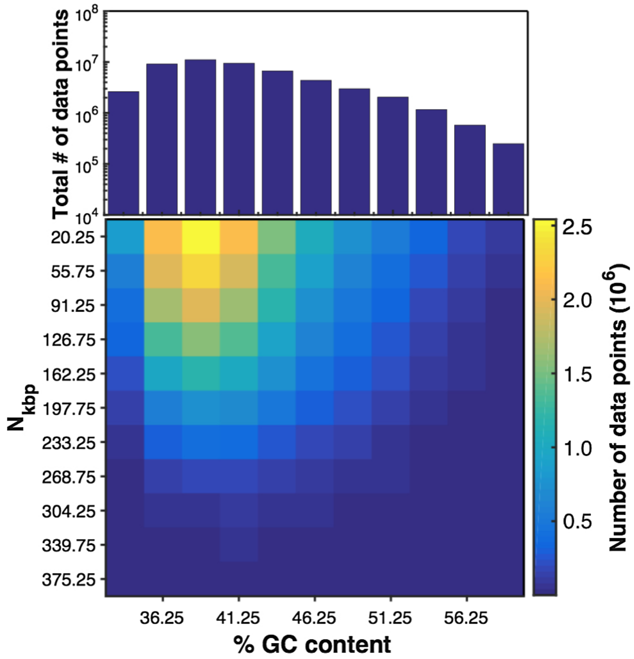

FIG. 2. Heat map of the number of measurements of extension using a bin size of 2.5% for percent GC content and 35.5 kbp for the number of kilobase pairs between nick sites, N_{kbp}. The tick labels on the left y axis of N_{kbp} and on the bottom x axis of percent GC content indicate the midpoints of the bins. The upper histogram presents the total number of data points in each percent GC content bin.

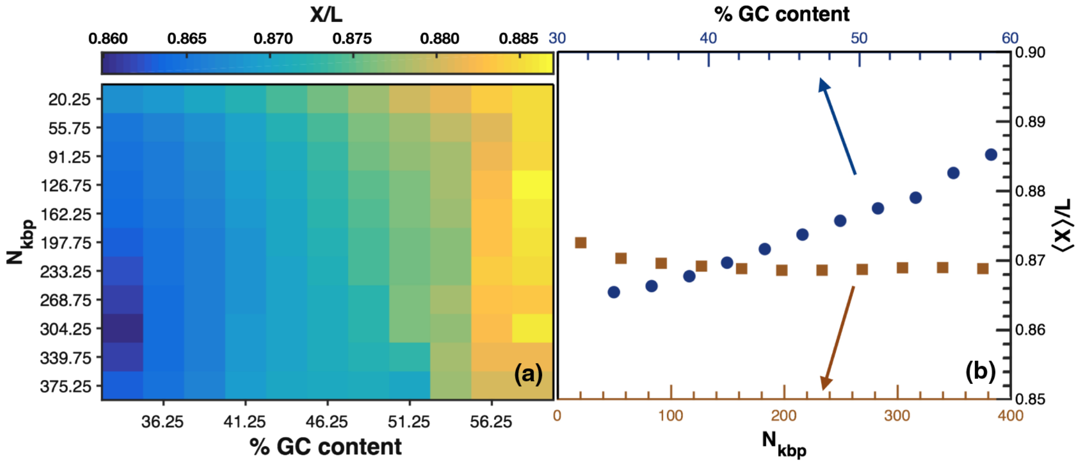

FIG. 3. (a) Heat map of the fractional extension using a bin size of 2.5% for percent GC content and 35.5 kbp for the number of kilobase pairs between nick sites, N_{kbp}. The tick labels indicate the midpoints of the bins. (b) Average fractional extension as a function of percent GC content (blue circles) and N_{kbp} (brown squares). The notation $\langle X\rangle$ indicates averaging over either percent GC content or N_{kbp}.

extension X/L, where X is the DNA extension measured between a pair of nick sites, assuming that the contour length L can be obtained from the 0.34 nm rise in B-DNA. While high levels of YOYO intercalation can increase L [5], the effect should be small at our low dye loading. We will address any systematic errors introduced by this assumption later.

An analysis of variance (ANOVA) and Tukey's minimum significant difference test for the data in Fig. 3(a) indicate that the increase in the average factional extension $\langle X\rangle/L$ as the percent GC content increases [blue circles in Fig. 3(b)] is statistically significant [24]. In contrast, $\langle X\rangle/L$ when binned by N_{kbp} is not statistically different [brown squares in Fig. 3(b)] [24]. Further statistical analysis [24] indicates that the results are independent of the number of nicking sites on a given molecule.

We thus proceed by binning the data only with respect to the percent GC content. Figure 3(b) shows that the change of the average fractional extension $\langle X\rangle/L$ vs percent GC content is small, around 2%. However, this small change is crucial to our genomic strategy. Genome mapping is required to obtain the measurements of L from the DNA sequence. The mapping method is robust to such small changes in extension, since it is a *de novo* method that relies on a pattern recognition [36]. Drawing a statistically meaningful conclusion, though, requires precise measurements of $\langle X\rangle/L$. Figure 2 indicates that each of these percent GC content bins contains between 10^5 and 10^7 measurements. As a result, the standard error of the average extension, $\langle X\rangle$, within a given percent GC content bin is very small.

Simulations of channel-confined wormlike chains [37,38] indicate that, for the fractional extensions in Fig. 3(b), the chain lies within the Odijk regime [10]. The corresponding fractional extension is predicted to be [10,39]

$$\langle X\rangle/L = 1 - 0.18274(D_{\mathrm{eff}}/l_p)^{2/3}, \quad (1)$$

where D_{eff} is the effective channel size available to the chain. For very small channels, such as those used here, the exact value of D_{eff} is not obvious due to the electrostatic interactions between the DNA and the channel walls [3,40]. However, we would expect those interactions to be independent of the sequence. To proceed, we adopt the standard approximation [37] of $D_{\mathrm{eff}} = D - w$, where $w = 7.6$ nm is the Stigter effective width [41] for our 48 mM buffer. As was the case with L, we will address any systematic errors from this assumption shortly. Inverting Eq. (1) yields the persistence length.

The sequence dependence of the DNA persistence length can be explained by modeling the DNA as a statistical terpolymer, illustrated in the inset in Fig. 4 and described in more detail in Supplemental Material [24]. The particular sequence of the DNA is replaced by an effective sequence where a G—C bond is replaced by S (strong hydrogen bonding) and an A—T bond is replaced by W (weak hydrogen bonding). The bending energy depends not on each base itself but on the sequence of dinucleotide pairs [11]: E_{SS}, E_{SW}, and E_{WW}. Previously, Hogan, LeGrange, and Austin measured these bending energies by triplet state anisotropy decay [11]. We constrain the present model by the ratio of the bending energies obtained in these experiments: $E_{SW}/E_{SS} = 1.4/2.9$ and $E_{WW}/E_{SS} = 0.82/2.9$ [11]. The persistence length at large length scales emerges from the local bending energies. As such, the relevant bending energy is the weighted average of the dinucleotide pairs in the sequence:

$$E = \sum_{i,j} p_{ij} E_{ij}, \quad (2)$$

where $(i,j) \in (S,W)$. Denoting the percent GC content (i.e., the probability of locating a G or C base) by γ, the probabilities p_{ij} of observing particular dinucleotide pairs in a statistical terpolymer are $p_{WW} = (1-\gamma)^2$, $p_{SW} = p_{WS} = \gamma(1-\gamma)$, and $p_{SS} = \gamma^2$, leading to the bending energy $E = E_{WW}(1-\gamma)^2 + 2E_{SW}\gamma(1-\gamma) + E_{SS}\gamma^2$. Assuming that the surface moment of inertia I_s is independent of the sequence, the intrinsic persistence length is given by $l_{p,0} = EI_s/k_BT$ [11]. The polyelectrolyte theory [7,42] further requires that the persistence length include an electrostatic contribution $l_{p,\mathrm{el}}$ due to the screening of backbone charges by the counterions in solution. We assume that all sequences are affected by electrostatics in the same manner, since they arise from the acidic backbone. By fitting to experimental data for λ-DNA, Dobrynin [7] obtained the empirical formula

$$l_p[\mathrm{nm}] = l_{p,0} + l_{p,\mathrm{el}} = 46.1 + \frac{1.9195}{\sqrt{I[\mathrm{M}]}}, \tag{3}$$

where I is the ionic strength. Using $\gamma = 0.4986$ for λ-DNA yields $E_{SS}I_s/k_BT = 82.2$ nm [24]. As a result, the statistical terpolymer model predicts [24]

$$l_p[\mathrm{nm}] = (23 + 33\gamma + 26\gamma^2) + \frac{1.9195}{\sqrt{I[\mathrm{M}]}}. \tag{4}$$

This is the key result of our analysis and extends Dobrynin's result for the GC-even genome of λ-phage DNA to the range of sequences commonly found in human DNA.

Figure 4 shows that Eq. (4) (dashed line) captures the trend in persistence length as a function of percent GC content. As noted previously, there are systematic errors due to the intercalation of YOYO dye (which affects L) and the DNA-wall electrostatic interactions (which affect D_{eff}). It is also possible that there is an additional source of systematic error from the effect of intercalation on the persistence length, but there is a growing body of systematic experimental work [43,44] indicating that intercalation does not affect the persistence length. These systematic errors should affect all sequences in the same manner, so they would shift the prediction of the model up or down but would not change the curvature. Indeed, Fig. 4 shows that we can bring the model into agreement with the experiments by assuming $D_{\mathrm{eff}} = 30.1$ nm (open circles in Fig. 4), which is certainly within reason based on the uncertainty in the DNA-wall interactions [3,40] and the accuracy of the SEM characterization of such a large array of channels.

To check the accuracy of assuming a random sequence, we also computed the dinucleotide composition between pairs of nick sites from the DNA sequences that lie within a given percent GC content bin and then recomputed the predictions of the model by replacing the probabilities in Eq. (2) with those data. Figure 4 shows that accounting for the exact DNA sequence (orange diamonds in Fig. 4), rather than assuming a random sequence with a particular averaged percent GC content (dashed line in Fig. 4), hardly affects the result.

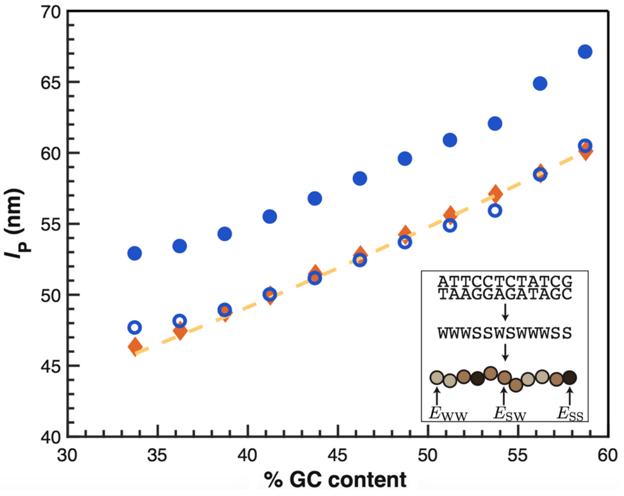

FIG. 4. Persistence length as a function of the percent GC content. Blue circles are experimental data using $D_{\mathrm{eff}} = D - w = 33.4$ nm (solid circles) and $D_{\mathrm{eff}} = 30.1$ nm (open circles). The dashed line is the statistical terpolymer model prediction in Eq. (4), and the orange diamonds are the model predictions using the average dinucleotide composition in each percent GC content bin. Inset: Statistical terpolymer model. The DNA sequence is converted first into a sequence of strong (G—C) and weak (A—T) hydrogen bonds. The persistence length is computed from the resulting sequence of dinucleotide pairs (WW, SW, or SS) based on their respective bending energies E_{ij}, where $i, j = (S, W)$.

We also examined whether the accuracy of the model could be improved with the ten-dinucleotide model of Geggier and Vologodskii [12] but found that it did not agree with our data [24]. This outcome is expected, as the data set used to parameterize that model specifically excluded sequences with a strong intrinsic curvature, which are scattered throughout the human genome.

One untested assumption in our model is the exclusive incorporation of sequence effects into the intrinsic persistence length. It is relatively straightforward, albeit tedious, to test this assumption by repeating the present experiments at different ionic strengths [45,46]. We are optimistic that such experiments will validate Eq. (4), as previous experiments on confined DNA [46] provide convincing evidence that the dependence on the ionic strength is correct and electrostatic interactions should govern long-range interactions.

In summary, we have demonstrated that the persistence length of long DNA has a remarkable dependence on the underlying sequence. We are optimistic that the model proposed in Eq. (4) will prove useful for the quantitative analysis of DNA-based experiments.

This work was supported by NIH R01-HG006851. We thank Damini Gupta for assistance in the data analysis. J. G. R. and H. C. are employees of BioNano Genomics, which is commercializing nanochannel genome mapping.

H.-M. C. and J. G. R. contributed equally to this work.

*dorfman@umn.edu

[1] T. T. Perkins, D. E. Smith, and S. Chu, Science **264**, 819 (1994); **276**, 2016 (1997); D. E. Smith, T. T. Perkins, and S. Chu, Phys. Rev. Lett. **75**, 4146 (1995); E. S. G. Shaqfeh, J. Non-Newtonian Fluid Mech. **130**, 1 (2005).

[2] J. O. Tegenfeldt, C. Prinz, H. Cao, S. Chou, W. W. Reisner, R. Riehn, Y. M. Wang, E. C. Cox, J. C. Sturm, P. Silberzan, and R. H. Austin, Proc. Natl. Acad. Sci. U.S.A. **101**, 10979 (2004); W. Reisner, K. J. Morton, R. Riehn, Y. M. Wang, Z. Yu, M. Rosen, J. C. Sturm, S. Y. Chou, E. Frey, and R. H. Austin, Phys. Rev. Lett. **94**, 196101 (2005); L. Dai, C. B. Renner, and P. S. Doyle, Adv. Colloid Interface Sci. **232**, 80 (2016).

[3] W. Reisner, J. N. Pedersen, and R. H. Austin, Rep. Prog. Phys. **75**, 106601 (2012).

[4] K. D. Dorfman, S. B. King, D. W. Olson, J. D. P. Thomas, and D. R. Tree, Chem. Rev. **113**, 2584 (2013).

[5] K. D. Dorfman, Rev. Mod. Phys. **82**, 2903 (2010).

[6] C. Bustamante, J. F. Marko, E. D. Siggia, and S. Smith, Science **265**, 1599 (1994).

[7] A. V. Dobrynin, Macromolecules **38**, 9304 (2005).

[8] E. 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).

[9] W. F. Reinhart, J. G. Reifenberger, D. Gupta, A. Muralidhar, J. Sheats, H. Cao, and K. D. Dorfman, J. Chem. Phys. **142**, 064902 (2015).

[10] T. Odijk, Macromolecules **16**, 1340 (1983).

[11] M. Hogan, J. LeGrange, and B. Austin, Nature (London) **304**, 752 (1983).

[12] S. Geggier and A. Vologodskii, Proc. Natl. Acad. Sci. U.S.A. **107**, 15421 (2010).

[13] G. S. Freeman, D. M. Hinckley, J. P. Lequieu, J. K. Whitmer, and J. J. de Pablo, J. Chem. Phys. **141**, 165103 (2014).

[14] B. Audit, C. Vaillant, A. Arneodo, Y. d'Aubenton Carafa, and C. Thermes, J. Mol. Biol. **316**, 903 (2002).

[15] T. J. Richmond and C. A. Davey, Nature (London) **423**, 145 (2003).

[16] E. Segal, Y. Fondufe-Mittendorf, L. Chen, A. Thåström, Y. Field, I. K. Moore, J.-P. Z. Wang, and J. Widom, Nature (London) **442**, 772 (2006).

[17] G. S. Freeman, J. P. Lequieu, D. M. Hinckley, J. K. Whitmer, and J. J. de Pablo, Phys. Rev. Lett. **113**, 168101 (2014).

[18] C. Rivetti, M. Guthold, and C. Bustamante, J. Mol. Biol. **264**, 919 (1996).

[19] P. Cong, L. Dai, H. Chen, J. R. C. van der Maarel, P. S. Doyle, and J. Yan, Biophys. J. **109**, 2338 (2015).

[20] D. MacDonald, K. Herbert, X. Zhang, T. Polgruto, and P. Lu, J. Mol. Biol. **306**, 1081 (2001).

[21] J. S. Mitchell, J. Glowacki, A. E. Grandchamp, R. S. Manning, and J. H. Maddocks, J. Chem. Theory Comput. **13**, 1539 (2017).

[22] L. Fetters, D. Lohse, T. Richter, T. Witten, and A. Zirkelt, Macromolecules **27**, 4639 (1994).

[23] D. R. Tree, A. Muralidhar, P. S. Doyle, and K. D. Dorfman, Macromolecules **46**, 8369 (2013).

[24] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevLett.119.227802 for (i) additional experimental methods, (ii) an ANOVA analysis of Fig. 3 and additional statistical analysis, (iii) a comparison with the model of Ref. [12], and (iv) a detailed derivation of the statistical terpolymer model, which includes Refs. [25–32].

[25] A. C. Y. Mak, Y. Y. Y. Lai, E. T. Lam, T. P. Kwok, A. K. Y. Leung, A. Poon, Y. Mostovoy, A. R. Hastie, W. Stedman, T. Anantharaman, W. Andrews, X. Zhou, A. W. C. Pang, H. Dai, C. Chu, C. Lin, J. J. K. Wu, C. M. L. Li, J. W. Li, A. K. Y. Yim *et al.*, Genetics **202**, 351 (2016).

[26] J. M. Zook, D. Catoe, J. McDaniel, L. Vang, N. Spies, A. Sidow, Z. Weng, Y. Liu, C. E. Mason, N. Alexander, E. Henaff, A. B. McIntyre, D. Chandramohan, F. Chen, E. Jaeger, A. Moshrefi, K. Pham, W. Stedman, T. Liang, M. Saghbini *et al.*, Sci. Data **3**, 160025 (2016).

[27] P. Licinio and J. C. O. Guerra, Biophys. J. **92**, 2000 (2007).

[28] C. Yoon, G. G. Privé, D. S. Goodsell, and R. E. Dickerson, Proc. Natl. Acad. Sci. U.S.A. **85**, 6332 (1988).

[29] M. Dlakic and R. E. Harrington, J. Biol. Chem. **270**, 29945 (1995).

[30] I. Brukner, S. Susic, M. Dlakic, A. Savic, and S. Pongor, J. Mol. Biol. **236**, 26 (1994).

[31] A. A. Travers, Phil. Trans. R. Soc. A **362**, 1423 (2004).

[32] J. Bednar, P. Furrer, V. Katritch, A. Z. Stasiak, J. Dubochet, and A. Stasiak, J. Mol. Biol. **254**, 579 (1995).

[33] J. G. Reifenberger, K. D. Dorfman, and H. Cao, Analyst **140**, 4887 (2015).

[34] H. Cao, A. R. Hastie, D. Cao, 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**, 1 (2014).

[35] S. Laib, R. M. Robertson, and D. E. Smith, Macromolecules **39**, 4115 (2006).

[36] A. Valouev, Ph.D. thesis, University of Southern California, 2006.

[37] Y. Wang, D. R. Tree, and K. D. Dorfman, Macromolecules **44**, 6594 (2011).

[38] A. Muralidhar, D. R. Tree, and K. D. Dorfman, Macromolecules **47**, 8446 (2014).

[39] T. W. Burkhardt, Y. Yang, and G. Gompper, Phys. Rev. E **82**, 041801 (2010).

[40] G. K. Cheong, X. Li, and K. D. Dorfman, Phys. Rev. E **95**, 022501 (2017).

[41] D. Stigter, Biopolymers **16**, 1435 (1977).

[42] T. Odijk, J. Polym. Sci., Polym. Phys. Ed. **15**, 477 (1977); J. Skolnick and M. Fixman, Macromolecules **10**, 944 (1977).

[43] K. Günther, M. Mertig, and R. Seidel, Nucleic Acids Res. **38**, 6526 (2010).

[44] B. Kundukad, J. Yan, and P. S. Doyle, Soft Matter **10**, 9721 (2014).

[45] W. Reisner, J. P. Beech, N. B. Larsen, H. Flyvbjerg, A. Kristensen, and J. O. Tegenfeldt, Phys. Rev. Lett. **99**, 058302 (2007).

[46] C. C. Hsieh, A. Balducci, and P. S. Doyle, Nano Lett. **8**, 1683 (2008).