A multicolour polymer model for the prediction of 3D structure and transcription in human chromatin

Massimiliano Semeraro[†],[1] Giuseppe Negro[* †],[2] Antonio Suma,[1]
Giuseppe Gonnella,[1] Peter R. Cook,[3] and Davide Marenduzzo[2]

[1] *Dipartimento Interateneo di Fisica, Università degli Studi di Bari and INFN, Sezione di Bari, via Amendola 173, Bari, I-70126, Italy*
[2] *SUPA School of Physics and Astronomy, University of Edinburgh, Peter Guthrie Tait Road, Edinburgh EH9 3FD, UK*
[3] *Sir William Dunn School of Pathology, University of Oxford, OX1 3RE, UK*

Within each human cell, different kinds of RNA polymerases and a panoply of transcription factors bind chromatin to simultaneously determine 3D chromosome structure and transcriptional programme. Experiments show that, in some cases, different proteins segregate to form specialised transcription factories; in others they mix together, binding promiscuously the same chromatin stretch. Here, we use Brownian dynamics simulations to study a polymer model for chromosomes accounting for multiple types ("colours") of chromatin-binding proteins. Our multi-colour model shows the spontaneous emergence of both segregated and mixed clusters of chromatin-bound proteins, depending mainly on their size, thereby reconciling the previous experimental observations. Additionally, remarkable small-world networks emerge; in these, positive and negative correlations in activities of transcription units provide simple explanations of why adjacent units in large domains are co-transcribed so often, and how one eQTL (expression quantitative trait locus) can up-regulate some genes and down-regulate others. We also explain how local genome edits induce distant omnigenic and pangenomic effects, and develop ways to predict activities of all transcription units on human chromosomes. All results point to 1D location being a key determinant of transcription, consistently with the conservation of synteny seen between rapidly-evolving enhancers and their more stable target genes.

INTRODUCTION

Microscopy and high-throughput sequencing have uncovered a rich hierarchy of 3D folding in eukaryotic nuclei that ranges from loops of tens of bp up to hundreds of kbp, through topologically-associating domains (TADs) of $100 - 1000$ kbp, AB compartments, heterochromatic aggregates, and on to chromosomal territories [1–3]. An outstanding question is whether, and to what extent, this rich multi-scale organisation is functionally related to, or even driven by, transcription [4]. On the one hand, it is widely believed that TADs remain largely invariant in cells with very different transcriptional programs – which points to little role for transcription in determining structure [5] (for an opposing view, see [6]). On the other hand, clusters of active polymerases – called phase-separated condensates, hubs, and transcription factories [4, 7–10] – locally stabilise surrounding clouds of loops, thereby providing an example of a structural unit with a clear functional role.

Brownian-dynamics simulations point to a simple and generic mechanism – the bridging-induced attraction – that spontaneously drives formation of many higher-order structures [11, 12]. Reversible binding of multivalent transcription factors (TFs) – or TF:polymerase (TF:pol) complexes – to transcription units (TUs) scattered along a string of non-binding beads inevitably leads to clustering of bound TF:pol complexes without any additional energy input; in turn, these clusters organise TADs and AB compartments. Such microphase separation of active polymerases and TFs depends on positive feedback and is eventually arrested by entropic costs associated with crowding and looping more and more DNA [13]. When these simulations involve 2 different kinds (or "colours") of TF (e.g., red and green ones) binding specifically to red and green TUs, resulting clusters often contain bound TFs of just one colour, rather than mixtures [12]. Formation of such distinct clusters mimics the presence of distinct factories that specialize in transcribing different sets of genes. For example, active forms of RNA polymerases II and III are each housed in distinct nucleoplasmic factories that make genic and snRNA transcripts, respectively [7, 14, 15]. Similarly, distinct ERα, KLF1, and NFκB factories specialize in transcribing genes involved in the estrogen response, globin production, and inflammation [16–18]. One important consequence of this organisation is the creation of 3D networks, in which genes sharing the same TFs are co-transcribed in the same clusters, and so give contacts detected by techniques like Hi-C (a high-throughput variant of chromosome conformation capture), genome architecture mapping, and pore-C [2, 19–21].

The existence of clusters prompts the question: does any one cluster mainly contain just one kind of TF, or many different ones? In contrast to results cited above supporting the former possibility, others point to the latter. For example, chromatin immuno-precipitation (ChIP) shows that a promoter of a typical active human gene binds many different TFs, and that highly-

occupied targets (HOTs) are bound promiscuously by many TFs irrespective of whether or not they encode the specific binding motif for those factors [22–24]. Additionally, single-cell transcriptional profiling points to expression levels varying continuously as cells differentiate into other cell types, which points to a complex interplay between many factors, rather than a few acting as binary switches [25, 26].

Here, we simulate strings of beads (representing short chromosome fragments and whole human chromosomes in HUVEC and GM12878 cells) binding spheres of different colours (representing different TFs or TF:pol complexes). One key result is that specialised (demixed) and mixed clusters are not mutually exclusive: both emerge spontaneously depending on the 1D pattern of binding sites on strings, and this enables us to reconcile the apparently contrasting observations cited above. The size of emerging clusters correlates with the degree of mixing; specialised one-colour clusters are typically smaller. When we make the reasonable assumption that a bead is transcribed when it binds a TF, we find that cluster size and degree of mixing both determine transcriptional activities. Additionally, when we model just two colours of TF:pol complex that bind respectively to cell-type-invariant and cell-type-specific TUs in strings mimicking whole human chromosomes, we find transcriptional profiles that correlate well with those obtained by GRO-seq (global run-on sequencing [27, 28]).

MATERIALS AND METHODS

Chromatin fibres are modelled as bead-and-spring polymers [9, 11, 12, 29–38], where each monomer (diameter σ) represents 3 kbp packed into a 30 nm sphere [9, 11, 35]. Different TFs (or TF:pol complexes) are modelled as differently-coloured spheres (also with diameter σ) able to bind (when in an "on" state) to cognate sites of the same colour that are scattered along the polymer. Each TF and TF:pol switches between "off" (non-binding) and "on" (binding) states to reflect the post-translational modifications that occur in many TFs. Polymer beads are either non-binding ("heterochromatic"), weakly-binding ("euchromatic"), or strongly-binding (containing cognate sites). TFs bind non-specifically to all weakly-binding beads, and strongly only to TUs of the same colour. The system evolves in a cubic simulation domain with periodic boundary conditions through constant-temperature Langevin dynamics that are integrated numerically by the LAMMPS simulation package [39]. Hydrodynamic interactions are neglected for computational efficiency, and averages are evaluated over 100 independent runs for each case. The TF volume fraction of each colour is set to $\sim 3\ 10^{-5}$, and the polymer volume fraction to $\sim 2\ 10^{-3}$. More information about the model can be found in the SI.

Several quantities are monitored to describe the system's behavior. Mean transcriptional activity is measured as the fraction of time that a TU is "transcriptionally active" (i.e., within 2.25σ of a TF) in 100 simulations, and so represents a population average (each simulation run may be thought of as a different cell). This quantity is compared with experimental data on transcriptional activity, obtained via GRO-seq – a method providing a genome-wide average readout of ongoing transcription of both genic and non-genic TUs in cell populations [27, 28]. The mean transcriptional Pearson correlation between all pairs of TUs is also evaluated, and a graphical overview of this feature is provided via the Pearson correlation matrix. We also analyse clusters/factories of bound and spatially-proximate TFs, count the number of TFs of similar colour in each cluster, and introduce a demixing coefficient

$$\theta_{\rm dem} = \frac{n x_{i,max} - 1}{n - 1}, \tag{1}$$

where n is the number of colors, and $x_{i,max}$ the largest fraction of same-coloured TFs in a single TF cluster. If $\theta_{\rm dem} = 1$, this means that a cluster contains only TFs of one colour and so is fully demixed; if $\theta_{\rm dem} = 0$, the cluster contains a mixture of TFs of all colors in equal number, and so is maximally demixed. More details can be found in the SI.

We consider two different types of string, one with $M = 3000$ beads (or 9 Mbp) which is referred to as a "toy" string, and a second representing a whole human chromosome. Chromosomes are initialised in both cases as random walks: for more details, see SI.

Toy model

The toy model is built by placing one yellow, red, or green TU every 30 weakly-binding beads, giving a total of 100 TUs of all types in a string of 3000 beads [11]. Various different sequences of TU colour down the string are considered. In one – the "random" string – TU colours are chosen randomly (see Figure 1a and SI for the specific sequence generated). In a second and third – the "1-pattern" and "6-pattern" strings – TU colors follow a repeating pattern (red, then yellow, then green) 1 or 6 times (see Figure 4).

For the random string, we monitor how the system responds to different perturbations. Local "mutations" are inspired by editing experiments performed using CRISPR/Cas9 [40]. One to four mutations are mimicked by switching selected yellow beads inside a cluster of consecutive yellow TUs (between TUs 1920 to 2070) to red ones (Figure 2). Thus, conversion of TU bead 1980 gives a string with 1 mutation, of 1950 and 1980 gives 2 mutations, of 1950 to 2010 gives 3 mutations, and 1950 to 2040 gives 4 mutations. Global perturbations are in-

spired by experiments reducing global levels of TFs using auxin-induced degrons [41]. Here, we study the effects of reducing the concentration of yellow TFs by 30%.

Human chromosomes

Our reference case for whole human chromosome simulations in the main text is the mid-sized human chromosome HSA 14 (107 Mbp), coarse-grained into $M = 35784$ beads. For Figure 6, weakly- and strongly-binding beads are identified (using ENCODE data [42] for human umbilical vein endothelial cells, HUVECs) by the presence of H3K27ac modifications and DNase-hypersensitivity sites (DHSs) in the 3 kbp region corresponding to that bead – as these are good markers of open chromatin and active TUs (both genic and non-genic), respectively. For Figure 6, TUs are split into ones only active in HUVECs and others ("house-keeping" ones) that are also active in H1-hESC cells (again using DHS sites and ENCODE data). Then, if a TU appears in both HUVECs and H1-hESCs, it is marked as housekeeping and coloured red; if it appears only in HUVECs it is marked as HUVEC-specific and coloured green. This allows an intuitive and simple multicolour model of HUVECs to be constructed. All remaining beads (which are not either weakly-binding or TUs) are non-binding. This approach represents a generalisation of the DHS model described in [9], so we call it the multicolour DHS model.

We also consider HSA 18 (80 Mbp, 26026 beads) and 19 (58 Mbp, 19710 beads) in HUVECs, chosen as they represent gene-poor and gene-rich chromosomes, respectively. Additionally, we consider HSA 14 in the B-lymphocyte line GM12878 (again, colours are chosen by combining DHS data for GM12878 and H1-hESCs). H3K27ac and DHS data is again from ENCODE.

For human chromosomes, transcriptional-activity data obtained from simulations and GRO-seq are compared in two ways [9]. First, we rank activities of each TU, and build a two-dimensional histogram providing an overview of the agreement between the two sets of ranks. Second, we quantify Spearman's rank correlation coefficient between numerical and experimental data (see Supplemental Information for more details).

RESULTS

Toy model with three colours

We first simulate binding of yellow, red, and green TFs to a string representing 3 Mbp. TU beads are positioned regularly in this string, but are coloured randomly yellow, red, or green (we refer to this case as the random string). TFs also bind reversibly and strongly to beads of the same colour, and weakly to all others (Figure 1A).

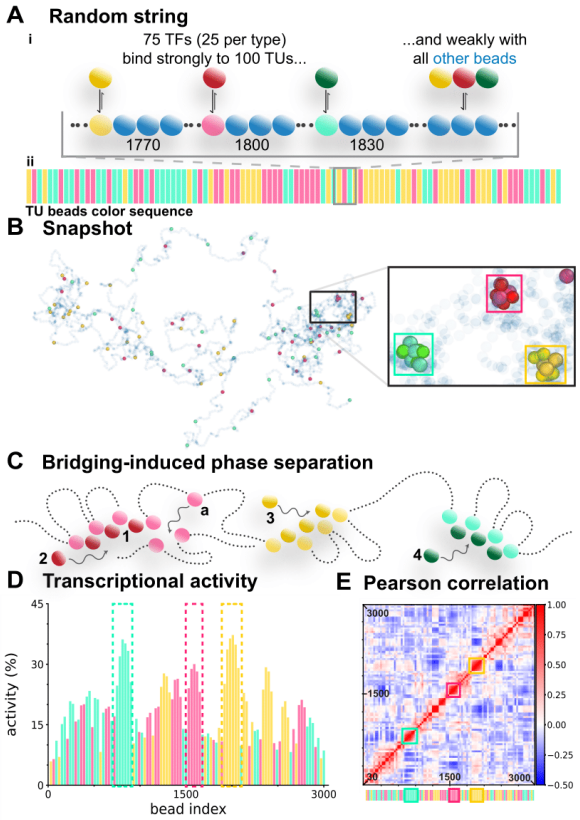

FIG. 1. **Toy model, with TUs coloured randomly (the random string)**. **(A)** Overview. (i) Yellow, red, and green TFs (25 of each colour) bind strongly (when in an on state) to 100 TUs beads of the same colour in a string of 3000 beads (representing 3 Mb), and weakly to blue beads. TU beads are positioned regularly and coloured randomly, as indicated in one region of the string. TFs switch between off and on states at rates $\alpha_{off} = 10^{-5}\ \tau_B^{-1}$ and $\alpha_{on} = \alpha_{off/4}$ (τ_B Brownian time, which one can map to $0.6-6\ 10^{-3}\ s$, see SI). (ii) The sequence of bars reflects the random sequence of yellow, red, and green TUs (blue beads not shown). **(B)** Snapshot of a typical conformation obtained after a simulation (TFs not shown). Inset: enlargement of boxed area. TU beads of the same colour tend to cluster and organize blue beads into loops. **(C)** Bridging-induced phase separation drives clustering and looping. Local concentrations of red, yellow, and green TUs and TFs might appear early during the simulation (blue beads not shown). Red TF 1 – which is multivalent – has bound to two red TUs and so forms a molecular bridge that stabilizes a loop; when it dissociates it is likely to re-bind to one of the nearby red TUs. As red TU 2 diffuses through the local concentration, it is also likely to be caught. Consequently, positive feedback drives growth of the red cluster (until limited by molecular crowding). Similarly, the yellow and green clusters grow as yellow TF 3 and green TF 4 are captured. **(D)** Bar heights give transcriptional activities of each TU in the string (average of 100 runs each lasting $8\ 10^5 \tau_B$). A TU bead is considered to be active whilst within $2.24\sigma \sim 6.7 \times 10^{-9} m$ of a TF:pol complex of similar colour. Dashed boxes: regions giving the 3 clusters in the inset in **(B)**. **(E)** Pearson correlation matrix for the activity of all TUs in the string. TU bead number (from low to high) is reported on axes, with pixel colour giving the Pearson value for each bead pair (bar on right). Bottom: reproduction of pattern shown in **(A,ii)**. Boxes: regions giving the 3 clusters in the inset in **(B)**.

After running a Brownian-dynamics simulation, Figure 1B shows a typical 3D conformation found in the steady state. Remarkably, clusters of TUs and TFs with distinct colours appear and disappear spontaneously, as seen previously using a single-colour model [9, 11, 12]. Such clustering is driven by the positive feedback illustrated in Figure 1C; it depends critically on TFs being able to form molecular bridges that anchor loops.

We now assume that the spheres represent TF:pol complexes, and make the reasonable assumption that a TU bead is transcribed if it lies within 2.25 diameters (2.25σ) of a complex of the same colour; then, the transcriptional activity of each TU is given by the fraction of time that the TU and a TF:pol lie close together. Figure 1D reports the mean activity profile down the string; TUs with the lowest activities are flanked by differently-coloured TUs, while those with the highest activities are flanked by similarly-coloured TUs (dashed rectangles in Figure 1D). As expected, a single-colour model with the same TU placement leads to a flat activity profile (Fig. S1A). Clearly, close proximity in 1D genomic space favours formation of similarly-coloured clusters.

We next examine how closely transcriptional activities of different TUs correlate [43]; the Pearson correlation matrix for all TUs is shown in Figure 1E. Correlations between neighbouring TUs of similar colour are often positive and strong, resulting in square red blocks along the diagonal (coloured boxes in Figure 1E highlight the 3 clusters shown in the zoom in Figure 1B). This effect is again due to the self-assembly of clusters containing neighbouring TUs of the same colour. In contrast, neighbours with different colours tend to compete with each other for TF:pols, and so down-regulate each other to yield smaller correlations. Correlations are more trivial in the single-color counterpart of Figure 1, where the matrix yields only a positive-correlation band along the diagonal (Fig. S1B). These results provide simple explanations of two mysterious effects – the first being why adjacent TUs throughout large domains tend to be co-transcribed so frequently [44]. The second concerns how expression quantitative trait loci (eQTLs) work. Current models see them doing so post-transcriptionally in highly-convoluted ways [9, 45], but we have argued that any TU can act as an eQTL directly at the transcriptional level [9]. Here, we see individual TUs up-regulate some TUs and down-regulating others – defining features of eQTLs that can lead to genetic effects like "transgressive segregation" [46].

Local mutations. These simulations are inspired by editing experiments performed using CRISPR/Cas9 [40]. We choose the most active region in the random string – one containing a succession of yellow TUs – and "mutate" $1-4$ of these TUs by recolouring them red (Figure 2A). Typical snapshots show red mutants are often ejected from yellow clusters (Figure 2Bi), or cluster together to leave their wild-type neighbours in isolation (Figure 2Bii). These changes are reflected in activity profiles (Figure 2C; arrows indicate mutations). As the number of mutations in the cluster increase, activities of yellow beads in that cluster decrease (Figure 2D), and new red clusters often emerge (Figure 2B,ii; Figure 2Ciii).

To confirm that 4 mutations in a yellow cluster often lead to the development of a red cluster, we monitor cluster dynamics over time. Figure 2Ei illustrates a typical kymograph illustrating changes in activity of all TUs in the wild-type; yellow, red, and green pixels mark activity of respective TUs, and black ones inactivity. In this particular simulation, a yellow cluster in the region that will be mutated (marked by the blue rectangle) is present during the first quarter of the time window; it then disappears to reappear half-way through the window and then persists until the end. In the string with 4 mutations, a yellow cluster is never seen; instead, different red clusters appear and disappear (Figure 2Eii). Pearson correlation matrices provide complementary information: the yellow cluster in the wild-type yields a solid red block indicating strong positive correlations (Figure 2Fi), but this block fragments in the string with 4 mutations (Figure 2Fii); mutations also induce subtle long-distance correlations between TUs. These results confirm that local arrangements of TUs on the genetic map determine the extent to which any particular TU will cluster and so become active.

Variations in TF concentration. These simulations are inspired by experiments reducing global levels of TFs using auxin-induced degrons [41]; we reduce the concentration of yellow TFs binding to the random string by 30% (Figure 3A). As expected, transcriptional activity falls both globally and locally (see yellow dotted rectangles in Figure 3B and C). Surprisingly, activity of a nearby cluster of red TUs (numbers 1080, 1110, 1170, 1200, and 1530 to 1650) increases by 50% (red dotted rectangles in Figure 3B and C). This effect is specific, in the sense that there is little effect on green clusters (e.g., compare Figure 1D with Figure 3B). We attribute this to a now-reduced steric competition for 3D space by yellow neighbours – fewer yellow clusters are present to stunt growth of nearby red ones.

Comparison of correlation matrices (compare Figure 1E and Figure 3D) shows a reduction in range and strength of positively-correlations for all yellow TUs and and an opposite effect for red ones. Again, TUs in green clusters are little affected, as they happen to be further away in 1D genomic space from yellow TUs in this particular string. Overall, these results show there are many statistically-significant correlations in activities both near and far away on the genetic map – much like those seen between singly-coloured TUs considered earlier [9].

Effects of 1D TU patterns on transcriptional activity. To better understand local effects uncovered with the random string, we now compare different toy

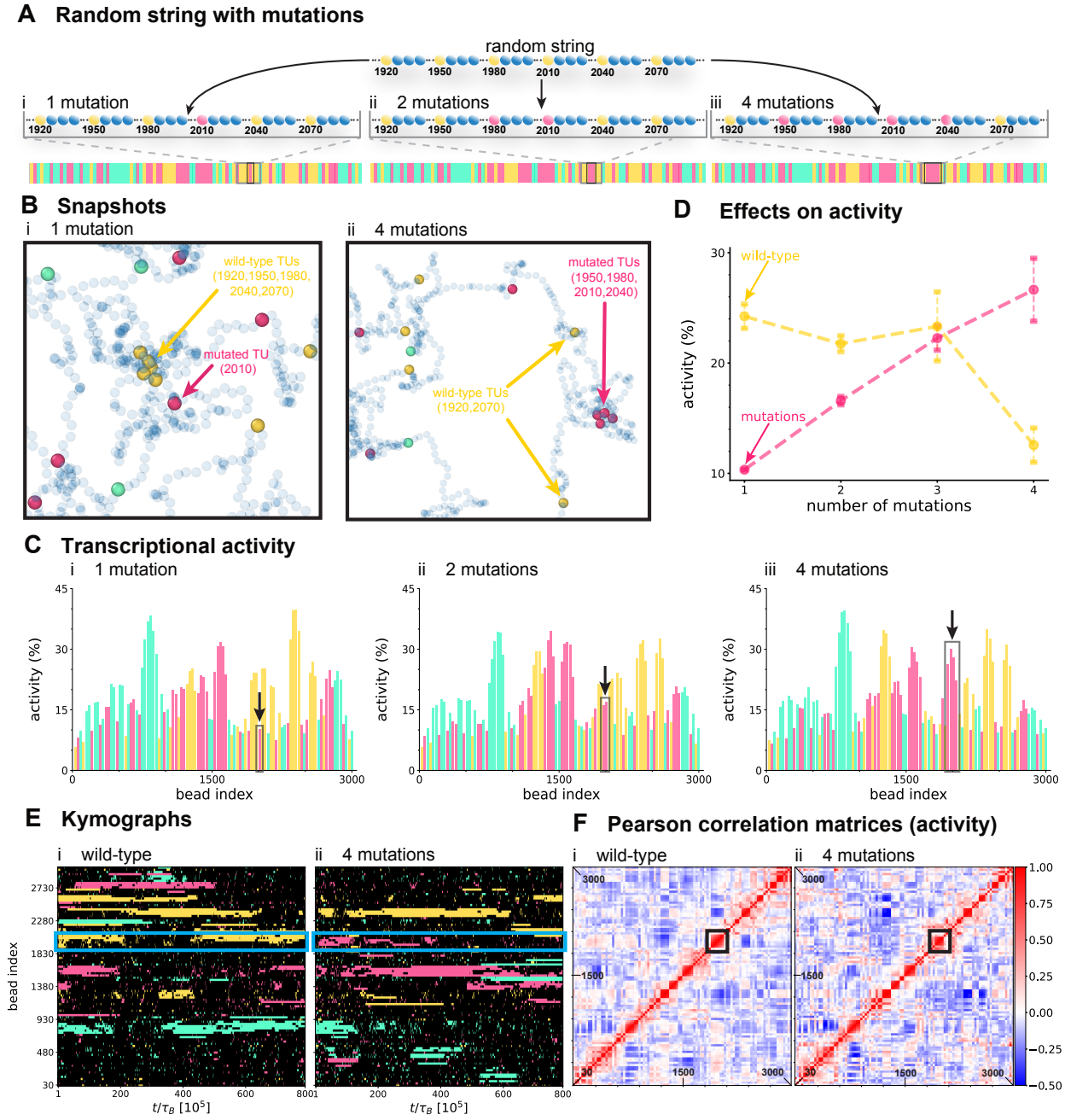

FIG. 2. **Simulating effects of mutations**. Yellow TU beads 1920, 1950, 1980, 2010, 2040 and 2070 in the random string have the highest transcriptional activity. 1-4 of these beads are now mutated by recolouring them red. **(A)** The sequence of bars reflects the sequence of yellow, red, and green TUs in random strings with 1, 2 and 4 mutations (blue beads not shown). Black boxes highlight mutant locations. **(B)** Typical snapshots of conformations with **(i)** one, and **(ii)** 4 mutations. **(C)** Transcriptional-activity profiles of mutants (averages over 100 runs, each lasting 8 $10^5\tau_B$). Bars are coloured according to TU colour. Black boxes: activities of mutated TUs. **(D)** Activities (+/- SDs) of wild-type (yellow) and different mutants. 3 mutations: TUs 1950, 1980 and 2010 mutated from yellow to red. **(E)** Typical kymographs for **(i)** wild-type and **(ii)** 4-mutant cases. Each row reports the transcriptional state of a TU during one simulation. Black pixels denote inactivity, and others activity; pixels colour reflects TU colour. Blue boxes: region containing mutations. **(F)** Pearson correlation matrices for wild-type and 4-mutant cases. Black boxes: regions containing mutations (mutations also change patterns far from boxes).

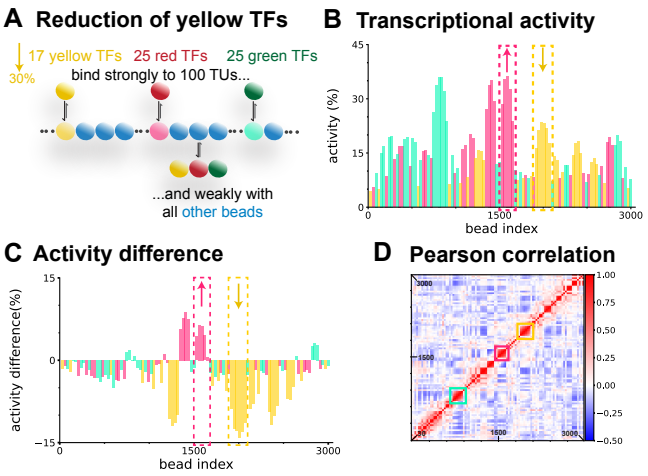

FIG. 3. **Reducing the concentration of yellow TFs reduces the transcriptional activity of most yellow TUs while enhancing the activities of some red TUs. (A)** Overview. Simulations are run using the random string with the concentration of yellow TFs reduced by 30%, and activities determined (means from 100 runs each lasting 8 $10^5\tau_B$). **(B)** Activity profile. Dashed boxes: activities fall in the region containing the biggest cluster of yellow TUs seen with 100% TFs, as those of an adjacent red cluster increase. **(C)** Differences in activity induced by reducing the concentration of yellow TFs. **(D)** Pearson correlation matrix. Boxes: regions giving the 3 clusters in **Fig. 1B, inset**.

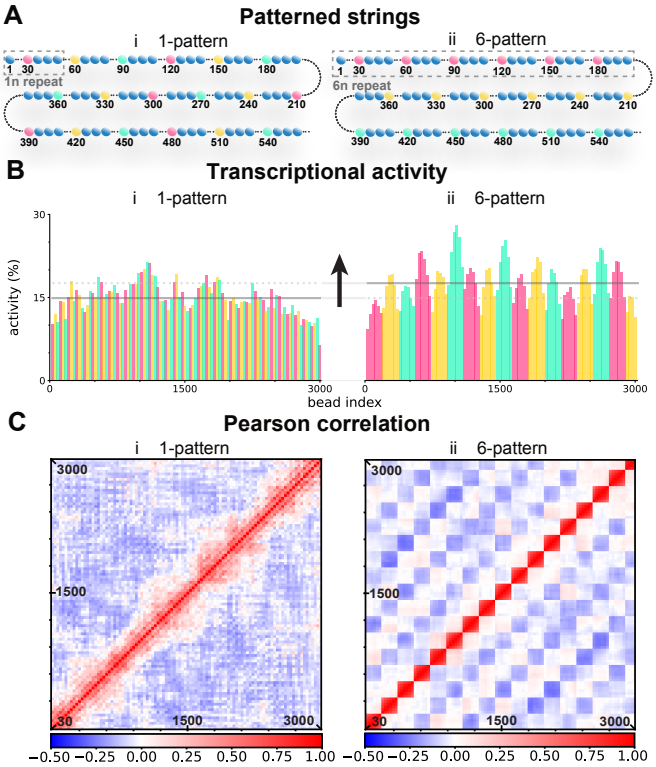

FIG. 4. **Clustering similar TUs in 1D genomic space increases transcriptional activity. (A)** Simulations involve toy strings with patterns (dashed boxes) repeated 1 or 6 times. Activity profiles plus Pearson correlation matrices are determined (100 runs, each lasting 8 $10^5\tau_B$). **(B)** The 6-pattern yields a higher mean transcriptional activity (arrow highlights difference between the two means). **(C)** The 6-pattern yields higher positive correlations between TUs within each pattern, and higher negative correlations between each repeat.

strings with regular and repeating patterns of colored TUs (Figure 4). Two results are apparent. First, activities (Figure 4Bii) in the 6-pattern case are higher overall (compare horizontal dotted lines), and more variable (compare activities of the two central TUs within each repeat with peripheral ones) relative to the 1-pattern case (Figure 4Bi). This is consistent with positive additive effects acting centrally within each 6-pattern repeat, coupled to competitive negative effects of flanking and differently-coloured repeats at the edges. Second, the 6-pattern also has a Pearson correlation matrix (Figure 4Cii) that is highly-structured, with a checkerboard pattern; red blocks on the diagonal indicate high positive correlations (so the 1D 6-pattern clearly favours 3D clustering). [Such a checkerboard pattern is not seen with a single-color model that has a correlation matrix with one red continuous diagonal when TUs are regularly spaced (Figure $S1$).] Additionally, blue off-diagonal blocks indicate repeating negative correlations that reflect the period of the 6-pattern. These results show how strongly TU position in 1D genomic space affect 3D clustering and activity, and that these effects depend on inclusion of more than one colour.

Emergent regulatory networks. We have seen many positive and negative correlations between activities of TUs in the random string (Figure 1). We now select significant correlations from Pearson correlation matrices (those which are > 0.2, Figure 5A) to highlight emergent interaction networks [9]. In such networks, nodes represent each TU from first to last (other beads are not shown), and edges indicate positive (black) or negative (grey) correlations in activities of node pairs. Even for the toy random string, these networks prove to be very complex (Fig. S2A). They are also "small-world" (i.e., most nodes can be reached from other ones by a few steps [9, 47]). Given this complexity, we now consider simplified versions. Thus, in Figure 5Ai, only interactions between red TUs are shown (the first red TU is at position 60, the last at position 2910, and interactions between different colours are not depicted). As expected, activities of most red TUs are positively correlated with those of nearby TUs. Conversely, negative correlations connect distant TUs, as found in the single-color model [9]; as we have seen, binding of red TFs to any red cluster reduces the number available to bind elsewhere.

In Figure 5Aii, we consider just interactions between red TUs and green TUs. Remarkably, close-range positive correlations (black edges) are still seen between TU

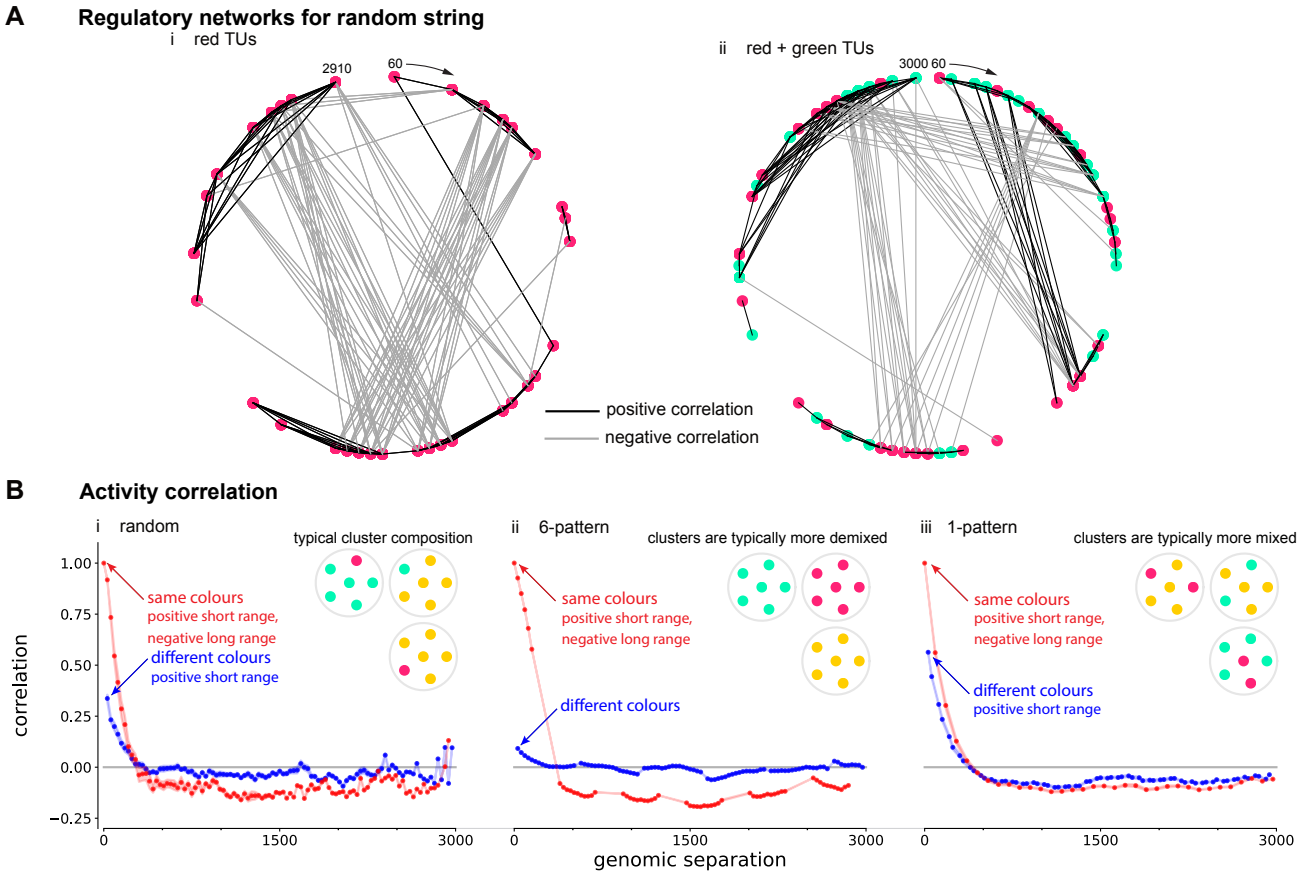

FIG. 5. **TU regulatory networks and demixing.** Simulations are run using the toy models indicated, and complete regulatory networks constructed from Pearson correlation matrices. **(A)** Simplified network given by the random string. TUs from first (bead 30) to last (bead 3000) are shown as peripheral nodes (coloured according to TU); black and grey edges denote statistically-significant positive and negative correlations, respectively (above a threshold of 0.2, corresponding to a p-value $\sim 5\ 10^{-2}$). The complete network consists of $n = 100$ individual TUs, so that there are $n_c = \binom{100}{2} = 4950$ pairs of TUs couples; we find 990 black and 742 gray edges. Since p-value$\cdot n_c = 223$, most interactions (edges) are statistically significant. Networks shown here only correlations (i) between red TUs, and (ii) between red and green TUs. (ii) **(B)** Average correlation (shading shows +/-SD, and is usually less than line/spot thickness) as a function of genomic separation for the (i) random, (ii) 6-, and (iii) 1-pattern cases. Correlation values at fixed genomic distance are taken from super-/sub-diagonals of Pearson matrices. Red dots give mean correlation between TUs of the same color (3 possible combinations), and blue dots those between TUs of different colors (4 possible combinations). Cartoons depict contents of typical clusters to give a pictorial representation of mixing degree (as this determines correlation patterns); see SI for exact values of $\theta_{\rm dem}$.

pairs that no longer bind TUs of the same colour. We suggest this is due to the presence of weakly-binding beads. Specifically, a red cluster organises a surrounding cloud of weakly-binding beads, and these will bind some green TFs that – in turn – bind green TUs. In contrast to the same-colour network in Fig. 5Ai, there are now more long-range positive correlations, showing that the presence of multiple colors enriches the emerging network.

To obtain further quantitative insight into these subtle yet remarkable correlations, we compute the average of those between same- and different-colour TUs as a function of genomic separation (Figure 5B). For the random string, same-colour correlations switch from clearly positive to slightly negative at about 300 beads (Figure 5Bi, red curve). Differently-coloured correlations yield a broadly-similar switch, although positive and negative values are weaker (Figure 5Bi, blue curve). The 6-pattern gives qualitatively similar trends, with the magnitude of differently-coloured correlations dampened further (Figure 5Bii). In contrast, the 1-pattern string yields largely overlapping curves (Figure 5Biii). These results illustrate how the sequence of TUs on a string can strikingly affect formation of mixed clusters; they also provide an explanation of why activities of human TUs within genomic regions of hundreds of kbp are positively correlated [48].

To quantify the extent to which TFs of different colours share clusters, we introduce a demixing coefficient, $\theta_{\rm dem}$ (defined in Equation 1), which can vary between 0 and 1. If $\theta_{\rm dem} = 1$, a cluster contains only TFs of one colour (and so is fully demixed); if $\theta_{\rm dem} = 0$, it contains both red

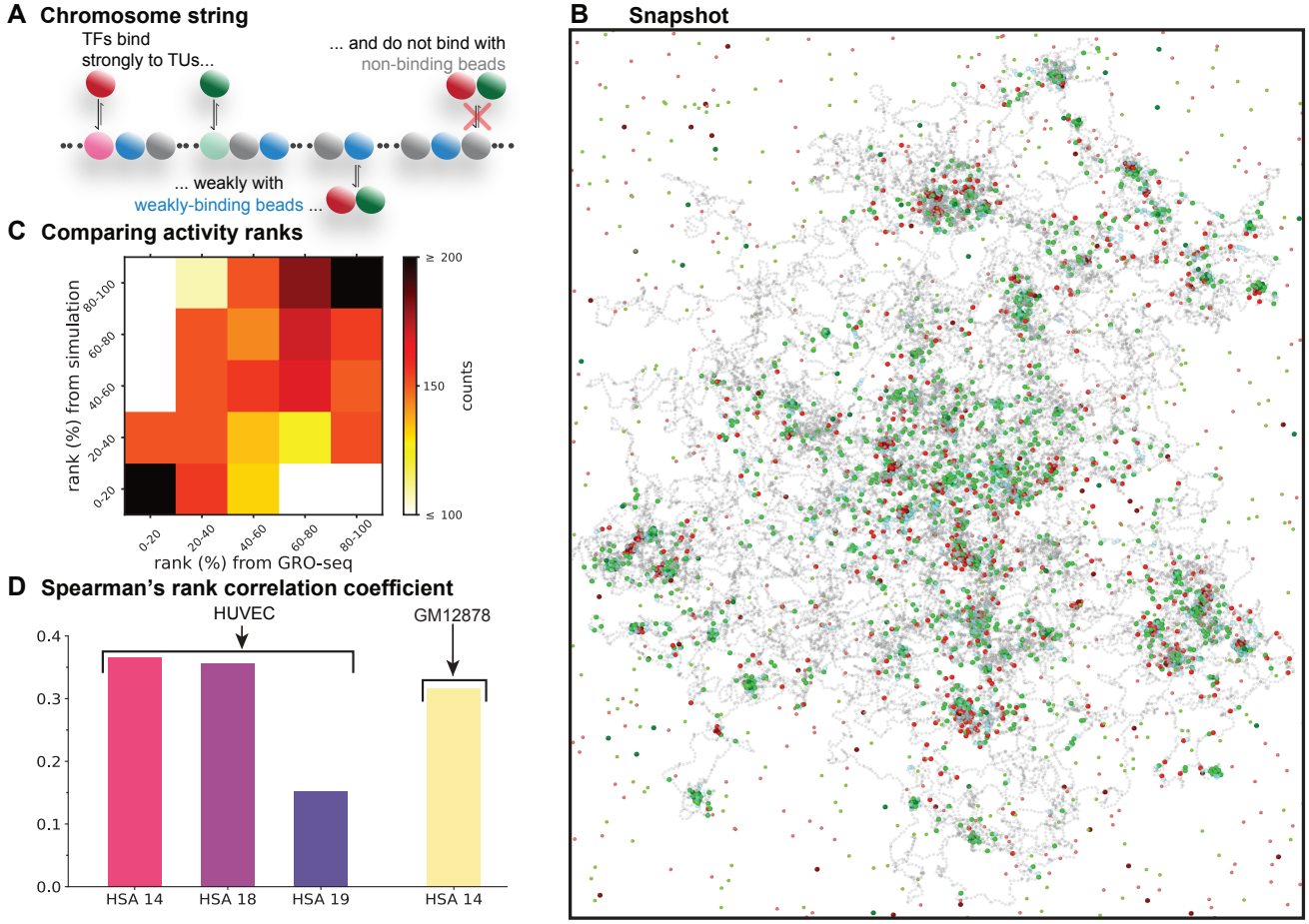

FIG. 6. **Comparison of transcriptional activities of TUs on different human chromosomes determined from simulations and GRO-seq. (A)** Overview of panels (A-C). The 35784 beads on a string representing HSA14 in HUVECs are of 4 types: TUs active only in HUVECs (red), "house-keeping" TUs – ones active in both HUVECs and ESCs (green), "euchromatic" ones (blue), and "heterochromatic" ones (grey). Red and green TFs bind strongly to TUs of the same colour, and weakly to euchromatin; neither binds to heterochromatin. **(B)** Snapshot of a typical conformation, showing both specialized and mixed clusters. **(C)** TU activities seen in simulations and GRO-seq are ranked from high to low, binned into quintiles, and activities compared. **(D)** Spearman's rank correlation coefficients for the comparison between activity data obtained from analogous simulations and GRO-seq for the chromosomes and cell types indicated.

and green TFs in equal numbers (and so is fully mixed). Intuitively, one might expect $\theta_{\rm dem}$ to fall as the number of adjacent TUs of similar colour in a string fall; this is what is seen with the 6- and 1-patterns – strings with the most and least numbers of adjacent TUs of similar colour, respectively (Fig. S2B; shown schematically by the cluster cartoons in Figure 5B).

Our results then show that in cases where same- and different-colour correlations overlap (as in the 1-pattern string), clusters are more mixed (have a larger value of $\theta_{\rm dem}$). Instead, in cases where same- and different-color correlations diverge, or are more different (as in the 6-pattern string), then clusters are typically unmixed, and so have a larger value of $\theta_{\rm dem}$ (Fig. S2B). Mixing is facilitated by the presence of weakly-binding beads, as replacing them with non-interacting ones increases demixing and reduces long-range negative correlations (Fig. S6).

Therefore, the sequence of strong and weak binding sites along strings determines the degree of mixing, and the types of small-world network that emerge. If eQTLs also act transcriptionally in the way we suggest [9], we predict that down-regulating eQTLs will lie further away from their targets than up-regulating ones. More generally, we suggest that the presence of multiple TF colours provides a powerful pathway to enrich and modulate transcriptional regulation.

Simulating whole human chromosomes

Transcriptional activity and comparison with GRO-seq data. We now simulate human chromosome 14 (HSA 14) in HUVECs, with individual beads in the string coloured appropriately (Figure 6A). Thus,

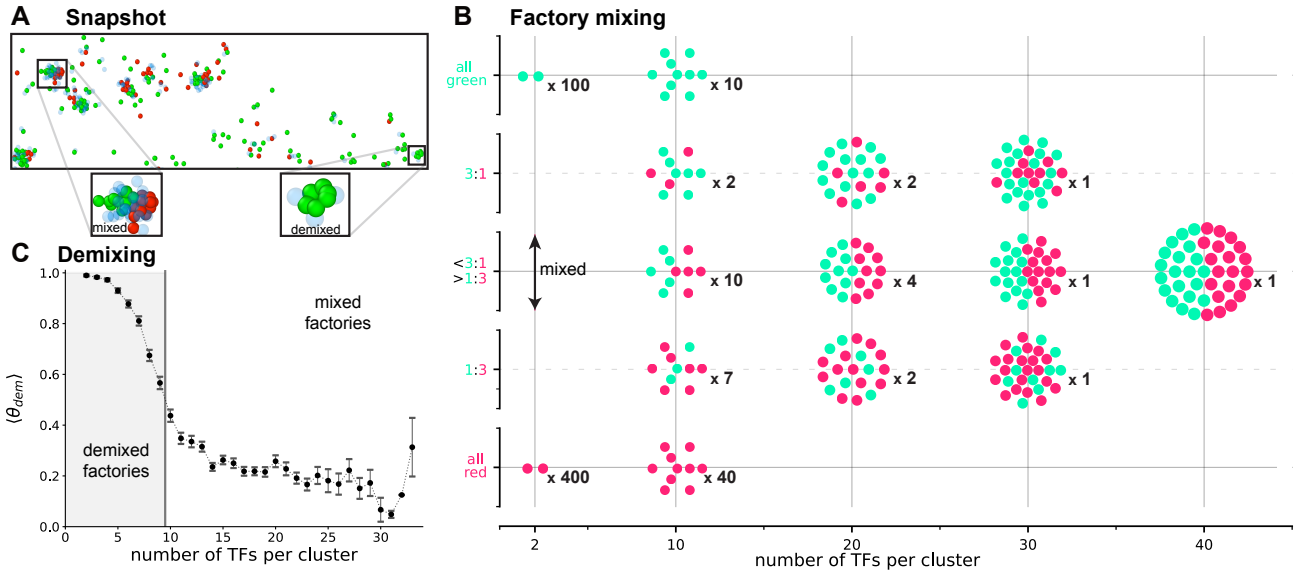

FIG. 7. **Small clusters tend to be unmixed, large ones mixed.** After running one simulation for HSA 14 in HUVECs, clusters are identified. **(A)** Snapshot of a typical final conformation (TUs, non-binding beads, and TFs in off state not shown). Insets: a large mixed cluster and a small demixed one. **(B)** Example clusters with different numbers of TFs/cluster (2, 10, 20, 30, 40) chosen to represent the range seen from all-red to all-green (with 3 intervening bins). Black numbers: observed number of clusters of that type seen in the simulation. **(C)** Average of the demixing coefficient $\theta_{\rm dem}$ (error bars: SD). Values of 1 and 0 are completely demixed and completely mixed respectively. Grey area: demixed regime where $\theta_{\rm dem}$ is > 0.5.

TUs transcribed uniquely in HUVECs are coloured red, housekeeping TUs (i.e., ones also expressed in a stem cell, namely H1-hESCs) are green, euchromatic regions blue, and heterochromatic ones grey. Figure 6B shows a typical snapshot; red and green clusters again form spontaneously. We next determine transcriptional activities, rank them in order from high to low, and compare binned rank orders with those obtained experimentally by GRO-seq (Figure 6C); most counts lie along the diagonal, meaning there is a good agreement between the two data sets. More quantitatively, Spearman's rank correlation coefficient is 3.66 10^{-1}, which compares with 3.24 10^{-1} obtained previously using a single-colour model [9]. In both cases the estimated uncertainty is of order 10^{-3} (mean and SD obtained using the bootstrap technique over 100 trials); consequently, use of an additional color provides a statistically-significant improvement (p-value $< 10^{-6}$, 2-sided t-test).

Activity predictions are also improved compared to the one-colour model with HSA 18 and HSA 19 in HUVECs, plus HSA 14 in GM12878 (Fig. 6D). However, Spearman's rank coefficient for gene-poor HSA 18 is about twice that for gene-rich HSA 19; this may be due to additional regulatory layers in regions with high promoter density. These results confirm that our multicolour polymer model generates strings that can mimic structures and functions found in whole chromosomes.

Specialized and mixed clusters. Inspection of snapshots shows 1-colour clusters tend to be smaller than mixed (2-colour) ones (Figure 7A). To quantify this, we count numbers and types of TFs in individual clusters (Figs. 7B and Fig. S6). Clusters with just two bound TFs never contain both colours; conversely, those with > 20 bound TFs never contain just one colour (Figure 7B). We also measure the average value of the demixing coefficient, $\theta_{\rm dem}$ (Materials and Methods). The cross-over point between fully mixed and demixed (where the average value of $\theta_{\rm dem} = 0.5$) occurs when there are ~ 10 TFs per cluster (Figure 7C): notably, this is similar to the average number of productively-elongating pols seen experimentally in a transcription factory [4]. Similar results are obtained for different cell types, or chromosomes (see, e.g., Fig. S4 for the case of HSA 14 in GM12878).

The transition between specialized (demixed) and mixed factories with increasing size can be explained as follows. Two red TFs in an unmixed cluster might stabilise 3 loops, and so bring into close proximity only a few non-specific binding sites that could bind a green TF. In contrast, 10 red TFs in a cluster will stabilise many loops that inevitably bring into close proximity many non-specific binding sites – and this makes it highly likely that some green TFs will also bind nearby to create a mixed cluster. This mixing transition provides a way to reconcile observations that some clusters are unmixed (like factories rich in polymerases II and III), and others highly mixed (like HOTs).

Finally, as for the toy model, the balance between mixing and demixing determines correlation patterns.

For example, activity patterns of same- and differently-colored TUs in the whole chromosome (Fig. S7) are much like those in the 1-pattern model (Figure 5Biii). We attribute this to $\sim$ 78% TFs being in mixed clusters ($\theta_{\rm dem} < 0.5$), and and so inevitably the resulting interactions will dominate the pattern seen.

DISCUSSION AND CONCLUSIONS

We use coarse-grained simulations of long polymers representing human chromosomes to study inter-relations between 3D structure and transcriptional dynamics. Unlike previous work [9, 12], we adopt a multi-colour model where different "colours" of TF:pol complex bind to different genic and non-genic TUs. We confirm that TF:pols spontaneously self-assemble into clusters (Figure 1), and obtain various striking results. First, when small, these clusters typically contain TFs of just one colour; these are reminiscent of the specialized transcription factories found in the nucleoplasm that contain active forms of just pol II or pol III – but not both [49]. When large, they are typically mixed (Figure 7C); this provides a mechanistic basis for the formation of HOTs, where many different TFs bind promiscuously and weakly to segments of open chromatin that are often devoid of cognate binding cites [22–24]. Consequently, this simple mechanism provides an appealing resolution of the puzzling observation that specialized factories and HOTs can occur in the same cell.

Second, we see remarkable positive and negative correlations in the transcriptional activities of different TUs. For example, activities of same-colour and nearby TUs tend to be strongly positively correlated, as such TUs tend to co-cluster (Figure 5). Conversely, activities of similar TUs lying far from each other on the genetic map are often weakly negatively correlated, as the formation of one cluster inevitably sequesters some TFs to reduce the number available to bind elsewhere. Taken together, these results provide simple explanations of why adjacent TUs throughout large domains tend to be co-transcribed so frequently [48], and how one eQTL might up-regulate some TUs and down-regulate others – and so lead to "transgressive segregation" [46]. More generally, they suggest that transcriptional correlations can be modulated by varying TU position and by tilting the balance between weakly-binding and non-binding sites in fibres.

Third, we can predict effects of local mutations and genome edits that often induce distant omnigenic and pangenomic effects uncovered by genome-wide association studies ([9, 45]. For example, mutations that switch a binding site of one TF to another can convert a cluster of one colour into another (Figure 2). Similarly, global effects of knocking down TF levels are easily assessed (Figure 3).

Fourth, we also predict transcriptional activities of all TUs (both genic and non-genic) on whole human chromosomes by including cell-type-invariant and cell-type-specific TFs (Figure 6). We find this yields a better correlation with GRO-seq experimental data than a single-colour model (where just one TF binds to all TUs similarly). This result underscores the importance of including different TFs in polymer models.

Finally, all our results point to the importance of the 1D pattern of TUs and TF-binding sites on chromosomes in determining activity. In other words, 1D location is a key feature determining transcriptional patterns, and so cell identity. We speculate this is why relative locations of active regulatory elements are so highly conserved. For instance, despite human enhancers evolving much more rapidly than their target protein-coding genes, the synteny between the two (over distances up to 2 Mbp) is highly conserved [50, 51].

In the future, it would be of interest to include many more types of TF and TU both within this and other frameworks like HiP-HoP [30] and ones incorporating epigenetic modifications [13, 52]. From a theoretical point of view, we hope our results will stimulate work aimed at understanding inter-relations between structure and function.

ACKNOWLEDGEMENTS

The work has been performed within the HPC-EUROPA3 Project (INFRAIA-2016-1-730897), with the support of the EC Research Innovation Action under the H2020 Programme. We acknowledge funding from MIUR Project No. PRIN 2020/PFCXPE, and from the Wellcome Trust (223097/Z/21/Z).

DATA AVAILABILITY

All the experimental data used in this paper are available in the ENCODE database [42]. Simulations are performed using the open source software LAMMPS. All the custom scripts used for the simulations presented here will be shared on reasonable request to the corresponding author.

Conflict of interest statement.

None declared.

[1] Ana Pombo and Niall Dillon, "Three-dimensional genome architecture: players and mechanisms," Nat. Rev. Mol. Cell Biol. **16**, 245–257 (2015).

[2] Erez Lieberman-Aiden, Nynke L. van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, Bryan R. Lajoie, Peter J. Sabo, Michael O. Dorschner, Richard Sandstrom, Bradley Bernstein, M. A. Bender, Mark Groudine, Andreas Gnirke, John Stamatoyannopoulos, Leonid A. Mirny, Eric S. Lander, and Job Dekker, "Comprehensive mapping of long-range interactions reveals folding principles of the human genome," Science **326**, 289–293 (2009).

[3] Jesse R. Dixon, Siddarth Selvaraj, Feng Yue, Audrey Kim, Yan Li, Yin Shen, Ming Hu, Jun S. Liu, and Bing Ren, "Topological domains in mammalian genomes identified by analysis of chromatin interactions," Nature **485**, 376–380 (2012).

[4] P. R. Cook and D. Marenduzzo, "Transcription-driven genome organization: a model for chromosome structure and the regulation of gene expression tested through simulations," Nucleic Acids Res. **46**, 9895–9906 (2018).

[5] Jesse R Dixon, David U Gorkin, and Bing Ren, "Chromatin domains: the unit of chromosome organization," Mol. Cell **62**, 668–680 (2016).

[6] M Jordan Rowley, Michael H Nichols, Xiaowen Lyu, Masami Ando-Kuri, I Sarahi M Rivera, Karen Hermetz, Ping Wang, Yijun Ruan, and Victor G Corces, "Evolutionarily conserved principles predict 3d chromatin organization," Mol. Cell **67**, 837–852 (2017).

[7] Argyris Papantonis and Peter R Cook, "Transcription factories: genome organization and gene regulation," Chemical Reviews **113**, 8683–8705 (2013).

[8] Patrick Cramer, "Organization and regulation of gene transcription." Nature **573**, 45–54 (2019).

[9] CA Brackley, Nick Gilbert, Davide Michieletto, Argyris Papantonis, MCF Pereira, PR Cook, and Davide Marenduzzo, "Complex small-world regulatory networks emerge from the 3d organisation of the human genome," Nat. Commun. **12**, 1–14 (2021).

[10] Michael Chiang, Chris A Brackley, Catherine Naughton, Ryu-Suke Nozawa, Cleis Battaglia, Davide Marenduzzo, and Nick Gilbert, "Gene structure heterogeneity drives transcription noise within human chromosomes," bioRxiv (2022).

[11] Chris A. Brackley, Stephen Taylor, Argyris Papantonis, Peter R. Cook, and Davide Marenduzzo, "Nonspecific bridging-induced attraction drives clustering of dna-binding proteins and genome organization," Proc. Natl. Acad. Sci. USA **110**, E3605–E3611 (2013).

[12] Chris A. Brackley, James Johnson, Steven Kelly, Peter R. Cook, and Davide Marenduzzo, "Simulated binding of transcription factors to active and inactive regions folds human chromosomes into loops, rosettes and topological domains," Nucleic Acids Res. **44**, 3503–3512 (2016).

[13] C. A. Brackley, B. Liebchen, D. Michieletto, F. L. Mouvet, P. R Cook, and D. Marenduzzo, "Ephemeral protein binding to dna shapes stable nuclear bodies and chromatin domains," Biophys. J. **28**, 1085–1093 (2017).

[14] Frank W Albert and Leonid Kruglyak, "The role of regulatory variation in complex traits and disease," Nat. Rev. Genet. **16**, 197–212 (2015).

[15] P. R. Cook, *Principles of nuclear structure and function* (Wiley New York, 2001).

[16] Melissa J Fullwood, Mei Hui Liu, You Fu Pan, Jun Liu, Han Xu, Yusoff Bin Mohamed, Yuriy L Orlov, Stoyan Velkov, Andrea Ho, Poh Huay Mei, *et al.*, "An oestrogen-receptor-α-bound human chromatin interactome," Nature **462**, 58–64 (2009).

[17] Stefan Schoenfelder, Tom Sexton, Lyubomira Chakalova, Nathan F Cope, Alice Horton, Simon Andrews, Sreenivasulu Kurukuti, Jennifer A Mitchell, David Umlauf, Daniela S Dimitrova, *et al.*, "Preferential associations between co-regulated genes reveal a transcriptional interactome in erythroid cells," Nat. Genet. **42**, 53–61 (2010).

[18] Argyris Papantonis, Takahide Kohro, Sabyasachi Baboo, Joshua D Larkin, Binwei Deng, Patrick Short, Shuichi Tsutsumi, Stephen Taylor, Yasuharu Kanki, Mika Kobayashi, *et al.*, "Tnfα signals through specialized factories where responsive coding and mirna genes are transcribed," EMBO J. **31**, 4404–4414 (2012).

[19] Jie Chen, III Hero, Alfred O., and Indika Rajapakse, "Spectral identification of topological domains," Bioinformatics **32**, 2151–2158 (2016), https://academic.oup.com/bioinformatics/article-pdf/32/14/2151/19568527/btw221.pdf.

[20] Robert A Beagrie, Antonio Scialdone, Markus Schueler, Dorothee CA Kraemer, Mita Chotalia, Sheila Q Xie, Mariano Barbieri, Inês de Santiago, Liron-Mark Lavitas, Miguel R Branco, *et al.*, "Complex multi-enhancer contacts captured by genome architecture mapping," Nature **543**, 519–524 (2017).

[21] Gabrielle A Dotson, Can Chen, Stephen Lindsly, Anthony Cicalo, Sam Dilworth, Charles Ryan, Sivakumar Jeyarajan, Walter Meixner, Cooper Stansbury, Joshua Pickard, *et al.*, "Deciphering multi-way interactions in the human genome," Nat. Commun. **13**, 1–15 (2022).

[22] Celine Moorman, Ling V Sun, Junbai Wang, Elzo de Wit, Wendy Talhout, Lucas D Ward, Frauke Greil, Xiang-Jun Lu, Kevin P White, Harmen J Bussemaker, *et al.*, "Hotspots of transcription factor colocalization in the genome of drosophila melanogaster," Proc. Natl. Acad. Sci. USA **103**, 12027–12032 (2006).

[23] Joseph W Foley and Arend Sidow, "Transcription-factor occupancy at hot regions quantitatively predicts rna polymerase recruitment in five human cell lines," BMC Genom. **14**, 1–17 (2013).

[24] Ruggero Cortini and Guillaume J Filion, "Theoretical principles of transcription factor traffic on folded chromatin," Nat. Commun. **9**, 1–10 (2018).

[25] Jun Ding, Nadav Sharon, and Ziv Bar-Joseph, "Temporal modelling using single-cell transcriptomics," Nat. Rev. Genet. **23**, 355–368 (2022).

[26] Rasa Elmentaite, Cecilia Domínguez Conde, Lu Yang, and Sarah A Teichmann, "Single-cell atlases: shared and tissue-specific cell types across human organs," Nat. Rev. Genet. , 1–16 (2022).

[27] Leighton J Core, André L Martins, Charles G Danko, Colin T Waters, Adam Siepel, and John T Lis, "Analysis of nascent rna identifies a unified architecture of initiation regions at mammalian promoters and enhancers," Nat. Genet. **46**, 1311–1320 (2014).

[28] Antonio Jordán-Pla, Maria E Pérez-Martínez, and José E Pérez-Ortín, "Measuring rna polymerase activity genome-wide with high-resolution run-on-based methods," Methods **159**, 177–182 (2019).

[29] Chris A. Brackey, Davide Marenduzzo, and Nick Gilbert, "Mechanistic modeling of chromatin folding to understand function," Nature Methods **17** (2020), 10.1038/s41592-020-0852-6.

[30] Adam Buckle, Chris A Brackley, Shelagh Boyle, Davide Marenduzzo, and Nick Gilbert, "Polymer simulations of

heteromorphic chromatin predict the 3d folding of complex genomic loci," Mol. Cell **72**, 786–797 (2018).

[31] C. A. Brackley, J. Johnson, D. Michieletto, A. N. Morozov, M. Nicodemi, P. R. Cook, and D. Marenduzzo, "Non-equilibrium chromosome looping via molecular slip-links," Phys. Rev. Lett. **119**, 138101 (2017).

[32] Simona Bianco, Andrea M Chiariello, Carlo Annunziatella, Andrea Esposito, and Mario Nicodemi, "Predicting chromatin architecture from models of polymer physics," Chromosome Res. **25**, 25–34 (2017).

[33] Mario Nicodemi and Ana Pombo, "Models of chromosome structure," Current Opinion in Cell Biology **28**, 90–95 (2014), cell nucleus.

[34] Mattia Conte, Andrea M. Chiariello, Alex Abraham, Simona Bianco, Andrea Esposito, Mario Nicodemi, Tommaso Matteuzzi, and Francesca Vercellone, "Polymer models of chromatin imaging data in single cells," Algorithms **15** (2022), 10.3390/a15090330.

[35] Massimiliano Semeraro, Giuseppe Negro, Antonio Suma, Giuseppe Gonnella, and Davide Marenduzzo, "3d polymer simulations of genome organization and transcription across different chromosomes and cell types," (2022).

[36] Guido Tiana, Assaf Amitai, Tim Pollex, Tristan Piolot, David Holcman, Edith Heard, and Luca Giorgetti, "Structural fluctuations of the chromatin fiber within topologically associating domains," Biophys. J. **110**, 1234–1245 (2016).

[37] Luca Giorgetti, Rafael Galupa, Elphège P Nora, Tristan Piolot, France Lam, Job Dekker, Guido Tiana, and Edith Heard, "Predictive polymer modeling reveals coupled fluctuations in chromosome conformation and transcription," Cell **157**, 950–963 (2014).

[38] Martina Crippa, Yinxiu Zhan, and Guido Tiana, "Effective model of loop extrusion predicts chromosomal domains," Phys. Rev. E **102**, 032414 (2020).

[39] A. P. Thompson, H. M. Aktulga, R. Berger, D. S. Bolintineanu, W. M. Brown, P. S. Crozier, P. J. in 't Veld, A. Kohlmeyer, S. G. Moore, T. D. Nguyen, R. Shan, M. J. Stevens, J. Tranchida, C. Trott, and S. J. Plimpton, "LAMMPS - a flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales," Comp. Phys. Comm. **271**, 108171 (2022).

[40] Stefanie L. Morgan, Natasha C. Mariano, Abel Bermudez, Nicole L. Arruda, Fangting Wu, Yunhai Luo, Gautam Shankar, Lin Jia, Huiling Chen, Ji-Fan Hu, Andrew R. Hoffman, Chiao-Chain Huang, Sharon J. Pitteri, and Kevin C. Wang, "Manipulation of nuclear architecture through crispr-mediated chromosomal looping," Nat. Comm. **8**, 15993 (2017).

[41] Jing Luan, Guanjue Xiang, Pablo Aurelio Gómez-García, Jacob M Tome, Zhe Zhang, Marit W Vermunt, Haoyue Zhang, Anran Huang, Cheryl A Keller, Belinda M Giardine, *et al.*, "Distinct properties and functions of ctcf revealed by a rapidly inducible degron system," Cell reports **34**, 108783 (2021).

[42] ENCODE Project Consortium, "An integrated encyclopedia of dna elements in the human genome," Nature **489**, 57–74 (2012).

[43] Barak A Cohen, Robi D Mitra, Jason D Hughes, and George M Church, "A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression," Nat. Genet. **26**, 183–186 (2000).

[44] Nick Gilbert, Shelagh Boyle, Heike Fiegler, Kathryn Woodfine, Nigel P Carter, and Wendy A Bickmore, "Chromatin architecture of the human genome: gene-rich domains are enriched in open chromatin fibers," Cell **118**, 555–566 (2004).

[45] Evan A Boyle, Yang I Li, and Jonathan K Pritchard, "An expanded view of complex traits: from polygenic to omnigenic," Cell **169**, 1177–1186 (2017).

[46] Rachel B Brem and Leonid Kruglyak, "The landscape of genetic complexity across 5,700 gene expression traits in yeast," Proceedings of the National Academy of Sciences **102**, 1572–1577 (2005).

[47] Duncan J Watts and Steven H Strogatz, "Collective dynamics of small-worldnetworks," Nature **393**, 440–442 (1998).

[48] Laurence D Hurst, Csaba Pál, and Martin J Lercher, "The evolutionary dynamics of eukaryotic gene order," Nat. Rev. Genet. **5**, 299–310 (2004).

[49] Meng Xu and Peter R Cook, "Similar active genes cluster in specialized transcription factories," J. Cell Biol. **181**, 615–623 (2008).

[50] Camille Berthelot, Diego Villar, Julie E Horvath, Duncan T Odom, and Paul Flicek, "Complexity and conservation of regulatory landscapes underlie evolutionary resilience of mammalian gene expression," Nat. Ecol. Evol. **2**, 152–163 (2018).

[51] Alexandre Laverré, Eric Tannier, and Anamaria Necsulea, "Long-range promoter–enhancer contacts are conserved during evolution and contribute to gene expression robustness," Genome Res. **32**, 280–296 (2022).

[52] Davide Michieletto, Enzo Orlandini, and Davide Marenduzzo, "Polymer model with epigenetic recoloring reveals a pathway for the de novo establishment and 3d organization of chromatin domains," Phys. Rev. X **6**, 041047 (2016).