Systematic identification of interchromosomal interaction networks supports the existence of specialized RNA factories

Systematic identification of inter-chromosomal interaction networks supports the existence of specialized RNA factories

Borislav Hrisimirov Hristov[1], William Stafford Noble[1,2], and Alessandro Bertero[3,*]

[1]Department of Genome Sciences, University of Washington, Seattle, USA
[2]Paul G. Allen School of Computer Science and Engineering, University of Washington, Seattle, USA
[3]Molecular Biotechnology Center "Guido Tarone", Dept. of Molecular Biotechnology and Health Sciences, University of Turin, Torino, Italy
*Correspondence: alessandro.bertero@unito.it

Abstract

Most studies of genome organization have focused on intra-chromosomal (*cis*) contacts because they harbor key features such as DNA loops and topologically associating domains. Inter-chromosomal (*trans*) contacts have received much less attention, and tools for interrogating potential biologically relevant *trans* structures are lacking. Here, we develop a computational framework that uses Hi-C data to identify sets of loci that jointly interact in *trans*. This method, trans-C, initiates probabilistic random walks with restarts from a set of seed loci to traverse an input Hi-C contact network, thereby identifying sets of *trans*-contacting loci. We validate trans-C in three increasingly complex models of established *trans* contacts: the *Plasmodium falciparum var* genes, the mouse olfactory receptor "Greek islands", and the human RBM20 cardiac splicing factory. We then apply trans-C to systematically test the hypothesis that genes co-regulated by the same *trans*-acting element (i.e., a transcription or splicing factor) co-localize in three dimensions to form "RNA factories" that maximize the efficiency and accuracy of RNA biogenesis. We find that many loci with multiple binding sites of the same DNA-binding proteins interact with one another in *trans*, especially those bound by factors with intrinsically disordered domains. Similarly, clustered binding of a subset of RNA-binding proteins correlates with *trans* interaction of the encoding loci. We observe that these *trans*-interacting loci are close to nuclear speckles. These findings support the existence of *trans* interacting chromatin domains (TIDs) driven by RNA biogenesis. Trans-C provides an efficient computational framework for studying these and other types of *trans* interactions, empowering studies of a poorly understood aspect of genome architecture.

Running title: Identification of *trans* interacting DNA domains

Introduction

Mammalian interphase chromosomes are exquisitely folded in three dimensions to enable precise regulation of gene expression (reviewed in Hafner and Boettiger 2023). The study of such organization has been greatly advanced by sequencing-based chromosome conformation capture (3C) technologies, chiefly Hi-C (Lieberman-Aiden et al. 2009), and by orthogonal imaging approaches (Jerkovic and Cavalli 2021). A rapidly growing body of evidence indicates that while a sizeable portion of 3D genome architecture is relatively invariant across cell types, specific dynamic changes play a critical role in regulating gene expression in different cell types (Tan et al. 2023; Schaeffer and Nollmann, 2023; Winick-Ng et al. 2021; Duan et al. 2021) and in disease (Krumm and Duan, 2019; Zheng and Xie, 2019).

Most of our current understanding of 3D genome architecture centers around chromatin folding within individual chromosomes, that is, on intra-chromosomal or *cis* contacts. These contacts give rise to a variety of

hierarchical features at different genomic scales, including different types of DNA loops (i.e., cohesin-mediated looping and promoter-enhancer pairing), topologically associating domains (TADs; sub-megabase domains of preferential self-interaction; Dixon et al. 2012), and A/B compartments (chromosome-wide segregation of active/inactive chromatin resulting from sparse intra-TAD interactions driven mainly by phase separation of heterochromatin; Hildebrand and Dekker 2020). In contrast, interactions across different chromosomes (inter-chromosomal or *trans* contacts) are poorly understood.

Chromosome-wide *trans* genome architecture of non-holocentric chromosomes in eukaryotic species exhibits two non-mutually exclusive features: Rabl-like configuration (i.e., featuring centromere clustering, telomere clustering, and/or a telomere-to-centromere axis) and chromosome territories (Hoencamp et al. 2021). The latter is typical of mammalian chromosomes, which tend to occupy distinct domains of the interphase nucleus (Cremer and Cremer, 2010). Although chromosome territories limit the possibility for *trans* contacts — compared to an alternative model of "spaghetti" DNA fibers (Longo and Roukos, 2021) — they do not represent hard boundaries: regions that overcome territorial topological restrictions engage with each other within mRNA and tRNA factories, polycomb domains, the nucleolus, nuclear speckles, and potentially other nuclear subcompartments (Fig 1A; Bhat et al. 2021). Some of these *trans* contacts involve specific loci whose interactions are important to gene regulation in enhancer hubs (Monahan et al. 2019), transcription factories (Osborne et al. 2004, 2007; Papantonis et al. 2012), and splicing factories (Bertero et al. 2019). Despite these and a few other examples, whose discovery was serendipitous or informed by domain-specific prior knowledge (De Wit et al. 2013; Takizawa et al. 2008; Ito et al. 2016), the systematic discovery of functional *trans* interactions is currently very challenging.

One reason for this difficulty is that in a typical Hi-C matrix the number of reads from *trans* contacts is 2 to 4 times smaller than that from *cis* contacts, depending on cell type and assay type. Furthermore, the number of possible pairs of loci that can interact in *trans* is also much larger than the number of possible pairs of loci that can interact in *cis*. Collectively, therefore, *trans* contact data is typically quite sparse. Most importantly, there is a lack of robust statistical and computational approaches to confidently identify reproducible *trans* contacts. In particular, available methods are limited to the identification of pairwise *trans* contacts (Cook et al. 2020) or large patterns of *trans* contacts across broad subnuclear structures (Joo et al. 2023). There remains an important knowledge gap in detecting smaller, specific sets of *trans* contacts (cliques) that could underlie important local regulations of DNA and RNA biochemistry.

In this manuscript, we overcome this limitation by providing a computational framework that systematically finds sets of jointly interacting loci from Hi-C data. The method, trans-C, takes as input a Hi-C contact map as well as, optionally, one or more seed loci and uses a random walk with restart algorithm to identify sets of *trans*-contacting loci. Trans-C provides a powerful way to uncover and measure various types of *trans* interactions, empowering both discovery and hypothesis-driven studies of genome structure-function relationships.

Results

Trans-C randomly walks through the Hi-C graph

Our goal is to algorithmically identify, from a given set of Hi-C data, a collection of genomic loci that exhibit strong *trans* interactions with each other (i.e., a "clique"). We represent the Hi-C data as a matrix, referred to as a "contact map", in which each axis corresponds to the complete genome and entries in the matrix represent Hi-C contact counts (Fig 1B). In practice, the genomic axes are discretized using fixed-width bins. The bin size is thus inversely proportional to the effective resolution of the contact map. The contact map can be thought of as the adjacency matrix of a corresponding Hi-C graph, in which nodes are genomic loci (bins) and edges are weighted by the corresponding contact counts (Fig 1C). Our goal is thus to find dense subgraphs in this Hi-C graph.

The problem of dense subgraph discovery arises in many application domains and consequently has been very widely studied. Depending on the exact formulation and the notion of density, theoretical computer science has shown that the problem complexity ranges from easily solved in linear time *via* a max flow algorithm (Khuller and Saha, 2009) to computationally intractable (NP-hard; Charikar, 2000). Common

techniques to approximate the latter case, to which our specific problem belongs, are greedy approaches, which iteratively select the best option available at the moment without guaranteeing that this strategy will bring the global optimal result, (Charikar, 2000) and semi-random models, which account for model errors by incorporating both adversarial and random choices in instance generation (Bhaskara et al. 2010). Trans-C approaches the discovery task of finding strongly interacting loci in *trans* using a random walk with restart algorithm (Fig 1C). This general approach has been applied successfully in domains as diverse as web searching (Gibson et al. 2005), protein remote homology detection (Weston et al. 2004), and gene functional prediction (Mostafavi et al. 2008). Prior to the random walk operation, trans-C performs three pre-processing steps on the provided Hi-C contact map. First, to control for sequencing and accessibility biases, Hi-C counts are ICE-normalized (Imakaev et al. 2012). Second, the resulting matrix is processed using a binomial model to estimate interaction p-values based on an empirical null model that accounts for potential biases arising from chromosomal territorialization (i.e., small, gene-rich chromosomes generally occupying the nuclear interior and interacting more with each other than with large, gene-poor chromosomes, and *vice versa*; Lieberman-Aiden et al. 2009; Bertero et al. 2019). This step allows the algorithm to focus on interactions that stand out from the noise. Third, the negative log p-values are used as weights for the network edges and subsequently refined using a "donut filter" (Rao et al. 2014) to highlight points that are local maxima. The post-processed Hi-C interaction matrix is finally represented as a weighted graph, in which each node corresponds to a bin and the weight on each edge corresponds to the negative log p-value computed in the previous step. Trans-C then carries out a random walk with restarts algorithm which exploits global patterns of connectivity on the graph. Each walk is initiated from a randomly selected "seed" locus and moves from a node to a neighboring one probabilistically based on the weight of the edge. A parameter α controls the probability that the walk will restart at a new, randomly selected seed locus. Mathematically, as an infinite number of walks are performed, the frequency with which each node is visited converges to a stationary distribution. This can be computed analytically using the Perron-Frobenius theorem. We use the stationary distribution to obtain a ranked list of *trans* interacting bins (Fig 1D), because the most frequently visited nodes are the ones that interact most strongly with the seed loci. Highly ranked genes are most likely to be functionally related with the seed loci, and therefore a putative clique is obtained by extracting the top ranked loci.

Trans-C uncovers the clustering of *var* genes critical for *P. falciparum* immune evasion

Having developed trans-C, we set out to test its ability to uncover known sets of loci that interact together in *trans* in three different organisms. First, we focused on the protozoan *Plasmodium falciparum*, the parasite responsible for the most lethal form of malaria. The three-dimensional organization of the *P. falciparum* genome is strongly associated with gene expression (Ay et al. 2014), particularly for genes involved in pathogenesis, immune evasion, and master regulation of gene expression (Bunnik et al. 2018). Among these are the variant antigenic repertoire (*var*) genes, a family of 60 virulence genes responsible for the antigenic variation of the parasite and evasion of the host immune system. Only a single *var* gene is active at a given time, the other *var* genes being maintained in a perinuclear cluster of heterochromatic telomeres (Fig 2A; Duffy et al. 2017). This cluster is an excellent test case to validate the ability of trans-C to uncover a group of biologically important genes that co-localize in 3D from Hi-C data.

To this end, we examined Hi-C for two stages of the *P. falciparum* life cycle, trophozoite and schizont, both of which are characterized by *trans* contacts between *var* genes (Ay et al. 2014). To visually highlight the *var* cluster (Gardner et al. 2002), we extracted the bins containing *var* genes and also drew 60 bins at random from the full set of genomic loci. The submatrix of *trans* contacts formed by the concatenation of the two sets of bins showed a striking contrast between the *var* and non-*var* loci, as anticipated (Fig 2B). Next, we selected three *var* genes at random to act as seed loci and examined whether trans-C (with $\alpha = 0.5$) could automatically identify the remaining 57 *var* gene loci. For comparison, we used a method based on a greedy heuristic which iteratively selects the bins that interact most strongly with the selected loci (Supplemental Methods). For each approach, we plotted a receiver operating characteristic (ROC) curve, in which each element is a genomic bin, labeled as positive (*var* gene) or negative (other loci; Fig 2C and Supplemental Fig S1A). In both *Plasmodium* life stages, trans-C quickly found the majority of the *var* genes by ranking their corresponding bins highly: of the top 50

bins, 28 contained a *var* gene, and all 60 *var* genes were recovered within the top 280 bins. Trans-C outperformed the greedy heuristic baseline, with an area under the ROC curve (AUROC) of 0.94 compared to 0.88 for the trophozoite analyses (Fig 2C), and similar findings in schizont (Supplemental Fig S1A). This demonstrates that a random walk approach is more suited to the task of identifying *trans* cliques even in the context of a clear example.

As a negative control, we run trans-C starting from three seed loci that were randomly re-selected until a trio could be identified so that its *trans* subnetwork showed the same or greater total interaction strength as the one for the three *var* genes previously used as seeds. We repeated this procedure, which henceforth will be referred to as "matched random control", for a total of 1000 times in order to estimate empirical p-values for the *var* genes-associated subnetwork: this proved extremely significant (p-value = 3×10^{-165} and 3×10^{-171} for trophozoite and schizont, respectively; Fig 2C and Supplemental Fig S1A). Visualization of the *var* genes-associated *trans* interaction subnetworks for the top 40 loci ranked by trans-C in both plasmodium life cycle stages showcased the remarkable intricacy of highly significant contacts (Fig 2D and Supplemental Fig S1B).

Identification of Greek islands regulating the expression of mouse olfactory receptor genes

To further validate trans-C we turned to the mouse and its larger diploid nuclear genome. In mouse olfactory sensory neurons (mOSN), chromatin regions associated to olfactory receptor gene clusters from 18 chromosomes make specific and robust inter-chromosomal contacts that increase in strength with differentiation (Lomvardas et al. 2006; Markenscoff-Papadimitriou et al. 2014; Monahan et al. 2017). These contacts are orchestrated by intergenic olfactory receptor enhancers that form a multi-chromosomal super-enhancer driving the monoallelic and stochastic expression of a single mouse olfactory receptor gene (Fig 3A; Monahan et al. 2019). The mOSN-specific *trans* contacts are arguably the strongest *trans* contacts in a mammalian genome known to date. The regions involved in such interactions were dubbed "Greek islands", since they are sprinkled across the chromosomes as the tiny islands are in the Mediterranean sea. Importantly, in horizontal basal cells (HBCs), the quiescent stem cell progenitors of mOSNs, these inter-chromosomal contacts are absent.

We applied trans-C to mOSN Hi-C data (Monahan et al. 2019), randomly selecting five Greek islands from the previously reported list of 63 to use as seeds in order to measure the ability of trans-C to uncover the remaining 58. Besides running a matched random control, as a biological negative control we used HBC Hi-C data. As expected, trans-C successfully found the Greek islands in mOSNs (AUROC = 0.93; p-value = 6×10^{-194}; Fig 3B and Supplemental Table S1), while it failed to do so effectively in HBCs (AUROC = 0.71; Supplemental Fig S2A). At a false positive rate of 10%, 95% of known Greek islands were identified in mOSNs, though we speculate that some of the false positives may actually represent previously unidentified Greek islands.

To visually verify whether trans-C detected specific inter-chromosomal contacts, we selected the top 60 predicted bins from the ranked stationary distribution (30% of which are known Greek islands). For each pair of loci from this set of 60, we extracted from the Hi-C data a 21 by 21 matrix centered at their interaction, and then averaged these matrices (Fig 3C). The resulting contact heatmap exhibited very strong punctuated signal in the middle, suggesting that the top 60 loci ranked by trans-C form specific interactions that are not driven by larger, non-specific "neighborhood" features. Visualization of the Greek islands-associated *trans* interaction subnetwork for the top 40 loci ranked by trans-C in mOSNs revealed a very dense network that greatly increases in significance when HBCs differentiate in mOSNs (Fig 3D). Collectively, trans-C efficiently pinpoints *trans* cliques even in a complex eukaryotic genome.

We also used the Greek island data set in mOSNs to evaluate how strongly the behavior of trans-C depends on its primary parameter, the random walk restart probability α. We varied α between 0 (no restart) to 1 (restart after every step) in small increments. We observed that the performance of trans-C was stable in the range [0.3, 0.7], while it deteriorated significantly in the two extremes when it approached 0 or 1 (Supplemental Fig S2B). This behavior is expected theoretically: when α is close to 0 the random walk restarts infrequently and so its stationary distribution becomes less dependent on the seeding bins and is mostly determined by the topology of the Hi-C graph. At the extreme, when $\alpha = 0$ the walk is "memoryless" and entirely

independent of the starting seed loci. On the other hand, when α is close to 1 there is little or no exploration along the graph. In this setting, the Hi-C data is essentially ignored, and consequently no discoveries can be made. As an additional benchmarking, we evaluated the performance of trans-C with respect to the number of *trans* reads in the input Hi-C matrix. For this, we run trans-C with α = 0.5 using 100% of the *trans* contacts (≈436 M) *versus* decreasing sub-samples down to ≈87 M (Supplemental Fig S2C). Trans-C maintained a comparable performance when 80% of the matrix was used, and an acceptable performance at 60% sub-sampling, while 40% and 20% of contacts proved insufficient, as could be anticipated. Lastly, we re-benchmarked trans-C against the greedy heuristic: also in the context of Greek island discovery in mOSNs, our algorithm delivered a larger AUROC (0.93 *versus* 0.92; Supplemental Fig S2D).

Dissecting the RBM20 splicing factory during cardiomyocyte differentiation

We next sought to explore the sensitivity of trans-C in a more challenging model in the human genome. We previously identified a network of gene loci that increase their association inter-chromosomally during cardiac development of human pluripotent stem cells (hPSCs) and are targets of the muscle-specific splicing factor RBM20 (Fig 4A). Functional experiments indicated that the main RBM20 target, the large *TTN* pre-mRNA (which contains over 100 binding sites for RBM20), nucleates RBM20 foci. Secondary RBM20 targets interact in *trans* with *TTN* at RBM20 foci, which maximizes the efficiency of their alternative splicing (Bertero et al. 2019). We therefore dubbed the network a "*trans* interacting chromatin domain" (TID) and the resulting structure a "splicing factory". Of note, however, the cumulative interaction score of the TID calculated from shallow Hi-C data (≈90 M contacts) was only modestly enriched compared to a null model (p-value = 0.05). Thus, these interactions are less easily detected by Hi-C, and are likely to be much more transient in nature compared to those involving the Greek islands.

We set out to test whether trans-C would re-identify the RBM20 TID in an independent, more deeply sequenced Hi-C dataset of hPSC differentiation into the cardiac lineage (≈3 billion read pairs per sample; Zhang et al. 2019). Besides various progenitors and early hPSC-derived cardiomyocytes (hPSC-CMs), this dataset also contains late hPSC-CMs obtained after 80 days of in *vitro* differentiation. Moreover, older hiPSC-CMs were FACS-purified using an expression reporter for the mature marker ventricular myosin light chain 2 (MLC-2v; *MYL2* gene). We first attempted to recover the *trans* network of 16 RBM20 target genes from our original report, using five of them (*TTN, CACNA1C, CAMK2D, KCNIP2, CAMK2G*) as seeds for trans-C. Figure 4B shows the ROC curve for day 0 (hPSCs), 15 (early CMs) and 80 (late CMs). The best performance was achieved using Hi-C data from day 80 (AUROC = 0.84, p-value = 5×10^{-122}; Supplemental Table S2); second was day 15 (AUROC = 0.78, p-value = 2×10^{-105}; Supplemental Fig S3B); and last was day 0 (AUROC = 0.75, p-value = 6×10^{-101}; Supplemental Fig S3A). The improvement in ROC area as differentiation advances is in line with the important role of RBM20 in cardiac maturation (Guo et al. 2012). RBM20 is not expressed at day 0, moderately expressed at day 15, and maximally expressed at day 80. We note, however, that the performance at day 0 was better than random, suggesting that some structure that brings the loci close together is present even in undifferentiated cells. Benchmarking of trans-C against a greedy heuristic demonstrated a strong increase in performance for late CMs (AUROC 0.84 *versus* 0.72; Supplemental Fig S3C), highlighting the advantages of the approach particularly to find *trans* cliques that do not stand out strongly from the noise.

Encouraged by these results, we decided to use trans-C to expand our knowledge of the RBM20 TID. Our original list of 16 genes was not the result of an unbiased search but rather reflected our prior knowledge of RBM20 biology: these 16 genes were the known splicing targets of RBM20 in both human and rat hearts that were also upregulated in hPSC-CMs. As an alternative strategy to identify genes involved in the RBM20 TID in an unbiased fashion, we hypothesized that such genes would encode for transcripts most strongly bound by RBM20 and thus enriched within the splicing factory. To test this hypothesis we leveraged a recent dataset that measured RBM20 binding to RNAs using enhanced UV crosslinking and immunoprecipitation (eCLIP; Van Nostrand et al. 2016).

We used RBM20 eCLIP data from hPSC-CMs (Fenix et al. 2021) and counted the number of peaks that fall in each genomic bin (mapping RNAs to the encoding DNA loci). We selected the five bins with the most eCLIP peaks, which contained the genes *TTN, SLC8A1, OBSCN, NEAT1*, and *LBD3*. Using these as seed loci, we ran

trans-C with $\alpha = 0.5$ on Hi-C matrices from differentiating hPSC-CMs (Zhang et al. 2019). Our goal was to test whether trans-C would uncover the remaining 202 bins with eCLIP peaks. We note that this experimental setup is very different from the previous one. Here, we used Hi-C data to find binding sites in an orthogonal eCLIP dataset. Moreover, only a single seed locus, *TTN*, was shared between this analysis and the one reported in Figure 4B. The resulting ROC curves (Fig 4C) show the same trend: best performance was at day 80 (AUROC 0.82; p-value = 4×10^{-120}; Supplemental Table S3), second at day 15 (AUROC 0.79; p-value = 1×10^{-106}; Supplemental Fig S3E), and last at day 0 (AUROC 0.72; p-value = 4×10^{-98}; Supplemental Fig S3D), consistent with biological expectations.

Next, we performed a second performance recall analysis in the same *trans* subnetwork identified by trans-C from RBM20 eCLIP data, but in which we restricted the list of RBM20 targets to those whose RNA is bound by RBM20 on at least three sites and is differentially spliced in hPSC-CMs with RBM20 knocked out (Fenix et al. 2021). The resulting list of 45 high confidence RBM20 targets was efficiently recovered in day 80 hPSC-CMs, with AUROC = 0.84 and a p-value of 2×10^{-125} (Supplemental Fig S3F), a performance improvement compared to the full list of RBM20 bound loci.

Similarly to our observation for *trans* contacts between the Greek islands (Fig 3C), the aggregated contact frequency heatmap for loci involved in the RBM20-associated *trans* interaction subnetworks identified from RBM20 eCLIP data showed a clear punctuated pattern, supporting the spatial specificity of these interactions (Fig 4D). Visualization of this subnetwork showed that it is quite dense and that it clearly increases in significance when hPSC differentiate in CMs, and even more when CMs mature (Fig 4E). Similar results were obtained for the subnetwork identified from established RBM20 targets (Supplemental Fig S3G).

Since the ENCODE Project generated a large number of Hi-C matrices for the left ventricle (LV; The ENCODE Project Consortium 2012), we used this model to both evaluate the reproducibility of trans-C and determine whether the RBM20 splicing factory could be identified in adult, fully mature cardiomyocytes. We ran trans-C starting from the same five seed bins prioritized using RBM20 eCLIP data, and compared the resulting list of ranked bins: the Pearson's correlation for analyses in the 10 biologically independent LV samples was very high (range 0.77 – 0.85; Supplemental Fig S3H), while negative control analyses in non-cardiac samples showed a low correlation (range 0.50 – 0.70; Supplemental Fig S3H). The AUROC for recovering high-confidence RBM20-bound mRNAs in LV Hi-C data was high (range 0.73 – 0.79), further supporting the existence of a measurable RBM20-associated *trans* clique also *in vivo*.

In all, we conclude that trans-C captures even weak and/or unstable yet biologically meaningful *trans* subnetworks associated with RNA biogenesis.

Loci strongly bound by DNA-binding proteins often exhibit significant interactions in *trans*

The identification of dense subnetworks of *trans* Hi-C contacts that are enriched for RBM20 targets supports our hypothesis that RNA biogenesis influences 3D chromatin organization by bringing into proximity co-regulated nucleic acids, so as to maximize the efficacy and specificity of their processing (Fig 5A; Bertero 2021). We specifically propose that, as in the case of RBM20, "RNA factories" arise from the clustered binding of *trans*-acting factors to one or more core co-regulated genes and/or their encoded transcripts. These, in turn, recruit accessory targets of the same factors. This hypothesis predicts the existence of both transcription factories specialized for certain transcription factors (TFs) and/or chromatin regulators, and other RNA factories specialized for various RNA-binding and regulatory proteins. We set out to test this hypothesis systematically using trans-C, as an example of its potential applicability to address biological questions.

First, we focused on DNA-binding proteins (DBPs), hypothesizing that the genes most strongly bound by a given DBP would be associated with strong TIDs. To test this notion, we used the most deeply sequenced Hi-C dataset reported to date: an ultra-deep Hi-C map of human GM12878 lymphoblastoid cells (Harris et al. 2023). The ENCODE Project produced chromatin immunoprecipitation sequencing (ChIP-seq) data for 110 DBPs in this cell line (The ENCODE Project Consortium 2012), providing an ample resource to test our hypothesis in a systematic manner. For each ChIP-seq dataset we counted the number of peaks in each genomic bin.

First, we took the 40 bins with the most peaks for each DBP and calculated the weight of the subnetworks formed by these bins. The distribution of this subnetwork weight across all 110 DBPs is shown in pink in Figure 5B. For comparison, we randomly drew 1000 sets of 40 bins and plotted the distribution of their weight in gray. Clearly, the subnetworks of loci selected based on ChIP-seq peak density formed stronger interactions in *trans* than random sets of loci. This is a first important hint that many DBPs may be indeed involved in specific *trans* contact networks.

Second, for each DBP individually, we formed a seed by selecting the five bins with the most peaks from its corresponding ChIP-seq track, and we ran trans-C to identify a set of potential interactors in *trans*. We took the top 40 predicted bins for a given DBP, and observed that these bins were enriched with ChIP-seq peaks not only for the DBP that spawned the seed, but also for ChIP-seq peaks of other DBPs (Supplemental Fig S4A). This is not surprising because many DBPs act in concert, and many loci contain proximal binding sites of several DBPs (Ibarra et al. 2020). When examining the weights of subnetworks formed by trans-C ("Top 5 DBP-bound seeds", orange in Fig 5B), we noted that they were heavier on average than the subnetworks based on the ChIP-seq signal only ("Top 40 DBP-bound loci", pink). This observation validates that trans-C finds loci that interact even more strongly in *trans* with the seed bins than just the bins most bound by the respective DBP.

We also assessed how well trans-C can build dense subnetworks when it is seeded from biologically unrelated loci. To that end, we first drew 1000 times five random loci to use as seeds and ran trans-C (Fig 5B; green). The subnetworks it built were significantly weaker than the DBP-based ones. This is likely due to the fact that a randomly drawn seed set likely includes loci that are not interacting with one another, while the loci in the DBP-based seed tend to have strong interactions in *trans*. Thus, in order to establish a more stringent baseline, for each DBP we performed 1000 matched random controls with seeds of five random bins for each DBP (Fig 5B, yellow). When using this matched random seed, the subnetworks that trans-C found were once again weaker on average than the ones it found using DBP-based seed (Mann-Whitney *U* test p-value = 0.006). At an individual level, the weight of subnetworks for 53% (58 out of 110) DBPs identified from the DBP-based seeds were significantly stronger than those from matched random seeds (p-value < 0.05; Fig 5C). Visualization of the two strongest subnetworks, associated to the TFs PAX5 and MAX, and the two most significant subnetworks compared to their matched random controls, linked to the TF FOS and chromatin regulator MLLT1, demonstrated that these are interconnected with strong significance (Fig 5D and Supplemental Fig 4D).

Next, we performed an additional control where matched random seeds were selected from a network that was previously randomly shuffled to remove all specific signals resulting from inter-chromosomal structure: as could be expected, the resulting cliques were on average the weakest recovered by trans-C (Fig 5B; light blue), and individual comparisons for DBP-associated cliques were all significant compared to this type of control (Supplemental Fig S4B). This confirms that inter-chromosomal genome architecture is far from random, and that trans-C identifies signals much stronger than random noise. In all, these observations confirm the common sense conception that the inter-chromosomal interactions of biologically unrelated loci are mostly noise, while providing more rigorous support to the hypothesis that co-regulated loci are often enriched for *trans* contacts (Bertero 2021).

To provide additional validation, we examined the strength of the DBP-associated cliques in an orthogonal dataset based on split-pool recognition of interactions by tag extension (SPRITE; Quinodoz et al. 2022), a proximity ligation-independent method to detect higher-order interactions within the nucleus. Specifically, we processed a SPRITE interaction matrix for GM12878 cells (Quinodoz et al. 2018), and evaluated whether the cliques trans-C identified using the Hi-C data exhibited strong interactions in this orthogonal SPRITE dataset. All of the 110 DBP-based cliques showed much higher weight than a background of 1000 randomly drawn sets of loci (p-value = 6×10^{-48}, Mann-Whitney *U* test; Supplemental Fig S5A). To use a more stringent background model, we statistically assessed whether the trans-C-derived subnetworks for each of the analyzed DBPs were stronger than their "matched random" controls in the orthogonal SPRITE data (Supplemental Fig S5B). We observed that the majority (30 of 58) of the DBP-based subnetworks that were significantly stronger than their matched random controls in the Hi-C data were also significantly stronger than their matched random controls in the SPRITE data (Supplemental Fig S5C). Notably, our top five most significant cliques were all significant in the SPRITE data. These findings suggest that trans-C reliably identifies cliques that exhibit strong interactions in *trans* also when these are measured by proximity ligation-independent sequencing approaches.

Lastly, we evaluated whether the loci that are part of a trans-C clique are physically closer to one another. To that end, we used orthogonal imaging measurements with multiplexed error-robust fluorescence *in situ* hybridization (MERFISH; Su et al. 2020). We computed the *trans*-interaction proximity matrix for loci in the IMR-90 fibroblast cell line, and downloaded the corresponding Hi-C matrix along with the ChIP-seq tracks for 16 DBPs available in the ENCODE portal. Because the MERFISH study involved only 1041 loci, we devised a two-fold experiment. First, we subsetted the Hi-C matrix to the loci surveyed in the MERFISH study and ran trans-C using the five loci most bound by a given DBP as a seed. We calculated the average *trans*-proximity in the MERFISH dataset for the loci in the subnetworks trans-C found using the Hi-C data, and compared them to 1000 randomly selected sets of loci. We found that the trans-C cliques exhibited significantly higher proximity in the orthogonal imaging dataset (p-value = 3×10^{-11}, Mann-Whitney *U* test; Supplemental Fig S6A). Second, we ran trans-C on the full Hi-C matrix with the five loci most bound by a given DBP as a seed to obtain a ranking of all loci, and then we looked at the proximity of the 40 MERFISH loci that were ranked closest to the top of the trans-C ranking *versus* the proximity of the 40 MERFISH loci that were ranked closest to the bottom. We observed that for all DBPs the trans-C top-ranked MEFISH loci had significantly higher proximity than the bottom-ranked ones (p-value = 3×10^{-5}, binomial test; Supplemental Fig S6B). These data extend the cross-validation of trans-C predictions of DBP-associated *trans* cliques from Hi-C data with orthogonal methods, including those based on imaging.

DNA-binding-protein-associated *trans* cliques are proximal to nuclear speckles

Our hypothesis is that DBP-associated cliques represent specialized RNA factories. To test this, we examined the nuclear localization of loci involved in *trans* interacting subnetworks identified by trans-C. Specifically, we asked whether they are near nuclear speckles, membraneless subnuclear organelles involved in various aspects of DNA and RNA metabolism (Galganski et al. 2017). To that end, we turned to data generated by tyramide signal amplification sequencing (TSA-seq; Chen et al. 2018). TSA-seq is an experimental protocol that provides a "cytological ruler" for estimating mean chromosomal distances from nuclear landmarks genome-wide. We used the $\log_2$ fold change of TSA-seq signal compared to an input control from the lymphoblastoid K562 cell line (Chen et al. 2018). This measurement captures the distance to a target protein from loci genome-wide, with higher values corresponding to shorter distances and lower values to longer distances. First, for each DBP subnetwork we calculated the average TSA-seq signal strength when probing the SON protein, which plays a crucial role in the formation of nuclear speckles. Indeed, the SON TSA-seq score is proportional to the cytological distance of genes from nuclear speckles, and that it can be even calibrated to estimate mean distance in micrometers (Chen et al. 2018). Next, as a control, for each DBP we took the 40 bins ranked lowest by trans-C but containing at least one ChIP-seq peak, and we compared their average SON TSA-seq score to that of the trans-C-identified subnetwork (Fig 5E). We observed a statistically significant shift to higher values for the trans-C subnetworks, supporting the conclusion that these loci are closer to nuclear speckles. An analogous analysis leveraging on TSA-seq data for phosphorylated SC35 (p-SC35), a splicing factor that also marks nuclear speckles, led to similar results (Fig 5F). We noticed that cliques that were significantly stronger than their matched random controls appeared to be particularly close to nuclear speckle markers, particularly SON (Fig 5E). Indeed, the SON TSA-seq score was significantly higher for these cliques compared to the other cliques (Fig 5G). Collectively, these observations suggest that several DBP-associated cliques identified by trans-C represent RNA factories located at nuclear speckles.

Examining the distribution of the trans-C subnetwork weights identified for different DBPs (Fig 5B, orange) we noticed a bimodal distribution, indicating that some DBPs are associated with stronger TIDs. This bimodality did not correlate with differences in the expression level of the two groups of DBPs, nor in their preference to bind to loci in the A or B compartments (Supplemental Fig S4C). Intrinsically disordered regions (IDRs) within proteins, which lack a defined tertiary structure and are thus prone to self-aggregation, are emerging as an important mediator of subcellular condensates involved in multiple aspects of cell function, including nuclear regulations (Hirose et al. 2023; Wright and Dyson 2015). We thus investigated the correlation between the intrinsic disorder in DBP structure and the strength of the trans-C subnetworks they are associated with. For this analysis, we took the DBPs whose seeds gave rise to the strongest and weakest

subnetworks (top and bottom 25% along the Y axis in Fig 5C; Supplemental Table S4). We calculated the average intrinsically disordered protein (IDP) score for each DBP in the two groups (Meszaros et al. 2018), and plotted them in Figure 5H. The difference between the two groups was statistically significant (Mann-Whitney *U* test p-value = 1.4×10^{-5}), suggesting that the DBPs with more intrinsically disordered regions form stronger interactions in *trans*. In all, trans-C allowed us to identify a large set of DBP enriched for IDR regions that are involved in strong TIDs proximal to nuclear speckles and that may thus be important in efficient transcriptional regulation of their target genes.

Selected RNA-binding proteins are associated to significant nuclear speckle-proximal *trans* cliques

Encouraged by the results on DBP subnetworks, we also examined whether RNA-binding proteins (RBPs) are generally associated with TIDs. Only a few RBP binding profiles are available for GM12878. Thus, for this analysis we turned to K562 cells, another human lymphoblastic line, for which ENCODE reports 139 eCLIP datasets and a deep Hi-C matrix. Similar to our DBP analysis, we counted the number of peaks in each genomic bin for each RBP (mapping RNAs to the encoding DNA loci). Then, for each RBP individually, we formed a seed by selecting the five bins with the most peaks and ran trans-C to identify a set of potential interactors in *trans*. To form a null model per RBP, we repeatedly drew 1000 contact frequency-matched random seeds. We report the average total weight of the matched random seeds compared to the RBP seed in Figure 6A. Most RBP subnetworks built by trans-C were comparatively as dense as those from the corresponding matched random control, lying broadly along the y = x line. Nevertheless, several outliers were notably denser. To assess this observation quantitatively we performed a signed ranked test per RBP and FDR controlled the corresponding p-values. Thirteen proteins had corrected p-values lower than 0.05, which we considered as a significance threshold (indicated in red in Fig 6A and listed in Supplemental Table S5). Distinctly from DBPs, RBPs associated with significantly stronger trans-C subnetworks were not characterized by higher IDP scores (Supplemental Fig S7A), indicating that other characteristics may explain their specific behavior in *trans* genome organization.

We repeated the analyses of TSA-seq data and observed a very strong and significant global correlation between the trans-C subnetworks and both SON and p-SC35 signal (Fig 6B and Supplemental Fig S7B). This indicates that the trans-C eCLIP subnetworks are close in cytological distance to nuclear speckles. This same trend manifested for POL1RE TSA-seq signal, a measurement of proximity to transcription factories (Supplemental Fig S7C). We observed the opposite result when we examined the Lamin A TSA-seq signal, indicating that the trans-C subnetworks are significantly further away from the nuclear lamina (Figure 6C). This finding is in line with the notion that sites of active RNA biogenesis are localized in the euchromatic nucleoplasm and away from heterochromatin regions associated with the nuclear lamina.

Visualization of significant RBP-associated subnetworks showed that these are noticeably interconnected (Fig 6D), more so than DBP-associated subnetworks of similar strength (Fig 5D); on the other hand, the individual pairwise interactions were less significant, possibly due to a more transient nature. In all, trans-C allowed us to identify a subset of RBPs associated with nuclear speckle- and transcription factory-proximal TIDs that may contribute to gene regulation.

Discussion

Potential interactions between pairs of loci on different chromosomes occupy 90-95% of the pairwise 3D DNA contact space (Supplemental Table S6) and a sizable fraction of experimentally measured interactions (Supplemental Table S7). Although in certain species whose nucleus is characterized by chromosome territories — such as humans and other mammals — a large fraction of *trans* contacts are likely nonspecific, illuminating an even small fraction of specific and functional inter-chromosomal interactions may provide important advances in our understanding of nuclear mechanisms such as transcription and splicing. In this context, trans-C is an important step towards refined analytical methods to probe the *trans* contact space for functional gene networks.

The study of *trans* contacts requires statistical methods designed for the specific task at hand. Approaches devised for the analysis of *cis* interactions control for some biases that are not applicable to *trans* ones, such as correction for the linear genomic distance between the interacting loci. To date, most analytical tools for Hi-C data have been limited to *cis* interactions (Lin et al. 2019). Recent network-based strategies to study inter-chromosomal interactions from bulk and single cell Hi-C (Kaufmann et al. 2015; Bulathsinghalage and Liu, 2020; Joo et al. 2023) proposed probabilistic models that focus on identifying large patterns of *trans* contacts (i.e., those involved in nuclear speckles and nucleoli) rather than small sets of interactions linked to a specific process. Trans-C, in contrast, controls for chromosome territory biases to identify cliques that "stand out" from other *trans* interactions resulting from the random intermixing of neighboring chromatin domains. Given that the combinatorial number of possible sets is astronomical (4.7×10^{131} for sets of 40 loci in the human genome binned at 100 kb resolution), this problem cannot be solved directly. Trans-C addresses this challenge by applying a random walk algorithm to obtain a highly reproducible, approximate solution.

We first validate the ability of trans-C to detect known examples of functional *trans* contacts. We find that it outperforms a simple greedy heuristic even in the case of the small haploid genome of *P. falciparum* (22.9 Mb), which is characterized by remarkable *trans* contacts among *var* genes. In more complex and larger mammalian genomes, trans-C not only identifies with high precision the mOSN Greek islands, but also the less striking example of *trans* contacts represented by the RBM20 splicing factory. Thus, trans-C may find applicability across nuclear genomes with different sizes and types of organization, and *trans* contacts of varying strength.

We demonstrate the broader utility of trans-C by using it to systematically search for *trans* cliques around loci most strongly bound by one of many DBPs or RBPs. These analyses support the existence of a large number of statistically significant TIDs readily measurable from Hi-C data, particularly in the case of intrinsically disordered DBPs. Orthogonal analyses of TSA-seq data confirm that such loci are proximal to nuclear speckles. The concept of "bookmarked" transcription factories, Pol II clusters that are specifically enriched for a set of DBPs and their target loci, was proposed over a decade ago (Cook 2010). However, examples of this phenomenon have been sparse (reviewed in Bertero 2021). Our analysis of 110 DBPs provides an important piece of evidence to support this model for over 50% of such DBPs, including leukemia-associated TFs (i.e., PAX5, MAX, and FOS) and chromatin regulators (i.e., MLLT1; Sigvardsson 2023; Zhou et al. 2018). Nevertheless, a mechanistic dissection of these leads will be required to further validate this model.

The few RBPs associated with significant TIDs are involved in a wide variety of functions. Not only do we identify several factors involved in major and minor spliceosomes (PRPF8 and BUD13), but we also identify alternative splicing regulators (ZRANB2), a multifunctional RNA processing factor (TARDBP), a component of the RNA exosome complex (EXOSC10), a ribosomal protein (RPS11), and even a DNA helicase involved in homologous recombination (WRN). We speculate that these factors exemplify a wide range of chromatin structures involving both *cis* and *trans* interactions that regulate not only transcription but also other aspects of nucleic acid biology such as DNA replication and repair, or various aspects of RNA biogenesis. In line with this hypothesis, recent evidence published during the revision of our manuscript supports the notion that genome organization around nuclear speckles drives mRNA splicing efficiency (Bhat et al. 2024). Notably, several of the RBPs highlighted by our trans-C analysis are known to be mutated in severe human monogenic diseases: PRPF8 in retinitis pigmentosa (McKie et al. 2001), WRN in Werner syndrome (Yu et al. 1996), and TARDBP in amyotrophic lateral sclerosis (Sreedharan et al. 2008). Moreover, mutations in ZRANB2 have been linked to unfavorable prognosis in breast and liver cancer (Tanaka et al. 2020), while RPS11 has been shown to be a key player in poor outcomes of glioblastoma patients (Dolezal et al. 2018). Whether disorganization of *trans* genome architecture is implicated in the pathogenesis of these diseases is an interesting topic for future investigation.

Using trans-C we confirmed the existence of significant RBM20-associated *trans* cliques in both hPSC-CMs from a different laboratory and *in vivo* samples of the human left ventricle. These findings support the physiological relevance of muscle-specific inter-chromosomal splicing factories involving RBM20. We have previously shown that preventing *TTN* transcription disrupts RBM20 clustering, decreases the proximity of RBM20 targets to the *TTN* locus, and impairs their RBM20-dependent alternative splicing in *trans* (Bertero et al. 2019). *TTN* is the most commonly mutated gene in both familial and sporadic dilated cardiomyopathy (DCM; Herman et al. 2012; Kayvanpour et al. 2017). While RBM20 is mutated in only ⍰2% of DCM patients, it leads to

a particularly aggressive disease characterized by conduction system disorders (≈30%), malignant ventricular arrhythmia (≈44%) and a rapid progression to heart failure (Kayvanpour et al. 2017; Refaat et al. 2012). We and others recently showed that RBM20 mutations in the RS domain hotspot lead to nuclear mislocalization of RBM20 and severe changes in gene expression (Fenix et al. 2021; Schneider et al. 2020). We speculate that these or other mutations in *RBM20*, *TTN*, and/or other RBM20-associated targets may lead to disease in part through disruption of inter-chromosomal genome architecture. Trans-C will be a useful instrument in testing this hypothesis.

Although we validate and apply trans-C using seed loci selected from *a priori* hypotheses about inter-chromosomal genome architecture, either related to specific genes or to a general mechanism, trans-C can be run using all possible sets of seeds of a given size to conduct discovery. However, this approach is computationally challenging since the number of sets of possible seeds can become combinatorially very large. Moreover, the strongest cliques may not necessarily be the most biologically interesting, as showcased by our example for *RBM20*, which would not have stood out in an unbiased analysis of all hPSC-CM cliques.

It is important to point out that while trans-C identifies sets of loci that significantly interact with one another, this does not rule out the possibility that some of these interactions may be biologically unrelated. Indeed, in many cases, if we select seed bins at random, requiring only that they display *trans* interactions comparable in strength to those involving genes most strongly bound by DBPs or RBPs, we observe that trans-C sometimes identifies strong cliques. This observation is in line with the understanding that in mammals active chromatin tends to be situated at the periphery of chromosome territories (Di Pierro et al. 2016, 2017; Cheng et al. 2020; Su et al. 2020). Thus, while statistical assessment of trans-C results can allow the identification of cliques that are significantly stronger than matched random controls, a sizable portion of the signal is likely to nevertheless arise from compartmental interactions. Accordingly, predictions of novel cliques should be confirmed *via* orthogonal methods or experimentally validated for their biological significance, particularly if trans-C is applied to discovery research with no *a priori* hypothesis.

Another limitation to keep in mind is that the performance of trans-C analyses is strongly dependent on the sequencing depth of the Hi-C matrices. This could represent a bottleneck, since the generation of ultra-deep matrices not only requires substantial resources, but also large enough cell numbers to capture a sufficient number of contacts for each locus. For rare samples, this may be infeasible even if the economic resources were available. When challenged by this situation, a compromise would be to reduce the resolution of genomic binning at the expense of increased noise and more complex biological interpretation of results.

Overall, our work focuses on poorly studied between-chromosome contacts and provides an efficient computational framework for identifying potentially biologically important sets of loci that interact in *trans*. We demonstrate the flexibility and sensitivity of trans-C and provide examples of how our approach can be used to identify candidate gene sets for subsequent hypothesis-driven studies. Application of trans-C to the growing number of Hi-C datasets from the ENCODE (The ENCODE Project Consortium 2012) and 4D Nucleome consortia (Reiff et al. 2022) will reveal novel cell- or disease state-specific *trans* networks. We also provide preliminary evidence that trans-C also allows exploration of SPRITE data (Quinodoz et al. 2018); minor adaptations of the approach will enable investigation of other proximity-ligation independent assays, such a GAM (Beagrie et al. 2017), and will collectively offer the potential to accurately characterize inter-chromosomal architecture at varying spatial resolutions.

Methods

Overview

The full mathematical formulation of trans-C is reported in the Supplemental Methods. In short, trans-C takes as input a Hi-C contact matrix H of interaction counts and an initial set S of loci of interest ("seed loci"); after processing, it outputs a set of loci U (containing S) that interact strongly together in *trans*. In practice, we model the Hi-C interaction matrix H as a weighted graph $G = (V,E,W)$, in which nodes V correspond to the genomic loci (bins) of H, edges E between pairs of nodes correspond to interactions between their respective loci, and weights W on the edges reflect the strength of the interactions represented by the edges. For instance, the weight w_{ij} on edge e_{ij} between loci i and j corresponds to the Hi-C matrix entry h_{ij}. The goal of trans-C is to

find a subset of loci that exhibit strong inter-chromosomal contacts. To solve this problem, which is computationally intractable to solve exactly, we employ a random walk with restart algorithm over the graph G. In essence, this reformulates the problem as a dense subgraph optimization.

Random walk with restart

The random walk traverses the graph by moving probabilistically from one node to another. The walk is initiated from a specified set S of seed loci. The goal of the random walk is to highlight the nodes that are strongly connected to those in S (Hristov et al. 2020). At each step, with a fixed probability α the walk restarts from a randomly selected seed locus, and with probability $1 - \alpha$ the walk moves to a neighboring node picked probabilistically based upon the weights W. Specifically, if $N(i)$ are the nodes that i interacts with, then the walk goes from node i to node $j \in N(i)$ with probability proportional to w_{ij}. That is, for any node i, if at time t the walk is at i, then we calculate the probability p_{ij} that it will transition to node j at time $t + 1$ using only W and α. Hence, the random walk is fully described by a stochastic transition matrix P with entries p_{ij}. Importantly, this stochastic matrix P has certain mathematical properties (Supplemental Methods) which guarantee that, by the Perron-Frobenius theorem, the random walk converges. That is, the probability of the walk being at any given node at time t is constant as $t \to \infty$. This probability π, known as the "stationary distribution" of the walk, can be analytically computed. Further, the probability π_i reflects how well the node i is connected to the seed nodes because more strongly connected nodes are more frequently visited. The loci that have the largest probabilities are most frequently visited and, therefore, are more likely to be relevant because they are strongly connected to the seed loci. We use these probabilities as scores to rank all loci and include the top ℓ loci in U, where ℓ is a user-specified parameter. In this work, we use $\ell = 40$ unless otherwise stated.

Data pre-processing

Prior to running the trans-C random walk algorithm, we perform three pre-processing steps on the Hi-C matrix to ensure that the weights W on the edges are not influenced by many of the biases common in Hi-C data.

First, we normalize the matrix using the iterative correction and eigenvalue decomposition (ICE) procedure (Imakaev et al. 2012; Servant et al. 2012). This procedure iteratively normalizes rows and columns of the matrix, equalizing their sum. We note that we carry out this procedure on the entire Hi-C matrix, including *cis* and *trans* contacts.

Second, we adjust the matrix entries to account for the fact that chromosomes tend to occupy chromosome territories; as a result, some pairs of chromosomes interact more frequently. We do this by using a binomial model to estimate interaction p-values based on an empirical null model that accounts for this territorialization (Supplemental Methods).

Third, we process each matrix entry using a "donut filter" as previously described (Rao et al. 2014). This step allows us to emphasize points that are local maxima in the contact map.

Matched random seed

We run trans-C with 1000 matched random seeds to generate a background model of cliques that are seeded at biologically unrelated loci that form a starting network of comparable strength to the loci of interest. This background model allows us to assess statistically (by Mann-Whitney U test) whether the clique trans-C found using the original seed is significantly stronger than the matched random background. Specifically, the procedure is as follow:

1. Given a seed S of b loci $S = (s_1, s_2, \ldots, s_b)$, calculate the strength of that seed $\text{score}(S) = \sum_{i,j \in S} w_{i,j}$
2. Repeat 1000 times:
    2.1. draw sets of b random loci $R = (r_1, r_2, \ldots, r_b)$ until $\text{score}(R) \geq \text{score}(S)$
    2.2. use the set R as a seed to run trans-C to find a clique trans-C(R)

2.3. add trans-C(R) to the background list of cliques obtained from a "matched random seed"

3. Assess statistically whether trans-C(S) is significantly stronger than the matched random background

Clique visualization

To visualize the cliques we use the Cytoscape software version 3.10.1(Shannon et al. 2003).

Datasets

Validation experiments relied on Hi-C data from three organisms available publicly as either MCOOL or HIC files: *P. falciparum* trophozoite and schizont stages, binned at 10 kb resolution (GEO id: GSE126074; Bunnik et al. 2018); mouse olfactory sensory neurons, binned at 250 kb resolution as in the original analyses (4DN Portal id: 4DNFI3M6726I; Monahan et al. 2019); human cardiomyocyte differentiation from embryonic stem cells (4DN Portal id: 4DNFIT5YVTLO, 4DNFIIOUG5RF, and 4DNFI8RH55DO; Zhang et al. 2019). For this last dataset, we used the cooltools package (Abdennur et al. 2024) to extract from each MCOOL file the interaction counts for its corresponding Hi-C matrix binned at 10 kb resolution. Then we aggregated the Hi-C matrix to 100 kb resolution by summing the counts in each ten consecutive bins of size 10 kb. RBM20 eCLIP data were previously reported (GEO id: GSE175886; Fenix et al. 2021), and analyzed as described below using 100 kb bins. Human left ventricle and unrelated tissue controls Hi-C were obtained from the ENCODE portal (ids: ENCFF193CQL, ENCFF546TZN, ENCFF341WOY, ENCFF033WGK, ENCFF294GFP, ENCFF294GFP, ENCFF251VFA, ENCFF556RLR, ENCFF591MHA, ENCFF004YZQ; The ENCODE Project Consortium 2012), and binned at 100 kb resolution.

Discovery analyses involved ChIP-seq data for 110 human DBPs in the GM12878 cell line and 139 eCLIP for RBPs in the K562 cell line from the ENCODE portal (ids listed in Supplemental Tables 4 and 5). We used the IDR thresholded peaks provided by ENCODE. We split the human linear genome in 100 kb bins, and for a given protein t counted the number of peaks in each bin, producing a count vector C_t. For the DBP analysis in the GM12878 cell line we used an ultra-deep sequenced Hi-C matrix (ENCODE id: ENCSR410MDC; Harris et al. 2023), which contains 3.7 billion *trans* contacts. We performed our analysis at 100 kb resolution, which results in non-zero counts for 85% of all pairwise *trans* contacts. For the RBP analysis in the K562 cell line we used an intact Hi-C matrix (ENCODE id: ENCFF621AIY), which has 360 million *trans* contacts and was binned at 100 kb resolution.

For the SPRITE analysis, we downloaded the processed SPRITE interaction matrix in GM12878 cells (4DN Portal id: 4DNFIUOOYQC3), and normalized it using the steps described above as done for the Hi-C matrices, binning at 100 kb resolution. For the imaging analysis, we downloaded the computed (x,y,z) coordinates (files: genomic-scale.tsv and genomic-scale-with transcription-and-nuclearbodies.tsv available at https://zenodo.org/records/3928890) of 1041 loci studied by MERFISH (Su et al. 2020), and we used the scripts provided by the authors to compute the *trans*-interaction proximity matrix. Since the study was done in the IMR-90 fibroblast cell line, we used a corresponding Hi-C matrix (ENCODE id: ENCFF281ILS) and ChIP-seq data (ENCODE ids: ENCFF483ERE, ENCFF459DPT, ENCFF470FUH, ENCFF770ISZ, ENCFF585XWV, ENCFF567GON, ENCFF124ORZ, ENCFF170WDS, ENCFF718BQI, ENCFF566MPI, ENCFF890WEE, ENCFF150MNG, ENCFF453XKM, ENCFF786CKM, ENCFF448ZOJ, ENCFF699YDJ). For the nuclear speckles analysis, we used TSA-seq data in K562 cells (4DN Portal ids: 4DNFI2WK5IVI, 4DNFI1WULK53, 4DNFII37TNR5, 4DNFIXWDLHDL).

Software Availability

The trans-C code and the custom scripts used for data processing and figure preparation are available in the https://github.com/Noble-Lab/trans-C repository as well as a Supplemental Code file.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

We thank Irene Farabella for critical advice on MERFISH data analyses and Łukasz Truszkowski for insightful discussions and manuscript proofreading. The study has received funding from the European Research Council (ERC) under the European Union's Horizon Europe research and innovation programme (Grant agreement No. 101076026; Project Acronym TRANS-3; AB). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them. We also acknowledge financial support from the National Institutes of Health (award UM1HG011531; WSN) and the Giovanni Armenise-Harvard Foundation (Career Development Award 2021; AB).

Authors' contributions

B.H.H. performed all of the analyses and wrote the first draft of the manuscript. W.S.N. supervised the analyses, edited the manuscript, and obtained funding. A.B. conceptualized the study, co-supervised the analyses, assembled the final figures, edited the manuscript, and obtained funding.

References

Abdennur N, Abraham S, Fudenberg G, et al (2024) Cooltools: enabling high-resolution hi-c analysis in python. *PLOS Computational Biology* **20**(5):e1012067

Ay F, Bunnik EM, Varoquaux N, et al (2014) Three-dimensional modeling of the *P. falciparum* genome during the erythrocytic cycle reveals a strong connection between genome architecture and gene expression. *Genome Research* **24**:974–988

Beagrie RA, Scialdone A, Schueler M, et al (2017) Complex multi-enhancer contacts captured by genome architecture mapping. *Nature* **543**(7646):519–524

Bertero A (2021) Rna biogenesis instructs functional inter-chromosomal genome architecture. *Frontiers in Genetics* **12**:645863

Bertero A, Fields PA, Ramani V, et al (2019) Dynamics of genome reorganization during human cardiogenesis reveal an RBM20-dependent splicing factory. *Nature Communications* **10**(1):1538

Bhaskara A, Charikar M, Chlamtac E, et al (2010) Detecting high log-densities: an O($n^{1/4}$) approximation for densest k-subgraph. In: *Proceedings of the forty-second ACM symposium on Theory of computing*, pp **201–210**

Bhat P, Honson D, Guttman M (2021) Nuclear compartmentalization as a mechanism of quantitative control of gene expression. *Nature Reviews Molecular Cell Biology* **22**(10):653–670

Bhat P, Chow A, Emert B, et al (2024) Genome organization around nuclear speckles drives mrna splicing efficiency. *Nature* pp **1–9**

Bulathsinghalage C, Liu L (2020) Network-based method for regions with statistically frequent interchromosomal interactions at single-cell resolution. *BMC Bioinformatics* **21**(14):1–15

Bunnik EM, Cook KB, Varoquaux N, et al (2018) Changes in genome organization of parasite-specific gene families during the *Plasmodium* transmission stages. *Nature Communications* **15**(9):1910

Charikar M (2000) Greedy approximation algorithms for finding dense components in a graph. In: International workshop on approximation algorithms for combinatorial optimization, *Springer*, pp **84–95**

Chen Y, Zhang Y, Wang Y, et al (2018) Mapping 3d genome organization relative to nuclear compartments using tsa-seq as a cytological ruler. *J Cell Biol* **217**(11):4025–4048

Cheng RR, Contessoto VG, Lieberman Aiden E, et al (2020) Exploring chromosomal structural heterogeneity across multiple cell lines. *Elife* **9**:e60312

Cook KB, Hristov BH, Le Roch KG, et al (2020) Measuring significant changes in chromatin conformation with accost. *Nucleic acids research* **48**(5):2303–2311

Cook PR (2010) A model for all genomes: the role of transcription factories. *Journal of molecular biology* **395**(1):1–10

Cremer T, Cremer M (2010) Chromosome territories. Cold Spring Harb Perspect Biol 2(3):a003889

De Wit E, Bouwman BA, Zhu Y, et al (2013) The pluripotent genome in three dimensions is shaped around pluripotency factors. *Nature* **501**(7466):227–231

Di Pierro M, Zhang B, Aiden EL, et al (2016) Transferable model for chromosome architecture. *Proceedings of the National Academy of Sciences* **113**(43):12168–12173

Di Pierro M, Cheng RR, Lieberman Aiden E, et al (2017) De novo prediction of human chromosome structures: Epigenetic marking patterns encode genome architecture. *Proceedings of the National Academy of Sciences* **114**(46):12126–12131

Dixon JR, Selvaraj S, Yue F, et al (2012) Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**(7398):376–380

Dolezal JM, Dash AP, Prochownik EV (2018) Diagnostic and prognostic implications of ribosomal protein transcript expression patterns in human cancers. *BMC Cancer* **18**:1–14

Duan A, Wang H, Zhu Y, et al (2021) Chromatin architecture reveals cell type-specific target genes for kidney disease risk variants. *BMC biology* **19**(1):1–13

Duffy MF, Tang J, Sumardy F, et al (2017) Activation and clustering of a plasmodium falciparum var gene are affected by subtelomeric sequences. *The FEBS Journal* **284**(2):237–257

Fenix AM, Miyaoka Y, Bertero A, et al (2021) Gain-of-function cardiomyopathic mutations in rbm20 rewire splicing regulation and re-distribute ribonucleoprotein granules within processing bodies. *Nature Communications* **12**(1):6324

Galganski L, Urbanek MO, Krzyzosiak WJ (2017) Nuclear speckles: molecular organization, biological function and role in disease. *Nucleic acids research* **45**(18):10350–10368

Gardner MJ, Hall N, Fung E, et al (2002) Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* **419**(6906):498–511

Gibson D, Kumar R, Tomkins A (2005) Discovering large dense subgraphs in massive graphs. In: *Proceedings of the 31st international conference on Very large data bases*, pp **721–732**

Guo W, Schafer S, Greaser ML, et al (2012) Rbm20, a gene for hereditary cardiomyopathy, regulates titin splicing. *Nature Medicine* **18**(5):766–773

Hafner A, Boettiger A (2023) The spatial organization of transcriptional control. *Nature Reviews Genetics* **24**(1):53–68

Harris HL, Gu H, Olshansky M, et al (2023) Chromatin alternates between a and b compartments at kilobase scale for subgenic organization. *Nature communications* **14**(1):3303

Herman DS, Lam L, Taylor MR, et al (2012) Truncations of titin causing dilated cardiomyopathy. *New England Journal of Medicine* **366**(7):619–628

698 Hildebrand EM, Dekker J (2020) Mechanisms and functions of chromosome compartmentalization. *Trends in*
699 *biochemical sciences* **45**(5):385–396

700 Hirose T, Ninomiya K, Nakagawa S, et al (2023) A guide to membraneless organelles and their various roles in
701 gene regulation. *Nature Reviews Molecular Cell Biology* **24**(4):288–304

702 Hoencamp C, Dudchenko O, Elbatsh AM, et al (2021) 3d genomics across the tree of life reveals condensin ii as a
703 determinant of architecture type. i **372**(6545):984–989

704 Hristov BH, Chazelle B, Singh M (2020) A guided network propagation approach to identify disease genes that
705 combines prior and new information. In: *International Conference on Research in Computational Molecular*
706 *Biology*, Springer, pp **251–252**

707 Ibarra IL, Hollmann NM, Klaus B, et al (2020) Mechanistic insights into transcription factor cooperativity and its
708 impact on protein-phenotype interactions. *Nature Communications* **11**(1):124

709 Imakaev M, Fudenberg G, McCord RP, et al (2012) Iterative correction of Hi-C data reveals hallmarks of
710 chromosome organization. *Nature Methods* **9**:999–1003

711 Ito K, Sanosaka T, Igarashi K, et al (2016) Identification of genes associated with the astrocyte-specific gene
712 gfap during astrocyte differentiation. *Scientific reports* **6**(1):23903

713 Jerkovic I, Cavalli G (2021) Understanding 3d genome organization by multidisciplinary methods. Nature
714 Reviews *Molecular Cell Biology* **22**(8):511–528

715 Joo J, Cho S, Hong S, et al (2023) Probabilistic establishment of speckle-associated inter-chromosomal
716 interactions. *Nucleic Acids Research* p **211**

717 Kaufmann S, Fuchs C, Gonik M, et al (2015) Inter-chromosomal contact networks provide insights into
718 mammalian chromatin organization. *PloS one* **10**(5):e0126125

719 Kayvanpour E, Sedaghat-Hamedani F, Amr A, et al (2017) Genotype-phenotype associations in dilated
720 cardiomyopathy: meta-analysis on more than 8000 individuals. *Clinical Research in Cardiology* **106**:127–139

721 Khuller S, Saha B (2009) On finding dense subgraphs. In*: International colloquium on automata, languages, and*
722 *programming*, Springer, pp **597–608**

723 Krumm T, Duan Z (2019) Understanding the 3D genome: Emerging impacts on human disease. *Seminars in Cell*
724 *& Developmental Biology* **90**:62–77

725 Lieberman-Aiden E, van Berkum NL, Williams L, et al (2009) Comprehensive mapping of long-range
726 interactions reveals folding principles of the human genome. *Science* **326**(5950):289–293

727 Lin D, Bonora G, Yardımcı GG, et al (2019) Computational methods for analyzing and modeling genome
728 structure and organization. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine* **11**(1):e1435

729 Lomvardas S, Barnea G, Pisapia DJ, et al (2006) Interchromosomal interactions and olfactory receptor choice.
730 *Cell* **126**(2):403–413

731 Longo GM, Roukos V (2021) Territories or spaghetti? chromosome organization exposed. *Nature Reviews*
732 *Molecular Cell Biology* **22**(8):508–508

733 Markenscoff-Papadimitriou E, Allen WE, Colquitt BM, et al (2014) Enhancer interaction networks as a means
734 for singular olfactory receptor expression. *Cell* **159**(3):543–557

735 McKie AB, McHale JC, Keen TJ, et al (2001) Mutations in the pre-mrna splicing factor gene prpc8 in autosomal
736 dominant retinitis pigmentosa (rp13). *Human Molecular Genetics* **10**(15):1555–1562

M´esz´aros B, Erd˝os G, Doszt´anyi Z (2018) Iupred2a: context-dependent prediction of protein disorder as a function of redox state and protein binding. *Nucleic Acids Research* **46**(W1):W329–W337

Monahan K, Schieren I, Cheung J, et al (2017) Cooperative interactions enable singular olfactory receptor expression in mouse olfactory neurons. *Elife* **6**:e28620

Monahan K, Horta A, Lomvardas S (2019) Lhx2-and ldb1-mediated trans interactions regulate olfactory receptor choice. *Nature* **565**(7740):448–453

Mostafavi S, Ray D, Warde-Farley D, et al (2008) Genemania: a real-time multiple association network integration algorithm for predicting gene function. *Genome Biology* **9**:1–15

Osborne CS, Chakalova L, Brown KE, et al (2004) Active genes dynamically colocalize to shared sites of ongoing transcription. *Nature Genetics* **36**(10):1065–1071

Osborne CS, Chakalova L, Mitchell JA, et al (2007) Myc dynamically and preferentially relocates to a transcription factory occupied by igh. *PLoS Biology* **5**(8):e192

Papantonis A, Kohro T, Baboo S, et al (2012) Tnfα signals through specialized factories where responsive coding and mirna genes are transcribed. *The EMBO journal* **31**(23):4404–4414

Quinodoz SA, Ollikainen N, Tabak B, et al (2018) Higher-order inter-chromosomal hubs shape 3D genome organization in the nucleus. *Cell* **174**(3):744–757

Quinodoz SA, Bhat P, Chovanec P, et al (2022) Sprite: a genome-wide method for mapping higher-order 3d interactions in the nucleus using combinatorial split-and-pool barcoding. *Nature protocols* **17**(1):36–75

Rao SSP, Huntley MH, Durand N, et al (2014) A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **59**(7):1665–1680

Refaat MM, Lubitz SA, Makino S, et al (2012) Genetic variation in the alternative splicing regulator rbm20 is associated with dilated cardiomyopathy. *Heart Rhythm* **9**(3):390–396

Reiff SB, Schroeder AJ, Kırlı K, et al (2022) The 4d nucleome data portal as a resource for searching and visualizing curated nucleomics data. *Nature Communications* **13**(1):2365

Schaeffer M, Nollmann M (2023) Contributions of 3d chromatin structure to cell-type-specific gene regulation. *Current opinion in genetics & development* **79**:102032

Schneider JW, Oommen S, Qureshi MY, et al (2020) Dysregulated ribonucleoprotein granules promote cardiomyopathy in rbm20 gene-edited pigs. *Nature medicine* **26**(11):1788–1800

Servant N, Varoquaux N, Lajoie BR, et al (2015) HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biology* **16**:259

Shannon P, Markiel A, Ozier O, et al (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research* **13**(11):2498–2504

Sigvardsson M (2023) Transcription factor networks link b-lymphocyte development and malignant transformation in leukemia. *Genes & Development* **37**(15-16):703–723

Sreedharan J, Blair IP, Tripathi VB, et al (2008) Tdp-43 mutations in familial and sporadic amyotrophic lateral sclerosis. *Science* **319**(5870):1668–1672

Su JH, Zheng P, Kinrot SS, et al (2020) Genome-scale imaging of the 3d organization and transcriptional activity of chromatin. *Cell* **182**(6):1641–1659

Takizawa T, Gudla PR, Guo L, et al (2008) Allele-specific nuclear positioning of the monoallelically expressed astrocyte marker gfap. *Genes & development* **22**(4):489–498

Tan J, Shenker-Tauris N, Rodriguez-Hernaez J, et al (2023) Cell-type-specific prediction of 3d chromatin organization enables high-throughput in silico genetic screening. *Nature biotechnology* pp **1–11**

Tanaka I, Chakraborty A, Saulnier O, et al (2020) Zranb2 and syf2-mediated splicing programs converging on ect2 are involved in breast cancer cell resistance to doxorubicin. *Nucleic Acids Research* **48**(5):2676–2693

The ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**:57–74

Van Nostrand EL, Pratt GA, Shishkin AA, et al (2016) Robust transcriptome-wide discovery of rna-binding protein binding sites with enhanced clip (eclip). *Nature Methods* **13**(6):508–514

Weston J, Elisseeff A, Zhou D, et al (2004) Protein ranking: from local to global structure in the protein similarity network. *Proceedings of the National Academy of Sciences* **101**(17):6559–63

Winick-Ng W, Kukalev A, Harabula I, et al (2021) Cell-type specialization is encoded by specific chromatin topologies. *Nature* **599**(7886):684–691

Wright PE, Dyson HJ (2015) Intrinsically disordered proteins in cellular signalling and regulation. *Nature reviews Molecular cell biology* **16**(1):18–29

Yu CE, Oshima J, Fu YH, et al (1996) Positional cloning of the werner's syndrome gene. *Science* **272**(5259):258–262

Zhang Y, Li T, Preissl S, et al (2019) Transcriptionally active herv-h retrotransposons demarcate topologically associating domains in human pluripotent stem cells. *Nature Genetics* **51**(9):1380–1388

Zheng H, Xie W (2019) The role of 3d genome organization in development and cell differentiation. *Nature Reviews Molecular Cell Biology* **20**(9):535–550

Zhou J, Ng Y, Chng WJ (2018) Enl: structure, function, and roles in hematopoiesis and acute myeloid leukemia. Cellular and Molecular *Life Sciences* **75**:3931–3941

Main Figures Legends

Figure 1. The trans-C algorithm. **(A)** Schematic of typical inter-chromosomal genome organization in mammals. Inter-chromosomal (*trans*) interactions mainly involve genomic domains that extrude from chromosome territories and engage with a variety of membraneless structures involved in gene regulation. **(B)** A Hi-C matrix captures the contact frequency of loci in a genome-wide fashion. Besides intra-chromosomal (*cis*) contacts, specific loci can exhibit strong inter-chromosomal (*trans*) contacts among themselves. **(C)** Trans-C employs a random walk algorithm that traverses the Hi-C contact graph choosing to move to a node (bin) probabilistically based on the strength of the edge (interaction). **(D)** The output is a list of loci ranked by how frequently they are visited during the random walk: more frequently visited loci interact more strongly in *trans* as a clique.

Figure 2. Trans-C identifies the *var* genes cluster in *Plasmodium falciparum*. **(A)** Schematic of *P. falciparum* in the trophozoite stage of its red blood cell life cycle, with a zoomed in view of the nucleus highlighting its Rabl-like structure and the clustering of the *var* genes in a repressive heterochromatic cluster. **(B)** Contact heat map comparing *trans* contact counts among all 60 *var* genes *versus* 60 randomly selected 10 kb bins. *Cis* contacts are grayed out. **(C)** Performance evaluation of trans-C-mediated identification of *var* gene clustering. We plot the receiver operating characteristic (ROC) curve for the trophozoite life stage of *P. falciparum*. The *var* genes are uncovered by the random walk algorithm of trans-C with high area under ROC curve (AUROC; 0.94). The cumulative distribution is statistically significant (p-value = 3×10^{-165}) from a null model of 1000 random walks performed from seeds selected randomly but with an equal or greater collective interaction strength (matched random; line reports the average and shaded area the 95% confidence interval). We also report the

performance of a simpler greedy heuristic. **(D)** Visualization of the *var* genes-associated *trans* clique identified by trans-C in *P. falciparum* trophozoite. Nodes are color-coded by chromosome and sequentially numbered based on their relative position along each chromosome (expressed in Mb). Edges are color-coded based on *trans* interaction significance (*cis* contacts are not plotted). The seed loci for the random walk are indicated by solid black lines around the nodes.

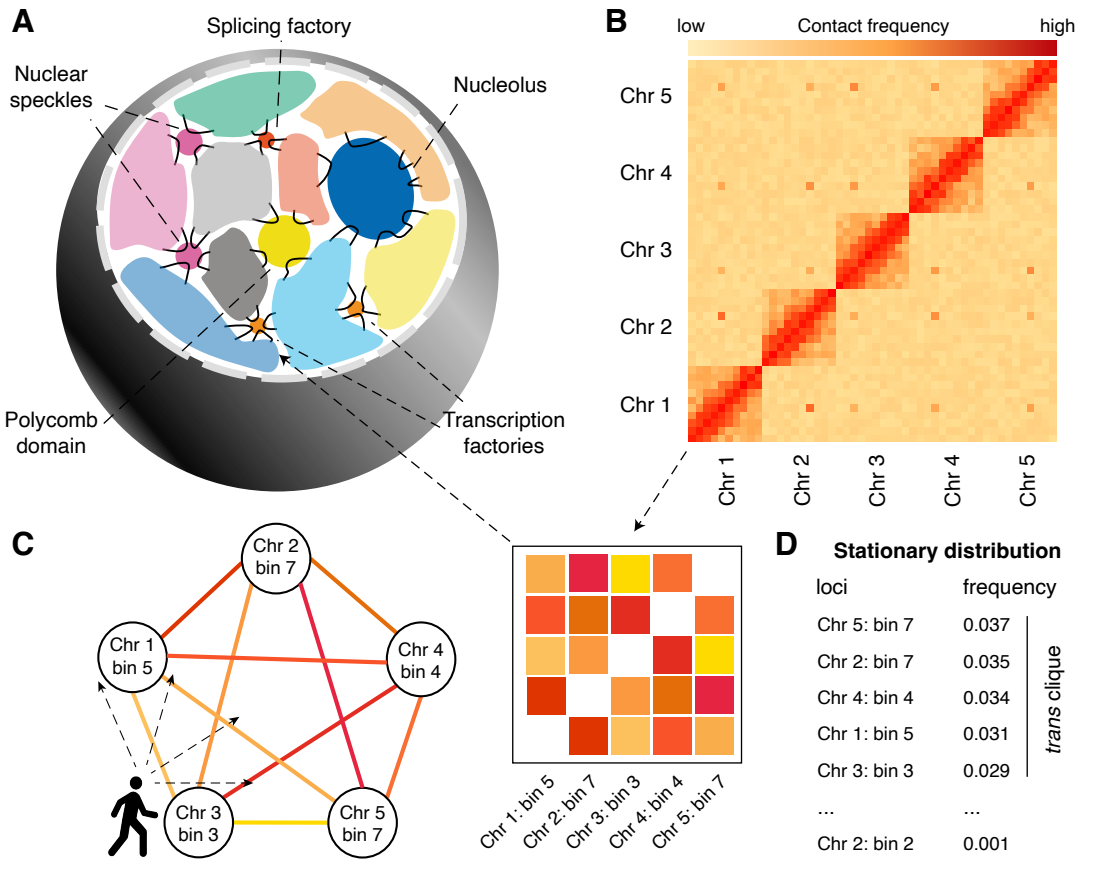
Figure 3. Trans-C identifies the Greek islands cluster in mouse olfactory sensory neurons. **(A)** Schematic of *trans* contacts in a mouse olfactory sensory neuron (mOSN). The Greek islands form a multi-enhancer hub that is segregated from the inactive olfactory receptor (OR) genes. **(B)** Performance evaluation of trans-C-mediated identification of Greek islands clustering. We plot the ROC curve for $\alpha = 0.5$ in mOSNs *versus* their progenitors (horizontal basal cells, HBCs). Trans-C correctly identifies Greek islands clustering specifically in mOSNs in a way that is statistically significant (p-value = 6×10^{-194}) *versus* a matched random seed null model (average and 95% confidence interval from 1000 runs). **(C)** Aggregated heatmap of *trans* contacts among the top 60 loci selected by trans-C in mOSNs. Each square in the grid represents an average 250 kb bin in a Hi-C matrix of 21 × 21 bins centered at each interacting pair of loci (reference). The exhibited spot-like structure highlights the highly specific nature of the inter-chromosomal interactions of the Greek islands. **(D)** Visualization of the Greek island-associated *trans* clique identified in mOSNs by trans-C, showcasing the increased significance of loci interactions after differentiation of HBCs, plotted as described for Figure 2D.

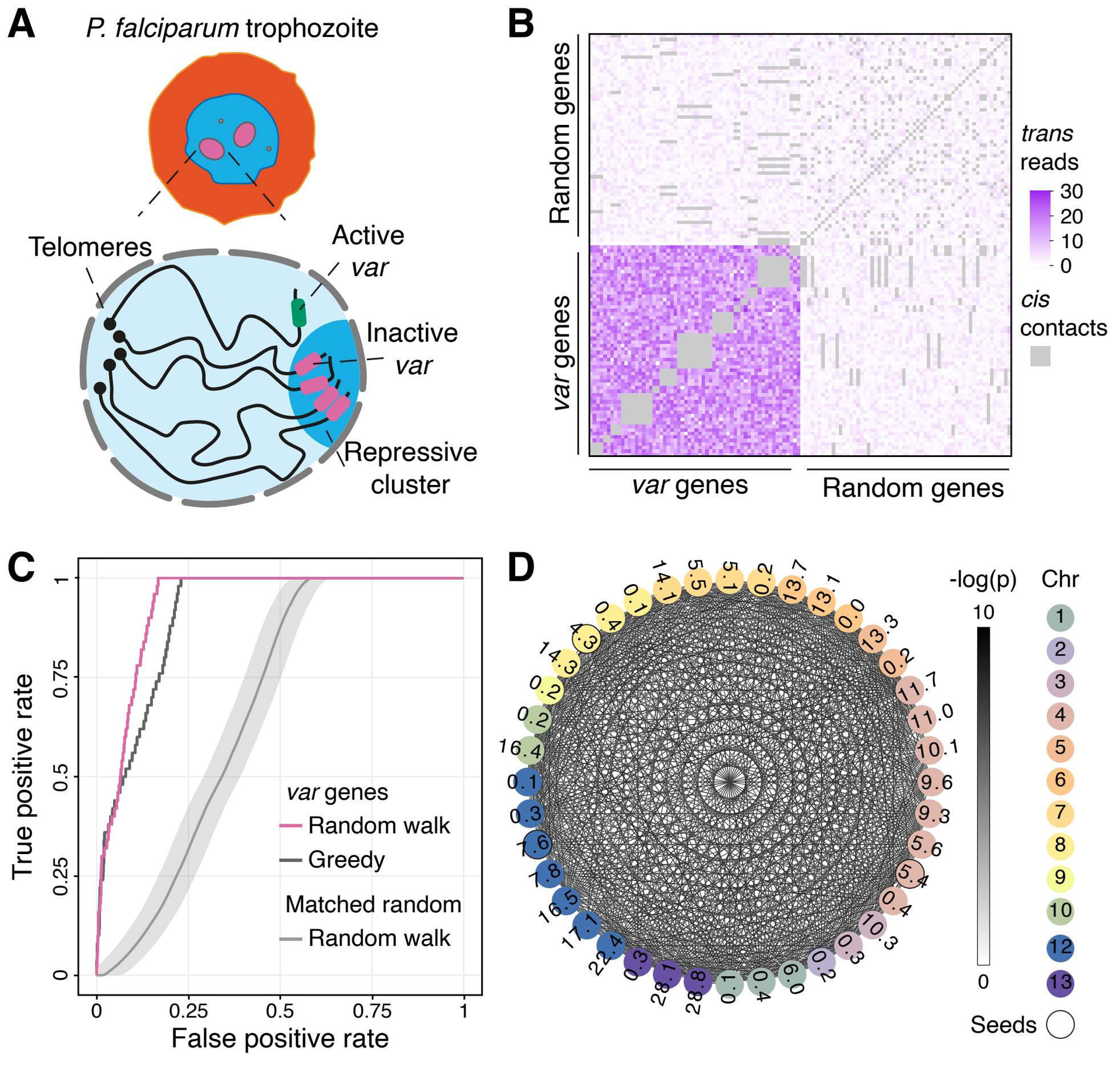
Figure 4. Trans-C identifies the RBM20 splicing factory in human cardiomyocytes. **(A)** Schematic of the RBM20 splicing factory, a muscle-specific inter-chromosomal structure organized by the *TTN* pre-mRNA. This pre-mRNA binds to more than 100 copies of RBM20 and nucleates foci that engage with other RBM20 targets to promote their alternative splicing (blue arrows). **(B)** Performance evaluation of trans-C in uncovering the RBM20 splicing factory in early (day 15) *versus* late (day 80) cardiomyocytes (CMs) differentiated from human pluripotent stem cells (hPSC; also analyzed as "day 0" baseline control). Results for late CMs are statistically significant (p-value = 5×10^{-122}) *versus* a matched random seed null model (average and 95% confidence interval from 1000 runs). Seed loci and ROC curves are based on a list of established RBM20 targets (Bertero et al. 2019). **(C)** Similar to B, but seed loci and ROC curves are based on loci directly bound by RBM20 as determined by eCLIP; p-value = 4×10^{-120}. **(D)** Aggregated heatmap of *trans* contacts between the top sixty loci selected by trans-C in late CMs starting from eCLIP data. Each square in the grid represents an average 100 kb bin in a Hi-C matrix of 41 × 41 bins centered at each interacting pair of loci extracted from the Hi-C data (reference). The denser region in the middle reveals the specific nature of the *trans* interactions at the RBM20 splicing factory. **(E)** Visualization of the RBM20-associated *trans* clique identified by trans-C in late CMs starting from eCLIP data, showcasing the increased significance of loci interactions during hPSC differentiation and CM maturation, plotted as described for Figure 2D.

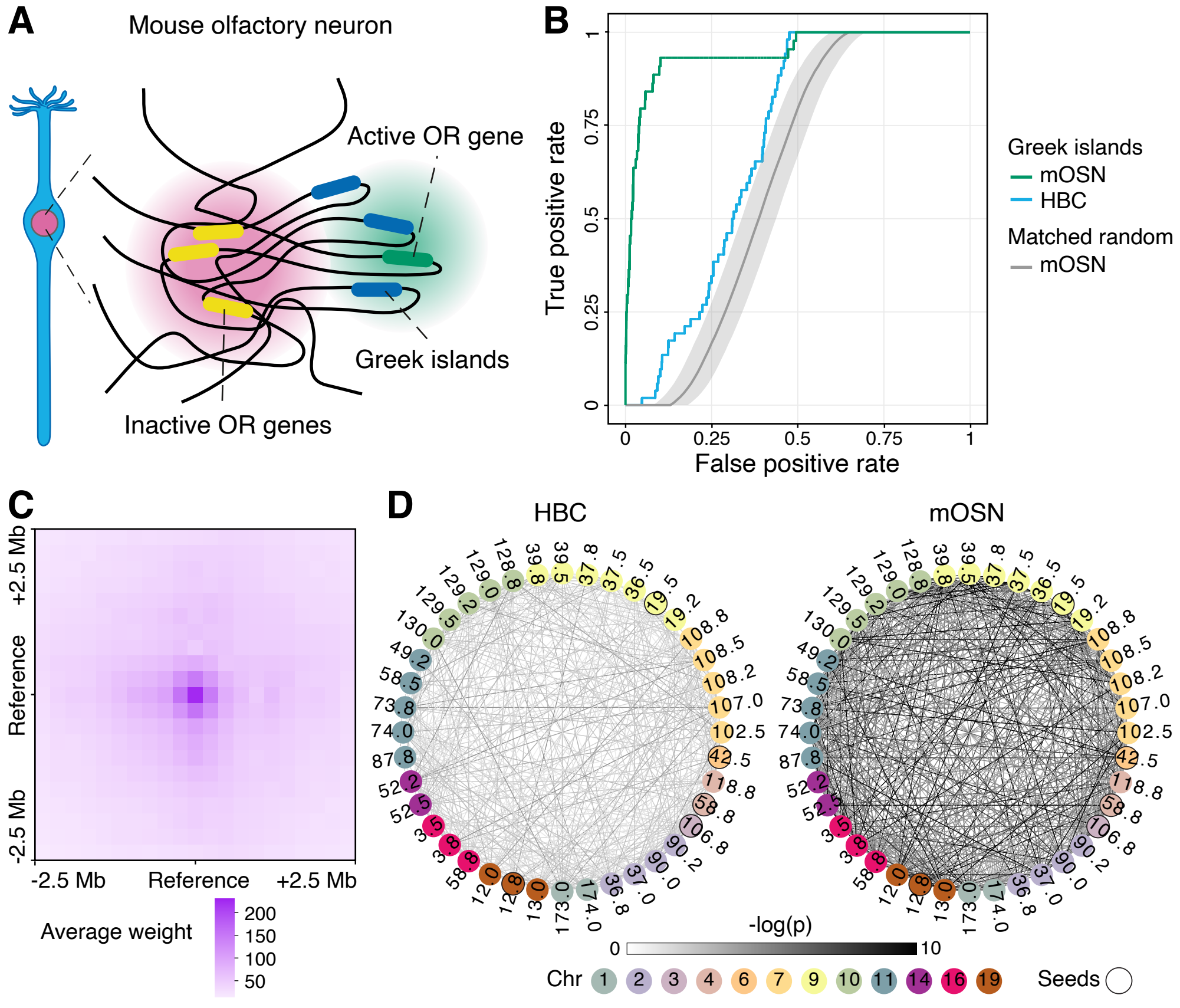
Figure 5. Trans-C identifies DNA-binding protein-associated *trans* cliques proximal to nuclear speckles. **(A)** Schematic of the mechanistic hypothesis for the formation of specialized RNA factories involving *trans* interacting chromatin domains. Multiple copies of *trans*-acting regulatory factors (i.e., transcription or splicing factors) bind to core nucleic acids, aggregate to form new clusters and/or enrich pre-existing ones, and recruit accessory co-regulated nucleic acids. RNA factories promote the efficacy and accuracy of RNA biogenesis processes (thicker black arrows). **(B)** Trans-C-identified subnetworks in lymphoblastoid cells built from loci characterized by strong binding of 110 DBPs have dense contacts. We plot the distribution of subnetwork weights for six types of sets of forty loci: (1 - pink) loci with the highest number of ChIP-seq peaks for a given DBP; (2 - gray) randomly drawn loci; (3 - orange) top loci ranked by trans-C from a seed of five loci with the highest number of ChIP-seq peaks for a given DBP; (4 - yellow) top loci ranked by trans-C from a random seed of five loci whose starting subnetwork weight was matched to the seed of group 3; (5 - green) top loci ranked by trans-C from a seed of five randomly drawn loci; and (6 - light blue) top loci ranked as for group 4 but starting from an interaction matrix that has been randomly shuffled. On average, sets seeded from loci most

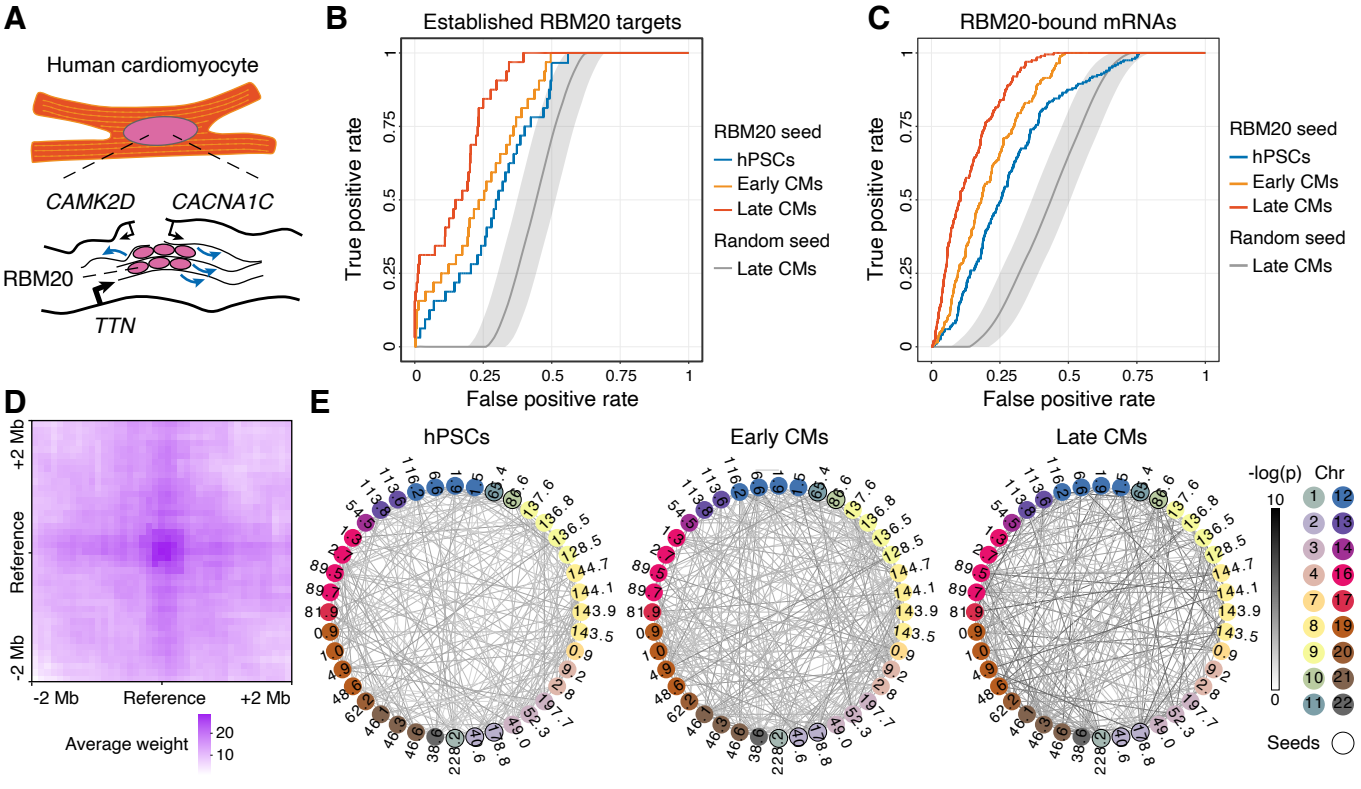
strongly bound by DBPs interact more strongly in *trans* than any of the other five types of sets of loci, including the stringent "matched random" control (p-values by Mann-Whitney *U* test). **(C)** For each DBP analyzed in B, we compare the weights of subnetworks obtained with trans-C from "Top 5 DBP-bound" seeds (single data point) and "Matched random" seeds (average of 1000 subnetworks ± standard deviation). In red are comparisons with significantly different weights (p-value < 0.05 after FDR correction). Shaded areas highlight the top and bottom quartile of DBP-based subnetwork weights. **(D)** Visualization of selected significant DBP-associated *trans* cliques in lymphoblastoid cells, plotted as described for Figure 2D (PAX5 & MAX: strongest cliques; FOS & MLLT1: highest fold change of clique strength over average strength of cliques in the matched random null model). **(E,F)** Proximity to nuclear speckles of loci within trans-C-identified cliques, measured as the average SON and p-SC35 TSA-seq signal for the corresponding genomic regions. For each subnetwork, the signal is compared with that of an equal number of loci at the opposite end of the trans-C ranking. DBP-based cliques are overall significantly more proximal to both SON and p-SC35 than matched control sets (p-values by Mann-Whitney *U* test). Cliques that are significantly stronger compared to the null model in the analysis from panel C (Sign.) are in red, while non-significant ones (N.s.) are in blue. **(G)** The proximity to SON for significant *versus* non-significant cliques from panel C is significantly different by Mann-Whitney *U* test. **(H)** The strongest DBP-based subnetworks correspond to DBPs with a higher intrinsically disordered protein (IDP) score. We plot the IDP scores for DBPs resulting in the bottom and top quartile of DBP-based subnetworks from panel C. The difference is statistically significant by Mann-Whitney *U* test.

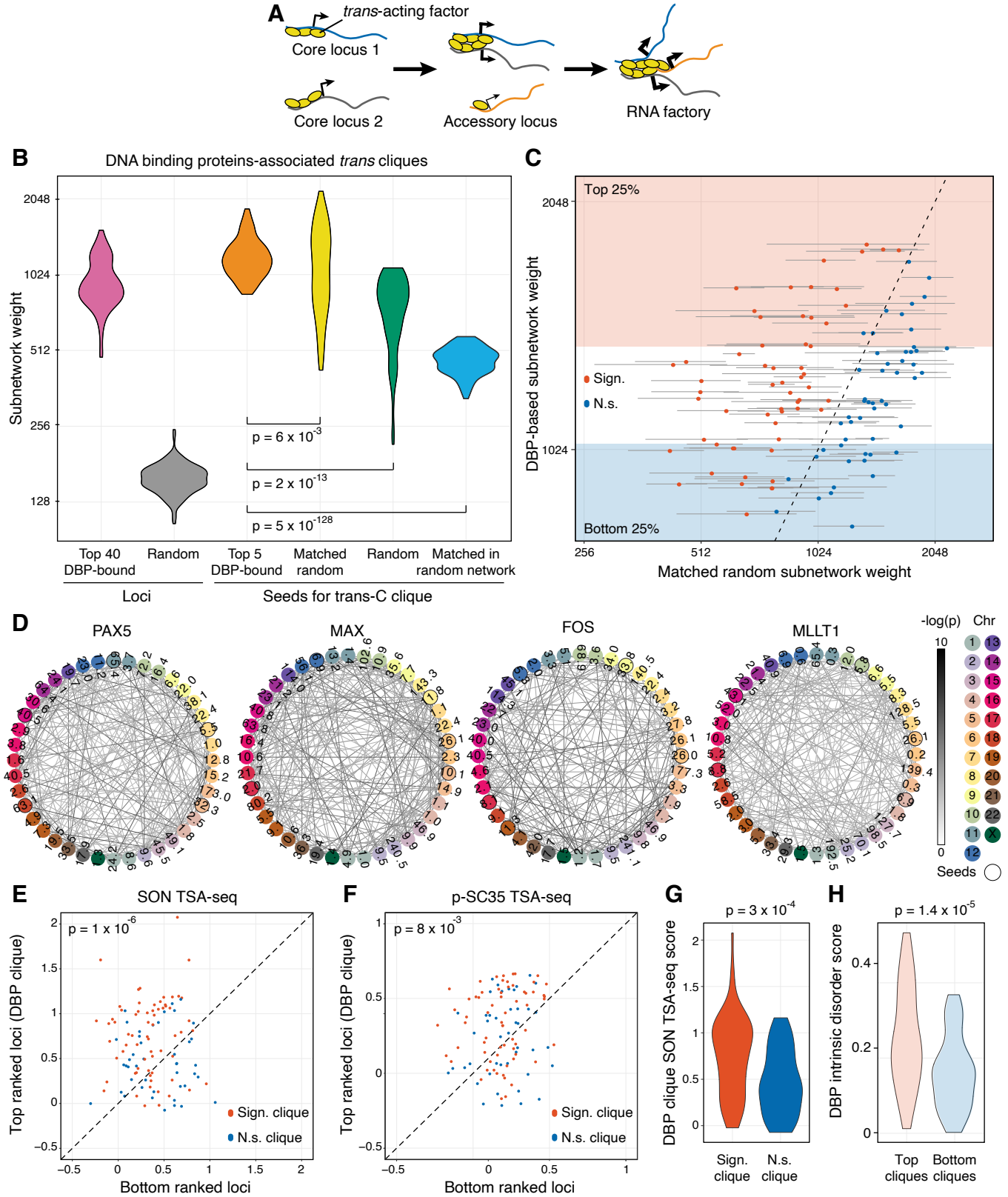
Figure 6. Trans-C identifies RNA-binding protein-associated *trans* cliques proximal to nuclear speckles. **(A)** A subset of trans-C-identified subnetworks in lymphoblastoid cells built from loci characterized by strong binding of RBPs have denser contacts than the corresponding matched random null model. We plot the weight of a RBP-based subnetwork and the average weight of 1000 matched random seed subnetworks (error bars correspond to the standard deviation) for 139 RBPs. In red and listed by name are those with significant p-values after FDR correction. **(B,C)** RBP-associated cliques identified by trans-C are significantly closer to nuclear speckles (stronger SON TSA-seq signal) and significantly further away from the nuclear lamina (weaker Lamin A TSA-seq signal) than matched control sets at the opposite end of the trans-C rankings. **(D)** Visualization of selected significant RBP-associated *trans* cliques in lymphoblastoid cells, plotted as described for Figure 2D.

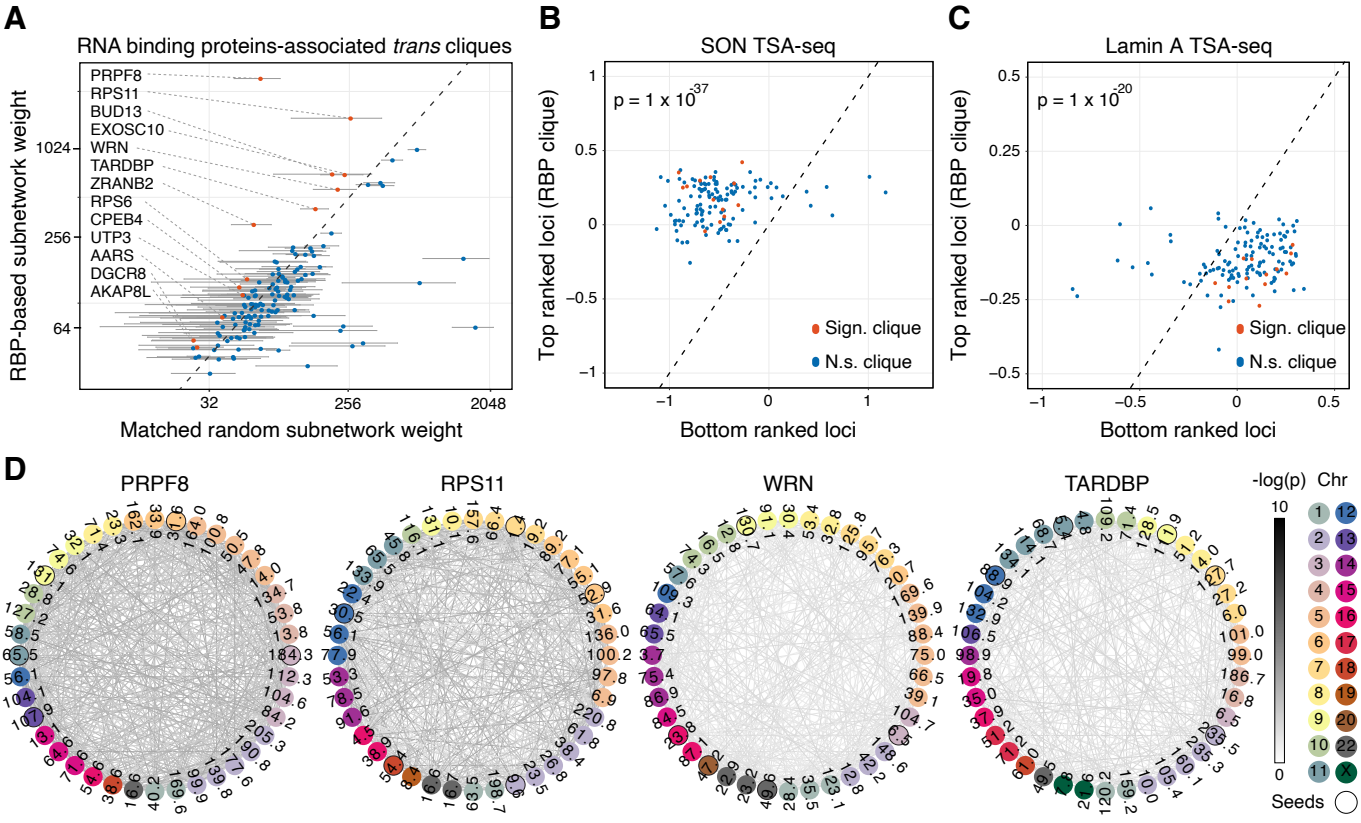

A
Splicing factory
Nuclear speckles
Nucleolus
Polycomb domain
Transcription factories
B
low
Contact frequency
high
Chr 5
Chr 4
Chr 3
Chr 2
Chr 1
Chr 1
Chr 2
Chr 3
Chr 4
Chr 5
C
Chr 2 bin 7
Chr 1 bin 5
Chr 4 bin 4
Chr 3 bin 3
Chr 5 bin 7
Chr 1: bin 5
Chr 2: bin 7
Chr 3: bin 3
Chr 4: bin 4
Chr 5: bin 7
D
Stationary distribution
loci
frequency
Chr 5: bin 7 0.037
Chr 2: bin 7 0.035
Chr 4: bin 4 0.034
Chr 1: bin 5 0.031
Chr 3: bin 3 0.029
trans clique
... ...
Chr 2: bin 2 0.001


A
P. falciparum trophozoite
Telomeres
Active
var
Inactive
var
Repressive
cluster
B
Random genes
var genes
trans
reads
30
20
10
0
cis
contacts
var genes
Random genes
C
True positive rate
False positive rate
0
0.25
0.5
0.75
1
var genes
Random walk
Greedy
Matched random
Random walk
D
-log(p)
10
0
Chr
1
2
3
4
5
6
7
8
9
10
12
13
Seeds


A
Mouse olfactory neuron
Active OR gene
Greek islands
Inactive OR genes
B
True positive rate
False positive rate
0
0.25
0.5
0.75
1
Greek islands
mOSN
HBC
Matched random
mOSN
C
+2.5 Mb
Reference
-2.5 Mb
-2.5 Mb
Reference
+2.5 Mb
Average weight
200
150
100
50
D
HBC
mOSN
-log(p)
0
10
Chr
1
2
3
4
6
7
9
10
11
14
16
19
Seeds


A
Human cardiomyocyte
CAMK2D
CACNA1C
RBM20
TTN
B
Established RBM20 targets
True positive rate
False positive rate
RBM20 seed
hPSCs
Early CMs
Late CMs
Random seed
Late CMs
C
RBM20-bound mRNAs
True positive rate
False positive rate
RBM20 seed
hPSCs
Early CMs
Late CMs
Random seed
Late CMs
D
+2 Mb
Reference
-2 Mb
-2 Mb
Reference
+2 Mb
Average weight
20
10
E
hPSCs
Early CMs
Late CMs
-log(p)
10
0
Chr
Seeds

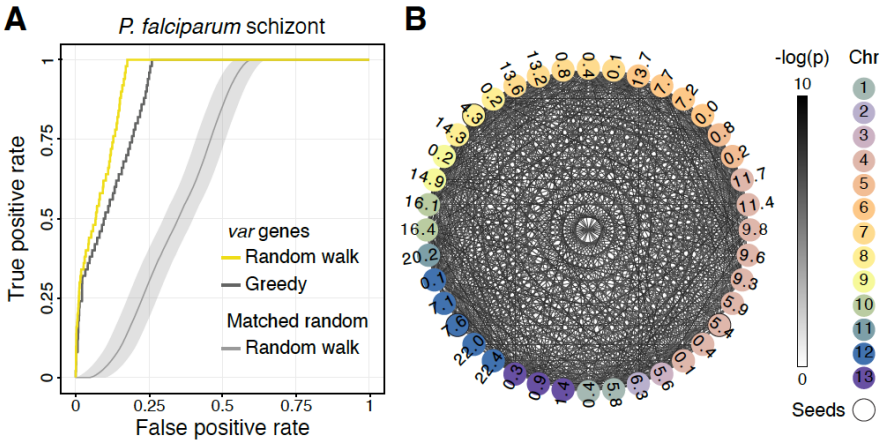

A
trans-acting factor
Core locus 1
Core locus 2
Accessory locus
RNA factory
B
DNA binding proteins-associated trans cliques
Subnetwork weight
2048
1024
512
256
128
p = 6 x 10-3
p = 2 x 10-13
p = 5 x 10-128
Top 40 DBP-bound
Random
Top 5 DBP-bound
Matched random
Random
Matched in random network
Loci
Seeds for trans-C clique
C
Top 25%
Bottom 25%
Sign.
N.s.
DBP-based subnetwork weight
Matched random subnetwork weight
256
512
1024
2048
D
PAX5
MAX
FOS
MLLT1
-log(p)
Chr
Seeds
E
SON TSA-seq
p = 1 x 10-6
Top ranked loci (DBP clique)
Bottom ranked loci
Sign. clique
N.s. clique
F
p-SC35 TSA-seq
p = 8 x 10-3
G
p = 3 x 10-4
DBP clique SON TSA-seq score
Sign. clique
N.s. clique
H
p = 1.4 x 10-5
DBP intrinsic disorder score
Top cliques
Bottom cliques


A
RNA binding proteins-associated *trans* cliques
PRPF8
RPS11
BUD13
EXOSC10
WRN
TARDBP
ZRANB2
RPS6
CPEB4
UTP3
AARS
DGCR8
AKAP8L
RBP-based subnetwork weight
1024
256
64
32
256
2048
Matched random subnetwork weight
B
SON TSA-seq
p = 1 x 10^-37
Top ranked loci (RBP clique)
Bottom ranked loci
Sign. clique
N.s. clique
C
Lamin A TSA-seq
p = 1 x 10^-20
Top ranked loci (RBP clique)
Bottom ranked loci
Sign. clique
N.s. clique
D
PRPF8
RPS11
WRN
TARDBP
-log(p)
10
0
Chr
Seeds

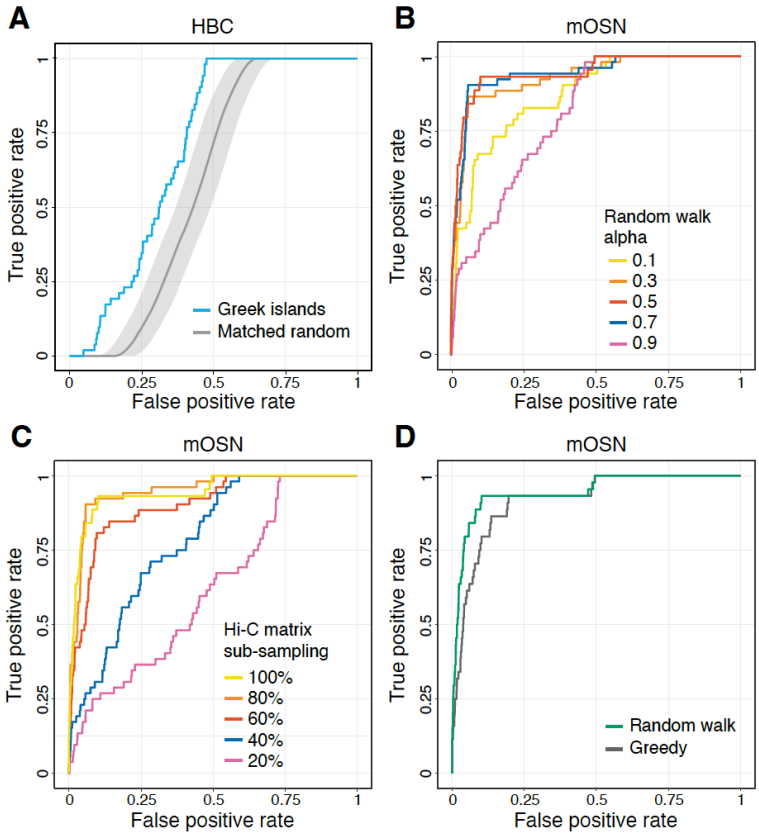
Supplemental Material
Systematic identification of inter-chromosomal interaction networks supports the existence of specialized RNA factories

Borislav Hrisimirov Hristov, William Stafford Noble, Alessandro Bertero

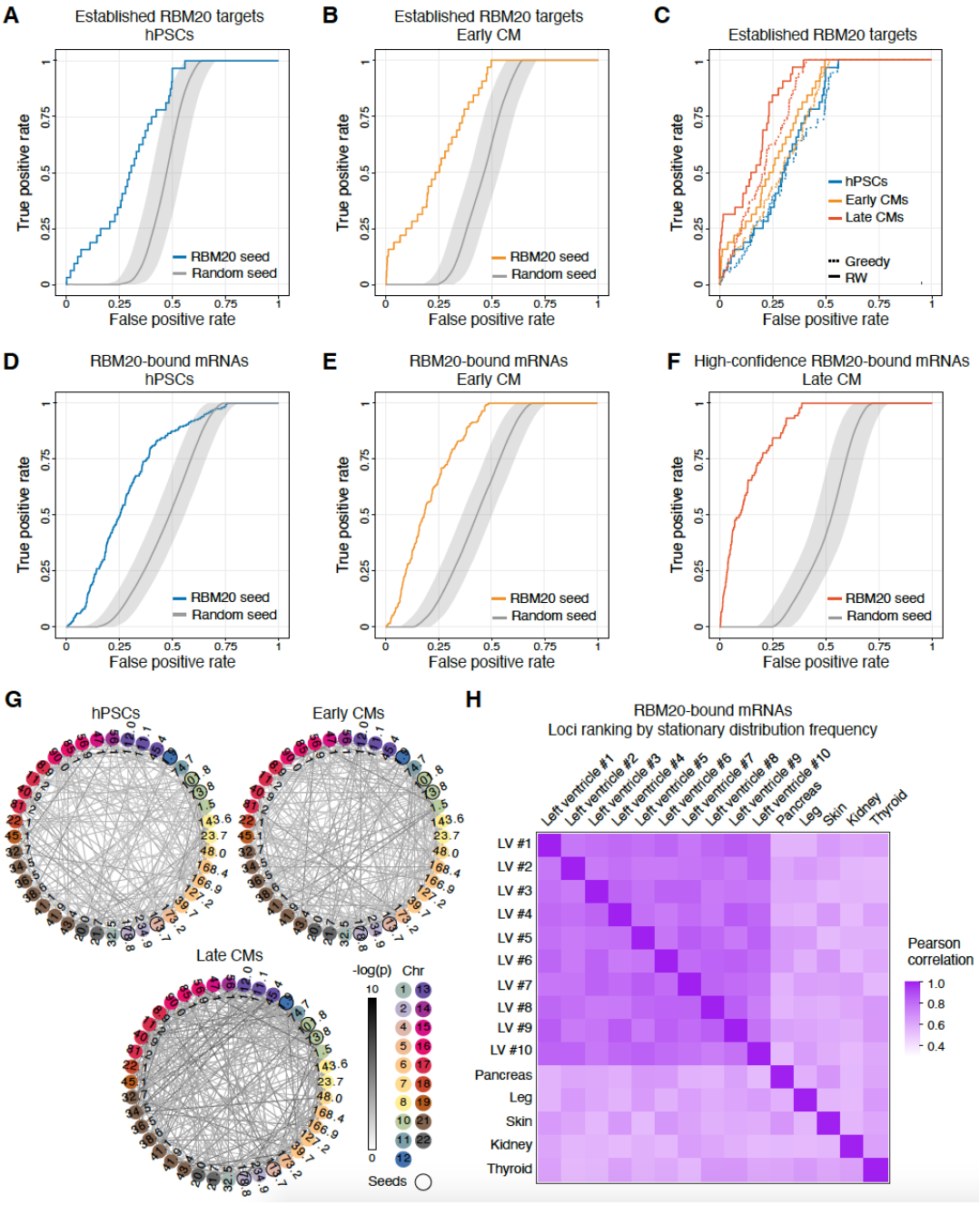
Supplemental Figures



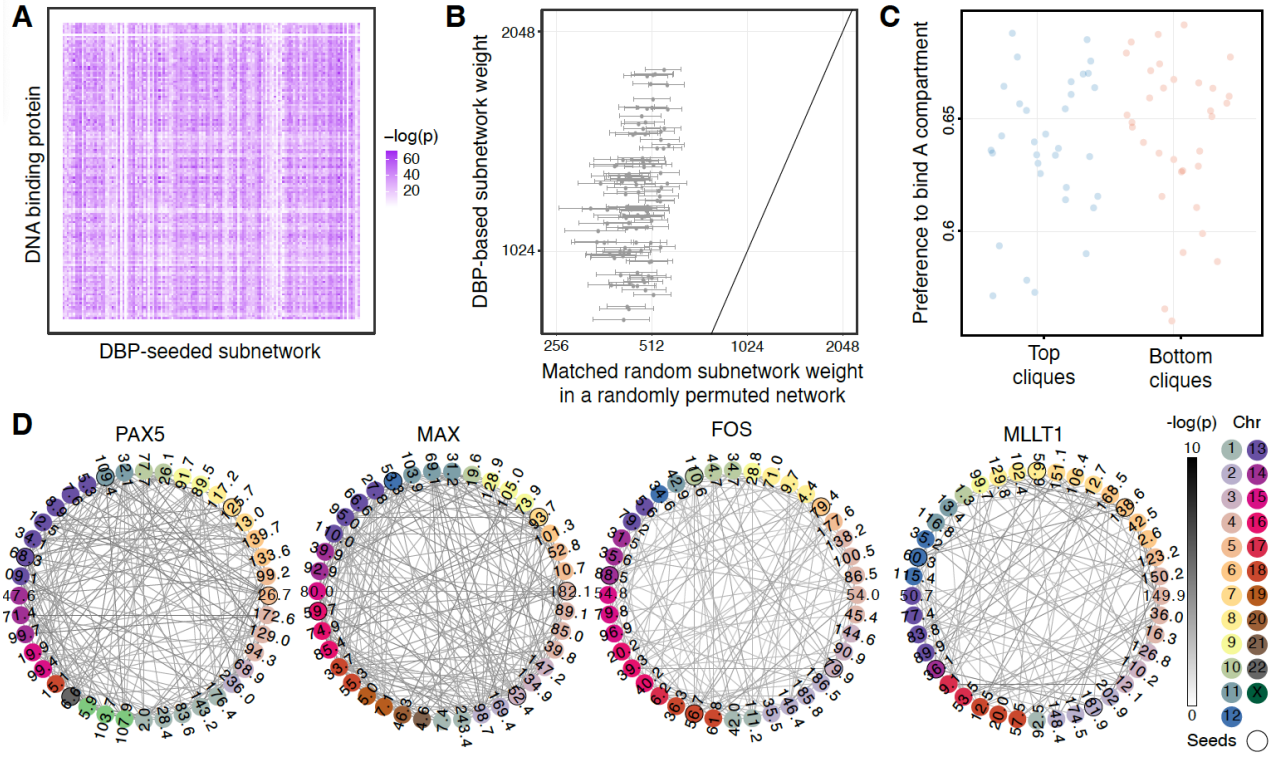
Supplemental Figure S1: Related to Figure 2. **(A)** Performance evaluation of trans-C-mediated identification of *var* gene clustering in schizont stage *P. falciparum* (AUROC 0.93); p-value $= 3 \times 10^{-171}$ *versus* a matched random seed null model (average and 95% confidence interval from 1000 runs). **(B)** Visualization of the *var* genes-associated *trans* clique identified by trans-C in *P. falciparum* schizont, plotted as described for Figure 2D.



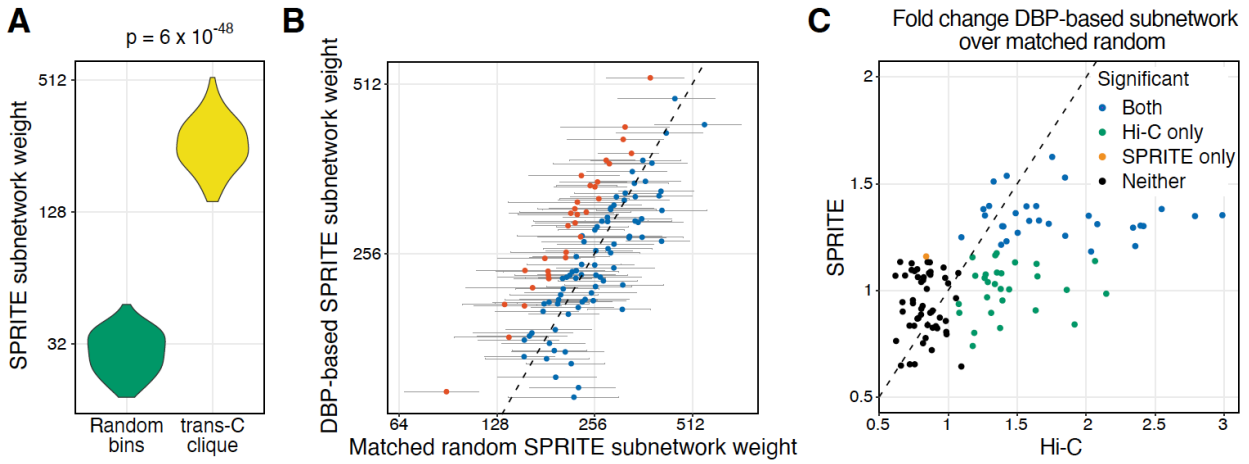
Supplemental Figure S2: Related to Figure 3. Performance evaluation of trans-C in recovering the Greek islands in: **(A)** HBCs, p-value $= 8 \times 10^{-31}$ *versus* a matched random seed null model (average and 95% confidence interval from 1000 runs). **(B)** mOSNs, running trans-C with different values of the parameter alpha. **(C)** mOSNs, comparing the results with those obtained using as input various sub-samples of the Hi-C matrix. **(D)** mOSNs, comparing it with a simpler greedy heuristic.



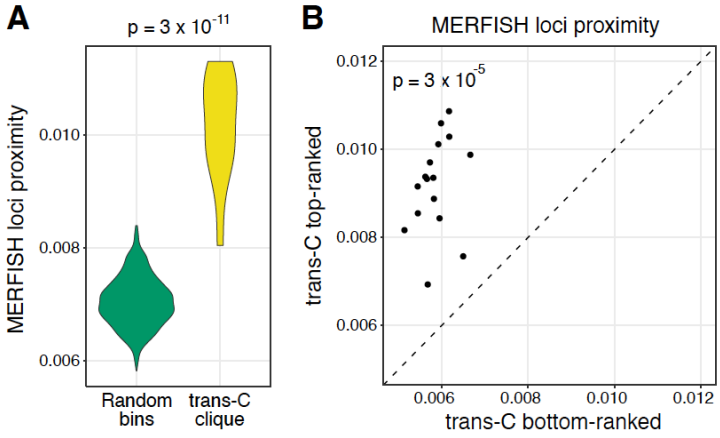
Supplemental Figure S3: Related to Figure 4. **(A-C)** Performance evaluation of trans-C in recovering established RBM20 targets starting from a subset of them in: (A) hPSCs, p-value $= 6 \times 10^{-101}$ *versus* a matched random seed null model (average and 95% confidence interval from 1000 runs); (B) early CMs, p-value $= 2 \times 10^{-105}$ *versus* the same type of control; (C) hPSCs, early CMs, and late CMs (see Fig. 4B) compared to a simpler greedy heuristic. **(D,E)** Same analyses as panels A and B, respectively, but evaluating the recovery of RBM20-bound mRNAs starting from a subset of those most bound (p-value $= 4 \times 10^{-98}$ and 1×10^{-106} for D and E, respectively). **(F)** Same analysis as panels D and E except in late CMs and using a recall list of 45 high confidence RBM20 targets (>2 binding sites & differentially spliced in RBM20 KO) in late CMs; p-value $= 2 \times 10^{-125}$. **(G)** Visualization of the RBM20-associated *trans* clique identified by trans-C in late CMs starting from established RBM20 targets, showcasing the increased significance of loci interactions during hPSC differentiation and CM maturation; plotted as described for Figure 2D. **(H)** Reproducibility evaluation of trans-C in ranking loci starting from eCLIP data. We report the Pearson's correlation of ranked loci stationary distributions for 10 Hi-C matrices of left ventricle and 5 unrelated tissues used as controls.



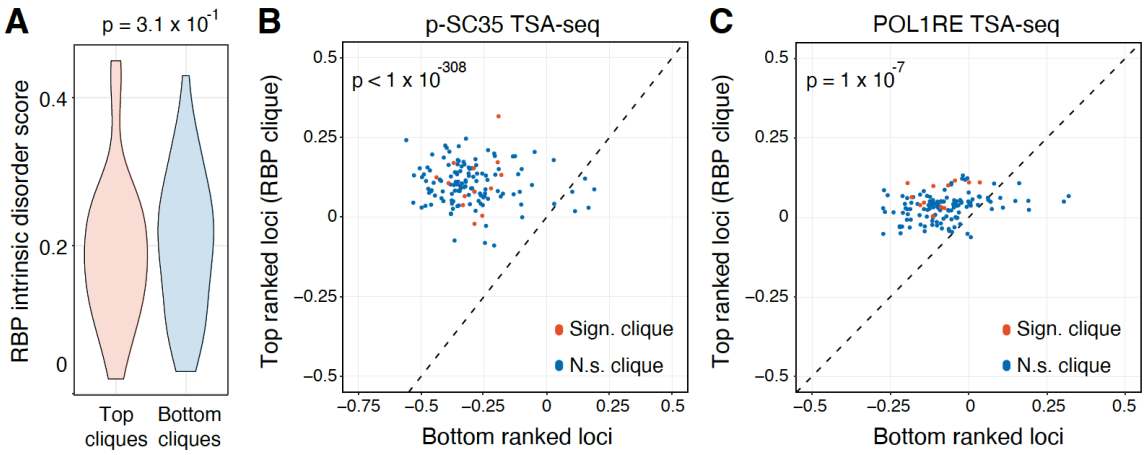
Supplemental Figure S4: Related to Figure 5. **(A)** Enrichment for DBP ChIP-seq peaks in subnetworks built by trans-C from DBP-based seeds (refer to Fig. 5B, orange plot). Each row corresponds to a DBP and each column to a DBP-based subnetwork built by trans-C. For each DBP-based subnetwork we report enrichment for peaks of other DBPs using a hypergeometric test, and report the negative logarithmic p-value in the corresponding cell. **(B)** For each DBP analyzed in Fig. 5B, we compare the weights of subnetworks obtained with trans-C from "Top 5 DBP-bound" seeds (single data point) and "Matched in random network" seeds (average of 1000 subnetworks ± standard deviation). All comparisons are significantly different (p-value < 0.05 after FDR correction). **(C)** Refer to Fig. 5B, orange plot. DBPs that yielded the top and bottom quantiles of subnetwork strength show no difference in their preference to bind loci in the A or B compartments. The y-axis measures the fraction of bins bound by a DBP that are in A compartment. **(D)** For each of the DBPs analyzed in Figure 5D, we show one of its corresponding matched random controls, plotted as described for Figure 2D. These subnetworks are noticeably less dense than the subnetworks trans-C obtained using DBP-bound loci as seed.



Supplemental Figure S5: Related to Figure 5. **(A-B)** For each DBP analyzed in Figure 5, we compare the weights in SPRITE data of subnetworks originally obtained from Hi-C data using trans-C from "Top 5 DBP-bound" seeds to that of random sets of loci (A) or "Matched random" controls (B; average of 1000 subnetworks, error bars correspond to the standard deviation). For B, in red are comparisons with significantly different weights (p-value < 0.05 after FDR correction). **(C)** For each DBP analyzed, we calculate the fold change of the subnetwork weights obtained with trans-C from "Top 5 DBP-bound" seeds and their corresponding "Matched random" seeds (average of 1000 subnetworks). We plot this ratio for the Hi-C data (x-axis) and SPRITE (y-axis) and color the subnetworks based on the dataset(s) they are significant in. The top subnetworks are significant in both datasets (blue dots).



Supplemental Figure S6: Related to Figure 5. **(A)** We compute the proximity of 16 DBP-associated trans-C cliques obtained using the IMR-90 fibroblasts Hi-C submatrix, subsetted to only the loci measured by MERFISH, and compare it to the average proximity of randomly selected sets of loci in the same MERFISH dataset. **(B)** For each DBP, we compare the proximity of the 40 loci measured by MERFISH ranked closest to the top of the trans-C raking on the full IMR-90 fibroblasts Hi-C matrix to that of the 40 loci measured by MERFISH and ranked closest to the bottom of the trans-C ranking.



Supplemental Figure S7: Related to Fig. 6. **(A)** We plot the IDP scores for RBPs resulting in the bottom and top quartile of RBP-based subnetworks from Figure 6A. The difference is not significant by Mann-Whitney U test. RBP-associated cliques identified by trans-C are significantly closer to **(B)** transcription factories (stronger POL1RE TSA-seq signal) and **(C)** nuclear speckles (stronger p-SC35 TSA-seq signal) than matched control sets at the opposite end of the trans-C rankings.

Supplemental Methods

Overview

Trans-C takes as input a Hi-C contact matrix H of interaction counts and an initial set S of loci of interest ("seed loci") and outputs a set of loci U (containing S) that interact strongly together in *trans*. We refer to U and its associated edges as a "dense subgraph." We model the Hi-C interaction matrix H as a weighted graph $G = (V, E, W)$ with nodes V corresponding to the genomic loci (bins), edges E between pairs of loci, and weights W on the edges reflecting the strength of the interactions represented by the edges. Thus, the weight w_{ij} on edge e_{ij} between loci i and j corresponds to the contact matrix entry h_{ij}. Our goal is to find a subset of loci that exhibit strong inter-chromosomal contacts. We propose two methods to solve this problem, one that uses a random walk with restart and a second that formulates the problem as a dense subgraph optimization and solves it using a fast greedy heuristic.

H	Hi-C matrix
$h_{i,j}$	Contact count between loci i and j
G	Graph corresponding to H
V	Set of all genomic loci
E	Set of edges in G
W	Set of weights on the edges of G
$e_{i,j}$	edge connecting loci i and j
$w_{i,j}$	weight associated with edge $e_{i,j}$
U	Set of all genomic loci
S	Set of seed loci
ℓ	User-specified size of desired subgraph
p	Inner radius of the donut filter
q	Oter radius of the donut filter
C	Vector of peak counts
ν_j	Boolean indicating wether locus i is in set U
η_{ij}	Boolean indicating wether locus i interacts with j
Δ_k	Change in the clique density score by adding loci k
π_j	The stationary distribution the random walk converges to

Random walk with restart

Our first solution carries out random walks with restarts over the graph G and then uses the results of the random walks to select the nodes in U. The walk is initiated from the set of seed loci S and its goal is to highlight the nodes that are strongly connected to those in S. At each step, with probability α the walk restarts from a randomly selected seed locus, and with probability $1-\alpha$ the walk moves to a neighboring locus picked probabilistically based upon W. Specifically, if $\mathcal{N}(i)$ are the loci i interacts with, then the walk goes from locus i to locus $j \in \mathcal{N}(i)$ with probability proportional to $w_{i,j} / \sum_{k \in \mathcal{N}(i)} w_{i,k}$. That is, if at time t the walk is at locus i, then the probability that it transitions to locus j at time $t+1$ is

$$p_{ij} = (1-\alpha)\frac{\eta_{ij} w_{i,j}}{\sum_{k \in N(i)} w_{i,k}} + \alpha \frac{\nu_j}{|U|}$$

where $\eta_{ij} = 1$ if $j \in \mathcal{N}(i)$ and 0 otherwise, and $\nu_j = 1$ if $j \in U$ and 0 otherwise. Hence, the guided random walk is fully described by a stochastic transition matrix P with entries p_{ij}. This stochastic matrix is non-negative and by the Perron-Frobenius theorem it has a right eigenvector π corresponding to eigenvalue 1. Therefore, $\pi P^t = \pi$, and the random walk converges. That is, the probability of the walk being at node i at time t is constant as $t \to \infty$. This probability is specified by the ith element of π and π, known as stationary distribution of the walk, can be efficiently computed. Further, the probability π_i reflects how well the node i is connected to the seed nodes as more strongly connected nodes are more frequently visited. We obtain a score for each locus j by finding the jth element of π. The loci that have the highest scores are most frequently visited and, therefore, are more likely relevant as they are strongly connected to the seed loci. We use these

scores to rank all loci and include the top ℓ loci in U, where ℓ is a user-specified parameter. In this work, we use $\ell = 40$ unless otherwise stated.

Greedy heuristic

Our second solution builds a ranked list of loci in U in a greedy fashion. In this approach, we formalize our goal as finding $U \in V$ such that the subgraph induced by $|U|$ is dense; i.e.,

$$\text{score}(U) = \sum_{i,j \in U} w_{i,j}$$

is large. We note that when we constrain the size of U, the problem is computationally hard as it can be reduced to the maximum clique problem, which is NP-complete. As in the previous approach, we assume that the user specifies an initial set S of seed loci, as well as the desired subgraph size ℓ. Thus, formally, the optimization problem we aim to solve is

$$\max_{|U|=\ell, S \subset U \in V} \text{score}(U) \tag{1}$$

We propose to maximize Equation 1 using a greedy algorithm. The procedure begins by adding the seed loci S to the initially empty set U. Then, in each step the heuristic expands U by examining all loci not currently in U and selecting to add to U the one that yields the largest increase in Equation 1. Mathematically, the greedy step finds

$$\max_{k \in |V| \setminus |U|} \Delta_k = \text{score}(|U \cup k|) - \text{score}(U) = \sum_{x \in U} w_{x,k}$$

Ties are broken randomly. The greedy selection proceeds as long as $\Delta_k > 0$ and $|U| < 40$. In practice, in the calculation of Δ_k we exclude the single strongest interaction between k and U. We do so because we do not want a single very large $w_{k,x}$ to dominate Δ_k; instead, our aim is that all loci in U interact strongly.

Data pre-processing

Prior to searching a given Hi-C matrix for dense subgraphs, we perform three pre-processing steps.

First, we normalize the Hi-C matrix using the iteractive correction and eigenvalue decomposition (ICE) procedure (Imakaev et al., 2012). This procedure iteratively normalizes rows and columns of the matrix, producing as output a matrix in which the marginal values are all equal to a specified constant. We carry out the procedure on the entire Hi-C matrix, including *cis* and *trans* contacts, using the Python package "iced" (Servant et al., 2015).

Second, we adjust the matrix entries to account for the fact that chromosomes tend to occupy specific regions of the nucleus, called *chromosome territories*, and as a result some pair of chromosomes interact more frequently. For each pair of chromosomes L and M ($L \neq M$) we find the total number of interactions between any locus i in L and j in M: $T_{L,M} = \sum_{\forall i \in L; \forall j \in M} h_{ij}$. If T is the total number of trans-interactions in H, then we rescale contact count h_{ij} as $h'_{ij} = h_{i,j} * T / T_{L,M}$. During this step, we set all *cis* contacts (i and j are on the same chromosome) to zero.

Third, we process H using a "donut filter" to emphasize points that are local minima in the 2D contact map (Rao et al., 2014). Given a *trans* contact (i, j), we define its donut background as the set of all loci that are at least p loci away from (i, j) but no further than q loci away and which do not lie along the i or j axes. Intuitively, p is the radius of the hole of the donut centered at (i, j), q is the outer radius of the donut, and the donut has been sliced in four pieces along the i and j axes. Mathematically,

$$\text{DN}(i,j) = \frac{1}{\text{DN}_{p,q}} \left(\sum_{a=i-q}^{i+q} \sum_{b=i-q}^{j+q} h_{a,b} - \sum_{a=i-p}^{i+p} \sum_{b=i-p}^{j+p} h_{a,b} - \sum_{a=i-q}^{i-p-1} h_{a,j} - \sum_{a=i+p+1}^{i+q} h_{a,j} - \sum_{b=j-q}^{j-p-1} h_{b,i} - \sum_{b=j+p+1}^{j+q} h_{b,i} \right)$$

where we divide the sum by the total number of loci $\text{DN}_{p,q}$ in the donut to obtain the average strength of interactions in the donut. The enrichment for the contact (i, j) with respect to its local background can then be calculated as $h_{i,j}/\text{DN}(i, j)$. In practice, we select $p = 2$ and $q = 20$, and we set $w_{i,j} = h'_{i,j}/\text{DN}(i, j)$.

This weight $w_{i,j}$ is finally placed on the edge between nodes i and j in the graph G to reflect the normalized strength of interaction between loci i and j.

We calculate a TSA-seq score per bin by aggregating the processed $\log_2$ fold change values given at a single nucleotide resolution from the 4DN Portal. We define the TSA-seq score for given clique as the average TSA-seq score of each loci of the clique.

Supplemental Tables

Supplemental tables are available online:

Supplemental Table S1: Greek islands-based clique

Supplemental Table S2: Established RBM20 targets-based clique

Supplemental Table S3: RBM20-bound mRNAs-based clique

Supplemental Table S4: DBPs-based cliques

Supplemental Table S5: RBPs-based cliques

Supplemental Table S6: Intra- *versus* inter-chromosomal space in different species

Supplemental Table S7: Statistics of the Hi-C datasets analyzed in the study