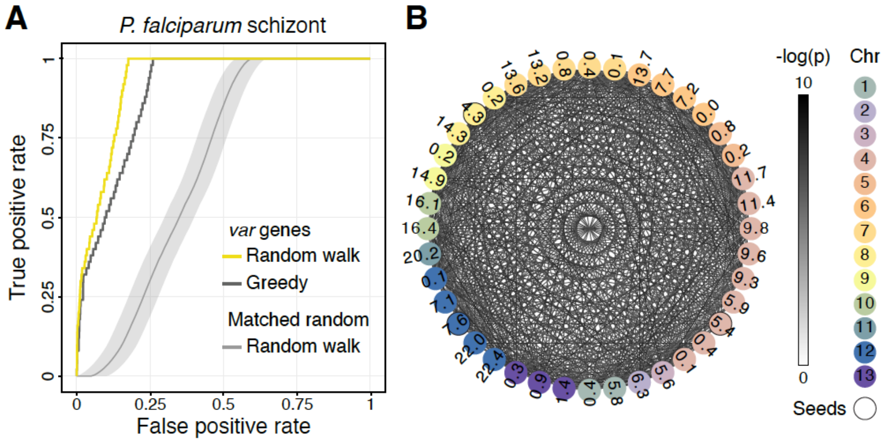
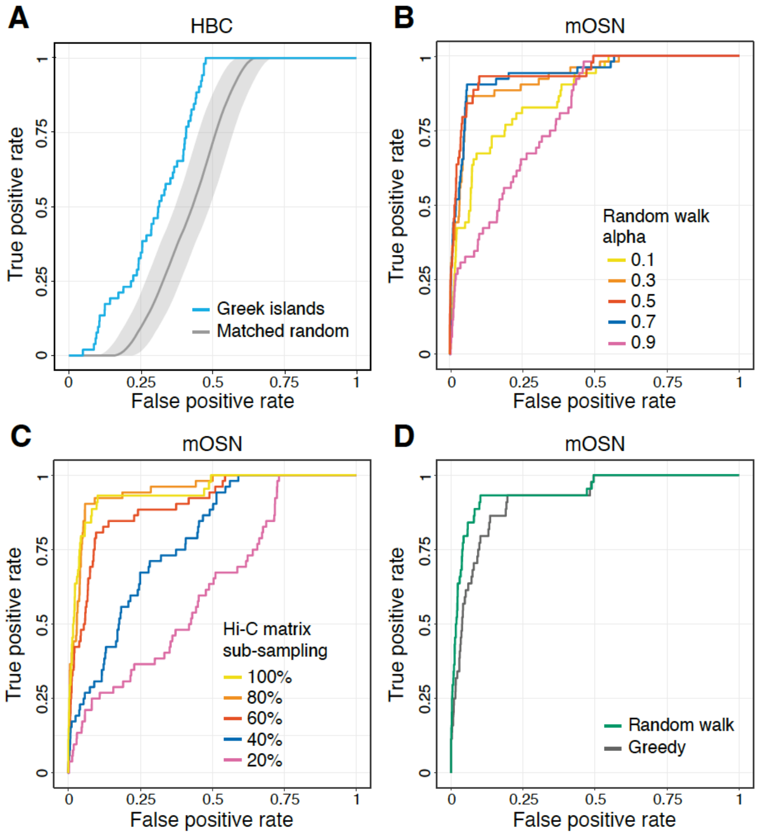
# Supplemental Material
# Systematic identification of inter-chromosomal interaction networks supports the existence of specialized RNA factories

Borislav Hrisimirov Hristov, William Stafford Noble, Alessandro Bertero

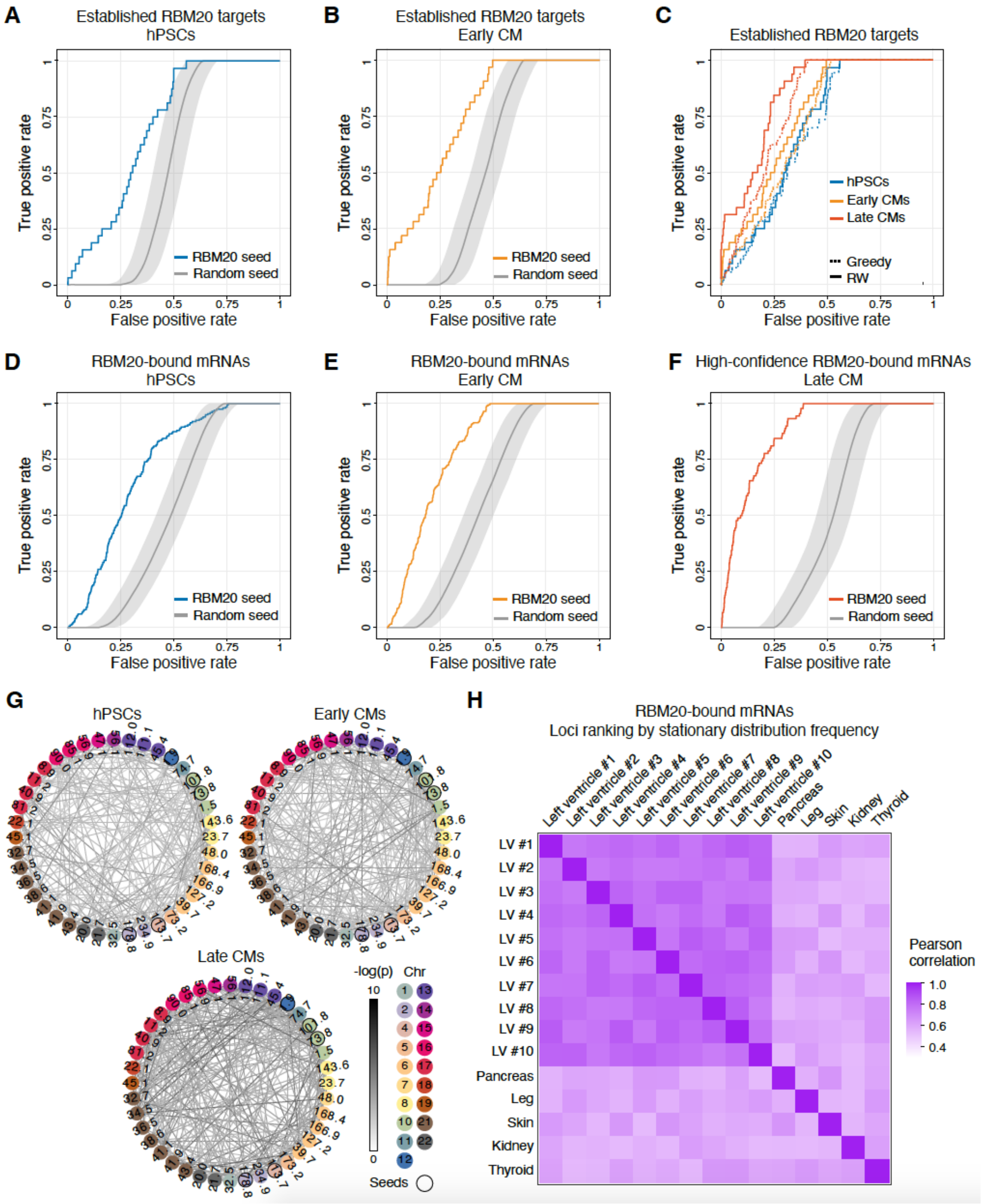
# Supplemental Figures



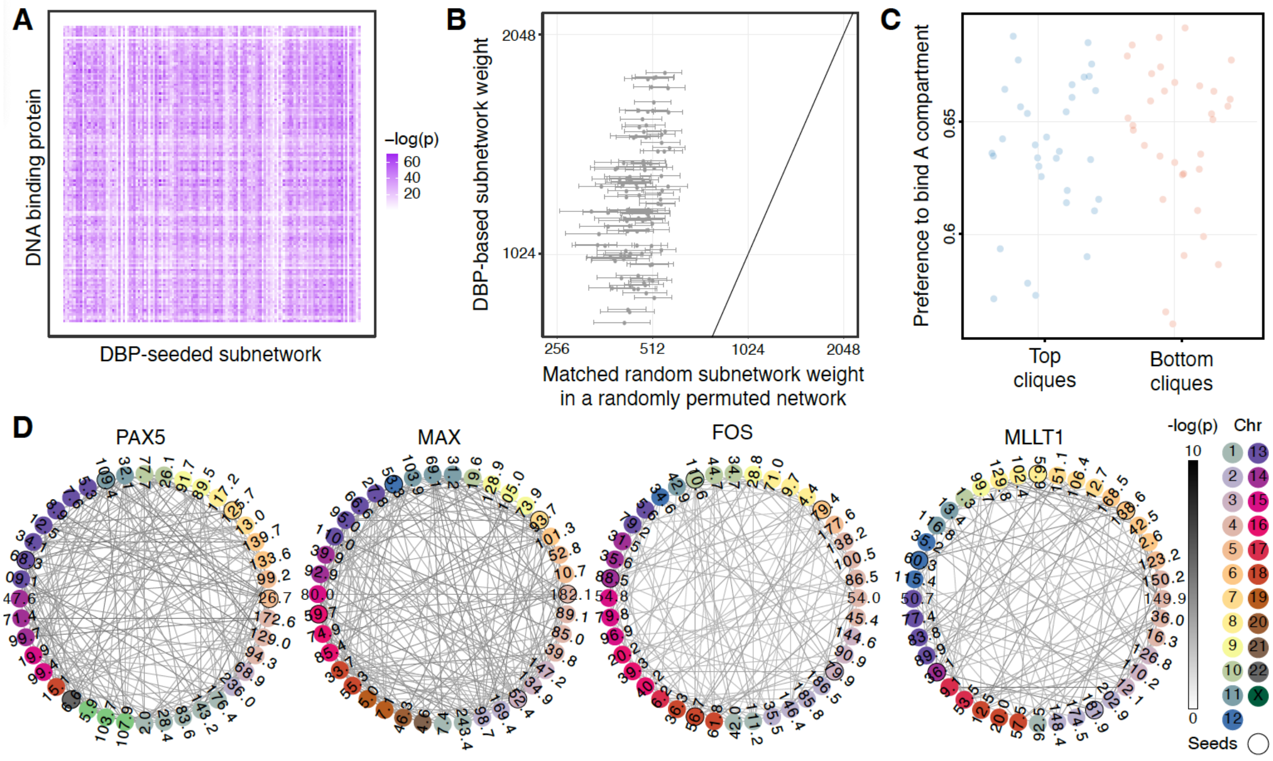
Supplemental Figure S1: Related to Figure 2. **(A)** Performance evaluation of trans-C-mediated identification of *var* gene clustering in schizont stage *P. falciparum* (AUROC 0.93); p-value = $3 \times 10^{-171}$ *versus* a matched random seed null model (average and 95% confidence interval from 1000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



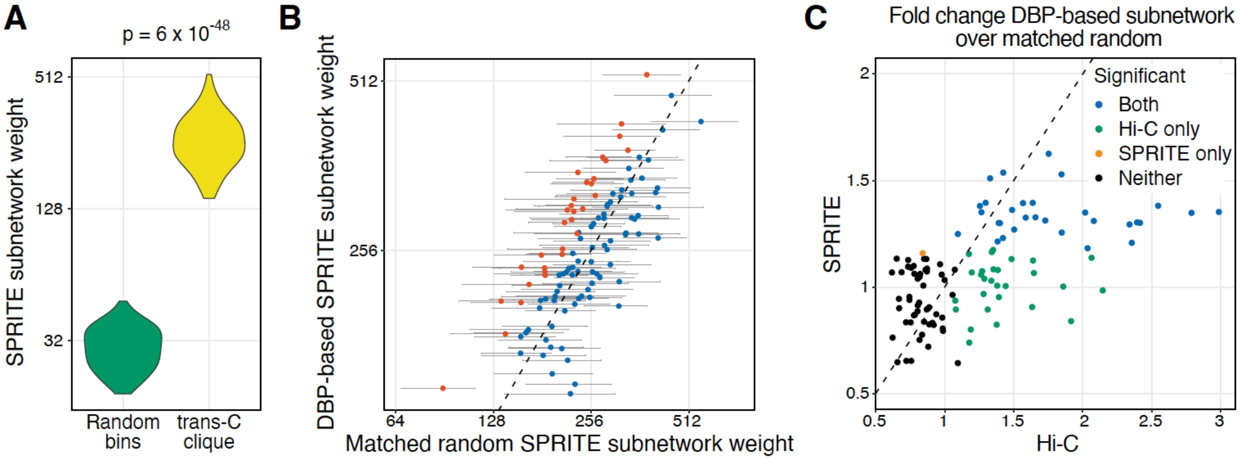
Supplemental Figure S2: Related to Figure 3. Performance evaluation of trans-C in recovering the Greek islands in: **(A)** HBCs, p-value = $8 \times 10^{-31}$ *versus* a matched random seed null model (average and 95% confidence interval from 1000 runs). **(B)** mOSNs, running trans-C with different 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.



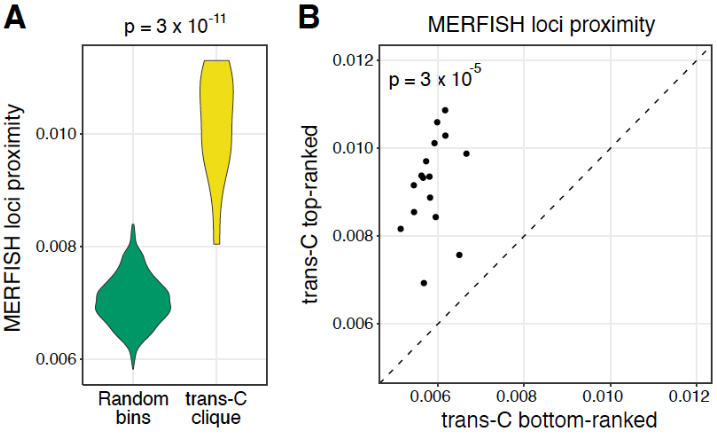
Supplemental Figure S3: Related to Figure 4. **(A-C)** Performance evaluation of trans-C in recovering established RBM20 targets starting from a subset of them in: (A) hPSCs, p-value $= 6 \times 10^{-101}$ *versus* a matched random seed null model (average and 95% confidence interval from 1000 runs); (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 $1 \times 10^{-106}$ 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.



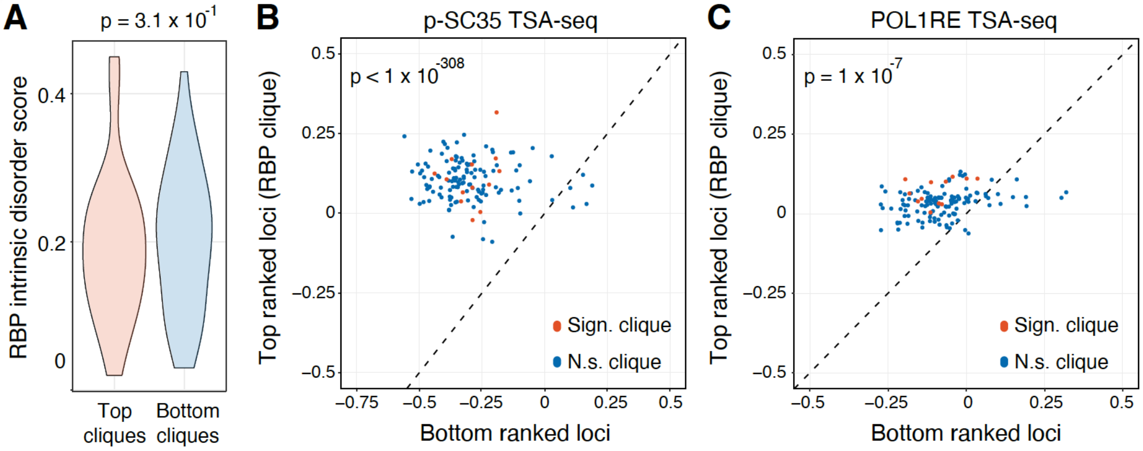
Supplemental Figure S4: Related to Figure 5. **(A)** Enrichment for DBP ChIP-seq peaks in subnetworks built by trans-C from DBP-based seeds (refer to Fig. 5B, orange plot). Each row corresponds to a DBP and each column to a DBP-based subnetwork built by trans-C. For each DBP-based subnetwork we report 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 |
| $\ell$ | 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 |
| $\nu_j$ | Boolean indicating wether locus $i$ is in set $U$ |
| $\eta_{ij}$ | Boolean indicating wether locus $i$ interacts with $j$ |
| $\Delta_k$ | Change in the clique density score by adding loci $k$ |
| $\pi_j$ | 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 $\alpha$ 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 $\pi$ 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 $i$th element of $\pi$ and $\pi$, known as stationary distribution of the walk, can be efficiently computed. Further, the probability $\pi_i$ 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 $j$th element of $\pi$. 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 $\ell$ loci in $U$, where $\ell$ 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 $\ell$. 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 $\Delta_k$ 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 $\Delta_k$; 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