Supporting Information

Reisner et al. 10.1073/pnas.1007081107

SI Text

Supplementary Materials and Methods. ***Fabrication.*** The devices were fabricated on fused silica wafers (HOYA) via a three-stage process combining electron beam and UV contact lithography. A 200-μm-long array of 120–150-nm-wide nanochannels spaced 2 μm apart was defined using electron beam lithography (JEOL) in ZEP520A resist (1) and then transferred to the silica substrate via CF_4:CHF_3 reactive ion etching (RIE). Contact UV lithography was then used to expose a 350-μm-long and 50-μm-wide nanoslit in photoresist running perpendicular to the nanochannel array. The slit was etched using RIE to a depth of 30 nm. Where the slit and the nanochannel array intersected, an array of nanogroove structures was thus formed; see Fig. 1. In order to introduce buffer into the nanoslit and nanochannels, a last UV contact lithography and etching step was used to define a 50-μm-wide microchannel 1 μm deep in U-shaped arms adjoining the nanoslit and nanochannels (Fig. 1). Loading holes were sandblasted in the reservoirs (eight total), and the chip was sealed using direct silica-silica bonding to a 150-μm-thick fused silica cover glass (Valley Design) so that high numerical aperture oil immersion objectives could be used. Etch depths were measured using a profilometer.

Materials. The experiments were performed with λ-phage DNA (48.5 kbp, $L = 16.5$ μm, New England BioLabs), T4GT7 DNA (166 kbp, $L = 56.4$ μm, Nippon Gene), T7 DNA (39.9 kbp, $L = 13.6$ μm, Yorkshire Bioscience), and a BAC construct from chromosome 12 (RP11-125C7, 152 kbp, $L = 51.7$ μm, position 12q21.31). The BAC contains an 11.6-kbp cloning vector (pBACe3.6). The DNA was dyed with YOYO®-1 fluorescent dye (Invitrogen) at a concentration of 1 dye molecule per every 5 base pairs. The running buffer consisted of 0.05 × TBE (4.5 mM Tris, 4.5 mM boric acid, and 0.1 mM EDTA) plus 10 mM NaCl, diluted with formamide (Sigma) to the volume fraction specified. In addition, we used an antiphotobleaching system consisting of a reducing agent, 3% β-mercaptoethanol and an oxygen-scavenging system 4 mg/mL β-D-glucose, 0.2 mg/mL and 0.04 mg/mL catalase (added to loading buffer, which was then diluted with formamide). The single-molecule measurements were conducted with a fluorescence videomicroscopy system incorporating a Nikon Eclipse TE2000 inverted microscope, 100× N.A. 1.4 oil immersion objective and an EMCCD camera (Andor iXon and Photometrics Cascade II).

BAC Preparation. BAC clones were cultured in LB medium and 12.5 μg/mL chloramphenicol at 37 °C overnight in a shaking incubator. Then 1.5 mL of the culture was transferred to an eppendorf tube and spun down. The supernant was discarded, and the procedure was repeated with another 1.5-mL culture using the same tube. The pellet was resuspended in 250 μL of P1 buffer (50 mM TrisCl, pH 8.0, 10 mM EDTA, 100 μg/mL RNase A) and left on ice for 10 min followed by addition of 250 μL of P2 Buffer [200 mM NaOH, 1% SDS (wt/vol)]. The tube was then inverted 10–15 times and left on ice for 5 min. Next, 350 μL of P3 solution (3.0 M NaAc, pH 4.8) was added, and the tube was inverted 10–15 times immediately and left on ice for 15 min. Then the tube was spun at 19,800 g for 10 min at room temperature. The supernatant (approximately 850 μL) was transferred to a new tube and treated with RNase A (15 μL; 10 mg/mL) at 37 °C for 30–45 min.. Following the phenol/chloroform extraction, the upper layer was transferred to a new tube. P3 Buffer was added at volume of 1/9th of the transferred amount, followed by addition of cold isopropanol at a final concentration of 50%. The tube was inverted smoothly 10–15 times and placed in −80 °C for a minimum of 30 min before being centrifuged at 4 °C for 30 min. The supernatant was discarded and 1 mL of 70% alcohol was added. Following a 1-min centrifugation, the alcohol was removed and the dried pellet was resuspended in 25–30 μL of distilled water.

DNA Loading Protocol. The molecules are brought from the microchannels into the nanochannels with a burst of high pressure (Fig. S1*A*). The molecules introduced in the nanochannels are then concentrated in the nanogrooves via the following procedure. Equally distributed positive pressure is applied to the four reservoirs adjoining the nanochannels, forcing buffer to circulate through the nanochannels and out into the nanoslit (Fig. S1 *A* and *B*). This flow pattern will cause nanochannel confined molecules on either side of the nanoslit to be symmetrically driven into the nanogrooves. As the nanoslit region is more confined than the nanogrooves (Fig. S1*B*), in order to escape from the nanogrooves into the nanoslit the molecules must cross through an entropic barrier (2). While at high enough pressure the flow will be sufficient to overcome the barrier, and molecules will be forced out of the nanogrooves, below a certain pressure threshold DNA will remain trapped in the nanogrooves. Consequently, for applied pressures below this threshold, the effect of the circulating flow will be to concentrate molecules in the center of the nanogroove array (Fig. S1*C* and Movie S1). While it is possible to run the devices with only the initial loading step, the concentration protocol consistently maximizes the number of molecules available for imaging in the microscope field of view. The physics of DNA transport across nanogroove arrays will be discussed in depth in an upcoming publication (3).

Time-Trace Rescaling. Once raw movies of denatured molecules are acquired, we normalize the time-trace plots of all molecules so that averaged single-molecule barcode profiles can be obtained. The first step is to align the molecule center of mass across all frames. We accomplish this by using correlation of the ith frame with the initial frame to obtain the translational offset of the ith frame relative to the initial frame. The second step is to "smooth out" longitudinal thermal fluctuations in the contour density that create a local distortion of the barcode structure. While using a single dilation factor to normalize the profiles works well, it is possible to improve the procedure by using local dilation/contraction factors. Thus, instead of using a single overall adjustment to normalize the molecule extension between plots, we use local adjustments, so that different positions along the molecule profile can receive different adjustments.

In practice we create a piecewise linear map M, defined by a series of dilation factors d_k, the slopes of the individual linear components of the map [so that M is a function of the d_k, for example, $M(d_k)$]. The map $M(d_k)$ will then operate on the profile $P_i(x_j)$ at the ith frame to create a profile $P'_i(x_j,d_k)$ (x_j is the jth pixel of the profile). The parameters d_k are chosen to minimize the least-squared difference Δ between the profile $P_i(x_j)$ and a template profile taken to be the profile at frame $i = 1$ ($P_1(x_j)$):

$$\Delta = \sum_{j=1}^{N} [P'_i(x_j,d_k) - P_1(x_j)]^2. \qquad \textbf{[S1]}$$

Using custom code written in Matlab, this procedure is applied to all frames, creating the rescaled time series shown in Fig. S2 *A* and *B*. Averaging over the rescaled frames, we obtain the

average barcode profile normalized to the local expansion/contraction present in the first frame.

We need, however, to obtain the average profile reflecting the true equilibrium conformation of the chain. This can be accomplished as follows: During the rescaling we save the maps $M_i(d_k)$ relating the profiles for each frame to the initial profile at frame $i = 1$. From these saved maps we can compute the average map $\langle M(d_k)\rangle$: This map relates the true equilibrium profile to the profile at $i = 1$. In order to correctly normalize the averaged rescaled profile to the true equilibrium chain conformation, we simply apply the inverse of this map $[\langle M(d_k)\rangle^{-1}]$.

Barcode Alignment. In order to create a consensus barcode, it is necessary to align the profiles. The first step is to find the translational overlap and profile orientation that maximizes the correlation between a profile and a template profile. We then apply a global dilation to minimize the squared difference between the profile and template (as a precaution to avoid forcing agreement, we do not apply local dilation maps to align profiles taken for different molecules). The dilation accounts for any overall difference in the profile scaling, typically adjusting the relative scaling of the two profiles by less than 10%. Examples of alignments are shown in Supplementary Fig. S2 *C–E*.

Barcode Registration with Theory. In this procedure, the theoretical barcode is created using an extension per base pair estimated from the measured stretching of λ-DNA and a "best-guess" helicity. A single experimental profile is aligned to the theory using the method described above for aligning experimental profiles to experimental profiles (Fig. 4 *C* and *D*). We term this experimental profile, aligned to theory, the "template." Additional experimental profiles are then aligned to the template to create a consensus barcode. We intentionally do not choose to align all the individual experimental profiles to theory in order to create a consensus profile that is based purely on alignments within the experimentally determined set of profiles. Finally, we refine the theoretical profile by a least-squares fit of the theory to the aligned consensus profile, updating the best-guess helicity to a value determined via the secondary fitting procedure. If the template used to construct the consensus is correctly aligned to the theory, then the resulting consensus will also have the correct registration with the true sequence.

Global BAC Alignment. The global alignment of RP11-125C7 to chromosome 12 was performed automatically by finding the position of the BAC that minimized the least-squared difference between the profile and the calculated genomic melting map. Let $P_T(x_j)$ be the theoretical profile and $P_{\exp}(x_j)$ be the experimental profile (with N pixels). The average of the experimental profile is denoted $\langle P_{\exp}\rangle$. Let the average of the theoretical profile, over a certain sequence window from $i+1$ to $i+N$ (equal in length to the experimental profile) be denoted $\langle P_T\rangle_{i,N}$. Then define the local mean subtracted profiles, normalized to the local profile standard deviations:

$$\Delta P_{\exp} = \frac{P_{\exp} - \langle P_{\exp}\rangle}{\langle (P_{\exp} - \langle P_{\exp}\rangle)^2\rangle}, \qquad \textbf{[S2]}$$

$$\Delta P_T = \frac{P_T - \langle P_T\rangle_{i,N}}{\langle (P_T - \langle P_T\rangle_{i,N})^2\rangle_{i,N}}. \qquad \textbf{[S3]}$$

With these definitions the estimator is defined as

$$\Delta(i) = \frac{1}{2N}\sum_{j=1}^{N}[\Delta P_T(i,x_{i+j}) - \Delta P_{\exp}(x_j)]^2 \qquad \textbf{[S4]}$$

with the index $i = 1\cdots n-N$, where n is the length of the calculated sequence. The use of profiles with mean local signal levels removed, normalized to the local standard deviation, ensures that the fitting procedure is most sensitive to the details of variation along the profiles (which provides the specificity of the match). The estimator was calculated for a range of helicities (i.e., temperatures) across the melting transition. The best-fitting helicity was found by choosing the value that led to a global minimum of Δ. The measured stretching of the BAC-DNA was used to calibrate the extension per base pair. The global minumum of Δ along chromosome 12, which corresponds to the best-fitting BAC position, is shown in Fig. S3.

Statistical Significance of Fit. To determine the statistical significance of the given identification, we need to determine how many matches of equivalent quality we would expect to obtain randomly at sequence positions unrelated to the underlying BAC sequence. By "equivalent quality," we mean matches that would lead to an equivalent value of the least-squares estimator used (Eq. **S4**). Thus, the key question is to evaluate the distribution of estimator values expected for our search (i.e., at all positions across the sequence, at varying values of temperature). This distribution of estimator values constitutes the "background" to our search. If, based on this distribution, we find that the estimator value for the best-fit parameters should have a very low corresponding background, then the estimator value is significant (i.e., has a low probability of resulting from a random misalignment). We determine the background distribution on the estimator by sampling the estimator at sequence positions and helicity values across our search range (omitting the region of DNA that gives the best fit). The distribution of the estimator, shown in Fig. S4, is to close approximation a Gaussian centered on 1. Note that, using the definition of $\Delta P_{\exp}$ and ΔP_T (Eqs. **S2** and **S3**) the estimator (Eq. **S4**) can be written

$$\Delta(i) = 1 - \frac{1}{N}\sum_{j=1}^{N}\Delta P_T(i,x_{i+j})\Delta P_{\exp}(x_j). \qquad \textbf{[S5]}$$

The Gaussian arises via the central limit theorem as the sum contains essentially N random numbers, except when there is a match (where it is close to 1, so that $\Delta \sim 0$). Upon closer examination of the fit, it appears that the Gaussian fit fails in the distribution tails: This deviation can be captured by a quartic correction term:

$$P(\Delta) = Ae^{\left(-\frac{(\Delta-\Delta_o)^2}{2\sigma^2} - \frac{(\Delta-\Delta_o)^4}{24\eta^4}\right)}. \qquad \textbf{[S6]}$$

A best fit with $\Delta_o = 1$ yields $\sigma = 0.240 \pm 0.001$ and $\eta = 0.225 \pm 0.001$. The quartic correction likely arises due to effects of finite N: The fragment length is not quite large enough to remove lingering remnents of the original distribution of the $\Delta P_T(i,x_{i+j})$ $\Delta P_{\exp}(x_j)$. We can use Eq. **S6** to estimate the probability that we would randomly obtain fits with extremely small values of Δ. In particular, we find that the best fit of the experimental profile to theory yields a $\Delta_f = 0.131$ (see Fig. S3). Integrating Eq. **S6** from 0 to $\Delta_f = 0.131$ yields an expected $10^{-(2.4\pm0.1)}$ matches of equivalent or higher quality. The match we obtain is clearly significant. Moreover, this result suggests that the match would be significant even for a search performed across the entire human genome (with a 23× larger search-space).

Sequences and Melting Probability Profiles. Sequences used were downloaded from the NCBI GenBank and the UCSC genome browser (Assembly hg17). In particular, the T4GT7 sequence was obtained from T4 by deletion of a 3.256-bp segment between sites 165,255 and 168,510. Melting probability profiles were then calculated from the sequences using the Web site www.stitchprofiles.uio.no documented in refs. 4 and 5. The melting probability map for chromosome 12 was too large to be obtained from the server and was generously provided by the Hovig group.

1. Rai-Choudhury P, ed (1997) *SPIE Handbook of Microlithography, Micromachining and Microfabrication* (SPIE, Bellingham, WA), Vol 1, Sect 2.7.2.4.
2. Han J, Turner SW, Craighead HG (1999) Entropic trapping and escape of long DNA molecules at submicron size constriction. *Phys Rev Lett* 83:1688–1691.
3. Mikkelsen MB, Reisner W, Flyvbjerg H, Kristensen A. Propagation modes of pressure-driven DNA in nanogroove arrays: Complex dynamical behavior in a nanofabricated device. Manuscript submitted to Phys. Rev. Lett.
4. Tøstesen E, Jerstad GI, Hovig E. (2005) Stitchprofiles.uio.no: Analysis of partly melted DNA conformations using stitch profiles. *Nucl Acid Res* 33:W573–W576.
5. Tøstesen E, Liu F, Jenssen TK, Hovig E. (2003) Speed-up of DNA melting algorithm with complete nearest neighbor properties. *Biopolymers* 70364–376.

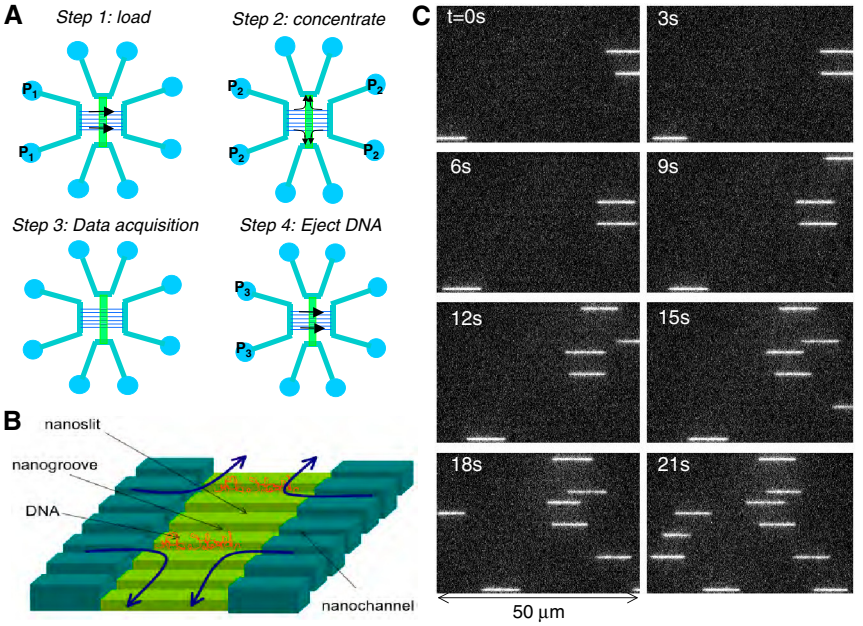


Fig. S1. Device loading protocol. (*A*) Molecules are loaded (1) with pressure $P_1 \sim 2$ bar and resulting DNA velocity $v_{DNA} \sim 250$ μm/s; then concentrated (2) with P2 % 0.3 bar ($v_{DNA} \sim 5$ μm/s), imaged (3) in equilibrium and ejected (4) with $P_3 \sim 1$ bar. (*B*) Three-dimensional schematic of the circulating flow pattern (purple arrows) created by applying equally distributed pressure to the four reservoirs adjoining the nanochannel array. The effect of this flow is to symmetrically drive molecules from the nanochannels into the nanogrooves where they will remain trapped by the entropic barrier between the nanogrooves and nanoslit. (*C*) Time series of λ-DNA being concentrated in nanogrooves (see Movie S1).

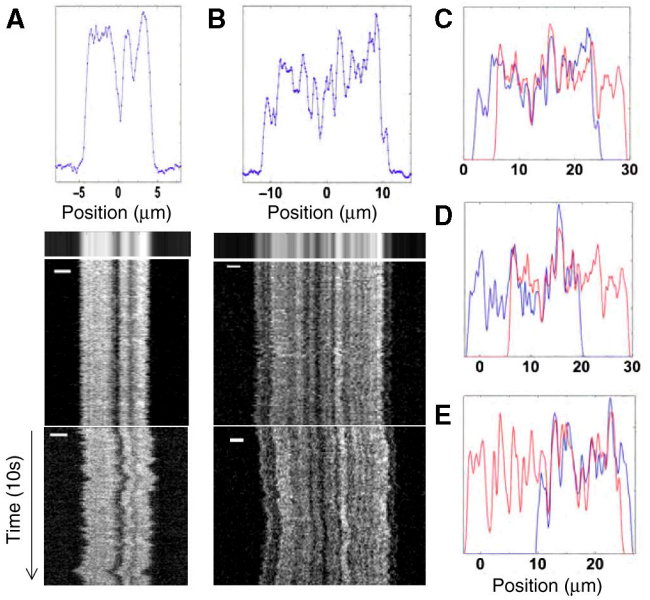


Fig. S2. Time-trace rescaling and correlation analysis. (*A*) An example of the rescaling procedure for λ-phage DNA. (*B*) Rescaling procedure for a BAC RP11-125C7 molecule. (*A* and *B*: *Bottom*) Raw time trace of molecule (integrated intensity transverse to channel for each recorded frame). (*A* and *B*: *Middle*) Rescaled time trace. (*A* and *B*: *Top*) Intensity profile obtained by averaging over rescaled frames. Barcode shown below plot is graphed data displayed as a grayscale plot. (*C* and *D*) Examples of RP11-125C7 (blue) molecules aligned by correlation to template (red) of identical sequence. (*E*) Raw T4GT7 fragment (blue) aligned by correlation. The scale bar in all images is 2 μm.

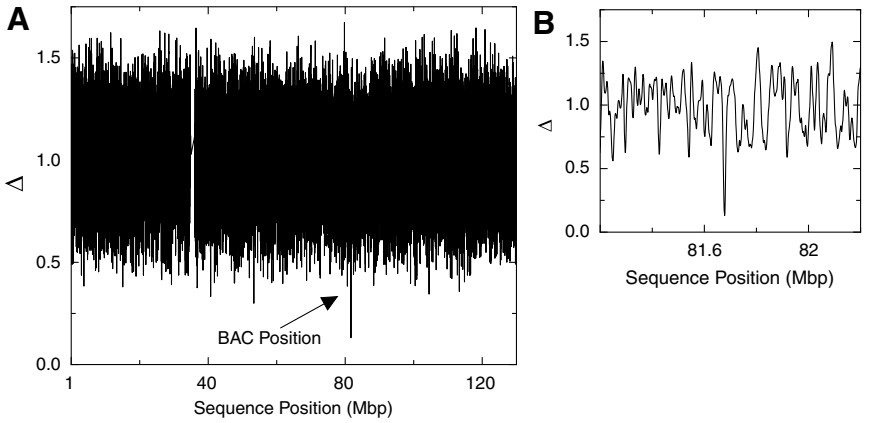


Fig. S3. Estimator as function of sequence position. (*A*) The squared difference Δ between the experimental profile for the BAC and the theoretical profile as a function of sequence position at the best-fitting temperature (59 °C). The fitting position of the BAC corresponds to the clear minimum in the figure: This minimum is, in fact, a global minimum across the entire search space, including profiles calculated at multiple temperatures (corresponding to helicity values ranging from 20 to 90%). Note that the break in the Δ fluctuations, evident at around 37 Mbp, is caused by a gap in the sequence assembly close to the centromere, e.g., the sequence is unknown in this region. The melting probability map assigns values of 1 to the unknown gap region (which are removed from the analysis). (*B*) Estimator plotted vs. sequence position for 1 Mbp of sequence about best BAC fitting position.

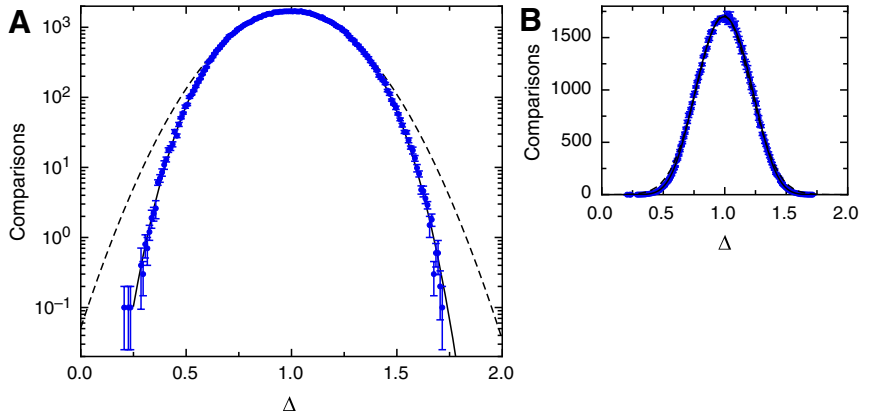
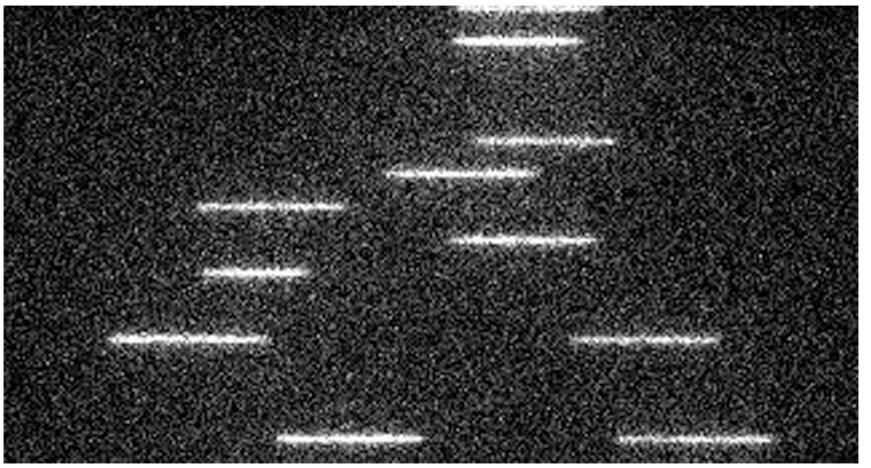


Fig. S4. Distribution of estimator. (*A*) The distribution of the squared difference Δ between the experimental profile for the BAC and the theoretical profile (on a log-linear scale). (*B*) The distribution of Δ on a linear scale. The y axis shows the total number (counts per bin) of comparisons of the experimental profile to theory made over a search range that includes a variation in helicity of 20–90% and all of chromosome 12. The data were initially binned at a $\delta\Delta = 10^{-3}$ and then averaged over bins of size $\delta\Delta = 10^{-2}$: The error bars show the error on the mean count per bin. The dashed curve is a Gaussian fit; the bold curve is a Gaussian fit with a quartic correction term.



Movie S1. Movie of λ-DNA being concentrated in nanogroove array (see Fig. S1). The movie duration is 34 s; the movie field is 51 × 27 μm. The buffer is 0.05× TBE +10mM NaCl diluted with 50% formamide (by volume). The DNA is stained with YOYO®-1 at a ratio of 1 dye molecule to every 5 bp.
Movie S1 (AVI)