5818–5828 *Nucleic Acids Research, 2017, Vol. 45, No. 10* *Published online 23 March 2017*
doi: 10.1093/nar/gkx198

Genome urbanization: clusters of topologically co-regulated genes delineate functional compartments in the genome of *Saccharomyces cerevisiae*

Maria Tsochatzidou[1], Maria Malliarou[1], Nikolas Papanikolaou[1], Joaquim Roca[2] and Christoforos Nikolaou[1,*]

[1]Computational Genomics Group, Department of Biology, University of Crete, Herakleion 70013, Greece and [2]Molecular Biology Institute of Barcelona (IBMB), Spanish National Research Council (CSIC), Barcelona 08028, Spain

Received January 16, 2017; Revised March 03, 2017; Editorial Decision March 14, 2017; Accepted March 15, 2017

ABSTRACT

The eukaryotic genome evolves under the dual constraint of maintaining coordinated gene transcription and performing effective DNA replication and cell division, the coupling of which brings about inevitable DNA topological tension. DNA supercoiling is resolved and, in some cases, even harnessed by the genome through the function of DNA topoisomerases, as has been shown in the concurrent transcriptional activation and suppression of genes upon transient deactivation of topoisomerase II (topoII). By analyzing a genome-wide transcription run-on experiment upon thermal inactivation of topoII in *Saccharomyces cerevisiae* we were able to define 116 gene clusters of consistent response (either positive or negative) to topological stress. A comprehensive analysis of these topologically co-regulated gene clusters reveals pronounced preferences regarding their functional, regulatory and structural attributes. Genes that negatively respond to topological stress, are positioned in gene-dense pericentromeric regions, are more conserved and associated to essential functions, while upregulated gene clusters are preferentially located in the gene-sparse nuclear periphery, associated with secondary functions and under complex regulatory control. We propose that genome architecture evolves with a core of essential genes occupying a compact genomic 'old town', whereas more recently acquired, condition-specific genes tend to be located in a more spacious 'suburban' genomic periphery.

INTRODUCTION

The distribution of genes in the genome of eukaryotes is highly non-random. Early genome-wide transcriptome analyses showed the expression of genes to correlate with their linear order along the genome (1). Although it was later shown that this was due to the clustering of constitutive genes (2), such spatial associations have since been used to provide the theoretical framework for links between gene expression and chromatin structure (3) and the inference of protein–protein interaction patterns (4). Non-random gene distribution is also evident in the ontological enrichments of gene neighborhoods, with functionally related genes being found in linear proximity more often than expected by chance (5,6).

The selective pressures underlying gene localization are thus of unequal intensity and diverse nature and a number of seemingly irrelevant characteristics may shape the overall genome architecture through evolution (7). Among those, DNA supercoiling plays a prominent role. The structure of the eukaryotic nucleus is affected by a number of processes such as DNA replication, RNA transcription and the constant ebb and flow of gene activation and repression. These are imposing topological constraints in the form of supercoiling, both types of which (positive and negative) may be found in localized areas of the eukaryotic genome (8). It was recently shown that such structurally-defined areas may form part of extended 'supercoiling domains', where chromatin conformation correlates with the density of topoisomerases I and II (9). The connection between topological attributes and gene expression appears to be so strong, that in *Drosophila melanogaster* regions of negative supercoiling, created through the inhibition of topoisomerase I, show increased nucleosome turnover and recruitment of RNA-PolII molecules positively correlating with transcription levels (10). Accumulated positive supercoiling, on the

*To whom correspondence should be addressed. Tel: +30 2810 394361; Fax: +30 2810 394404; Email: cnikolaou@biology.uoc.gr

other hand, precludes the formation of transcription initiation complexes (11,12), a fact indicative of the association between topological constraints and gene expression.

In the budding yeast (*S. cerevisiae*), the organization of genes in linear space has also been attributed to common regulatory mechanisms (13). Yeast's distinguishing genomic feature is the overall gene density, with genes covering ∼70% of the total genome (14). Despite its reduced size of only 12 Mbp, the transcription dynamics of its genome is highly complex, with genes being expressed in tandem and in operon-like transcripts, with varying sizes of gene upstream and downstream regions (15). Transcription directionality in such a highly streamlined genome also plays a crucial role in the regulatory process, with a number of bidirectional promoters (16) exerting control over coupled gene pairs. The interplay between DNA structure and gene regulation is manifest in a number of cases where gene expression is modulated through three-dimensional (3D) loops formed at gene boundaries (17). Thus, even in a small eukaryotic genome, there is a strong association between gene organization (in both linear and 3D space) and gene expression.

The response to topological stress has been shown to be shaped by specific structural properties of yeast promoters (18). In this work, we sought to investigate how the response to the accumulation of topological stress may extend beyond single gene promoters to affect broader genomic regions. Starting from a genome-wide transcription run-on (GRO) experiment, conducted shortly (15 min) after the thermal inactivation of topoII, we explored the formation of clusters of genes that are differentially affected and then went on to assess a number of related functional and structural preferences. We were able to detect intricate associations between DNA topology and the distribution of genes in linear order and to show how the two may be linked to other organizational characteristics such as gene spacing, transcriptional directionality and the 3D organization of the yeast genome. Our results are suggestive of a subtle dynamics of evolution of genome architecture, which we describe as 'Genome Urbanization' and according to which the relative position of genes in the nucleus reflects a broader functional, structural and regulatory compartmentalization.

MATERIALS AND METHODS

Genome-wide transcription run-on data

Data were obtained from a genome-wide transcription run-on (GRO) experiment conducted in triplicates on a yeast strain lacking topoisomerase I and carrying a thermosensitive topoisomerase II (JCW28–top1Δ, top2ts). Cultures were incubated with a calibrated volume of hot medium for 10 min at 37°C and then for 5 min at 30°C in the presence of ^{33}P-UTP. GRO was conducted as described in (19) and data were analyzed as previously described in (18).

Gene clustering

Starting from an initial dataset of differential GRO values for 5414 yeast protein coding genes (Supplementary Figure S1 and Supplementary File 1), gene clusters were defined as the uninterrupted regions spanning the genomic space from the first to the last segment in an all-positive (upregulated) or all-negative (downregulated) gene series (Figure 1A and B). Clusters of ≥7 genes were selected on the basis of a permutation analysis as suggested in (7). This was performed by conducting 10 000 random permutations of gene order while keeping the same GRO values. We used functions from the BedTools Suite (20) to control for unaltered gene sizes and chromosomal distributions. Size distributions (in number of genes) of the derived clusters were calculated alongside the mean values and standard deviation of those obtained for the 10 000 random gene sets. We then compared the observed values with the expected under randomness asking that the observed value be at least greater than the mean of the 10 000 permutations by two standard deviations. Clusters with ≥7 genes occurred in less than 0.1% of the simulations (bootstrap value $P = 0.0008$) and were divided into upregulated and downregulated, depending on the mean GRO value of all genes in each cluster (Supplementary File 2).

Gene cluster co-expression index

The co-regulation of genes in the clusters was assessed through the adjusted correlation scores (ACS), as obtained for the complete set of yeast genes from the SPELL database (21). ACS values represent weighted correlation values for a large number of genome-wide expression profiles. As a measure of co-expression in a gene cluster, we calculated the mean ACS of all genes within the confines of the cluster.

Positional enrichments of gene clusters in linear dimension

Genomic coordinates for yeast centromeres were obtained from the Saccharomyces Genome Database (SGD) (http://www.yeastgenome.org/locus/centromere). Cluster-centromere distances were calculated as the sequence length between the most proximal cluster boundary to the central point of the centromeric coordinates. Distances were then scaled with the size of the chromosomal arm extending from the central point of the centromere to the chromosome's boundary, so as to be represented in a range of 0 (i.e. overlapping the centromere) to 1 (i.e. lying at the edge of the corresponding chromosomal arm). Distances from autonomously replicating sequences (ARS) were calculated in the same way based on a compiled list of 829 yeast ARS published in OriDB (http://cerevisiae.oridb.org/) (22).

Three-dimensional positional enrichments of gene clusters

We obtained the normalized frequency measurements of a yeast 3C experiment (23). In order to define regions of increased intra-chromosomal interactions, we used the insulation profile approach described in (24), where an aggregate score of contact frequencies is calculated along the diagonal of an interaction map. By setting an upper limit of insulation score equal to the bottom 5%-percentile we were able to define 86 insulation domains at 10 kb resolution. We then compared these domains for overlaps with the defined gene clusters and long-terminal repeat (LTR) regions, obtained from SGD (http://downloads.yeastgenome.

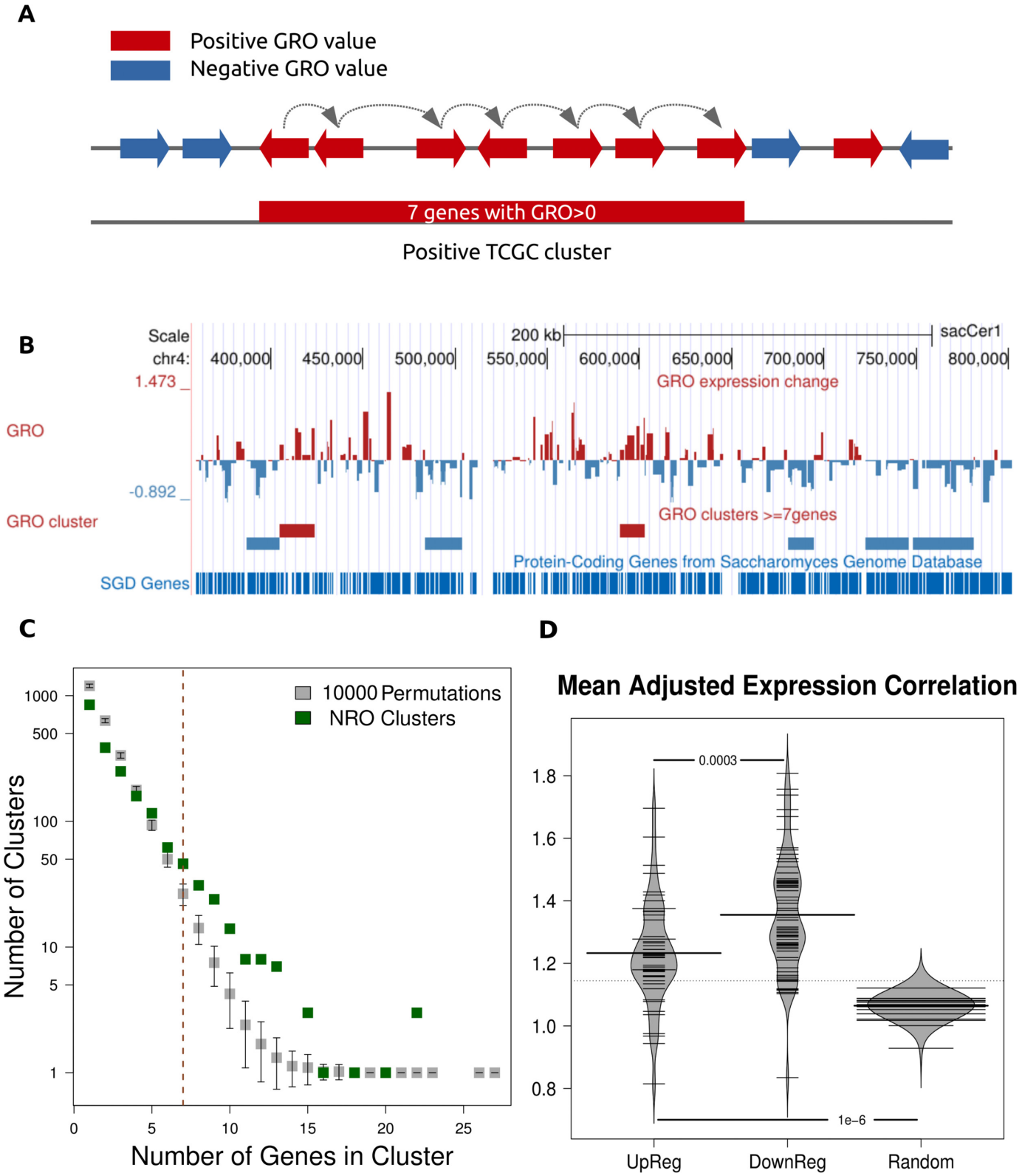

Figure 1. DNA topological stress-responsive genes are clustered non-randomly. (**A**) Schematic representation of cluster definition. Contiguous genes with similar (positive or negative) GRO values were joined in gene clusters, which were defined as the genomic region spanning the chromosomal space from the farthest upstream to the farthest downstream gene boundary. (**B**) Location of genes and clusters with GRO values in part of chromosome IV. (**C**) Distribution of number of genes in clusters. Real clusters show a skewed distribution toward larger sizes as compared to the mean of 10 000 random permutation of GRO values. Differences are significant for gene numbers ≥ 6. (**D**) Mean adjusted correlation scores (ACS) for up- and downregulated gene clusters are significantly higher than the genomic average (randomly selected clusters of equal size and number), indicating increased co-regulation within the confines of the defined clusters.

org/curation/chromosomal_feature/SGD_features.tab), as LTRs have been known to be associated with barrier regions between chromosomal domains of interaction.

At a second level we used the classification of yeast chromosomal regions in network communities described in (25). We calculated the enrichment of our up- and downregulated gene clusters separately for the 13 distinct level-1 communities (Supplementary Table S7 from (25)). Overlap enrichments were calculated, in all cases, on the basis of an observed over expected ratio of overlaps between the two sets of genomic coordinates and was statistically assessed on the basis of 1000 random permutations of cluster coordinates. Overlaps with a bootstrap P-value ≤ 0.01 were deemed significant (Supplementary File 3).

Functional and regulatory enrichment

We employed a modified gene set enrichment analysis at gene cluster level to analyze concerted over-representations of gene ontology (GO) terms (www.geneontology.org). Enrichment was calculated based on a hypergeometric test for each gene cluster and controlled for multiple comparisons at a 5% False Discovery Rate (FDR) (26). GO terms with significant enrichment (adjusted P-value ≤ 0.05) in at least one of the two types of gene clusters (up- or downregulated) were recorded.

Conserved transcription factor binding sites (TFBS) were obtained from the UCSC Genome Browser's Transcriptional Regulatory Code track. These corresponded to a compendium of 102 transcriptional regulators based on a combination of experimental results, cross-species conservation data for four species of yeast and motifs from the literature, initially compiled by (27) and updated by (28). Enrichment in TF binding was calculated as in the case of chromosomal communities described above. Enrichments were assessed as ratios of observed over expected overlaps and P-values were obtained as bootstrap values from 1000 random permutations of cluster coordinates.

Gene and intergenic space size and direction of transcription

We used genomic coordinates downloaded from UCSC (SGD/saCcer2). Intergenic distances were calculated as the full length of regions spanning the genomic space between two consecutive genes, using transcription initiation and termination as boundaries, regardless of gene transcription direction. We assigned to each gene a mean intergenic space length to be the arithmetic mean of the lengths of gene upstream and downstream intergenic regions. For genes at chromosomal boundaries, one of the two intergenic regions was set to be equal to the distance from the gene boundary to the corresponding chromosomal start/end.

For the analysis of gene directionality, each chromosome was scanned in overlapping 11-gene windows and for each step we recorded: the full list of 11 GRO values, mean GRO value of the central 7 genes and gene lengths and mean intergenic space lengths for all genes. The top/bottom 200 non-overlapping clusters in terms of mean GRO value were analyzed at the level of gene and intergenic spacer lengths (Figure 3A). We used the same list to obtain patterns of gene directionality as arrays of seven genes (Figure 3B). Sizes of 11 and 7 genes respectively were restricted by our TCGC analysis. As our clusters contained on average 9 genes (s.d = 1.51, with all clusters bearing ≥ 7 genes), sizes above 11 genes would significantly exceed the mean size of the observed clusters. GRO values of the central gene were analyzed for three characteristic patterns corresponding to (i) co-directional genes (central five genes transcribed in the same direction) (ii) the central gene being a member of a divergent or (iii) a convergent gene pair. These characteristic patterns were chosen, among various combinations of gene directionality, as the corresponding numbers of cases were large enough to allow for statistical comparisons.

Sequence conservation and TFBS density

Sequence conservation was calculated as aggregate phastCons scores (29) obtained from UCSC and based on a multiple alignment of seven *Saccharomyces* species. Mean conservation was taken as the mean phastCons score for a given region. For each cluster we removed intergenic space and calculated the mean aggregate phastCons score for all genes in the cluster. TFBS density was calculated as the percentage of the length of each TCGC overlapping with conserved TFBS as compiled in (27).

Gene cluster directionality conservation index

We obtained orthologous gene coordinates for *Saccharomyces paradoxus* and *Saccharomyces mikatae* from the Yeast Gene Order Browser (YGOB) (30). For each genomic region of *S. cerevisiae* we calculated the ratio of genes retaining their position and direction of transcription in the other two species. A value of 1/N, N being the number of genes in the region, was added to the score if both the gene's position and direction was maintained in the other two species. This led to measure of directionality conservation on a scale of 0 (no retention of direction) to 1 (absolute retention of direction). The contour map of Figure 3C was formed by splitting the two-dimensional space in a 10×10 grid and assigning each bin with the proportion of clusters falling in the corresponding sequence/direction conservation value range (bins of 0.1 for each). The final value assigned to each of the 10×10 bin was the log2(ratio) of up/downregulated cluster frequency. Values >0 corresponded to an enrichment of up- and values <0 to an enrichment of downregulated clusters.

RESULTS

Non-random clustering of topologically co-regulated genes

We first sought to define domains with concordant response to DNA topological stress in the form of gene clusters of contiguous GRO values (Figure 1A and B and 'Materials and Methods' section). In total there were 116 clusters with more than 7 genes and 180 clusters containing 6 or more genes, which were deemed highly significant on the basis of a permutation test (Figure 1C and 'Materials and Methods' section). Of these significantly long (≥ 7 genes) clusters, 50 comprised exclusively upregulated genes and 66 exclusively downregulated ones (median number of genes = 8

for both types, Supplementary File 2). In total, the clusters comprised 1074 genes (∼20% of the total).

Given that we measure topological stress in transient heat shock conditions, we wanted to see if the clustering effect we observe could be attributed to the temperature shift. We employed an identical clustering approach in gene expression profiles obtained upon heat shock stress conditions as published in a landmark paper (31). Even though a certain degree of clustering is observed (Supplementary Figure S2) the degree of overlap between the two conditions is insignificant. In fact, comparison of our torsional stress condition with a variety of nutrient, environmental and chemical stresses showed limited similarity in the gene expression patterns (Supplementary Figure S3). Furthermore, applying the same clustering approach in eight different stress conditions yielded numbers of clusters that were in all cases smaller and containing <70% of the genes contained in our topological-stress gene clusters (Supplementary Figure S4). Hence, the observed strong clustering tendency appears to be a characteristic property of the topologically induced/suppressed genes.

Genes belonging to the topological stress-induced genes showed a significant tendency to be co-regulated. By analyzing weighted gene expression correlations based on the largest compendium of gene expression experiments in yeast (32), we found topologically induced clusters to be have significantly greater ACS compared to a random selection of gene clusters (Figure 1D). Based on the way they were defined, we chose to refer to them as 'Topologically co-regulated gene clusters' (TCGCs) and went on to characterize them in terms of various properties.

Positional preferences of topologically co-regulated gene clusters in linear chromosomes

The chromosomal distribution of TCGCs (Figure 2A) suggests a non-random localization along the genome. Upregulated gene clusters tend to be found toward the outer boundaries of linear chromosomes, while downregulated ones show a tendency for their center, often in close proximity to the centromeres. In some cases, clusters appear to assemble in super-clusters as in the case of the right arm of chromosome 12 or the left arms of chromosomes 6 and 7. A straight-forward analysis of TCGC distance from the centromeres showed statistically significant opposing preferences for the up- and downregulated gene clusters to be located away from and close to centromeres respectively ($P \leq 0.05$, Supplementary Figure S5).

The process of DNA replication is tightly connected to DNA supercoiling. We sought to examine differences in the positions of TCGCs compared to DNA replication origins (ARS). We found downregulated clusters to be preferentially located away from DNA replication origins (ARS) (Mann–Whitney U-test $P \leq 0.0003$ compared to a random set of equal number of clusters). Even though this may be related to a lack of ARS sites in close proximity to the centromeres, the great discrepancy between down- and upregulated TCGCs to overlap DNA replication origins (Fisher's test, $P = 0.000381$, OR $= 0.175$) is characteristic of strong opposite positional preferences between the two types of clusters. This avoidance may be explained on the basis of an optimization strategy throughout the course of evolution, as genes that are more severely affected by topological stress will tend to be located far from DNA replication origins.

Opposing functional and regulatory preferences in different types of TCGCs

TopoII is essential for yeast cells and its prolonged deactivation is bound to cause a general shutdown of cellular activity. The fact, however, that a significant proportion of yeast genes respond to its transient deactivation with increased transcription levels indicates the existence of a positive effect for a subset of cellular functions. In a previous study (18) we have shown that the regulatory and functional properties of the topo-affected genes is radically different from those induced or repressed upon heat shock. In fact, the stress imposed by the deactivation of topoII is very particular, when compared to a number of nutrient, chemical or environmental stresses (Supplementary Figure S3). This prompted for a detailed functional analysis of the topologically defined gene clusters.

A functional enrichment analysis at the level of GO (Figure 2B) shows extensive differences between the two types of TCGCs, a fact indicative of their nuclear compartmentalization being echoed in their functional roles. Three main clusters are apparent: (i) functions enriched in both types of clusters include secondary metabolism and DNA maintenance. (ii) GO terms that are enriched in upregulated clusters and depleted in downregulated ones, represent functions related to cellular transport, the metabolism of cofactors and general stress response. (iii) Downregulated-specific GO terms contain basic cellular functions associated with RNA transcription and processing, translation and the nuclear environment. A general pattern suggests that the localization of clusters among chromosomes is also reflected in their functions with upregulated gene clusters being mostly enriched in peripheral functions, unrelated to the core molecular processes, as opposed to the downregulated ones that are associated with basic cell functions related to RNA production, processing and protein synthesis.

Figure 2C highlights the transcription factors whose binding sites are found more or less frequently than expected by chance for both types of TCGCs (see also Supplementary Table S1). Downregulated gene clusters tend to be mostly depleted of TFBS, which is partly explained by the fact that they are enriched in constitutively expressed genes and thus subject to less complex regulation. From a previous analysis at the level of gene promoters on the same dataset we know downregulated genes to be enriched in essential functions, with stable expression levels and mostly depleted of TATA-boxes (18). Upregulated TCGCs, on the other hand, reflect a more complex regulation pattern, with significant enrichments for factors related to chromatin structure, DNA surveillance and amino acid transport (Supplementary Table S1 and relevant discussion). This positional-functional compartmentalization is also reflected on a number of structural attributes of these clusters, discussed in the following.

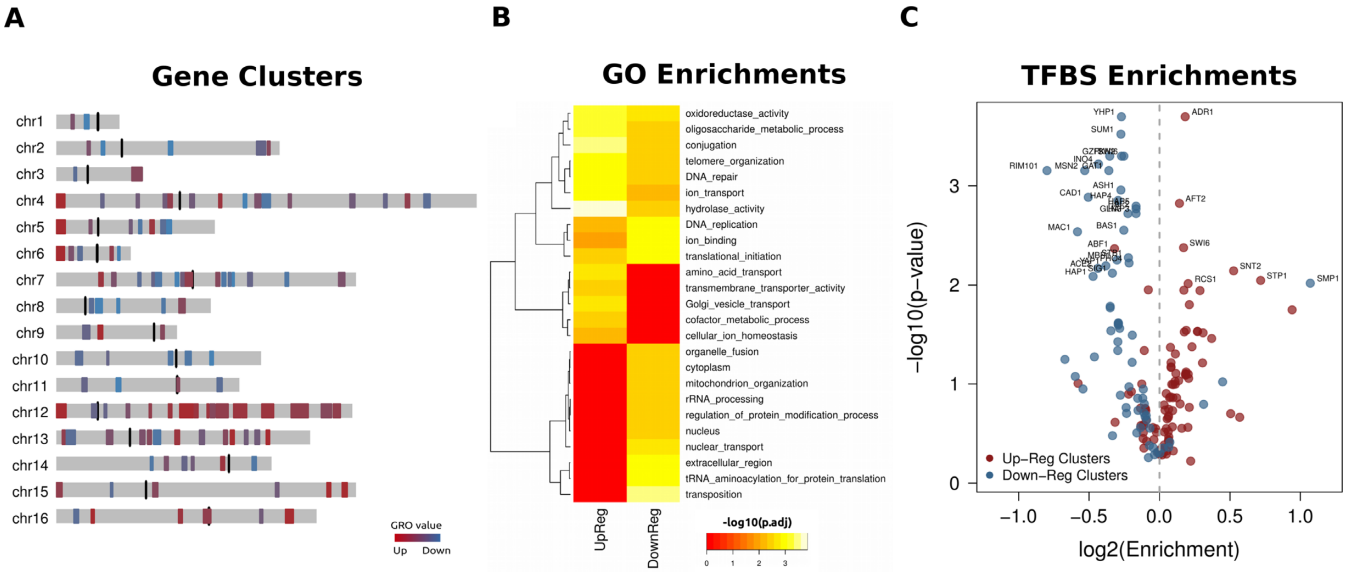

Figure 2. Functional enrichment and regulatory modes in topologically co-regulated clusters. (**A**) Distribution of 116 topologically co-regulated gene clusters (TCGCs) in the yeast genome. (**B**) GO term enrichment heatmap of TCGCs of both types. Enrichments were calculated based on a modified Gene Set Enrichment Analysis (26). Only GO terms with an adjusted *P*-value ≤ 0.05 (at 5% FDR) for at least one of the two TCGCs types are reported. (**C**) Volcano plot showing enrichments of transcription factor binding sites (TFBS) for 102 different transcriptional regulators compiled by (27). Enrichments are shown as log2-based observed/expected ratios. Values >0 indicate enrichment and values <0 indicate depletion (see 'Materials and Methods' section). *P*-values correspond to 1000 permutations for each transcriptional regulator.

Gene spacing and directionality of transcription in TCGCs

During transcription, DNA torsional stress accumulates with different sign ahead of and behind the gene's transcription start site. This makes the size of both the gene and the preceding intergenic spacer, as well as the relative direction of transcription in relation to adjacent genes, highly relevant for the dissipation of topological tension. The effect of topoII deactivation has been shown to be generally independent from the size of the majority of yeast genes (33), but it is strongly inhibitory in the case of long transcripts (34). The situation is very different when one looks, instead, into the surrounding intergenic space. When we ranked the complete set of yeast genes according to their GRO values and plotted them against the mean size of intergenic spacers we found a a clear positive correlation (Supplementary Figure S6, *P*-value $\leq 10^{-12}$) between the log-size of the intergenic regions and the observed GRO value. This is highly indicative of transcription-induced topological stress being more readily dissipated in genes with long upstream (and downstream) regions. It is worth noticing, that this is another characteristic property of topologically-induced gene clusters. When comparing the intergenic spacer sizes of up- versus downregulated gene clusters a significant difference was found for TCGCs (*t*-test *P*-value $\leq 10^{-6}$, Heat-shock clusters *P*-value = 0.09).

The association between DNA topology and structural genomic features is expected to be more pronounced in the series of adjacent genes with similar GRO values. In order to study the effect of intergenic space in co-regulated gene clusters, we employed a more relaxed criterion in the definition. We thus obtained all possible arrays of seven contiguous genes, ranked them according to their mean GRO value and kept the top and bottom 200 non-overlapping such arrays as upregulated and downregulated clusters. These contained the complete set of our TCGCs but also a number of additional gene clusters that showed consistent behavior in their response to topological stress, although not entirely positive or negative in terms of GRO value. We then expanded these clusters on either side in order to comprise 11 genes each (see 'Materials and Methods' section for details) and compared the average intergenic space along them as shown in Figure 3A. Upregulated clusters showed intergenic regions of significantly increased size compared to the genomic average (which is about 660 bp), an increase that, moreover, appeared to be inflated toward the central genes in the cluster. Genes in downregulated clusters were, on the other hand, flanked by much shorter intergenic regions and did so consistently, with little fluctuation. This striking discrepancy is not observed in the case of heat-shock or hyperosmotic stress induced clusters (Supplementary Figure S8). Besides their reduced potential for resolving topological stress, shorter intergenic regions provide shorter available genomic space for transcription factors, which may account for the marked under-representation of TFBS in downregulated gene clusters (Supplementary Figure S10B).

Figure 3A is strongly indicative of the impact of genomic architecture on the maintenance of topological equilibrium in the nucleus. Genes flanked by shorter intergenic spacers will be more prone to the accumulation of supercoiling on either side of the transcription bubble and therefore more sensitive to the lack of topoII, while genes that allow for the dissipation of topological strain into longer, untranscribed, nearby regions are predictably more resilient.

Synergistic effects between neighboring genes may be accentuated by the directionality of transcription of consecutive genes. Gene clusters with more 'streamlined' directionality patterns are expected to be able to accommodate DNA

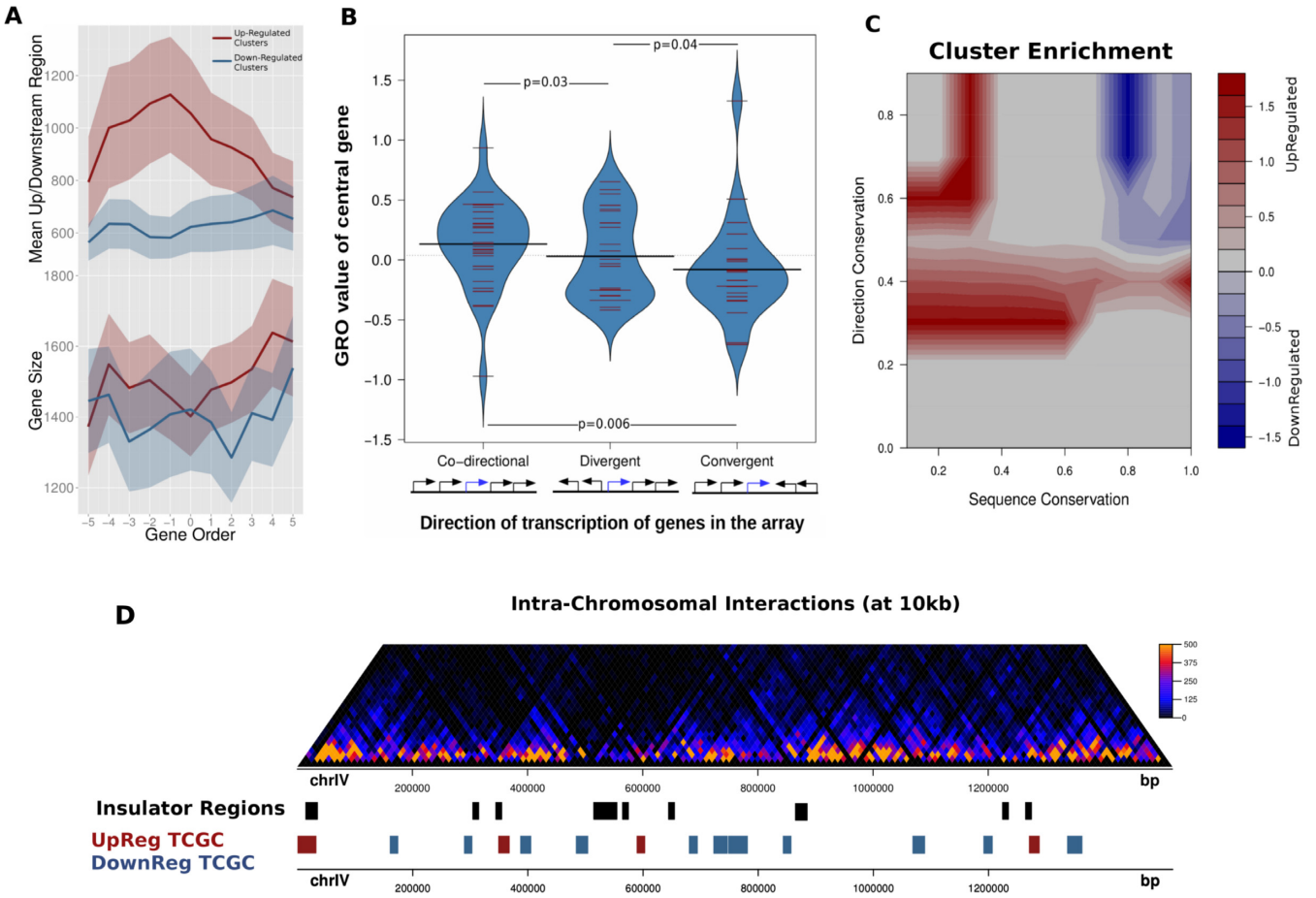

Figure 3. Topologically co-regulated gene clusters share distinct preferences for intergenic space and transcription directionality. (**A**) Top: mean intergenic region length for clusters of 11 consecutive genes. Each line corresponds to the mean values calculated for the top/bottom 200 clusters based on the central 7 GRO values (see 'Materials and Methods' section for details). Bottom: same analysis for mean gene length. Shaded bands correspond to 95% confidence intervals. (**B**) Distribution of GRO values of genes lying in the center of five-gene clusters with different directionality patterns defined on the basis of transcriptional direction (N-co-directional = 36, N-divergent = 29, N-convergent = 25). *P*-values calculated on the basis of a Mann–Whitney U test. (**C**) Contour heatmap of enrichment of different types of TCGCs in areas of mean sequence conservation (as above, x-axis) and transcriptional direction index (y-axis), defined as the proportion of genes retaining relative gene position and directionality in two closely related species. Enrichments were calculated as log2-ratios of upregulated/downregulated clusters having values in a 10 × 10 value grid (see 'Materials and Methods' section). (**D**) Intra-chromosomal interactions map of chromosome IV and corresponding domains of insulation (black). Upregulated clusters (red) show a significantly increased tendency to be located in proximity to insulation regions, when compared to downregulated ones.

supercoiling in a more effective manner, using alternating positive and negative supercoiling to 'propel' transcription. In order to test this hypothesis, we searched our gene cluster dataset for specific patterns of gene directionality. We split clusters in three categories depending on whether the central gene in the cluster (i) formed part of a series of co-directional transcriptional units or (ii) was belonging to a pair of divergently or (iii) convergently transcribed genes. We then compared the GRO values of the central gene in each category. The results (Figure 3B and Supplementary Figure S7) are indicative of a mild, yet significant association between gene directionality patterns and the response to topoII deactivation. Genes lying midway in clusters of co-directional transcription have in general higher GRO values, while genes belonging to convergent pairs have difficulty in dealing with topological tension. Divergently transcribed genes lie somewhere in the middle in terms of sensitivity as reflected in their average GRO values. Again, this appears to be a distinctive property of TCGCs. Identical analysis conducted for clusters obtained under a variety of stresses shows no differences between up- and downregulated clusters in terms of intergenic spacer sizes or gene directionality patterns (Supplementary Figures S8 and 9). Thus even if a certain clustering tendency may exist under various stress conditions, the functional and structural constraints described herein are characteristics predominantly imposed by DNA topology.

Different conservation constraints in TCGCs

In order to investigate how the properties described above may be constrained through evolution, we performed an analysis of conservation at two levels. First, we analyzed the mean sequence conservation per cluster as aggregate phastCons scores (29), obtained from a genome-wide alignment of six *Saccharomyces* species (35). Average sequence conservation (excluding intergenic space) was negatively corre-

lated with the mean GRO value for the 116 TCGCs (Supplementary Figure S10A, $P \leq 0.01$), confirming that downregulated clusters are significantly more constrained in terms of sequence conservation. Increased conservation for downregulated gene clusters does not come as a surprise given their functional preferences described in previous sections. Genes in upregulated clusters on the other hand appear to be under more moderate sequence constraint, a fact which could be indicative of their less essential role or their more recent acquisition through gene duplication (36).

We next turned to more complex conservation features that also take into account synteny relationships, reflected upon the position and transcriptional direction of genes in related species. We made use of data from the Yeast Gene Order Browser (YGOB; http://wolfe.gen.tcd.ie/ygob) (30), which contains a detailed catalog of orthologous genes between a number of yeast species. We collected all orthologous gene pairs between *S. cerevisiae* and two of its closest species in the *sensu strict*o complex, *S. paradoxus* and *S. mikatae*. We analyzed them separately for up- and downregulated clusters by calculating a simple measure of 'directional conservation' ('Materials and Methods' section). Given that syntenic regions are by definition under sequence constraint downregulated clusters were, as expected, characterized by both high sequence and directional conservation as may be seen in Figure 3C. What was rather interesting was the corresponding position of genes in upregulated clusters in the same two-dimensional constraint space. While we already knew that sequence constraints were more relaxed in these regions, we found a significant proportion of genes with high values of directional conservation, suggesting that upregulated gene clusters tend to maintain the directionality patterns even under milder sequence constraints. It thus seems, that keeping a co-directional gene layout confers a relative advantage to genomic regions that are otherwise less conserved in terms of sequence.

Topologically co-regulated gene clusters associate with different components of the three-dimensional genome structure

The eukaryotic nucleus is organized in three-dimensions, where chromosomes interact in space forming intra and inter-chromosomal domains (37) and which largely affect the processes of genome replication and transcription. Even though, the 3D organization of yeasts does not share the complexity of higher eukaryotes with topologically-associated domains and nuclear compartments, it maintains aspects of organization such as 'globules' that represent regions of increased intra-chromosomal interactions (23,38). Having observed strong positional preferences of TCGCs in linear dimension, we went on to examine whether these may be reflected in the higher-order 3D structure of the genome. We used intrachromosomal interaction data from a 3C experiment (23) to define regions of increased intra-chromosomal interactions separated by insulating regions (see 'Materials and Methods' section). We found downregulated gene clusters to be predominantly occupying regions spanning the high-interaction frequency regions, while upregulated ones were found to be significantly more proximal to insulating boundaries (Figure 3D, Fisher's test P-value $= 0.0182$, Supplementary Figure S11). This propensity is indicative of a preference of clusters that are positively affected by topological stress to occupy less entangled and more 'open' parts of the chromatin and fits well with the rest of their characteristic properties described up to now. LTR regions are known to be associated with intra-chromosomal domain boundaries and were also found to be enriched in our defined insulating regions (enrichment > 1.5, P-value < 0.001, Supplementary Table S2). There was, however, no significant overlap of our TCGCs with LTRs, a fact indicative of an LTR-independent preference of upregulated clusters to lie between high-interaction-frequency genomic domains. That this tendency is primarily linked to DNA topology, is further supported by the fact that a similar degree of enrichment was not found for any other of the studied stress conditions (Supplementary Table S2).

Differences in the distribution of TCGCs in the 3D nucleus were also picked up after an analysis at a higher level using a partition of the yeast genome in chromosomal networks (25) (see 'Materials and Methods' section). We found downregulated TCGCs to be preferentially located in the center of the nucleus, described in the model of (25) as an extensive 'community' of pericentromeric interchromosomal interactions. Upregulated ones, on the other hand, were mostly found enriched in the periphery, which is constituted by the subtelomeric regions and the right arm of chromosome 12 (Supplementary Figure S12 and Supplementary File 3).

DISCUSSION

The existence of topologically co-regulated gene clusters (TCGCs) implies that eukaryotic genes may be synergistically orchestrated in gene neighborhoods with particular characteristics. By persistently analyzing the defined gene clusters at various levels, we were able to outline an overarching pattern, according to which the yeast genome may be broadly divided in two compartments that have, in time, assumed radically different architectures and operational roles. Downregulated clusters, preferentially located toward the center of the nucleus, consisted of highly conserved genes associated with essential functions. These are expectedly shut-down by the accumulation of supercoiling in the process of a general suspension of topologically 'expensive' processes such as the transcription of rRNA. Upregulated clusters, on the contrary, predominantly comprise stress-responsive genes, whose functions may be required to dampen or even reverse topological stress. Their structural organization, with long intergenic spacers and gene transcriptional co-directionality may further enable these areas of the genome to even harness DNA supercoiling in order to achieve increased transcription levels.

Such compelling disparity at all studied levels points toward a general pattern of genome architecture. This very much resembles an urbanization process, that has over evolution demarcated an 'old-town' at the centromeric part of the nucleus, formed by tightly crammed ancient genes and a 'suburban genome' at the chromosomal outskirts, where newly acquired genes occupy greater spaces with an ordered directionality that resembles tract housing (Figure 4). This 'Genome Urbanization' is echoed in various genomic features that we have discussed in the context of TCGCs. When

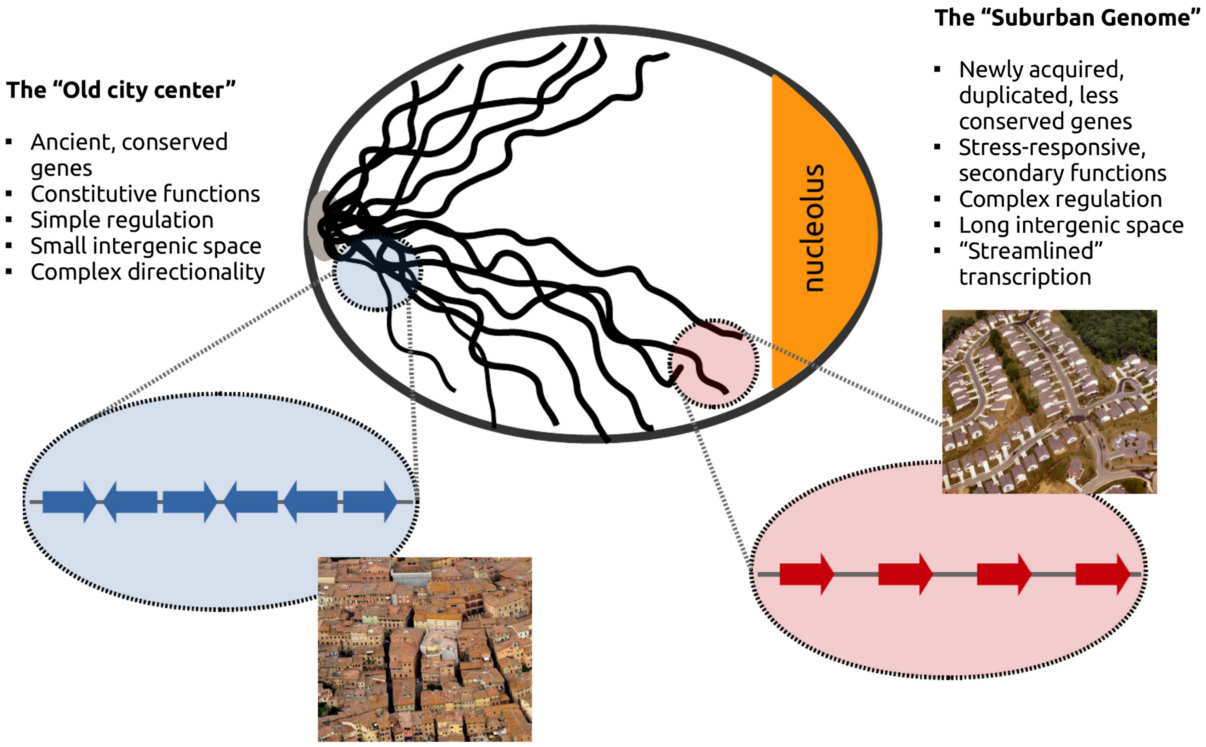

Figure 4. Genome Urbanization. Positional preferences of topologically co-regulated gene clusters reflect structural, regulatory and functional compartmentalization. Genome Urbanization in *Saccharomyces cerevisiae*. A schematic of the yeast interphase nucleus is shown based on the Rabl configuration (46). Pericentromeric regions correspond to what we call the 'Old city center' with enrichment in gene clusters downregulated under topoII deactivation. The genome in these areas may be compared to the crammed houses of a medieval town separated by narrow, intertwined alleys. Genes in the 'old town' are more conserved, associated with essential functions and located within tighter genomic spaces with fewer TFBS and entangled directionality. Genomic regions at the nuclear periphery are resembling a 'suburban landscape' where more recently acquired (and less conserved) genes are spaced in co-directional operon-like arrays, separated by longer intergenic sequences, reminiscent of the tract housing of modern city suburbia.

looking at the sequence conservation of genes as a function of their distance from the centromere we find a weak negative correlation, with the 5% most distant genes being significantly less conserved than the 5% most proximal ($n = 638$, t-test P-value $= 0.005$). Similar discrepancy is observed when looking at the intergenic space length ($n = 508$, t-test P-value $< 10^{-6}$). Gene clusters that match the properties of our TCGCs may be found in the literature, with the most notable example being the the DAL cluster (Degradation of Allantoin), located in the subtelomeric region of chromosome IX and containing a string of genes, whose directionality has been conserved among yeast species (39). Such examples suggest that the functional/structural properties described herein represent natural characteristics of the genome.

It thus appears that the division of the genome in domains with specific 'architectural' characteristics may well extend beyond DNA topology. Our findings indicate that the Genome Urbanization scheme is likely a general feature, that allows the nucleus to dissipate DNA topological stress more effectively, but whose functions are likely to extend to gene functionality (40), regulation programs (16,41) and genome evolution (42).

A particularly important element to consider is that of transcriptional plasticity. The over-representation of stress responsive genes in upregulated clusters points toward an organization of the genome, in which genes that need to readily modulate their expression levels according to environmental conditions are preferentially located in particular genomic 'niches'. Recent works have provided interesting links between plasticity and genomic features that resemble the ones we find to be hallmarks of the 'suburban genome', namely non-essentiality, complex regulation and gene duplication (43). The size of the intergenic space between genes has also been shown to widely shape expression variability (44).

The concept of 'Genome Urbanization' may extend to more complex eukaryotes, albeit not in a straight-forward manner. The size, gene density and evolutionary dynamics of the unicellular *S. cerevisiae* make the delineation of domains more clear-cut, while the complexity of gene-sparse genomes from multicellular organisms with the requirements for spatio-temporal expression patterns is bound to be reflected upon a more entangled genome architecture (45) compared to the less complex yeast genome conformation (46). The advent of new experimental approaches for the study of genome conformation in three dimensions provides a solid framework for testable hypotheses that will expand our understanding of the evolution of genome organization.

ACCESSION NUMBER

Raw data for both wild-type and thermosensitive strains are deposited at GEO with Accession Number GSE16673.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors thank Despina Alexandraki and Babis Spilianakis for a critical reading of an initial draft of the manuscript and Roderic Guigó for fruitful discussions. We thank two anonymous reviewers for suggesting control analyses, which have added to the clarity of the presented concepts.

FUNDING

University of Crete Small-Scale Research Grant [4274 to C.N.]. Funding for open access charge: Plan Nacional de I+D+I of Spain Grant Number: BFU2015-67007-P to J.R.
Conflict of interest statement. None declared.

REFERENCES

1. Caron,H., van Schaik,B., van der Mee,M., Baas,F., Riggins,G., van Sluis,P., Hermus,M.C., van Asperen,R., Boon,K., Voûte,P.A. *et al.* (2001) The human transcriptome map: clustering of highly expressed genes in chromosomal domains. *Science*, **291**, 1289–1292.
2. Lercher,M.J., Urrutia,A.O. and Hurst,L.D. (2002) Clustering of housekeeping genes provides a unified model of gene order in the human genome. *Nat. Genet.*, **31**, 180–183.
3. Batada,N. and Hurst,L. (2007) Evolution of chromosome organization driven by selection for reduced gene expression noise. *Nat. Genet.*, **39**, 945–949.
4. Dandekar,T., Snel,B., Huynen,M. and Bork,P. (1998) Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem. Sci.*, **23**, 324–328.
5. Lee,J.M. and Sonnhammer,E.L.L. (2003) Genomic gene clustering analysis of pathways in eukaryotes. *Genome Res.*, **13**, 875–882.
6. Tiirikka,T., Siermala,M. and Vihinen,M. (2014) Clustering of gene ontology terms in genomes. *Gene*, **550**, 155–164.
7. Hurst,L.D., Pál,C. and Lercher,M.J. (2004) The evolutionary dynamics of eukaryotic gene order. *Nat. Rev. Genet.*, **5**, 299–310.
8. Ljungman,M. and Hanawalt,P.C. (1992) Localized torsional tension in the DNA of human cells. *Proc. Natl. Acad. Sci. U.S.A.*, **89**, 6055–6059.
9. Naughton,C., Avlonitis,N., Corless,S., Prendergast,J.G., Mati,I.K., Eijk,P.P., Cockroft,S.L., Bradley,M., Ylstra,B. and Gilbert,N. (2013) Transcription forms and remodels supercoiling domains unfolding large-scale chromatin structures. *Nat. Struct. Mol. Biol.*, **20**, 387–395.
10. Teves,S.S. and Henikoff,S. (2014) Transcription-generated torsional stress destabilizes nucleosomes. *Nat. Struct. Mol. Biol.*, **21**, 88–94.
11. Roca,J. (2011) Transcriptional inhibition by DNA torsional stress. *Transcription*, **2**, 82–85.
12. Joshi,R.S., Piña,B. and Roca,J. (2010) Positional dependence of transcriptional inhibition by DNA torsional stress in yeast chromosomes. *EMBO J.*, **29**, 740–748.
13. Kruglyak,S. and Tang,H. (2000) Regulation of adjacent yeast genes. *Trends Genet.*, **16**, 109–111.
14. Goffeau,A., Barrell,G., Bussey,H., Davis,R.W.W., Dujon,B., Feldmann,H., Galibert,F., Hoheisel,J.D., Jacq,C., Johnston,M. *et al.* (1996) Life with 6000 genes. *Science*, **274**, 546–567.
15. David,L., Huber,W., Granovskaia,M., Toedling,J., Palm,C.J., Bofkin,L., Jones,T., Davis,R.W. and Steinmetz,L.M. (2006) A high-resolution map of transcription in the yeast genome. *Proc. Natl. Acad. Sci. U.S.A.*, **103**, 5320–5325.
16. Xu,Z., Wei,W., Gagneur,J., Perocchi,F., Clauder-Münster,S., Camblong,J., Guffanti,E., Stutz,F., Huber,W. and Steinmetz,L.M. (2009) Bidirectional promoters generate pervasive transcription in yeast. *Nature*, **457**, 1033–1037.
17. Tan-Wong,S.M., Zaugg,J.B., Camblong,J., Xu,Z., Zhang,D.W., Mischo,H.E., Ansari,A.Z., Luscombe,N.M., Steinmetz,L.M. and Proudfoot,N.J. (2012) Gene loops enhance transcriptional directionality. *Science*, **338**, 671–675.
18. Nikolaou,C., Bermúdez,I., Manichanh,C., García-Martinez,J., Guigó,R., Pérez-Ortín,J.E. and Roca,J. (2013) Topoisomerase II regulates yeast genes with singular chromatin architectures. *Nucleic Acids Res.*, **41**, 9243–9256.
19. García-Martínez,J., Aranda,A. and Pérez-Ortín,J.E. (2004) Genomic run-on evaluates transcription rates for all yeast genes and identifies gene regulatory mechanisms. *Mol. Cell*, **15**, 303–313.
20. Quinlan,A. and Hall,I. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.
21. Hibbs,M., Hess,D.C., Myers,C.L., Huttenhower,C., Li,K. and Troyanskaya,O.G. (2007) Exploring the functional landscape of gene expression: directed search of large microarray compendia. *Bioinformatics*, **23**, 2692–2699.
22. Siow,C.C., Nieduszynska,S.R., Müller,C.A. and Nieduszynski,C.A. (2012) OriDB, the DNA replication origin database updated and extended. *Nucleic Acids Res.*, **40**, D682–D686.
23. Duan,Z., Andronescu,M., Schutz,K., McIlwain,S., Kim,Y.J., Lee,C., Shendure,J., Fields,S., Blau,C.A. and Noble,W.S. (2010) A three-dimensional model of the yeast genome. *Nature*, **465**, 363–367.
24. Crane,E., Bian,Q., McCord,R.P., Lajoie,B.R., Wheeler,B.S., Ralston,E.J., Uzawa,S, Dekker,J and Meyer,B.J. (2015) Condensin-driven remodelling of X chromosome topology during dosage compensation. *Nature*, **523**, 240–244.
25. Hoang,S. and Bekiranov,S. (2013) The network architecture of the Saccharomyces cerevisiae genome. *PLoS One*, **8**, e81972.
26. Chouvardas,P., Kollias,G. and Nikolaou,C. (2016) Inferring active regulatory networks from gene expression data using a combination of prior knowledge and enrichment analysis. *BMC Bioinformatics*, **17**(Suppl. 5), 181.
27. Harbison,C.T., Gordon,D.B., Lee,T.I., Rinaldi,N.J., Macisaac,K.D., Danford,T.W., Hannett,N.M., Tagne,J.B., Reynolds,D.B., Yoo,J. *et al.* (2004) Transcriptional regulatory code of a eukaryotic genome. *Nature*, **431**, 99–104.
28. de Boer,C.G. and Hughes,T.R. (2012) YeTFaSCo: a database of evaluated yeast transcription factor sequence specificities. *Nucleic Acids Res.*, **40**, D169–D179.
29. Siepel,A., Bejerano,G., Pedersen,J.S., Hinrichs,A.S., Hou,M., Rosenbloom,K., Clawson,H., Spieth,J., Hillier,L.W., Richards,S. *et al.* (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.*, **15**, 1034–1050.
30. Byrne,K.P. and Wolfe,K.H. (2005) The Yeast Gene Order Browser: combining curated homology and syntenic context reveals gene fate in polyploid species. *Genome Res.*, **15**, 1456–1461.
31. Gasch,A.P., Spellman,P.T., Kao,C.M., Carmel-Harel,O., Eisen,M.B., Storz,G., Botstein,D. and Brown,P.O. (2000) Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell*, **11**, 4241–4257.
32. Gordân,R., Murphy,K.F., McCord,R.P., Zhu,C., Vedenko,A. and Bulyk,M.L. (2011) Curated collection of yeast transcription factor DNA binding specificity data reveals novel structural and gene regulatory insights. *Genome Biol.*, **12**, R125.
33. Pedersen,J.M., Fredsoe,J., Roedgaard,M., Andreasen,L., Mundbjerg,K., Kruhøffer,M., Brinch,M., Schierup,M.H., Bjergbaek,L. and Andersen,A.H. (2012) DNA topoisomerases maintain promoters in a state competent for transcriptional activation in Saccharomyces cerevisiae. *PLoS Genet.*, **8**, e1003128.
34. Joshi,R.S., Piña,B. and Roca,J. (2012) Topoisomerase II is required for the production of long Pol II gene transcripts in yeast. *Nucleic Acids Res.*, **40**, 7907–7915.
35. Kellis,M., Patterson,N., Endrizzi,M., Birren,B. and Lander,E.S. (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*, **423**, 241–254.
36. Fischer,G., Neuvéglise,C., Durrens,P., Gaillardin,C. and Dujon,B. (2009) Evolution of gene order in the genomes of two related yeast species. *Genome Res.*, **11**, 2009–2019.

37. Gibcus,J.H. and Dekker,J. (2013) The hierarchy of the 3D genome. *Mol. Cell*, **49**, 773–782.
38. Mizuguchi,T., Fudenberg,G., Mehta,S., Belton,J.-M., Taneja,N., Folco,H.D., FitzGerald,P., Dekker,J., Mirny,L., Barrowman,J. *et al.* (2014) Cohesin-dependent globules and heterochromatin shape 3D genome architecture in S. pombe. *Nature*, **516**, 432–435.
39. Wong,S. and Wolfe,K.H. (2005) Birth of a metabolic gene cluster in yeast by adaptive gene relocation. *Nat. Genet.*, **37**, 777–782.
40. Gehlen,L.L.R., Gruenert,G., Jones,M.B., Rodley,C.D., Langowski,J. and O'Sullivan,J.M. (2012) Chromosome positioning and the clustering of functionally related loci in yeast is driven by chromosomal interactions. *Nucleus*, **3**, 370–383.
41. Zhu,J., Zhang,B., Smith,E.N., Drees,B., Brem,R.B., Kruglyak,L., Bumgarner,R.E. and Schadt,E.E. (2008) Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. *Nat. Genet.*, **40**, 854–861.
42. Sugino,R.P. and Innan,H. (2012) Natural selection on gene order in the genome reorganization process after whole-genome duplication of yeast. *Mol. Biol. Evol.*, **29**, 71–79.
43. Lehner,B. (2010) Conflict between noise and plasticity in yeast. *PLoS Genet.*, **6**, e1001185.
44. Bajić,D. and Poyatos,J.F. (2012) Balancing noise and plasticity in eukaryotic gene expression. *BMC Genomics*, **13**, 343.
45. Bagadia,M., Singh,A. and Sandhu,K.S. (2016) Three dimensional organization of genome might have guided the dynamics of gene order evolution in eukaryotes. *Genome Biol. Evol.*, **8**, 946–954.
46. Taddei,A. and Gasser,S.M. (2012) Structure and function in the budding yeast nucleus. *Genetics*, **192**, 107–129.