CrossMark click for updates

Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes

Adrian L. Sanborn[a,b,c,1], Suhas S. P. Rao[a,d,1], Su-Chen Huang[a], Neva C. Durand[a,2], Miriam H. Huntley[a,2], Andrew I. Jewett[a,2], Ivan D. Bochkov[a], Dharmaraj Chinnappan[a], Ashok Cutkosky[a], Jian Li[a,b], Kristopher P. Geeting[a], Andreas Gnirke[e], Alexandre Melnikov[e], Doug McKenna[a,f], Elena K. Stamenova[a,e], Eric S. Lander[e,g,h,3], and Erez Lieberman Aiden[a,b,e,3]

[a]The Center for Genome Architecture, Baylor College of Medicine, Houston, TX 77030; [b]Center for Theoretical Biological Physics, Rice University, Houston, TX 77030; [c]Department of Computer Science, Stanford University, Stanford, CA 94305; [d]School of Medicine, Stanford University, Stanford, CA 94305; [e]Broad Institute of MIT and Harvard, Cambridge, MA 02139; [f]Mathemaesthetics, Inc., Boulder, CO 80306; [g]Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139; and [h]Department of Systems Biology, Harvard Medical School, Boston, MA 02115

Contributed by Eric S. Lander, September 18, 2015 (sent for review July 27, 2015; reviewed by Frank Alber, Ido Amit, Roger D. Kornberg, Corina E. Tarnita, and Shing-Tung Yau)

We recently used in situ Hi-C to create kilobase-resolution 3D maps of mammalian genomes. Here, we combine these maps with new Hi-C, microscopy, and genome-editing experiments to study the physical structure of chromatin fibers, domains, and loops. We find that the observed contact domains are inconsistent with the equilibrium state for an ordinary condensed polymer. Combining Hi-C data and novel mathematical theorems, we show that contact domains are also not consistent with a fractal globule. Instead, we use physical simulations to study two models of genome folding. In one, intermonomer attraction during polymer condensation leads to formation of an anisotropic "tension globule." In the other, CCCTC-binding factor (CTCF) and cohesin act together to extrude unknotted loops during interphase. Both models are consistent with the observed contact domains and with the observation that contact domains tend to form inside loops. However, the extrusion model explains a far wider array of observations, such as why loops tend not to overlap and why the CTCF-binding motifs at pairs of loop anchors lie in the convergent orientation. Finally, we perform 13 genome-editing experiments examining the effect of altering CTCF-binding sites on chromatin folding. The convergent rule correctly predicts the affected loops in every case. Moreover, the extrusion model accurately predicts in silico the 3D maps resulting from each experiment using only the location of CTCF-binding sites in the WT. Thus, we show that it is possible to disrupt, restore, and move loops and domains using targeted mutations as small as a single base pair.

genome architecture | molecular dynamics | CTCF | chromatin loops | CRISPR

The human genome is over 2 m long, yet it must fold up to fit inside a nucleus that is only a few microns wide. At the smallest scale, this folding is well characterized: dsDNA helices wrap around histone proteins, forming a nucleosome every ∼200 bp (a beads-on-a-string configuration known as the "10-nm fiber") (1). At larger scales, the physical structure of chromatin is more mysterious.

One common hypothesis is that the 10-nm fiber is organized into a higher order structure known as the "30-nm fiber," which has been observed in vitro but not in vivo (2). In the most well-known model, individual nucleosomes are wound about a central cavity that runs axially along the fiber's length. Every six nucleosomes (roughly 1 kb) correspond to a full turn about this axial cavity, creating a solenoidal structure with a 30-nm diameter.

Another common notion, dating back to the 1970s, is that the human genome is partitioned into domains. These studies have relied on many experimental modalities, such as chromatin sedimentation (3); fluorescence microscopy (4); and, in the past several years, genome-wide DNA proximity ligation data generated using Hi-C (5–8). The internal structure of domains is not well understood (5–10).

A third feature of chromatin folding is the formation of loops, which bring pairs of genomic sites that lie far apart along the linear genome into close spatial proximity (8, 11). Many aspects of chromatin looping are poorly understood, including how loops form.

We recently reported new contact maps of the human genome with a resolution of 1 kb (8). These maps were created by using in situ Hi-C, which couples DNA-DNA proximity ligation in intact nuclei (nuclear ligation assay) (12) with high-throughput sequencing (Fig. 1*A*). The maps, containing over 15 billion contacts, allowed us to annotate over 9,000 contact domains (median length = 185 kb), which are contiguous genomic intervals in which there is an enhanced probability of contact among all loci. The maps also allowed us to annotate nearly 10,000 loops. These loops typically lie

Significance

When the human genome folds up inside the cell nucleus, it is spatially partitioned into numerous loops and contact domains. How these structures form is unknown. Here, we show that data from high-resolution spatial proximity maps are consistent with a model in which a complex, including the proteins CCCTC-binding factor (CTCF) and cohesin, mediates the formation of loops by a process of extrusion. Contact domains form as a byproduct of this process. The model accurately predicts how the genome will fold, using only information about the locations at which CTCF is bound. We demonstrate the ability to reengineer loops and domains in a predictable manner by creating highly targeted mutations, some as small as a single base pair, at CTCF sites.

Author contributions: E.L.A. conceived of this project; A.L.S. led the development of all mathematical results; A.L.S. performed all simulations using a pipeline created by A.I.J. and A.C.; S.S.P.R. led the development of the 3D genome engineering pipeline; S.S.P.R. led the development of Hi-C^2, based on initial experiments by A.G. and A.M.; S.S.P.R., S.-C.H., and K.P.G. performed CRISPR experiments; S.S.P.R., S.-C.H., K.P.G., and E.K.S. performed Hi-C and Hi-C^2 experiments; S.-C.H., I.D.B., D.C., and E.K.S. performed microscopy experiments; J.L. analyzed fractal curves; D.M. constructed the Inside-Out Hilbert curve; A.L.S., S.S.P.R., N.C.D., M.H.H., E.S.L., and E.L.A. analyzed data; and A.L.S., S.S.P.R., E.S.L., and E.L.A. prepared the manuscript.

Reviewers: F.A., University of Southern California; I.A., Weizmann Institute; R.D.K., Stanford University School of Medicine; C.E.T., Princeton University; and S.-T.Y., Harvard University.

The authors declare no conflict of interest.

Freely available online through the PNAS open access option.

Data deposition: The data reported in this paper have been deposited in the Gene Expression Omnibus (GEO) database, www.ncbi.nlm.nih.gov/geo (accession no. GSE74072).

[1]A.L.S. and S.S.P.R. contributed equally to this work.

[2]N.C.D., M.H.H., and A.I.J. contributed equally to this work.

[3]To whom correspondence may be addressed. Email: lander@broadinstitute.org or erez@erez.com.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1518552112/-/DCSupplemental.

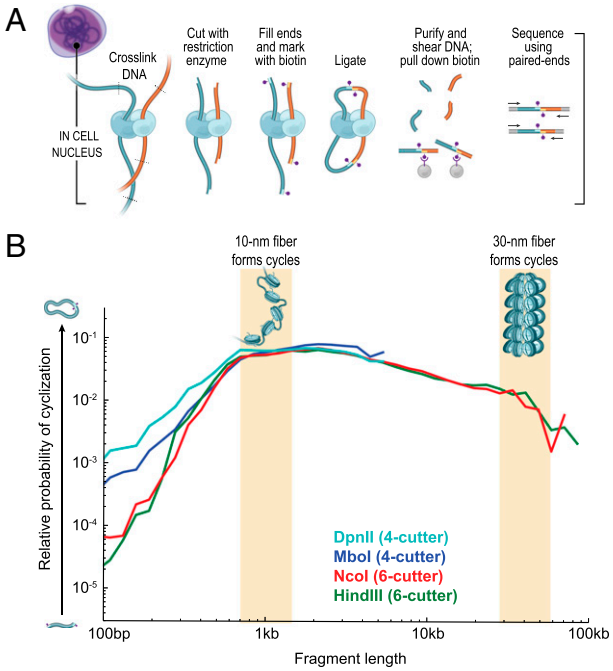


Fig. 1. Chromatin is bendable at the kilobase scale. (*A*) In situ Hi-C maps DNA-DNA contacts occurring in intact nuclei. Reprinted with permission from ref. 8. (*B*) Probability that a restriction fragment will bend to form a cycle as a function of fragment length. Results are shown for four restriction enzymes. The 30-nm fiber predicts a peak around 30 kb (*Right*, yellow shading), whereas the 10-nm fiber is consistent with the peak observed around 1 kb (*Left*, yellow shading).

between CCCTC-binding factor (CTCF) motifs in the convergent orientation (i.e., the motifs point toward one another), suggesting a "convergent rule" for loop formation. Notably, we found that many contact domains are also "loop domains," that is, contact domains whose boundaries are demarcated by the end points of a loop.

Here, we use our new maps to explore the physical structure of chromatin fibers, contact domains, and loops. First, we demonstrate that chromatin fibers are bendable at the kilobase scale, casting doubt on the widespread existence of 30-nm fibers in vivo. Next, we combine Hi-C data, molecular dynamics simulations, and a novel analog of McKean's dimension-doubling theorem for Brownian motion (13) to explore how chromatin fibers fold inside contact domains. After considering a series of models, we find that the data are best explained by a model where loops form through the extrusion of flexible chromatin fibers by a CTCF- and cohesin-associated complex. This model has many appealing features, and explains why loops tend not to overlap and why they only form between convergent CTCF motifs.

Finally, we use clustered regularly interspaced short palindromic repeat (CRISPR)-mediated genome editing to manipulate CTCF motifs at loop anchors (14). In all 13 cases examined, we find that the convergent rule correctly predicts which loops will disappear. Moreover, the extrusion model accurately predicts in silico the contact maps resulting from these loop engineering experiments, including the conditions under which domains disappear.

Results

Chromatin Is Bendable at the Kilobase Scale, Far Less Stiff than Predictions Based on a 30-nm Fiber. The stiffness of a fiber can be characterized by the Kuhn length, the minimum fiber length such that it is possible for the beginning and the end of the fiber segment to point in the same direction. Published estimates suggest a Kuhn length in the range of 30–60 kb for a 30-nm fiber (15).

To measure the Kuhn length of human chromatin in vivo experimentally, we examined the tendency of cross-linked chromatin fragments, formed during Hi-C's restriction digestion step, to form single-fragment DNA cycles during the subsequent proximity ligation step (16). Restriction fragments shorter than 200 bp (the size of a nucleosome) rarely formed cycles, suggesting that they were too stiff. Cyclization probability increased sharply for 200- to 800-bp fragments, and was relatively constant thereafter (Fig. 1*B*). The results were similar regardless of the restriction enzyme used in Hi-C (MboI, DpnII, HindIII, and NcoI), and with and without cross-linking or harsh detergents (*SI Appendix*, Fig. S1). These results suggest a Kuhn length of roughly 1 kb for chromatin fibers.

The results of cyclization analysis were consistent with two other approaches. First, we measured the probability, $I(s)$, of contact between two loci as a function of the genomic distance, s, between them. $I(s)$ is maximal at the Kuhn length of a polymer and decreases monotonically as s increases. Using our in situ Hi-C data, we were able to measure $I(s)$ reliably for the human genome at all distances larger than 5 kb (i.e., distances much longer than the typical 4-cutter fragment). The function declines monotonically at all distances probed, implying that the Kuhn length of chromatin is less than 5 kb (*SI Appendix*, Fig. S2*A*). Second, 40-kb-long loops were visually obvious in our initial report. At the specific loci involved, kilobase-length chromatin fibers must thus be capable of bending.

The Kuhn length observed in our data (≈1 kb) is incompatible with previous estimates of the Kuhn length for the 30-nm fiber. These results suggest that 30-nm fibers, if they exist, are rare in human nuclear chromatin in vivo. One important caveat should be noted: Our Hi-C protocol might disrupt the structure of the 30-nm fiber (e.g., due to use of nonphysiological ion concentrations.)

Measurements of Contact Probability Using Genome-Wide Averages Are Inconsistent with an Ordinary Polymer at Equilibrium. In our original Hi-C study (5), we characterized the behavior of chromatin at the megabase scale by analyzing the contact probability function, $I(s)$, described above, based on Hi-C data, analytical estimates, and in silico studies. The data for human chromatin showed a power-law relationship of the form $I(s) \propto s^{-\gamma}$ between 500 kb and 7 Mb, with $\gamma = 1.08$. We showed that values of γ can be used to discriminate among distinct polymer states. Specifically, $\gamma = 1.08$ is inconsistent with the classic structure of a globular polymer at equilibrium (an "equilibrium globule," which has $\gamma = 1.5$), but is consistent with a dense, scale-invariant, isotropic, long-lived polymer state known as the fractal globule (5). Because the fractal globule's unknotted topology makes it easier to access individual genomic loci, it furnishes an appealing model for chromatin structure.

When we repeated the above analysis on our new, kilobase-resolution maps, we observed a scaling of $\gamma = 1.27$ between 300 kb and 3 Mb (*SI Appendix*, Fig. S2*A*). This slightly higher value is consistent with our previous conclusion that chromatin does not fold into an equilibrium globule, and falls within the range we predicted for a fractal globule (17).

Genome-Wide Measurements of Chromatin Folding Inside Individual Contact Domains Reveal a Polymer State Characterized by $\gamma \approx 0.75$. In our original study, we could not discern local folding features at scales smaller than ∼1 Mb. With our new maps, which contain 200- to 1,000-fold more data, we had the opportunity to study folding within contact domains, which are contiguous genomic intervals in which there is an enhanced probability of contact among all loci (Fig. 2*A*). We found that folding measurements differ sharply within contact domains vs. across contact domains.

We began by calculating $I_{same}(s)$ using our genome-wide averaging technique, but only including pairs of loci that were in the same contact domain. We obtained a markedly lower value: $\gamma = 0.76$ (Fig. 2*B*).

Next, we measured the decline in contact probability with distance relative to a fixed DNA locus. For loci longer than 50 kb, the results were highly reproducible (*SI Appendix*, Fig. S2*G*). We focused on 1,057 distinct 50-kb loci, each situated at the midpoint of a contact domain. The resulting contact probability plots consistently exhibited two distinct regimes. Contact frequency within a domain exhibited a power-law decline whose γ was centered on 0.75 (SD = 0.05) and always smaller than 1 (Fig. 2*B* and *SI Appendix*, Fig. S2*D*). Outside

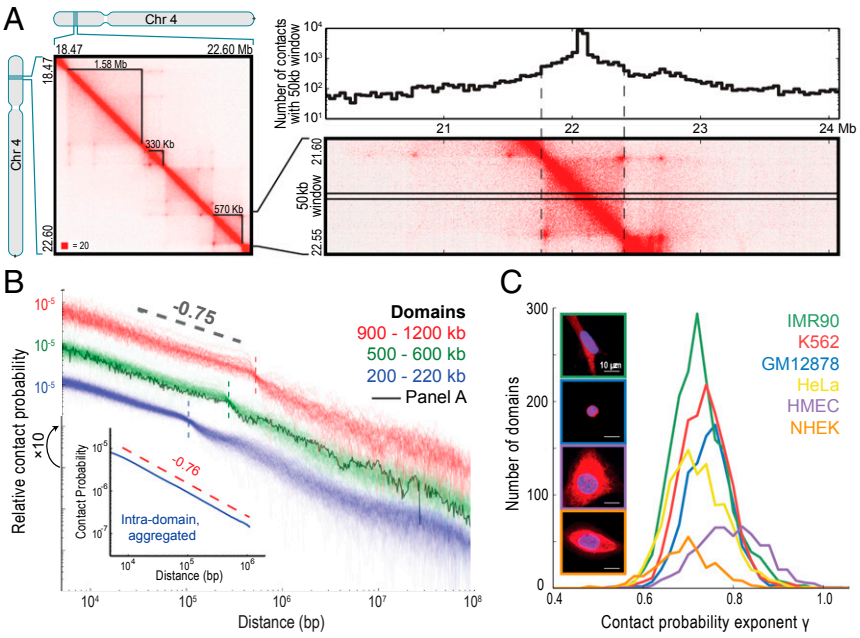


Fig. 2. Contact domains exhibit a contact probability scaling with $\gamma \approx 0.75$. (*A*, *Left*) Contact domains from a region on chromosome 4 of GM12878 lymphoblastoid cells. (*A*, *Right*) Number of contacts (*Top*) incident on a 50-kb window at the center of a domain (*Bottom*). (*B*) Contact probability vs. genomic distance for 473 individual domains, measured with respect to a 50-kb locus at the domain's center. A power law (reference slope of −0.75, gray dashed line) is consistently observed inside domains, whose boundary is indicated by a vertical dashed line. A single black line shows contact probability for the domain from *A*. Domains are grouped by size; each group is vertically shifted by an order of magnitude for visual clarity. (*Inset*) Contact probability vs. genomic distance, excluding pairs of loci that lie in different contact domains. A power-law ($\gamma = 0.76$) is seen over two orders of magnitude. (*C*) Histogram of γ values for contact domains across six human cell types. (*Inset*) Representative microscopy images (maximum Z projections) of four cell types, showing chromatin (blue, DAPI stain) and cytoplasm (red, CellTracker CMTPX dye [Thermo Fisher]). (Scale bar: 10 μm.)

the domain, the power-law regularity was replaced by a more heterogeneous monotonic decline (Fig. 2*B* and *SI Appendix*, Fig. S2*E*).

Our findings suggest that because the frequency of contact between two loci declines markedly when a contact domain boundary is crossed, $I(s)$, which is calculated predominantly using pairs of loci separated by such a boundary, tends to overestimate γ for contact domains (*SI Appendix*, Fig. S2*B*).

To check whether γ depended on nuclear volume, we compared four human cell types using confocal microscopy and in situ Hi-C. Despite threefold variation in nuclear volume (from smallest to largest, GM12878: 237 ± 84 μm^3, IMR90: 381 ± 157 μm^3, NHEK: 440 ± 90 μm^3, HMEC: 728 ± 307 μm^3), the intradomain γ measurements were indistinguishable (Fig. 2*C* and *SI Appendix*, Table S1). The results were also similar in different nuclear compartments (A/B) (5) and subcompartments (A1/A2/B1/B2/B3) (8) (*SI Appendix*, Fig. S2*C*), in mouse (CH12-LX lymphoblasts), and in experiments with and without cross-linking (*SI Appendix*, Table S2).

Because site-directed recombination relies on the spatial proximity of DNA sites, we reexamined published experiments probing the relationship between flippase recombination frequency and genomic distance in humans (18). Interestingly, we noted a power-law scaling with $\gamma = 0.75$ (*SI Appendix*, Fig. S2*F*).

Taken together, these results suggest that chromatin in contact domains is characterized by $\gamma \approx 0.75$. Next, we sought to understand the implications of this exponent.

New Mathematical Theorem Indicates That Chromatin Folding Inside Contact Domains Is Not Strictly Fractal. A difficulty in interpreting γ is the uncertainty about which γ values are consistent with a fractal globule. Approximate methods and physical simulations suggested values ranging from 1 to 1.33 (5, 17). However, no rigorous bounds are known. We therefore sought to derive such bounds.

We proved mathematically that γ must lie between 1 and 2 for any fractal structure. To prove this result, we analyzed mathematical functions (denoted f) that continuously map (in other words, fold) the unit segment [0,1] into a higher dimensional space. We focused on fractal curves, which are generated by applying a simple folding rule to a simple initial state and repeating ad infinitum. When the folding rule is applied identically at all scales, the resulting fractal curves have no characteristic length scale. Because such curves continuously transform a 1D segment into a higher dimensional object, they have been of interest to mathematicians ever since the first space-filling curves, which map the unit segment onto the unit square, were discovered by Giuseppe Peano (the "Peano curve" in 1890) and David Hilbert (the "Hilbert curve" in 1891). If the repetition process is terminated after a finite number of steps, the resulting curve corresponds to a physically realizable polymer chain. For this reason, finite iterations of fractal curves are often used to model the fractal globule (5). By bounding the values of γ that can be obtained from fractal curves, we can test whether our experimental data are consistent with a strict fractal globule.

When characterizing a fractal curve, a commonly used measure is the Minkowski (or "box-counting") dimension, denoted dim(X), which generalizes the common notion of dimension to non-integer values. Just as the number of line segments with width $1/N$ needed to cover the 1D unit segment scales as N^1 and the number of squares with width $1/N$ needed to cover the 2D unit square scales as N^2, dim(X) is defined so that the number of boxes with width $1/N$ needed to cover X scales as $N^{\dim(X)}$. For instance, the Minkowski dimension of a crumpled sheet of paper (≈2.51) provides a measure of its packing density (19). The Minkowski dimension of Great Britain's coastline (≈1.25) is a measure of its roughness (20). The Minkowski dimension can also be less than 1: The set of points in [0,1] without an odd digit in their decimal expansion (i.e., 0.86, 0.22222) has dimension 0.699.

We proved that folding the 1D unit segment [0,1] into a d dimensional fractal curve scales the Minkowski dimension of all subsets of [0,1] by a factor of d; that is, any k-dimensional subset of the unit segment will fold into a $k * d$ dimensional shape. Our results can be summarized in the following theorem and corollary, whose proofs appear in *SI Appendix*:

> Theorem: For any self-similar fractal curve f([0,1]), $\dim f(X) = d \cdot \dim X$ for any $X \subseteq [0,1]$.*
>
> Corollary: The contact probability of a fractal curve satisfies $I(s) \propto s^{-\gamma}$ with $\gamma = 2 - (d_{surf}/d)$, where s is the linear distance along the curve, d_{surf} is the Minkowski dimension of the curve's surface (i.e., its surface roughness), and d is the Minkowski dimension of the curve as a whole.

An illustration of the theorem is the 2D Dragon curve, which doubles the Minkowski dimension of all subsets in its domain (Fig. 3*A*). Mathematically, our result is a deterministic analog of Henry

*The proof is in two parts. First, we show that any fractal curve f is a *1/d-Hölder* function, which gives an upper bound on dim X. Next, we construct a push-forward measure on $f(X)$, which gives a lower bound on dim X. Both bounds are the same, thus giving the exact value. Full proofs are provided in *SI Appendix*.

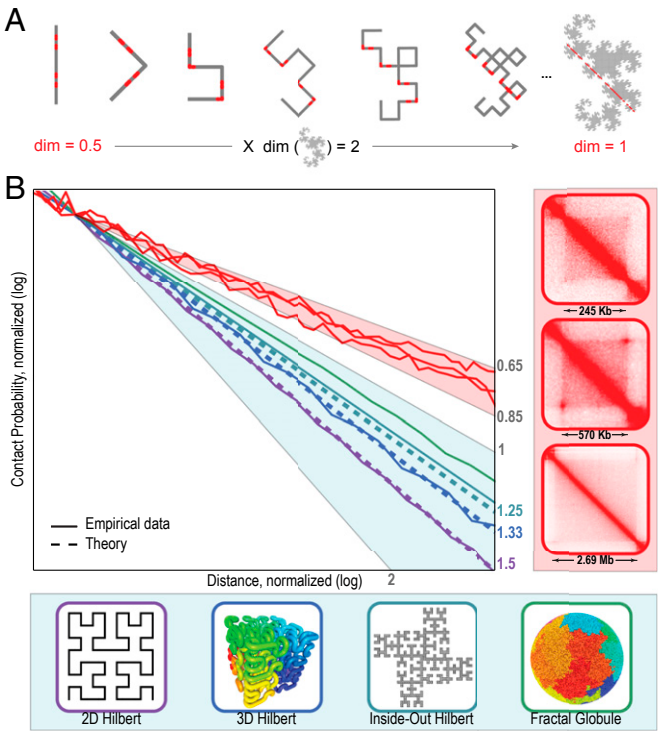


Fig. 3. New mathematical theorem demonstrates that chromatin folding inside contact domains is not strictly fractal. (*A*) Successively applying a simple folding rule transforms a 1D line segment (*Left*) into a 2D Dragon curve (*Right*). Our theorem reveals that just as the curve doubles the dimension of the line segment, it doubles the dimension of all subsets of the line segment. Thus, when we intersect the Dragon curve with a line to create a 1D feature, the corresponding points in the original segment must have a Minkowski dimension of 1/2. A corollary of this theorem makes it possible to calculate the contact probability scaling exponent γ for any fractal curve. (*B*) Contact probability vs. distance for various structures. Our corollary predicts that γ for fractal curves satisfies $\gamma = 2 - (d_{surf}/d)$, where d_{surf} is the dimension of the curve's surface and d is the dimension of its interior. (*Left*) Comparing simulations (solid lines) with theoretical predictions (dashed lines). Two-dimensional Hilbert curve (purple; $d_{surf} = 1$, $d = 2$, $\gamma = 1.5$), 3D Hilbert curve (blue; $d_{surf} = 2$, $d = 3$, $\gamma = 1.33$), inside-out Hilbert curve, rank 3 (teal; $d_{surf} = 1.5$, $d = 2$, $\gamma = 1.25$), and fractal globule (green). As the rank of an inside-out Hilbert curve increases, its boundary becomes nearly 2D and its γ draws asymptotically close to 1. (*Right, Top* to *Bottom*) Three contact domains: chromosome 12: 46.2–46.4 Mb, chromosome 4: 21.8–22.4 Mb, and chromosome 5: 2.1–4.8 Mb.

McKean's well-known "dimension-doubling" theorem, which states that Brownian motion doubles the dimension of subsets (13).

The corollary may be illustrated by measuring γ for classic fractal curves, such as the 2D Hilbert curve ($d_{surf} = 1$, $d = 2$, $\gamma = 3/2$; Fig. 3*B*, purple), the 3D Hilbert curve ($d_{surf} = 2$, $d = 3$, $\gamma = 4/3$; Fig. 3*B*, blue), and many others (*SI Appendix*, Fig. S3 and Table S3). The corollary also implies that for curves with extremely rough surfaces (d_{surf} close to d), γ can draw arbitrarily close to unity. Because no such curves are known, we generalized the Hilbert curve, constructing a class of "inside-out" Hilbert curves (Fig. 3*B*, teal) whose boundaries are arbitrarily rough and whose γ values come arbitrarily close to 1 (*SI Appendix*, Figs. S6 and S7).

Because $0 \leq d_{surf}/d < 1$, the corollary proves that γ for a fractal curve must lie between 1 and 2. Thus, our measurements of $\gamma \approx 0.75$ inside contact domains (Fig. 3*B*, red) are inconsistent with the hypothesis that contact domains tend to form fractal globules.

Physical Simulations Suggest That $\gamma = 0.75$ Is Consistent with an Unknotted, Nonequilibrium State That Is Anisotropic Rather than Fractal. Another way of exploring the significance of a particular value of γ is by computationally modeling chromatin as a polymer comprising many identical monomers, each of which represents a fixed number of bases. By simulating the dynamics of a condensing polymer chain and the surrounding mixture under various physical assumptions, it is possible to test whether a particular set of conditions leads to a realistic γ value.

In our original models, we simulated a condensation process in which the collapse of the polymer was driven by external forces (i.e., the crowding of a stretch of chromatin by other components of the nucleoplasm). Through an excluded volume interaction, these components crush the polymer chain. Such forces can be modeled using a potential function that attracts all monomers equally toward a single point (5). Because this potential has no characteristic length scale, the resulting dynamics are scale-invariant, and the polymers collapse isotropically into a fractal globule.

Notably, our earlier models did not examine the effects of interactions among the monomers on polymer condensation. Attractive forces between individual nucleosomes have been observed in vitro by many groups (2), and effective attractions between monomers are seen in all polymer globules, arising when a polymer is immersed in a poor solvent (21). Therefore, in the present study, we incorporated attractive forces between monomers using the classic Lennard–Jones potential. This model of intermolecular attractions was originally developed to study van der Waals effects and is commonly used to describe the attractive forces between nucleosomes (15, 22). We examined a class of systems in which both intermonomer attractions and external crushing forces are present during condensation. The ratio of these forces is given by a single parameter, R, which we varied across eight orders of magnitude.

We probed these systems using Langevin dynamics simulations, in which random collisions between the solvent and the polymer are accounted for implicitly through parameters for viscosity and temperature. We ran our simulations using LAMMPS (23), accelerated using graphical processing units (24). Each monomer represented 1 kb (matching the above estimates of Kuhn length). The polymer chain contained up to 10,000 monomers, or 10 Mb. For each condition, we ran at least 100 simulations from randomized starting configurations.

Our simulations revealed a family of nonequilibrium states (Fig. 4*A*). When internal forces are weak ($R << 1$), polymer condensation closely resembles the isotropic dynamics observed with pure external forces ($R = 0$) and results in a fractal globule. Because intermonomeric attractions decay as monomers move apart, internal forces introduce a length scale into the system. When they are sufficiently strong ($R >> 1$), condensation transitions into an anisotropic regime. First, tiny globules form along an extended chain; tension along the chain then causes the globules to concatenate in a linear fashion (Fig. 4*B*). This model of polymer condensation was first postulated by de Gennes (21). The resulting state, which we dub a "tension globule," is not scale-invariant. Instead, it contains long intervals in which genomic position is correlated with spatial position along a linear axis (*SI Appendix*, Fig. S8 *D* and *E*).

Importantly, the values of γ differ depending on the regime. When R is small, $\gamma(R)$ is slightly larger than unity, consistent with our earlier fractal globule simulations. When R is large, $\gamma(R)$ is roughly 0.72, consistent with our observations for contact domains (Fig. 4*C* and *SI Appendix*, Fig. S8*A*). In-between, $\gamma(R)$ exhibits intermediate values. Interestingly, all of the states in this family are dense and largely unknotted (*SI Appendix*, Fig. S8*F*).

Our findings were robust to variations in numerous simulation parameters: chain length and initial configuration, solvent temperature, viscosity, and simulation time (*SI Appendix*, Fig. S9 and Table S4). They were also robust to the mechanism underlying the internal forces. They did not change significantly when we replaced the Lennard–Jones potential with a Yukawa potential, a model of screened electrostatic forces, in which the attractions decay much more rapidly with distance (*SI Appendix*, Fig. S10 and Table S4).

Our simulations again suggest that the structure of nuclear chromatin inside contact domains is not consistent with a fractal globule. However, the γ value is consistent with a tension globule resulting from condensation amid strong internal attractions between monomers.

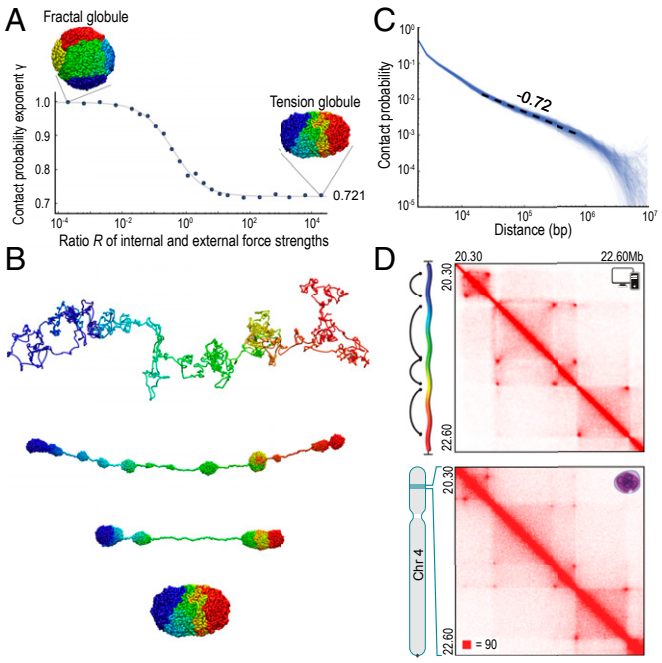


Fig. 4. Value of γ obtained using Hi-C is consistent with a tension globule in which loops form by diffusion. (*A*) Value of γ for a polymer after condensation varies as the ratio of internal to external forces, *R*, is changed. (*B*) Condensation of a 10-Mb tension globule over time. (*C*) Contact probability vs. distance for 450 simulated tension globules with a length of 10 Mb. In each case, a power law is seen for distances between 20 kb and 800 kb. Mean γ = 0.73, SD = 0.07. (*D*) Simulation of a region of chromosome 4 in GM12878 (chromosome 4: 20.3–22.6 Mb). The experimental data exhibit four loop domains (*Left*) and can be recapitulated (*Right*) using a tension globule containing four loops (black arcs) which is tethered at both ends (*SI Appendix*). Contact domains form spontaneously.

Condensation of a Tension Globule Results in Spontaneous Formation of Contact Domains Between the Anchors of a Loop. One of the most surprising features of our in situ Hi-C maps is that contact domains often correspond to loops, that is, the domain boundaries lie at the loop's two anchor loci.

We used our physical simulations to explore the effects of bringing together two anchor points followed by condensation into a tension globule. Notably, the formation of a loop led to enhanced contact frequency between all pairs of loci in the interval demarcated by the two loop anchors (i.e., a contact domain) (Fig. 4*D*). These contact domains exhibited values of γ (0.77 ± 0.08) that match our experimental observations (0.75 ± 0.05; *SI Appendix*, Fig. S11*B*).

Simulation Results for a Tension Globule Match Intradomain Distances Measured by 3D-FISH. We next examined whether the tension globule model recapitulates the spatial distances observed experimentally. We studied four pairs of loci using 3D-FISH (8). Each pair lay in a single domain; the genomic distance between the loci ranged from 320 kb to nearly 1 Mb. We measured at least 50 3D distances for each locus pair. We compared these values with distributions obtained using our tension globule simulations (*SI Appendix*, Fig. S13*B*). The simulated distributions matched the experimental distributions as closely [Kolmogorov–Smirnov (K-S) statistic = 0.15] as experimental replicates matched one another (K-S statistic = 0.18).

Taken together, our findings show that the tension globule model accounts for three important features of genome folding observed in our 3D maps: the value of the contact probability scaling γ, the formation of contact domains between loop anchors, and the distribution of 3D distances between pairs of loci.

Tension Globule Model Fails to Explain Other Key Features of 3D Folding. Although the tension globule model can explain the general scaling properties of our data, it does not explain many other facets, particularly those facets related to loop formation. In a tension globule model, loop formation would occur via diffusion. Looping proteins (e.g., CTCF) would initially bind to DNA anchor motifs. When diffusion brings two anchors into close spatial proximity, the proteins would dimerize, forming a chromatin loop.

This diffusive process inevitably leads to loops that overlap (i.e., a point in the interior of one loop is anchored to a point outside the loop), thus creating chromatin entanglements. By contrast, our experimental data show that pairs of loops rarely overlap. It is also hard to understand, in a diffusive model, why the CTCF motifs at pairs of loop anchors must lie in the convergent orientation. We therefore considered alternative models.

Loop Formation by Extrusion Complexes Would Explain Key Features of 3D Folding. Nasmyth proposed a model based on an "extrusion complex" containing two DNA-binding subunits that are physically tethered together (25, 26). This complex is loaded onto chromatin at a single locus. Initially, its subunits are bound to nearby DNA elements, forming a tiny chromatin loop. Next, DNA is extruded through the subunits such that the two tethered subunits move in opposite directions with respect to the genome: one forward and one reverse. As a result, the loop continues to grow, without knotting. Eventually, the extrusion complex dissociates from DNA (Fig. 5*A*, *i*–*iii*).

We explored the behavior of extrusion complexes in our simulations as follows. As before, we began with a condensing polymer. Extrusion complexes are bound to the polymer at a density that depends on their concentration and dissociate at a rate that depends on their processivity. The complexes cannot pass through one another; if they collide, one must fall off.

We then added one novel feature to Nasmyth's model (19), based on our observations about the role of CTCF motifs. We assume that each subunit of the extrusion complex recognizes the presence of a particular motif on a particular DNA strand, such as an appropriately oriented CTCF motif, by binding tightly and halting the extrusion process through the subunit. We therefore designated certain monomers in our simulated polymer as anchors, assigning each a forward or reverse orientation. In our simulations, the forward subunit's progress may be halted by a forward anchor, but not by a reverse anchor; the reverse subunit may be halted by a reverse anchor, but not by a forward anchor (Fig. 5*A*, *iv*).

We began by simulating a condensing polymer containing a pair of convergent anchors 1 Mb apart. When an extrusion complex landed between the anchors, it began extruding a loop until its subunits eventually arrived at the anchor monomers. At this point, the extrusion came to a halt, yielding a "persistent loop" between the anchors (i.e., a loop that was present for a protracted period) (Fig. 5*B* and *SI Appendix*, Fig. S12*C*). This extrusion process did not prevent condensation and resulted in a globular state.

When we examined the resulting contact maps, we made three observations. First, a prominent peak was present between the two anchors, reflecting the formation of a persistent loop. Second, there was enhanced contact frequency between all pairs of loci lying between the two anchors, forming a contact domain. Third, the contact domains exhibited power-law scalings in contact probability with values of γ (0.72 ± 0.06) that match our experimental observations (0.75 ± 0.05; Fig. 5*C*). These findings reflect the equilibrium for a condensing polymer immersed in a solvent containing extrusion complexes and did not depend on the condensation forces (external or internal) or the initial condition (*SI Appendix*, Fig. S12 *A* and *B*).

Next, we examined whether the extrusion model recapitulates pairwise spatial distances observed using 3D-FISH. Repeating our previous analysis (*SI Appendix*, Fig. S13*C*), we found that the simulated distributions matched the experimental distributions almost as closely (K-S statistic = 0.16) as experimental replicates matched one another (K-S statistic = 0.18).

Thus, the extrusion model, like the tension globule model, is consistent with the contact probability scaling γ, the formation of

contact domains between loop anchors, and the distribution of 3D distances between pairs of loci. The extrusion model explains a much wider range of observations, however. In particular, it provides a natural explanation for why loops overwhelmingly form between convergent CTCF sites. Because loops that occur in the same cell cannot overlap under the extrusion model, the model also explains why we rarely see overlapping loops in our experimental data.

Extrusion Model Can Recapitulate Hi-C Experimental Results, Given the Locations of CTCF Binding. Next, we explored whether the extrusion model could recapitulate Hi-C experimental results in silico using CTCF ChIP-sequencing (ChIP-Seq) data alone.

We began by simulating the folding of a 2.3-Mb target region on chromosome 4 (20.3–22.6 Mb). Our algorithm created an in silico representation of the region as a uniform polymer, adding forward and reverse anchors at peaks of CTCF binding observed in experimental ChIP-Seq data for the region. The strength of each anchor (the likelihood that the corresponding subunit would halt) reflected the amplitude of the ChIP-Seq peak. Anchor orientation was assigned based on the strand of the CTCF motif associated with the peak. The inputs did not include Hi-C data.

We simulated the behavior of the model polymer in a solvent containing extrusion complexes. Strikingly, the resulting contact matrix closely resembled the contact matrix obtained using Hi-C experiments (Fig. 5*D*; Pearson's $r = 0.964$). Peaks and contact domains were observed in the same positions, and appropriate γ values were obtained inside each contact domain. The results were similar for other regions (*SI Appendix*, Fig. S12*D*).

We also sought to simulate the target region using the tension globule model. As before, we identified peaks in CTCF ChIP-Seq data, and assigned each peak an orientation based on the strand of the CTCF motif associated with the peak. However, to obtain contact matrices that resembled our experimental results, we had to impose a number of ad hoc penalties that do not correspond to any natural processes in 3D diffusion: Loops were only allowed between pairs of convergent peaks, and the likelihood of such a loop depended on (*i*) the strength of the peaks, (*ii*) the distance between the peaks, and (*iii*) the number and strength of intervening CTCF peaks. Even so, the simulations did not provide as good a fit as the extrusion model (*SI Appendix*, Fig. S11*E*; Pearson's $r = 0.922$).

Network of Loops Contains Hundreds of Isolated Cliques and Is Consistent with a Model in Which Consecutive Loops Can Form Simultaneously by Extrusion. We then explored higher order relationships among the location of loops by constructing a "loop network" for GM12878 lymphoblastoid cells (8). The network's nodes are genomic loci containing at least one loop anchor. Edges indicate loops.

We were particularly interested in "isolated cliques." This network motif comprises a set of $n \geq 3$ loci such that any pair in the set is connected by a loop (i.e., the set is a "clique") but none of the loci are connected by loops with loci outside the set (i.e., the set is "isolated") (Fig. 6).

Isolated cliques were dramatically enriched in the loop network. For instance, in GM12878, we observed 69 isolated cliques with three nodes (sixfold enrichment relative to an ensemble of randomized control networks), 16 with four nodes (28-fold), and 1 with five nodes (161-fold). If we allowed for a small number of loops ($\leq N - 2$) between loci inside the clique and loci outside, these numbers increase: 145 cliques with three nodes (threefold enrichment), 86 with four nodes (12-fold), 5 with five nodes (14-fold), and 1 with six nodes (41-fold).

Typically, the clique loci were positioned one after another in the genome, with no other loop anchors intervening (in 64% of cases). The first clique locus (in genomic coordinates) typically contained a bound CTCF motif in the forward orientation (97%, fourfold enrichment). The last clique locus contained a bound CTCF motif in the reverse orientation (98%, fourfold enrichment). The middle clique loci typically contained CTCF motifs in both orientations (70%, 8.5-fold enrichment). These findings are consistent with the convergent rule for CTCF looping.

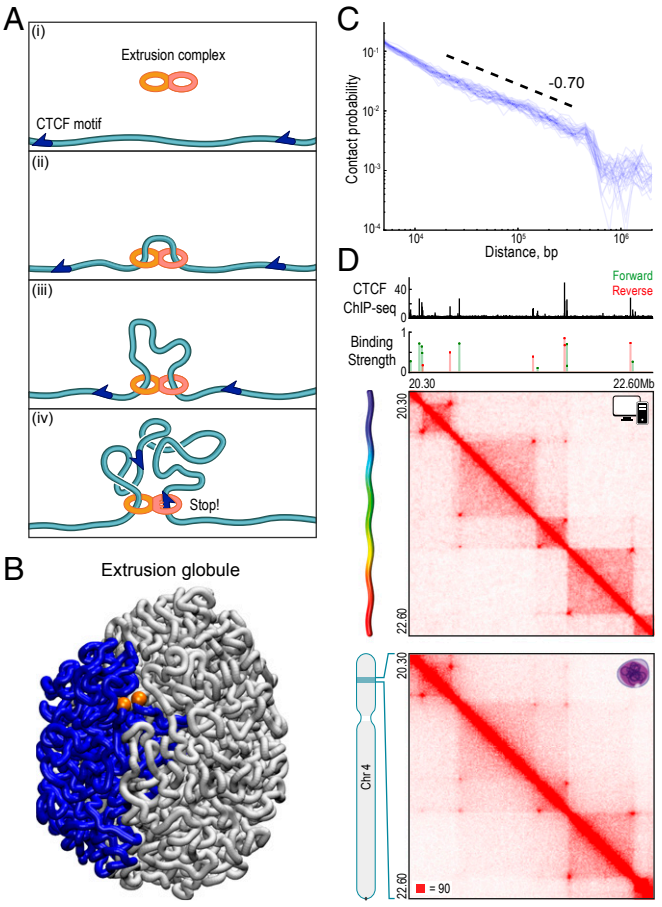


Fig. 5. Model based on loop extrusion makes it possible to recapitulate Hi-C maps accurately using only CTCF ChIP-Seq results. (*A*, *i* and *ii*) Extrusion complex loads onto the fiber at a random locus, forming an extremely short-range loop. (*A*, *iii*) As the two subunits move in opposite directions along the fiber, the loop grows and the extruded fiber forms a domain. (*A*, *iv*) When a subunit detects a motif on the appropriate strand, it can stop sliding. Unlike diffusion, extrusion cannot mediate co-location of motifs on different chromosomes. (*B*) Three-dimensional rendering of a 3-Mb extrusion globule from the ensemble described below. Convergent CTCF anchors (orange spheres) lead to an unknotted loop spanning a compact, spatially segregated contact domain (highlighted in blue). (*C*) Contact probability vs. distance for 12 domains with a length of 1 Mb, created in silico using loop extrusion, measured from a 100-kb locus at the center of the domain. In each case, a power law is seen for distances between 5 kb and 400 kb. Mean $\gamma = 0.72$, SD = 0.06. (*D*) We use loop extrusion to model a 2.3-Mb region on chromosome 4 of GM12878. CTCF ChIP-Seq signals are normalized and converted into binding probabilities for the simulated extrusion complex. Each peak is assigned a forward (green) or reverse (red) orientation based on the strand of the underlying CTCF motif. Extrusion simulations yield an ensemble of 3D polymer configurations; contact maps for the simulated ensemble (*Top*) recapitulate the features observed in our kilobase-resolution Hi-C experiments (*Bottom*), including the position of domains and loops.

Isolated cliques may reflect simultaneous colocation of all clique loci at a single spatial hub within the same cells, forming a chromatin rosette (9, 27). Alternatively, it may be that some, but not all, clique loops are simultaneous. However, another possibility is that all of the loops in our isolated cliques tend to occur in distinct cells. We cannot distinguish between these alternatives using pairwise contact maps from cell ensembles because such maps do not reflect higher order colocation patterns.

However, we noticed that when a middle locus contained a unique motif in each orientation, the motifs were almost always arranged in a divergent configuration: The first motif was in the reverse orientation, pointing toward the preceding clique locus,

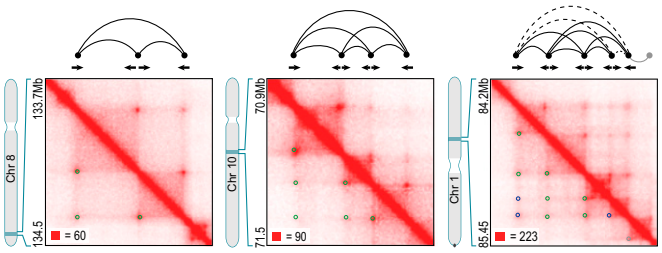


Fig. 6. Analysis of loop networks reveals many isolated cliques, consistent with a model in which consecutive loops can form simultaneously by extrusion. Cliques of size three (*Left*), four (*Middle*), and five (*Right*), shown as network representations (*Above*) and in the Hi-C contact map (*Below*). Nodes correspond to loop anchor loci; edges and open circles indicate a loop called in (8) (solid green), or using a more relaxed threshold (dashed blue). The five-clique exhibits an additional loop (gray) connecting a clique locus to a locus outside the clique. The loop anchor CTCF motifs are indicated; each middle clique locus contains a pair of CTCF motifs in the divergent orientation.

and the second motif was in the forward orientation, pointing toward the subsequent clique locus (89% vs. 11% in the convergent orientation). Because the extrusion model predicts that the genomic intervals inside simultaneous chromatin loops cannot overlap, the overwhelming bias toward the divergent configuration would be expected only if consecutive loops in our cliques (e.g., loops A-B and B-C in a clique comprising A, B, and C) form simultaneously by extrusion.

Genome Editing of CTCF Motifs Disrupts Corresponding Loops, Consistent with the Convergent Rule. We next sought to study the formation of loops experimentally. We used CRISPR/Cas9-based genome editing (14) to modify specific CTCF motifs and explore the effects on loop structure. To avoid allelic heterogeneity, we used HAP1, a human haploid cell line.

We generated an in situ Hi-C map of WT HAP1 cells, with 1.1B reads (*SI Appendix*, Table S6), in which we annotated 8,334 loops and 4,332 contact domains. We chose to study a target region containing three loci: A (chromosome 8: 133.9 Mb), B (134.2 Mb), and C (134.5 Mb). Each pair of these three loci forms loops with one another (A-B, B-C, and A-C). CTCF sites are present at each locus in accordance with the convergent rule: Locus A has a forward-oriented CTCF motif (dubbed A/Forward), locus B has a reverse-oriented CTCF motif (B/Reverse) followed by a forward-oriented motif (B/Forward), and locus C has a reverse-oriented motif (C/Reverse). All three loops are associated with contact domains.

If our loop anchor motif annotation is accurate, deleting the A/Forward site would disrupt the A-B and A–C loops. To test this hypothesis, we performed genome editing by using two guide RNAs designed to flank either side of the A/Forward motif (*SI Appendix*, Table S7). We grew a clonal population of cells carrying a 17-bp deletion spanning the A/Forward motif.

To study the effects of this and other genome editing experiments, we developed an inexpensive way to monitor Hi-C results only in the target region by performing HYbrid Capture on the in situ Hi-C library (28). We validated this method, dubbed "Hi-C^2," by applying it to WT HAP1, capturing a 2-Mb region (chromosome 8: 133–135 Mb). The results were equivalent to the results obtained using ordinary in situ Hi-C (*SI Appendix*, Fig. S14).

When we performed Hi-C^2 on our A/Forward mutant cell line, we observed that, as predicted, the A-B and A-C loops were disrupted (Fig. 7*A*).

Next, we examined the effect of disrupting the motifs at the B locus. We tested five predictions of the convergent CTCF rule: (*i*) Disruption of B/Forward (a 159-bp deletion) will eliminate the B-C loop alone (and not affect the loop between A and B, which, by the convergent rule, must be anchored at B/Reverse); (*ii*) disruption of B/Reverse (142 bp) will eliminate the A-B loop; (*iii*) inversion of B/Forward will eliminate the B-C loop; (*iv*) if B/Reverse is disrupted, inverting B/Forward will eliminate the B-C loop, but restore the A-B loop at a slightly offset anchor position (corresponding to the position of the inverted B/Forward rather than B/Reverse); and (*v*) inversion of both sites will not disrupt either loop, but will offset the positions of both anchors. The experimental data matched these predictions in every case (Fig. 7*A* and *SI Appendix*, Fig. S16).

We next examined a similar region on chromosome 1, spanning D (180.5 Mb), E (180.8 Mb), and F (181.1 Mb). We tested four predictions of the convergent rule: (*i*) Disruption of E/Reverse (a 16-bp deletion) will eliminate the D-E loop; (*ii*) disruption of E/Forward (10 bp) will eliminate the E-F loop; (*iii*) inversion of E/Forward will eliminate the E-F loop; and (*iv*) simultaneously disrupting E/Forward (7 bp) and E/Reverse (16 bp) will eliminate the D-E and E-F loops, leaving only the D-F loop. The experimental data matched these predictions in every case (Fig. 7*B*).

Finally, we targeted a similar region on chromosome 5, spanning G (31.6 Mb), H (31.8 Mb), and I (32.1 Mb). We inserted a single base pair into the G/Forward motif (chromosome 5: 31,581,788). Strikingly, both the G-H loop and the G-I loop disappeared, but the H-I loop remained (Fig. 7*C*).

The convergent rule, combined with our loop anchor motif annotation, correctly predicted the looping pattern in all 13 mutants (Fig. 7 and *SI Appendix*, Fig. S15). When we examined the effect of removing the corresponding CTCF motifs in our extrusion simulations, the contact maps predicted by the simulations closely matched those contact maps obtained experimentally (Fig. 7). These results show that it is possible to reengineer loops in a targeted fashion.

Our experiments also shed light on which loops in a clique occur in the same cells. For example, if all clique loops occurred simultaneously, then A-B would still be in close proximity even after deleting B/Reverse, because the A-C and B-C loops would sequester A and B near C/Reverse. The fact that we can disrupt the A-B loop without affecting the A-C and B-C loops suggests that the A-C and B-C loops do not occur in the same cells. Similarly, our ability to disrupt the B-C loop alone suggests that the A-C and A-B loops do not occur in the same cells. Taken together, our data are consistent with a model in which consecutive loops (i.e., A-B and B-C) tend to occur simultaneously in some cells, whereas the larger loop (A-C) tends to form in other cells (Fig. 7*D*).

Contact Domains Can Form Between Consecutive Loop Anchors That Do Not Loop to One Another. Interestingly, disrupting certain anchor motifs eliminated a loop, but not the associated contact domain. For instance, deletion of the B/Forward motif disrupted the B-C loop, but not the B-C contact domain. Similar behavior was observed in all nine experiments involving the removal of a single anchor motif at a middle locus.

We found that our computational simulations produced similar results for the extrusion model (but not for the tension globule model). The reason was because even after we eliminate B/Forward, an extrusion complex landing in the B-C interval tends to remain within the B-C interval. At one end, the complex tends to halt at the C/Reverse motif; at the other end, its progress tends to be impeded by the forward subunit of the extrusion complexes positioned at B/Reverse (whenever this site is engaged in looping with A/Forward) (Fig. 7*E*). Because the extrusion complex is excluded from adjacent intervals, it tends to bring points within the B-C interval together, forming a contact domain. We dubbed this configuration an "exclusion domain."

The extrusion model predicts that an exclusion domain will remain when one of the middle anchors is deleted, but not if both are deleted, as in the case of the mutant lacking both E/Forward and E/Reverse. In this case, an exclusion domain will not form between D and E because there is no E-F loop to interfere with the sliding of extrusion complexes landing in the D-E interval; similarly, an E-F exclusion domain will not form because there is no D-E loop. These predictions are precisely what is seen in our Hi-C^2 experiments.

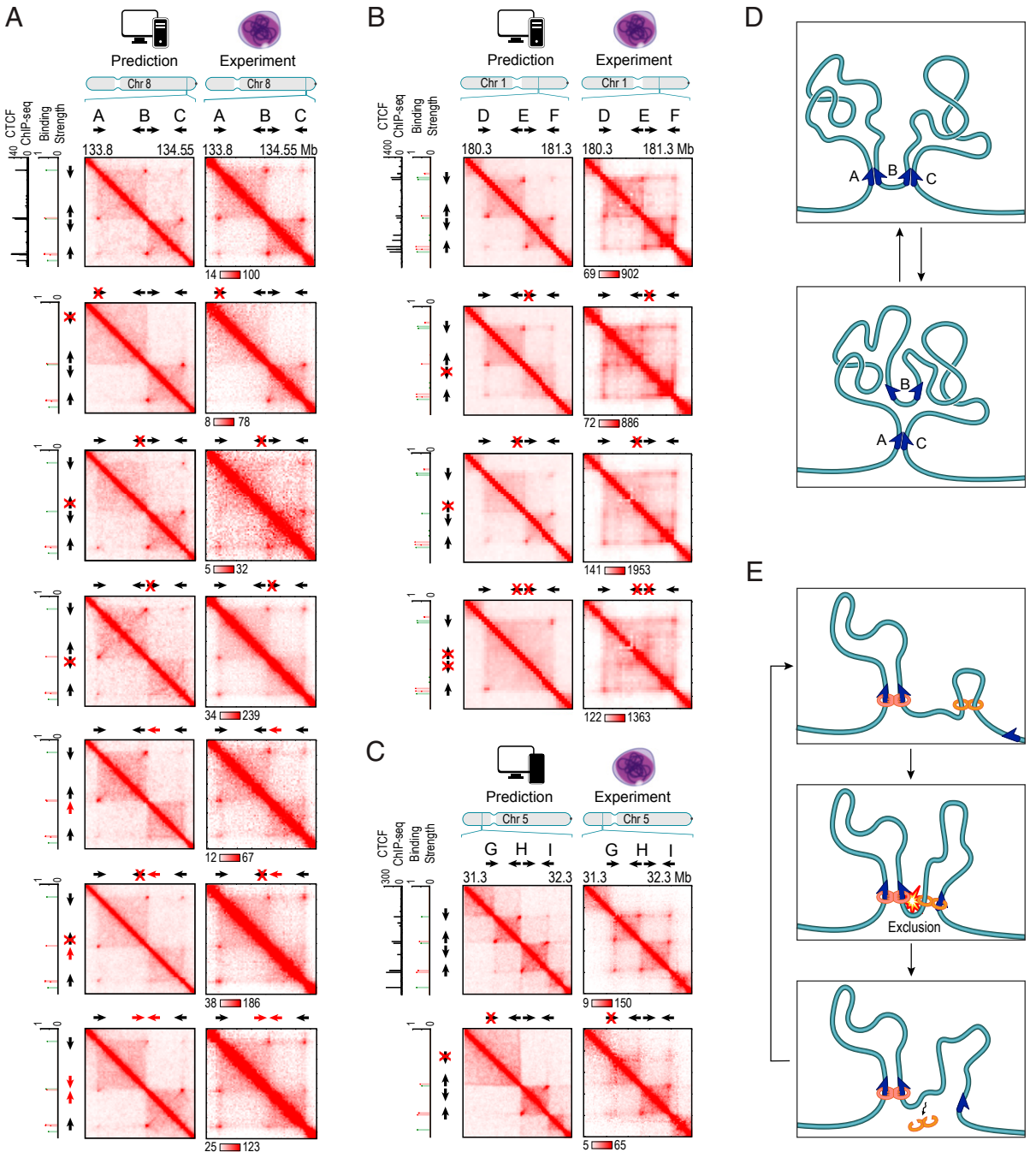


Fig. 7. Genome editing of CTCF motifs allows reengineering of loops in accordance with the convergent rule; the resulting contact maps can be predicted in silico using extrusion simulations. (*A*) Results of CRISPR/Cas9-based genome editing experiments at chr8:133.8–134.55 Mb in HAP1 cells. Extrusion simulations (*Left*) and experimental data (*Right*) are shown. (*A*, first row) Contact map for the WT locus, calculated using in silico simulations (*Left*), closely matches the map observed using Hi-C^2 experiments (*Right*). (*A*, second row) Deletion of A/Forward eliminates the A-B and A-C loops and the contact domain boundary at locus A. The predictions of our in silico simulations (*Left*) closely match the contact map observed using Hi-C^2 experiments (*Right*). All parameters in this and subsequent simulations of mutant regions use exactly the same parameters as the simulations of the corresponding WT contact map. The only difference in the mutant simulation is the modification of the appropriate CTCF-binding site (in this case, deletion of A/Forward). (*A*, third row) Deletion of B/Reverse eliminates the A-B loop. (*A*, fourth row) Deletion of B/Forward eliminates the B-C loop. (*A*, fifth row) Inversion of B/Forward eliminates the B-C loop. (*A*, sixth row) Simultaneous deletion of B/Reverse and inversion of B/Forward eliminates the B-C loop. (*B*) Similar series of results for chromosome 1 (180.3–181.3 Mb). Notably, the elimination of one loop anchor motif at the middle locus fails to eliminate either the D-E or E-F contact domain. When both loop anchor motifs are eliminated, both the D-E and E-F contact domains disappear. (*C*) We disrupted a forward CTCF motif by inserting a single base at chromosome 5: 31,581,788. Two loops are disrupted. The domain boundary moves to a nearby, weak CTCF site. Because the binding at this new site was weaker than the threshold value, this new boundary was not predicted by our extrusion simulations. (*D*) Our data suggest that the region shown in *A* is typically found in one of two states in wild-type cells. In the first state, both the A-B and B-C loop domains are present, but the A-C loop domain is absent. In the second, only the A-C loop domain is present. The data suggests a similar decomposition for the region in *B*. (*E*) Extrusion can explain the formation of exclusion domains. In this example, an extrusion complex forms a loop between adjacent motifs in the convergent orientation. Downstream, a second CTCF motif in the reverse orientation is unoccupied. Obstructed on both sides, extrusion complexes landing in the interval between the two reverse motifs tend to remain inside the interval. This leads to the formation of a domain.

The extrusion model also predicts that exclusion domains will occur in the WT genome. To identify such domains, we examined regions containing three consecutive loci, A, B, and C, such that A loops to B and to C, but B and C do not loop to one another. We found 986 such cases. An A-B contact domain was annotated in 399 cases (41%), consistent with the frequency of contact domains at loops genome-wide (38%), given the stringent threshold applied for calling domains (8). Despite the lack of an apparent loop between B and C, a B-C contact domain was seen in 158 cases (16%), a 6.3-fold enrichment (*SI Appendix*, Fig. S17). The reduced frequency is likely due to the fact that exclusion domains tend to be less pronounced.

Discussion

Our results illuminate the structure of chromatin at multiple scales: chromatin fibers, contact domains, and loops.

At the smallest scale, the winding of DNA around histones has long been known to form the flexible 10-nm fiber. This fiber is widely believed to coil into the larger, stiffer, 30-nm fiber, although recent studies using microscopy, electron spectroscopy, and X-ray scattering have failed to find evidence for 30-nm fibers in vivo (29, 30). Our Hi-C data allow us to measure the Kuhn length, or bendability, of chromatin fibers. We find that chromatin fibers are highly bendable, with a Kuhn length of roughly 1 kb. This value is far smaller than what would be expected for a 30-nm fiber (30–60 kb) (15), suggesting that 30-nm fibers, if they exist, are rare in intact chromatin. Our findings suggest that at the scale of the typical gene (~15 kb), chromatin is highly flexible. This flexibility is also compatible with (and essential for) loop formation via extrusion.

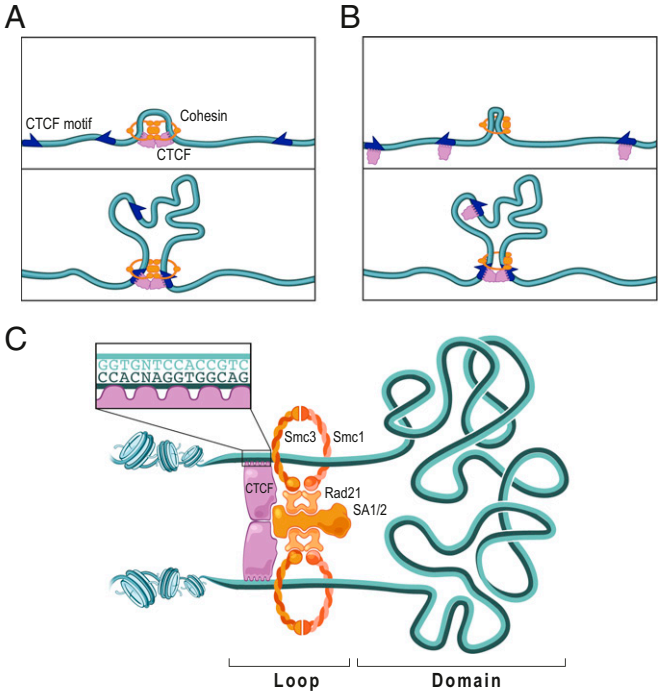


Fig. 8. We hypothesize that loops are formed during interphase by an extrusion mechanism comprising CTCF and cohesin. Here, we illustrate possible models for the extrusion complex. (*A*) In one model, the complex includes two DNA-binding subunits, each comprising a cohesin ring and a CTCF protein. When the complex is loaded onto DNA, a tiny loop forms. The two subunits engage the chromatin fiber in an antisymmetrical fashion, with their CTCF proteins facing the outside of the loop, scanning opposite DNA strands. The loop expands without knotting as the subunits slide in opposite directions. The interior of the loop forms a contact domain. When the CTCF proteins find a target motif on the appropriate strand, they can bind, arresting the progress of the subunit. Eventually, the extrusion complex dissociates. (*B*) In a second model, the sliding of cohesin alone leads to extrusion. Independently, CTCF proteins bind to their motif in an oriented fashion. When the cohesin ring encounters a CTCF protein, the extrusion process either continues or halts, depending on the orientation of CTCF. (*C*) Detailed view of the model in *A*. Other models are possible. Notably, it is unclear how many CTCF proteins and cohesin rings participate in a single extrusion complex, or whether the complex is part of a larger structure. All extrusion models predict that focal chromatin interactions mediated by CTCF must be intrachromosomal.

In our original Hi-C study (5), we probed the physical structure of chromatin at the megabase scale by calculating the relationship between the 1D distance separating two loci, s, and the probability of physical contact between them, $I(s)$. Because the size of our dataset was limited, our calculation was a genome-wide average. For values of s between 500 kb and 7 Mb, we found power-law behavior: specifically, $I(s) \propto s^{-\gamma}$ with $\gamma = 1.08$. This value of γ is inconsistent with an ordinary condensed polymer at equilibrium (for which $\gamma = 1.5$) but is consistent with a fractal globule.

In our recent Hi-C experiments at kilobase resolution (8), we observed a large number of contact domains (median length = 185 kb) that partition the genome. In the present study, we explore the structure of chromatin inside these domains by exploiting the vastly higher resolution of our new maps to calculate $I(s)$, in a locus-specific fashion, genome-wide. The contact probability exhibits a power-law behavior at fine scale, but with a different exponent, $\gamma = 0.75$, than the exponent observed from our low-resolution, genome-wide average.

This value is inconsistent with an ordinary polymer at equilibrium. To determine rigorously whether such a value could be consistent with a fractal globule architecture for individual domains, we proved a novel mathematical theorem describing how the Minkowski (fractal) dimension of a set changes when the set is mapped using a fractal curve. Our result is analogous to a well-known theorem of McKean for Brownian motion. As a corollary, we find that values of γ inside a fractal globule must lie between 1 and 2, implying that $\gamma = 0.75$ is also inconsistent with a fractal globule. We illustrate our corollary by constructing a novel variant of the famous Hilbert curve: Our inside-out Hilbert curve snakes through a 2D shape with arbitrarily rough fractal boundaries, achieving γ close to 1. Our findings highlight the potential of genomic questions to catalyze mathematical discoveries seemingly unrelated to biology.

Another way of interpreting values of γ is by using physical simulations to identify polymer states with similar γ values. In our original Hi-C study, we showed that a polymer that was crushed by external forces naturally folds into a fractal globule with a value of $\gamma = 1$. In the present work, we considered the possibility that internal forces, attractions between pairs of monomers, may also play a role. We found that varying the ratio of internal and external forces results in a family of possible structures. At one extreme, when external forces dominate, the result of the condensation process is symmetrical, yielding a classic fractal globule with $\gamma = 1$. At the other extreme, when internal forces dominate, tension arises along the polymer chain, leading to anisotropic condensation with $\gamma = 0.72$. Thus, the value of γ observed in these tension globules closely matches the value of γ observed in Hi-C contact domains. The tension globule model is consistent with two other important observations: the formation of contact domains inside loops and the 3D distance between pairs of loci as measured by 3D-FISH.

However, the tension globule model does not explain many aspects of our data. These limitations emerge from the putative model of loop formation in a tension globule, in which diffusion brings loop anchors together. Such a model makes it difficult to understand why loops tend not to overlap or why CTCF motifs at pairs of loop anchors must lie in the convergent orientation.

To overcome these limitations, we explored a different model of loop formation based on a proposal by Nasmyth (25), who hypothesized that loops form during metaphase chromosome condensation through the action of an extrusion complex comprising two tethered DNA-binding subunits, each of which extrudes DNA as they slide, relative to the genome, in opposite

directions. He suggested that such a process might involve cohesin proteins, which form a tripartite ring that can slide along DNA and chromatin. To date, no direct evidence has been observed in support of this model.

By means of physical simulations, we show that loop formation by extrusion results in the formation of a contact domain between the loop anchors, yields a γ value that closely matches the experimental data, and accurately estimates pairwise spatial distances measured using 3D-FISH. Moreover, the extrusion model can recapitulate the results of Hi-C experiments at short range (<2 Mb) in silico, including the position of peaks and contact domains, using only data about CTCF-binding sites from ChIP-Seq.

The extrusion model also explains two key features not explained by the tension globule. First, the extrusion model provides a natural explanation for why loops must lie between convergent CTCF motifs. Second, the model explains the strong tendency of loops not to overlap: Whereas simple diffusion is likely to result in substantial overlap, extruded loops cannot overlap. The extrusion model also has advantageous topological properties. Whereas simple diffusion could give rise to intrachromosomal knots and interchromosomal entanglements, extrusion would create neither intrachromosomal knots (facilitating chromatin accessibility during interphase) nor interchromosomal entanglements (facilitating chromosome condensation and segregation during metaphase) (*SI Appendix*, Fig. S12*F*).

For these reasons, we strongly favor the extrusion model.

It is interesting to consider possible structures for an extrusion complex. One such structure would comprise two cohesin rings and two CTCF proteins (Fig. 8 *A* and *C*), with the complex loaded onto DNA via loading of the cohesin rings at adjacent DNA sites and the simultaneous binding of the CTCF proteins nearby. Each CTCF/cohesin pair would then serve as a DNA-binding subunit. The subunits would engage DNA in an antisymmetrical fashion, sliding in opposing directions (one forward and one reverse) until the presence of a CTCF motif on the appropriate strand is detected. In fact, ChIP-Seq data provide evidence of antisymmetrical positioning of the CTCF and cohesin proteins with respect to chromatin loop anchors. Despite the fact that both proteins are associated with the same DNA motif, close examination of ChIP-Seq peaks at the ends of loops shows that CTCF tends to lie toward the outside of a loop, whereas RAD21 and SMC3 are positioned ∼20 bp closer to the loop interior (*SI Appendix*, Fig. S18). This result could indicate that the cohesin ring trails behind the CTCF protein as they slide along DNA.

Other structures are possible. Extrusion complexes might contain CTCF but not cohesin, with cohesin binding only after a loop has been formed. Alternatively, extrusion complexes may include cohesin, but not CTCF. In such a model, the complex would encounter CTCF proteins bound at their target sites and would either continue or halt depending on their orientation (Fig. 8*B*).

We also demonstrate that it is possible to reengineer loops and domains by modifying the CTCF motifs that lie at loop anchors in accordance with the convergent rule. We show that inserting even a single base pair can eliminate multiple loops and domains, affecting genome folding at the megabase scale. Moreover, our extrusion model accurately predicts the Hi-C contact map of an engineered region, including the positions of loops and domains, using only binding sites for CTCF in WT cells as input.

In some cases, our experiments suggest that particular loops tend to occur simultaneously. Specifically, our analyses suggest that both cliques studied in detail using CRISPR appear to fold into one of two spatial states: In some cells, only the A-C loop domain is present, whereas in other cells, both the A-B and B-C loop domains are simultaneously present and the A-C loop domain is absent. These findings suggest the possibility that overlapping contact domains may reflect alternative folding states within a cell population.

Our models do not address one important feature of Hi-C data: the observation that contact domains fall into at least two compartments and six subcompartments, each comprising domains that exhibit similar chromatin modifications and long-range contact patterns. Compartmentalization, seen in humans and many other species, manifests as a plaid arrangement in Hi-C maps. Our simulations do not recapitulate this phenomenon, indicating that other mechanisms are responsible for positioning each contact domain in its nuclear neighborhood.

The ability to read out the 3D structure of a genome is improving rapidly. As shown by our genome-editing experiments, it may now be possible not only to read genome folding patterns but also to write them.

ACKNOWLEDGMENTS. We thank S. Koch, C. McMullen, A. Tripathy, S. Batra, and J. Onuchic for discussions; S. Knemeyer for figure assistance; N. Tarazi for animations; and I. Dodd for insights on flippase recombination. Horizon Genomics assisted with the generation of several engineered cell lines. This study was supported by National Science Foundation (NSF) Grant PHY-1308264 (to A.L.S.), NSF Grant PHY-1427654, NIH New Innovator Award 1DP2OD008540-01, Cancer Prevention Research Institute of Texas Scholar Award R1304, a McNair Medical Institute Scholar Award, and the President's Early Career Award in Science and Engineering, and funding from the Welch Foundation, International Business Machines, and Nvidia (to E.L.A.).

1. Kornberg RD, Lorch Y (1999) Twenty-five years of the nucleosome, fundamental particle of the eukaryote chromosome. *Cell* 98(3):285–294.
2. Hansen JC (2002) Conformational dynamics of the chromatin fiber in solution: Determinants, mechanisms, and functions. *Annu Rev Biophys Biomol Struct* 31(1):361–392.
3. Hartwig M (1982) The size of independently supercoiled domains in nuclear DNA from normal human lymphocytes and leukemic lymphoblasts. *Biochim Biophys Acta* 698(2):214–217.
4. Zehnbauer BA, Vogelstein B (1985) Supercoiled loops and the organization of replication and transcription in eukaryotes. *BioEssays* 2(2):52–54.
5. Lieberman-Aiden E, et al. (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 326(5950):289–293.
6. Dixon JR, et al. (2012) Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* 485(7398):376–380.
7. Sexton T, et al. (2012) Three-dimensional folding and functional organization principles of the Drosophila genome. *Cell* 148(3):458–472.
8. Rao SSP, et al. (2014) A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* 159(7):1665–1680.
9. Sachs R, Engh G, Trask B, Yokota H, Hearst J (1995) A random-walk/giant-loop model for interphase chromosomes. *Proc Natl Acad Sci USA* 92(2):2710–2714.
10. Mateos-Langerak J, et al. (2009) Spatially confined folding of chromatin in the interphase nucleus. *Proc Natl Acad Sci USA* 106(10):3812–3817.
11. Schleif R (1992) DNA looping. *Annual Rev Biochem* 61(1):199–223.
12. Cullen KE, Kladde MP, Seyfred MA (1993) Interaction between transcription regulatory regions of prolactin chromatin. *Science* 261(5118):203–206.
13. McKean HP (1955) Hausdorff-Besicovitch dimension of Brownian motion paths. *Duke Math J* 22:229–234.
14. Cong L, et al. (2013) Multiplex genome engineering using CRISPR/Cas systems. *Science* 339(6121):819–823.
15. Wedemann G, Langowski J (2002) Computer simulation of the 30-nanometer chromatin fiber. *Biophys J* 82(6):2847–2859.
16. Wang JC, Davidson N (1966) On the probability of ring closure of lambda DNA. *J Mol Biol* 19(2):469–482.
17. Aiden EL (2010) Evolution and the emergence of structure. PhD thesis (Harvard University, Cambridge, MA).
18. Ringrose L, Chabanis S, Angrand PO, Woodroofe C, Stewart AF (1999) Quantitative comparison of DNA looping in vitro and in vivo: chromatin increases effective DNA flexibility at short distances. *EMBO J* 18(23):6630–6641.
19. Gomes M (1987) Paper crushes fractally. *J Phys A* 20:283–284.
20. Mandelbrot B (1967) How long is the coast of britain? Statistical self-similarity and fractional dimension. *Science* 15(3775):636–638.
21. de Gennes PG (1985) Kinetics of collapse for a flexible coil. *J Phys Lett-Paris* 46(14):639–642.
22. Langowski J, Heermann D (2007) Computational modeling of the chromatin fiber. *Sem in Cell & Dev Bio* 18(5):659–667.
23. Plimpton S (1995) Fast parallel algorithms for short-range molecular dynamics. *J Comp Phys* 117(1):1–19.
24. Brown WM, Wang P, Plimpton SJ, Tharrington AN (2011) Implementing molecular dynamics on hybrid high performance computers - short range forces. *Comp Phys Comm* 182(4):898–911.
25. Nasmyth K (2001) Disseminating the genome: joining, resolving, and separating sister chromatids during mitosis and meiosis. *Annu Rev Genet* 35:673–745.
26. Alipour E, Marko JF (2012) Self-organization of domain structures by DNA-loop-extruding enzymes. *Nucl Acids Res* 40(22):11202–11212.
27. León P, Macaya G (1983) Properties of DNA rosettes and their relevance to chromosome structure. *Chromosoma* 88(4):307–314.
28. Gnirke A, et al. (2009) Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nature Biotech* 27(2):182–189.
29. Nishino Y, et al. (2012) Human mitotic chromosomes consist predominantly of irregularly folded nucleosome fibres without a 30-nm chromatin structure. *EMBO J* 31:1644–1653.
30. Ricci MA, Manzo C, García-Parajo M, Lakadamyali M, Cosma MP (2015) Chromatin fibers are formed by heterogeneous groups of nucleosomes in vivo. *Cell* 160(6):1145–1158.

Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes – SI Appendix

Adrian L. Sanborn[a,b,c,1], **Suhas S. P. Rao**[a,d,1], **Su-Chen Huang**[a], **Neva C. Durand**[a,2], **Miriam H. Huntley**[a,2], **Andrew I. Jewett**[a,2], **Ivan D. Bochkov**[a], **Dharmaraj Chinnappan**[a], **Ashok Cutkosky**[a], **Jian Li**[a,b], **Kristopher P. Geeting**[a], **Andreas Gnirke**[e], **Alexandre Melnikov**[e], **Doug McKenna**[a,f], **Elena K. Stamenova**[a,e], **Eric S. Lander**[e,3], **Erez Lieberman Aiden**[a,b,e,3]

[a]The Center for Genome Architecture, Baylor College of Medicine, Houston, TX 77030, USA; [b]Center for Theoretical Biological Physics, Rice University, Houston, TX 77030, USA; [c]Department of Computer Science, Stanford University, Stanford, CA 94305, USA; [d]School of Medicine, Stanford University, Stanford, CA 94305, USA; [e]Broad Institute of MIT and Harvard, Cambridge, MA 02139, USA; [f]Mathemaesthetics, Inc., Boulder, CO 80306, USA; [1]A.L.S. and S.S.P.R. contributed equally to this work; [2]N.C.D., M.H.H., and A.I.J. contributed equally to this work; [3] To whom correspondence may be addressed. Email: lander@broadinstitute.org or erez@erez.com.

Table of Contents

I. Experimental Methods
- a. Hi-C
  - i. Previously generated Hi-C data
  - ii. New Hi-C data sets
- b. Nuclear Volume
- c. 3D DNA FISH
- d. CRISPR workflow
  - i. Experimental Design
  - ii. Guide RNA and HDR template design
  - iii. Cell culture and transfection
  - iv. Mutation strategy
  - v. in situ Hi-C on mutated cell lines
- e. Hi-C^2
  - i. Probe design
  - ii. Probe construction
  - iii. Hybrid selection
  - iv. Hi-C^2 data processing

II. Computational analysis of Hi-C maps
- a. Measurements of chromatin flexibility
  - i. Cyclization probability measurements
  - ii. Cyclization probability is consistent across compartments
  - iii. Contact probability decays at distances greater than 5kb
  - iv. Cyclization probability data are inconsistent with the 30-nm fiber
  - v. Experimental ionic conditions are comparable to physiological conditions
- b. Measurements of local contact probability
  - i. Definition of contact probability
  - ii. Local contact probability scalings are highly reproducible
  - iii. Contact probability within domains exhibits a power law with $\gamma \approx 0.75$
  - iv. Directional contact probability exhibits similar values of γ
  - v. Values of γ are consistent across compartments
  - vi. Aggregate measures of contact probability consistently overestimate γ
- c. Fractal models are inconsistent with intra-domain contact probability
- d. Analysis of chromatin loop networks
  - i. Hubs are detected are (nearly) isolated cliques in the loop network
  - ii. An extended list of loops accounts for false negative loops
  - iii. Hubs often involve a series of consecutive loops
  - iv. Hubs are consistent with looping between pairs of convergent CTCF motifs
  - v. Middle loci of hubs show strong enrichment for pairs of divergent CTCF motifs
  - vi. All observed effects are highly enriched over suitable null models
  - vii. An "exclusion domain" often forms between two loci that loop to a common third locus

III. Polymer simulations
- a. Molecular dynamics simlations
  - i. Polymer properties
  - ii. Simulation dynamics
  - iii. Simulations of tension globules
  - iv. Simulations of fractal globules
  - v. Simulations of tension globule loop domains
  - vi. Simulations of loop extrusion
- b. Simulated CTCF binding locations
  - i. Assigning extrusion complex binding strengths from CTCF ChIP-Seq tracks
  - ii. Assigning tension globule loops from CTCF ChIP-Seq tracks
- c. Analysis of the tension globule
  - i. Tension globule contact probability matches experimental observations
  - ii. Contact probability scalings are robus to changes in simulation parameters
  - iii. The tension globule exhibits a linear axis
  - iv. The tension globule is unknotted
  - v. Tension vs. fractal structure depends on ratio of internal and external forces
  - vi. Contact probability of simulated loop domains matches experimental observations
- d. Analysis of the extrusion model
  - i. Contact probability of extruded domains matches experimental observations
  - ii. The extrusion model accurately recapitulates Hi-C contact maps, including loops and domains

iii. The extrusion model accurately predicts changes in 3D structure after CRISPR editing
iv. These properties are robust to changes in the attractive forces and the initial configuration
v. The extrusion model predicts that loops will be intrachromosomal and tend to be short
vi. Loop extrusion promotes an unentangled, unknotted topology
e. The tension globule and the extrusion model are consistent with 3D DNA FISH measurements
f. SMC3 and RAD21 are positioned on loop anchors 20bp towards the loop interior
g. Monte Carlo simulations of large fractal globules using Confined-BFACF
IV. Mathematical theory
a. Introduction
b. Construction of self-similar curves
c. Dimension scaling theorem for self-similar curves
i. Overview
ii. An upper bound on dim $f(X)$
iii. A lower bound on dim $f(X)$
d. Derivation of contact probability critical exponent for self-similar paths
i. Dimension of the contact map derives from surface dimension
ii. Contact probability measures the contact map dimension
iii. Appendix: proofs
e. Novel curves
i. Construction of space-filling curves using tiling
ii. Rough-boundaried space-filling curves
iii. Inside-Out Hilbert curves

I. Experimental Methods

a. Hi-C

Most of the Hi-C datasets used in our analyses were generated using either the *in situ* Hi-C protocol or the dilution Hi-C protocol in our previous study (8) (GEO accession GSE63525). All of the Hi-C data was processed using the computational pipeline described in full detail in (8). In brief, the pipeline uses BWA (S1) to map each read end separately to the b37 reference genome, removes duplicate and near-duplicate reads, removes reads that map to the same fragment, and filters the remaining reads based on mapping quality score. Contact matrices were generated at base pair delimited resolutions of 2.5Mb, 1Mb, 500kb, 250kb, 100kb, 50kb, 25kb, 10kb, and 5kb, as well as fragment-delimited resolutions of 500f, 200f, 100f, 50f, 20f, 5f, 2f, and 1f. All Hi-C contact matrices were normalized using a matrix balancing algorithm (S2) in order to correct for coverage biases. All analyses performed in this study were performed on the KR-normalized data. The locations of contact domains and loops used in our analyses were annotated as described in (8).

i. Previously generated Hi-C data

The primary Hi-C map used in our analyses was a one-kilobase resolution map consisting of over 4.9 billion contacts, generated in the GM12878 human lymphoblastoid cell line using the MboI restriction enzyme. All analyses involving contact domains in this map were performed using a list of 9,274 domains annotated via the Arrowhead algorithm. Hi-C maps and domain annotations generated in the IMR90, K562, HeLa, HMEC, and NHEK cell lines using *in situ* Hi-C and the MboI restriction enzyme were also used for supplementary analysis. In addition, *in situ* Hi-C data sets generated in GM12878 using the DpnII restriction enzyme and dilution Hi-C data sets generated in the GM12878 cell line using the dilution Hi-C protocol and the HindIII or NcoI restriction enzymes were also used for analysis of single-fragment cyclization probabilities. Native Hi-C data sets (with no crosslinking) generated in GM12878 using the MboI restriction enzyme were also used in the cyclization probabilities analysis as a control. All of these Hi-C datasets were previously reported in (8) and have been previously uploaded to GEO accession GSE63525. All domain and loop annotations from these previously reported maps that were used in this study are also available at GEO accession GSE63525.

ii. New Hi-C datasets

We also generated 6 additional Hi-C datasets using the *in situ* Hi-C protocol, varying the duration of crosslinking. In brief, the *in situ* Hi-C protocol involves crosslinking cells with formaldehyde, permeabilizing them with nuclei intact, digesting DNA with a suitable 4-cutter restriction enzyme (in our experiments we used MboI), filling the 5'-overhangs while incorporating a biotinylated nucleotide, ligating the resulting blunt-end fragments, shearing the DNA, capturing the biotinylated ligation junctions with streptavidin beads, and analyzing the resulting fragments with paired-end sequencing. In order to test the possibility that the strength of formaldehyde crosslinking could affect the contact probability scalings observed, we performed experiments using 1, 5, or 10 minutes of crosslinking with 1% formaldehyde at 22C or 37C. For each experiment, we computed the genome-wide contact probability scaling (defined below) at distances from 30kb to 300kb and compared to contact probability of the primary map (crosslinked at 22C for 10 minutes). The values of γ obtained are listed in Table S2.

For our genome engineering experiments, we generated 5 *in situ* Hi-C libraries (using the same protocol from (8)) from Hap1 wild-type cells that were deeply sequenced to create our genome-wide wild-type Hap1 Hi-C map. We also generated 63 *in situ* Hi-C libraries from Hap1 wild-type and various mutant cell lines that we performed hybrid selection on to generate Hi-C^2 libraries (see section I.e below for information on the Hi-C^2 protocol).

All new in situ Hi-C libraries generated as part of this study are detailed in Table S6.

b. Nuclear Volume

All cell lines were cultured according to the supplier's instructions. Adherent cell lines were grown on the surface of pre-cleaned glass slides in the appropriate growth media, then stained with 2M CellTracker Red CMTPX fluorescent dye (Life Technologies, C34552) in fresh serum-free medium for 45 minutes at 37C, 5% CO2. Suspension cell lines were harvested by centrifugation at 300xG for 5 minutes and re-suspended in serum-free medium with CellTracker Red for staining, as above. After incubation, medium with dye was removed and cells were again incubated in fresh medium for 30 minutes. Cells were then washed with 1X phosphate-buffered saline (PBS) and fixed in 4% formaldehyde in 1X PBS for 15

minutes at ambient temperature. After fixation, cells were washed twice with 1X PBS. Suspension cell lines were diluted to a concentration of 1×10^6 cells/mL in 1X PBS and were allowed to attach to the surface of poly-L-lysine coated microscope slides. Nuclei were stained with DAPI (100ng/mL) for 30 minutes at ambient temperature in the dark. Slides were imaged with either an LSM 700 or an LSM 780 laser scanning inverted confocal microscope (Zeiss), 40X oil objective, ZEN 2011 imaging software (Zeiss), at 405nm and 594nm. Nuclear and cytoplasmic volumes were determined using Volocity v.6 software (Perkin Elmer). The intensity threshold of each channel was kept constant for all measurements.

c. 3D DNA FISH

3D DNA fluorescence *in situ* hybridization (FISH) was performed essentially as described in (8), based on the Oligopaints method of (S3). Briefly, probes were designed with OligoArray as a pool of 32bp sequences tiling nine 30kb target loci (3 per region at a density of 9-16 probes per kilobase), flanked by 21bp primer sequences unique to each locus (Table S5). The 74bp full sequences were ordered as an oligonucleotide pool (CustomArray, Inc.). Forward primers were synthesized with a 5' conjugated fluorophore (Alexa Fluor 488, ATTO 565, or Alexa Fluor 647) and purified by HPLC (Integrated DNA Technologies). IMR90 cells were cultured on Lab-Tek II glass chamber slides (Thermo Fisher Scientific). Further details on experimental procedures (including measures to minimize bias due to chromatic aberration), microscopy, and image processing can be found in (8).

d. CRISPR workflow

In brief, our CRISPR workflow consisted of the following steps: (i) identifying chromatin loops using HiCCUPS (8), (ii) identifying unique, correctly oriented CTCF motifs within loop anchors (as defined in (8)), (iii) rationally designing a CRISPR guide RNA or multiple guide RNAs to cut within or around the CTCF motif while optimizing for cutting efficiency and minimizing off-target effects, (iv) optionally designing homology directed repair (HDR) templates to specifically invert or replace the CTCF motif, (v) transfecting cells with the Cas9 and the guide RNA(s) (and optionally the HDR template), (vi) sorting single transfected cells via fluorescence-activated cell sorting (FACS), (vii) growing up and genotyping clonal populations of cells, (viii) selecting clonal cell lines with mutations disrupting the CTCF motif (or in the case of HDR, the specific desired mutation), (ix) performing *in situ* Hi-C on the selected mutated cell lines, and (x) performing hybrid selection on the *in situ* Hi-C libraries for a region around the targeted CTCF motif to generate Hi-C^2 libraries that can easily and cheaply be sequenced to read off the effects of our mutations on genome folding.

While we performed our CRISPR experiments in the Hap1 cell line to be able to read off the effects of our mutations without having to worry about allelic heterozygosity, our CRISPR workflow is easily adaptable to other cell lines, as long as one has a reasonable Hi-C map to identify loops and guide predictions (in fact, since our extrusion model is very good at predicting genome folding from CTCF ChIP-Seq alone, in many cases the requirement for Hi-C data may not even be essential). The steps in our workflow are described in detail below.

i. Experimental Design

Three regions containing triple-hubs (three loci A, B and C with all pair-wise loops present) were chosen for thorough dissection. Since we did not have CTCF or cohesin ChIP-Seq data for Hap1, the regions were chosen such that they showed extremely similar patterns of chromatin folding to GM12878 and IMR90, so that ChIP-Seq data from those cell lines could be used to identify precise motifs in loop anchors to target as well as to simulate folding in the regions (see section III.d.iii below).

The three hubs were chosen such that unique anchors (as defined in (8)) were present at least at the middle loop anchor and ideally at one of the upstream or downstream loop anchors as well. Motifs in loop anchors were identified using FIMO (S4) using the CTCF motif position weight matrices (PWMs) from Kim, et al (S5) and Schmidt, et al (S6). The hubs were chosen such that all loops were clearly anchored by correctly oriented motifs. Motifs to target via CRISPR were only chosen if they were clearly unique among the correctly oriented motifs in a ChIP-Seq binding site (i.e. there was only one motif present or only one motif that was clearly the strongest match when compared against both PWMs and in the case of the middle loop anchor, the reverse CTCF motif corresponding to the A-B loop was upstream of the forward CTCF motif corresponding to the B-C loop).

Targeted motifs and regions are listed in Table S7 and S8.

ii. Guide RNA and HDR template design

Guide RNAs were designed using one of two strategies: (i) a single guide RNA was designed to cut inside the target CTCF motif or (ii) two guide RNAs were designed to cut flanking both sides of the target CTCF motif.

Prospective guide RNAs were screened using the cutting efficiency scoring scheme from Doench, et al (S7) and the off-target scoring scheme from Hsu, et al (S8). Wherever possible, guides with cutting efficiency scores of 0.4 or lower were avoided, and guide RNAs with scores of lower than 0.25 were discarded altogether. Wherever possible, guides ranked as high quality guides by the Hsu off target assessment algorithm were used. In a few cases, where no high quality guide was identified or when the cutting efficiency as ranked by the Doench, et al algorithm was extremely low, a mid-quality guide (with respect to off-targets) was used.

Guide RNAs for all mutations are listed in table S7.

All the HDR templates used in this study were ssODNs (S9), either 200bp (IDT ultramers) or 100bp (Invitrogen custom DNA oligonucleotides) in size. They were designed such that they contained the 20bp CTCF motif inverted (or a new 20bp CTCF motif), flanked by homology arms either 90bp or 40bp in size.

iii. Cell culture and transfection

This culture and transfection protocol was used for the 8 mutated cell lines generated entirely within our lab; 5 of the mutated cell lines were ordered from Horizon Genomics and created using their proprietary methods (see Table S7). The experimental design and guide RNA design for all 13 experiments was conducted as described above.

Hap1 cells (Horizon Genomics) were cultured according to manufacturer's conditions. 24 hours before transfection, 0.9M Hap1 cells were plated in each well of a 6 well plate. After 24 hours, when the cells were roughly 60% confluent, the cells were transfected with the pSpCas9(BB)-2A-GFP (px458) plasmid from the Zhang lab (S9). Guide RNAs were cloned into the plasmid using the protocol provided at `http://www.genome-engineering.org/crispr/wp-content/uploads/2014/05/CRISPR-Reagent-Description-Rev20140509.pdf` (14).

The Hap1 cells were transfected (in antibiotic free media) with 3µg of DNA using Turbofectin according to manufacturer's instructions (a 3:1 ratio of Turbofectin to DNA was used; 9µl of Turbofectin for 3µg of DNA). For single guide

RNAs, 3µg of the Cas9-gRNA plasmid was used. For double guide RNA mediated deletions, 1.5µg of each Cas9-gRNA plasmid was used. For HDR, either 1.5µg of Cas9-gRNA plasmid and 3µl 10µM 200bp ssODN or 1.875µg Cas9-gRNA plasmid and 3.75µl 10µM 100bp ssODN were used. For HDR experiments, the culture media was supplemented with 0.1µM SCR7 (S10, S11) 12-24 hours after transfection.

24-48 hours after transfection, GFP+ cells were sorted via FACS (PI was also added to filter for dead cells). Transfection efficiencies were usually between 5 and 10%. Populations of 500-10,000 cells were screened for gRNA cutting efficiency or for HDR efficiency to judge roughly how many clones would need to be screened. Single cells were sorted into individual wells of a 96-well plate and allowed to grow for 10-14 days. After that, roughly 32-96 clones were screened per transfection.

iv. Mutation strategy

Deletions were obtained either via a single guide RNA-mediated cut within the CTCF motif or via two guide RNAsmediated double strand breaks on either side of the CTCF motif. In the case of the single guide RNA mediated cuts, clones were screened for mutations that were as small as possible, but also highly likely to completely disrupt CTCF binding (as judged by the strength of the motif match before and after mutation). Mutations that were likely to completely abrogate CTCF binding were selected for expansion. Mutations generated via two double strand breaks were all generated by Horizon Genomics and clones containing the region between the two guide RNAs either cut out or inverted were selected for expansion. Clones targeted with HDR were screened for the 20bp inversion or 20bp replacement and successfully targeted clones were selected for expansion.

v. in situ Hi-C on mutated cell lines

Expanded mutant clones were crosslinked as in (8) and subsequently *in situ* Hi-C was performed on the pellets as described in (8) and summarized above. On average, 4.3 *in situ* Hi-C libraries were generated per mutated cell line for a total of 56 *in situ* Hi-C libraries.

e. Hi-C^2

i. Probe design

To design probes targeting a particular region for HYbrid Capture Hi-C (Hi-C^2), we first identified all restriction sites within the target region. Since Hi-C ligation junctions occur between restriction sites, we designed our bait probe sequences to target sequences within a certain distance of the (MboI) restriction sites present in our target region. Specifically, we performed a first pass, scanning all 120bp sequences with one end within 80bp of a restriction site and selecting, for each restriction end (i.e. both upstream and downstream of the restriction site), the closest 120bp sequence to the restriction site that had fewer than 10 repetitive bases (as determined by the repeat masked hg19 genome downloaded from UCSC) and had between 50% and 60% GC content. If there was no probe satisfying those criteria, the closest probe with between 40% and 70% GC content but satisfying all the other above criteria was retained. The GC content bounds were chosen based on the hybridization bias data presented in (28).

After the first pass, we removed one probe from any pair of probes that overlapped. We then identified any gaps in the probe coverage (intervals larger than 110bp) and identified any restriction sites falling within those gaps. We then searched for additional 120bp probes with a looser set of criteria: For each restriction site within a gap, we scanned all 120bp sequences with one end within 110bp of a restriction site and selected the closest sequence to the restriction site that had fewer than 20 repetitive bases and had between 40% and 70% GC content. After the second pass, we once again identified gaps of at least 110bp in the probe coverage. For gaps that fell within 5kb windows in the target region that were covered by fewer than 5 probes, we performed a third probe design pass. For each restriction site within these low coverage gaps, we scanned all 120bp sequences with one end within 110bp of a restriction site and selected the closest sequence to the restriction site that had fewer than 25 repetitive bases and had between 25% and 80% GC content.

After all three passes, we identified 3107 probes covering region 1 (chr8:133-135Mb; 1.55 probes/kb), 2666 probes covering region 2 (chr1:179.8-181.8Mb; 1.33 probes/kb), and 2497 probes covering region 3 (chr5:31-33Mb; 1.25 probes/kb). 15bp primer sequences (unique for each region) were appended to either end of the 120bp probe sequence in order to allow for synthesis of all probes together in one oligo pool and subsequent amplification of region-specific sub-pools (see below).

ii. Probe construction

We obtained custom synthesized pools of 150bp (120bp + 15bp primer sequence on either end) single stranded oligodeoxynucleotides from CustomArray, Inc. (Bothell, WA). The oligonucleotides were of the form TCGCGCCCATAACTCN$_{120}$CTGAGGGTCCGCCTT for Region 1, ATCGCACCAGCGTGTN$_{120}$CACTGCGGCTCCTCA for Region 2, and CCTCGCCTATCCCATN$_{120}$CACTACCGGGGTCTG for Region 3. Region-specific sub-pools were first amplified from the overall CustomArray oligo pool using the following mix and PCR profile:

2µl oligo pool (160 ng)
6µl Primer 1 (10µM)
6µl Primer 2 (10µM)
36µl H2O
50µl 2X Phusion master mix
100µl TOTAL

Amplify for 10-18 cycles using the following PCR profile:

98C for 30s
  98C for 10s
  55C for 30s
  72C for 30s cycle 10-18 times
72 for 7min
hold at 4C

where Primer 1 was CTGGGATCGCGCCCATAACTC for Region 1, CTGGGAATCGCACCAGCGTGT for Region 2, and CTGGGACCTCGCCTATCCCAT for Region 3 and Primer 2 was CGTGGAAAGGCGGACCCTCAG for Region 1, CGTGGATGAGGAGCCGCAGTG for Region 2, CGTGGACAGACCCCGGTAGTG for Region 3.

After the initial amplification of the region-specific sub-pool, a 1X SPRI clean up was performed on the 162bp PCR product to remove primers and primer-dimers. We then performed a second PCR amplification to add a T7 promoter, using the following mix and PCR profile:

2µl first PCR product
12µl Primer 1-T7 (10µM)
12µl Primer 2 (10µM)
74µl H2O
100µl 2X Phusion master mix
200µl TOTAL

Amplify for 12-18 cycles using the following PCR profile:

98C for 30s
98C for 10s
55C for 30s
72C for 30s cycle 12-18 times
72 for 7min
hold at 4C

where Primer 1-T7 was `GGATTCTAATACGACTCACTATAGGGTCG CGCCCATAACTC` for Region 1, `GGATTCTAATACGACTCACTATAG GGATCGCACCAGCGTGT` for Region 2, and `GGATTCTAATACGACT CACTATAGGGCCTCGCCTATCCCA` for Region 3.

After the second PCR, once again, we performed a 1X SPRI clean up to purify the 182bp PCR product. We then used the purified second PCR product as the template in a MAXIScript T7 transcription reaction (Ambion) as follows:

Xµl purified DNA template (1ug)
10µl T7 enzyme mix
10µl 10X transcription buffer
5µl 10mM ATP
5µl 10mM CTP
5µl 10mM GTP
4µl 10mM UTP
1µl 10mM Biotin-16-UTP
Yµl H2O
100µl TOTAL

After incubating the reaction for at least 90 minutes at 37^{o}C, we added 1µl of TURBO DNase 1 and incubated at 37^{o}C for 15 minutes to remove template DNA. We added 1µl of 0.5M EDTA to stop the reaction and removed unincorporated nucleotides and desalted the RNA by purifying using a Zymo Oligo Clean and Concentrator column (following manufacturer's instructions). Our RNA yield was typically 5-15µg of RNA per reaction, so we measured the concentration of the RNA prior to the column cleanup using a Qubit RNA assay in order to determine whether to use one or two columns (the capacity of one of the Zymo columns is 10µg). For long-term storage of the RNA probes, we added 1U/µl of SUPERase-In RNase inhibitor (Ambion) and stored at -80C.

iii. Hybrid selection

Final *in situ* Hi-C libraries were assessed for quality using the metrics outlined in Rao and Huntley, et al (8). High quality libraries of sufficient complexity were selected for hybrid capture. 500ng of Hi-C library was used as the pond for the hybrid selection reaction; libraries were diluted to a concentration of 20ng/µl (i.e. 25µl of library was used). For a few libraries that were under 20ng/µl in concentration, as low as 250ng total was used (still in 25µl).

For the hybridization reaction, 25µl of pond was mixed with 2.5µg (1µl) of Cot-1 DNA (Invitrogen) and 10µg (1µl) of salmon sperm DNA (Stratagene). The DNA mixture was heated to 95^{o}C for 5 minutes and then held at 65^{o}C for at least 5 minutes. After at least 5 minutes at 65^{o}C, 33µl of prewarmed (65^{o}C) hybridization buffer (10X SSPE, 10X Denhardt's buffer, 10mM EDTA, and 0.2% SDS) and 6µl of RNA probe mixture (500ng of RNA probes, 20U of SUPERase-In RNase inhibitor; prewarmed at 65^{o}C for 2 minutes) were added to the DNA library for a total volume of 66µl. This mixture was incubated at 65^{o}C in a thermocycler for 24 hours.

After 24 hours at 65^{o}C, 50µl of streptavidin beads (Dynabeads MyOne Streptavidin T1, Life Technologies) were washed three times in 200µl of Bind-and-Wash buffer (1M NaCl, 10mM Tris-HCl, pH 7.5, and 1mM EDTA) and then resuspended in 134µl of Bind-and-Wash buffer. The beads were added to the hybridization mixture and incubated for 30 minutes at room temperature (with occasional mixing to prevent the beads from settling). After 30 minutes, the beads were separated with a magnet and the supernatant discarded. The beads were then washed once with 200µl low-stringency wash buffer (1X SSC, 0.1% SDS) and incubated for 15 minutes at room temperature. After 15 minutes, the beads were separated on a magnet and the supernatant discarded. The beads were then washed three times in high-stringency wash buffer (0.1X SSC, 0.1%SDS) at 65^{o}C for 10 minutes, each time separating the beads with a magnet and discarding the supernatant.

After the last wash, the DNA was eluted off the beads by resuspending in 50µl of 0.1M NaOH and incubating for 10 minutes at room temperature. After 10 minutes, the beads were separated on a magnet and the supernatant was transferred to a fresh tube with 50µl of 1M Tris-HCl, pH 7.5 (to neutralize the NaOH).

To desalt the DNA, we performed a 1X SPRI cleanup using 3X concentrated SPRI beads (taking 3 volumes of SPRI bead/solution mix, separating on a magnet, discarding 2 volumes of SPRI solution and resuspending the beads in the remaining 1 volume). We eluted the DNA in 22.5µl of 1X Tris buffer (10mM Tris-HCl, pH 8.0).

In order to prep the Hi-C^2 library for sequencing, we added 25µl of 2X Phusion and 2.5µl of Illumina primers and amplified the library for 12-18 cycles. After PCR, we performed two 0.7X SPRI cleanups to remove primers, etc. and then quantified the libraries for sequencing.

iv. Hi-C^2 data processing

Hi-C^2 libraries were sequenced to a depth of between 600K-60M reads (on average, 7.8M reads). All data was initially processed using the pipeline published in our previous study (8); however, additional processing was needed to properly normalize the Hi-C^2 data.

Normalization is an important problem to address in the analysis and interpretation of all proximity ligation experiments. We have previously shown that matrix balancing with the KR algorithm is an effective tool for properly normalizing Hi-C data (8). However, one requirement of the KR algorithm is the requirement of a square symmetric matrix. As hybrid selection strongly enriches for certain rows of the matrix corresponding to the target region, there are large regions of the overall matrix that are extremely sparse (entries corresponding to interactions between two non-target loci). As a result, performing KR matrix balancing on the overall matrix generated by a Hi-C^2 experiment does not efficiently correct both first-order hybrid selection target-enrichment biases and second-order hybridization biases within the target region.

To deal with this, we utilized the high resolution genome-wide in situ Hi-C map of wild-type of Hap1 we had already generated. Since all genome-editing perturbations were made within the region targeted using Hi-C^2, for every Hi-C^2 dataset, we spiked in data from the genome-wide wild-type Hap1 map corresponding to regions of the chromosome-wide matrix where both loci fall outside of the target region. Spiked data was added such that the average coverage of a locus in the overall chromosome-wide matrix was equal to the average coverage of loci within the target region. By spiking in data from the wild-type map where we expect to see no change (since there were no perturbations), we could remove the first-order bias from hybrid-selection target enrichment, and use KR matrix balancing on the entire chromosome-wide matrix (which is no longer extremely sparse) to correct the second-order hy-

bridization biases. Several different flavors of this normalization scheme were implemented yielding extremely similar results; they are described below and the results of the various normalizations are shown in Figure S14. Method e. below is the method used for all Hi-C^2 data shown in the main figures of this study.

a. Raw gap-filling: For a given resolution, the average intrachromosomal coverage of the loci within the target region (defined as the entire interval tiled by probes not specifically the loci that were covered by a probe) was calculated from the raw uncorrected Hi-C^2 matrix. Similarly, the average intrachromosomal coverage of all loci was calculated from the raw uncorrected genome-wide Hap1 wild-type Hi-C map. A matrix consisting of all entries corresponding to two loci that were both outside the target region was constructed from the raw uncorrected genome-wide Hap1 Hi-C map. This matrix was multiplied by the ratio of the average coverage of loci within the target region in the Hi-C^2 data to the average coverage of all loci from the genome-wide Hap1 wild-type Hi-C data and then summed with the Hi-C^2 matrix (thereby filling in the extremely sparse areas of the Hi-C^2 matrix). This summed matrix was then corrected with the KR matrix balancing algorithm. The resulting normalization factors were used as correction factors for the Hi-C^2 data.

b. KR gap-filling: The KR gap-filling normalization was performed similarly to the method described above, but to avoid corrected Hi-C biases and Hi-C^2 biases together, the method above was performed on KR normalized data. Specifically, the KR correction factors derived from the genome-wide Hap1 wild-type Hi-C map were used to perform an initial correction of the Hi-C^2 data. After the initial correction, the average intrachromosomal coverage of the loci within the target region (defined as the entire interval tiled by probes not specifically the loci that were covered by a probe) was calculated from the Hi-C^2 matrix. Similarly, the average intrachromosomal coverage of all loci was calculated from the corrected genome-wide Hap1 wild-type Hi-C map. A matrix consisting of all entries corresponding to two loci that were both outside the target region was constructed from the raw uncorrected genome-wide Hap1 Hi-C map. This matrix was multiplied by the ratio of the average coverage of loci within the target region in the Hi-C^2 data to the average coverage of all loci from the genome-wide Hap1 wild-type Hi-C data and then summed with the Hi-C^2 matrix (thereby filling in the extremely sparse areas of the Hi-C^2 matrix). This summed matrix was then corrected with the KR matrix balancing algorithm. The resulting normalization factors were used as correction factors for the Hi-C^2 data.

c. Raw gap-filling with rescaling: Filling in the sparse areas of the Hi-C^2 matrix corrects for first order target enrichment biases from hybrid capture to some extent, but does not account for the fact that differential enrichments may be present for entries of the matrix corresponding to one on-target loci and one off-target loci vs. entries corresponding to two on-target loci. To address this, before performing gap-filling as in the above methods, we first calculated for each locus in the target region, the ratio of the number of contacts formed between the locus and off-target loci to the number of contacts formed between the locus and other on-target loci using the genome-wide Hap1 wild-type Hi-C data. We then calculated the same ratio using the Hi-C^2 data. The ratio of these ratios provided a scaling factor for each on-target locus which we used to scale all entries in the Hi-C^2 matrix corresponding to contacts between the on-target locus and off-target loci. After performing this correction, we followed the method from method a, i.e. a matrix consisting of all entries corresponding to two loci that were both outside the target region was constructed from the raw uncorrected genome-wide Hap1 Hi-C map. This matrix was multiplied by the ratio of the average coverage of loci within the target region in the Hi-C^2 data (using the rescaled Hi-C^2 data) to the average coverage of all loci from the genome-wide Hap1 wild-type Hi-C data and then summed with the Hi-C^2 matrix (thereby filling in the extremely sparse areas of the Hi-C^2 matrix). This summed matrix was then corrected with the KR matrix balancing algorithm. The resulting normalization factors were used as correction factors for the Hi-C^2 data.

d. KR gap-filling with rescaling: This method is the same as method c., except that as in method b., the Hi-C^2 data was initially corrected with the KR factors derived from the Hap1 genome-wide wild-type Hi-C matrix and the KR corrected wild-type Hi-C data was used for gap-filling.

e. Raw gap-filling with rescaling and thresholding: We noticed that for a few very sparse (under-covered) rows in the Hi-C^2 data, our normalization methods would actually overcorrect, leading to highly-covered streak artifacts in the data. In order to remove these artifacts, we added a final filtering step, where loci with a normalization factor (C) of less than 0.33 (where $M_{i,j}$ is divided by C_i and C_j to get the corrected entry $M^*_{i,j}$) were thresholded so that their normalization factors were raised to 0.33 (this was implemented after the KR matrix balancing was run, not as a constraint during the running of the algorithm). The threshold of 0.33 was chosen based on empirical observation of rows that led to streaky artifacts. This method is the same as method c. except with the aforementioned thresholding.

f. KR gap-filling with rescaling and thresholding: This method is the same as method d. except with the addition of the thresholding described in method e.

Method e. was used for all Hi-C^2 data described in the main text and shown in the main figures of this study. (It was chosen because it makes no assumptions about underlying biases in the data except for the assumption that biases for loci that were not targeted for perturbation or for hybrid selection have identical biases to those present in the wild-type Hi-C data.)

II. Computational analysis of Hi-C maps

a. Measurements of chromatin flexibility

i. Cyclization probability measurements

At the smallest scale, models of chromatin structure rely on an estimate of the Kuhn length of a chromatin fiber (S12). Polymer theory predicts that higher order structures can only form at scales an order of magnitude larger than the Kuhn length. Because direct estimates of chromatin flexibility in vivo have not previously been available, inferences about the Kuhn length of chromatin have been based on theoretical, computational, and *in vitro* models (2, 15, 22, S13, S14).

If L_K is the Kuhn length of nuclear chromatin, measured in base pairs, single fragments of chromatin of length L will form few cycles when $L < L_K$ and many cycles when $L > L_K$. To empirically measure L_K, we filtered for Hi-C contacts formed by the two ends of a single chromatin fragment. We employed two criteria. First, for each restriction enzyme, we examined only contacts occurring between two ends of a single restriction fragment, assuming full cutting of the enzyme. Second, depending on the forward or reverse orientation of the alignment of each read end to the template, four "contact orientations" are possible. Contacts occurring between two ends of

a single fragment are necessarily "outer" contacts, characterized by reverse alignment of the upstream read and forward alignment of the downstream read. At small distances, outer contacts predominantly correspond to single fragments. Thus, we further restricted our analysis to outer contacts.

Cyclization probability at length L was computed as the number of single-fragment outer contacts of length L divided by the total number of restriction fragments of length L, binned logarithmically. Because contact maps are aggregated over samples from many cells, cyclization probability is a relative measure. Thus, we normalized to a probability distribution. Cyclization probability was computed on five Hi-C maps: two maps using restriction enzymes HindIII and NcoI, which recognize a six-basepair sequence and have cutting sites on average every 3.6kb; two maps using restriction enzymes MboI and DpnII, which recognize a four-basepair sequence and have cutting sites on average every 420bp; and one "native" map using MboI and no cross-linking preparation (8) (Fig 1B, Fig S1B)

ii. Cyclization probability is consistent across compartments

We also computed cyclization probability with single-fragment outer contacts grouped by each of the five subcompartments as identified in (8). Cyclization counts and restriction fragment lengths were compared against subcompartment boundaries and partitioned. Fragments that spanned more than one subcompartment were discarded; fewer than 0.1% of fragments were discarded in this manner. Cyclization probability was computed as described above, summing over only cyclization counts and restriction fragment lengths within each subcompartment. Plots of cyclization probability as function of fragment length were extremely consistent across all five subcompartments (Fig S1A).

In order to search for variation in local flexibility, we additionally examined cyclization probability of the MboI experiment in 1Mb windows across the genome. We consistently observed a steep rise in probability for distances less than 1kb and a flattening at distances larger than 1kb, consistent with genome-wide and compartment-wide averages. We noticed some variability in cyclization probability at distances larger than 1kb ? the probability in some regions showed slight increase as a function of distance while the probability in many regions remained flat for large distances, perhaps reflecting local variability. However, this analysis significantly noisier since few cycles are formed at large distances within each individual window.

iii. Contact probability decays at distances greater than 5kb

Genome-wide contact probability was computed (as described in Section II.b.i) on the primary GM12878 Hi-C map (Fig S2A). The contact probability decay is reliable at distances larger than the typical restriction fragment, which is around 2kb for maps using the MboI restriction enzyme (due to cutting inefficiency). We observe a contact probability decay at distances larger than 5kb, suggesting that $L_K < 5$kb.

iv. Cyclization probability data are inconsistent with the 30-nm fiber

The persistence length is the length of a polymer at which the tangent vectors at the two ends of the polymer become uncorrelated. For a worm-like chain, the Kuhn length is twice the persistence length. Coarse-grain computer simulations of the 30-nm fiber have estimated its persistence length to be between 120nm and 265nm across a range of parameters (15, S15-17). At a linear fiber density of 6 or 7 nucleosomes per 11nm, 200bp per nucleosome, this is equivalent to a persistence length between 15kb and 30kb, suggesting a Kuhn length between 30kb and 60kb.

The 30-nm fiber is significantly less flexible than a 10-nm fiber or a 10-nm fiber compacted into a polymer melt. Because the 30-nm fiber is characterized by a helical structure with six nucleosomes, or roughly 1.2kb, per turn, it is only flexible at length scales an order of magnitude larger, and is clearly inconsistent with our measurements of chromatin flexibility. More generally, our data are inconsistent with any repeating helical structure since cyclization probability is flat at distances larger than 1kb whereas repeating structures should show characteristic spikes at regular intervals. A view of chromatin as a melt of 10-nm fibers is consistent with many recent studies (29, 30, S18-20).

v. Experimental ionic conditions are comparable to physiological conditions

Ionic concentrations are known to affect fiber flexibility, with high concentration causing greater fiber compaction and decreased flexibility. *In situ* Hi-C experiments are conducted at high salt concentrations (10mM NaCl + 10mM $MgCl_2$) comparable to physiological conditions and simulations of 30-nm fibers described above.

b. Measurements of local contact probability

i. Definition of contact probability

The contact probability of two points separated by linear distance s was computed as the ratio $I(s) = I_{actual}(s)/I_{possible}(s)$. The numerator $I_{actual}(s)$ is computed as the number of contacts observed occurring at distance s. The denominator $I_{possible}(s)$ is computed as the possible number of contacts occurring at distance s; for a chain of length N, $I_{possible}(s) = N - s - 1$. All contact probability plots are displayed on log-log axes with distance s binned logarithmically.

Contact probability may be computed across the whole genome or within a specific window. For genome-wide chromatin contact probability, $I_{actual}(s)$ and $I_{possible}(s)$ were each aggregated over all chromosomes. For contact probability within a specified window, $I_{actual}(s)$ and $I_{possible}(s)$ were computed as the number of actual Hi-C contacts and possible contacts at distance s with at least one end in the chosen region.

Contact probability often exhibited a power law $I(s) \sim s{-}\gamma$, or "contact probability scaling", within a range of values of s. We measured γ as the slope of the best-fit line on $I(s)$, when plotted on log-log axes, within a chosen range of distances.

ii. Local contact probability scalings are highly reproducible

Local contact probability scaling exponents were highly reproducible across biological replicates. We partitioned chromosome 1 into 4,517 50kb windows and measured γ for three different distance regimes (10-100kb, 10-350kb, 10-1000kb) in each window and on each of the primary and replicate GM12878 Hi-C maps. Values of γ obtained from the primary and replicate maps were highly correlated in each regime (Pearson coefficients 0.93, 0.97, 0.98 respectively, Fig S2E).

iii. Contact probability within domains exhibits a power law with $\gamma \approx 0.75$

To measure γ within contact domains, we computed contact probability in a 50kb window at the center of each domain (Fig 2A). To obtain reliable measurements of intra-domain behavior, domains containing other domains in their interiors were excluded. For the GM12878 map, all such domains of size 200kb or larger were used; for all other cell types, all domains

of size 300kb or larger were used. Contact probability measured in this manner for all domains in the GM12878 map with sizes 200-220kb, 500-600kb, and 900-1200kb are plotted in Figure 2C.

For each domain of size L, we measured the contact probability exponent γ between distances of 10kb and $L/2-50\text{kb}-20\text{kb}$. All contacts in this distance range are necessarily intra-domain and 20kb from the domain boundary, avoiding boundary effects. Distributions of γ were consistent across domains of different sizes (Fig 2D), as well as across 6 additional cell types (Fig 2E, Table S1).

As an additional measurement of γ, we computed aggregated intra-domain contact probability, taking all contact pairs with both ends inside a single domain. All domains larger than 100kb that did not contain sub-domains were included, for a total of 5,265 domains. The intra-domain contact probability exhibited a clear scaling with $\gamma = 0.76$ between 10kb and 1Mb (Figure 2B). Notably, by using only intra-domain contacts in this aggregate measure, the scaling extended to distances larger than in genome-wide aggregate measurements.

iv. Directional contact probability exhibits similar values of γ

To assess whether local contact probability exponents varied substantially within domain, we compared exponents on different windows within single domains. Specifically, for a domain with left and right endpoints A and B, we picked two consecutive 50kb windows $R_1 = [(A+B)/2-50\text{kb}, (A+B)/2]$ and $R_2 = [(A+B)/2, (A+B)/2+50\text{kb}]$ flanking the domain center. On R_1 we computed the "leftward" contact probability, counting only actual and possible contacts (s_1, s_2) such that $s_1 < s_2$ and s_2 lies in R_1, and measured the intra-domain scaling exponent γ_1. Analogously, on R_2 we computed the "rightward" contact probability, counting only actual and possible contacts (s_1, s_2) such that $s_1 < s_2$ and s_1 lies in R_2, and measured the intra-domain scaling exponent γ_2. In this manner, the two measurements contained no common contacts. We found that γ_1 and γ_2 were uncorrelated and the distributions of each were similar to distributions of our original measurements of γ through the domain center (Fig S2F).

v. Values of γ are consistent across compartments

To assess the variability of γ between nuclear subcompartments, we computed γ for each of 1,730 domains that were more than 80% contained in a single subcompartment. We examined the five subcompartments as identified in (8) (excluding B4 which occupies a very small proportion of the genome). Contact probability scalings were extremely consistent across all subcompartments (Fig S2C), including both active compartments A1 (mean $\gamma = 0.743$) and A2 (mean $\gamma = 0.725$), as well as all three inactive compartments B1 (mean $\gamma = 0.746$), B2 (mean $\gamma = 0.768$), and B3 (mean $\gamma = 0.774$).

While values of γ were consistent between nuclear subcompartments, we note that domains in different nuclear subcompartments display different size distributions. Domains in A type compartments tend to be smaller (A1 median=140kb, mean = 190kb; A2 median=145kb, mean=200kb) than domains in B type compartments (B1 median=200kb, mean=265kb; B2 median=330kb, mean=425kb; B3 median=280kb, mean=360kb).

vi. Aggregate measures of contact probability consistently overestimate γ

We found that the existence of domains skews the value of γ when it is measured on aggregate contact probability over distances of several hundred kilobases to several megabases. Specifically, because contacts are enhanced between two loci in the same domain and depleted between loci in different domains, windows with many domains exhibit a steeper drop-off in contacts. To demonstrate this, we partitioned the genome into 567 windows of length 5Mb and measured the contact probability scaling at distances of 300Kb to 3Mb. When we plotted the values of γ obtained against the number of domains intersecting the window, a clear dependence emerged, where regions with many small domains exhibited larger values of γ (Fig S2B).

c. Fractal models are inconsistent with intra-domain contact probability

Interpreting the contact probability exponent γ requires a full understanding of which values of γ may be achieved by which structures. The γ value of the fractal globule can be studied in terms of mathematical fractal curves, which are scale-free and densely packed, like the fractal globule. Because they are computationally tractable, they are often used to approximate the fractal globule (S21). Approximate numerical studies of fractal curves have suggested that the fractal globule can exhibit values of γ ranging from 1 to 1.33 (5, 17). However, no rigorous bounds have been derived.

Here we prove mathematically that any fractal (regular, scale-free) structure must have γ between 1 and 2. These results are illustrated in Figure 3B and are derived in full mathematical rigor in Section IV. Notably, contact domains have exponent $\gamma \approx 0.75$, inconsistent with the fractal globule model.

Domain contact probability in Figure 3B was measured in a 50kb window through the domain center, as in Figure 2B. Each trace in Figure 3B was renormalized to allow direct comparison of contact probability for Hi-C domains and fractal curves. Linear distances (x-direction) were normalized to place the start and end of the scaling regime at the left and right boundaries of the plot: contact domain scalings ran from 10kb to half the domain size; fractal curve and fractal globule scalings ran through the whole length with one point excluded at each end to avoid boundary effects. Contact probabilities (y-direction) were renormalized so that all traces pass through a single point in the upper left.

d. Analysis of chromatin loop networks

i. Hubs are detected as (nearly) isolated cliques in the loop network

Chromatin hubs, collections of loci that simultaneously colocalize in the nucleus, are thought to play a crucial role in gene regulation and chromatin packaging. Chromatin hubs manifest clearly in Hi-C contact maps as collections of loci with loops forming between all pairs of loci (8).

We identified hundreds of chromatin hubs in GM12878 by converting Hi-C loop calls (as annotated by HiCCUPS (8)) into a "loop network" where each node in the network corresponds to a genomic locus and each edge in the network corresponds to a chromatin loop. Due to variability in HiCCUPS, which is performed at 5kb and 10kb resolution, we grouped any two loop anchors within 20kb into a single locus with a single corresponding node in the network. An edge is drawn between two loci if there exists a Hi-C loop between a loop anchor in each of the loci. Hubs are then detected in the network as isolated cliques.

ii. An extended list of loops accounts for false negative loops

The loops reported in (8) use conservative parameters to avoid false positive loops, and therefore may have some missing false negative loops. However, failing to identify any one of $N(N-1)/2$ loops will prevent a hub of size N from being identified. Thus, we allow for a margin when detecting hubs of size

4 or larger. Specifically, for a hub of size $N \geq 4$, we allowed up to $N-2$ to be called from an extended list of chromatin loops. This extended list was detected from the GM12878 Hi-C map using HiCCUPS (8) with much more relaxed thresholds. (Specifically, the allowed false discovery rate (FDR) in the Benjamini-Hochberg FDR control procedure was increased from 10% to 50% and thresholds for the lower left, vertical, and horizontal filters were reduced to 1. Loop calls from 5kb, 10kb, and 25kb resolution were combined.) Significantly decreasing threshold stringency would typically confound analysis by increasing the number of false positives; however, this effect is dramatically mitigated here since any loop in the extended list that is not formed between genomic loci in the original list may be discarded. In this way, the number of allowed loops is increased from 9,273 to 13,295, significantly reducing the false negative rate of the loop list.

Using this approach, we detected 69 isolated cliques of size 3, 16 cliques of size 4, and 1 clique of size 5. We found that allowing for a small number of loops ($\leq N-2$) between loci inside the clique and loci outside the clique allowed us to call more hubs without lowering quality. By doing so, we identified 145 cliques of size three, 86 cliques of size four, 5 cliques of size five, and 1 clique of size 6.

iii. Hubs often involve a series of consecutive loops

The detected hubs were often formed between consecutive loop anchors. Of the 237 detected hubs, 152 of them involved only consecutive loop anchors; i.e. all loop anchor loci between the first and last loop anchor of the hub belong to the hub itself.

iv. Hubs are consistent with looping between pairs of convergent CTCF motifs

Our previous study demonstrated that loops typically lie between convergently-oriented DNA motifs that bind a complex containing CTCF and cohesin (8). We find that the detected hubs are strongly consistent with this property. A list of the most probable CTCF binding sites was assembled by identifying the best match to the consensus CTCF binding motif (S5) within each peak in the ENCODE CTCF ChIP-Seq data from GM12878 and the orientation each site was determined by whether the binding motif was on the forward or reverse strand (8). We then matched each hub locus with any CTCF binding site that was within 15kb of any loop anchor in the locus. The first hub locus, which forms several downstream loops, was associated with a forward-oriented motif in 231 of 237 hubs. The last hub locus, which forms several upstream loops, was associated with a reverse-oriented motif in 232 of 237 hubs.

v. Middle loci of hubs show strong enrichment for pairs of divergent CTCF motifs

Of the 336 middle hub loci, 235 were associated with motifs of both orientations. When we examined middle loci that were associated with exactly one motif of each orientation, 59 of the pairs were in the divergent configuration (a reverse motif followed by a forward motif) while only 7 of the pairs were in the convergent configuration (89.4% versus 10.6%). We observed an even stronger effect when we generalized this measurement to middle loci with more than two motifs. In this analysis, a set of CTCF motifs at middle hub loci was counted as divergent when it contained one or more reverse motifs followed by one or more forward motifs. Similarly, a set of CTCF motifs at middle hub loci was counted as convergent when it contained one or more forward motifs followed by one or more reverse motifs. We found that 145 middle hub loci were associated with a divergent arrangement of CTCF motifs while only 12 were associated with a convergent arrangement of CTCF motifs (92.4% versus 7.6%).

vi. All observed effects are highly enriched over suitable null models

Hubs detected from the Hi-C loop list were significantly enriched relative to a randomly shuffled control. From the loop network and the extended loop network, we computed the distribution of loops $L(d)$ as a function of genomic distance d. From a given loop network, a shuffled network was generated with number of nodes and edges equal to the original network, but with randomly chosen edges. The probability of drawing edge (a, b) was proportional to $\deg(a)\cdot\deg(b)\cdot L(|a-b|)$. In this manner, we computed 10,000 shuffled networks of the original loop list. In addition, we computed 10,000 corresponding extended networks by shuffling the extended loop list and adding the appropriate number of new edges.

On each of 10,000 randomly shuffled networks with corresponding extended networks, we ran our hub detection algorithm. When cliques were required to be perfectly isolated, we detected an average of only 10.8 hubs of size three, 0.571 hubs of size four, and 0.008 hubs of size five. When cliques were allowed N-2 outside loops, we detected an average of only 42.5 hubs of size three, 6.95 hubs of size four, 0.362 hubs of size five, and 0.024 hubs of size six.

Enrichment of CTCF binding orientations was computed relative to the null model of a randomly chosen locus on the genome. That is, the percentage of hub loci within 15kb of a forward CTCF motif, a reverse CTCF motif, or CTCF motifs in both orientations was compared with the percentage of the same observations occurring for a randomly chosen locus in the genome.

vii. An "exclusion domain" often forms between two loci that loop to a common third locus

We observed in our genome editing experiments that deletion of the first or last locus (A or C) in a 3-clique (A, B, C) disrupted both the corresponding loops and the contact domains. However, deletion of one of the two anchor motifs at the middle locus (B) disrupted the corresponding loop but did not eliminate the domain spanned by the loop. This behavior is predicted by the extrusion simulations: for example, existence of the B-C loop excluded extrusion complexes between A and B from passing B. We dubbed this configuration an "exclusion domain".

Upon closer examination, we observed exclusion domains throughout the wild type genome. We examined pairs of loops (A, B) and (C, D) such that either the upstream loop anchors (A and C) or the downstream loop anchors (B and D) coincide (within 20kb). Assume for clarity that loci A and C coincide and B lies upstream of D; other situations are analogous. We filtered out cases that overlap with a compartment boundary (specifically, we require no compartment flip to occur between min(A, C) − 20kb and D + 20kb) in order to rule out the influence of compartments on domain formation.

We then examined the frequency of overlap of each of the three regions A-B, A-D, and B-D with the list of contact domains reported in (8); overlap was defined by a distance of $min(50\text{kb}, 0.2\times\text{domain size})$. These frequencies were compared with a control in which contact domains were reshuffled in the chromosome, averaged over 100 replicates. Since loops are formed between A-D and A-B, we observed domains in about 40% of cases, consistent with previous observations of loop domain formation. Specifically, in 986 cases, we observed 399 domains (40.5%) between A-B, a 14.4-fold enrichment, and 404 domains (41.0%) between A-D, a 12.4-fold enrich-

ment. Notably, we observed significant enrichment of contact domains formed between B-D despite the lack of a loop: a total of 158 cases (16.0%) or a 6.3-fold enrichment. Because there is no loop between B-D but B and D each loop to the same distal locus, these are exclusion domains.

To ensure that the detection of exclusion domains was not driven by false positive loops, we repeated the analysis and filtered out all cases in which a loop formed between B-D (within a 30kb threshold) in an expanded list of loops containing 99% more loops. These loops were detected in GM12878 using HiC-CUPS (8) but with relaxed thresholds (FDR rate of 10% and lower left, vertical, and horizontal thresholds of 1.3, 1.4, and 1.5). Enrichment of exclusion domains remained: out of 818 cases we observed 99 exclusion domains (12.1%), a 4.7-fold enrichment. In this case, we observed 333 (40.7%) domains at A-B (14.5-fold) and 329 (40.2%) domains at A-D (12.4-fold).

We also repeated the analysis and filtered for loops associated with CTCF motifs oriented appropriately for the exclusion effect. Specifically, when X and Y looped downstream to Z we required that X and Y were associated with (within a 20kb window of) forward CTCF motifs and Z was associated with a reverse CTCF motif; when X and Y looped upstream to Z we required that X and Y were associated with reverse CTCF motifs and Z was associated with a forward CTCF motif. Enrichment was similar. When false negative loops were simultaneously filtered, in 444 cases we observed 59 (13.3%) exclusion domains (5.1-fold enrichment), 189 (42.6%) domains between A-B (14.9-fold), and 190 (42.8%) domains between A-D (13.2-fold).

III. Polymer simulations

It has been known for decades that the higher-order folded structure of chromatin plays an important role in the biological function of the cell, particularly through partitioning into chromatin domains (S22) and loops (4, S23). However, the exact physical nature of this folding has been difficult to characterize because the principles governing this higher-order organization were entirely unclear. Possible mechanisms underlying chromatin compaction have included looping on a backbone (9, S24), looping through diffusion (S25, S26), compaction through confinement (5, 10), and supercoiling (S27). Other approaches use chromatin contact maps to define a population of physical models but do not reveal biological principles underlying the 3D organization (S28, S29). However, these models are appropriate only at scales larger than chromatin domains. Currently, no convincing models exist for chromatin at the scale of domains. Previous models have been limited by inexact experimental measurements of the folded state, typically measurements of distance distributions using 3D-FISH or coarse Hi-C maps.

Using our new kilobase-resolution Hi-C maps, we are able to comprehensively examine the folding within contact domains genome wide. Here we show using molecular dynamics simulations that the tension globule and the extrusion model are consistent with the folding of chromatin within these domains.

a. Molecular dynamics simulations

i. Polymer properties

Coarse-grained molecular dynamics simulations were used to investigate the properties of collapsing polymers. Chains of identical monomers were simulated under Brownian-like conditions using Langevin dynamics. Monomers were taken to have mass $m = 1$ and radius $\sigma = 1$. Each monomer represents 1kb of DNA. Polymers up to 10Mb (10,000 monomers) were simulated.

Successive monomers in the chain were held together by stiff harmonic bonds described by $U_{bond}(r) = (k_{bond}/2) \cdot (r - r_0)^2$ with resting length $r_0 = 1/\sqrt{2}$ (distance units) and $k_{bond} = 1000$ (energy units). All other pairs of atoms interact using pairwise forces, denoted by $U_{pair}(r)$, whose details depend upon the type of simulation being run (see below). Monomer volume exclusion was implemented by a repulsive Lennard-Jones potential with $\epsilon_{LJ} = 1$ (energy units) and $\sigma = 1$ (distance units).

No 3-body or 4-body bond-angle or torsion-angle forces were used. Due to repulsive interactions between the first and third of three consecutive monomers, angle between consecutive bonds was typically 90 degrees or larger.

Polymers were initialized as a random walk with step size $L_0 = 3 \cdot r_0$ and then statically minimized (with velocity zeroed at each step) to relax to regular bond lengths r_0. This resulted in more extended random initial states. Changing the expansion factor had negligible effect on the properties of the final collapsed states (Fig S9C).

ii. Simulation dynamics

The equations of motion were integrated using Langevin dynamics with a time step of 0.005 in units of τ_m, where $\tau_m = \sqrt{m\sigma^2/\epsilon}$. Parameters such as the temperature T (in units of ϵ/k_B) and the damping coefficient, $\zeta = 1/t_{damp}$ (t_{damp} measures the time over which momentum decays by a factor of 2, in units of τ_m) were varied to explore a wide variety of behavior. The number of time steps was chosen to ensure full collapse, and compactness of the final states was verified by measuring monomer density as a function of distance from the center of mass. All simulations were run using LAMMPS (23). Images of polymer collapse were rendered in Visual Molecular Dynamics (S30).

iii. Simulations of tension globules

In tension globule simulations and extrusion model simulations, all pairs of atoms experienced an attractive short-range force described by the Lennard-Jones 6-12 potential

$$U_{pair}(r) = 4\epsilon_{LJ}((\sigma/r)^{12} - (\sigma/r)^{6})$$

with $\epsilon_{LJ} = 1$ and $\sigma = 2^{-1/6}$ so that $U_{pair}(r)$ is minimized at $r = 1$ and cutoff distance of $r_{cut} = 2.5$ (distance units). This mimics the effect of immersing the polymer in a poor solvent (21, S31-32).

Tension globule simulations were characterized by moderate temperature and over-damping; i.e. an environment in which momentum does not significantly accumulate. The tension globules shown in Figure 4B-C were simulated with $N = 10000$ monomers, temperature $T = 1$, and damping coefficient, $\zeta = 1/t_{damp}$ with $t_{damp} = 35$ for 510,000 time steps (Sim #1 in Table S4).

We also simulated tension globules with pairwise monomer interactions modeled by a Yukawa potential instead of Lennard-Jones (Fig S10, Table S4C), a model of screened electrostatic interactions, similar to observed inter-nucleosomal forces (S33-36). The classic form of the Yukawa potential has energy proportional to e^{-mr}/r, where screening constant m determines how quickly the force decays. To interpolate between repulsive Lennard-Jones forces for $r < \sigma$ and Yukawa

forces for $r > \sigma$ we used the form

$$U_{pair}(r) = \begin{cases} \epsilon\left(\frac{b}{a-b}\left(\frac{\sigma}{r}\right)^a - \frac{a}{a-b}\left(\frac{\sigma}{r}\right)^b\right) & r < \sigma \\ \beta\epsilon\left(\frac{b}{a-b}\left(\frac{\sigma}{r}\right)^a - \alpha\frac{a}{a-b}\left(\frac{\sigma}{r}\right)^b e^{-m(r-\sigma)}\right) & r \geq \sigma \end{cases}$$

where $a = 12$, $b = 1$, and β and α are chosen to match the boundary condition at $r = \sigma$, namely, $\beta = (a-b)/(\alpha a - b)$ and $\alpha = 1/(1 + m\sigma/b)$.

iv. Simulations of fractal globules

Fractal globules were generated using molecular dynamics simulations by crushing a random polymer with global forces. In addition to repulsive Lennard-Jones, an external harmonic force was added to pull each monomer inward toward the origin $U_{ext}(r) = (k_{ext}/2) \cdot r^2$. These forces do not depend on the relative position between atoms, only their position relative to the origin. This is mathematically equivalent to adding a long-range attractive quadratic term $(k_{ext}/4N) \cdot r^2$ between all pairs of monomers. This compressive force mimics the effects of crowding experienced by chromosomes confined within the space of the cell nucleus. Simulations driven predominantly by external forces exhibited qualitatively different collapses as well as values of γ near 1.0 (Fig 4A, Fig S10).

v. Simulations of tension globule loop domains

Tension globule simulations incorporating loops used Lennard-Jones potential with parameters identical to simulation #1 in Table S4. A harmonic bond with strength $k_{loop} = 1000$ and resting length $r_0 = 1$ was introduced between several select pairs of monomers to form a loop at the beginning of simulation. In addition, a harmonic bond with strength $k_{tether} = 50$ and resting length r_{tether} (variable depending on polymer length, ranging from 50 to 300) was introduced between the first and last monomers in order to tether the start and end of the polymer apart. After the initial static minimization step, an additional equilibration step was run for 500,000 timesteps, incorporating only repulsive Lennard-Jones forces and temperature $T = 1$, to relax the looping contour into a natural state. Afterwards, standard tension globule collapse was simulated. Looping monomers for the simulations displayed in Figure 4D were chosen to match the positions of loops observed in Hi-C data. Simulations were run on a 3Mb polymer with $r_{tether} = 100$, and the contact map was computed by aggregating over 530 globules.

vi. Simulations of loop extrusion

In the extrusion model, an extrusion complex with two binding domains binds to two locations on the chromosome, and each domain walks in opposite directions along the chromatin fiber, pulling pairs of distal loci into close proximity. In simulation, we represent the extrusion complex by a harmonic bond with coefficient $k_{bond} = 10$ and length 1.0. The extrusion complex is initially bound to two consecutive monomers and sliding outwards every T_0 timesteps; i.e. the bond shifts from monomers (a, b) to monomers $(a-1, b+1)$. The specific value of T_0 does not have a substantial impact on the final result and is typically taken to be 200 or 500 timesteps. The sliding dynamics may be made stochastic by assuming a probability p of stepping outwards, a probability q of stepping inwards, and a probability $1-p-q$ of remaining in place, and the results will be essentially the same as long as p is substantially larger than q to drive the overall movement outwards. Similarly, the leftward and rightward movements may be decoupled with negligible effect on outcome. For simplicity, we assume the extrusion complex takes regular, equally sized steps outwards.

Extrusion complexes may interact with oriented CTCF motifs, represented as special monomers that cause the extrusion complex to halt and fix in position with some probability. These motifs are oriented: the downstream binding domain may be bind to and halt at a reverse-oriented motif with some probability but will be unaffected by a forward-oriented motif, while the upstream binding domain may be bind to and halt at a forward-oriented motif with some probability but will be unaffected by a reverse-oriented motif. As input to the extrusion simulation, a list of two binding strengths for each monomer (one for each orientation) is required; almost all the values will be 0. As we show in Section III.b.i, probabilities can be calculated in a simple, principled manner from CTCF ChIP-seq data of a particular region of the genome, and the resulting simulations recapitulate the contact map of that region very well.

The extrusion complexes are bound to the polymer at a density that depends on their concentration. Our extrusion simulations typically contained 2,000 to 3,000 monomers and 8 to 15 extrusion complexes. Extrusion complexes cannot pass each other; if two binding domains of two different extrusion complexes are adjacent on the polymer, one of the colliding complexes dissociates. A mobile complex may collide with a halted complex and dissociate, but a halted complex cannot dissociate in this manner (Fig S12C). In addition to dissociation due to collisions, the extrusion complexes can be made to dissociate at a fixed rate that depends on their processivity. At every extrusion step, each complex, including halted ones, has a halted probability $P_{dissociate}$ of randomly falling off. All simulations shown here have $P_{dissociate} = 0$, but allowing small values of $P_{dissociate}$ (e.g. < 0.002, or processivity > 500 steps) will not change results significantly. Simulations are run for a large number of extrusion steps (typically 4,000) and thus represent a steady state of extrusion dynamics.

Most extrusion simulation presented use similar conditions as the tension globule, with inter-monomeric attractive forces modeled by the Lennard-Jones potential, high viscosity, and moderate temperature. Results are robust to changes in the physics of the system (Fig S12A).

b. Simulated CTCF binding locations

i. Assigning extrusion complex binding strengths from CTCF ChIP-seq tracks

In order to model the 3D structure of a region in GM12878 using extrusion in a simple, principled way, we used the ENCODE CTCF ChIP-Seq data (S37) from GM12878 (tracks from the Broad Institute or Stanford University) to directly define the oriented binding strengths. First, a genomic region of length N kilobases was chosen to be modeled by a simulated polymer with length N, and the CTCF ChIP-seq signal s (binned at 1kb) was converted into a binding strength; i.e. a probability of binding defined by $P = 1 - \lambda/(s - s^*)$ for $s > s^*$ and $P = 0$ otherwise, where s^* is the median ChIP-seq signal in the region and λ is a normalization constant. Since ChIP-seq signal is variable between regions and the absolute value of the signal is not directly interpretable as a binding strength, the normalization constant λ is chosen to yield binding strengths in an appropriate range. Low values of λ cause extrusion to halt at a large number of weak ChIP-seq peaks, preventing proper extrusion coverage, while high values of λ causes extrusion to occur irrespective of ChIP-seq-defined boundaries, preventing the formation of distinct loops and domains. Second, a list of the most probable CTCF binding sites was assembled by identifying the best match to the

consensus CTCF binding motif (S5) within each peak in the CTCF ChIP-seq track, and the orientation of each site was determined by whether the binding motif was on the forward or reverse strand. Each non-zero halting probability was then oriented according to the orientation of the nearest CTCF binding site within 5kb.

ii. Assigning tension globule loops from CTCF ChIP-seq tracks

We also explored automatically assigning loops for tension globule simulations based on CTCF ChIP-seq signals. Oriented binding probabilities were computed as described above. A loop was formed between the two loci A and B (with $A < B$) proportional to the forward probability of A times the reverse probability of B times the probability A did not bind to any other locus between A and B. In simulation, 10 loops were chosen according to this distribution and the endpoints of the polymer were tethered 150 units apart (Fig S11E).

c. Analysis of the tension globule

i. Tension globule contact probability matches experimental observations

For each globule, two monomers were said to be in contact if they were separated by less than $\rho = 1.5$ (distance units). With this definition, we generated contact maps (Fig S8B) and computed contact probability (as described above) aggregated over many replicates of the same simulation. Figure 4C shows contact probability of 450 individual tension globules. While changing the value of ρ affected the total number of contacts measured per globule, it did not significantly affect the contact probability scaling exponent (Fig S9).

ii. Contact probability scalings are robust to changes in simulation parameters

We performed numerous additional simulations, each with at least 100 replicates, to verify that our observations were robust to changes in simulation parameters. Changes to the damping strength, temperature, initial extent, polymer length, and simulation time had no significant effect on γ (Fig S9, Table S4). Qualitatively, simulations with temperature $T < 0.7$ not collapse fully, settling into an extended ?stringy? state, and globules with $T \gg 2.0$ retained too much thermal energy to fully compact. After polymers were fully collapsed into a globular state, running the simulation for additional time did not significantly affect the value of γ.

Tension globule collapse due to Yukawa forces was qualitatively similar, displaying sub-globule formation and tension-driven linear compaction (Fig S10). Because the attractive forces decay less quickly with distance, the tension forces are stronger and values of γ ranged between 0.7 and 0.6. Exact value of γ as well as the distance range of the scaling depended on the screening strength parameter m. (Fig S10, Table S4).

iii. The tension globule exhibits a linear axis

In the later stages of collapse, tension-driven forces cause sub-globules to compact in an anisotropic manner, generating a visually apparent linear axis. Sub-regions of the tension globule tend to form flat slices perpendicular to the linear axis (Fig S8E). Strikingly, the axial and lateral views of the tension globule are similar to images of metaphase chromatin (S38).

We measured the predominance of the linear axis by computing the typical angle between displacement vectors of regions separated by genomic length L. Specifically, for each genomic length $L < N/3$, we subdivided the polymer into N/L regions of length L, and computed the center of mass of each region. Then, for any three consecutive regions with centroids A, B, and C respectively, we computed the angle between the vectors AB and BC, and averaged all such angles. This measurement distinguished the internal structure of the tension and fractal globule (Fig S8D). Fractal globules have displacement angles around 90 degrees or less for all values of L, while tension globules have obtuse displacement angles for large values of L.

iv. The tension globule is unknotted

To evaluate the topological state of the tension globule, we computed the determinant of the Alexander polynomial, a knot invariant. A value of 1 indicates an unknot; large values reflect high levels of knottedness. The Alexander polynomial determinant was computed in Knotplot using the "alex" command, with the two endpoints of the globule joined along the exterior to create a closed contour. Fig S8G shows the values obtained for 100 tension globules compared to 100 equilibrium globules (8), with each globule containing 4,000 monomers. Tension globules are comparatively highly unknotted.

v. Tension vs. fractal structure depends on ratio of internal and external forces

Simulations incorporating both internal inter-monomeric Lennard-Jones forces U_{pair} and external crowding forces U_{ext} showed a transition between fractal and tension globule behavior (Fig 4A). The relative strengths of these forces were varied by changing the values of ϵ_{LJ}, the Lennard-Jones coefficient, and k_{ext}, the external force coefficient. In order to renormalize the force strengths to have the same potential energy at unit distance, the ratio of internal to external force strengths was computed as $R = \epsilon_{LJ}/(k_{ext}/2)$. The values of ϵ_{LJ} and k_{ext} were each varied between 1 and 0.0001 in order to achieve values of R between 0.0002 and 2000. All simulations used 10Mb polymers and $t_{damp} = 30$, and the contact probability scaling was measured between 15kb and 1Mb.

Fractal globules are necessarily simulated in the absence of temperature because introduction of thermal fluctuations produces a length scale and prevents full compaction of the globule. On the other hand, low-temperature simulations with strong Lennard-Jones forces do not collapse fully and instead stagnate in an extended, stringy position. To accurately transition between the two regimes, simulation temperature was varied together with the Lennard-Jones coefficient as $T = min(1.0, 2\epsilon_{LJ})$. Exact parameter values and scaling exponents are listed in Table S4B.

The fractal-tension transition was also simulated using the Yukawa potential for intermonomeric forces (Fig S10). The crushing force was fixed in strength at $k_{ext} = 10$ while the Yukawa coefficient ϵ_{Yukawa} varied in strength from 0.001 to 1. Weakly-screened Yukawa forces ($m = 0.2$) were chosen. Simulations with weakly-screened Yukawa forces do not require thermal fluctuation for full collapse, so all simulations were performed with $T = 0$ and $t_{damp} = 10$. Exact parameter values are listed in Table S4C.

vi. Contact probability of simulated loop domains matches experimental observations

Loop domains of size 500kb and 1Mb were simulated as described in Section III.a.v. Simulations of 500kb loops used 10Mb polymers with 18 loops separated by 50kb of spacing. Simulations of 1Mb loops used 10Mb polymers with 9 loops separated by 50kb of spacing. Contact probability was aggregated over N = 50 globules and computed on 100kb windows running through the center of each domain. By running multiple replicates, contact probability scalings were computed for a total of 36 500kb loop domains and 27 1Mb loop domains

(Fig S11B). Mean γ value of 0.768 (standard deviation 0.077) was in excellent agreement with intra-domain values of γ measured from Hi-C data.

d. Analysis of the extrusion model

i. Contact probability of extruded domains matches experimental observations

To determine the contact probability exponent of the extrusion model, we simulated 3Mb polymers each containing three loop domains of size 980kb separated by 20kb. These domains were formed using binding strengths of 0.9 and placed at 10kb/1010kb/2010kb (forward) and 990kb/1990kb/2990kb (reverse). A total of 140 replicates were run for 4,000 extrusion steps and contacts were measured after 2,000 and 4,000 steps. Every set of 20 replicates was aggregated, contact probability was measured through a 100kb locus at the center of each of the three domains (Fig. 5C), and values of γ were found to be 0.72 ± 0.06 between distances of 20kb and 400kb (the edge of the domain) for these 42 simulated extrusion domains, in excellent agreement with intra-domain values of γ measured from Hi-C data.

While a large proportion of Hi-C domains are associated with loops at their corners, domains may appear in many contexts; e.g. without associated loops or at compartment flips (8). The values of γ remain similarly around 0.75 across the wide variety of domains observed. Note that a contact probability scaling of $\gamma \approx 0.75$ for loop-less domains is also well-explained by the extrusion model. Indeed, loop extrusion simulations will exhibit scalings around 0.75 even in the absence of CTCF-binding motifs that cause loops to form.

ii. The extrusion model accurately recapitulates Hi-C contact maps, including loops and domains

We computed binding strengths from CTCF ChIP-seq tracks of four regions of the genome from GM12878 cells and simulated those regions polymers undergoing loop extrusion. Specifically, we simulated chr 4: 20.3-22.6Mb; chr 3: 62-64Mb; chr 5: 110.45-111.75Mb; chr 7: 15-17.5Mb. Simulations were run with temperature $T = 2.0$ and $t_{damp} = 10$ for 200,000 time steps to collapse the polymer from an extended state and then 4,000 extrusion steps (each 200 time steps). At least 100 replicates were run for each region. Domains and loops are evident in the contact map after around 500 extrusion steps. Contact maps were aggregated for simulations at 1,000, 2,000, 3,000, and 4,000 extrusion steps. (Substantial reorganization occurs due to continuous extrusion, so the structures at these times are not similar.) In all cases, the contact maps of the simulated polymers accurately recapitulated the contact maps observed from Hi-C (Fig 5D, Fig S12D).

iii. The extrusion model accurately predicts changes in 3D structure after CRISPR editing

The extrusion model accurately recapitulates Hi-C contact maps by assigning extrusion complex binding strengths from CTCF ChIP-seq data with a single normalization constant. From this wild-type model, we were able to generate de novo predictions for the Hi-C contact map after CTCF motifs have been deleted or reversed using genome-editing. Focusing on three regions containing triples of loci which form three loops, we generated predictions for 13 different deletions and inversions of the CTCF at the loci (Table S4E). In every single case, the extrusion model correctly predicted the positioning of loops and domains in the corresponding genome-editing experiments (Fig 7, Fig S14-16, Section I.d). Experiments were performed in Hap1 cells for which high-quality CTCF ChIP-seq tracks were not available, so wild-type simulations used GM12878 ChIP-seq data.

All simulation parameters were tuned to the wild-type maps, and predictions were made simply by altering the extrusion complex binding strengths in a manner similar to the CRISPR experiment. Deletion of a CTCF motif was simulated by altering the binding strength at that location to 0. Inversion of a CTCF motif was simulated by flipping the orientation of the motif at that location.

Note that on chromosome 1, the GM12878 ChIP-seq data showed CTCF binding at two locations near 181.1Mb that are not evident in the Hap1 wild-type contact maps, so the binding strengths of these two loci were dampened in our simulations of the wild-type recapitulation and all predictions for that region (Fig 7B). It is possible that this CTCF signal corresponds to a subdivision of the domain between E and F into two subdomains; however, this is not entirely resolvable by our current Hi-C maps. Because this procedure was followed identically for simulations of wild-type and edited conditions, it does not affect the validity of the *in silico* predictions.

iv. These properties are robust to changes in the attractive forces and the initial configuration

We find that the state achieved in the extrusion model predominantly results from the extrusion dynamics and not from the exact physics of the system. Changes in viscosity and temperature did not affect results (Fig S12A). Notably, loop extrusion produced very similar contact maps (Fig S12A) and contact probability scaling (Fig S12B) when performed using inter-monomeric attractive forces or using a global attractive potential. (Extrusion alone is not sufficient, however, as some attractive forces are required for collapse.) Moreover, because the extrusion simulations represent a steady state, the results described are identical when the simulation is started from an extended chain or a collapsed state, such as a tension globule or a fractal globule (Fig S12A).

v. The extrusion model predicts that loops will be intra-chromosomal and tend to be short

The extrusion model makes particular predictions about the distribution and locations of loops. First, the sliding of the extrusion complex necessitates that the resulting loops are formed between two loci on the same chromosome. This is in contrast to a model in which loops are formed when anchors come into close contact through random diffusion, which would allow for inter-chromosomal looping. Indeed, in our Hi-C maps we have observed no loops occurring between two distinct chromosomes (8).

Second, in the extrusion model, genomic distances at which loops are formed is limited by the processivity of the extrusion complex. Thus, although loops could hypothetically form between any two loci on the same chromosome, the extrusion model predicts that the loops formed will tend to be relatively short. Indeed, median length of loops observed in our GM12878 Hi-C map was around 275kb (8).

vi. Loop extrusion promotes an unentangled, unknotted topology

Classically, loops of DNA were thought to occur through a diffusive process, much like numerous other mechanisms in the cell, that randomly brought the loop anchors into close proximity in 3D. However, uncontrolled diffusion and subsequent looping would drive the chromatin fiber towards a highly entangled state. In contrast, loop extrusion promotes a disentangled state by forming a small loop and growing the size of the loop locally as the extrusion complex processes across the DNA (Fig S12F). We observe that two well-mixed domains in

simulation tend to un-mix when extrusion complexes repeatedly act upon the polymer. Loop extrusion in the cell may play a crucial role in keeping the chromatin fiber unentangled and locally accessible.

Loop extrusion could also play a role in controlling or reducing the knottedness of the chromatin fiber. Note that if the actual loading of the extrusion complex creates a knot (e.g. a trefoil knot), the knot would persist, but not get complexified, during the subsequent extrusion. In general, sliding of the extrusion will not change the knottedness, in contrast to diffusion-based looping, and would instead slide knots along the chromatin fiber. In fact, this suggests a potential mechanism for unknotting DNA through synergistic action of the extrusion complex and topoisomerase II. As loop extrusion slides and tightens existing knots, purely random passage of double-stranded DNA by topoisomerase II would drive the fiber towards an unknotted state.

e. The tension globule and the extrusion model are consistent with 3D DNA FISH measurements

Physical 3D distances between 30kb-wide loci were measured experimentally using three-color 3D DNA FISH as described in Section I.c. Three different regions were measured in IMR90, yielding 3D distance distributions between 5 pairs of loci at 4 genomic distances: 320kb, 490kb, 545kb, and 1,090kb. All pairs of loci lie within a single contact domain. Locations of loci are listed in Table S5.

Tension globule distance distributions were measured from simulation #1 (Table S4A) while extrusion model distance distributions were measured from simulation #79 (Table S4E). For each genomic distance, several pairs of monomers with the appropriate genomic separation were chosen randomly from each simulation and the 3D distances were measured.

The diameter of a single monomer was chosen to represent a length of 0.65µm to produce the best fit between experimental and simulated distance histograms. The two-sample Kolmogorov-Smirnov (K-S) test was used to determine how well each pair of distributions match. The mean K-S test statistic was 0.15 for the tension globule vs. experimental data (Fig S13B), 0.19 for the extrusion model vs. experimental data (Fig S13C), and 0.18 for the two different FISH pairs at the same genomic distance.

f. SMC3 and RAD21 are positioned on loop anchors 20bp towards the loop interior

A list of the most probable CTCF binding sites was assembled as described in Section III.b.i. From this list, we identified CTCF sites that are the unique CTCF motif within the loop anchor (8), and partitioned these loop anchor CTCF sites into two groups corresponding to the upstream or the downstream loop anchors. Next, we plotted ChIP-seq signals for SMC3 (SYDH TFBS) and RAD21 (two replicates from HAIB TFBS and SYDH TFBS) relative to the CTCF binding sites, aggregated over all sites in each group (S37). As previously demonstrated, there is a strong enhancement of interactions of both proteins near CTCF binding sites. However, we also observed an additional bias of the interaction position: interaction of SMC3 or RAD21 is shifted 20bp towards the interior of the loop (Fig S18). When averaged over both loop anchors or over all CTCF motifs, SMC3 or RAD21 interactions show no such positional bias. Interactions of CTCF protein at CTCF sites (UTA TFBS) show a slight shift towards the exterior of the loop.

g. Monte Carlo simulations of large fractal globules using Confined-BFACF

Studying densely packed polymers is an area of interest as it related to many important biological entities, for example DNA in a nucleus. One common way to study densely packed polymers is to model the polymers using molecular dynamics simulations. However, these simulations can get cumbersome - the computational complexity is often $O(N^2)$, making studying long chains (with $N = 104$ to 106) in three spatial dimensions rather difficult.

An alternative approach is to model polymers as self-avoiding walks on a lattice. While real polymers are obviously not constrained such that their monomers lie on a lattice, nevertheless it is thought that these lattice walks share many properties with real polymers. We further restrict our exploration here to self-avoiding lattice polygons, i.e. polymers whose ends are connected, so that the topology of the polymers may be specified.

To generate these polymers, we use a novel algorithm called Confined-BFACF, which we adapted from the previously described Markov Chain Monte Carlo algorithm BFACF (S39, S40). Confined-BFACF was used to generate the fractal globule shown in Figure 3B. In this algorithm, each iteration of BFACF accepts a lattice self-avoiding polygon (SAP) as input and outputs a SAP. At every iteration, the algorithm picks an edge of the current SAP at random and proposes to move that edge by one unit in one of the 4 directions perpendicular to the edge. To keep this edge connected with the rest of the polymer, the vertices connected to the edge may then need to be deleted, or new ones will need to be added, depending on the result of the translation. Edge translations fall into three categories: negative moves, where two vertices must be deleted; positive moves, where two vertices must be added; and neutral moves, where the net change in number of vertices is zero. The probability with which the algorithm chooses a direction to move the edge in is pre-computed by first determining which of the three categories such a move would result in (positive, negative, or neutral), and weighing the direction probabilities accordingly. The parameter z determines the relative probability of the algorithm choosing positive, negative, or neutral moves: $p(\text{positive}) = z^2/(1+3z^2)$, $p(\text{negative}) = 1/(1+3z^2)$, and $p(\text{neutral}) = (1+z^2)/(2+6z^2)$ (a fifth option of not moving the edge at all is given such that the probabilities add to 1).

Once a move is proposed, the algorithm checks to see if the move violates self-avoidance, in which case the move is rejected and the previous SAP is retained for that iteration. When running Confined BFACF, the SAP is initially placed in the center of a containing volume. At each iteration, the algorithm checks if the proposed move will cause the SAP to go outside the containing volume, in which case the move is rejected.

To create large unknotted spherical globules, a small unknot (typically 6 monomers) is placed in the center of a confining spherical volume. The algorithm is run with a bias towards growing, typically with $z = 5$, and the algorithm is run until the polymer has filled up the confining volume. The BFACF algorithm preserves the topology of the polygon (S41), so the final configuration of the polymer, though contorted and random, is still topologically unknotted.

IV. Mathematical theory

a. Introduction

Loosely speaking, a *fractal* is a mathematical object which can be broken down into parts with each part similar to the original whole. A fractal exhibits fine structure at arbitrarily small scales as a result of its self-similarity. Beyond their fascinating and beautiful mathematical properties, fractals have found numerous applications in modeling natural phenomena and data compression.

Many fractals are constructed through an iterative process. This construction may be understood geometrically, where a finite path is chosen to form a *base motif*, a *rewrite rule* is defined which positions multiple shrunken copies of the base motif and connects them into a more complex path, and this process is iterated ad infinitum. For example, this iterative construction of the Dragon curve is shown in Figure 3A. Section 5 discusses these constructions in a manner that avoids mathematical abstraction and may be instructive for the general reader.

In Section 2, we present a general mathematical construction of self-similar curves. For example, the same Dragon curve can be defined via the iterative action of two contracting similarities mapping $\mathbb{R}^2 \to \mathbb{R}^2$:

$$\begin{aligned}\Phi_0(x,y) &= \rho_{-\pi/4}(x/2, y/2),\\ \Phi_1(x,y) &= \rho_{\pi/4}(x/2, y/2) + (\sqrt{2}/2, \sqrt{2}/2),\end{aligned}$$

where $\rho_\theta : \mathbb{R}^2 \to \mathbb{R}^2$ is a counterclockwise rotation of angle θ around the origin. Beginning with the compact set $A_0 = \{0\} \times [0,1] \subset \mathbb{R}^2$, we recursively define

$$A_k = \Phi_0(A_{k-1}) \cup \Phi_1(A_{k-1}).$$

The first few iterations exhibit increasingly rough behavior (Fig 3A). The sets $\{A_k\}_{k=0}^{\infty}$ converge to a compact set $\mathcal{A}$, a *self-similar set*. For each k, A_k is parameterizable by a continuous curve $f_k : I \to A_k$, where $I := [0,1]$, and the limit set $\mathcal{A}$ is traced out by the limit function $f = \lim f_k$. This continuous map $f : I \to \mathcal{A}$ is an example of a *self-similar curve*.

In fact, the Dragon curve is also a *space-filling curve* since the path it traces forms a two-dimensional set. This iterative method was used by Giuseppe Peano in 1890 to create the first space-filling curve. The *Peano curve* traces out a two-dimenisonal square, defying conventional wisdom at the time which held that higher-dimensional sets were "too large" to be covered by lower-dimensional sets by a continuous mapping.

It is useful to extend the notion of *dimension* to non-integral values in order to understand and compare complicated fractal sets. For example, the boundary of the Dragon curve (Fig 3A) has zero two-dimensional area, yet it is so finely folded that it has infinite one-dimensional length. Thus, in some sense it is larger than 1D but smaller than 2D.

Based on this intuition, there are two common ways to measure non-integral dimensions: the *box-counting dimension* (also known as the *Minkowski dimension*) and the *Hausdorff dimension*. These measures are consistent with our usual notion of integral dimension – that is, $\mathbb{R}^n$ has box-counting and Hausdorff dimension both equal to n – and can be used to compare the "sizes" of many different fractal sets.

The box-counting dimension is more commonly used in applied settings because it can be easily measured on data. The general notion is to perform some measurement $N_\delta(X)$ on X "at the scale δ" and see how the measurement varies as δ grows small. If the measurement exhibits a power law $N_\delta(X) \sim \delta^{-d}$, then the box-counting dimension is defined to be d. Typically, $N_\delta(X)$ counts the number of boxes in an overlaid grid with side length δ that intersect the set X.

The Hausdorff dimension similarly measures the size of a set at infinitesimally small scales. However, it is more sensitive, using infinite covers of sets with diameter at most δ, making the Hausdorff dimension more difficult to empirically calculate and cleaner to manipulate theoretically. First, the s-dimensional Hausdorff measure of a set X, denoted $\mathcal{H}^s(X)$, is defined as

$$\lim_{\delta \to 0} \inf \left\{ \sum_{i=1}^{\infty} \text{Diameter}(U_i)^s : \{U_i\} \text{ is a } \delta\text{-cover of } X \right\}.$$

Then there is some critical value s_0 for which $\mathcal{H}^s(X) = \infty$ when $s < s_0$ and $\mathcal{H}^s(X) = 0$ when $s > s_0$. The value s_0 is defined to be the *Hausdorff dimension* of X.

The box-counting and Hausdorff dimensions are often equal for many tractable sets, including any self-similar set. An rigorous overview of the properties of the box-counting and Hausdorff dimensions can be found in a number of standard texts (S43, S44).

After understanding the dimension of the image of a self-similar curve, it is natural to ask next about dimensions of subsets of the curve. For example, what is the dimension of the set of points in $[0,1]$ which map to a given line intersecting the Dragon curve? (Fig 3A) Yet, in the 125-year history of self-similar curves, it is not known how they transform dimensions of sets other than the entire unit interval. In general, dimension behaves unpredictably for arbitrary smooth maps.

Transformation of Hausdorff dimension has been explored greatly in the field of stochastic processes. In 1955, McKean proved a dimension scaling result for two-dimensional Brownian motion, showing that if B is a Brownian motion path in $\mathbb{R}^2$ and $X \subset (0,\infty)$, then

$$\dim_H B(X) = 2 \cdot \dim_H X$$

almost always (13).

In Section 3 we prove an original theorem, a deterministic analog of McKean's theorem, showing that self-similar curves scale the dimension of any subset of I. That is, if the self-similar curve $f : I \to \mathbb{R}^n$ has d-dimensional image, then

$$\dim f(X) = d \cdot \dim X$$

for any $X \subset I$, where dim denotes either Hausdorff or box-counting dimension.

Measurements of a contact probability power law $I(s) \sim s^{-\gamma}$ have emerged as a crucial tool for linking measurements of spatial contacts of nuclear DNA to models of its folded structure. The fractal globule model has been shown to exhibit values of γ between 1 and 1.2, but it is unclear what range of exponents can be achieved by general fractal models of folding. By computing contacts of finite self-similar paths, we apply our dimension scaling result to characterize contact probability scalings of a wide range of fractal structures (Fig S3, Table S3).

In Section 4, we derive an explicit formula for γ for a self-similar curve f. First, we show that the contact probability exponent effectively measures the box-counting dimension of the contact map. Next, we apply our dimension scaling result to compute the dimension of the contact map in terms of the packing density $d_i = \dim_B \text{Image}(f)$ and the surface roughness $d_e = \dim_B \text{Exterior}(\text{Image}(f))$. Specifically, we show that finite self-similar paths converging to f will have contact probability exponent $\gamma = 2 - d_e/d_i$. Since $0 \le d_e < d_i$, any fractal model will have $1 < \gamma \le 2$ and thus cannot explain the exponents of $\gamma = 0.75$ observed for Hi-C contact domains.

The boundary case of $\gamma = 2$ can be achieved when $d_e = 0$, such as for the Sierpinski arrowhead curve (Fig S3). Values

of γ arbitrarily close to 1 are only achieved when the exterior dimension nearly matches interior dimension. In Section 5, we construct an original family of self-similar space-filling curves, dubbed "Inside-Out Hilbert curves", with arbitrarily rough boundaries and γ arbitrarily close to 1, the first curves of this kind.

b. Construction of self-similar curves

A *self-similar curve* is a curve which can be decomposed into parts each directly similar to the whole. Classically, self-similar curves are constructed by choosing a short lattice path to serve as a base motif and iteratively constructing exponentially longer paths. The self-similar curve is the mathematical limit of this process. Several examples are shown in Fig. S3. Here we outline a formal definition of self-similar curves in terms of iterated function systems.

Let X be a closed subset of $\mathbb{R}^n$ and let $|x|$ denote the vector length of $x \in X$. The map $\Phi : X \to X$ is called a *contracting similarity* if there is a constant c with $0 < c < 1$ such that $|\Phi(x) - \Phi(y)| = c \cdot |x - y|$ for all $x, y \in X$.

Definition 1. *Given $X \subset \mathbb{R}^n$ and $N \geq 2$, a family of contracting similarities $\{\Phi_0, \cdots, \Phi_{N-1}\}$ with $\Phi_i : X \to X$ for all i is called an iterated function system on X, abbreviated as IFS. An attractor of an IFS is a non-empty compact set $\mathcal{A} \subset X$ that satisfies*

$$\mathcal{A} = \bigcup_{i=0}^{N-1} \Phi_i(\mathcal{A}). \qquad [1]$$

In fact, every IFS uniquely defines an attractor $\mathcal{A}$. If $A_0 \subset X$ is any non-empty compact set and $A_{k+1} = \cup_{i=0}^{N-1} \Phi_i(A_k)$ for each $k \in \mathbb{N}$, then the sequence $(A_k)_{k\in\mathbb{N}}$ converges to $\mathcal{A}$ (S44, S45).

An IFS describes the self-similar structure of its unique attractor set $\mathcal{A}$. Here we define *self-similar curves*, a continuous surjective function $f : [0, 1] \to \mathcal{A}$ that respects the self-similarity imposed by the IFS. To construct a self-similar curve on $\mathcal{A}$, we first define a notion of indexing the self-similar set $\mathcal{A}$.

Let Ω_N denote the collection of all finite-length sequences of integers in $\{0, \cdots, N-1\}$; Ω_N^k denote the collection of sequences in Ω_N with length k; Ω_N^∞ denote the collection of all infinite sequences of integers in $\{0, \cdots, N-1\}$. Given $\omega = (i_1, i_2, \cdots) \in \Omega_N^\infty$, one can show that there is a unique point x_ω such that

$$\bigcap_{k=1}^{\infty} \Phi_{i_1} \circ \cdots \circ \Phi_{i_k}(\mathcal{A}) = \{x_\omega\}.$$

Define the *indexing function* $\rho_\mathcal{F} : \Omega_N^\infty \to \mathcal{A}$ that maps the *index* $\omega \in \Omega_N^\infty$ to the associated point $x_\omega \in \mathcal{A}$. Notice that

$$\Phi_j \circ \rho_\mathcal{F}(i_1, i_2, \cdots) = \rho_\mathcal{F}(j, i_1, i_2, \cdots) \qquad [2]$$

for any $(i_1, i_2, \cdots) \in \Omega_N^\infty$ and $j \in \{0, 1, \cdots, N-1\}$.

Construction 2. (Self-Similar Curve) Let $\mathcal{A}$ be the attractor of the IFS $\mathcal{F} = \{\Phi_0, \cdots, \Phi_{N-1}\}$ on $\mathbb{R}^n$. Subdivide the unit interval $I = [0, 1]$ into N parts given by subdivisions $0 = x_0 < x_1 < \cdots < x_N = 1$. Construct the IFS on I given by $S = \{\sigma_0, \cdots, \sigma_{N-1}\}$ where $\sigma_k(t) = (x_{k+1} - x_k)t + x_k$; that is, σ_k maps I to $[x_k, x_{k+1}]$. Consequently, I is the attractor of S.

Define a function $f_{S,\mathcal{F}} : I \to \mathcal{A}$ by requiring that $f(\rho_S(\omega)) = \rho_\mathcal{F}(\omega)$ for all $\omega \in \Omega_N^\infty$. With this definition, the value of $f_{S,\mathcal{F}}$ on $\{x_k\} = \sigma_{k-1}(I) \cap \sigma_k(I)$ is ambiguous. It is necessary to ensure that the "ends are connected." That is, write $a_0 := \rho_\mathcal{F}((0, 0, \cdots)) = \Phi_0(a_0)$ and $a_{N-1} := \rho_\mathcal{F}((N-1, N-1, \cdots)) = \Phi_{N-1}(a_{N-1})$, and then require $\Phi_k(a_{N-1}) = \Phi_{k+1}(a_0)$ for $k = 0, \cdots, N-2$. One can verify that $f_{(S,\mathcal{F})}$ is then well-defined and automatically continuous.

The function $f_{S,\mathcal{F}}$ constructed in this manner is said to be the *self-similar curve described by* $(S, \mathcal{F})$. This construction captures the notion of self-similarity through the iterated function systems since $f_{S,\mathcal{F}}$ restricted to $\sigma_k(I)$ behaves as $f_{S,\mathcal{F}}$ on I with $f_{S,\mathcal{F}} \circ \sigma_k \equiv \Phi_k \circ f_{S,\mathcal{F}}$. See Fig S4A for an example of an IFS that describes the Hilbert curve.

In many cases, there is a simple formula which computes the dimension of a self-similar set from the contracting ratios of the associated IFS. First, one must verify that the sets $\Phi_i(\mathcal{A})$ do not overlap significantly:

Definition 3. *The IFS $\{\Phi_0, \cdots, \Phi_{N-1}\}$ on $\mathbb{R}^n$ satisfies the open set condition if there exists a bounded, non-empty open set $\mathcal{U} \subset \mathbb{R}^n$ such that*

I. $\Phi_i(\mathcal{U}) \subset \mathcal{U}$ for all $i = 0, \cdots, N-1$;
II. $\Phi_i(\mathcal{U}) \cap \Phi_j(\mathcal{U}) = \emptyset$ when $i \neq j$.

Write the Hausdorff dimension of $X \subset \mathbb{R}^n$ as $\dim_H X$ and the Minkowski or "box-counting" dimension as $\dim_B X$. Write the s-dimensional Hausdorff measure of X as $\mathcal{H}^s(H)$. The following simple formula can be used to compute the dimension of the attractor set (S43):

Theorem 4. *Let $\mathcal{F} = \{\Phi_0, \cdots, \Phi_{N-1}\}$ be an IFS of contracting similarities on $\mathbb{R}^n$ that satisfies the open set condition. Suppose Φ_i has similarity ratio $R_i < 1$ and $\mathcal{A}$ is the attractor of $\mathcal{F}$. Let $s > 0$ be the unique solution to*

$$\sum_{i=0}^{N-1} R_i^s = 1. \qquad [3]$$

Then $\dim_H \mathcal{A} = \dim_B \mathcal{A} = s$ and $0 < \mathcal{H}^s(X) < \infty$.

c. Dimension scaling theorem for self-similar curves

i. Overview

We investigate how a fractal curve f transforms the dimension of subsets of the unit interval. It is known how to compute the dimension of the image of f using methods such as Theorem 4. We prove a formula relating the dimension of $f(X)$ to the dimension of X for a general set $X \subset [0, 1]$.

Hausdorff and box-counting dimensions often behave poorly under transformations – in general, dimension is not preserved under a homeomorphism. However, transformation of Hausdorff dimension has been explored greatly in the field of stochastic processes. In 1955, McKean proved a dimension scaling result for two-dimensional Brownian motion – if B is a Brownian path in $\mathbb{R}^2$ and $X \subset (0, \infty)$, then $\dim_H B(X) = 2 \cdot \dim_H(X)$ almost always (13). Since the image of the full Brownian path has dimension 2 almost always, McKean's result can be restated as

$$\dim_H B(X) = \dim_H \text{Image}(B) \cdot \dim_H X. \qquad [4]$$

In 1961, Blumenthal and Getoor conjectured that equation [4] holds when B is a Lévy processes, a generalization of Brownian motion (S46). While this conjecture turns out to be false in general, numerous researchers have since computed $\dim_H B(X)$ under various conditions on X, B, or both. In 2005, Khoshnevisan and Xiao showed a general formula for $\dim_H B(X)$ in terms of X (S47, Corollary 2.6).

In fact, a general class of deterministic fractal curves do transform dimensions in a uniform manner. In this section, we give a proof of this original result. The fractal curves under consideration are defined as follows:

Definition 5. *Let $f : I \to \mathbb{R}^n$ be a fractal curve described by the IFSs $S = \{\sigma_0, \cdots, \sigma_{N-1}\}$ on I and $\mathcal{F} = \{\Phi_0, \cdots, \Phi_{N-1}\}$ on $\mathbb{R}^n$ as in Construction 2. Let r_i be the scaling ratio of σ_i and let R_i be the scaling ratio of Φ_i, for $i = 0, \cdots, N-1$. We say that the fractal curve f is balanced if $\mathcal{F}$ satisfies the open set condition and there is some constant β such that*

$$r_i = R_i^\beta \quad \text{for } i = 0, \cdots, N-1. \qquad [5]$$

The uniform dimension scaling of fractal curves is characterized by the following theorem, a deterministic analogue to McKean's Theorem. The result applies to any subset of the unit interval that is a Borel set, a very general class that includes any sets that can be formed from open sets using the operations of countable union, countable intersection, or complementation.

Theorem 6. *Suppose that $f : I \to \mathbb{R}^n$ is a balanced fractal curve with image $\mathcal{A} = f(I)$, and let $d = \dim_H \mathcal{A} = \dim_B \mathcal{A}$. Then for all Borel sets $X \subset I$,*

$$\dim_H f(X) = d \cdot \dim_H X \quad \text{and} \quad \dim_B f(X) = d \cdot \dim_B X.$$

The proof is given below in two parts, separately proving an upper bound and a lower bound on the dimension of $f(X)$. We first prove an upper bound (Corollary 9) by showing that f is α-Hölder with $\alpha = 1/d$ and $d = \dim f(I)$. This arises naturally because the self-similar structure of f guarantees that small distances in the domain are mapped to correspondingly small distances in the range. Next, we prove a lower bound (Lemma 14) by using Frostman's Lemma to pick a subset $X_0 \subset X$ for which small balls in X_0 have small Hausdorff measure relative to their radius. We then show that the push-forward by f of the Hausdorff measure on X_0 gives a measure μ on $f(X_0)$ that satisfies the mass principle $\mu(U) \leq c|U|^s$ for all sufficiently small U and a particular s. This gives a lower bound on $\dim_H f(X_0)$ and thus a lower bound on $\dim_H f(X)$. The lower bound on $\dim_B f(X)$ follows similarly.

The condition that f is balanced is necessary for the uniform dimension scaling result to hold. Given distinct $\alpha, \beta < 1/2$, define $f_{\alpha,\beta} : I \to I$ as the fractal curve described by $(\mathcal{F}_\alpha, \mathcal{F}_\beta)$ as in Construction 2, where

$$\mathcal{F}_a := \{t \mapsto at, \;\; t \mapsto (1-2a)t + a, \;\; t \mapsto 1 - a(1-t)\}$$

for a $= \alpha, \beta$. If f is not required to satisfy equation [5], precomposing with $f_{\alpha,\beta}$ (Fig S4B) would change the way the function transforms dimension of specific subsets of I when $\alpha \neq \beta$.

In practice, one often wants to take a self-similar set $\mathcal{A}$, the attractor of some IFS $\mathcal{F}$, and trace it entirely as the image of a fractal curve. In this instance, the values of the R_i are given and one can define an IFS S on I with the appropriate r_i so that the fractal curve described by S and $\mathcal{F}$ is balanced. In any motif-based construction, the scaling ratios R_i are all equal and one can simply choose $r_i = 1/N$ for each i.

ii. An upper bound on $\dim f(X)$

A function $f : \mathbb{R}^m \to \mathbb{R}^n$ is *α-Hölder* if there exists some $c > 0$ such that $|f(x) - f(y)| \leq c|x-y|^\alpha$ for all $x, y \in \mathbb{R}^m$. This uniform upper bound on the way that f transforms distances gives an upper bound on the way it transforms dimensions (S43).

Lemma 7. *Suppose that $f : A \subset \mathbb{R}^m \to \mathbb{R}^n$ is α-Hölder. Then for any $X \subset A$,*

$$\begin{aligned} \dim_H f(X) &\leq \alpha^{-1} \cdot \dim_H X \\ \dim_B f(X) &\leq \alpha^{-1} \cdot \dim_B X. \end{aligned}$$

Given a sequence $\omega = (i_1, \cdots, i_k) \in \Omega_N^k$, define the shorthand notation $\Phi_\omega = \Phi_{i_1} \circ \cdots \circ \Phi_{i_k}$. Balanced fractal curves are naturally α-Hölder because the relation $f(\sigma_\omega(I)) = \Phi_\omega(\mathcal{A})$ for $\omega \in \Omega_N$ allows small distances in the domain of f to govern small distances in the range of f.

Lemma 8. *Suppose that $f : I \to \mathbb{R}^n$ is a balanced fractal curve with image $\mathcal{A}$, and let $d = \dim_H \mathcal{A} = \dim_B \mathcal{A}$. Then f is $1/d$-Hölder.*

Proof. Let f be a fractal curve described by the IFSs $\{\sigma_0, \cdots, \sigma_{N-1}\}$ on I and $\mathcal{F} = \{\Phi_0, \cdots, \Phi_{N-1}\}$ on $\mathbb{R}^n$. Let r_i and R_i be the scaling ratios of σ_i and Φ_i respectively, for $i = 0, \cdots, N-1$. Since f is balanced, let d' be the constant such that $r_i = R_i^{d'}$ for $i = 0, \cdots, N-1$. Then $\sum_{i=0}^{N-1} R_i^{d'} = 1$. Hence, by Theorem 4, $d' = \dim_H \mathcal{A} = \dim_B \mathcal{A} = d$.

We now show that f satisfies

$$|f(s) - f(t)| \leq c|s-t|^{1/d} \text{ for all } s, t \in [0,1]$$

for some constant c. Let $r = \min r_i$. Since $|f(s) - f(t)| \leq |\mathcal{A}|$ for all $s, t \in [0,1]$, it suffices to prove this equation for s, t such that $|s-t| < r$ if we choose $c \geq |\mathcal{A}| r^{-1/d}$.

Take $s, t \in I$ distinct with $|s-t| < r$. Truncate each $(i_1, i_2, \cdots) \in \Omega_N^\infty$ after the first index i_k such that

$$|s-t| \leq r_{i_1} \cdots r_{i_k} \leq |s-t| r^{-1}. \qquad [6]$$

(This is always possible since $|s-t| < r$.) Let $\mathcal{Q}$ denote the set of all (finite) sequences obtained in this way. Since every infinite sequence is truncated to some element of $\mathcal{Q}$, we have

$$I = \bigcup_{\omega \in \mathcal{Q}} \sigma_\omega(I).$$

Thus, there exist $\omega, \omega' \in \mathcal{Q}$ such that $s \in \sigma_\omega(I)$ and $t \in \sigma_{\omega'}(I)$. In fact, one can choose ω and ω' such that $\sigma_\omega(I)$ and $\sigma_{\omega'}(I)$ are either equal or adjacent intervals in I, since each such interval has length at least $|s-t|$. We conclude that $\sigma_\omega(I) \cap \sigma_{\omega'}(I) \neq \emptyset$ and hence $\Phi_\omega(\mathcal{A}) \cap \Phi_{\omega'}(\mathcal{A}) \neq \emptyset$. Since $f(s) \in \Phi_\omega(\mathcal{A})$ and $f(t) \in \Phi_{\omega'}(\mathcal{A})$,

$$\begin{aligned} |f(s) - f(t)| &\leq |\Phi_\omega(\mathcal{A})| + |\Phi_{\omega'}(\mathcal{A})| \\ &= |\mathcal{A}| \prod_{i \in \omega} R_i + |\mathcal{A}| \prod_{j \in \omega'} R_j \\ &= |\mathcal{A}| \prod_{i \in \omega} r_i^{1/d} + |\mathcal{A}| \prod_{j \in \omega'} r_j^{1/d} \\ &\leq 2|\mathcal{A}| r^{-1/d} \cdot |s-t|^{1/d} \end{aligned}$$

by equation [6]. □

Lemma 7 and Lemma 8 immediately imply

Corollary 9. *Suppose that $f : I \to \mathbb{R}^n$ is a balanced fractal curve with image $\mathcal{A} = f(I)$, and let $d = \dim_H \mathcal{A} = \dim_B \mathcal{A}$. Then for any set $X \subset [0,1]$,*

$$\dim_H f(X) \leq d \cdot \dim_H X \quad \text{and} \quad \dim_B f(X) \leq d \cdot \dim_B X.$$

iii. A lower bound on $\dim f(X)$

Our proof of the lower bound on the dimension of $f(X)$ is inspired by the proof of the dimension of attractors (Theorem 9.3 in (S43)). We use the mass distribution principle (Principle 4.2 in (S43)):

Lemma 10. *Let μ be a measure defined on X with $\mu(X) > 0$. Suppose that, for some $s > 0$, there exist $c, \epsilon > 0$ such that*

$$\mu(U) \leq c|U|^s \text{ for all } U \subset X \text{ such that } |U| \leq \epsilon.$$

Then $\mathcal{H}^s(X) \geq c^{-1}\mu(X)$ and $s \leq \dim_H X$.

The following Lemma by Frostman is often useful in constructing a measure based on the Hausdorff dimension of a set ((S48), Theorem 8.8).

Lemma 11. (Frostman) *Suppose that $X \subset \mathbb{R}^n$ is a Borel set with $0 < \mathcal{H}^s(X) \leq \infty$. Let $B_r(x)$ denote the ball of radius r centered at the point x. Then there exists a compact set $X_0 \subset X$ and a constant b such that $0 < \mathcal{H}^s(X_0) < \infty$ and*

$$\mathcal{H}^s(X_0 \cap B_r(x)) \leq br^s$$

for all $r > 0$ and $x \in \mathbb{R}^n$.

To each IFS $\mathcal{F}$ one can associate, by repeated sub-division, a canonical measure on its attractor $\mathcal{A}$ that respects its self-similar structure. Consider 'cylindrical' subsets of Ω_N^∞ defined by

$$\mathcal{O}_{i_1,\cdots,i_k} := \{\omega \in \Omega_N^\infty : \omega \text{ begins with } (i_1, \cdots, i_k)\}.$$

Let $s = \dim_H \mathcal{A}$ so that $\sum_{i=0}^{N-1} R_i^s = 1$, and define $\tilde{\mu}$ to be the measure on Ω_N^∞ that satisfies $\tilde{\mu}(\mathcal{O}_{i_1,\cdots,i_k}) = (R_{i_1} \cdots R_{i_k})^s$ for each $(i_1, \cdots, i_k) \in \Omega_N$. Then $\tilde{\mu}(\mathcal{O}_{i_1,\cdots,i_k}) = \sum_{i=0}^{N-1} \tilde{\mu}(\mathcal{O}_{i_1,\cdots,i_k,i})$, so $\tilde{\mu}(\Omega_N^\infty) = 1$. One can show that $\tilde{\mu}$ indeed extends to a measure on all of Ω_N^∞; it can be alternatively computed as

$$\tilde{\mu}(\mathcal{O}) = \inf\left\{\sum_\omega \tilde{\mu}(\mathcal{O}_\omega) : \mathcal{O} \subset \bigcup_\omega \mathcal{O}_\omega \text{ and } \omega \in \Omega_N\right\}.$$

Recall the indexing map associated with $\mathcal{F}$, $\rho_{\mathcal{F}} : \Omega_N^\infty \to \mathcal{A}$.

Definition 12. *The canonical measure μ on $\mathcal{A}$ associated with the IFS $\mathcal{F}$ is the push-forward of the measure $\tilde{\mu}$ on Ω_N^∞ by the indexing function ρ; that is,*

$$\mu(A) := \tilde{\mu}(\rho_{\mathcal{F}}^{-1}(A)) \quad \textit{for all } A \subset \mathcal{A}.$$

It is easily checked that the canonical measure μ satisfies $\mu(\mathcal{A}) = 1$ and that $\mu(\Phi_{i_1,\cdots,i_k}(\mathcal{A})) = (R_{i_1} \cdots R_{i_k})^s$ ((S43), Proposition 1.7).

To give the proof of Theorem 4, we require the following small result ((S43), Lemma 9.2):

Proposition 13. *Let $\{U_i\}$ be a collection of disjoint open subsets in $\mathbb{R}^n$. Suppose that each U_i contains a ball of radius a_1 and is contained in a ball of radius a_2. Then any ball with radius $r > 0$ intersects at most $(1 + 2a_2/r)^n (a_1/r)^{-n}$ of the closures $\overline{U_i}$.*

Proof. If the ball B with radius $r > 0$ meets $\overline{U_i}$, then U_i is contained in the ball B' concentric with B with radius $r + 2a_2$. Now if B meets exactly m of the $\overline{U_i}$, B' contains m disjoint balls of radius a_1. By comparing n-dimensional volumes of these balls, we find that $m(a_1)^n \leq (r + 2a_2)^n$ which proves the desired result. □

We complete the proof of Theorem 6 by showing the following result.

Lemma 14. *Suppose that $f : I \to \mathbb{R}^n$ is a balanced fractal curve with image $\mathcal{A} = f(I)$, and let $d = \dim_H \mathcal{A} = \dim_B \mathcal{A}$. Then for any Borel set $X \subset [0, 1]$,*

$$\dim_H f(X) \geq d \cdot \dim_H X \quad \textit{and} \quad \dim_B f(X) \geq d \cdot \dim_B X.$$

Proof. Let f be a fractal curve described by the IFSs $\{\sigma_0, \cdots, \sigma_{N-1}\}$ on I and $\mathcal{F} = \{\Phi_0, \cdots, \Phi_{N-1}\}$ on $\mathbb{R}^n$. Let r_i and R_i be the scaling ratios of σ_i and Φ_i respectively, for $i = 0, \cdots, N-1$. Since f is balanced, let d' be the constant such that $r_i = R_i^{d'}$ for $i = 0, \cdots, N-1$. By assumption $\sum_{i=0}^{N-1} r_i = 1$, so $\sum_{i=0}^{N-1} R_i^{d'} = 1$. Hence, by Theorem 4, $d' = \dim_H \mathcal{A} = \dim_B \mathcal{A} = d$.

Part (i): Hausdorff dimension.

Fix $s < \dim_H X$. Since $\mathcal{H}^s(X) = \infty$, by Frostman's Lemma there exists $X_0 \subset X$ compact and $b > 0$ such that $0 < \mathcal{H}^s(X_0) < \infty$ and

$$\mathcal{H}^s(X_0 \cap B_\rho(x)) \leq b\rho^s \text{ for all } x \in X \text{ and } \rho > 0. \qquad [7]$$

Define the measure μ on $f(X_0)$ to be the push-forward by f of $\mathcal{H}^s$ restricted to X_0; i.e.

$$\mu(A) := \mathcal{H}^s(f^{-1}(A) \cap X_0) \quad \text{for all } A \subset f(X_0).$$

We will bound $\mu(B_\rho(y))$ as a function of ρ for any $y \in f(X_0)$ and use Lemma 10 to show that $\dim_H f(X_0) \geq sd$.

Let B be a ball of radius ρ with $0 < \rho < 1$ and let $R = \min R_i$. Truncate each $(i_1, i_2, \cdots) \in \Omega_N^\infty$ after the first index i_k such that

$$R\rho \leq R_{i_1} \cdots R_{i_k} \leq \rho. \qquad [8]$$

Let $\mathcal{Q}$ denote the set of all (finite) sequences obtained in this way. Let $\mathcal{U}$ be an open set defined from the open set condition 3 for the IFS $\mathcal{F}$. Since the sets $\{\Phi_i(\mathcal{U}) : i = 0, \cdots, N-1\}$ are disjoint, applying the functions $\{\Phi_i\}$ recursively shows that, for $\omega, \omega' \in \Omega_N$, $\Phi_\omega(\mathcal{U})$ and $\Phi_{\omega'}(\mathcal{U})$ are disjoint if and only if neither ω nor ω' is a prefix of the other. Thus, $\{\Phi_\omega(\mathcal{U}) : \omega \in \mathcal{Q}\}$ is a collection of disjoint sets. It also follows easily from the open set condition that $\mathcal{A} \subset \overline{\mathcal{U}}$. Consequently,

$$\mathcal{A} \subset \bigcup_{\omega \in \mathcal{Q}} \Phi_\omega(\mathcal{A}) \subset \bigcup_{\omega \in \mathcal{Q}} \Phi_\omega(\overline{\mathcal{U}}).$$

Choose $a_1, a_2 \in \mathbb{R}$ so that $\mathcal{U}$ contains a ball of radius a_1 and is contained in a ball of radius a_2. For all $\omega = (i_1, \cdots, i_k) \in \mathcal{Q}$, the set $\Phi_\omega(\mathcal{U})$ contains a ball of radius $a_1(R_{i_1} \cdots R_{i_k})$ and therefore a ball of radius $a_1 R\rho$, and is contained in a ball of radius $a_2(R_{i_1} \cdots R_{i_k})$ and therefore a ball of radius $a_2\rho$. Let $\mathcal{Q}'$ denote the set of sequences $\omega \in \mathcal{Q}$ such that B intersects $\Phi_\omega(\overline{\mathcal{U}})$. By Proposition 13, $\mathcal{Q}'$ has at most $q_1 = (1 + 2a_2)^n (a_1 R)^{-n}$ elements.

Similarly, for some $\omega \in \mathcal{Q}$ and $\epsilon > 0$ consider a ball B_ω of radius $a_2\rho \cdot (1 + \epsilon)$ that contains $\Phi_\omega(\mathcal{U})$. By Proposition 13, B_ω intersects at most $(1 + 2/(1 + \epsilon))^n (a_1 R/a_2(1 + \epsilon))^{-n}$ of the sets $\Phi_{\omega'}(\overline{\mathcal{U}})$ for $\omega' \in \mathcal{Q}$. Since this is true for all $\epsilon > 0$, it follows that $\Phi_\omega(\overline{\mathcal{U}})$ is disjoint from all but $q_2 = 3^n (a_1 R/a_2)^{-n}$ of the sets $\Phi_{\omega'}(\overline{\mathcal{U}})$ for $\omega' \in \mathcal{Q}$.

For each $\omega \in \mathcal{Q}$, let $\mathcal{Q}'_\omega = \{\omega' \in \mathcal{Q} : \Phi_\omega(\overline{\mathcal{U}}) \cap \Phi_{\omega'}(\overline{\mathcal{U}}) \neq \emptyset\}$. For each $\omega \in \mathcal{Q}$, $\mathcal{Q}'_\omega$ has at most q_2 elements. Since $f \circ \sigma_{\omega'}(I) \subset \Phi_{\omega'}(\overline{\mathcal{U}})$, the set $f^{-1} \circ \Phi_\omega(\overline{\mathcal{U}})$ intersects $\sigma_{\omega'}(I)$ only if $\omega' \in \mathcal{Q}'_\omega$. Thus,

$$\begin{aligned} \mu(B) &\leq \sum_{\omega \in \mathcal{Q}'} \mu(\Phi_\omega(\overline{\mathcal{U}})) \\ &= \sum_{\omega \in \mathcal{Q}'} \mathcal{H}^s(f^{-1} \circ \Phi_\omega(\overline{\mathcal{U}}) \cap X_0) \\ &\leq \sum_{\omega \in \mathcal{Q}'} \sum_{\omega' \in \mathcal{Q}'_\omega} \mathcal{H}^s(\sigma_{\omega'}(I) \cap X_0). \end{aligned} \qquad [9]$$

For $\omega' = (i_1, \cdots, i_k) \in \mathcal{Q}$, $\sigma_{\omega'}(I)$ is an interval with length $r_{i_1} \cdots r_{i_k} = R_{i_1}^d \cdots R_{i_k}^d \leq \rho^d$ by equation [8]. By equation [7], $\mathcal{H}^s(\sigma_{\omega'}(I) \cap X_0) \leq b(\rho^d)^s$, so

$$\begin{aligned} \mu(B) &\leq \sum_{\omega \in \mathcal{Q}'} \sum_{\omega' \in \mathcal{Q}'_\omega} \mathcal{H}^s(\sigma_{\omega'}(I) \cap X_0) \\ &\leq \sum_{\omega \in \mathcal{Q}'} \sum_{\omega' \in \mathcal{Q}'_\omega} b(\rho^d)^s \\ &= q_1 q_2 b \cdot \rho^{ds}. \end{aligned} \qquad [10]$$

Since this is true for all balls B with radius ρ, $\dim_H f(X_0) \geq ds$ by Lemma 10. Because $X_0 \subset X$, we have shown that $\dim_H f(X) \geq \dim_H f(X_0) \geq ds$ for all $s < \dim_H X$. Consequently, $\dim_H f(X) \geq d \cdot \dim_H X$.

Part (ii): Box-counting dimension.

Fix $\rho > 0$. Let $\mathcal{B}$ be a collection of K balls, each of radius at most ρ, which covers $f(X)$. Define $\mathcal{Q} \subset \Omega_N$ exactly as in Part (i). Again, each ball $B \in \mathcal{B}$ intersects at most q sets in $\{\Phi_\omega(\overline{\mathcal{U}}) : \omega \in \mathcal{Q}\}$. Defining $\mathcal{Q}'$ to be the set of $\omega \in \mathcal{Q}$ such that $\Phi_\omega(\mathcal{U})$ intersects some $B \in \mathcal{B}$, we see that $\mathcal{Q}'$ has at most Kq_1 elements. Similarly define $\mathcal{Q}'_\omega$ for each $\omega \in \mathcal{Q}'$; each $\mathcal{Q}'_\omega$ has at most q_2 elements. From the reasoning in Part (i), it follows that $f(X) \subset \cup_{\omega \in \mathcal{Q}'} \Phi_\omega(\overline{\mathcal{U}})$ and

$$\begin{aligned} X &\subset f^{-1}(f(X)) \\ &\subset \bigcup_{\omega \in \mathcal{Q}'} f^{-1} \circ \Phi_\omega(\overline{\mathcal{U}}) \\ &\subset \bigcup_{\omega \in \mathcal{Q}'} \bigcup_{\omega' \in \mathcal{Q}'_\omega} \sigma_{\omega'}(I). \end{aligned} \quad [11]$$

As above, $\sigma_{\omega'}(I)$ is an interval of length at most ρ^d for all $\omega \in \mathcal{Q}$, so $\{\sigma_{\omega'}(I) : \omega \in \mathcal{Q}', \omega' \in \mathcal{Q}'_\omega\}$ is a cover of X using at most $q_1 q_2 K$ intervals with diameter at most ρ^d.

For $\delta > 0$ and a set A, let $N_\delta(A)$ denote the smallest number of balls of diameter at most δ required to cover A. Equation [11] shows that $N_{\rho^d}(X) \leq q_1 q_2 K$. Taking the minimum over all possible K, we have

$$N_{\rho^d}(X) \leq q_1 q_2 \cdot N_\rho(f(X)).$$

Consequently,

$$\begin{aligned} \dim_B X &= \lim_{\rho \to 0} \frac{\log N_{\rho^d}(X)}{-\log \rho^d} \\ &\leq \lim_{\rho \to 0} \frac{\log(q_1 q_2 \cdot N_\rho(f(X)))}{-\log \rho^d} \\ &\leq \lim_{\rho \to 0} d^{-1} \cdot \frac{\log N_\rho(f(X)) + \log q_1 q_2}{-\log \rho} \\ &= d^{-1} \cdot \dim_B f(X). \end{aligned}$$

□

d. Derivation of contact probability critical exponent for self-similar paths

Measurements of the contact probability exponent have emerged as a crucial tool for linking measurements of spatial contacts of nuclear DNA to models of its folded structure. The fractal globule model has been shown to exhibit a contact probability exponent ranging from -1 to -1.33, but it is unclear what range of exponents can be achieved by general fractal models of folding. By computing contacts of paths constructed by discrete sampling self-similar curves, we apply the theory derived above to characterize contact probability scalings of a wide range of fractal structures. Particularly relevant to chromatin models are motif-based *space-filling curves*, such as the Dragon curve or the Hilbert curve, in which dense packing leads to numerous self-contacts (Fig S3).

In this section, we derive an explicit formula for the contact probability scaling exponent for motif-based self-similar curves. Let f be such a curve, and set $d_i = \dim_B f(I) = \dim_H f(I)$, the interior dimension of the curve. We will also define a set E_f consisting of the points lying on the exterior of the curve, also known as the *dynamical boundary*, and define the exterior dimension of the curve to be $d_e = \dim_B E_f$. When f is space-filling, E_f is simply the topological boundary of the set $f(I)$. We will show that the contact probability of finite self-similar paths converging to f will scale with critical exponent $d_e/d_i - 2$.

i. Dimension of the contact map derives from surface dimension

For any curve $f : I = [0,1] \to \mathbb{R}^n$, one can compute the corresponding *contact map* $\mathcal{C}_f = \{(s,t) \in [0,1]^2 : f(s) = f(t), s \neq t\}$. The image of $\mathcal{C}_f$ is the set of points with multiple preimages, or

$$M_f := \{p \in \mathcal{A} : \exists (s,t) \in \mathcal{C}_f \text{ s.t. } f(s) = f(t) = p\},$$

which we call the *multiple points* of f. We assume henceforth that M_f is not empty.

When f is a balanced self-similar curve described by the iterated function systems $\mathcal{F} = \{\Phi_0, ..., \Phi_{N-1}\}$ and $S = \{\sigma_0, ..., \sigma_{N-1}\}$, the dimensions of $\mathcal{C}_f$ and M_f can be related using the following variant of the Theorem 6:

Lemma 15. *Let f be a balanced fractal curve with image $\mathcal{A}$, contact map $\mathcal{C}_f \subset I^2$, and multiple points $M_f \subset \mathcal{A}$. Define the map $g : \mathcal{C}_f \to M_f$ by $g : (s,t) \mapsto f(s) = f(t)$. Then for any Borel set $X \subset \mathcal{C}_f$,*

$$\begin{aligned} \dim_H g(X) &= \dim_H \mathcal{A} \cdot \dim_H X, \\ \dim_B g(X) &= \dim_B \mathcal{A} \cdot \dim_B X \end{aligned}$$

The proof of this result closely follows the proof of Theorem 6 and is given in the Appendix. As a consequence, $\dim_H \mathcal{C}_f = (\dim_H M_f)/(\dim_H \mathcal{A})$.

Because of self-similarity, M_f is dense in $\mathcal{A}$, so $\dim_B M_f = \dim_B \mathcal{A}$. That is, the box-dimension of the multiple points does not give any information about the prevalence of these self-contacts. Therefore, we consider the points which lie on the exterior of $\mathcal{A}$ that form external contacts when mapped into the interior of $\mathcal{A}$. Formally, choose $\mathcal{U} \subset \mathbb{R}^n$ such that $\mathcal{F}$ satisfies the open set condition with $\mathcal{U}$, and define the *exterior set* E_f to be

$$E_f = \{p \in \partial\mathcal{U} \cap \mathcal{A} : \exists \omega \in \Omega_N \text{ s.t. } \Phi_\omega(p) \in M_f\}.$$

We show in the Appendix that E_f is independent of the choice of $\mathcal{U}$. The exterior set, also known as the "dynamical boundary," generalizes the notion of boundary to self-similar sets. When f is *space-filling* – that is, it contains and is contained in open n-dimensional balls – E_f coincides with $\partial\mathcal{A}$, the topological boundary of $\mathcal{A}$ ((S49), Proposition 2.8).

In particular, since M_f is a countable union of images of E_f, $\dim_H E_f = \dim_H M_f$, and thus the dimension of the contact map can be computed as the ratio of the dimensions of the boundary and the interior. Furthermore, when E_f is compact, $\dim_B E_f = \dim_H E_f$ (see Proposition 17), enabling the boundary dimension to be computed with a box-counting measurement. We call this common value the *exterior dimension* of f, denoted d_e. Analogously, denote the *interior dimension* of f by $d_i = \dim_H \mathcal{A} = \dim_B \mathcal{A}$. Then Lemma 15 can be restated as $\dim_H \mathcal{C}_f = d_e/d_i$; we call this value the *contact dimension*.

Classic constructions of self-similar curves begin from a lattice path Γ_1 that serves as a base motif and use the self-similar rules to iteratively generate exponentially longer lattice paths that converge to a smooth self-similar curve. A subsequent path Γ_n is defined from Γ_{n-1} by placing a copy of Γ_1 at each node of Γ_{n-1}, suitably oriented so that the end of one copy of Γ_1 can be connected to the beginning of the next copy of Γ_1 by an edge on the sub-lattice. The number of nodes in the motif path specifies the number of self-similar copies – if Γ_1 has N nodes then Γ_n has N^n nodes – and the edges specify the

ordering and orientation of the copies. In this way, the base motif Γ_1 specifies an IFS, and the sequence of paths $\{\Gamma_n\}_{n=1}^{\infty}$ converges to a continuous curve described by the IFS. For an example of this process, see Section 4 or (S44).

In particular, the self-similar parts of a motif-based curve tile a lattice with only finitely many choices of rotations and reflections. Consequently, when f is a motif-based curve, E_f can be mapped into M_f via a bounded number of steps; i.e. there is some K such that

$$E_f = \{p \in \partial\mathcal{U} \cap \mathcal{A} : \text{ there exists } \omega \in \Omega_N^K \text{ such that } \Phi_\omega(p) \in M_f\}.$$

It follows that E_f is compact and therefore $\dim_B E_f = \dim_H E_f = d_e$.

With the same approach, we find that the contact dimension can also be measured using the box-counting dimension. If we examine the contact map away from the diagonal, $\mathcal{C}_f^j := \{(s,t) \in \mathcal{C}_f : s - t > 1/N^j\}$, we find from Lemma 15 that

$$\dim_B \mathcal{C}_f^j = \frac{\dim_B E_f}{\dim_B \mathcal{A}} = \frac{d_e}{d_i} \quad \text{for sufficiently large } j.$$

In the next section, we show that contact probability is effectively measuring $\dim_B \mathcal{C}_f^j$ for large j and derive the contact probability scaling as a function of d_e/d_i.

ii. Contact probability measures the contact map dimension

The contact probability of a curve measures the likelihood two points are in contact as a function of the distance along the curve between the points. Here we rigorously derive the contact probability scaling from the contact dimension.

Let f be a self-similar curve based on the length-N motif path Γ_1, and let $\{\Gamma_n\}_n$ be the iterative set of paths converging to f. For $a = 0, \cdots N^n - 1$, let $\Gamma_n(a)$ denote the ath node of Γ_n. We say that $\Gamma_n(a)$ and $\Gamma_n(b)$ are in contact when they are adjacent on the lattice, and define the discrete contact map of Γ_n by

$$\mathcal{I}_n = \{(a,b) : 0 \le a, b < N^n,\ \Gamma_n(a) \text{ and } \Gamma_n(b) \text{ are in contact}\}.$$

The contact map of Γ_n is a finite approximation to the infinite contact map $\mathcal{C}_f$ whose dimension was characterized above. Specifically, $(a,b) \in \mathcal{I}_n$ if and only if the square $[a/N^n, (a+1)/N^n] \times [b/N^n, (b+1)/N^n]$ has non-empty intersection with $\mathcal{C}_f$. (See Proposition 18.)

Define the *contact incidence* $I_n(k)$ to be the number of contact pairs in Γ_n at distance between N^{k-1} and $N^k - 1$, for $k = 1, ..., n$. This is the numerator of the contact probability measurement. The denominator is the "possible" number of contacts at distance between N^{k-1} and $N^k - 1$. Since the curve has length N^n, the possible number of contacts at distance ℓ is simply $N^n - \ell - 1$. Thus, contact probability, or $P_n(k)$, is given by the equation

$$P_n(k) = \frac{I_n(k)}{\sum_{\ell=N^{k-1}}^{N^k-1} N^n - \ell - 1}.$$

Contact incidence $I_n(k)$ counts the number of points lying inside a diagonal strip of the contact map of Γ_n. By dividing this strip into self-similar parts, we show in Lemma 19 in the Appendix that contact probability effectively makes a particular box-counting measurement of the contact map. Consequently, we find,

$$I_n(k) \sim (N^k)^{d_e/d_i - 1} \quad \text{for large } n.$$

Furthermore, it is clear that the denominator of $P(k)$ clearly scales as N^k. Therefore,

$$P_n(k) \sim (N^k)^{d_e/d_i - 2} \quad \text{for large } n.$$

We have proved that the contact probability for any motif-based curve obeys a power law with exponent equal to the contact dimension minus 2. That is, a plot of $\{(k, \log_N P(k))\}$ approaches a straight line with slope $\dim_B E_f / \dim_B \mathcal{A}_f - 2 = d_e/d_i - 2$ as n grows large.

This result is illustrated in Fig S3. We have constructed high iterations of numerous classic motif-based space-filling curves, and we find that our theory perfectly predicts the contact probability scaling measured from these paths (Table S3).

iii. Appendix: Proofs

Lemma 15. *Let f be a balanced fractal curve with image $\mathcal{A}$, contact portrait $\mathcal{C}_f \subset I^2$, and multiple points $M_f \subset \mathcal{A}$. Define the map $g : \mathcal{C}_f \to M_f$ by $g :\ (s,t) \mapsto f(s) = f(t)$. Then*

$$\begin{aligned} \dim_H g(X) &= \dim_H \mathcal{A} \cdot \dim_H X, \\ \dim_B g(X) &= \dim_B \mathcal{A} \cdot \dim_B X \end{aligned}$$

for any Borel set $X \subset \mathcal{C}_f$.

Proof. Let π denote the linear projection in $\mathbb{R}^2$ onto the x-axis. Then $g = f \circ \pi$. Since π is naturally α-Hölder with $\alpha = 1$, $\dim_H \pi(X) \le \dim_H X$ and $\dim_B \pi(X) \le \dim_B X$ for all $X \subset \mathcal{C}_f$. Theorem 6 then implies that $\dim_H g(X) \le \dim_H \mathcal{A} \cdot \dim_H X$ and $\dim_B g(X) \le \dim_B \mathcal{A} \cdot \dim_B X$.

The proofs of the reverse inequalities closely follow the proof of Theorem 6 with f replaced by g. In the proof of the Hausdorff dimension bound, the measure μ should be defined on M_f by

$$\mu(A) = \mathcal{H}^s(g^{-1}(A) \cap X_0) \ \ \forall A \subset M_f.$$

This is indeed a measure – clearly $\mu(\emptyset) = 0$ and $\mu(A) \le \mu(B)$ when $A \subset B$, and μ is countably additive since $g^{-1}(A)$ and $g^{-1}(B)$ are disjoint whenever A and B are disjoint.

The remainder of the proof follows in essentially the same manner as in Theorem 6. In the notation of this proof, since $g^{-1}(A) = (f^{-1}(A) \times f^{-1}(A)) \cap \mathcal{C}_f$, it follows that $g^{-1} \circ \Phi_\omega(\overline{\mathcal{U}}) \subset \sigma_\omega(I)^2$. Hence, equation **[9]** and equation **[11]** remain almost unchanged. Since $\sigma_\omega(I)^2$ is still a set with diameter at most ρ^d, equation **[10]** still holds. This shows that $\dim_H g(X) \ge \dim_H \mathcal{A} \cdot \dim_H X$ and $\dim_B g(X) \ge \dim_B \mathcal{A} \cdot \dim_B X$, as desired. □

Proposition 16. *The set E_f is independent of the choice of open set $\mathcal{U}$.*

Proof. Let $\mathcal{U}$ be any set for which $\mathcal{F}$ satisfies the open set condition. Then $\mathcal{A} \subset \overline{\mathcal{U}}$. If p satisfies the properties of an exterior point, then there exist $\omega, \omega' \in \Omega_N$, distinct and equal length, such that $p \in \Phi_\omega(\mathcal{A}) \cap \Phi_{\omega'}(\mathcal{A})$. Then $\Phi_\omega(\mathcal{U})$ and $\Phi_{\omega'}(\mathcal{U})$ are disjoint, so necessarily $p \in \partial\mathcal{U}$. □

Proposition 17. *If E_f is compact,* $\dim_B E_f = \dim_H E_f = \dim_H M_f$.

Proof. Suppose that f is described by the IFS $\mathcal{F} = \{(\Phi_0, \rho_0), ..., (\Phi_{N-1}, \rho_{N-1})\}$. Notice that E_f is *sub-self-similar* (S50); that is, E_f is compact and satisfies

$$E_f \subseteq \bigcup_{i=0}^{N} \Phi_i(E_f).$$

Then Theorem 3.5 in (S50) implies that the Hausdorff and Box-counting dimensions of E_f coincide. Since M_f is a countable union of images of E_f under the maps $\{\Phi_i\}$, $\dim_H E_f = \dim_H M_f$. □

Proposition 18. *Let f be a fractal curve which is self-similar in N equal parts. Construct Γ_n, the nth finite approximation to the curve f, consisting of N^n points on a lattice. Let $\mathcal{I}_n$ be the set of contact pairs in Γ_n; namely,*

$$\mathcal{I}_n = \{(a,b) : 0 \le a,b < N^n,\ \Gamma_n(a) \text{ and } \Gamma_n(b) \text{ are in contact}\}.$$

Then $(a,b) \in \mathcal{I}_n$ if and only if the square $[a/N^n,(a+1)/N^n] \times [b/N^n,(b+1)/N^n]$ has non-empty intersection with $\mathcal{C}_f$.

Proof. Suppose that f is described by the IFS $\{\Phi_0,...,\Phi_{N-1}\}$ in the image. Given integers a and b between 0 and N^n-1, let the base-N representations of a and b be given by

$$a = \alpha_1...\alpha_n$$

$$b = \beta_1...\beta_n$$

Then $(a,b) \in \mathcal{I}_n$ is equivalent to $\Phi_{(\alpha_1,...,\alpha_n)}(\mathcal{A}) \cap \Phi_{(\beta_1,...,\beta_n)}(\mathcal{A}) \neq \emptyset$, which is equivalent to the existence of $(s,t) \in \mathcal{C}_f$ with $a/N^n \le s \le (a+1)/N^n$ and $b/N^n \le t \le (b+1)/N^n$, as claimed. □

Lemma 19. *Let f be a motif-based self-similar curve. Construct Γ_n, the nth finite aproximation to f, and measure the contact incidence $I_n(k)$. For large n, there are constants α,β such that*

$$\alpha N^{(d_e/d_i-1)k} \le I_n(k) \le \beta N^{(d_e/d_i-1)k}$$

for large k.

Proof. Define $\mathcal{C}_f^j = \{(s,t) \in \mathcal{C}_f : s-t > 1/N^j\}$. Since f is motif-based, for some sufficiently large j_0, $\dim_B \mathcal{C}_f^{j_0} = d_e/d_i$. By replacing the N self-similarities describing f with N^{j_0} iterated self-similarities $\{\Phi_\omega : \omega \in \Omega_N \text{ has length } j_0\}$ which describe the same curve f, we may assume that $\dim_B \mathcal{C}_f^1 = d_e/d_i$. (In other words, replace the sequence of paths $\{\Gamma_n\}_{n=1}^\infty$ with $\{\Gamma_{j_0*n}\}_{n=1}^\infty$.)

Given some compact set X in the plane and $k \in \mathbb{N}$, we subdivide the plane into a grid of squares of side length N^{-k} with sides parallel to the x- and y-axes and the corner of one square at the origin. Let $B_k(X)$ equal the number of grid squares with side length N^{-k} which intersect X. Consider $\mathcal{I}_n$, the finite set of contacts of Γ_n. The number of points in the set

$$\mathcal{I}_n^{n-k} := \{(a,b) \in \mathcal{I}_n : N^k \le |a-b| < N^{k+1}\}$$

is exactly $2 \cdot I_n(k)$.

Set $K := N^{n-k}$ and define $2K-1$ disjoint sets in $\mathcal{I}_n^{n-k}$ as follows:

$$P_i := [(i-1)N^k, iN^k]^2 \cap \mathcal{I}_n^{n-k},$$

$$Q_i := ([iN^k,(i+1)N^k] \times [(i-1)N^k, iN^k]) \cap \mathcal{I}_n^{n-k},$$

where the P_i are defined for $i = 1,...,K$ and the Q_i are defined for $i = 1,...K-1$. Fig S4C shows an example of $\mathcal{I}_n^{n-1}$, $\{P_i\}$, and Q_1 when f is the Hilbert curve.

By self-similarity of the contact pairs, the sets $P_1,...,P_K$ are identical and equal to a scaled down copy of $\mathcal{I}_k^1$. Then by Proposition 18, the number of points in each P_i is exactly equal to $B_k(\mathcal{C}_f^1)$. Since $P_1,...,P_K$ are disjoint and contained in $\mathcal{I}_n^{n-k}$,

$$2 \cdot I_n(k) \ge \sum_{i=1}^{K} \#P_i = N^{n-k} \cdot B_k(\mathcal{C}_f^1).$$

Because f is motif-based, two sub-parts of the curve can only contact each other in finitely many different combinations. Specifically, define the contact map between consecutive sub-parts

$$Z_f = \bigcup_{k=1}^{\infty} \bigcup_{j=1}^{N^k-1} \{\ \phi_{k,j}((s,t)) \in \mathcal{C}_f^k\ :\ \frac{j}{N^k} \le s \le \frac{j+1}{N^k},\ \frac{j-1}{N^k} \le t \le \frac{j}{N^k}\},$$

where $\phi_{k,j}$ is chosen to map the sub-square $[j/N^k,(j+1)/N^k] \times [(j-1)/N^k, j/N^k]$ back to the unit square. (So $Z_f \subset [0,1]^2$.) Because consecutive sub-parts form contacts combinatorially in finitely many ways, Z_f can be written as a finite union over a subset of the sets above. Thus, Z_f itself has dimension at most d_e/d_i.

Again by Proposition 18, a point in Q_i corresponds to a box of side-length N^{-k} in the same location that intersects Z_f. Thus, the number of points in Q_i is at most $B_k(Z_f)$.

Now since $\mathcal{I}_n^{n-k}$ is contained in the union of the P_i and the Q_i, counting points gives

$$\begin{aligned} 2 \cdot I_n(k) &= \sum_{i=1}^{N^{n-k}} \#P_i + 2 \cdot \sum_{j=1}^{N^{n-k}-1} \#Q_i \\ &\le N^{n-k} \cdot B_k(\mathcal{C}_f^1) + 2(N^{n-k}-1) \cdot B_k(Z_f), \end{aligned}$$

where $\#P_i$ and $\#Q_i$ are the number of points in P_i and Q_i respectively. Combining the inequalities above shows

$$\begin{aligned} 2 \cdot I_n(k) &\ge N^{n-k} \cdot B_k(\mathcal{C}_f^1) \\ 2 \cdot I_n(k) &\le N^{n-k} \cdot B_k(\mathcal{C}_f^1) + 2(N^{n-k}-1) \cdot B_k(Z_f). \end{aligned}$$

Since $\mathcal{C}_f^1$ and Z_f both have box-counting dimension equal to d_e/d_i, $\log_N B_k(\mathcal{C}_f^1)$ and $\log_N B_k(Z_f)$ each scale linearly as a function of k with slope d_e/d_i. Thus, $\log_N I_n(k)$ scales with slope $d_e/d_i - 1$ as a function of k. □

e. Novel curves

i. Construction of space-filling curves using tiling

The IFS-based construction of self-similar curves presented in Section 2 may be interpreted nicely in terms of an iterative geometric construction. In this section, we walk through a general process for constructing self-similar space-filling curves.

First, the region chosen to be filled by the curve is subdivided into n parts each equal to the whole, giving a self-similar tiling of the original shape, also known as a *rep-tiling*. Second, a Hamiltonian path is chosen which passes through each tile; this forms the *base motif*. Third, a rewrite rule is determined using this base motif. Specifically, each of the n tiles is partitioned in the same manner as the original shape, and the base motif is drawn in each tile in a suitable orientation so that the end of the base motif in the ith tile may be joined directly to the beginning of the base motif in the $(i+1)$th tile, where the ordering of the tiles is also determined by the base motif. By iteratively applying the rewrite rule, arbitrarily long paths through the region are constructed, and these paths converge to a self-similar curve that fills the chosen region.

The rep-tiling, base-motif, and rewrite rule for the Hilbert curve is shown in Fig S5A. The region is a square, and the tiling consists of four squares of half the size. A path through the four squares is chosen as the base motif. Finally, one copy of this motif is placed in each square, suitably oriented, to define the rewrite rule. This approach may be generalized to higher dimensions, for example, to construct the 3D Hilbert curve (Fig S3).

Notice that it is possible to construct many different self-similar curves that fill the same space; for example, the Hilbert and Peano curves both fill the square. By subdividing the square into more sub-squares and choosing different paths through the sub-squares to form the base motif, a combinatorially unlimited number of such curves may be constructed.

Until now, all curves we have discussed have been unknotted. By subdividing a cube into sufficiently many parts and choosing a knotted base-motif, we can construct a space-filling curve that has arbitrarily many knots as the path is repeatedly iterated. We illustrate this with a construction, dubbed the *Gordian knot*, which subdivides the cube into 64 parts and choosing a base-motif which forms two trefoil knots (if the ends are joined) (Fig S5D). If the endpoints are joined along the outside to form a loop, the nth iteration contains $\sum_{i=0}^{n-1} 16\dot{6}4^i$ trefoil knots. Notably, the contact probability is unaffected by the knotted nature of this curve, as predicted by our theory (Fig S3).

It is possible to construct self-similar space-filling curves using tiles that overlap. The Levy Dragon is such an example: it has dimension equal to 2 and is self-crossing. Interestingly, although our theoretical results only apply to curves that are not self-crossing (as required by the open set condition), the Levy Dragon contact probability exponent satisfies the predictions of the theorem (Fig S3). Thus, there may be a generalization of our results for self-crossing curves.

ii. Rough-boundaried space-filling curves

By choosing an irregular region, a space-filling curve with a rough boundary may be constructed. However, the region must be chosen appropriately so that it self-interlocks and tiles space. In a general method, we generate a *reciprocal tesselation* by taking a simple tiling, such as the square tiling of the plane, and modifying the underlying tile by adding a bump to one side and a corresponding cavity on the other side (Fig S5B). This operation can be performed repeatedly, creating tiles with rougher boundaries that look nothing like squares.

We illustrate this approach in two- and three-dimensions. First, we create a very simple reciprocal tesselation by translating a single sub-square in a square tiling from the bottom of the tile to the top, and doing the same for a sub-square from left-to-right (Fig S5B). We then construct a Hamiltonian Path through this tile. However, we immediately face a hurdle: it turns out that there is no way to iterate this Hamiltonian path recursively, since the 2D rotational symmetry of the tile is destroyed by the change in tile shape. As a result, it is not possible to connect our Jigsaw motif to construct a motif-of-motifs. However, it is possible to compensate for this symmetry-breaking by constructing a second motif (Fig S5B). By using one or the other motif, it is possible to iterate this construction indefinitely as before. This yields a Peano curve whose boundary has a fractal dimension equal to $\log 12/\log 8 = 1.195$, and which exhibits the expected shift in scaling (Fig S3). The term "Jigsaw curve" is suggested by the shape of the resulting curve.

This technique may also be applied successfully in three dimensions to create the 3D Jigsaw curve. It has concavities and convexities on opposite faces, just like the jigsaw curve does on opposite sides. The tiling pattern ensures the boundary will be rough.

The base motif lattice is a 666 lattice with all six faces possessing either convexities or concavities (Fig S5C, second row). The three convex faces meet at a single vertex (the central vertex) and have their center four points replicated one unit outward, and the opposite faces, which are concave, meet at the opposite vertex (the anti-central vertex) and have their center four points deleted.

As before, because this shape is not fully rotationally symmetric, we require multiple base motifs. Specifically, there are twenty four base motif Hamiltonian paths, each of which begin at one of the eight corners and end at of the corresponding three adjacent corners. These may be specified by only three archetypal base motifs:

- Type 1 (Fig S5C, top left): All six Hamiltonian paths which either begin or end at the central vertex. They are symmetric via rotations and path reversals.
- Type 2 (Fig S5C, top middle): All six Hamiltonian paths that begin or end at anti- central vertex. They are symmetric via rotations and path reversals.
- Type 3 (Fig S5C, top right): All twelve Hamiltonian paths which are neither Type 1 nor Type 2.

Using these base motifs, a rewrite rule can be defined, as shown in Fig S5C.

Mathematically, the dimension of the surface of the 3D Jigsaw curve is $\log 44/\log 6 \approx 2.1120$. We computed the contact probability for the third iteration of the 3D Jigsaw curve possessing 10,077,696 monomers and confirmed the γ value predicted by our theory (Fig S3). As expected, the contact probability scaling is robust in the presence of fractal boundaries in three-dimensional space.

In the following section, reciprocal tesselations are applied to construct a family of curves in two dimensions with arbitrarily rough surfaces. For these constructions, we are able to use a single base-motif by choosing the tesselations to be rotationally symmetric.

iii. Inside-Out Hilbert curves

The Hilbert Curve (S51) is the limit in the plane of a sequence of self-avoiding approximation paths H^n (Fig S6, Row A) as $n \to \infty$. For $n > 0$ each path H^n is the composition $H^1 \circ H^{n-1}$, where $H^1 = H$ is called the curve's *motif*. The motif specifies where to place suitably oriented and connected copies of the previous approximation path to build the next approximation. When the unit square S is subdivided into a $2^n \times 2^n$ array of subsquares, path H^n passes through every subsquare (of side $1/2^n$) in S. In the limit, the self-similar, space-filling Hilbert Curve $H^{n\to\infty}$ has infinite length, is continuous, and reaches every point inside a two-dimensional unit square. The curve is fractal, but its filled area's square boundary—with dimension $D = 1$ and length 4—is not.

Many space-filling curves in the plane have boundaries with fractal dimension $D > 1$; for example the Dragon Curve has a boundary dimension $D = 1.5236+$. Can we construct a planar space-filling curve whose boundary's fractal dimension is arbitrarily close to 2? The answer is *yes*.

To construct motifs M_n for a sequence C_n of self-similar space-filling curves, each having a boundary with fractal dimension $D_n > D_{n-1}$ where $D_{n\to\infty} = 2$, we can turn the Hilbert Curve construction upon itself, as shown in Fig S6:

I. [Row A]. Choose an $n \geq 0$ and consider the Hilbert Curve approximation path H^n (e.g., for $n = 2$ consider the central column of Fig S6, starting at the top, where H^2 is circled).
II. [Row B]. Subdivide the unit square S into $(8 \cdot 2^n) \times (8 \cdot 2^n)$ subsquares. Call the lower left subsquare the *start* and the lower right subsquare the *end* (marked with dots). Along each of the four sides of subdivided S, use two copies of suitably scaled path H^n from Row A as a boundary "mold" by which 4×4 groups of subsquares are rearranged so as to extrude one half-side outward, and the other half-side inward (circled along the top for $n = 2$). The resulting shape,

called a *prototile*, has rotational (p4) wallpaper symmetry. The subsquare count necessarily remains the same, so each prototile, regardless of n, also has unit area.

III. [Rows C–F]. Construct a Hamiltonian path that visits all 64×2^{2n} subsquares in the prototile from *start* to *end*, using the following algorithm:

- [Row C] Beginning at *start*, follow the left-hand rule (shown in purple) along the inside of the prototile's boundary until the path hits the subsquare two subsquares above *end*.
- [Row D] Make a U-turn, then follow the right-hand rule back through unvisited subsquares until the path returns (as it must) to the subsquare directly to the right of *start*. Just under 3/4 of all the prototile's subsquares have now been visited.
- [Row E] Treat 2×2 groups of the remaining subsquares as super-subsquares (in green). Beginning to the right of *start*, construct a partial path P (in red) visiting these super-subsquares by following the right-hand rule until the path becomes adjacent to the central hub (four super-subsquares) of the prototile. Then change to the left-hand rule to include the hub, then change back to the right-hand rule upon leaving the hub. Continue until path P ends at the center of the super-subsquare whose lower right subsquare is *end*.
- [Row F] Construct $P \circ H$. In other words, "Hilbertize" partial path P by placing an H motif into each super-subsquare, suitably oriented to permit the sequence of Hs to be connected to form the final portion of the complete path from *start* to *end*.

The final constructed path is a self-avoiding motif M_n^1 for a space-filling curve whose approximation paths are also self-avoiding. M_n^2 is recursively composed of $64 \cdot 2^{2n}$ copies of M_n^1 scaled by $1/(8 \cdot 2^n)$, suitably oriented and connected according to the sequence of subsquares that M_n^1 specifies. Hence the limiting curve C_n $(= M_n^\infty)$ has a boundary with fractal dimension parameterized by n, as determined by the usual (edge substitution) Hausdorff formula

$$D_n = \frac{\log L_n}{\log (1/r_n)} \,,$$

where L_n is the extended-inward-and-outward edge-length of one side of S, measured in subsquare widths, and r_n is the similarity ratio (scaling factor) used to determine subsquare size. So we have

$$D_n = \frac{\log (8 \cdot 2^{2n})}{\log (8 \cdot 2^n)} = \frac{3 + 2n}{3 + n} \to 2 \,, \quad \text{as } n \to \infty \,.$$

The increasing boundary dimensions of the first few of these meta-Hilbert curves are:

$$\begin{aligned} D_0 &= \frac{3}{3} = 1 \\ D_1 &= \frac{5}{4} = 1.25 \\ D_2 &= \frac{7}{5} = 1.4 \\ D_3 &= \frac{9}{6} = 1.5 \\ D_4 &= \frac{11}{7} = 1.5714 \\ D_5 &= \frac{13}{8} = 1.625 \\ &\ldots \end{aligned}$$

Notice that boundary dimension D_4 slightly exceeds that of the Dragon Curve.

The area A_n of each space-filling curve C_n plainly (by symmetry) equals 1. But as n increases, this area, while remaining connected, becomes increasingly rarified as it is distributed in fractal tendrils throughout either side of a set of infinitely "fractalized" Hilbert approximation paths. In the limit, as $n \to \infty$, all tendril widths—and gaps between them—go to 0 so that the interior of C_∞ in some sense vanishes just as its boundary becomes space-filling, creating a new rotationally symmetric tiling region whose area is now not 1, but 2. The fractal fuzz that makes up the high-dimensional boundary collapses, leaving just a collection of standard Hilbert Curves, piecewise connected end-to-end, all with their usual linear boundaries. The first six cover the upper 3/4 of the limit set C_∞. The second six travel backwards, covering the exact same areas as the first six, but in opposite order and direction. (Like all space-filling curves, each Hilbert Curve is surjective, so here we have a sort of doubly surjective covering.) The final two Hilbert Curves cover the remaining 1/4 of the area along the bottom (Row E of Fig S6). A moving point that maps the unit interval to C_∞ covers these last two Hilbert Curves at half the instantaneous velocity it would cover the previous twelve Hilbert Curves. Hence C_∞ unexpectedly comprises just fourteen half-size Hilbert Curves.

The meta-motif construction technique described here works with any square-filling, generalized Hilbert Curve whose motif is a so-called *Greek key tour* (see Sloane's integer sequence A000532).

Because each C_n is a planar tiling region exhibiting p4 symmetry, and because each can be piecewise-connected in exactly the same places in a square array that self-similar sub-Hilbert Curves in a Hilbert Curve would be connected, one can create generally square, space-filled tiling regions having fractal boundaries whose dimension can be tuned—in theory, if not in practice—arbitrarily close to 2. (The number of prototile subsquares just a motif path must visit for $D_{297} = 1.99$ is enormous: $64 \times 2^{2 \cdot 297} = 2^{600}$.)

When generalized to three dimensions, a composite approximation path of this type might model a densely packed polymer bundle that exhibits a fuzzy external boundary at the same time as its interior evinces multiple self-similar regimes at different scales. However, no explicit three-dimensional analog currently exists.

S1. Li H and Durbin R (2010) Fast and accurate long-read alignment with Burrows-Wheeler Transform. *Bioinformatics*, 26, 589–595.
S2. Knight P and Ruiz D (2012) A fast algorithm for matrix balancing. *IMA J of Numer Anal*, 35.
S3. Beliveau, B et al. (2012) Versatile design and synthesis platform for visualizing genomes with Oligopaint FISH probes. *Proc Natl Acad Sci USA*, 109:21301-21306.
S4. Grant CE, Bailey TL, Noble WS (2011) FIMO: scanning for occurences of a given motif. *Bioinformatics* 27(7):1017-1018.
S5. Kim TH et al. (2007) Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. *Cell* 128(6):1231-1245.
S6. Schmidt D, et al. (2012) Waves of Retrotransposon Expansion Remodel Genome Organization and CTCF Binding in Multiple Mammalian Lineages. *Cell* 148, 335-348.
S7. Doench JG, et al. (2014) Rational design of highly active sgRNAs for CRISPR-Cas9-mediated gene inactivation. *Nat Biotechnol* 32(12): 1262-1267.
S8. Hsu PD, et al. (2013) DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat Biotechnol* 31(9):827-832.
S9. Ran FA, et al. (2013) Genome engineering using the CRISPR-Cas9 system. *Nat Protoc* 8(11):2281-2308.
S10. Chu V, et al. (2015) Increasing the efficiency of homology-directed repair for CRISPR-Cas9-induced precise gene editing in mammalian cells. *Nat Biotechnol* 33(5):543-548.
S11. Maruyama T, et al. (2015) Increasing the efficiency of precise genome editing with CRISPR-Cas9 by inhibition of nohomologous end joining. *Nat Biotechnol* 33(5):538-542.
S12. Rubinstein M, Colby RH (2003) *Polymer Physics* (Oxford University Press).
S13. Dekker J, Rippe K, Dekker M, Kleckner N (2002) Capturing Chromosome Conformation. *Science* 295, 1306-1311.
S14. Bystricky K, Heun P, Gehlen L, Langowski J, Gasser SM (2004) Long-range compaction and flexibility of interphase chromatin in budding yeast analyzed by high-resolution imaging techniques. *Proc Natl Acad Sci USA* 101, 16495?16500.
S15. Schiessel H (2003) The physics of chromatin. *J Phys Condens Matter* 15, R699-R774.
S16. Aumann F, Lankas F, Caudron M, Langowski J (2006) Monte Carlo simulation of chromatin stretching. *Phys Rev E* 73, 041927.
S17. Langowski, J (2006) Polymer chain models of DNA and chromatin. *Euro Phys J E* 19, 241-249.
S18. Fussner E, Ching R, Bazett-Jones D (2011) Living without 30nm chromatin fibers. *Trends in Biochem Sci* 36, 1?6.
S19. Fussner E et al. (2012) Open and closed domains in the mouse genome are configured as 10-nm chromatin fibres. *EMBO Rep* 13, 992?996.
S20. Joti Y et al. (2012) Chromosomes without a 30-nm chromatin fiber. *Nucleus* 3, 404-410.
S21. Schram R, Barkema G, Schiessel H (2013) On the stability of fractal globules. *J Chem Phys* 138, 224901.
S22. Goldman MA (1988) The chromatin domain as a unit of gene regulation. *Bioessays* 9, 50?55.
S23. Bulger M, Groudine M (1999) Looping versus linking: toward a model for long-distance gene activation. *Genes & development* 13(19), 2465-2477.
S24. Naumova N et al. (2013) Organization of the mitotic chromosome. *Science* 342, 948?953.
S25. Bohn M, Heermann D (2010) Diffusion-driven looping provides a consistent framework for chromatin organization. *PloS One* 5, e12218.
S26. Barbieri M et al. (2012) Complexity of chromatin folding is captured by the strings and binders switch model. *Proc Natl Acad Sci USA* 109, 16173?16178.
S27. Le TBK, et al. (2013) High-resolution mapping of the spatial organization of a bacterial chromosome. *Science* 342: 731-734.
S28. Kalhor R et al. (2012) Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. *Nat Biotechnol* 30(1):90-98.
S29. Giorgetti L et al. (2014) Predictive polymer modeling reveals coupled fluctuations in chromosome conformation and transcription. *Cell* 157.4:950-963.
S30. Humphrey W, Dalke A, Schulten K (1996) VMD: visual molecular dynamics. *J Mol Graph* 14, 33?38.
S31. Halperin A, Goldbart P (1999) Early Stages of Homopolymer Collapse. *Phys Rev E* 61(1): 565.
S32. Frisch T, Verga A (2002) Slow relaxation and solvent effects in the collapse of a polymer. *Phys Rev E* 66.
S33. Clark D, Kimura T (1990) Electrostatic mechanism of chromatin folding. *J Mol Bio* 211, 883896.
S34. Cui Y, Bustamante C (2000) Pulling a single chromatin fiber reveals the forces that maintain its higher-order structure. *Proc Natl Acad Sci USA* 97, 127-132.
S35. Luger K, Hansen J (2005) Nucleosome and chromatin fiber dynamics. *Current Opinion in Structural Biology* 15, 188-196.
S36. Chodaparambil et al. (2007) A charged and contoured surface on the nucleosome regulates chromatin compaction. *Nat Struc Mol Bio* 14(11):1105-1107.
S37. ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature* 489, 57?74.
S38. Marsden M, Laemmli U (1979) Metaphase chromosome structure: Evidence for a radial loop model. *Cell* 17, 849-858.
S39. Berg B, Foerster D (1981) Random paths and random surfaces on a digital computer. *Phys Lett B* 106(4): 323-326.
S40. Aragao de Carvalho C, Caracciolo S (1983) A new Monte-Carlo approach to the critical properties of self-avoiding random walks. *J de Phys* 44, 323-331.
S41. van Rensburg EJJ, Whittington SG (1991) The BFACF algorithm and knotted polygons. *J Phys A: Math Gen* 24, 5553.
S42. Dodd IB, Sneppen K (2011) Barriers and silencers: a theoretical toolkit for control and containment of nucleosome-based epigenetic states. *J Mol Bio* 414, 624-637.
S43. Falconer K (2003) *Fractal geometry: mathematical foundations and applications* (Wiley).
S44. Barnsley MF (1993) *Fractals Everywhere*, 2nd ed. Academic Press Professional.
S45. Hutchinson J (1981) Fractals and self similarity. *Indiana Univ Math J* 30(5):713-747.
S46. Blumenthal R, Getoor R (1961) Sample functions of stochastic processes with stationary independent increments. *J Math and Mech* 10, 493-516.
S47. Khoshnevisan D, Xiao Y (2005) Levy Processes: capacity and Hausdorff dimension. *The Annals of Probability* 33(3):841-878.
S48. Mattila P (1995) *Geometry of Sets and Measures in Euclidean Spaces: Fractals and Rectifiability*. Cambridge University Press.
S49. Moran M (1999) Dynamical boundary of a self-similar set. *Fundamenta Mathematicae* 160(1).
S50. Falconer K (1995) Sub-Self-Similar Sets. *Trans Amer Math Soc* 347(8):3121-3129.
S51. Hilbert D (1891) Uber die stetige abbildung einer linie auf ein flachenstuck. *Mathematische Annalen* 38, 459-460.

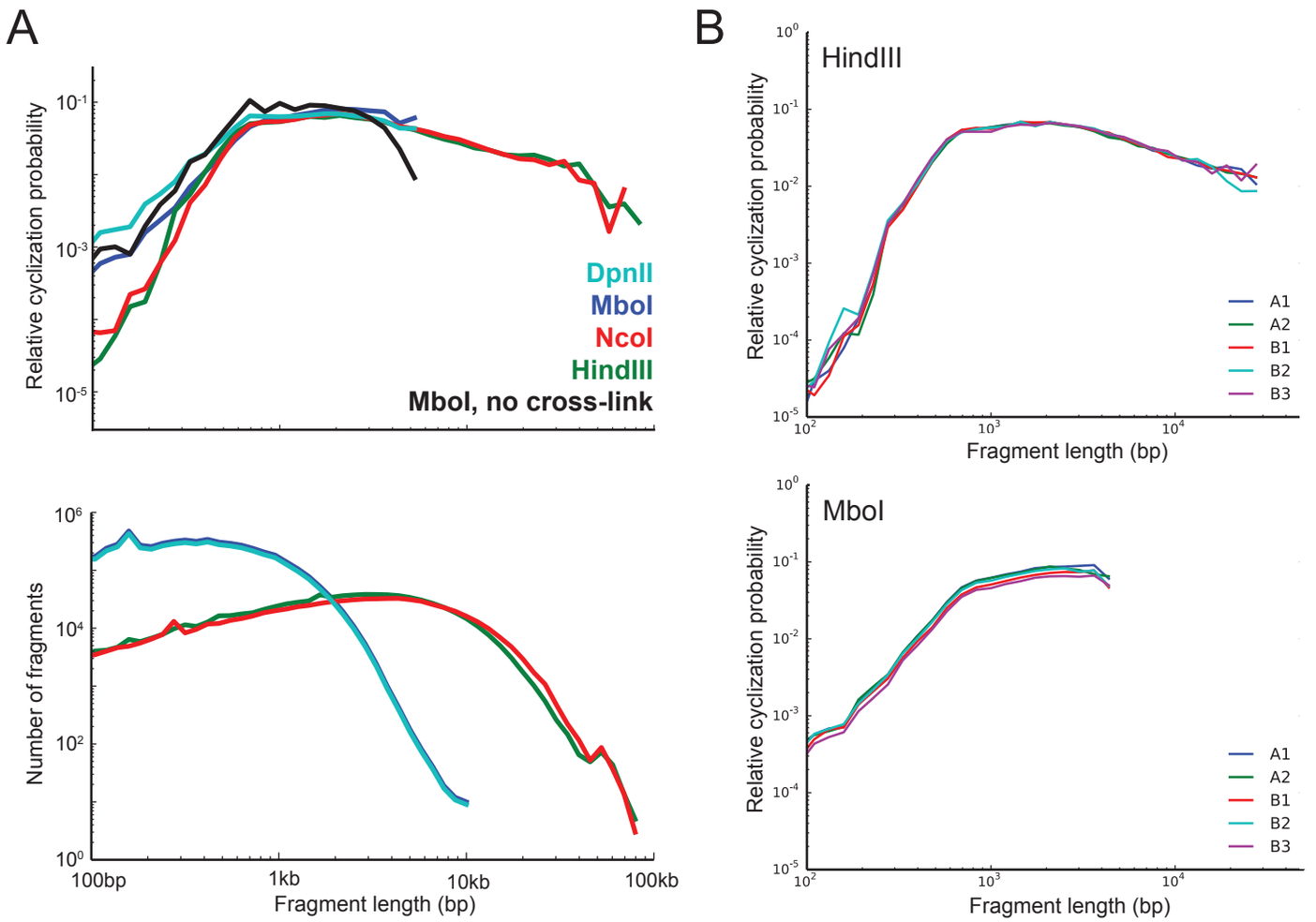


Fig. S1. Additional flexibility measurements
(A) *Above*: cyclization probability measured from a native Hi-C library in which no cross-linking was performed (8) was also consistent with cyclization measurements shown in Fig 1B. *Below*: distribution of restriction fragment lengths when the human genome is cut with DpnII (restriction site: GATC), MboI (GATC), NcoI (CCATGG), or HindIII (AAGCTT). **(B)** Flexibility measurements are consistent across the five nuclear compartments A1/A2, corresponding to active chromatin, and B1/B2/B3 corresponding to repressed chromatin (8). Relative cyclization probability within each compartment was plotted for Hi-C experiments using HindIII (top) and MboI (bottom) restriction enzymes.

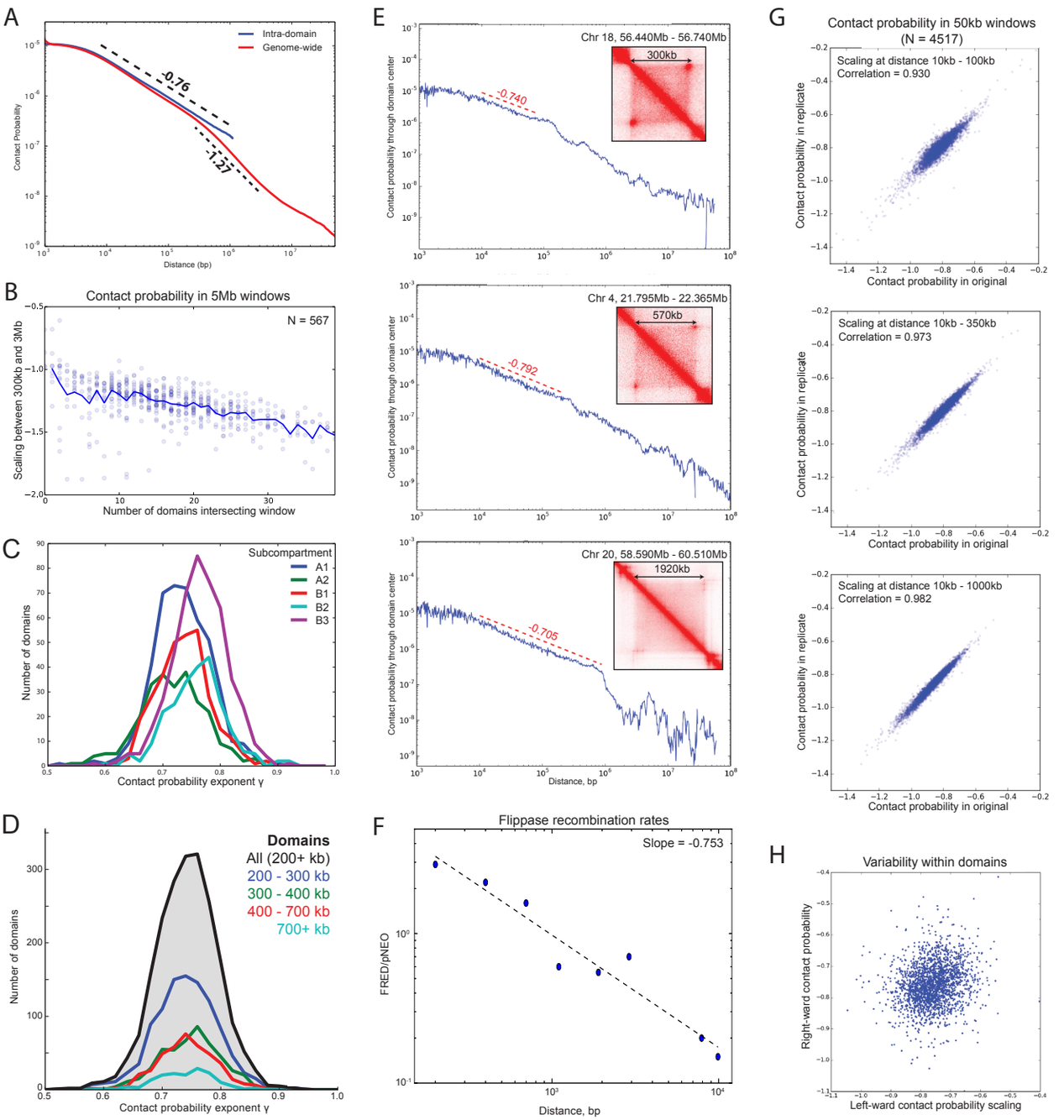


Fig. S2. Contact probability of in situ Hi-C maps.
(A) Contact probability of the primary Hi-C map (GM12878). Contact probability aggregated over intra-domain contacts (blue) exhibits a power law with $\gamma = 0.76$ between 10kb and 1Mb. Contact probability is aggregated over all contacts genome wide (red) has a scaling with $\gamma = 1.27$ between 300kb and 3Mb. **(B)** Power law exponents between 300kb and 3Mb for contact probability measured at 567 5Mb regions tiling the genome. Regions containing many domains exhibit steeper contact probability decay. **(C)** Histogram of domain contact probability scalings in GM12878 sorted by nuclear subcompartment. **(D)** Histogram of γ values observed inside 1057 high-confidence domains larger than 200kb throughout the genome. Average value is -0.75, standard deviation is 0.05. The value of γ does not depend on domain size. **(E)** Contact probability of a 50kb window through the center of three different domains. A power law with $\gamma \sim 0.75$ extends to the domain boundary, independent of domain size. Contacts often drop sharply at the domain boundary. **(F)** Measurements of Flippase recombination rates also exhibit a scaling around -0.75. Data is from (18), Figure 7C. Note that (S42) comments in passing about the presence of a power law scaling in this data, but does not provide further details. **(G)** Local contact probability scalings, measured on 50kb windows in three distance ranges (10-100kb, 10-350kb, 10-1000kb), are strongly correlated between replicates. **(H)** Left-ward and right-ward contact probability scalings in adjacent, non-overlapping windows at the center of domains are uncorrelated.

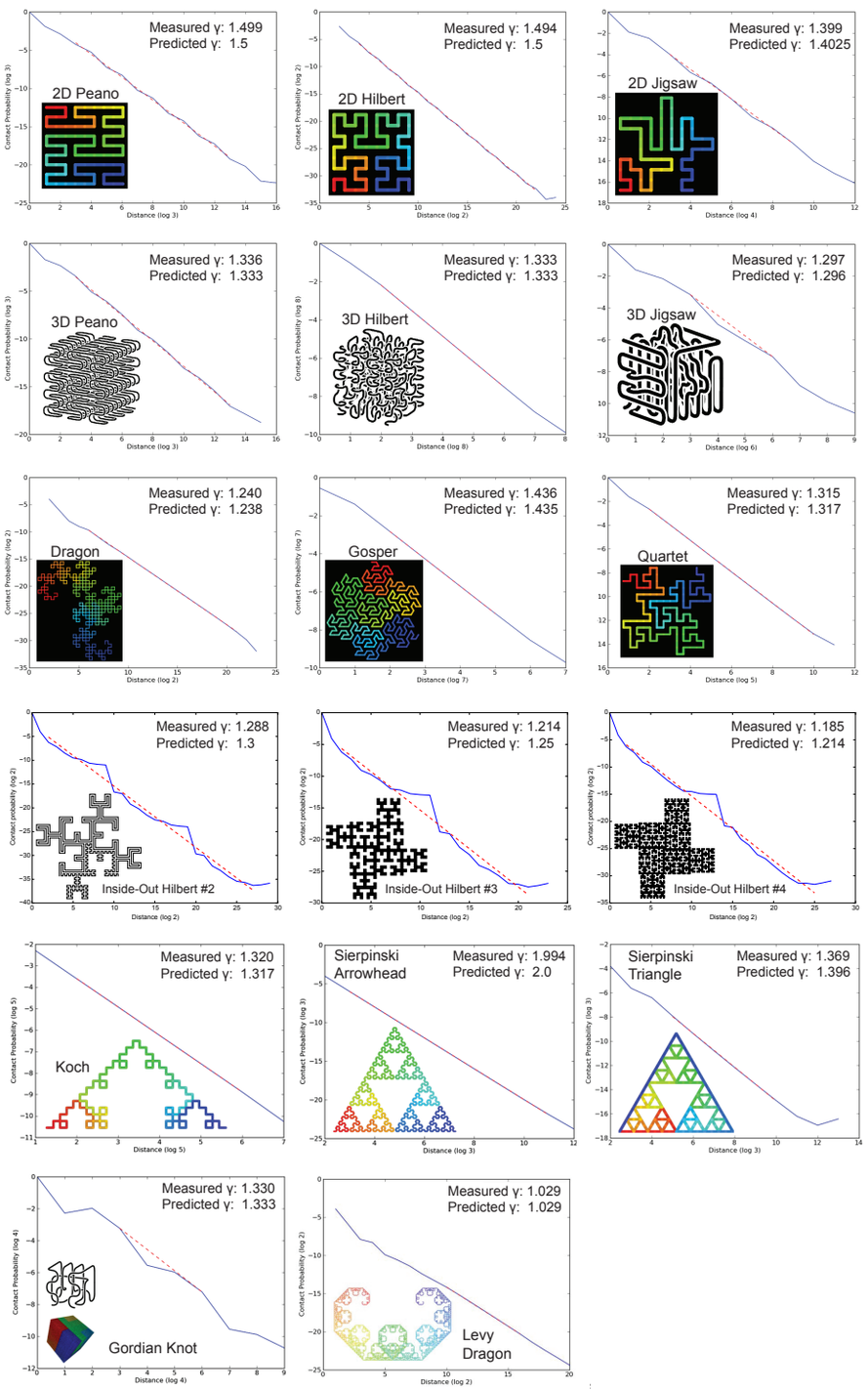


Fig. S3. Contact probability for seventeen self-similar curves, comparing measured γ to values predicted by our theory.

Contact probability plots. Each curve was iterated highly, often containing millions of segments. Left to right, top to bottom: 2D Peano curve, 2D Hilbert curve, 2D Jigsaw curve; 3D Peano curve, 3D Hilbert curve, 3D Jigsaw curve; Dragon curve, Gosper curve, Quartet curve; Inside-Out Hilbert #2, Inside-Out Hilbert #3, Inside-Out Hilbert #4; Koch curve, Sierpinski Arrowhead curve, Sierpinski triangle curve; 3D Gordian Knot, Levy Dragon.

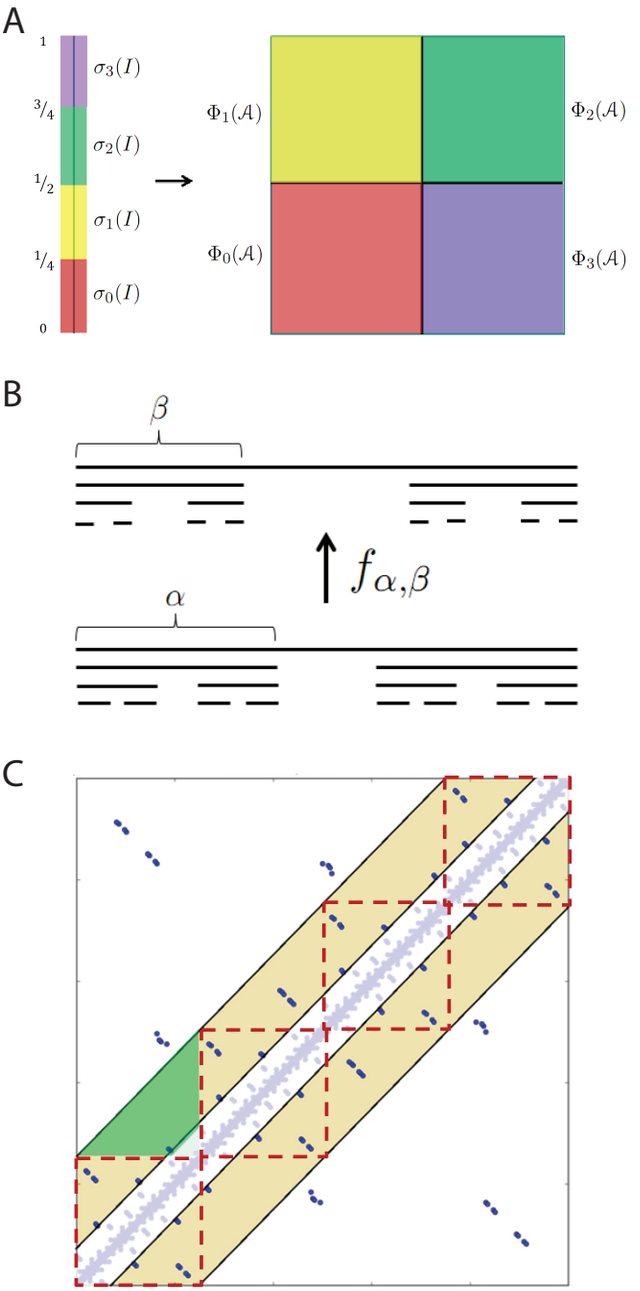


Fig. S4. Construction of the Hilbert curve using an iterated function system. **(A)** Subdivide the unit interval $I = [0, 1]$ into four equal parts. The interval $\sigma_i(I)$ is mapped to the square $\Phi_i(A)$ under the Hilbert curve. In general, by choosing different $\{\sigma_i\}$ and $\{\Phi_i\}$, many different self-similar curves can be defined. **(B)** The homeomorphism $f_{\alpha,\beta}$ maps the Cantor set with ratio α to the Cantor set with ratio β, though the two sets have different dimension. For this reason, the condition that a self-similar curve is balanced is necessary for the uniform dimension scaling result to hold. **(C)** Contact map of the Hilbert curve. In the notation of the proof of Lemma 3.5: J_n^{n-1} is highlighted in yellow, the sets $P_1, ..., P_4$ are outlined by the dotted red squares, and the set Q_1 is highlighted in green.

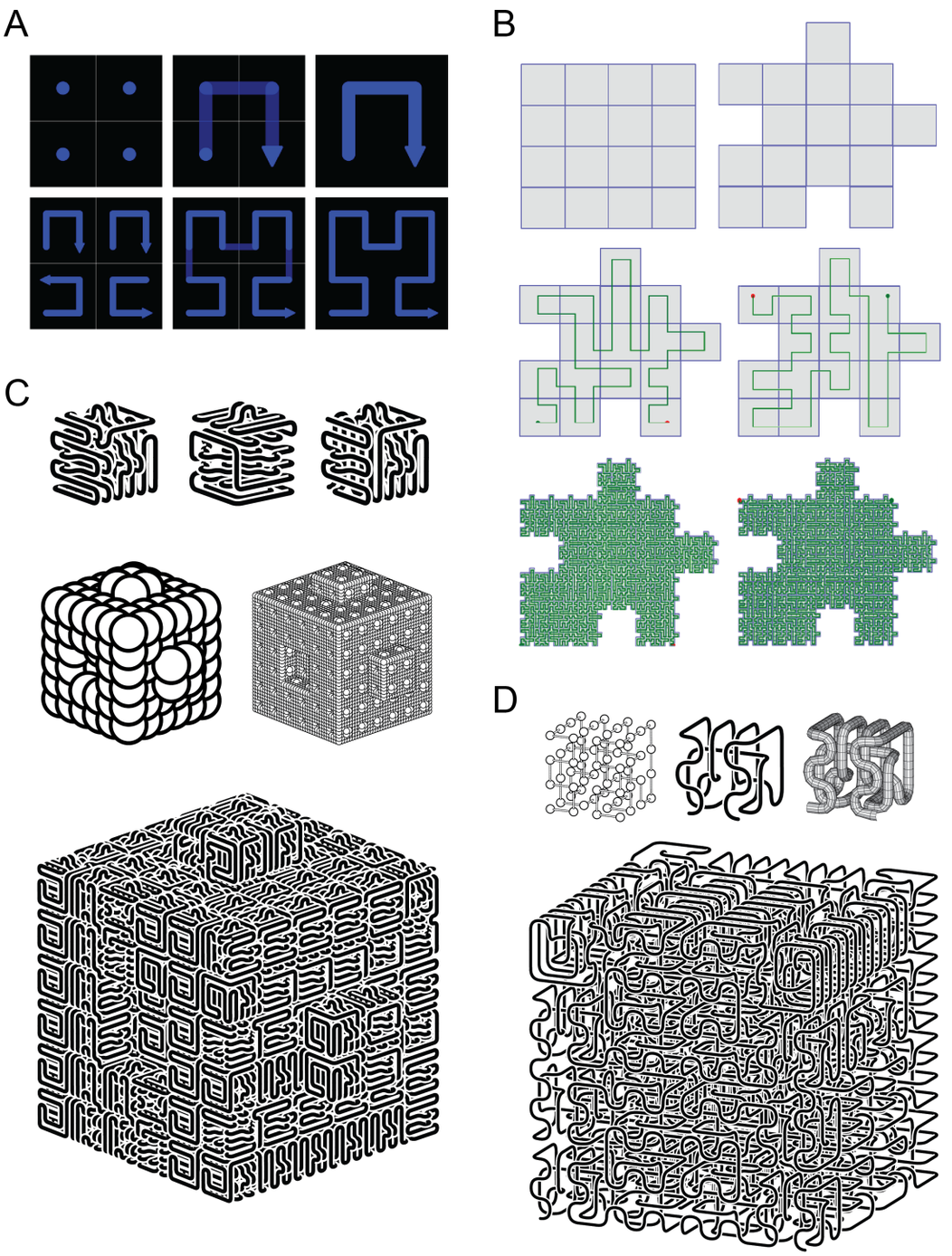


Fig. S5. Motif-based constructions of space-filling curves through rep-tiling. **(A)** Construction of the Hilbert curve. *Top*: The square is divided into four sub-tiles equal to the whole, and a Hamiltonian path is chosen through them. *Bottom*: one suitably-oriented copy of the base motif is drawn in each sub-tile and the end of one motif is connected to the beginning of the next. This defines a rewrite rule which can be iteratively applied to construct arbitrarily long paths. **(B)** *Top*: A simple reciprocal tessellation in which subsquares are moved from one side to the other, creating interlocking tiles. *Middle*: motifs for the 2D Jigsaw curve. Two motifs are required to compensate for the lost symmetry in the tile; by combining the two motifs, it is possible to iterate indefinitely. *Bottom*: second iteration; two different curves result depending on which motif is chosen initially. The start of each path is indicated by a green dot and the end by a red dot. **(C)** A 3D space-filling curve with three indentations, the 3D Jigsaw curve. This is the first example of a 3D space-filling curve with rough boundary – the surface dimension is 2.112. *Top*: three archetypal base-motifs. *Middle*: the first and second iterations drawn in a sphere-filling representation. *Bottom*: the second iteration, illustrating the rewrite rule. **(D)** A 3D space-filling curve, dubbed the "Gordian Knot", whose base-motif (shown above in three representations) contains a knot when the endpoints are joined. Higher iterations of the curve (second iteration shown below) contain arbitrarily many trefoil knots.

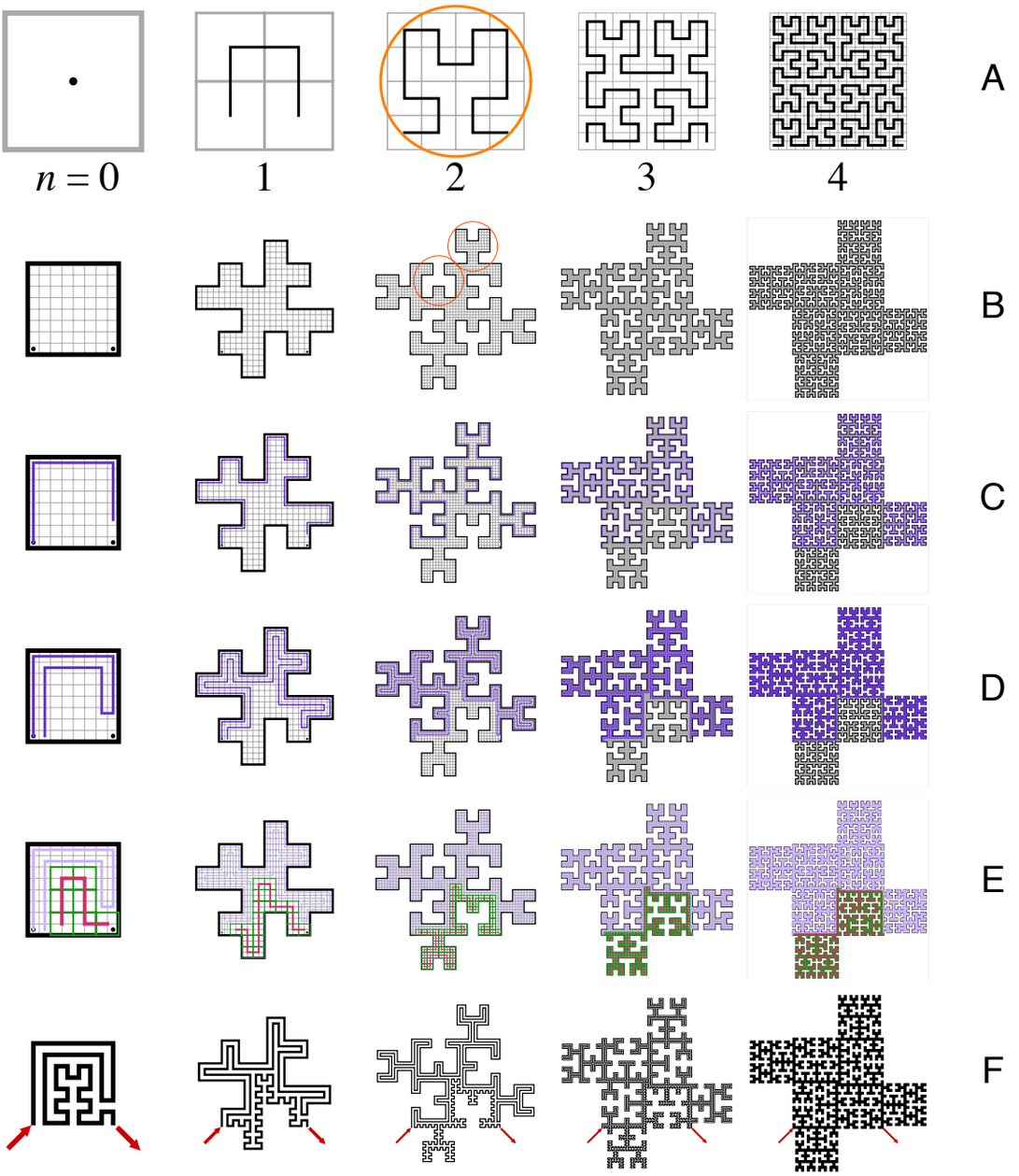


Fig. S6. Construction of the Inside-Out Hilbert Curve.
Each space-filling curve constructed this way has unit area, but the dimension of the resulting tile's border can be made arbitrarily close to 2.0, depending on which approximation path to the Hilbert Curve is used to build the prototile. **(Row A)** Pick a value of n and take the nth Hilbert Curve path. **(Row B)** Subdivide a unit square and mark the lower left and lower right subsquares as start and end points respectively. Along each side of the square, use two copies of the chosen Hilbert path to mold the square so that one half-side extrudes outwards and the other half-side is carved inwards. The resulting shape, called a prototile, has rotational wallpaper symmetry and unit area. **(Rows C - F)** Construct a Hamiltonian path that visits all subsquares in the prototile in four steps. The Inside-Out Hilbert Curve of order n can then be constructed iteratively using this path as the base motif. **(Row C)** Beginning at the start square, follow the left-hand rule along the inside of the prototile?s boundary until the path hits the subsquare two above the end square. **(Row D)** Make a U-turn, then follow the right-hand rule back through unvisited subsquares until the path returns to the subsquare directly to the right of the start square. **(Row E)** Treat 2×2 groups of the remaining subsquares as super-subsquares and draw a partial path through them. **(Row F)** Place a Hilbert base motif into each super-subsquare, suitably oriented, to complete the path. Path starts and ends at the red arrows.

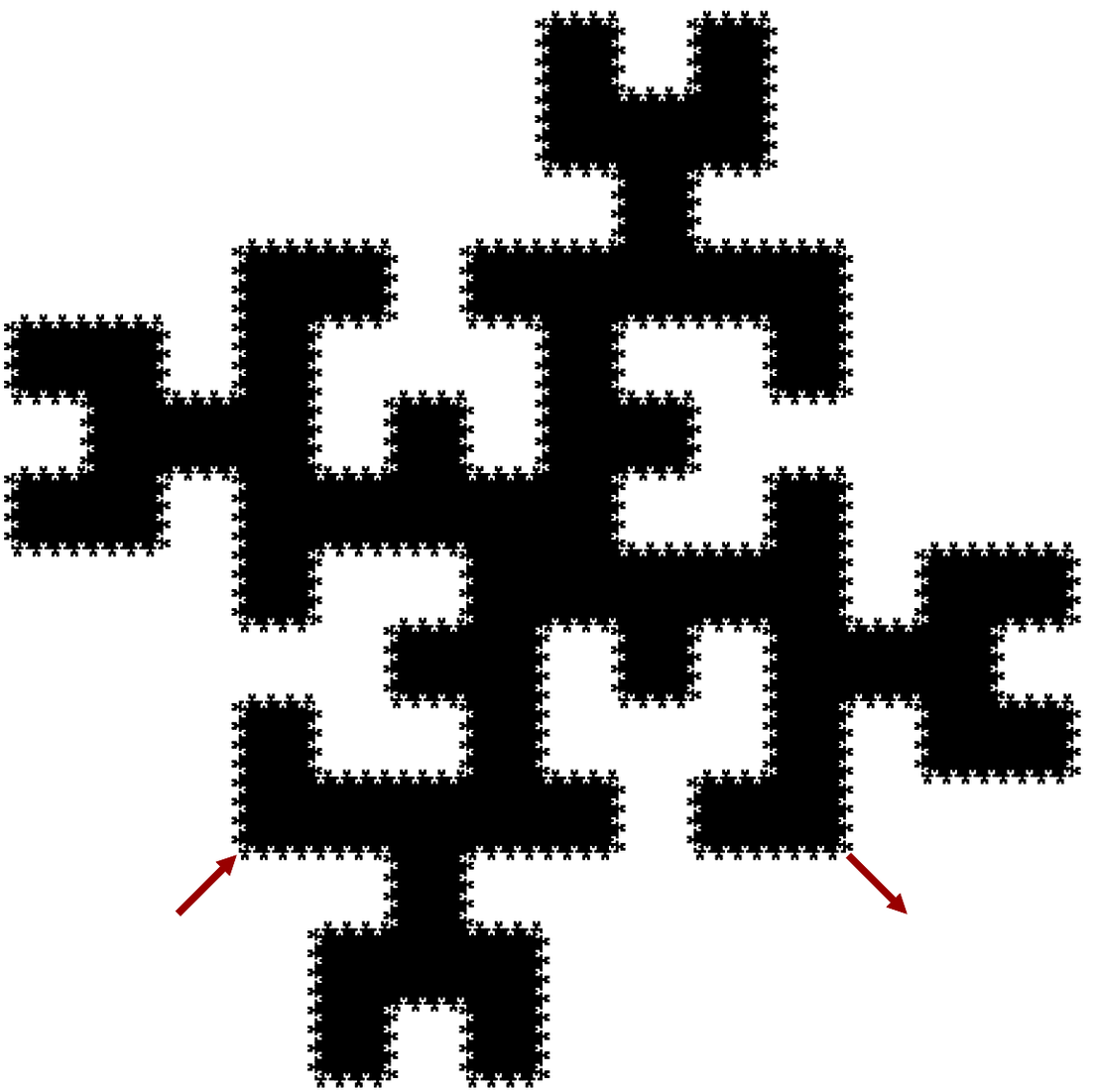
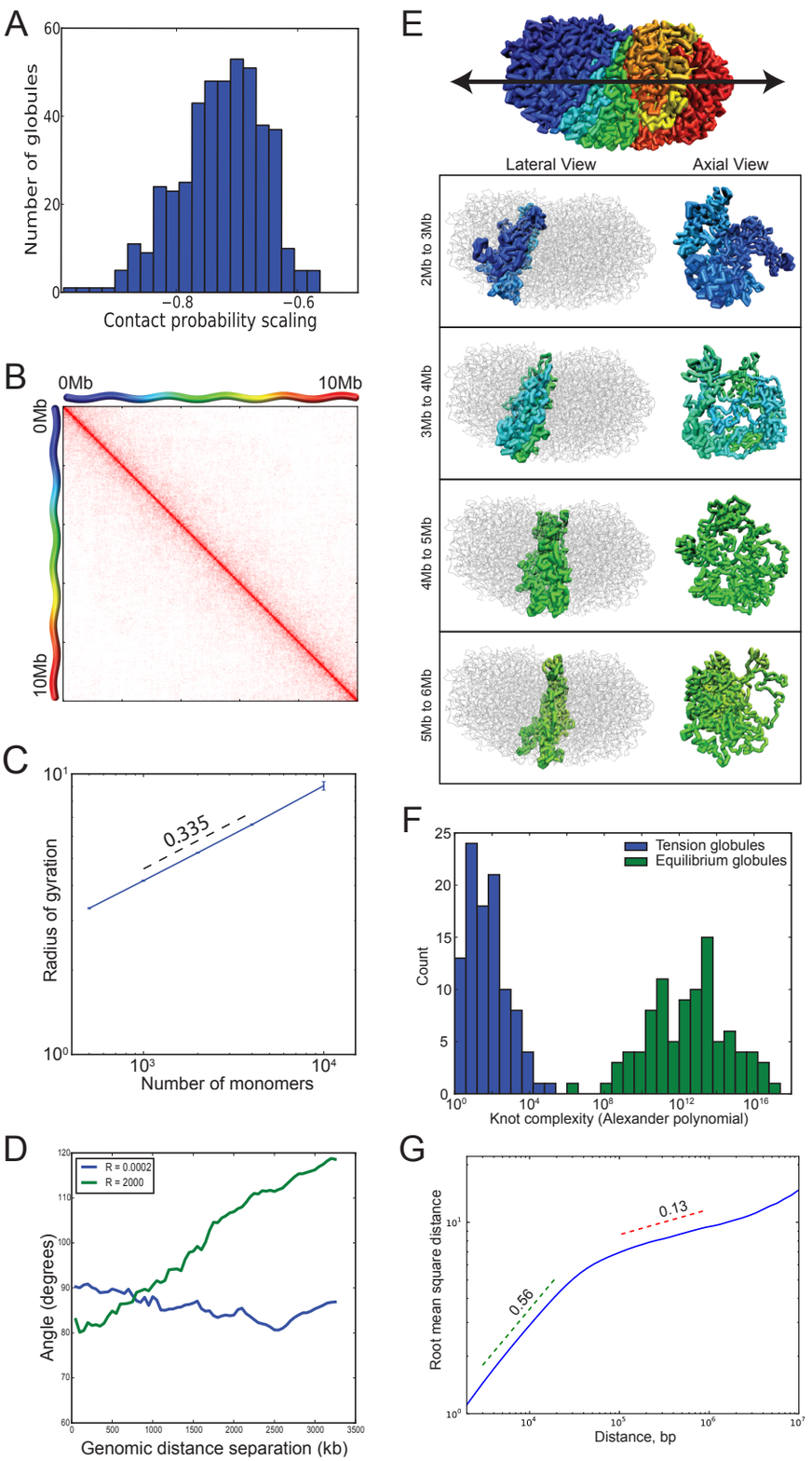


Fig. S7. The Inside-Out Hilbert Curve, second order, second iteration.
The self-avoiding, space-filling path containing 1,048,575 segments that outlines the second-order Inside-Out Hilbert Curve. Second iteration of the motif for $n = 2$ in Figure SS6. Path starts and ends at the red arrows. Zoom in to see additional detail.

Fig. S8.
Properties of the tension globule. **(A)** Histogram of contact probability exponents of 450 individual tension globules of length 10Mb (shown in Fig 4C) measured between distances of 15kb and 1Mb. **(B)** Contact map for the tension globule, aggregated over 450 simulations. **(C)** Average radius of gyration as a function of polymer length scales with exponent around 1/3, indicating a dense spherical structure. Error bars show standard deviation. **(D)** Tension globules have a linear axis at large scales. For tension globules ($R = 2000$) and the fractal globules ($R = 0.0002$) of length 10Mb, angles between the centroids of three consecutive regions of contour length L are plotted as a function of L, for $L < N/3$. At large distances, obtuse angles predominate in the tension globule. Angle measurements are averaged over each of the $N/L-2$ positions in the globule and over 100 simulation replicates. **(E)** A simulated 10Mb tension globule is shown, and 1Mb sub-regions are highlighted in lateral and axial views. Sub-regions tend to form flat slices, stacked along the linear axis. **(F)** The distribution of the determinant of the Alexander polynomial, a knot invariant which characterizes the degree of complexity of the knot, computed for 100 tension globules (blue) and 100 equilibrium globules (green). Globules have 4000 monomers each and endpoints are joined to create closed contours. Small values indicate lower complexity. **(G)** Root mean square end-to-end distance for the tension globule, as a function of genomic distance, exhibits scalings with exponent ≈ 0.56 between 2kb and 20kb and 0.13 between 100kb and 1Mb.



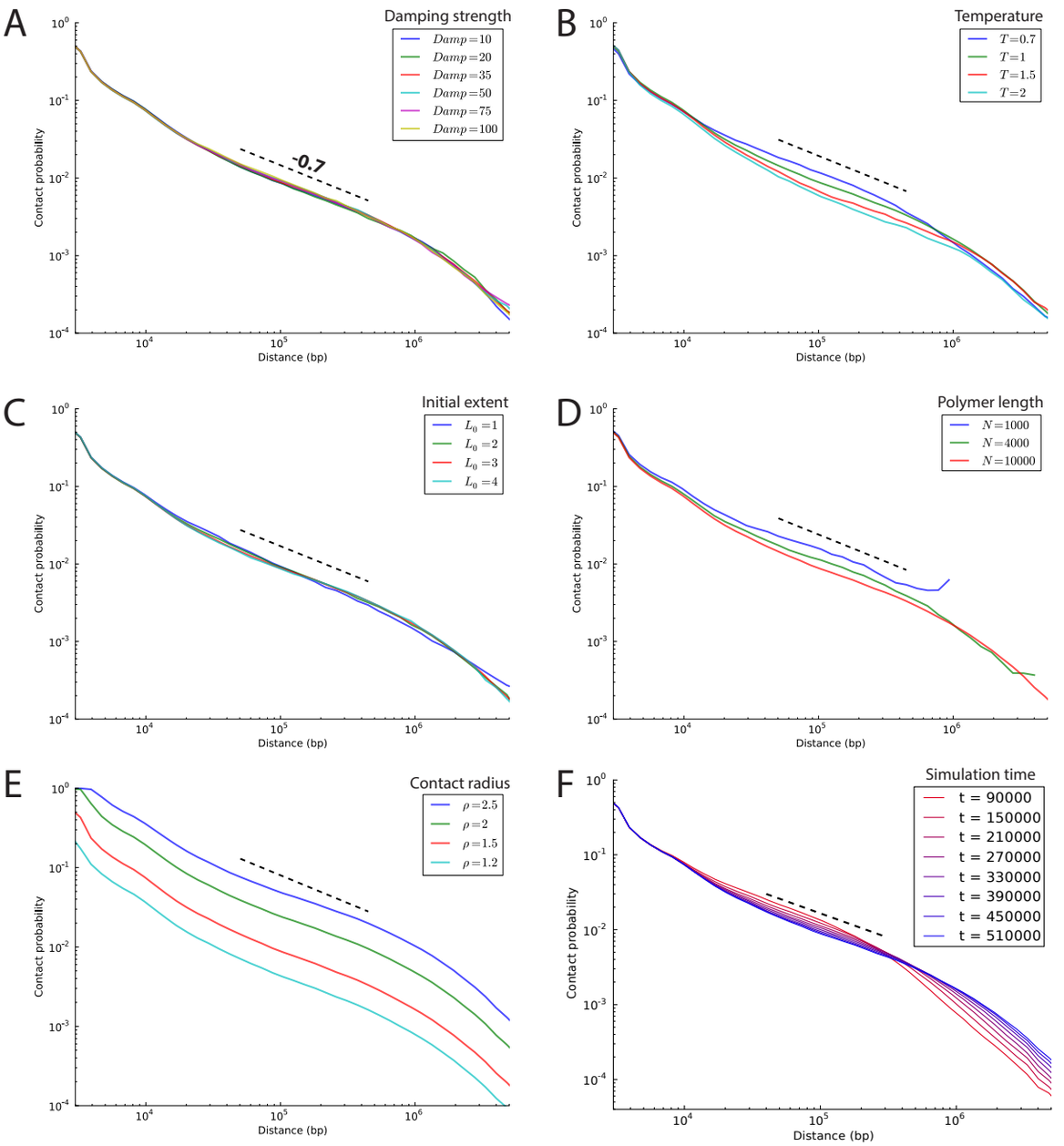


Fig. S9. Tension globule contact probability is robust to changes in simulation parameters. Contact probability plots for tension globules simulated with Lennard-Jones forces and varying parameter values. Dotted line in all plots shows a reference slope of -0.7. **(A)** Changes in damping strength, measured in time units, do not affect contact probability. **(B)** Increasing temperature decreases mid-range contacts but does not substantially affect the scaling. **(C)** Changes in the extendedness of the initial self-avoiding walk position do not affect contact probability. **(D)** Contact probability scalings are consistent for a range of polymer lengths. **(E)** Increasing the distance threshold for contacts increases the raw number of contacts counted but does not substantially affect the scaling. **(F)** Early in collapse, a scaling of $\gamma \approx 0.7$ emerges at short distances. After full collapse, the scaling extends to larger distances and does not change significantly with increased simulation time.

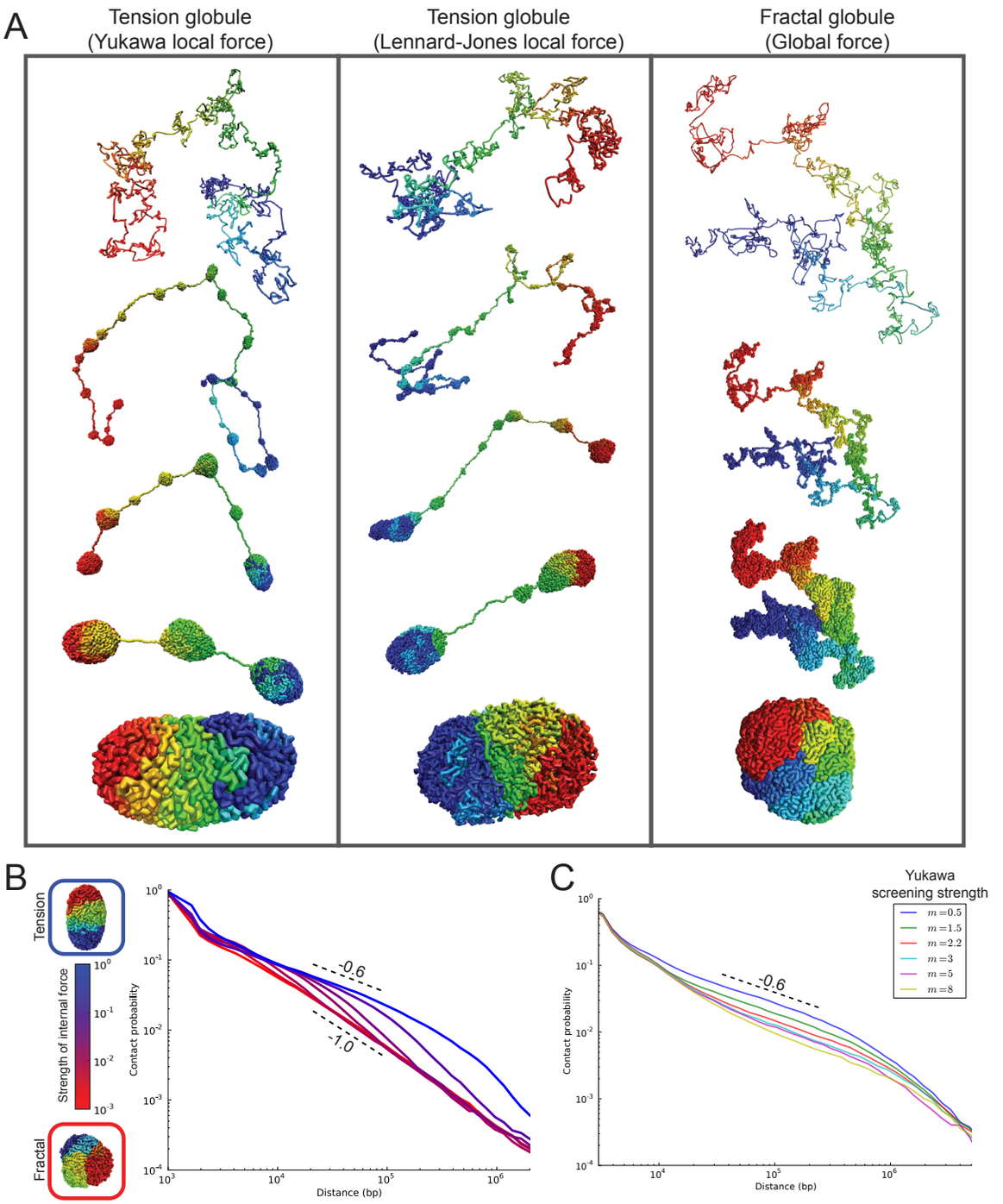


Fig. S10. Comparison of the fractal globule and tension globules with Lennard-Jones and Yukawa potentials. **(A)** Collapse process and contact probability scaling for the tension globule and the fractal globule. *Left:* tension globule formed with inter-monomeric attractive forces modeled by the Yukawa potential. *Middle:* tension globule formed with inter-monomeric attractive forces modeled by the Lennard-Jones potential. *Right:* fractal globule formed with a global crowding forces, modeled by a weak spring potential drawing all monomers towards a central position. **(B)** Contact probability of the fractal-tension transition simulated using Yukawa forces. Strength of the external force is held constant while strength of the Yukawa internal force is varied. When internal forces are weak, collapse is fractal (red, $\gamma \approx 1.0$); when internal forces are strong, collapse is tension-driven (blue, $\gamma \approx 0.6$). **(C)** Contact probability of tension globules with Yukawa forces. Scaling is consistent across a wide range of screening strengths.

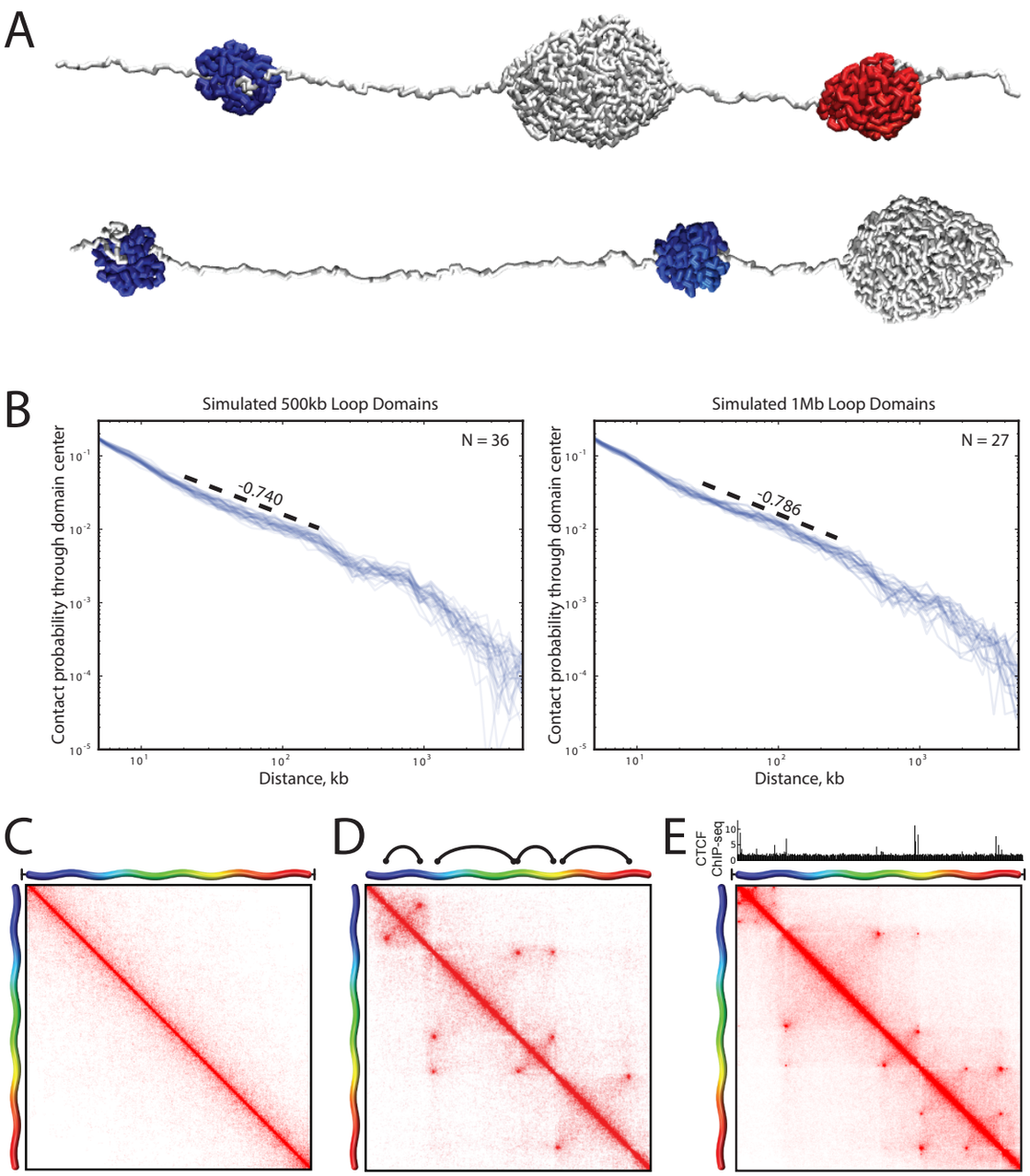


Fig. S11. Loops introduced during tension globule collapse cause domains to form between the loop anchors. **(A)** Two examples of tension globule loop simulations with length 3Mb and endpoints tethered 100 distance units apart. Loops are formed as shown in Fig 4D. In each structure, the region between select pairs of loop anchors is highlighted in color; between replicates, sub-globules form in different locations but are frequently anchored by loops. **(B)** Contact probability in a 100kb window through the center of simulated 500kb (left) and 1Mb (right) tension globule loop domains exhibit scalings of $\gamma \approx 0.75$ extending to the domain boundary, recapitulating Hi-C measurements. **(C)** Contact map for a 3Mb tension globule with tethered endpoints but no loops. Because sub-globules form at different locations, no domains emerge (compare to Fig S8B). **(D)** Contact map for non-tethered simulations with the same loops as Fig 4D. **(E)** Contact map for simulations in which loops were chosen with probabilities based on CTCF ChIP-seq data as described in Section III.b.ii.

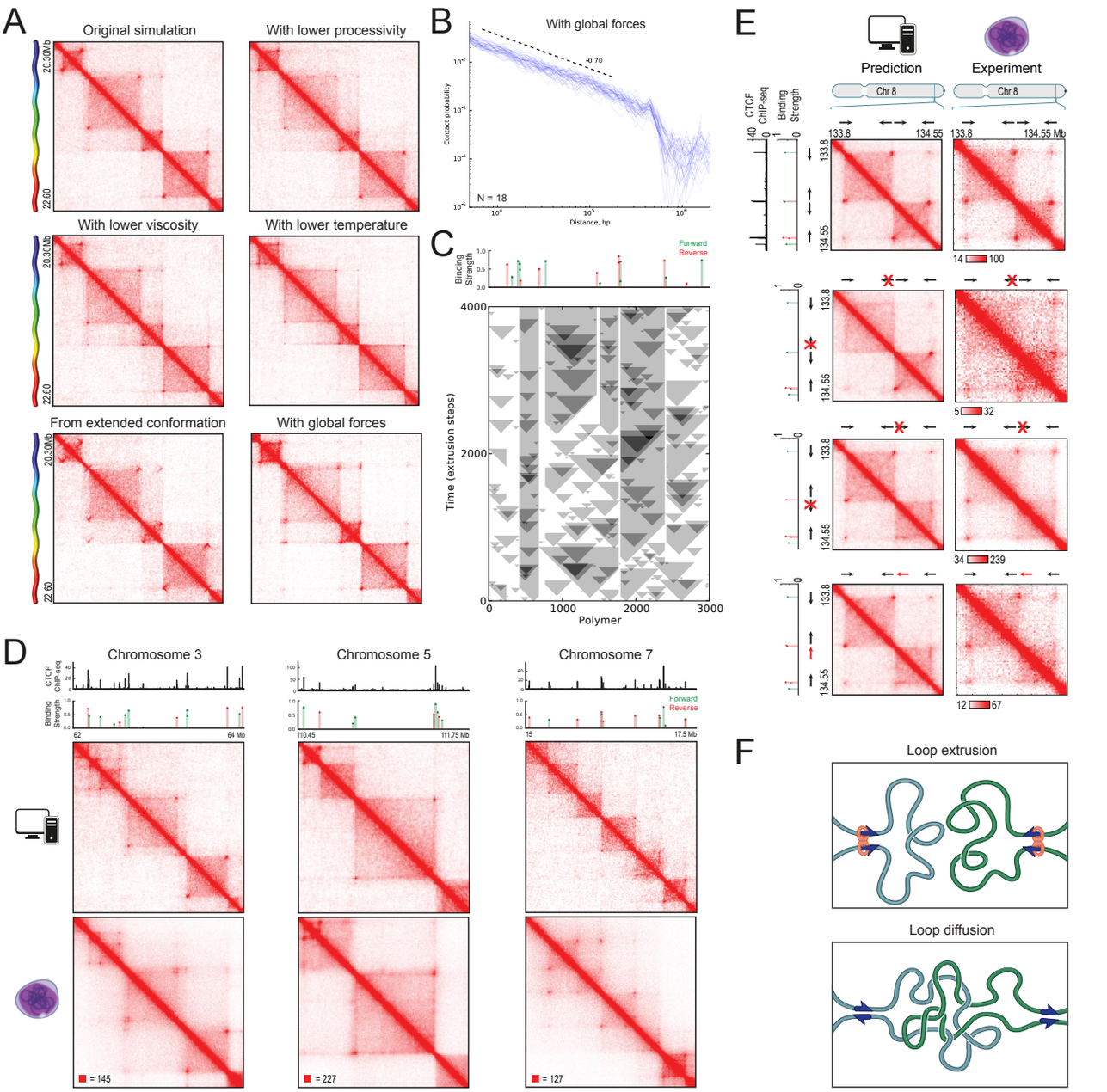


Fig. S12. The extrusion model robustly recapitulates Hi-C contact maps.
(A) Extrusion simulations performed with lower viscosity ($t_{damp} = 30$), with lower temperature ($T = 1.0$), starting from an extended polymer, or using a global potential instead of inter-monomeric forces ($T = 0$) are nearly identical to the original simulation shown in Fig 5D ($T = 2.0$, $t_{damp} = 10$). All simulations recapitulate Hi-C data well. **(B)** Contact probability in a 100kb slice through the center of 18 domains of size 980kb simulated using the extrusion model and global crowding forces. Although global forces produce a scaling around $\gamma \approx 1$ in the absence of other effects, loop extrusion yields a scaling with $\gamma \approx 0.7$. **(C)** A representative schematic of the loop extrusion dynamics of simulations shown in Fig 5D, showing the association, extrusion, binding, and dissociation of extrusion complexes over time. At time T, if an extrusion complex has bound to points A and B on the polymer, the line from (T, A) to (T, B) is shaded gray. **(D)** Contact maps of three additional regions on chromosomes 3, 5, and 7 were robustly recapitulated by the extrusion model, simulated directly from CTCF ChIP-seq signals. A ChIP-seq normalization constant was the only free parameter. **(E)** Contact maps of wild-type and three genome engineering experiments from Fig 7A. The simulations of wild-type and each engineered condition were produced before any CRISPR experimental data was available and were based only on the wild-type contact map. These are *de novo* predictions in the strictest sense. **(F)** *Above:* Loop extrusion produces unentangled domains and loops at convergently-oriented CTCF motifs. *Below:* In the classic loop diffusion model, loops form when freely diffusing loop anchors encounter each other in 3D. This behavior would cause the intervening DNA to be highly entangled and would not show any orientation preference at the loop anchor CTCF motifs.

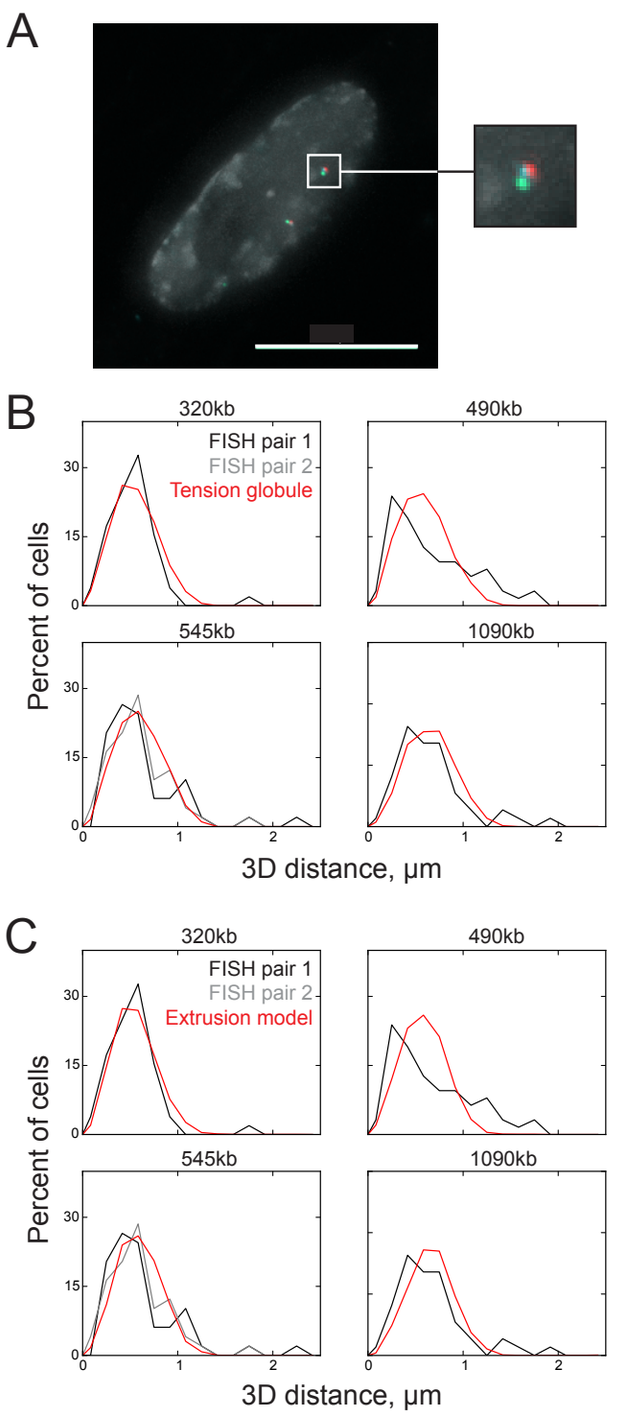


Fig. S13. Intra-domain distances measured by 3D-FISH match simulation results for both tension globules and the extrusion model. **(A)** Sample 3D-FISH measurement. Genomic loci are marked in red, blue, and green; DAPI stain is in gray, highlighting the nucleus. Scale bar is 15m. **(B, C)** Distributions of 3D distances between pairs of loci obtained experimentally using 3D-FISH (black, gray) and in simulated tension globules (B) or simulated extruded domains (C) for four different genomic distances from 320kb to 1090kb. Agreement of simulation to experiment (Kolmogorov-Smirnov statistic with tension globule: 0.15; K-S statistic with extrusion model: 0.16) is as good as agreement between two experimental measurements (K-S statistic: 0.18).

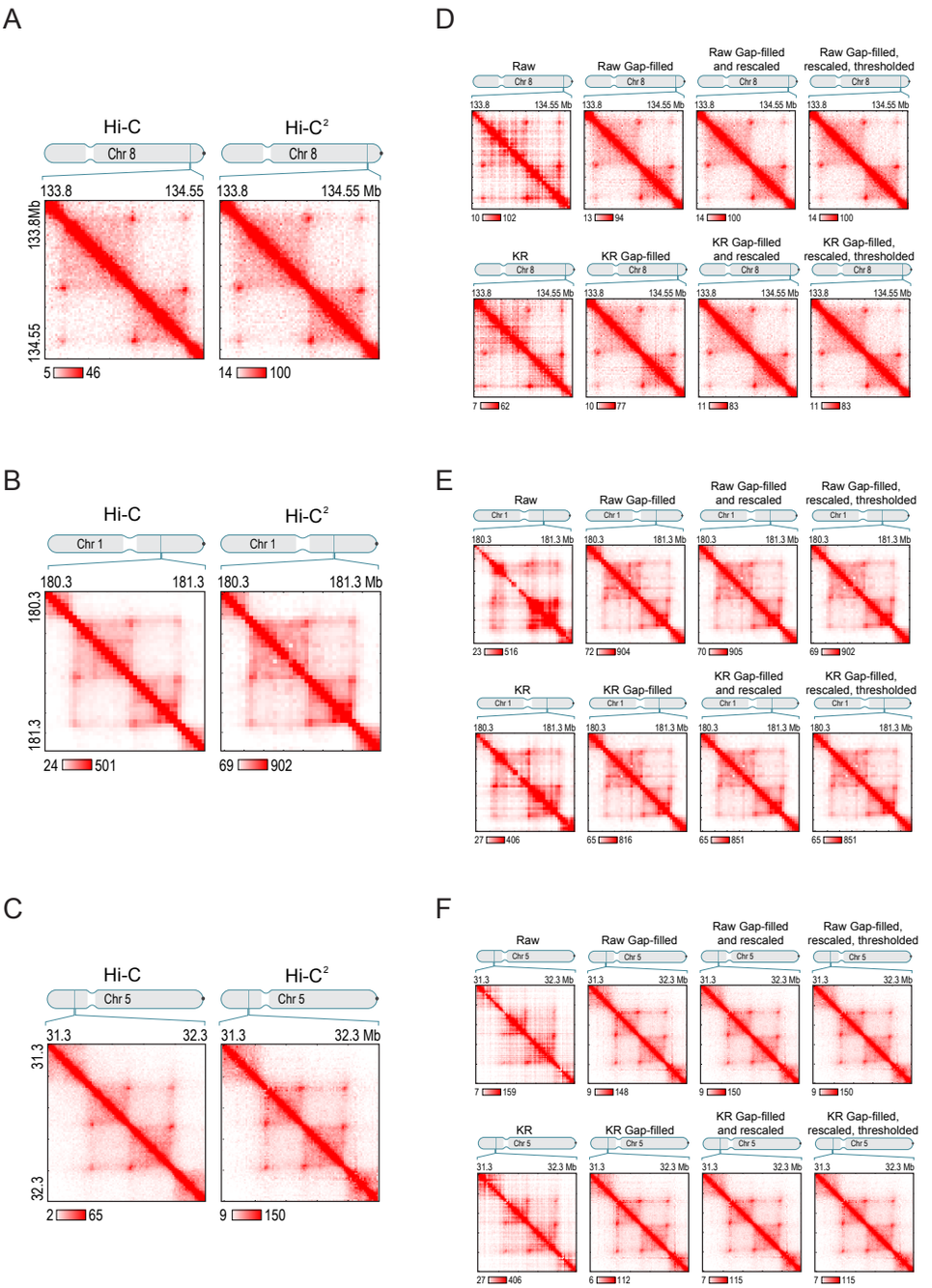


Fig. S14. Analysis and Validation of Hi-C^2.
(A,B,C) The *in situ* Hi-C contact map of all regions probed in this study (region 1: chr 8:133.8-134.55Mb, **(A)**; region 2: chr1:180.3-181.3Mb, **(B)**; region 3: chr5:31.3-32.3Mb, **(C)**) in wild-type HAP1 cells (left) closely resembles a Hi-C^2 contact map for the same region (right). **(D,E,F)** Hi-C^2 data shown for region 1 (D), region 2 (E), and region 3 (F) using different flavors of normalization. Raw unnormalized data is shown in the top left corner and labeled as "Raw". All different normalization methods are detailed in section I.e.iv. The method "Raw Gap-filled, rescaled, thresholded" was used for all Hi-C^2 data shown in the main figures.

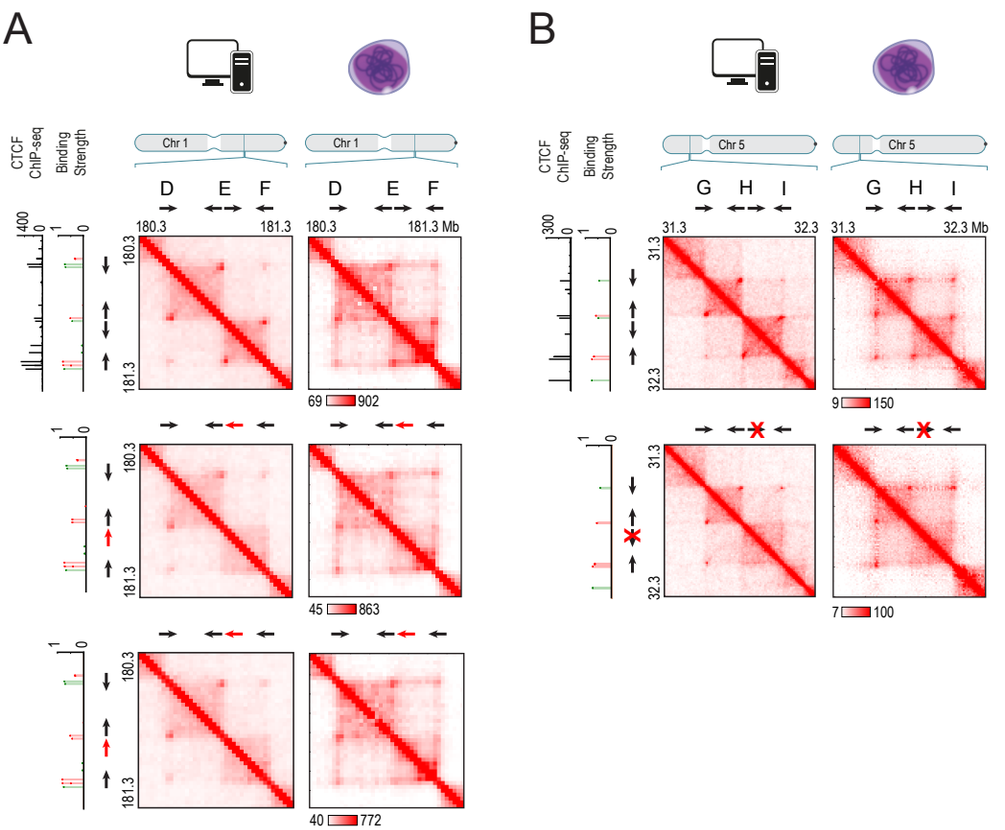
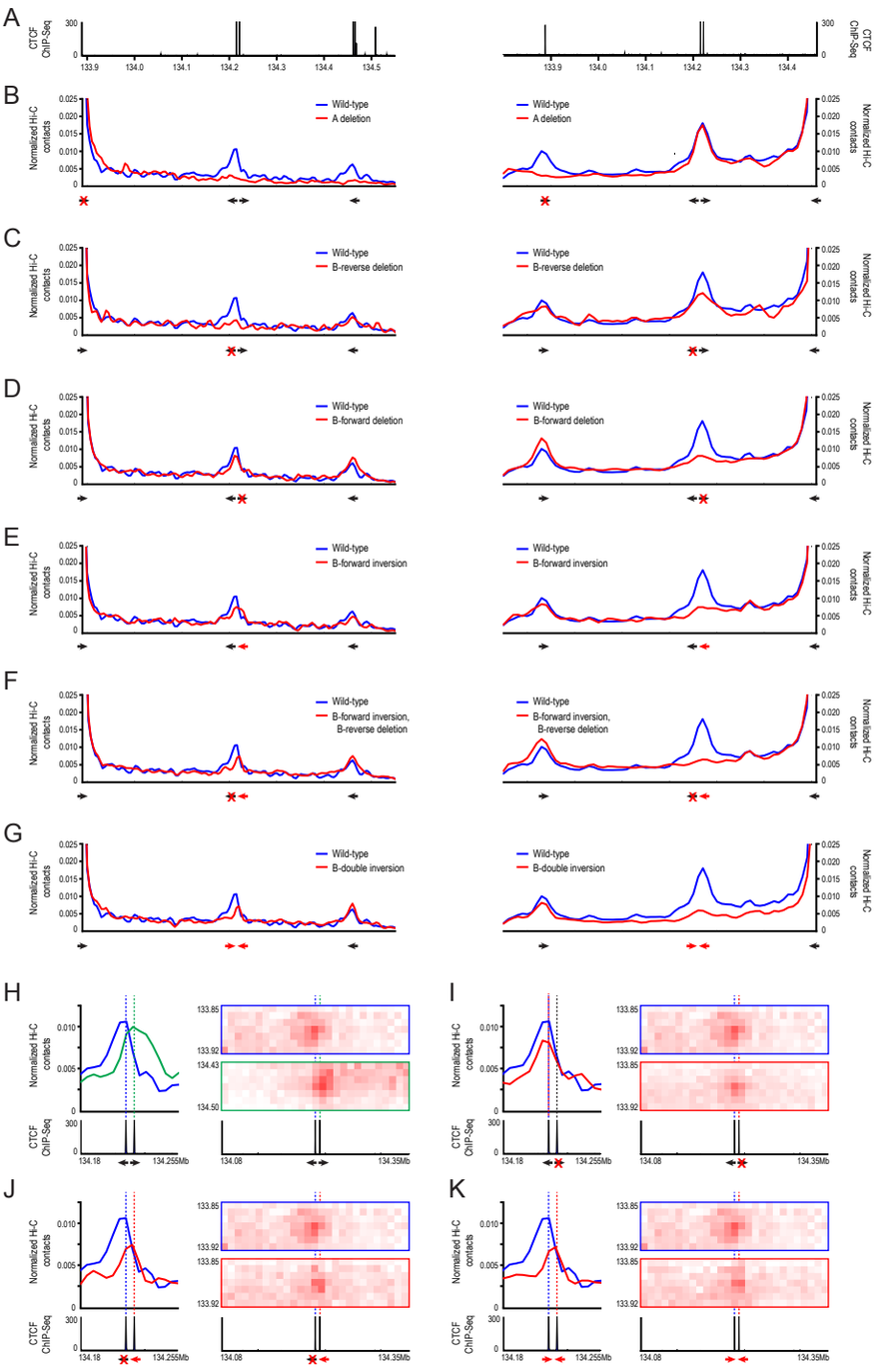


Fig. S15. Additional CRISPR mutants.
(A) Two additional CRISPR mutants for Region 2 (chr1:180.3-181.3 Mb). First row: The contact map for the wild-type locus, calculated using *in silico* simulations (left), closely matches the map observed using Hi-C^2 experiments (right). Second row: Replacement of the E/forward motif with the E/reverse motif eliminates the E-F loop. Third row: Inversion of the E/forward motif eliminates the E-F loop. **(B)** An additional CRISPR mutant for Region 3 (chr5:31.3-32.3 Mb). First row: The contact map for the wild-type locus, calculated using *in silico* simulations (left), closely matches the map observed using Hi-C^2 experiments (right). Second row: Deletion of the H/forward motif eliminates the H-I loop.

Fig. S16.
Hi-C and Virtual 4C experimental data reflect accuracy of predictions based on the convergent rule, including both movement of individual loop anchors and disruption of entire loops.
(A) GM12878 ChIP-Seq data is shown as a reference for region 1. **(B)** Virtual 4C plots of the A/forward deletion CRISPR mutant Hi-C2 data. For panels (B-G), the left column is always anchored at chr8:133,885,000 (locus A, 5kb resolution data shown) and the right column is always anchored at chr8:134,460,000 (locus C, 10kb resolution data shown). When A/forward is deleted, the A-B and A-C loops disappear, but the B-C loop is unaffected. **(C)** Deletion of B/Reverse eliminates the A-B loop. **(D)** Deletion of B/Forward eliminates the B-C loop. **(E)** Inversion of B/Forward eliminates the B-C loop. **(F)** Simultaneous deletion of B/Reverse and inversion of B/Forward eliminates the B-C loop. **(G)** Inversion of both B/Forward and B/Reverse does not eliminate loops. **(H)** Virtual 4C and heatmap blowouts of wild-type Hap1 Hi-C^2 data anchored at the A locus (blue) and at the C locus (green). The shift in the peak intensity of the A-B and B-C loop signals at B mirrors the 6kb distance between the B/reverse and B/forward motifs (ChIP-Seq). **(I)** Virtual 4C and heatmap blowouts of wild-type Hap1 Hi-C^2 data (blue) and B/forward deletion Hi-C^2 data, both anchored at the A locus. Both of the A-B loops shown are anchored at the B/reverse location and no shift in intensity of signal is seen. **(J)** Virtual 4C and heatmap blowouts of wild-type Hap1 Hi-C^2 data (blue) and B/reverse deletion + B/forward inversion Hi-C^2 data, both anchored at the A locus. In this case, the A-B loop in the mutant case forms between A and the inverted B/forward motif, and the shift in loop intensity to the new location is visible. **(K)** Virtual 4C and heatmap blowouts of wild-type Hap1 Hi-C^2 data (blue) and B/reverse and B/forward double inversion, both anchored at the A locus. In this case, the A-B loop in the mutant case forms between A and the inverted B/forward motif, and the shift in loop intensity to the new location is visible.



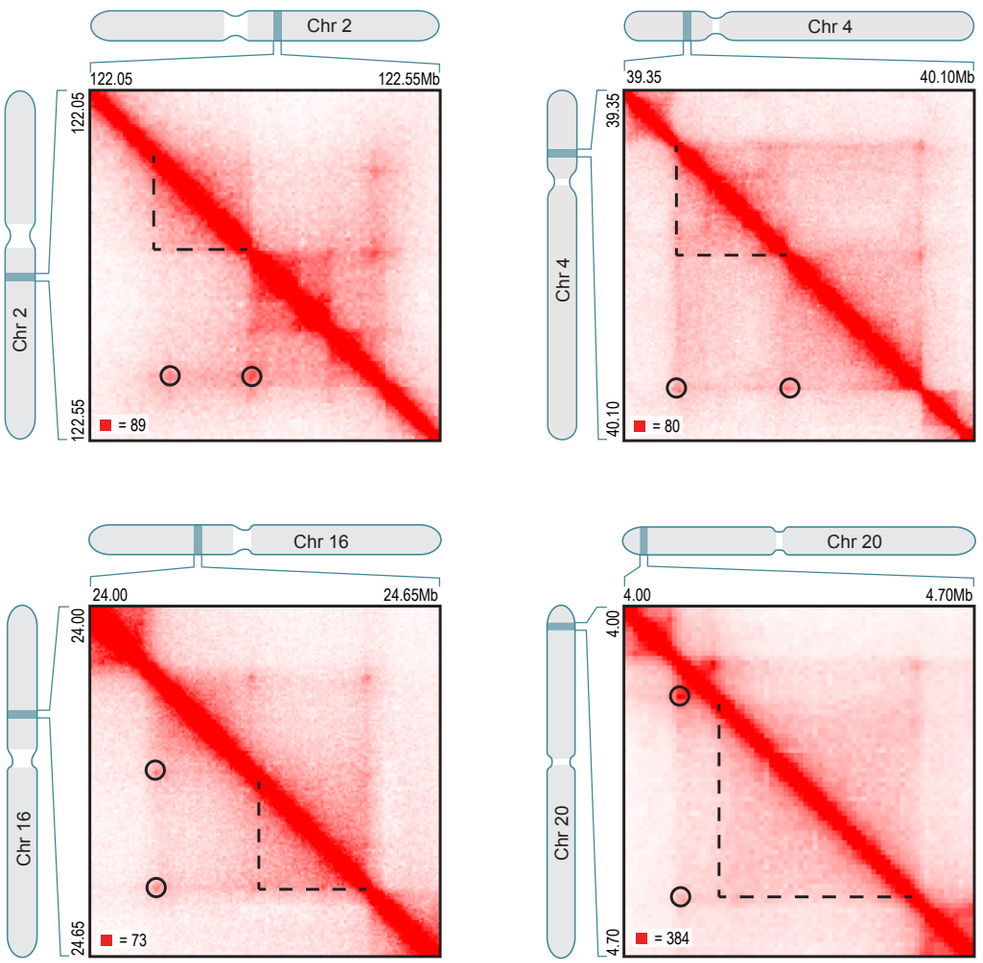


Fig. S17. Contact domains can form between consecutive loop anchors that do not loop to one another. When three consecutive loci, A, B and C, form two loops such that A loops to B and C but B and C do not loop to each other, the extrusion model predicts that a domain will still form between B and C. This is because an extrusion complex that lands between B and C will be excluded from the A-B region by the A-B loop; instead, it tends to bring points within the B-C interval together, forming an "exclusion domain." When we examined the wild-type GM12878 Hi-C map for such loci, we found that exclusion domains were prevalent: the B-C region in 158 cases coincided with an annotated contact domain, a 6.3-fold enrichment (loci near compartment flips were excluded). Four example exclusion domains are shown here, with the A-B and A-C loops circled and the exclusion domain outlined.

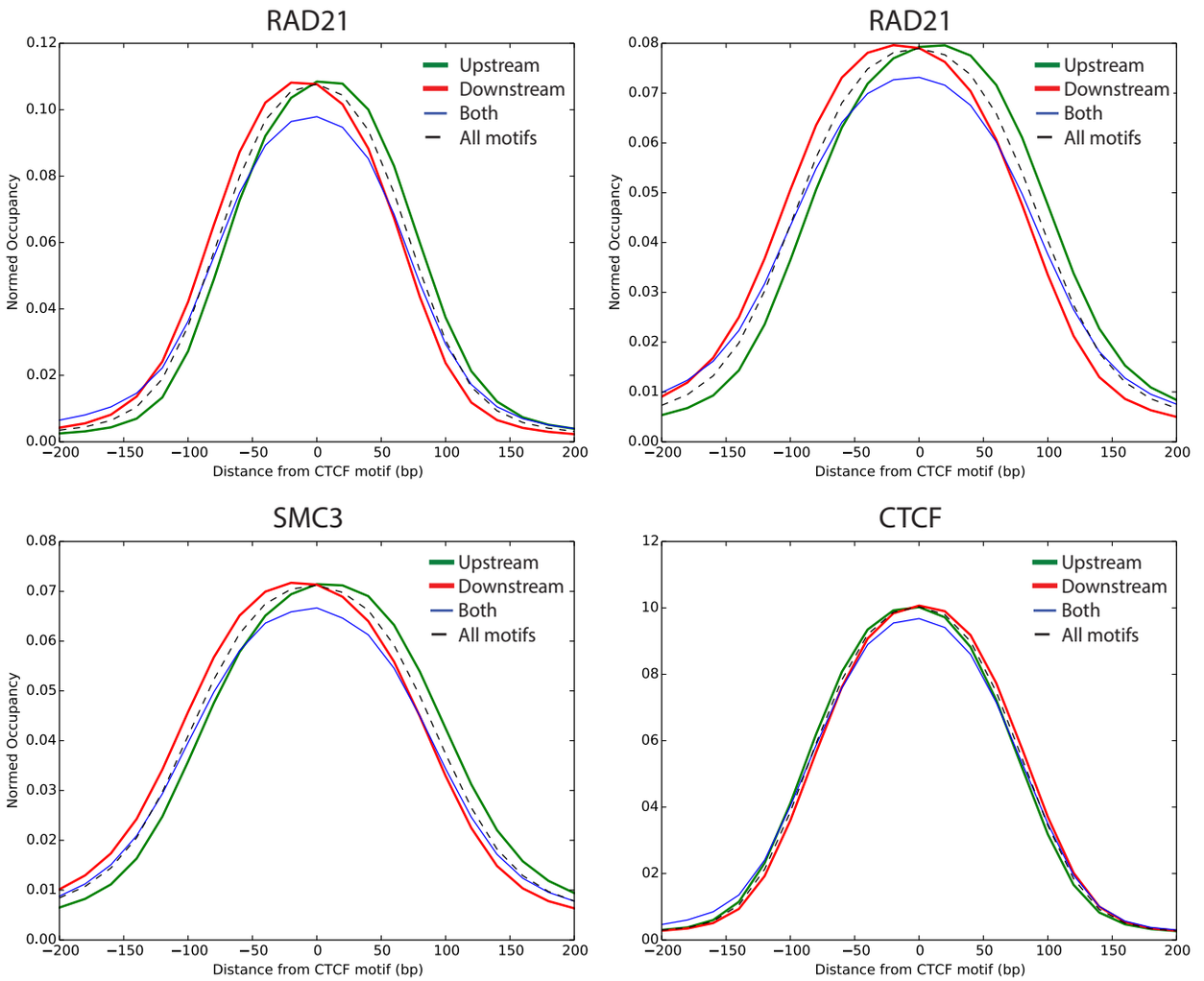


Fig. S18. Interactions of RAD21 and SMC3 at loop anchors are shifted ≈ 20bp towards the loop interior. ChIP-seq signal for RAD21 (two replicates), SMC3, and CTCF interactions aggregated over all unique CTCF motifs at upstream loop anchors, at downstream loop anchors, at both loop anchors, or all CTCF motifs. Interactions of RAD21 and SMC3 are shifted roughly 20bp towards the loop interior.