Computational Methods for Analyzing Gene Regulation in Model Organisms

Research Thesis

In Partial Fulfillment of the Requirements
for the Degree of Master of Science in Computer Science

Shay Ben-Elazar

Submitted to the Senate of
the Technion - Israel Institute of Technology

Cheshvan, 5772 Haifa
October, 2012

The Research Thesis Was Done Under The Supervision of Dr. Zohar Yakhini in the Faculty of Computer Science with collaboration of Dr. Itai Yanai in the Faculty of Biology.
The Generous Financial Help Of the Technion - Institute of Technology, Is Gratefully Acknowledged. We thank Merck for a small grant and the CGC for strains used in “A Genomic Bias for Genotype-environment Interactions in *C. elegans*.”.

Publications

- Grishkevich, Vladislav, Shay Ben-Elazar, Tamar Hashimshony, Daniel H Schott, Craig P Hunter, and Itai Yanai. "A Genomic Bias for Genotype-environment Interactions in C. Elegans." *Molecular Systems Biology* 8 (2012): 587. Impact factor: 8.626, ranked 24/289 in Biochemistry & Molecular biology.

- Shay Ben-Elazar, Zohar Yakhini and Itai Yanai. "Spatial localization of co-regulated genes is greater than genomic gene clustering in the S. cerevisae genome." (Submitted to Nucleic Acids Research. Impact factor: 8.026, ranked 26/289 in Biochemistry & Molecular biology).

Contents

Computational Methods for Analyzing Gene Regulation in Model Organisms 1
Publications 3
1 Abstract 1
2 Abbreviations 1
3 Introduction 2
3.1. Background 2
3.1.1. Gene-environment interactions 2
3.1.2. Chromatin structure 3
3.1.3. Open challenges in analyzing chromatin physical conformation 5
3.1.4. Overview 6
4 Materials and Methods 6
4.1. Gene-environment interaction 6
4.1.1. Compiling a Gene-environment dataset 6
4.1.2. Analyzing Gene-environment interactions 8
4.2. Chromatin structure 9
4.2.1. Compiling chromatin structure 9
4.2.2. Analyzing chromatin structure 10
4 Results 14
4.1. Gene-environment interactions 14
4.1.1. A comprehensive Gene-environment expression dataset 14
4.1.2. Detecting a myriad of expression patterns 14
4.1.3. Genomic properties correlate with a higher complexity of regulatory interactions 15
4.1.4. Intermediate expression of interaction genes 16
4.1.5. Interactions are caused by trans effects 18
4.2. Insights from chromatin structure 18
4.2.1. An unconstrained 1kb-resolution model of the yeast genome using natural neighbor interpolation and embedding 18
4.2.2. Statistical assessment of spatial functional enrichment controlled by genomic order 22
4.2.3. Widespread spatial regions enriched for TF targets 24
4.2.4. TFs whose gene targets are spatially enriched are highly expressed 28
4.2.5. Genome structure shows recurring patterns at large scales with no evidence of related functionality 29
5 Discussion 31
References 34
Supplementary material 39

List of Figures

Figure 1 A systematic examination of gene expression variation across genotypes and environments. 3
Figure 2 Genes with genotype-environment interactions show the hallmarks of highly regulated genes. 16
Figure 3 Genes with genotype-environment interactions following functional disruption of sid-1/haf-6 also show the hallmarks of highly regulated genes. 17
Figure 4 Studying genome structure using 3C at 1kb interpolated-resolution. 20
Figure 5 Comparing functional enrichment between the genomic and spatial regions of the genome. 23
Figure 6 Gene targets of the same TF generally spatially cluster in the yeast genome. 26
Figure 7 Comparing enrichment significance against a random model. 28
Figure 8 Gene expression is higher for genes in regions of functional co-localization. 29
Figure 9 Exploring the space of substructures in the genome. 30
Figure S1 Effect of cyclic permutation on spatial and genomic co-localization enrichments. 39
Figure S2 Cumulative sum of the eigenvalues associated with the linear embedding. 39
Figure S3 Effect of diagonal forcing on embedding. 40
Figure S4 Enrichment landscape for *SIP-4*. 40
Figure S5 The effect of a permutation on gene identities to the enrichment of co-localized targets of *GLN-3*. As is evident in this figure, there are no significant enrichments once running a permutation on the gene identities, indicating that the enrichment of gene co-localization is statistically significant and stems from non-random proximity. 41
Figure S6 Expression of genes which participate in many co-localized regions compared to genes which participate in few co-localized regions. 41
Figure S7 Correlation between expression and spatial organization of TF targets. 42
Figure S8 Distribution of expression levels at the four-cell stage of N2 under control conditions. 42
Figure S9 Reproducability of the microarray as estimated by replicates. 43
Figure S10 Spike-in control in the microarray data. 43
Figure S11 Comparison across datasets. 44
Figure S12 Determination of the ANOVA significance thresholds. 44
Figure S13 Expression profiles for all 198 identified genes with genotype-environment interactions. 45
Figure S14 Multivariate analysis of the four gene sets. 46
Figure S15 Genomic properties of genes with genotype-environment interactions. 46
Figure S16 Genes with long intergenic distances and mid-range expression levels are enriched for genotype-environment interactions. 47
Figure S17 Interaction genes in a functional set. 47

1 Abstract

This dissertation embodies two separate research projects with a common goal - exploring gene regulation. In Biology, gene regulation encompasses a broad field which attempts to describe the molecular interactions between various cellular factors that conspire to silence or activate the machinery in charge of compiling a gene from its source code – the DNA, to an executable thread – Protein, which in turn works in cohort with other active machinery in the cell to determine the organism's phenotype. In the first project, we examine the environment's' effect on gene regulation through the lens of evolution, comparing gene expression of 5 strains of the nematode *C. elegans* grown in 5 different mediums. We use robust statistical methods to show that highly regulated genes, as distinguished by intergenic lengths, motif concentration, and expression levels, are particularly biased towards genotype-environment interactions. Sequencing these strains, we find that genes with expression variation across genotypes are enriched for promoter SNPs, as expected. However, genes with genotype-environment interactions do not significantly differ from background in terms of their promoter SNPs. Collectively, these results suggest that the highly-regulated nature of particular genes predispose them for exhibiting genotype-environment interaction as a consequence of changes to upstream regulators. This observation may provide a deeper understanding into the origin of the extraordinary gene expression diversity present in even closely related species..
In the second project, we take a pragmatic approach and provide an analytical framework of exploring both the structure of DNA and of detecting spatial co-localization of genomic markers. We go on to deploy this framework and provide a 3D structural model of the S*accharomyces Cerevisae* genome, and use it to provide evidence of widespread co-localization of the targets of cellular factors, termed Transcription Factors (TFs). We also describe additional work aimed at exploring the space of structural conformations of the genome in an attempt to cluster chromatin conformations.

2 Abbreviations

3C – Chromatin Conformation Capture
MDS – Multidimensional Scaling
mHG – Minimum hypergeometric
MSE – Mean square error

TF – Transcription Factor

3 Introduction

3.1. Background

This thesis focuses on computational tools designed to studying gene regulation and their application in two different types of domains.

The relative ease by which DNA and RNA can be determined using sequencing technologies has revolutionized our understanding of gene regulation. On one hand, this has led to the identification of massive amounts of gene expression changes across different strains or species, and various environmental perturbations (1-5). On the other hand, the particular conformation of the genome, which determines in turn its regulatory state (6-10), is steadily being elucidated in different conditions (11,12) at specific loci (13,14). New experimental methods (15) now enable the systematic unbiased exploration of these observations. In particular, chromosomal conformation capture (3C) followed by high-throughput sequencing has produced a quantum leap in our ability to globally model genomic structure. Using this approach and its derivatives, the genomic structure of *S. Cerevisae*, *S. pombe*, *D. melanogaster*, and human has been determined for particular conditions.

3.1.1. Gene-environment interactions

A genotype-environment interaction occurs when the effect of a genetic locus on expression is different in magnitude or direction across environments (16). As an example of a genotype-environment interaction, consider a gene induced under heat relative to non-heat in one geographical isolate but uniformly expressed in both conditions in another isolate (Figure 1A). Intuitively, the interaction arises since the environmental expression profile across genotypes is not different by a global factor but rather different for particular environments. While, genomic sequences are now readily available, predicting the effect of specific mutations on gene expression profiles presents a formidable problem. An even bigger systems biology challenge is to predict the effect of a mutation for different environmental conditions, thereby predicting genotype-environment interactions at the level of gene expression.

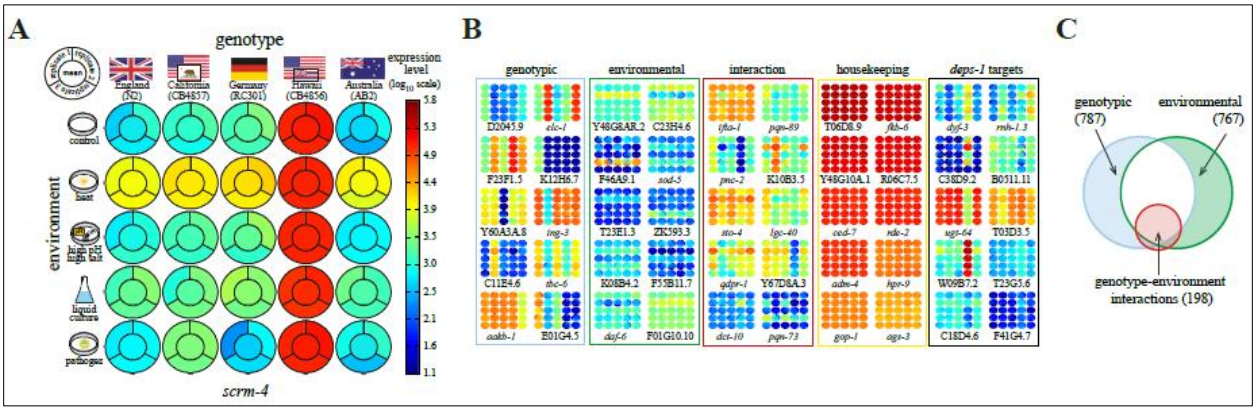

Figure 1 A systematic examination of gene expression variation across genotypes and environments.
(A) The measured gene expression levels across the 5 genotypes and 5 environments are shown for the scrm-4 gene. For each pairing, the colors in the periphery and center of the circle indicate the triplicate data and mean, respectively. Note the genotype-environment interaction. (B) Expression profiles for 50 other genes are shown in the same format. (C) Venn diagram indicating the number of genes with significant variation across genotypes (genotypic), environments (environmental), as well as genotype-environmental interactions (non-additive variation). These sets were delineated using two-way ANOVA with a threshold for significance established by randomization experiments (Figure S12).

Genotype-environment interactions have been identified at the level of a handful of genes and the genome, for single- and multi-cellular organisms, and across both strains and species (1,17-21). In particular, evidence has been provided for the notion that much of the observed gene expression variation within a species is due to changes at distant genomic positions (*trans* changes) (17-20,22). Furthermore, work in yeast has shown that genes with high expression plasticity tend to have a TATA-box in their promoter (4) and also a nucleosome occluded upstream region (23). However, it is not well understood how such *trans* effects targeting particular genes contribute to genotype-environment interactions. In the first work, we describe an investigation into the genomic properties of genes exhibiting genotype-environment interactions.

3.1.2. Chromatin structure

The initial analyses of 3C datasets have already led to insights into the structure of the genome, including the fractal nature of the human genome (24), the centromere co-localization and Rabl conformation in brewer's yeast (25), the proximity of functionally related genes in fission yeast (26), and the physical demarcation of chromosomal domains in *Drosophila* (27). The ability to measure genomic architecture in three dimensions (3D) provides an opportunity to address long

standing questions involving how genomic structure encodes the phenotype and addressing these will require new computational tools with an appropriate framework for analysis.

Of particular interest is the notion of nuclear transcription factories, and their role in establishing the regulatory states that underlie physiological stages. Most gene targets of *S. cerevisiae* transcription factors (TFs) have been determined with high confidence, revealing an average of 70 gene targets per TF (28,29). Coupling this data with genome structure enables the study of the co-localization of TF targets. For example, are the targets of the same TF co-localized to the same spatial arrangement as the transcription factory model suggests? Under which conditions does such co-localization occur? Previous analyses have addressed this question leading to contradictory results. Dai and Dai compared the number of interactions in different gene sets and observed statistical enrichment under the hypergeometric null model for interactions among TF targets (30). However, Witten and Noble argued that edges in the 3C interaction graph are not statistically independent, as was assumed by Dai and Dai, and as such co-localization events would be over-counted (31). To correct for this, Witten and Noble applied a re-sampling methodology under which no signal for TF target co-localization was detected.

Importantly, while the previous studies treated genomic proximity differently than spatial proximity, this was done by examining only inter-chromosomal distances. In additional, the spatial organization of the genome was not directly compared to the primary gene order in terms of their respective functional enrichment. This latter point is important since genomic analyses have revealed that neighboring genes tend to have similar expression profiles (32). Furthermore, genes with housekeeping functions in particular tend to be co-positioned along chromosomes (33). In particular, gene targets of the same TF are enriched for proximity in their genomic order (34). Thus, controlling for the genomic clustering is crucial for unbiased evidence regarding the degree to which the spatial clustering contributes to regulating functionally related genes.

Here we introduce a statistical framework for modeling chromatin structure and assaying the spatial proximity of functionally related genes while controlling for effects from linear co-localization along the genome. Our analysis is more subtle and flexible in refining gene sets for detecting the optimally clustered subset and defines enrichment environments more loosely based on this subset. Additionally, we apply a direct approach for controlling against results that may have emerged primarily from genomic proximity thereby focusing our results on the phenomenon of spatial co-localization. We applied this approach on a model of the genomic

structure generated using a method for the interpolation and the embedding of 3C data that circumvents observer bias by relying on a minimum set of assumptions. Our results indicate that for most TFs, the targets are significantly more co-localized in space than they are co-localized in genomic loci. We further found that TFs with spatially co-localized targets are also expressed higher under the same measurement condition, suggesting that regulatory activity is correlated with the presence of transcription factories. As more genomic structures are produced our method promises to be of importance to the study of transcription factories.

3.1.3. Open challenges in analyzing chromatin physical conformation

Recent attempts at modeling chromatin structure (24,26,35) have been prone to observer bias as state-of-the-art methods are based on solving a constrained optimization problem with a mostly arbitrary rule-set. The problem with such methods is that they tend to rely on an underdetermined set of equations with infinitely many possible solutions, or local minima, sometimes completely different from one another but with equal scores in their given target function. To resolve this issue, most approaches are to fall back to generating a torrent of such possible structures and comparing them for locally isomorphic patches. These patches are then heuristically assembled to a single structure. A different problem is systematically inspecting the co-localization of genomic annotations, e.g. functionally related genes, early replication genes, tRNA genes. The solutions for this problem tend to rely on the raw data, comparing the population of dissimilarities between the annotation group and the background. These methods are inherently prone to non-specific events and outliers and are not sensitive enough to detect significant effects that are localized to a particular subset of the annotation group. Additionally, these methods do not control for a known phenomenon which could potentially bias such observations of co-localization which stem from observed genomic clustering of functionally related genes which can arise from tandem duplications, for example.

Overall, the following major challenges in exploring gene conformation require consideration:

- A data-driven approach to genome modeling is required, along with a metric (such as MSE) to measure the quality of the model, and thus, the data.
- A need for a robust and sensitive statistic to measure an exact *P*-value for co-localization in the genome.

- An internal control for the known genomic clustering of genes.
- A method which can be shown to avoid false positives using a negative control.

3.1.4. Overview

The rest of this manuscript is divided into two main parts: in section 4 (Methods) we develop the framework for studying both Gene-environment datasets and characterizing genes by their 2-dimensional profiles which we apply to data we collected from *C. elegans*, and the framework to model and study the structure of genomes which we apply to a dataset published for *S. Cerevisae*. In section 5 (Results) we describe novel biological findings that were obtained by applying our method to biological data.

4 Materials and Methods

4.1. Gene-environment interaction

4.1.1. Compiling a Gene-environment dataset

4.1.1.1. Strains and conditions. The five *C. elegans* strains used in this study are previously collected geographical isolates. N2 was originally collected by L.N. Staniland from a mushroom compost near Bristol, England (36) and is the standard lab strain used in *C. elegans* research (37). CB4857 was collected from mushrooms in Claremont, California by E.M. Hedgecock (38). RC301 was collected in 1983 by R. Cassada from a compost heap in the Botanical Garden of the University of Freiburg in Germany (38). CB4856 was isolated from a pineapple field in Hawaii in 1972 by L. Hollen (38). AB2 was collected from soil in Adelaide, Australia by D. Riddle and A. Bird (38). The strains were propagated under control conditions: nematode growth medium (NG) with *B. subtilis* as a non-pathogenic food source. Embryos were collected by bleaching and ~2000 were placed into each of 5 conditions: 1) Control: 20°C with *B. subtilis* on NG plates; 2) Heat: 25°C with *B. subtilis* on NG plates, 3) pH/Salt/*E. coli*: 20°C with *E. coli* on high salt (4x regular NG) and high pH (8.5 relative to pH of 6 for NG) plates; 4) Liquid culture: 20°C with *B. subtilis* in S-medium in a shaker incubator; and 5) Pathogen: 20°C with *M. nematophilum* on NG plates. *B. subtilis* was used here as the standard food source in all

but one of the conditions since it is preferred by *C. elegans* relative to the *E. coli* OP50 strain (39).

4.1.1.2. Embryo collection and RNA processing. Four-cell stage embryos were isolated by mouth pipette (40). Each sample comprised 50 pooled embryos. For each genotype/environment combination there were triplicates, thus the dataset comprises 25*50*3 = 3,750 individually isolated embryos. RNA was isolated using Trizol as previously described. RNA was amplified using the Ambion MessageAmpII for two rounds in order to produce sufficient quantities for microarray analysis. mRNA was isolated, amplified, and hybridized along with Agilent Spike-ins onto one color microarrays as previously described (41).

4.1.1.3. Gene expression. We designed a custom 15K *C. elegans* microarray which was then manufactured by Agilent. The 60-mer probes were determined using OligoWiz2 (42) to target the coding region based upon the following factors: melting temperature, position along the transcript, folding potential, low-complexity in the sequence, and cross-hybridization to other coding sequences. The probes were also restricted against spanning splice junctions to avoid missing transcripts due to errors in gene structure predictions. For each gene, the best scoring probe with no significant match in other coding sequences (E-value < 0.001) was selected. This procedure yielded 16,831 gene-specific probes out of the total 20,074 genes searched. We then selected the 15,208 best scoring probes for the microarray. Data was extracted using Feature Extraction (Agilent). The raw data was normalized using quantile normalization. Analysis was done on $\log_{10}$ of the normalized data. The complete data set and array platforms have been deposited in the Gene Expression Omnibus with accession codes GSE34650 and GPL15046. The data is also available in Supplementary Table 5.

4.1.1.4. Genome sequencing of *C. elegans* strains. The strains RC301, CB4856, CB4857, and AB2 were sequenced so that together with the previously published strain, the N2 strain (43), the genomes of all examined strains were known. Genomic DNA was extracted by proteinase K digestion followed by two rounds of phenol-chloroform extraction, with an intermediate step of RNase A digestion in TE. Genomic DNA libraries were built using Illumina's standard paired-end protocol and 100x2 bp were sequenced on the Illumina HiSeq

2000 following the manufacturer's recommendations. The numbers of reads mapped to the N2 genome (Wormbase release 220) were: 119,071,331 (CB4857), 109,807,250 (RC301), 58,309,757 (CB4856) and 113,429,439 (AB2) with a coverage of 116X, 107X, 58X and 111X, respectively. SNP calling was performed using samtools utilities with the N2 genome as reference. SNPs with a variant quality score of at least 30 were selected. Overall 100,919 (CB4857), 85,776 (RC301), 184,912 (CB4856) and 98,415 (AB2) SNPs relative to N2 strain were detected. Probes on the microarray that were found to include SNPs in one or more of the strains were excluded from analysis. For this exclusion we used SNPs with all range of variant quality scores, i.e. even those with a quality score <30. 600 genes were excluded from analysis based upon this criterion. The complete sequencing data has been submitted to the NCBI SRA database with accession ID SRP011413.1 for the study. The accessions for the particular strains are SRS299995.1 (CB4857), SRS299996.1 (RC301), SRS299997.1 (CB4856), and SRS299999.1 (AB2). The SNPs in mpileup format are included as Table S6.

4.1.1.5. Gene properties. Intergenic distances and expression clusters were retrieved from Wormbase (44). Constitutively expressed genes were defined as those genes with a mean expression greater than 4 $\log_{10}$ units in all strains/conditions and an absolute expression range less than 0.2. Gene regulatory information in terms of the number of regulatory motifs per 1kb region of a genes' promoter was identified using the CISRED server (http://www.cisred.org). Motifs were required to have a *P*-value less than 0.05 and be conserved between *C. elegans* and *C. briggsae*.

4.1.2. Analyzing Gene-environment interactions

Gene-environment interactions were detected by applying a two-way ANOVA test per gene using Matlab's *anova2* function. Each gene is characterized by a 5x5x3 expression matrix where each entry is sampled under a specific condition for an orthologous variant of the gene from different strains of *C. elegans* performed in triplicates. ANOVA returns three *P*-values for the genes in question quantifying how much genes vary in response to environmental change, vary across evolutionary change (strain specific expression) or have Gene-environment interactions – expression which evolved specifically in a strain in response to an environment.

4.2. Chromatin structure

4.2.1. Compiling chromatin structure

4.2.1.1. Natural neighbor interpolation of 3C data. The raw frequency measurements provided by the yeast 3C experiment (25) was represented as a scattered sparse block matrix where each block corresponds to chromosomal pairs. Each read of a mapped paired-end insert is assigned to the mid-base of a restriction enzyme fragment in its unique genome location. Each block of the raw data matrix is then subjected to interpolation using a continuously differentiable C^1 interpolant. The natural neighbor interpolation method (45) was implemented at 1kb resolution using the *TriScatteredInterp* function in Matlab with the following modifications. First, the frequency of each position with itself was set to the highest observed frequency in the dataset. These measurements are not captured by the 3C method for technical reasons (25), but are required for the multi-dimensional scaling (MDS) in order preserve positive-definiteness. The results are robust to a wide range of different set diagonal frequencies (Figure S3). For each diagonal block matrix, "ghost points" (46) were added at 10% the distance of the chromosome size away and set to a frequency of zero. This enabled extrapolation near telomeres where there is little to no data. Finally, due to rounding errors in the interpolation the resulting matrix was non-symmetric which is resolved by averaging it with its transpose. The Voronoi tessellation, upon which natural neighbor interpolation relies, is shown in Figure 4A, where the colored domains are Voronoi cells. Each cell is generated by the intersect of all half-spaces imposed by the orthogonal separating planes between the point inside the cell and every other point separately.

4.2.1.2. Modeling genome structure. The interpolated contact frequency matrix was used as input for modeling the structure. The matrix was embedded to coordinates in an arbitrary 3-dimensional Euclidean space using non-linear metric multi-dimensional scaling (MDS, also referred to as principle coordinate analysis) (47). The three principle dimensions from the linear embedding were used as a starting reference for the genomic coordinates. Coordinates in these 3-dimensions were subjected to isotonic least-squares optimization. This approach attempts to minimize the deviation between the distances between coordinates in the resulting embedding from the distances provided as the input matrix, while also best preserving the order of pairwise distances. The target function queried which we attempt to minimize at each step of the

optimization process is the Kruskal stress-1 criterion (47), which measures relative deviations from the input matrix:

(1) $$stress\text{-}1 = \sqrt{\frac{\sum\sum(x_{ij}-d_{ij})^2}{\sum\sum(d_{ij})^2}}$$

where d_{ij} is the distance between coordinates i, j in the original input data, and x_{ij} is the distance between coordinates i, j in the resulting model. For the whole-genome embedding, we re-sampled the genome using 5kb resolution per coordinate. This lower resolution allowed the embedding process to converge at the whole-genome scale. To visualize this model at 1kb resolution, we use piecewise cubic Hermite interpolation, a C^1 interpolant for univariate data (48).

4.2.2. Analyzing chromatin structure

4.2.2.1. Functional enrichment of 3D and 1D loci. For each gene *g*, we compute the functional enrichment in 3D and 1D neighborhoods of *g*, genes which are proximal to *g*, according to the following metrics. All other genes are ordered separately according to:

1. Their interpolated contact frequency with respect to g (3D proximity to *g*),
2. Their genomic distance (1D) from *g*.

For any given TF we compute the minimum hypergeometric statistic (mHG) (49,50) for the enrichment of its target in both the 1D and 3D neighborhoods of *g*. Annotation data for TF targets was taken from a previous analysis (orfs_by_factor_p0.005_cons1 from (28)). Briefly, for a given ranked list of genes (for an example see Figure 5A), mHG finds a prefix of the list that maximizes the statistical enrichment of genes pertaining to an annotation set. The mHG *p*-value represents the likelihood of observing such an enrichment, at some prefix, under a null model (see (49,50)). We obtain a bound on the mHG *p*-value, per annotation term, and per centered gene *g* by multiplying the calculated mHG statistic by the number of genes in the annotation term. If we use $mHG - pval(\lambda)$ to denote the mHG value for a given binary vector, λ, then the bound described above is referred to as $\overline{mHG}(\lambda)$ where

(2) $$\overline{mHG}(\lambda) = B \cdot mHG(\lambda) \geq mHG - pval(\lambda)$$

With *B* indicating the number of 1's in the binary vector, λ.
To further correct for multiple testing across multiple binary vectors (annotation terms) these values are later Bonferroni-corrected. Since the process is applied on both the genomic and spatial orderings of genes, we limit the threshold search to the size of *g*'s chromosome which results in comparable *p*-values for the most enriched spatial and genomic environments centered around *g*. Hence, this implementation of mHG is partition limited as previously described (49,50). When analyzing peaks of enrichment (Figure 6A) we call for Matlab's *findpeaks* function. We limited the peak calling to a minimum distance of 10 from one another and a height of $-\log_{10}(0.05)$.

The last analysis we describe in the results compares the observed enrichment results, for a fixed given TF (specifically, the binary vector λ where its targets are true), to a background model. we took the following approach.
For each gene we compute:

(3) $$L(g) = -\log_{10}\left(mHG\left(3D(g)\right)\Big/_{mHG\left(1D(g)\right)}\right)$$

Where *3D(g)* is the λ vector reordered according to the spatial proximity of its corresponding genes to *g* (and truncated after the number of genes on g's chromosome). *1D(g)* is similarly calculated, only ordered according to the genomic proximity to *g*. Finally, *L(g)* was sorted across genes to produce L.
Separately, the same quantities were computed for each of 100 shuffled genomes (with gene identities randomly permuted). We denote $L(g)^k$ and L^k as the corresponding quantities which were computed for permutation *k*. Using these quantities, we compute Z-scores on each rank, *i*, of *L* in the following way:

(4) $$Z - score(i) = \frac{\left(L_i - \overline{L_i^k}\right)}{std\left(L_i^k\right)}$$

Where *Li* is the *i*-th value in *L*. $\overline{L_i^k}$ and $std\left(L_i^k\right)$ are the mean value and standard deviation of the i-th value of L^k across all *k* permutations, accordingly. This comparison is further exemplified in Figure 7A.

3.1.4.1. **A method for clustering structural genomic elements**. In this section we develop a method for the pairwise comparison of substructures in the genome which lay in the

nuclear space, given a complete genomic contact matrix - *A*. We first define the specific set of substructures of interest by focusing on substructures which are each defined by a ball of radius, ρ, centered on a locus in the genome, *c*. In practice we directly use the interpolated contact information, instead of a genome model, and therefore work with contact frequency threshold, θ, rather than radii. Such a structure is completely defined by the distances in the sub-matrix $A|_{c,\rho}$:

$$F(c,\rho) = \{l | A(c,l) \geq \theta\}$$
$$A|_{c,\rho} = \left[a_{i,j}\right] | i,j \in F(c,\rho) \quad (5)$$

i.e. $A|_{c,\rho}$ is the square sub-matrix of *A* which describes the contact frequencies between all loci in a ball of radius ρ as represented by contact $\geq \theta$, centered on *c*.

We will utilize in our shape comparison algorithm the observation that $A|_{c,\rho}$ is a block matrix composed of $|F(c,\rho)| \geq r \geq 1$ distinct stretches of consecutive loci along the genome, termed segments.

Consider $A|_{c^1,\rho^1}$,$A|_{c^2,\rho^2}$, two genomic structures centered upon c^1, c^2 with radii ρ^1, ρ^2 and which have r_1, r_2 segments, accordingly. If both substructures do not have the same number of points a structural alignment of pairs of points and the mean-square deviation between the structures cannot be defined. We assume W.L.O.G. that $|F(c^1,\rho^1)| < |F(c^2,\rho^2)|$, and to resolve this discrepancy we resample $A|_{c^1,\rho}$ by linearly interpolating each block, b^i, with a grid of size

$$G^i = \left\lfloor |b^i| * \frac{|F(c^2,\rho^2)|}{|F(c^1,\rho^1)|} \right\rfloor + m^i \quad (6)$$

Where m^i is used to round off remainders:

$$m^i = \left\lfloor \sum_{j=1}^{i-1} mod\left(|b^i| * \frac{|F(c^2,\rho^2)|}{|F(c^1,\rho^1)|}, 1 \right) \right\rfloor - \sum_{j=1}^{i-1} m^j \quad (7)$$

The resulting equal-sized shapes can then be compared by defining a pairing on their coordinates. Note, that the number of segments for each one of the structures does not change, and that both structure may still have a different such number. To find the optimal pairing of segments and coordinates which reduces the mean-square-error between the shapes we attempt to heuristically search for two permutations which determine the order and direction of segments in each of the shapes. Formally, $P_{i\in\{1,2\}}$ are permutations on the ordered-set $\{1, \ldots, r_i\}$ where P_i

determines the order and orientation (by inversion) of each segment. Finally, the distance function D between shapes $A|_{c^1,\rho}$,$A|_{c^2,\rho}$ is then defined as:

$$D\left(A|_{c^1,\rho}, A|_{c^2,\rho}\right) = \min_{P_1,P_2} \left\| P_1 A|_{c^1,\rho} P_1^{-1} - P_2 A|_{c^2,\rho} P_2^{-1} \right\|_F \quad (8)$$

I.e. the distance between two shapes is the Frobenius distance between their corresponding frequency matrices, where the minimum distance across all permutations which change the order, and thus pairing, of coordinates which belong to a consecutive interval on a chromosome. An example is shown in Figure 9A. We note that computing $D\left(A|_{c^1,\rho}, A|_{c^2,\rho}\right)$ precisely requires covering $O(2^{r_1+r_2} \cdot r_1! r_2!)$ permutations exhaustively. Our algorithm does complete this task for feasibly small number of segments. Specifically, if $r_1 \cdot r_2 \leq 5$ we cover all possible such permutations. For larger r_1, r_2 we employ the following search strategy using a Simulated Annealing approach:

Next permutation Computes $P_{1,2}(t)$ given $P_{1,2}(t-1)$, $r_{1,2}$ and temperature - t.

1: **for 1…*t*** – Annealing temperature**:**

2: define $S = \begin{cases} 1 & \frac{r_1}{r_1+r_2} \\ 2 & else \end{cases}$, $c_1 \sim U(0,1)$ and $c_2, c_3 \sim U([1, r_S] \cap Z^+)$

3: **if** $c_1 \leq 0.5$:

4: $P_S(t) = P_S(t-1)$ S.t. segment is c_2 inverted.

5: **else**

6: $P_S(t) = P_S(t-1)$ S.t. segments c_2, c_3 are transposed.

As for parameters to the Matlab annealing function, *simulannealbnd*, we set the stalling termination to 200 iterations, and regular termination to 800 iterations.

Using the above described distance function to systematically compare all pairs of loci in the genome required that we first reduce bias from overlapping conformations sampled in the genome. To do so, we use a greedy set-cover strategy, thereby eliminating all loci which are overlapping in their genomic coordinate composition by at least 30%. The resulting pair-wise distance matrix, *D*, is then clustered using CAST (51) with τ =15$^{\text{th}}$ percentile of the dissimilarities of *D*. To display a resulting cluster visually, we align the 3D embedding of

clustered shapes to a common Euclidean space using Procrustes alignment (52) and calculate the mean position of each coordinate *I* across members of the cluster.

4 Results

4.1. Gene-environment interactions

4.1.1. A comprehensive Gene-environment expression dataset

To study genotype-environment interactions at a genomic level, mRNA was collected from *C. elegans* embryos extracted from animals of five distinct geographical isolates (genotypes) examined in five conditions (environments) and subjected to microarray analysis. Each of the 25 genotype-environment combinations was assayed by a pool of 50 embryos collected individually at the four-cell stage, in triplicates. The four-cell stage is easy to identify morphologically and allows query of the composition of the large maternal mRNA dowry deposited in the embryo with low variability, therefore providing high sensitivity to detecting differences (40). The resulting dataset exhibited expected distributions of expression levels, high reproducibility across replicates, linear expression values of spiked-in transcripts, and congruence with a previous dataset (Figure S8-Figure S11).

4.1.2. Detecting a myriad of expression patterns

To systematically identify genes showing genotype-environment interactions, we invoked two-way ANOVA to compute the statistical significance of the variance across genotypes, environments, and their interaction. For example, the two-way ANOVA *P*-values for the *scrm-4* gene were 10^{-300} (across genotypes), 10^{-6} (across environments), and 10^{-3} (genotype-environment interaction), indicating a high significance for the observed changes across all three factors (Figure 1A). Figure 1B shows the expression of other genes exhibiting different patterns of variation. We filtered the dataset to score only those genes with a range of expression within the linear dynamic range of the microarray (2 to 5 $\log_{10}$ units, see Figure S10) and a minimum level of variation (0.5 $\log_{10}$ units, see Table S1 for robustness to this parameter). This filter reduced the set to 4,083 genes, of which 787 and 767 show significant variation across genotypes (but not environments) and environments (but not genotypes), respectively (Figure 1C), and henceforth

refer to these as genotypic and environmental genes. Consistent with previous work in yeast (4), we found that the set of genes that vary across genotypes and the set of genes that vary across environments significantly overlap ($P < 10^{-200}$, Hypergeometric distribution). Similarly, we used two-way ANOVA to define 198 genes with genotype-environment interactions (Figure 1C and Figure S13) and proceeded to query their defining properties.

4.1.3. Genomic properties correlate with a higher complexity of regulatory interactions

We first asked whether intergenic lengths might vary across sets of genes with particular expression patterns, since the intergenic distance upstream of a gene's coding region is a proxy for the length of the promoter (53). Thus, longer intergenic regions generally reflect a higher complexity in regulation (54). Constitutively expressed genes – defined as those with high expression without significant genotypic or environmental variation – have significantly shorter intergenic regions (Figure 2A), consistent with their potentially simple requirements for regulation (55) ($P<10^{-122}$, Kolmogorov-Smirnov test, henceforth KS-test). Genes showing environmental changes do not have a different intergenic lengths distribution than the background, while genotypic genes have slightly longer intergenic regions ($P<10^{-11}$, KS-test). This result suggests that an extensive promoter region may be a liability in terms of an inherent bias for producing aberrant expression patterns. Strikingly, interaction genes have intergenic regions that are significantly longer, suggesting complex regulation upon these genes ($P<10^{-7}$, KS-test). Consistently, we found a higher motif concentration in the 1kb promoter region immediately 5' of the coding region of interaction genes relative to that of all genes (Figure 2B, $P<0.039$, KS-test). The properties of intergenic length and motif concentration are significantly correlated ($P<10^{-16}$, correlation coefficient, Table S2) providing evidence for the notion that longer intergenic lengths indeed reflect increased regulation. These results implicate the interaction genes as a class of highly regulated genes in which the promoter sequence is longer and includes more motifs. Examining other genomic properties, we further found that interaction genes are also enriched in their nucleosome occupancy at the promoter region consistent with our observation of their high expression variability (Figure S14) (23).

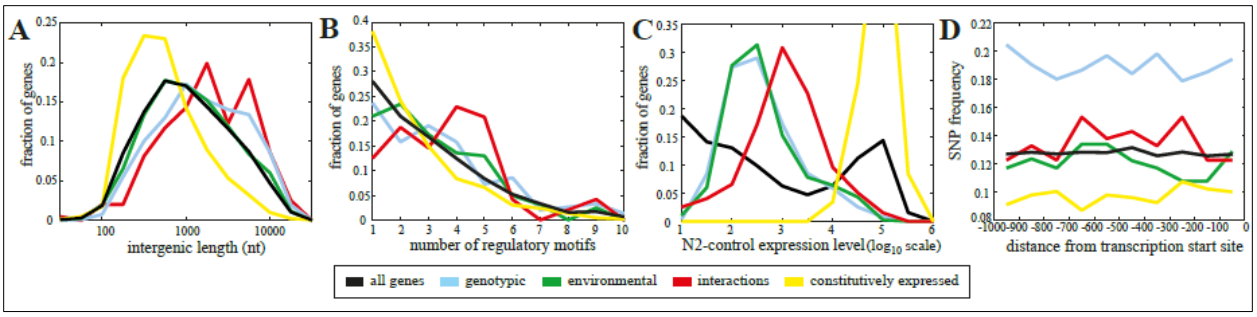

Figure 2 Genes with genotype-environment interactions show the hallmarks of highly regulated genes.
Distribution of (A) intergenic lengths, (B) motif concentration, and (C) expression levels for the indicated gene categories. Expression levels were defined according to the median across genotypes and environments. The plots indicate the normalized frequencies of the measurements across each gene set. (D) SNP analysis. For each gene set, the fraction of genes with at least one independent SNP across the strains is indicated for each 100 nucleotide promoter bin.

4.1.4. Intermediate expression of interaction genes

To further query the properties of the interaction genes we examined expression levels. Constitutively expressed genes were highly expressed (by their definition as highly and steadily expressed) while the genotypic and environmental genes had generally low expression (Figure 2C). By contrast, interaction genes occupied an intermediate position along this scale, expressed significantly higher than the environmental and genotypic genes ($P<10^{-19}$, KS-test). This predisposition towards higher expression provides additional support for the notion that interaction genes are under distinct regulation relative to the other gene classes. Since intergenic distance and basal expression levels may be thought of as proxies for highly regulated genes, we asked whether such a class of genes is enriched for genes with genotype-environment interactions. We defined a set of presumably highly regulated genes as those with long intergenic distance (>5kb) and a mid-range of expression (>2.5 and <3.5 $\log_{10}$ units); these two properties are only weakly correlated (Table S2). This set of 477 genes is enriched for genotype-environment interactions ($P<0.007$, hypergeometric distribution CDF), while lowly expressed genes (<2.5 $\log_{10}$ units) are depleted in interactions ($P<0.02$, hypergeometric distribution CDF). These trends are supported by the complete pattern of enrichments for interactions along the dimensions of intergenic distance and expression level as shown in Figure S15.

Changes in expression in the interaction genes may be due to local changes to the promoter (*cis*) (56) or to changes to either the regulators or remote regulatory regions (*trans*) (19). To distinguish between these we attempted to map the genomic changes that correlate with expression differences. We sequenced the four non-Bristol (N2) strains (see Supplementary

Information) and mapped single nucleotide polymorphisms (SNPs) across the strains to the motif-rich 1kb promoter region upstream of the start of translation of all genes. We first examined the number of promoter SNPs found in the constitutively expressed genes. These show a paucity of SNPs relative to all genes suggesting strong selection on maintaining the coherence of the promoter region ($P<10^{-11}$, KS-test relative to background, Figure 2D). Interestingly, the genotypic genes showed a higher SNP density, suggesting that a significant fraction of the changes in these genes are caused by local (*cis*) changes as opposed to changes to other factors that impinge upon its expression ($P<10^{-13}$, KS-test). However, the genes showing genotype-environment interactions (interaction genes), were not significantly distinguished in their SNP content ($P=0.93$, KS-test relative to all genes), suggesting that their expression changes are predominantly caused by *trans* effects.

If *trans* effects dominate genotype-environment interactions, our set of interaction genes are expected to be enriched for particular functions reflecting a coordinated change due to a common source. To test for this, we screened through sets of functionally related genes using Gene Ontology, Pfam, and Wormbase Expression Clusters, and queried for enrichment in similarity among the gene expression in our dataset. We found 16 gene sets with an enrichment for genotype-environment interactions ($P<0.01$, Table S3, hypergeometric distribution). One such gene set comprises the potential targets of the *deps-1* gene initially defined by the up regulation after *deps-1* loss of function (57). Of these potential targets, *scrm-4* was shown in Figure 1A with elevated expression in heat and Hawaiian and ten other genes from this set are shown in Figure 1B. These show striking interactions as also evidenced by the significant ANOVA interaction *P*-values associated with this gene set (Figure S16).

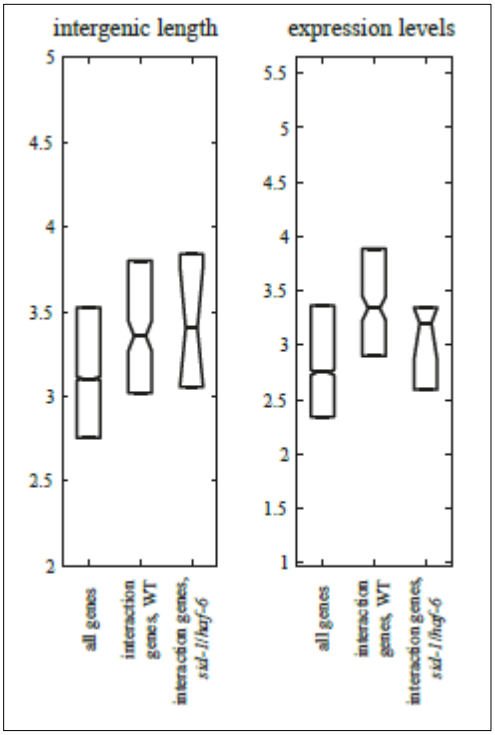

Figure 3 Genes with genotype-environment interactions following functional disruption of sid-1/haf-6 also show the hallmarks of highly regulated genes.
A. Distributions of intergenic distances, shown as boxplots, comparing the 12 genes with a genotype-environment interaction in the sid-1/haf-6 analysis (mutant and N2 strain across the five environments, P<0.005) with the background set and the 198 interaction genes in the geographical isolates analysis (Figures. 1,2).
B. The data for expression levels in the same format.

Interestingly, the *deps-1* gene itself does not show expression variation across the strains in our dataset suggesting that the difference in expression across strains may be post-transcriptional, or in a different co-regulator of these targets. The causal changes may also have occurred specifically in each of the targets, but this is unlikely since the promoters of *deps-1* targets do not show enrichment in SNPs relative to background (P<0.96, KS-test).

4.1.5. Interactions are caused by trans effects

Our results suggest that genes with long promoters and a mid-range level of expression have a disproportionately higher likelihood to develop genotype-environment interactions following *trans* changes. We next asked if a transgenic strain with introduced mutations will produce genotype-environment interactions with this same pattern. Therefore, we compared expression levels across the five conditions on the same microarray platform in triplicate for the N2 strain and a nematode strain deficient for *sid-1* and *haf-6* function in the N2 background (HC445). As expected, *sid-1* and *haf-6* transcripts were significantly reduced (P<10^{-200} and 10^{-70}, respectively). Querying the data for genotype-environment interactions we detected 12 genes with significant genotype (N2 vs. *sid-1/haf-6*) -environment interactions (P<0.005, two-way ANOVA, Table S4). Consistent with the above results, these 12 interaction genes also showed increased intergenic distances and higher expression on average (Figure 3). Although the P-value for the intergenic genes was greater than 0.1, when examining the 100 genes with the best P-values, we found a P<0.001. This independent analysis provides strong support for our findings from the geographically distributed strains that interaction genes are highly regulated and that the genotype-environment interaction is due to *trans* effects.

4.2. Insights from chromatin structure

4.2.1. An unconstrained 1kb-resolution model of the yeast genome using natural neighbor interpolation and embedding

The systematic analysis of genome structure and of 3D features of genome organization requires a coherent and comprehensive representation of the contacts between genomic loci. However, actual data resulting from 3C measurement assays are scattered across irregular genomic intervals. Thus, our first goal was to utilize the previously determined dataset (25) to study the

characteristics of the yeast genomic structure as it relates to function. To accomplish this we first set out to regularize and provide a uniformly spaced contact matrix. For this purpose we employed a natural neighbor interpolation to arrive at a 1kb resolution frequency matrix.

Since the median size of the intervals in the primary data is 1800bp (median restriction fragment length) (25) we chose to interpolate at a 1kb interval. This choice stemmed from the notion that the interpolated resolution must not greatly exceed that inherent in the primary data. We thus effectively binned the linear yeast genome to 12,071 regularly spaced 1kb coordinates. Figure 4A shows a representation of the raw data from the 3C measurement assay (25) such that each measured data point (pair of observed restriction fragments, represented by a black dot in Figure 4A) is mapped to the respective genomic loci in chromosome I. We note the sparseness of the data at some loci, as reflected by the large and irregular domains for many of the data points (see Methods), indicating the limited resolution of the data for the interaction between the respective loci. Related to this sparse sampling are the sharp discontinuities present in the data (Figure 4A). Figure 4B shows our implementation of a natural neighbor interpolation (see Methods) on the same data for chromosome I, which addresses this sparseness and sharpness by setting the local contact behavior to what would be expected of a continuously differentiable (smooth) curve. From the perspective of its differential geometry, a chromosome is expected to behave continuously due to its polymer structure and be differentiable due to the mechanical angular limitations imposed by its chemistry.

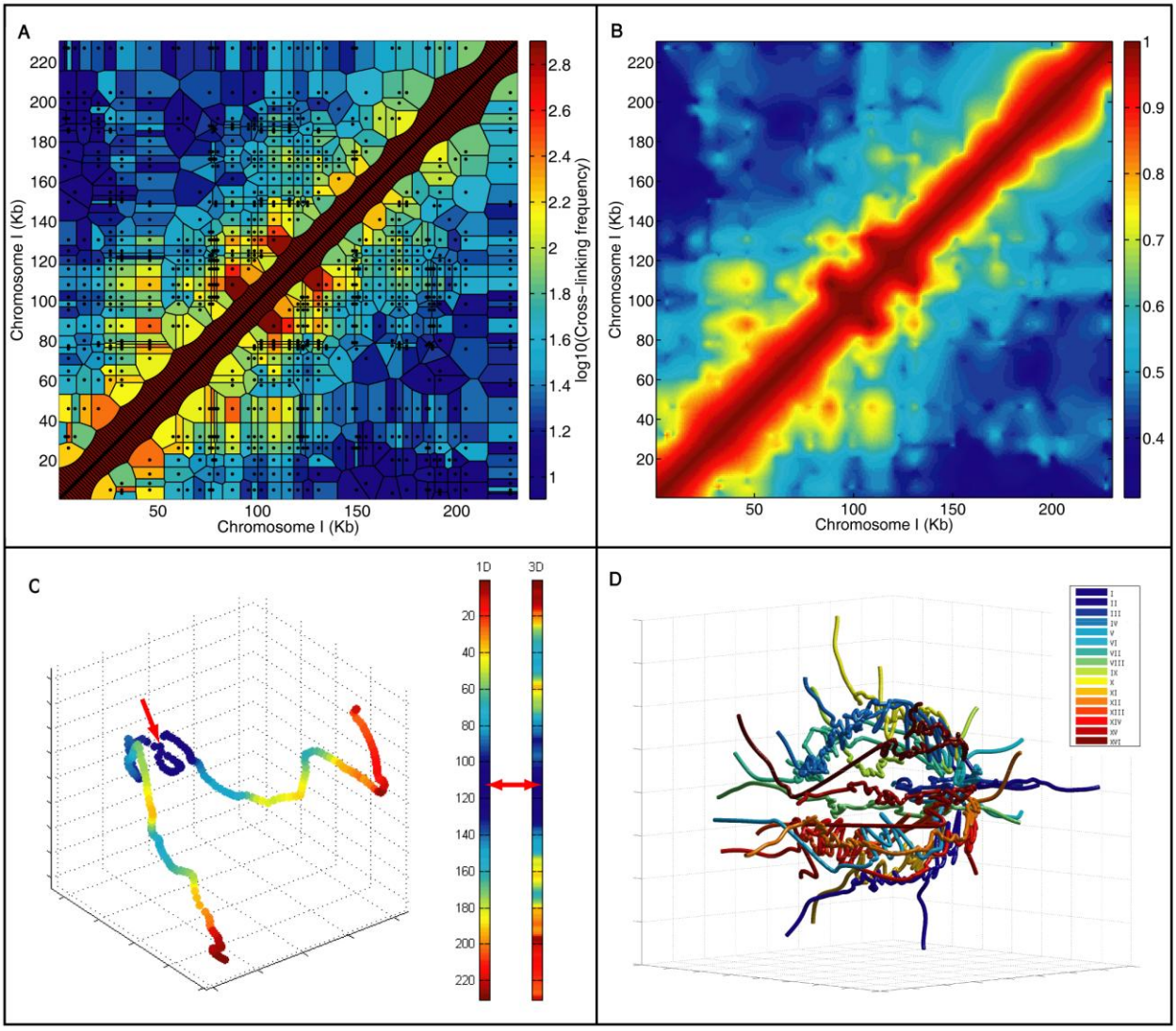

Figure 4 Studying genome structure using 3C at 1kb interpolated-resolution.
(A) 3C data for the S. cerevisiae chromosome I superimposed upon the estimated chromosomal relationships (tessellation cells) they represent. Black dots represent pairs of restriction fragment mid-points with evidence of cross-linking. Cell color indicates the observed frequency (effectively identical to a nearest neighbor interpolant). The diagonal areas are artificially inserted to overcome inherent lack of self-contacts in the method (see also Figure S3). (B) Natural neighbor interpolation of the 3C data at 1kb resolution. The colors indicate the likelihood of proximity of the genomic loci. (C) A 3D model of chromosome I generated using non-linear dimensionally reduction on the interpolated dataset shown in B. Color indicates proximity to the mid-point of the chromosome – marked with a red arrow. Note that the distance is not equivalent to the distance on the primary sequence (indicated by the left color bar) as the shape projects inwards. (D) A model of the yeast genome by non-linear dimensionally reduction as in C but extended to all chromosomes by sampling (see Methods). Note that the chromosomes lie at the periphery in a spherical fashion with the ends extended and centromeres joined.

In order to model the structure of the genome using the interpolated frequency matrix, we invoked a non-linear multi-dimensional scaling (47). This method is grounded in the well-established algebraic method of non-classical dimensionality reduction and yields a deterministic 3D view of the yeast genome using an unconstrained, and unsupervised methodology (see Methods). The linear embedding reduced the dimensionality of the dataset to orders-of-

magnitude-more dimensions than is expected of a shape measured in 3D space, reflecting the biological and measurement noise inherent in the 3C method (Figure S2). Applying this method on the intra-chromosomal interaction data of chromosome I resulted in a crescent-like curve, crumpled near the centromere (Figure 4C). Figure 4D shows the application of the method to the entire genome, resulting in a "water-lily" conformation of the chromosomes, consistent with other models proposed in the literature (25), with centromeres somewhat interwoven in one end, and chromosome arms extending outward. The quality of this embedding was quantified using the Kruskal stress-1 criterion (58). The resulting stress value of our model is 0.28, which we propose as a measure of the noisiness of the 3C data. This model is stable under small perturbations, as we show in Figure S3. In summary, our natural neighbor interpolation coupled with non-linear multidimensional scaling provides a natural 3D model of the genome at 1kb resolution.

The systematic analysis of genome structure and of 3D features of genome organization requires a coherent representation of the distances between genomic loci. However, measurements resulting from chromosome conformation capture experiments are scattered across irregular genomic intervals. Thus, our first goal in constructing a distance based description of the yeast genome (25) was to regularize and provide a uniformly spaced distance matrix. For this we employed natural neighbor interpolation to arrive at a 1kb resolution frequency matrix.
Since the median size of the intervals in the data is 1800bp (median restriction fragment length) (25) we chose to interpolate at a 1kb interval. This choice stemmed from the notion that the interpolated resolution must not greatly exceed that inherent in the primary data. We thus effectively binned the linear yeast genome to 12,071 regularly spaced 1kb coordinates. Figure 4A shows a representation of the raw data from the conformation capture experiment (25) such that each pair of observed restriction fragments is mapped to the respective genomic loci in yeast Chromosome I, represented by the black dots in Figure 4A. We note the sparseness of the data at some loci, as reflected by the large and irregular domains relating many of the data points, indicating the limited resolution of the data for the interaction between the respective loci. We further note the sharp discontinuities in the data. Figure 4B shows our implementation of natural neighbor interpolation on the same data for Chromosome I.

4.2.2. Statistical assessment of spatial functional enrichment controlled by genomic order

Using the structural model of the genome, we asked whether genes regulated by the same TF cluster together in the nuclear space. To address this question we developed a method for assessing the functional enrichment in a 3D environment. We designed the method based on three principles: 1. Direct comparison of any spatial enrichment with that observed for the linear genomic ordering, 2. Detection of enrichment of a subset rather than of correlation for the entire set (49,50), and 3. Detecting enrichment for variable-size environments, as the exact size of enriched regions was not known. The first was done to correct for the known functional co-localization of genes along the chromosomes (34). In the comparison, enrichment was favored over correlation as it is more sensitive at detecting signals at individual genomic locations, whereas genome-wide correlation methods will be dominated by noise and by effects outside of the scope of a possible transcription factory. As a statistical method we invoked the robust, sensitive and threshold-free minimum hypergeometric method (mHG) that has been successfully applied in other contexts (49,50,59,60). For each gene in the yeast genome, our method proceeds by ranking all other genes by either their genomic (linear) or their spatial (three-dimensional) distance to the gene (Figure 5A). Given a specific TF of inetrest, the mHG test is then applied to both of these two rankings in order to test whether the targets of that TF are enriched in the genomic and spatial neighborhoods of that gene (see Methods). Of particular interest are the most enriched environments, both in the genomic and in the spatial perspective, centered around a gene, as they can be compared on an equal setting. For any given locus, we quantify whether the spatial enrichment of targets is more significant than the genomic enrichment, for example, by examining the log odds ratio of the 3D and 1D enrichment *p*-values.

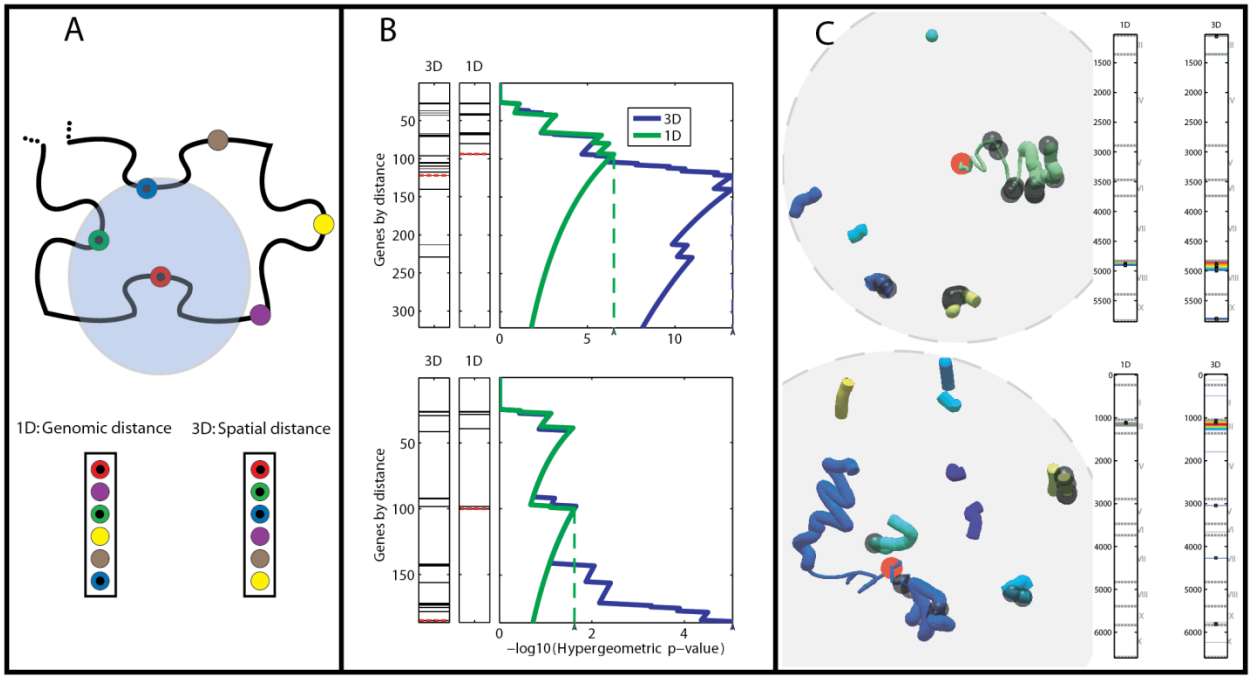

Figure 5 Comparing functional enrichment between the genomic and spatial regions of the genome.
(A) Two genomic distances. The schematic shows the gene neighborhood surrounding a particular gene (red). The neighboring genes may be ranked by their genomic proximity (left) or their spatial proximity (right). (B) Detecting areas of enrichment for TF-cohorts. In ranked gene lists, generated by either genomic or spatial proximity, the genes annotated as targets of a particular TF are indicated as black lines. The p-value of the enrichment of the targets for each threshold is indicated on the right. The threshold with the best p-value is indicated by the dashed line (see Methods). This analysis is shown for two genomic loci surrounding genes YCL012C and YHL050C respectively, and querying for targets of GLN3. (C) Local structures of the two loci examined in B. Colors indicate distinct yeast chromosomes. The red circles indicate the center gene around which co-localization was tested. The center genes shown are YCL012C (top) and both YHL050C and YHL050W-A (bottom). The content shown in each sphere is the environment which corresponds to the mHG threshold, dictated by the most enriched spatial environment for GLN3 targets. Bars on the right mark the loci along the linear genome which participate in the most enriched environment by both the genomic and spatial rankings. Black dots, both in the bars and the visualized structure, indicate gene targets of GLN3.

We demonstrate the method in Figure 5B with two specific loci in the yeast genome. In the first example (Figure 5B, top) we compare the enrichment of the targets of the TF *GLN3* in the linear genomic and spatial neighborhoods centered at *YCL012C* on chromosome VIII. The spatial enrichment, measured by the hypergeometric *p*-value, of the targets of *GLN3* increases (Figure 5B, blue line) as the radius of the ball examined (centered at *YCL012C*) is expanded (i.e., more genes at greater distances are included). In the very close neighborhood of *YCL012C* the enrichment is the same for both spatial and genomic proximity, suggesting that the genes most spatially proximate to *YCL012C* are identical to those proximate to it in the linear genomic order. Interestingly, as the number of genes included exceeds the first 100, the spatial enrichment becomes even more significant, surpassing the linear genomic enrichment. This enrichment then peaks for an environment containing ~125 genes (hypergeometric $P<10^{-12}$), after which the

addition of more distant genes diminishes the statistical significance. In comparison, the most significant enrichment based upon the genomic order alone is $P<10^{-5}$ obtained at a neighborhood that includes the nearest 80 genes. Thus, we conclude that for the environment centered on *YCL012C*, *GLN3* targets are significantly more highly enriched in space than along the linear genome.

A similar pattern is observed in the other example of *GLN3* targets when considering neighborhoods centered around *YHL050C* and *YHL050W-A*, whose transcription start sites map to the same 1kb region. For the first 140 genes added according to either genomic or spatial distance the enrichment is similar. However, as the spatial distance is allowed to increase, the enrichment sharply increases in contrast to the respective genomic distance enrichment (Figure 5B, bottom). The analysis is terminated at 200 genes, since the end of the chromosome is reached (chromosome III) and so the comparison with the linear genomic ordering is no longer possible for large neighborhoods. We note that when randomly shuffling the genomic positions of the genes we did not find any significant enrichment of co-localization, spatial or genomic, such as those shown in Figure 5B.

Examining the structural environments of the two genomic loci described above (Figure 5B) provided insight into the detected enrichments. Figure 5C shows the environments along with the corresponding genomic regions that are mapped to them. In both cases, regions from different chromosomes contribute to the significant spatial enrichment. The thin part of the chromosome on which the center gene (marked in red) is located indicates the interval with the most significant linear genomic enrichment around the center gene.

4.2.3. Widespread spatial regions enriched for TF targets

Our method allowed us to systematically test the spatial and genomic enrichments of TF targets surrounding each gene in the genome, as shown for *GLN3* targets in *YCL012C* (Figure 5B). The genomic landscape depicted in Figure 6A highlights the most significant spatial enrichment results surrounding each locus (marked in red) as well as the most significant linear genomic enrichment (marked in blue). The two specific regions shown in Figure 5C are noted with dashed

boxes. Strikingly, in many loci we observe significant spatial enrichment that is higher than that obtained for genomic order enrichment. To evaluate this result we employed two controls. First we tested whether a shuffled genomic ordering – maintaining the locations of the genes but randomizing their identities – would still lead to enrichment results, and found that as expected it does not (Figure S6). We also tested cyclic permutations of gene identities in each chromosome, and observed that the linear genomic enrichment is conserved (as clearly expected) while the spatial enrichment is eliminated (Figure S1).

To further quantify the observed higher spatial enrichment, compared to that obtained in linear genomic order, we first examined for each TF, the region with maximum enrichment at the 3D level and compared it with the 1D region that is most enriched. For *GLN3* the most significant 3D region has an associated *p*-value of 10^{-9}, while the most significant 1D region has a *p*-value of 10^{-8} (Figure 6A). Examining all 116 TF's, we found that 32 TFs have a more significant 3D region, while 6 have a more significant 1D region (Figure 6B). This indicates that when examining neighborhoods of genes, the 3D region captures more significant enrichment than an examination of solely the 1D order.

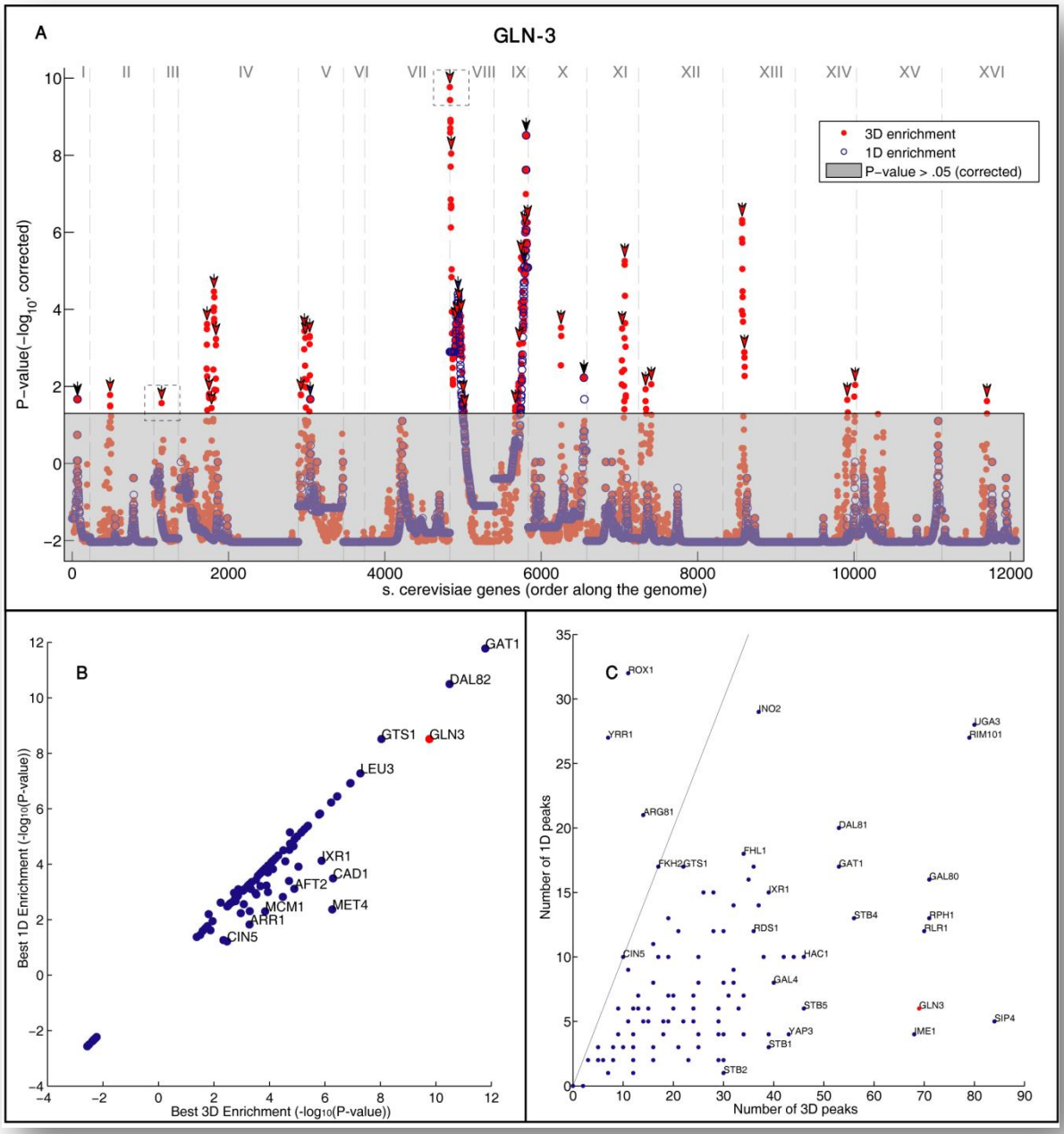

Figure 6 Gene targets of the same TF generally spatially cluster in the yeast genome.
(A) For each position in genome (x-axis, chromosomes are separated by vertical dashed lines), the p-value of the enrichment for GLN3 gene targets is shown (y-axis, -log10 of the mHG corrected p-value, see Methods). The enrichments values are shown for both the 3D (red) and 1D (blue) distances. Dotted boxes correspond to the environments shown in Figure 5B. Points in the grayed out region are below the significance threshold (P>0.05, mHG, corrected). Peaks over the significance threshold are indicated by arrows. Figure S5 shows the effect of running the same analysis on one random permutation of the target genes of GLN3 (B-C) Analysis on the gene targets of 107 TFs. GLN3 is marked in red. (B) A comparison between the maximal –log10 p-value for 3D and 1D enrichments for each examined TF. (C) A comparison of the number significant spatial (3D) and genomic (1D) regions (peaks; marked with arrows in A, see Methods) for each examined TF. The line indicates a unity relationship.

Next, we deployed a peak detection algorithm on the genomic landscape to identify distinct regions of locally maximal enrichment. We assigned each peak to either the 3D or 1D enrichment depending upon which is more significant, delineated to both in the case of a tie. Using *GLN3* again as an example, we detected 70 and 5 for the 3D enriched peaks and 1D enriched 1D peaks, respectively (Figure 6A, black arrows). A paired *t*-test on the 3D and 1D enrichments peaks indicated the significance of spatial enrichment ($P<10^{-6}$). Thus, for this transcription factor more enrichment is detected at the spatial level than in the genomic level, providing evidence for the tendency of the genome to co-localize its targets in transcription factories. Expanding these analyses to the rest of the TFs, we found an overall preponderance of 3D clusters relative to 1D clusters ($P<10^{-30}$ Kolmogorov-Smirnov test between the distributions of the number of peaks in 3D versus those in 1D). For some TFs this effect is particularly strong (Figure 6C), while for three TFs – ROX1, YRR1, and ARG81 – the signal is reversed; a more significant 1D clustering than 3D. SIP4 shows the most extreme spatial co-localization relative to genomic order (84 to 5, respectively, Figure S4). 64 of 117 TFs show a significant ($P<0.05$, FDR corrected, one-tailed two-sample *t*-test) enrichment of spatial (and 10 of 117 a significant enrichment of genomic) co-localization of their targets.

The peak analysis may be biased since we filter out genomically consecutive signals (1D) but not potentially overlapping 3D signals. To address this, we compared our observed enrichments to a suite of 100 genomes whose gene order has been shuffled using a ranking based approach (see Methods). Comparing with the randomly annotated genomes has the additional feature of direct *p*-value estimations without recourse to multiple testing corrections and parametric distribution assumptions. For *GLN3,* filtering for genes with two orders of magnitude more significant 3D to 1D and vice versa (non-grey region), the Z-scores indicate strong significance relative to the shuffled genomes (Figure 7A). Repeating this analysis for all of the available TFs, we found that for most TFs the Z-scores are positive, indicating that 3D enrichment is significantly greater than 1D enrichment when comparing to the random background model. Interestingly, some TFs show a wide bimodal distribution, indicating that the TF has both significant 1D and 3D regions of significant enrichment. We conclude that for most TFs we detect significant spatial co-localization of the targets.

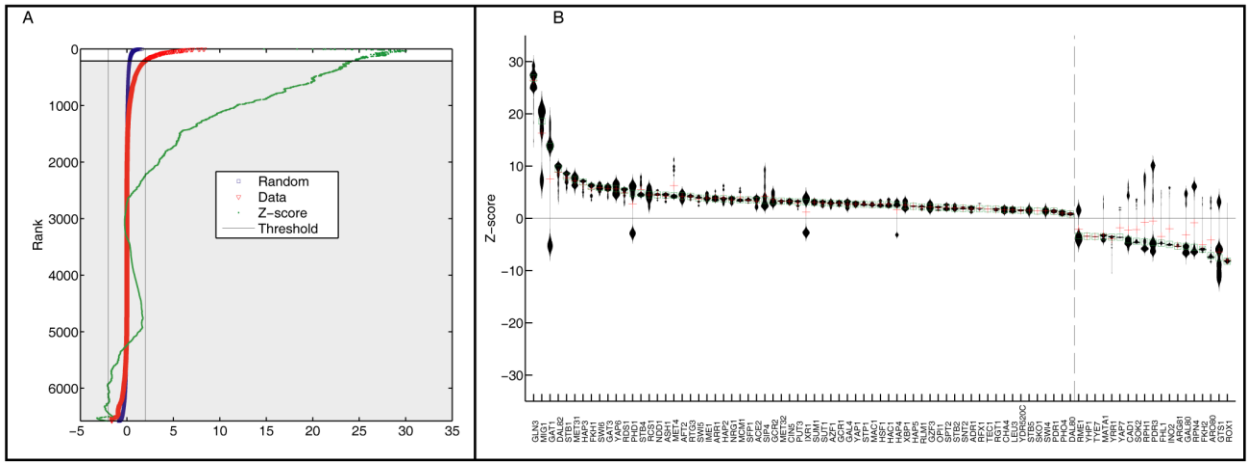

Figure 7 Comparing enrichment significance against a random model.
(A) Ranking of the relative 3D to 1D enrichment (-log odds) for GLN3 targets in the regions surrounding all genes is shown in red. The same is also shown for mean value of 100 gene order permutations is shown in blue. The Z-score is shown in green. Ranked indices with log odds which cross a significance threshold are used in downstream analysis (see Methods). (B) Showing the distribution of selected Z-scores for each TF. The dashed line separates the TFs which have positive median Z-score values from those with negative ones. TFs left of the line have more 3D enrichment than expected at random whereas TFs right of the line are ones with more 1D enrichment than expected by random.

4.2.4. TFs whose gene targets are spatially enriched are highly expressed

If the targets of a particular TF show significant co-localization in the genome, one would expect that TF to be functional under the conditions sampled for the genomic structure. A proxy for the function of a TF is its expression level, and thus we asked whether those TFs showing the strongest signals of co-localized targets are also more highly expressed (61).

We sorted TFs according to the ratio of spatial to genomic co-localization of their targets, an indication of their 3D co-localization. The expression of the top 50 TFs was then compared to that of the bottom 50. We detected a significant difference in expression (Kolmogorov-Smirnov $P<10^{-2}$) shown in Figure 8A. Overall, the correlation between the degree of target co-localization (significance of target co-localization *p*-value) and the gene expression was $r=0.25$ ($P<10^{-2}$). This correlation between expression levels and large-scale target co-localization suggests that as the cell regulates the expression of TFs, the genomic structure may rearrange to accommodate different transcription factories.

Finally, we queried for the spatial location of the apparent transcriptional factories. For each gene, we computed the number of instances in which a spatial region including that gene is enriched for TF gene targets more than for the genomic order, across the set of 107 TFs. Figure

8B shows these locations superimposed upon the genomic structure. We found that regions that are enriched for such 'transcriptional factories' indeed form distinct clusters. In particular, we observe a high degree of association of genes with transcription factories in the periphery, mainly located on chromosome II, and also on chromosome XV and chromosome XVI (Figure 8B). Comparing the expression of the set of genes highly associated with factories (>25 TF sets) relative to the genes only weakly associated with factories (<25 TF sets) we find that the former genes are more highly expressed (P<0.05, Kolmogorov-Smirnov, Figure S6). This provides further evidence that transcriptional factories generally correspond to transcriptionally active regions.

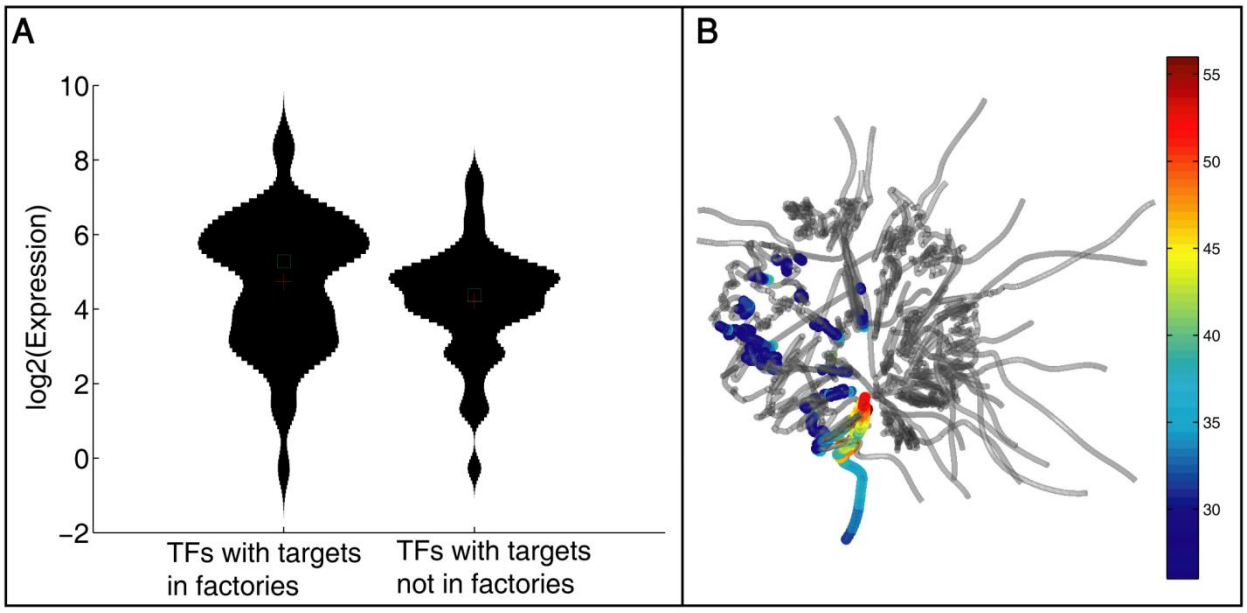

Figure 8 Gene expression is higher for genes in regions of functional co-localization.
(A) Violin plots of expression levels of TFs with and without spatially co-localized gene targets. The expression values are compared for the 50 TFs with the highest and lowest spatial localization score (-log of the ratio of t-test p-value comparing genomic and spatial enrichment co-localization). TFs with spatial co-localized targets have a significantly higher expression (P<0.01, KS-test). (B) Spatial locations of transcriptional factories. Superimposed on the genome structure, for each gene the color indicates the number of instances that the 3D structure is more significantly enriched in gene targets with respect to the linear order.

4.2.5. Genome structure shows recurring patterns at large scales with no evidence of related functionality

Covering the genome in semi-overlapping shapes (upto 30% identity in coordinates) and running our structural clustering algorithm has produced small (<40 instances) clusters of reoccurring conformations with ~150kb mean genomic coordinate content (Figure 9B). Clusters of structures

at this scale normally consist of a single chromatin fiber with one or few segments (Figure 9C). Reoccurring elements are found across the genome (Figure 9D). Extensive study into the properties of genes in each structural cluster has yielded no insight into functionality of such structural clusters.

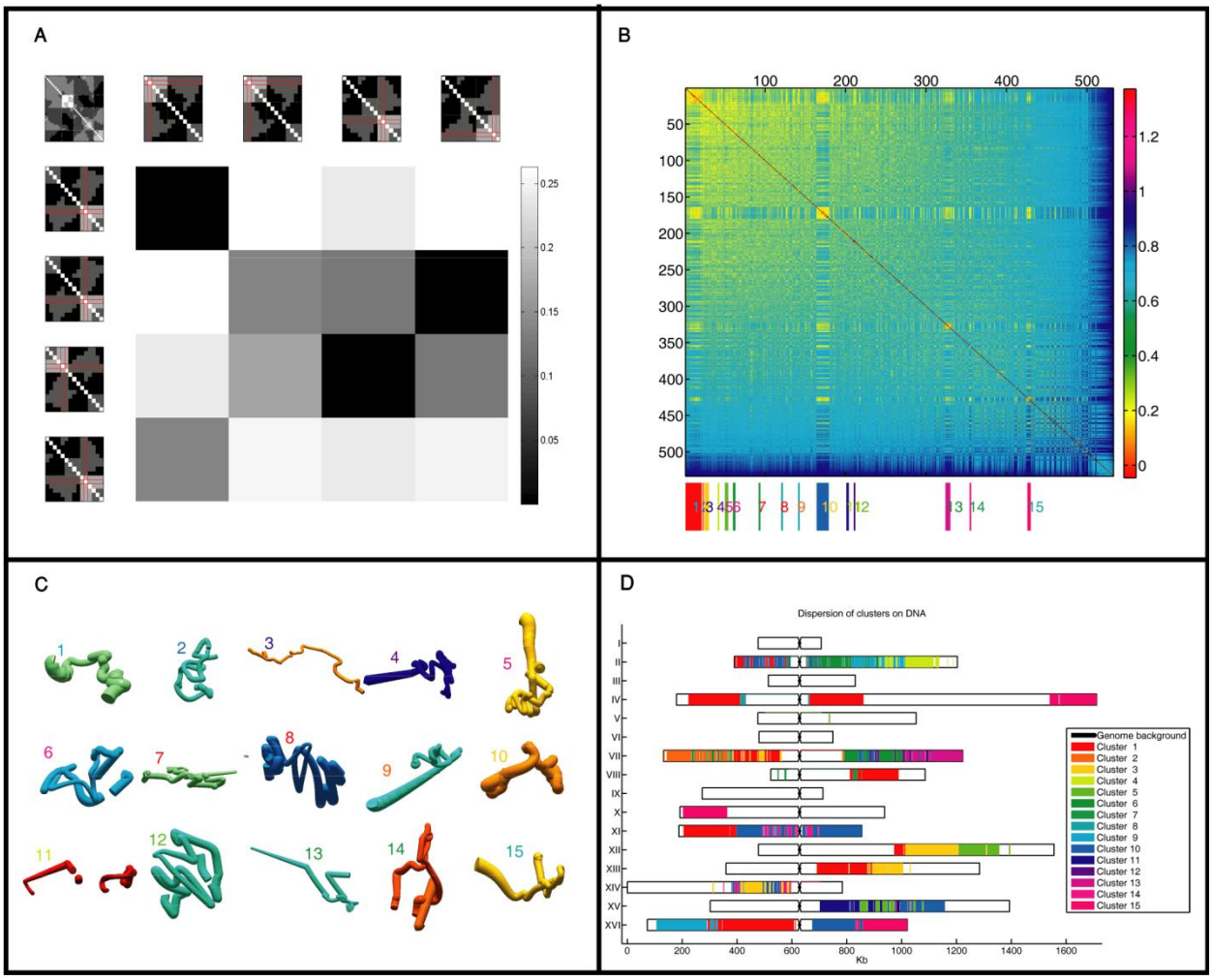

Figure 9 Exploring the space of substructures in the genome.
(A) Illustrating the method of matching conformations. Top left is a distance matrix for a sample of 35 coordinates spread across 4 chromosomes from the genome. Top row and left column are an example of two subsets of the 35 coordinates, each shown with 4 permutations on the order and direction of the 4 chromosomes. The matrix to the bottom right is the Frobenius distance between each pair of permuted subsets, indicating the importance of the order of coordinates when comparing a pair of conformations. (B) Clustergram generated by CAST with $\tau = 0.85$ of 530 selected shapes at $\rho = 20$ (see Methods) with the resulting distances among them. Clusters with more than three representatives are marked at the bottom and numbered 1 to 15. (C) Mean representation of the shapes in each of the 15 resulting clusters. Each cluster is graphically aligned to its centroid using the Procrustes algorithm (see Methods). The average location of each coordinate is shown as tubes generated using a frenet-frame. The width of the tube at each coordinate is a function of the standard deviation. (D) Each cluster is superimposed to the linear genome. Chromosomes are aligned at the centromere horizontally. Each position which belonged to a clustered shape was colored by the best representing shape around it at $\rho = 20$ (see Methods).

5 Discussion

In summary, we have provided evidence that genotype-environment interactions are enriched for highly regulated genes whose differential expression across strains and conditions is most likely due to *trans* effects. This ease by which new expression can arise in an environment-specific manner may itself be under selection. Such facilitated variation (62) for gene expression diversity may also explain in part the large amounts of divergent expression observed between species (41,63). We expect that future work will be directed towards generalizing the approach to developmental time-points, cell types, and conditions. These can be expected to provide an understanding of how genotype-environment interactions arise in the transcriptome; a readily assessable and quantifiable phenotype of the genetic material. However, gene expression in the fullest sense must include protein activity and contributions to fitness (64) and these provide a challenging goal for the greater understanding of the influence of the genotype and the environment on the organism.

Any advancement of biological methods to identify the precise structure of the genetic material throughout the life of an organism must be matched in rigor by the computational and statistical platforms that are used to interpret their measurement results. 3C has emerged as the most generalized method for establishing the structure of the genome in a systematic fashion (15). However, the statistical methods to make the most of the resulting data are only starting to be developed (24,25,65,66). Here, we report a novel approach to several aspects of the analysis of spatial conformation data. We model the structure of the *S. cerevisae* genome without the previously imposed assumptions (see below), thus capturing an unbiased representation of the data in 3D. Our method is based upon standard approaches in computational geometry, statistics and linear algebra (47), invoked here for the first time to the problem of genomic structure. We use the resulting contact matrix to ask whether functionally related genes are co-localized in the 3D structure. Using a rigorous and controlled statistical approach we provide evidence for this notion. In this section, we consider the advantages and limitations of all aspects of our methodology including the choice of interpolation and embedding procedures, internal reference to the one dimensional gene order as a control. Finally we discuss the notion of widespread transcriptional-control by spatially-defined factories.

Existing literature which addresses directly the problem of contact map completion in the context of 3C data relies either on a convolution with a fixed environment size (25,26,35) or a statistical background model to estimate either enrichment or depletion of observed contacts (27,66). Convolution based approaches lead to locally smoothed regions, while disproportionately distorting structures in data-sparse or outlier-rich regions. Both of these previously used approaches are dependent on a subjective choice of parameters such as the environment size and latent variables for statistical model. Since our method is fully reliant upon a complete contact map, we established a robust approach to generate a full contact map by interpolating missing data. We propose that the most appropriate interpolation method for completing 3C data is a modification of natural neighbor interpolation (C^1 family of interpolants). Natural neighbor interpolation is immune to the disadvantages inherent in nearest neighbor interpolation, where different genomic loci may optimally occupy the same position in space and tie-breaking scenarios are typically addressed in an arbitrary fashion. Further, natural neighbor is not as simplistic as bilinear interpolation, where only the two flanking data points in each dimension contribute to the interpolated value. Additionally, natural neighbor interpolation has been previously applied successfully for problems of smooth surface reconstruction (67) which relate to our problem in nature. Based upon a tessellated view of the data (see Methods), natural neighbor interpolation computes the weighted average of all the neighboring data points that can contribute to the information of the contact between the locations under interpolation. We note that our interpolation approach – and likewise all interpolations – does not necessarily yield inter-point contacts that mathematically qualify as a metric, and as such, the resulting contact map does not necessarily describe a structure residing in a Euclidean space precisely. To visualize the resulting interpolated contact map we attempted to generate a structural model which best captures the data.

Previous studies attempting to generate a structural model for chromatin used supervised rule sets, a random starting conformation, and optimization algorithms in order to fix each coordinate pair in its expected distance (if available) from one another (25,26,35). We propose that since such methods rely upon an underdetermined process, they cannot be rigorously applied to explore the most likely conformation. Our approach utilizes metric dimensionality reduction as a

starting point, which sets as a starting conformation the principle three dimensional outline of the shape. This outline is expected to capture the essence of the underlying geometry of the data. The optimization process preserves the order among contacts, maintaining the coherence of contacts in the resulting structure. Multidimensional scaling (MDS) is a classical algebraic and statistical approach which is well-established in the literature (47). MDS relies on a practical assumption and attempts to minimize the square error of inter-point distances while maintaining their order when comparing the input data to the resulting model. Our approach thus minimally intervenes with the underlying measurements applying a parsimonious genome modeling preferences.
We provide a solid statistical framework to determine enrichment in the spatial co-localization of genomic elements and apply it to detect a significant co-localization of TF targets. We also show a correlation between co-localization and higher expression of the targeting TF. Our results are thus consistent with previous studies attempting to link gene organization with control and regulation of transcription (11,12,14,68-72), and further extends previous systematic approaches to provide the imperative comparison to the genomic proximity of co-regulated genes.

Collectively, these results indicate that genome remains poised for the expression of co-regulated genes by adjusting their conformation to enrich for their co-localization. This conformation may likely have benefits in terms of the operations of an activated TF, which if shuttled to a region with enriched targets it will have a reduced number of possible gene targets to interact with by diffusion. This scenario would suggest that the mechanism for co-localization (whether active, or passively selected for), along with higher expression for the active TF, work in concert to regulate gene circuits and the interplay between them is crucial to understanding expression regulation.

Future directions will no doubt include a comprehensive analysis of co-localization of genomic elements to detect functional partitioning and to better characterize transcription factories. Additionally, it will be interesting to examine the extent of which these findings will be conserved across organisms and tissues. Single-cell based 3C methods – currently unavailable but sorely needed – will be able to produce a more accurate picture of genome structure, rather than a population-mean approach. Using sophisticated statistics for the detection of co-enrichment of ordinal measurements, similar methodology will surely be applied directly to non-

binary or thresholded experimental results, such as the ones from ChIP experiments to provide more unbiased views on annotated features.

Jointly, this manuscript describes innovative approaches to the study of gene expression. Unifying these methods is both the goal of deciphering some of the epigenetic principles of how genes are dynamically regulated and the rigor with which the applied computational frameworks were chosen. Collectively, these works provide compelling evidence for gene expression regulation by physical states of chromatin, and its imposing plasticity on the localization of genes. These results potentially add additional layers of complexity to the mechanisms of regulation which warrant further studies. The methods provided herein will no doubt provide a fertile basis for such studies and others as similar datasets surface with the onset of the next-generation sequencing revolution.

References

1. Tirosh, I., Wong, K.H., Barkai, N. and Struhl, K. (2011) Extensive divergence of yeast stress responses through transitions between induced and constitutive activation. *Proceedings of the National Academy of Sciences of the United States of America*, **108**, 16693-16698.
2. 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.
3. Rifkin, S.A., Kim, J. and White, K.P. (2003) Evolution of gene expression in the Drosophila melanogaster subgroup. *Nat Genet*, **33**, 138-144.
4. Tirosh, I., Weinberger, A., Carmi, M. and Barkai, N. (2006) A genetic signature of interspecies variations in gene expression. *Nat Genet*, **38**, 830-834.
5. Yanai, I., Graur, D. and Ophir, R. (2004) Incongruent expression profiles between human and mouse orthologous genes suggest widespread neutral evolution of transcription control. *Omics*, **8**, 15-24.
6. Peric-Hupkes, D., Meuleman, W., Pagie, L., Bruggeman, S.W., Solovei, I., Brugman, W., Graf, S., Flicek, P., Kerkhoven, R.M., van Lohuizen, M. *et al.* (2010) Molecular maps of the reorganization of genome-nuclear lamina interactions during differentiation. *Mol Cell*, **38**, 603-613.
7. Finlan, L.E., Sproul, D., Thomson, I., Boyle, S., Kerr, E., Perry, P., Ylstra, B., Chubb, J.R. and Bickmore, W.A. (2008) Recruitment to the nuclear periphery can alter expression of genes in human cells. *PLoS Genet*, **4**, e1000039.
8. Hiratani, I., Ryba, T., Itoh, M., Yokochi, T., Schwaiger, M., Chang, C.W., Lyou, Y., Townes, T.M., Schubeler, D. and Gilbert, D.M. (2008) Global reorganization of replication domains during embryonic stem cell differentiation. *PLoS Biol*, **6**, e245.
9. Meister, P., Towbin, B.D., Pike, B.L., Ponti, A. and Gasser, S.M. (2010) The spatial dynamics of tissue-specific promoters during C. elegans development. *Genes Dev*, **24**, 766-782.

10. Vastenhouw, N.L., Zhang, Y., Woods, I.G., Imam, F., Regev, A., Liu, X.S., Rinn, J. and Schier, A.F. (2010) Chromatin signature of embryonic pluripotency is established during genome activation. *Nature*, **464**, 922-926.
11. Chambeyron, S., Da Silva, N.R., Lawson, K.A. and Bickmore, W.A. (2005) Nuclear re-organisation of the Hoxb complex during mouse embryonic development. *Development*, **132**, 2215-2223.
12. Junier, I., Dale, R.K., Hou, C., Kepes, F. and Dean, A. (2012) CTCF-mediated transcriptional regulation through cell type-specific chromosome organization in the beta-globin locus. *Nucleic Acids Res*.
13. Schoenfelder, S., Sexton, T., Chakalova, L., Cope, N.F., Horton, A., Andrews, S., Kurukuti, S., Mitchell, J.A., Umlauf, D., Dimitrova, D.S. *et al.* (2010) Preferential associations between co-regulated genes reveal a transcriptional interactome in erythroid cells. *Nat Genet*, **42**, 53-61.
14. Zimmer, C. and Fabre, E. (2011) Principles of chromosomal organization: lessons from yeast. *J Cell Biol*, **192**, 723-733.
15. Sajan, S.A. and Hawkins, R.D. (2012) Methods for Identifying Higher-Order Chromatin Structure. *Annu Rev Genomics Hum Genet*.
16. Mackay, T.F., Stone, E.A. and Ayroles, J.F. (2009) The genetics of quantitative traits: challenges and prospects. *Nature reviews. Genetics*, **10**, 565-577.
17. Smith, E.N. and Kruglyak, L. (2008) Gene-environment interaction in yeast gene expression. *PLoS Biol*, **6**, e83.
18. Tirosh, I., Reikhav, S., Levy, A.A. and Barkai, N. (2009) A yeast hybrid provides insight into the evolution of gene expression regulation. *Science*, **324**, 659-662.
19. Wittkopp, P.J., Haerum, B.K. and Clark, A.G. (2004) Evolutionary changes in cis and trans gene regulation. *Nature*, **430**, 85-88.
20. Li, Y., Alvarez, O.A., Gutteling, E.W., Tijsterman, M., Fu, J., Riksen, J.A., Hazendonk, E., Prins, P., Plasterk, R.H., Jansen, R.C. *et al.* (2006) Mapping determinants of gene expression plasticity by genetical genomics in C. elegans. *PLoS Genet*, **2**, e222.
21. Gerke, J., Lorenz, K., Ramnarine, S. and Cohen, B. (2010) Gene-environment interactions at nucleotide resolution. *PLoS Genet*, **6**.
22. Wittkopp, P.J., Haerum, B.K. and Clark, A.G. (2008) Regulatory changes underlying expression differences within and between Drosophila species. *Nat Genet*, **40**, 346-350.
23. Tirosh, I. and Barkai, N. (2008) Two strategies for gene regulation by promoter nucleosomes. *Genome research*, **18**, 1084-1091.
24. Lieberman-Aiden, E., van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O. *et al.* (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, **326**, 289-293.
25. 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.
26. Tanizawa, H., Iwasaki, O., Tanaka, A., Capizzi, J.R., Wickramasinghe, P., Lee, M., Fu, Z. and Noma, K. (2010) Mapping of long-range associations throughout the fission yeast genome reveals global genome organization linked to transcriptional regulation. *Nucleic Acids Res*, **38**, 8164-8177.

27. Sexton, T., Yaffe, E., Kenigsberg, E., Bantignies, F., Leblanc, B., Hoichman, M., Parrinello, H., Tanay, A. and Cavalli, G. (2012) Three-dimensional folding and functional organization principles of the Drosophila genome. *Cell*, **148**, 458-472.
28. MacIsaac, K.D., Wang, T., Gordon, D.B., Gifford, D.K., Stormo, G.D. and Fraenkel, E. (2006) An improved map of conserved regulatory sites for Saccharomyces cerevisiae. *BMC bioinformatics*, **7**, 113.
29. 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.
30. Dai, Z. and Dai, X. (2012) Nuclear colocalization of transcription factor target genes strengthens coregulation in yeast. *Nucleic Acids Res*, **40**, 27-36.
31. Witten, D.M. and Noble, W.S. (2012) On the assessment of statistical significance of three-dimensional colocalization of sets of genomic elements. *Nucleic Acids Res*, **40**, 3849-3855.
32. Cohen, B.A., Mitra, R.D., Hughes, J.D. and Church, G.M. (2000) A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression. *Nat Genet*, **26**, 183-186.
33. 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.
34. Janga, S.C., Collado-Vides, J. and Babu, M.M. (2008) Transcriptional regulation constrains the organization of genes on eukaryotic chromosomes. *Proceedings of the National Academy of Sciences of the United States of America*, **105**, 15761-15766.
35. Tjong, H., Gong, K., Chen, L. and Alber, F. (2012) Physical tethering and volume exclusion determine higher-order genome organization in budding yeast. *Genome research*, **22**, 1295-1305.
36. Nicholas, W.L., Dougherty, E.C. and Hansen, E.L. (1959) Axenic cultivation of *Caenorhabditis briggsae* with chemically undefined supplements; comparative studies with related nematodes. *Ann. NY Acad. Sci.*, **77**, 218-236.
37. Brenner, S. (1974) The genetics of Caenorhabditis elegans. *Genetics*, **77**, 71-94.
38. Hodgkin, J. (1993) Molecular cloning and duplication of the nematode sex-determining gene tra-1. *Genetics*, **133**, 543-560.
39. Garsin, D.A., Villanueva, J.M., Begun, J., Kim, D.H., Sifri, C.D., Calderwood, S.B., Ruvkun, G. and Ausubel, F.M. (2003) Long-lived C. elegans daf-2 mutants are resistant to bacterial pathogens. *Science*, **300**, 1921.
40. Baugh, L.R., Hill, A.A., Slonim, D.K., Brown, E.L. and Hunter, C.P. (2003) Composition and dynamics of the Caenorhabditis elegans early embryonic transcriptome. *Development*, **130**, 889-900.
41. Yanai, I. and Hunter, C.P. (2009) Comparison of diverse developmental transcriptomes reveals that coexpression of gene neighbors is not evolutionarily conserved. *Genome research*, **19**, 2214-2220.
42. Wernersson, R., Juncker, A.S. and Nielsen, H.B. (2007) Probe selection for DNA microarrays using OligoWiz. *Nat Protoc*, **2**, 2677-2691.
43. Consortium. (1998) Genome sequence of the nematode C. elegans: a platform for investigating biology. *Science*, **282**, 2012-2018.
44. Yook, K., Harris, T.W., Bieri, T., Cabunoc, A., Chan, J., Chen, W.J., Davis, P., de la Cruz, N., Duong, A., Fang, R. *et al.* (2011) WormBase 2012: more genomes, more data, new website. *Nucleic Acids Res*, **40**, D735-D741.

45. Bobach, T.A. (2008), TU Kaiserslautern.
46. Bobach, T., Farin, G., Hansford, D. and Umlauf, G. (2009) Natural Neighbor Extrapolation Using Ghost Points. *Comput. Aided Des.*, **41**, 350–365.
47. Seber, G.A.F. (2004) *Multivariate Observations*. Wiley.
48. Carlson, R.E. and Fritsch, F.N. (1985) Monotone Piecewise Bicubic Interpolation. *SIAM Journal on Numerical Analysis*, **22**, 386-400.
49. Eden, E., Lipson, D., Yogev, S. and Yakhini, Z. (2007) Discovering motifs in ranked lists of DNA sequences. *PLoS Comput Biol*, **3**, e39.
50. Eden, E., Navon, R., Steinfeld, I., Lipson, D. and Yakhini, Z. (2009) GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC bioinformatics*, **10**, 48.
51. Ben-Dor, A., Shamir, R. and Yakhini, Z. (1999) Clustering gene expression patterns. *Journal of computational biology : a journal of computational molecular cell biology*, **6**, 281-297.
52. Goodall, C. (1991) Procrustes Methods in the Statistical-Analysis of Shape. *J Roy Stat Soc B Met*, **53**, 285-339.
53. Davidson, E.H. (2006) *The regulatory genome : gene regulatory networks in development and evolution*. Elsevier/Academic Press, Amsterdam ; Boston.
54. Shen-Orr, S.S., Pilpel, Y. and Hunter, C.P. (2010) Composition and regulation of maternal and zygotic transcriptomes reflects species-specific reproductive mode. *Genome Biol*, **11**, R58.
55. Grishkevich, V., Hashimshony, T. and Yanai, I. (2011) Core promoter T-blocks correlate with gene expression levels in C. elegans. *Genome research*, **21**, 707-717.
56. Wray, G.A. (2007) The evolutionary significance of cis-regulatory mutations. *Nature reviews. Genetics*, **8**, 206-216.
57. Spike, C.A., Bader, J., Reinke, V. and Strome, S. (2008) DEPS-1 promotes P-granule assembly and RNA interference in C. elegans germ cells. *Development*, **135**, 983-993.
58. Kruskal, J.B. (1964) Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, **29**, 1-27.
59. Straussman, R., Nejman, D., Roberts, D., Steinfeld, I., Blum, B., Benvenisty, N., Simon, I., Yakhini, Z. and Cedar, H. (2009) Developmental programming of CpG island methylation profiles in the human genome. *Nature structural & molecular biology*, **16**, 564-571.
60. Leibovich, L., Mandel-Gutfreund, Y. and Yakhini, Z. (2010) A structural-based statistical approach suggests a cooperative activity of PUM1 and miR-410 in human 3'-untranslated regions. *Silence*, **1**, 17.
61. James, N., Landrieux, E. and Collart, M.A. (2007) A SAGA-independent function of SPT3 mediates transcriptional deregulation in a mutant of the Ccr4-not complex in Saccharomyces cerevisiae. *Genetics*, **177**, 123-135.
62. Gerhart, J. and Kirschner, M. (2007) The theory of facilitated variation. *Proceedings of the National Academy of Sciences of the United States of America*, **104 Suppl 1**, 8582-8589.
63. Khaitovich, P., Weiss, G., Lachmann, M., Hellmann, I., Enard, W., Muetzel, B., Wirkner, U., Ansorge, W. and Paabo, S. (2004) A neutral model of transcriptome evolution. *PLoS Biol*, **2**, E132.
64. Feder, M.E. and Walser, J.C. (2005) The biological limitations of transcriptomics in elucidating stress and stress responses. *J Evol Biol*, **18**, 901-910.

65. Yaffe, E. and Tanay, A. (2011) Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nat Genet*, **43**, 1059-1065.
66. Rousseau, M., Fraser, J., Ferraiuolo, M.A., Dostie, J. and Blanchette, M. (2011) Three-dimensional modeling of chromatin structure from interaction frequency data using Markov chain Monte Carlo sampling. *BMC bioinformatics*, **12**, 414.
67. Boissonnat, J. and Cazals, F. (2000) Smooth Surface Reconstruction via Natural Neighbour Interpolation of Distance Functions. *Proceedings of the Sixteenth Annual Symposium on Computational Geometry*, 223–232.
68. de Laat, W. and Grosveld, F. (2003) Spatial organization of gene expression: the active chromatin hub. *Chromosome Res*, **11**, 447-459.
69. Ferrai, C., de Castro, I.J., Lavitas, L., Chotalia, M. and Pombo, A. (2010) Gene positioning. *Cold Spring Harb Perspect Biol*, **2**, a000588.
70. Palstra, R.J. (2009) Close encounters of the 3C kind: long-range chromatin interactions and transcriptional regulation. *Brief Funct Genomic Proteomic*, **8**, 297-309.
71. Steinfeld, I., Shamir, R. and Kupiec, M. (2007) A genome-wide analysis in Saccharomyces cerevisiae demonstrates the influence of chromatin modifiers on transcription. *Nat Genet*, **39**, 303-309.
72. Taddei, A., Schober, H. and Gasser, S.M. The budding yeast nucleus. *Cold Spring Harb Perspect Biol*, **2**, a000612.

Supplementary material

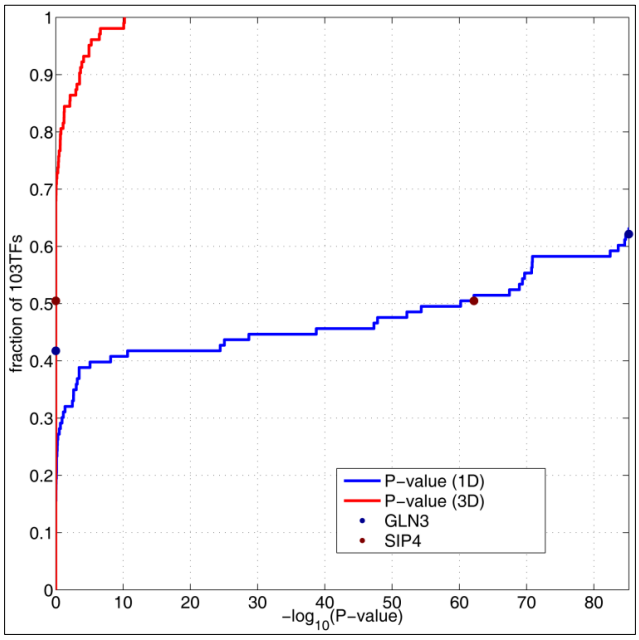

Figure S1 Effect of cyclic permutation on spatial and genomic co-localization enrichments.
Cyclically permuting the assigned annotations of TF gene targets of each chromosome conserves the relative genomic order of the annotations but disrupts the spatial order. This is evident by the ratio shift of the number of significant loci of 3D TF target enrichment vs. genomic TF target enrichment.

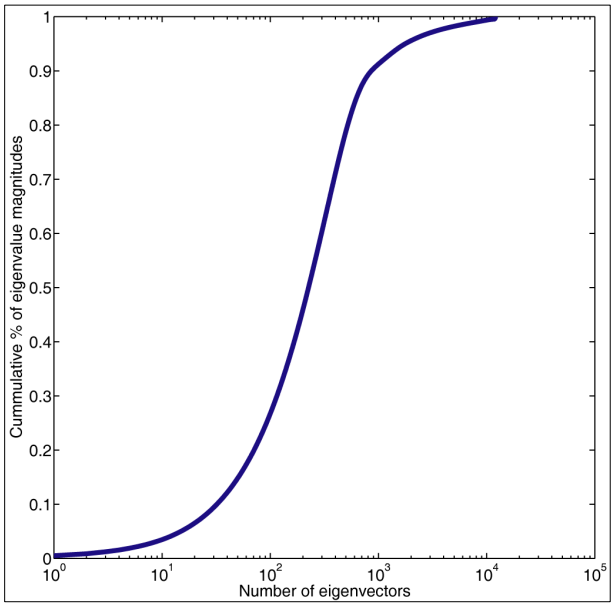

Figure S2 Cumulative sum of the eigenvalues associated with the linear embedding.
Eigenvalue magnitudes of the linear multidimensional scaling showing the underlying doubling dimension to be in the thousands, and not three as expected from a distance measurements of 3D shape.

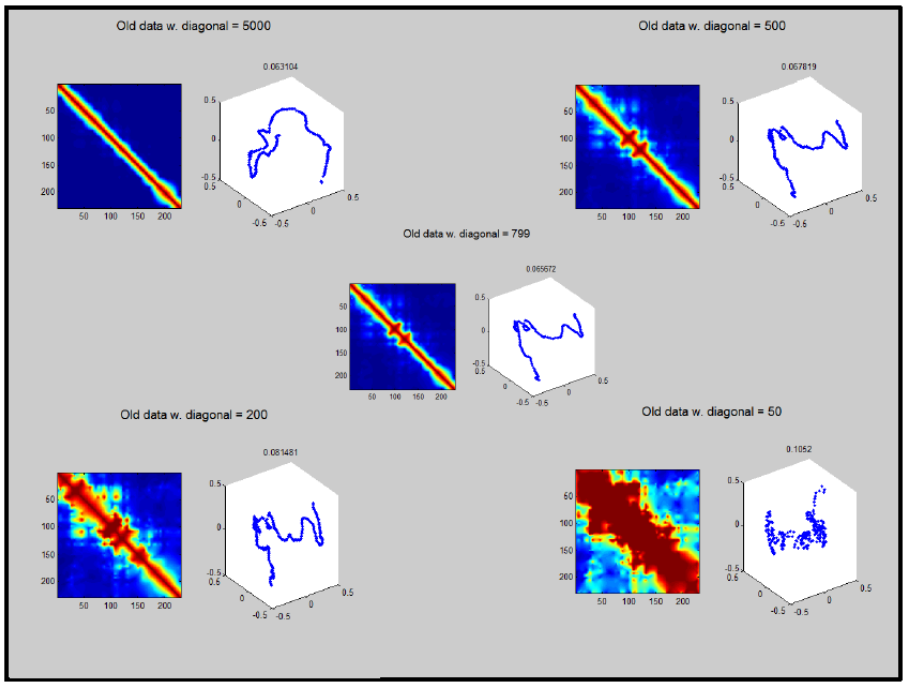

Figure S3 Effect of diagonal forcing on embedding.
Different non-linear embeddings of chromosome I show a decrease of ordinality conservation (Kruskal stress-1) with lower diagonal values prior to interpolation. Max frequency is selected as to not "drown out" local signals along the chromatin.

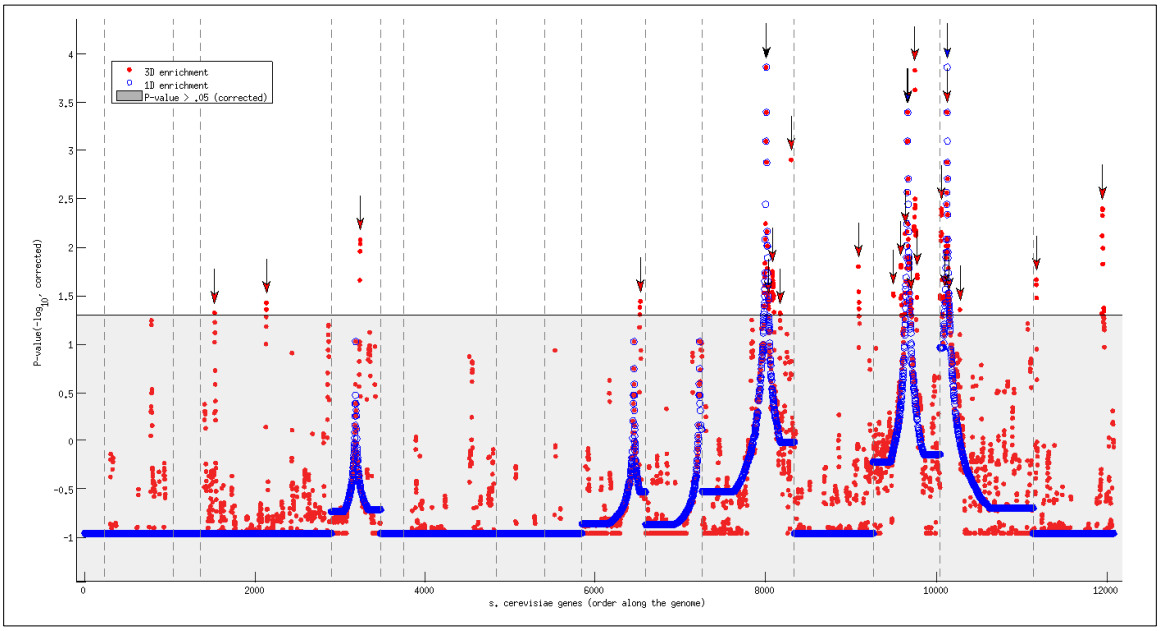

Figure S4 Enrichment landscape for *SIP-4*.
As one of the more extreme examples of 3D enrichment over 1D detected by the peak calling procedure, *SIP-4* shows a diverse pattern of enrichment.

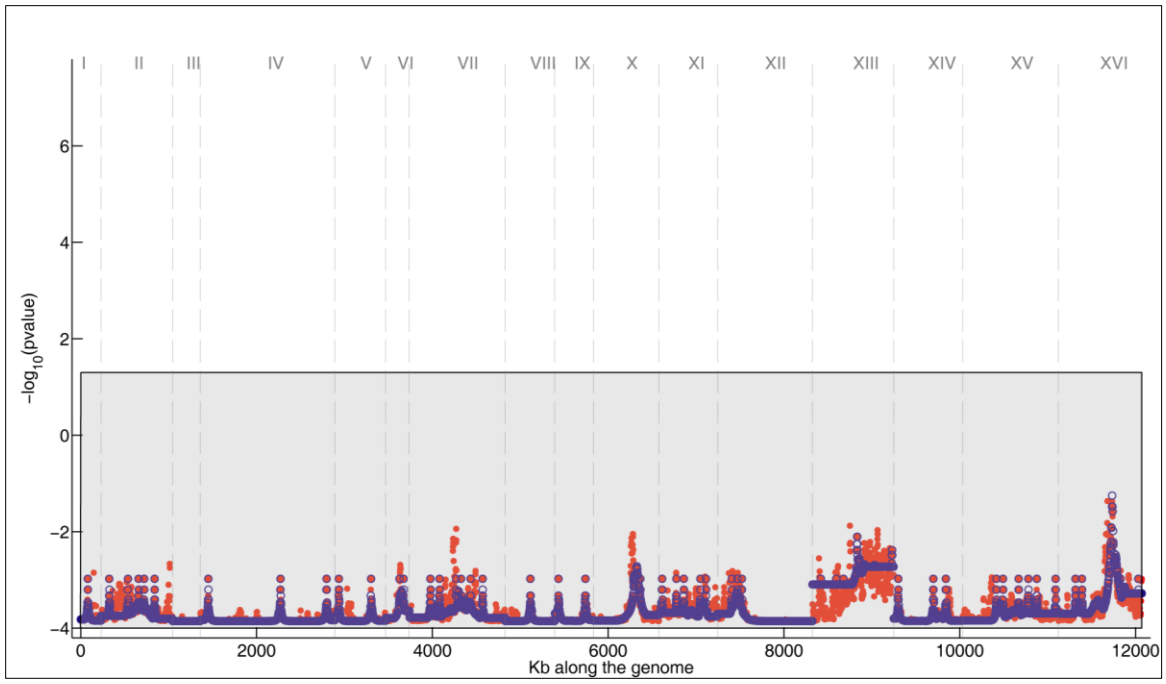

Figure S5 The effect of a permutation on gene identities to the enrichment of co-localized targets of *GLN-3*.
As is evident in this figure, there are no significant enrichments once running a permutation on the gene identities, indicating that the enrichment of gene co-localization is statistically significant and stems from non-random proximity.

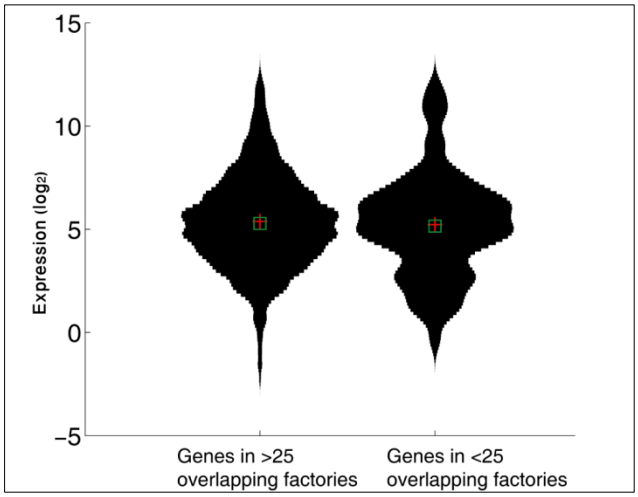

Figure S6 Expression of genes which participate in many co-localized regions compared to genes which participate in few co-localized regions.
The expression for the group of genes which participate in 25 or more regions with significant co-localization is higher than genes which participate in less than 25 regions.

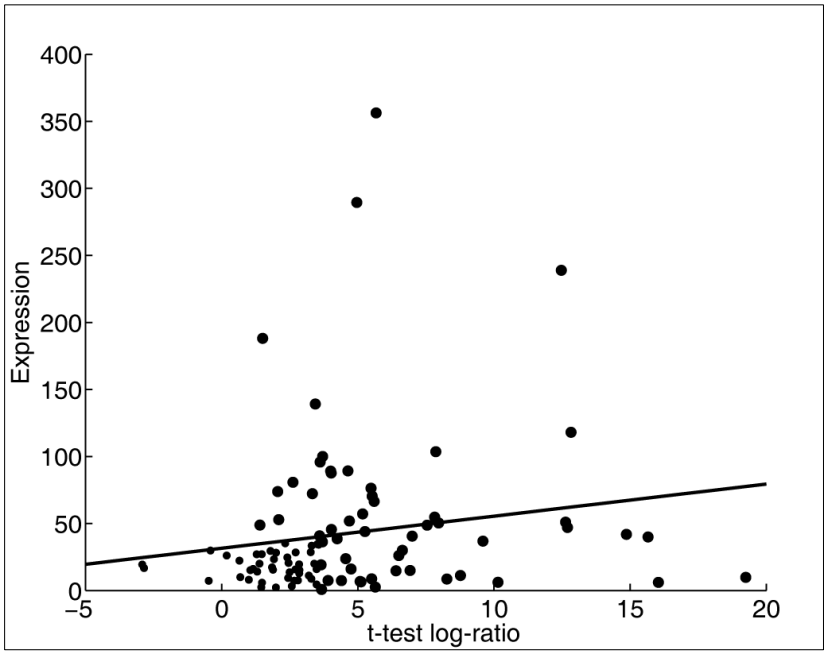

Figure S7 Correlation between expression and spatial organization of TF targets.
TFs with peaks that have a significant 3D over 1D enrichment of their targets (one sided paired *t*-test log-ratio, see Methods) also tend to have higher expression. Regression line of a positive correlation (r=0.25, P<10-2).

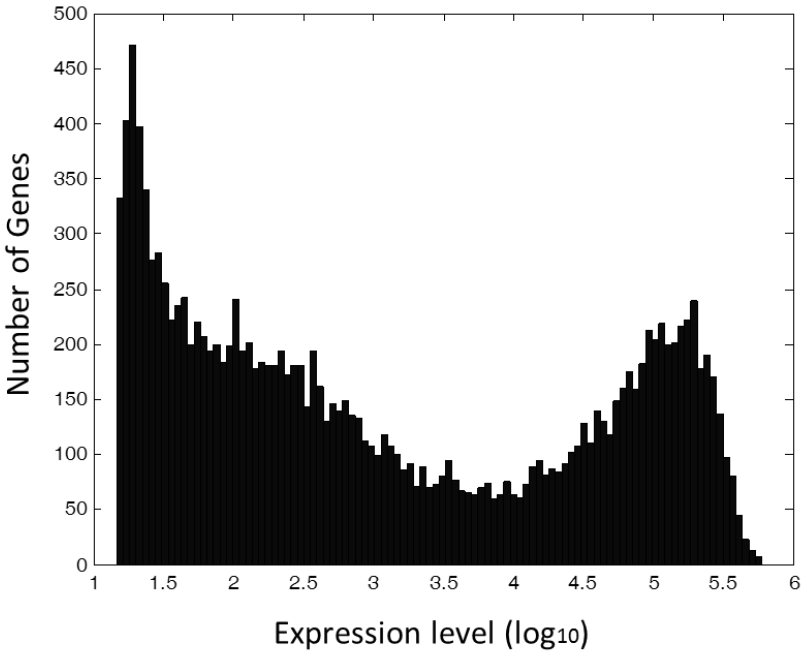

Figure S8 Distribution of expression levels at the four-cell stage of N2 under control conditions.

The four cell stage transcriptome is bimodal in its distribution of expression levels as previously described in Grishkevich et al. Genome Research 2011.

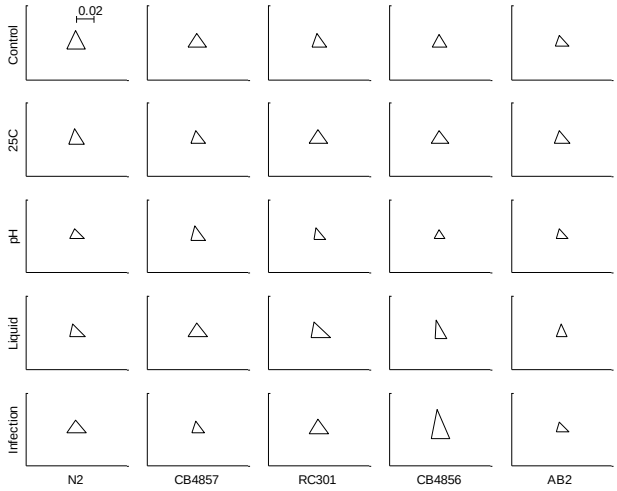

Figure S9 Reproducability of the microarray as estimated by replicates.
Comparing the 25 triplicate arrays for each strain/environment combination, we found a mean correlation of R=0.979 between replicates, highlighting the reproducibility of the dataset. The figure shows the correlations between each pair of the triplicates for each of the environments / strains combinations. The length indicates the expression divergence (1 – Pearson's correlation coefficient). The scale is shown in the N2 control element of the matrix.

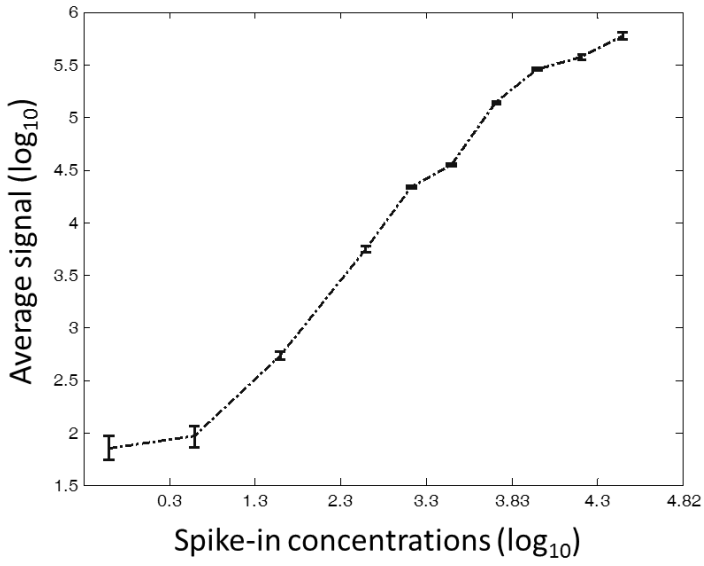

Figure S10 Spike-in control in the microarray data.
RNA from a Spike-in kit (Agilent) that included ten different species of RNA at various known concentrations was added to the total RNA from each sample. For each spiked-in RNA, the microarray contained 30 independent identical probes to measure the concentration. The figure shows for a representative microarray, the mean and standard deviation of the measured concentrations (across the 30 probes) relative to the known concentrations (x-axis). A linear range is observed between 2 and 5 log10 units. 89% of the genes probed on the microarray are within this range, and thus our data faithfully recovers linear increases in their expression levels.

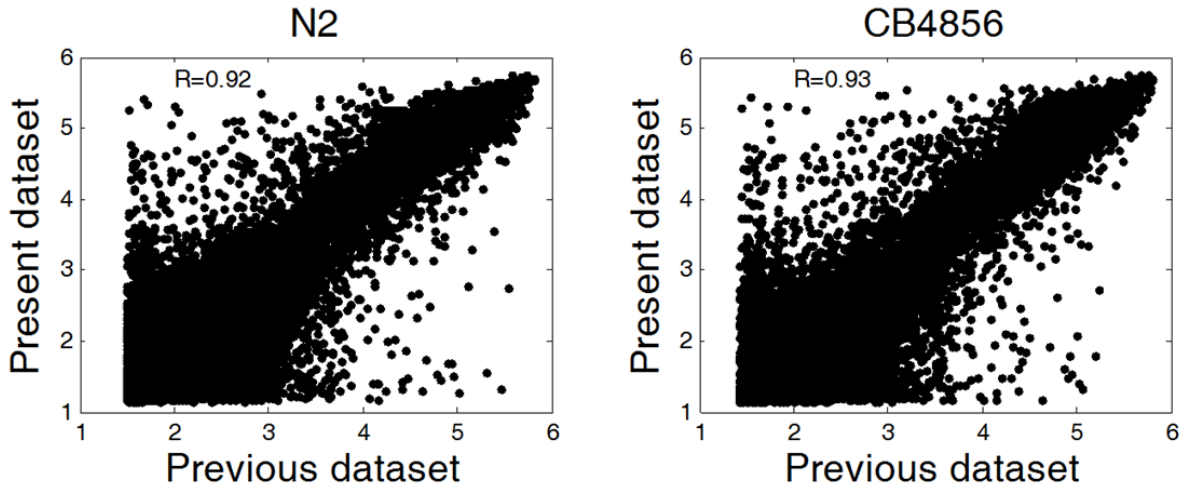

Figure S11 Comparison across datasets.
Two of the examined genotypes (N2 and CB4856) were previously assayed in a similar condition using microarrays, and the correspondence between the datasets was high (R=0.92) despite using slightly different lab conditions, microarrays design and probes. The control condition was different in terms of the bacterial food B. subtilis (presently) vs. E. coli OP50 (Yanai and Hunter, Genome Research 2011).

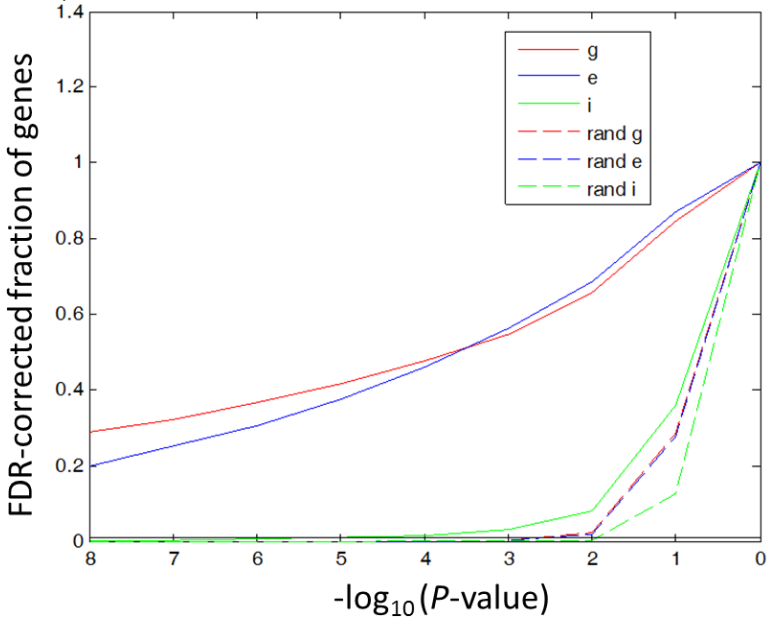

Figure S12 Determination of the ANOVA significance thresholds.
The plots show the cumulative fraction of genes found significant for each P-value cutoff. The randomized sets refer to a permutation of all replicate data (75 columns) per gene. For a false discovery of 0.01 the FDR corrected P-values were $5.1x10^{-4}$, $5x10^{-4}$, and $4.7x10^{-4}$, for the genotypic (g), environmental (e) and interaction (i) gene sets, respectively.

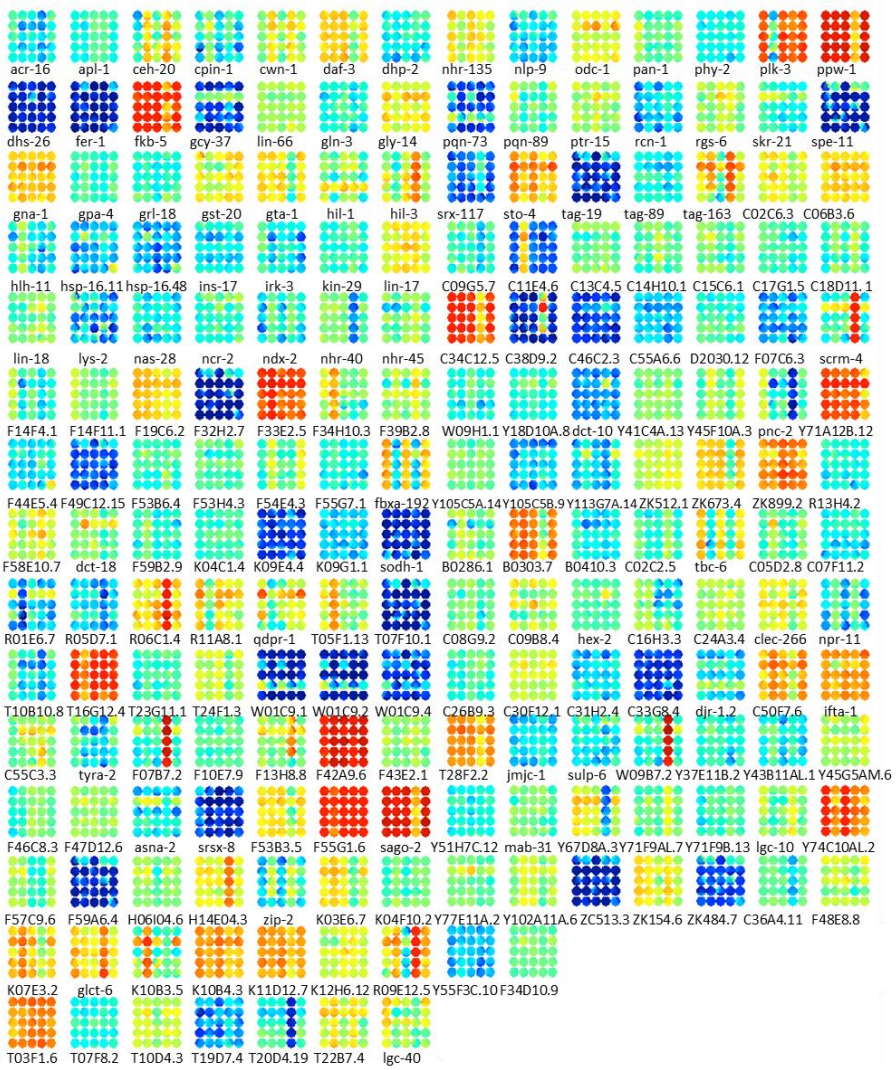

Figure S13 Expression profiles for all 198 identified genes with genotype-environment interactions.
The profiles are in the same format as Figure 1A. We point out the following classes of patterns: Class 1: changes are unique to one (or a few) combinations of the environment and the genotype. Examples of this are F44E5.4, hex-2, spe-11, C38D9.2, dct-10, hsp-16.11, F44F4.1, ifta-1, and W01C9.4. Class 2: For a particular genotype, variation across 9 conditions not found in the other genotypes. Examples of this are: cpin-1, plk-3, fkb-5, F54E4.3, sago-2, C11E4.6, fbxa-192, B0303.7. Class 3: For a particular environmentally induced expression there is particular variation across the genotypes. Examples are odc-1, gcy-37, R01E6.7, Y77E11A.2, asna-2. Class 4: There is variation across both the genotype and environment dimensions and in the intersection there is a non-additive change. Examples of this are scrm-4 (as in Fig. 1A), F07B7.2, and K10B3.5.

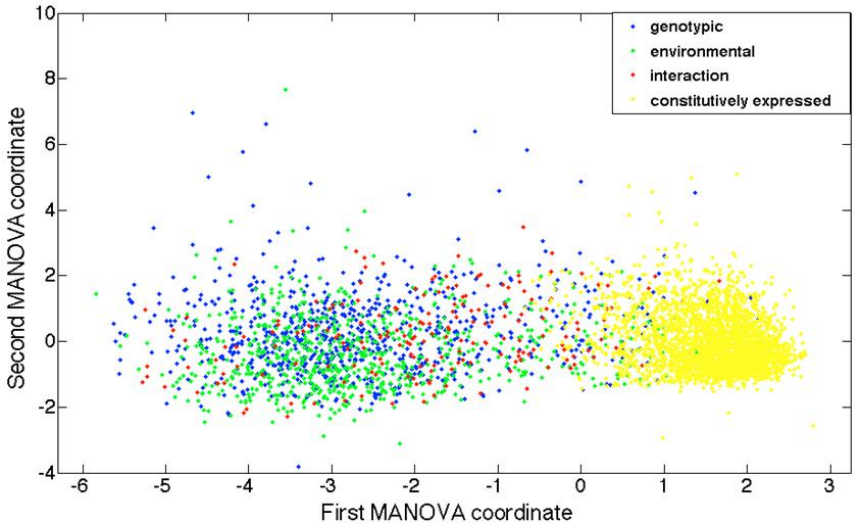

Figure S14 Multivariate analysis of the four gene sets.
MANOVA was invoked on the matrix with parameters for the intergenic distance, expression levels, number of SNPs and chromosomal location. CISRED motifs were excluded from the analysis as they are only available for a smaller subset of the genes; when included the results are similar. The figure shows a clear separation of the constitutively expressed genes and the others, as well as the interaction genes from the others. The distribution is similar to that found for expression levels and for intergenic regions (Figure 2).

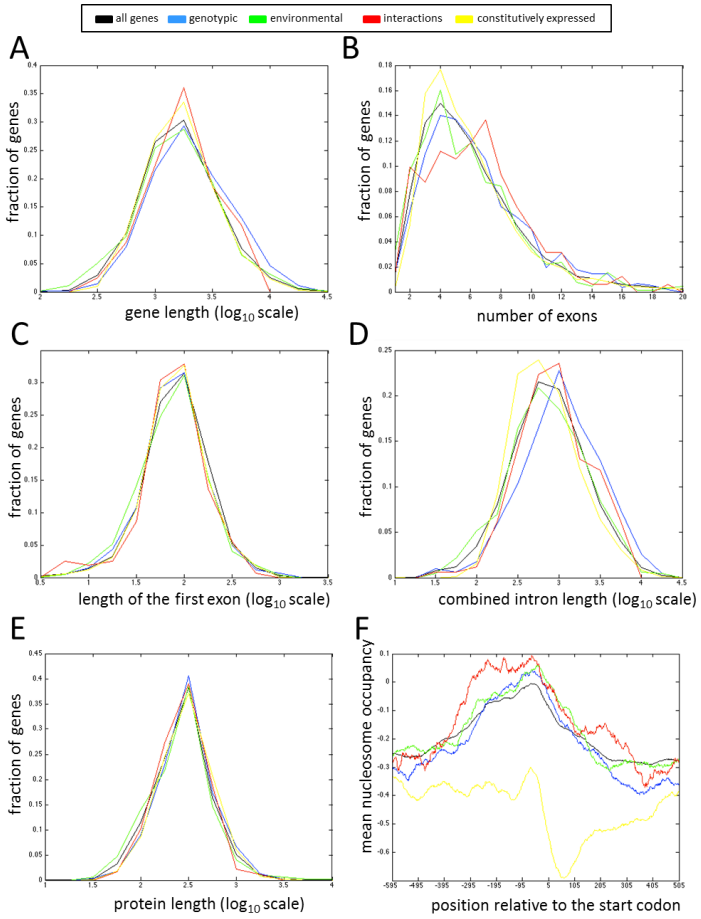

Figure S15 Genomic properties of genes with genotype-environment interactions.
Genes with genotype-environment interaction are not significantly different from the background in their gene length (A), number of exons (B), length of the first exon (C), combined intron length (D) and protein length (E), but have increased

nucleosome occupancy at their promoters (F) ($P<10^{\wedge-30}$, KS-test relative to all genes). Nucleosome data was obtained from Valouev, A., et al. (2008). Genome Research 18: 1051-1063.

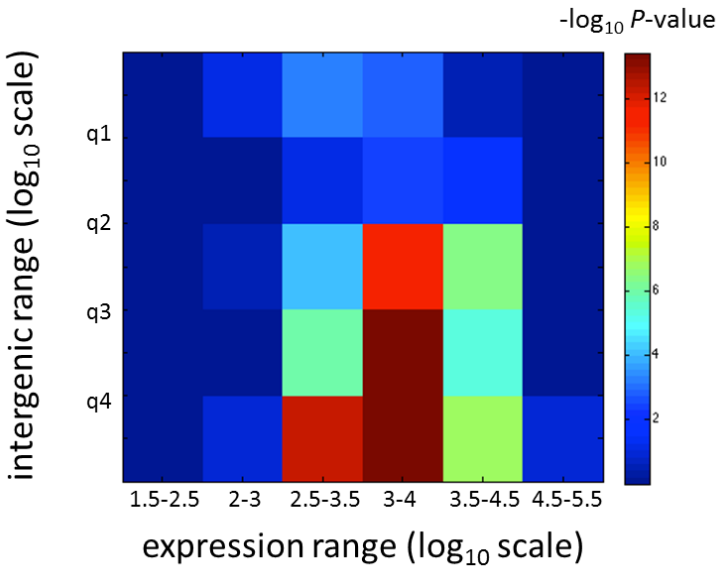

Figure S16 Genes with long intergenic distances and mid-range expression levels are enriched for genotype-environment interactions.
We divided genes into five populated bins with borders 0, 472, 928, 1,818, 4,214, and the maximum distance 57kb. We also binned genes according to their expression with steps of 0.5 log10 expression level units. For each set of genes with a particular combination of intergenic distance bin and expression level bin we compute the enrichment with genes with genotype-environment interactions. This is indicated in the graph according to the $-\log_{10}$ of the P-value of the enrichment as computed using the cumulative hypergeometric distribution.

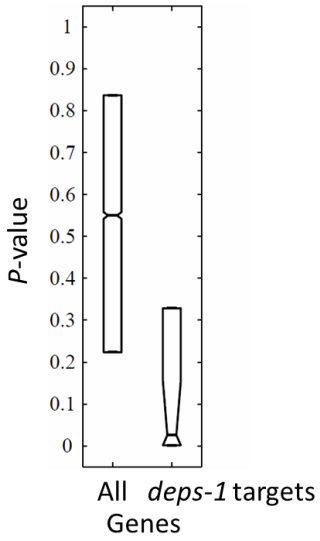

Figure S17 Interaction genes in a functional set.
Two-way ANOVA interaction P-values for the 16 deps-1 targets (Spike et al. 2008 Development, 135, 983-993). The expression data for ten of these is shown in Figure 1B.

Table S1 Sensitivity analysis to the threshold of expression difference used to define the filtered gene set. We selected 0.5 as the threshold for required expression changes but the results hold for all other thresholds.

Threshold value	*Quantile*	*Intergenic distances (P-val, interaction vs all, KS-test)*	*Expression levels (P-val, interaction vs all, KS-test)*
0.1521	0.1	2.37E-08	5.30E-18
0.223	0.2	4.00E-08	2.57E-19
0.313	0.3	1.06E-06	2.27E-20
0.4059	0.4	1.30E-07	8.95E-19
0.5025	0.5	1.80E-08	1.86E-19
0.5999	0.6	2.29E-07	1.07E-17
0.7144	0.7	7.41E-08	2.60E-16
0.8675	0.8	6.88E-08	1.08E-13
1.1103	0.9	3.43E-06	2.76E-07

Table S2 Pairwise correlations among the five examined genomic properties. See Table S3 regarding chromosomal position.

R (P)	*cisred*	*snps*	*Expression*	*Position in the chromosome*
Intergenic distances	*R*=0.2749 (*P*<3.07E-16)	*R*=-0.0681 (*P*<1.49E-05)	*R*=-0.0481 (*P*<0.0023)	*R*=0.0895 (*P*<1.24E-08)
cisred		*R*=-0.0639 (*P*<0.0625	*R*=-0.0471 (*P*<0.1695)	*R*=-0.0125 (*P*<0.716)
snps			*R*=-0.0056 (*P*<0.724)	*R*=0.2621 (*P*<2.40E-64)
Expression				*R*=0.064 (*P*<4.73E-05)

Table S3 Chromosomal location biases.
Each of the four gene sets was examined for enrichment/depletion within the chromosomal centers, using hypergeometric distribution. We found that interaction genes are not biased towards chromosomal centers. Genotypic genes are enriched (green) towards the gene poor regions suggesting that their increased intergenic sizes (Fig. 2a) may be due to their location in gene sparse chromosomal ends. Environmental and constitutively expressed genes are overrepresented in gene-dense chromosomal centers (red) reflecting their shorter intergenic regions. Definition of chromosomal partitions are as previously determined by Rockman and Kruglyak (PLoS Genetics, 2009).

	P-value
Genotypic	10E-100
Environmental	6E-8
Genotype-environment interaction	0.3892
Constitutively expressed	10E-100

Table S4 Functional gene sets with enrichment for interaction genes.
Gene sets were delineated using common GO terms, PFAM domains, and Wormbase expression clusters. For each gene set, we asked whether the genes had a coherent expression profile in our dataset as computed by comparing the distribution of their pairwise correlation coefficients relative to that of randomly paired genes. Enrichment was estimated by a Kolmogorov-Smirnov test between these two distributions. Enrichment for genes with interactions was computed using the hypergeometric distribution. Gene sets with enrichment for both coherent expression (P<10-4) and interaction genes (P<10-2) among the members of the set are shown in the table. See Table S8 for the same for the analogous table the genotypic and environmental gene sets.

Type	Name	Set ID	Enriched Coherent Expression (*P*-val)	Enriched Interactions (*P*-val)	Number of Genes in the Set	Number of genes in the Set with interactions (out of the 198, Fig. 1C)
GO	Nucleosome	786	5.02E-09	0.005518	52	3
GO	Nucleosome assembly	6334	4.97E-09	0.005148	51	3
Pfam	Leucine Rich Repeat	560	2.67E-05	0.001578	37	3
Pfam	Piwi domain	2171	1.13E-05	0.003654	23	2
Exp.Cluster	WBPaper00031477: up *deps-1* vs WT	443	7.04E-05	0.00124	16	2
Exp.Cluster	WBPaper00027111: *rrf-1*(pk1417) upregulated	101	1.32E-09	0.004112	147	6
Exp.Cluster	WBPaper00025032: Cluster 24	172	1.15E-07	0.001924	39	3
Exp.Cluster	WBPaper00032165: differentially expressed with age medoid 1	207	3.05E-17	0.003847	184	7
Exp.Cluster	WBPaper00025032: Cluster 12	218	3.80E-12	0.000601	76	5
Exp.Cluster	cgc4386:cluster 3 6	235	6.13E-06	0.004065	77	4
Exp.Cluster	cgc5767:cluster 5	236	1.24E-25	8.16E-08	119	11
Exp.Cluster	WBPaper00025032: Cluster 2	248	4.79E-93	1.23E-09	168	15
Exp.Cluster	WBPaper00025032: Cluster 17	294	1.65E-07	0.001144	34	3
Exp.Cluster	cgc5767: Expression class SE pi (122 min)	320	2.61E-08	7.31E-05	75	6
Exp.Cluster	cgc5767: Expression class SE pi (66 min)	335	1.59E-06	0.003544	46	3
Exp.Cluster	WBPaper00025032: PAL-1 target genes	92	9.85E-45	0.000139	141	8

Table S5 12 genes with significant genotype (N2 vs. sid-1/haf-6) -environment interactions (P<0.005, two-way ANOVA).

Gene	Description
hil-1	C. elegans HIL-1 protein; contains similarity to Pfam domain PF00538 linker histone H1 and H5 family contains similarity to Interpro domains IPR011991 (Winged helix-turn-helix transcription repressor DNA-binding), IPR005819 (Histone H5), IPR005818 (Histone H1/H5)
shw-3	C. elegans SHW-3 protein; contains similarity to Pfam domains PF02214 (K+ channel tetramerisation domain) , PF07885 (Ion channel) , PF00520 (Ion transport protein) contains similarity to Interpro domains IPR005821 (Ion transport), IPR000210 (BTB/POZ-like), IPR011333 (BTB/POZ fold), IPR003971 (Potassium channel, voltage dependent, Kv9), IPR003968 (Potassium channel, voltage dependent, Kv), IPR003131 (Potassium channel, voltage dependent, Kv, tetramerisation), IPR003091 (Voltage-dependent potassium channel), IPR013099 (Ion transport 2), IPR003974 (Potassium channel, voltage dependent, Kv3)
C09G5.7	contains similarity to Homo sapiens Isoform 1 of Neurofilament heavy polypeptide; ENSEMBL:ENSP00000311997
C17G1.5	
T10B10.8	contains similarity to Pfam domains PF01501 (Glycosyl transferase family 8) , PF11051 (Mannosyltransferase putative) contains similarity to Interpro domains IPR022751 (Alpha-mannosyltransferase), IPR002495 (Glycosyl transferase, family 8)
Y105C5A.14	
djr-1.2	C. elegans DJR-1.2 protein; contains similarity to Pfam domain PF01965 DJ-1/PfpI family contains similarity to Interpro domains IPR006287 (DJ-1), IPR002818 (ThiJ/PfpI)
klf-1	C. elegans KLF-1 protein; contains similarity to Interpro domains IPR015880 (Zinc finger, C2H2-like), IPR013087 (Zinc finger, C2H2-type/integrase, DNA-binding), IPR007087 (Zinc finger, C2H2-type).
arl-13	C. elegans ARL-13 protein; contains similarity to Pfam domains PF00025 (ADP-ribosylation factor family) , PF00503 (G-protein alpha subunit) contains similarity to Interpro domains IPR006687 (Small GTPase SAR1-type), IPR001019 (Guanine nucleotide binding protein (G-protein), alpha subunit), IPR006688 (ADP-ribosylation factor), IPR006689 (ARF/SAR superfamily)
Y44E3A.1	
ZK1055.2	contains similarity to Pfam domain PF03314 Protein of unknown function, DUF273 contains similarity to Interpro domain IPR004988 (Protein of unknown function DUF273)
C36A4.11	

Table S6 Functional gene sets with enrichment for the gene sets.
Same format as Table S4 for the genotypic and environmental groups, as well as the group of genes that are both genotypic and environmental but lacking interaction.

Name	Type	Set ID	Enriched Coherent Expression (*P*-val)	Enriched Interactions (*P*-val)	Number of Genes in the Set	Number of genes in the Set with members from the Group
Group: Genotypic genes						
ligand-dependent nuclear receptor activity	GO	4879	9.04E-32	0.002783486	199	20
axon	GO	30424	5.89E-07	0.007455402	53	7
Ligand-binding domain of nuclear hormone receptor	Pfam	104	2.30E-32	0.007419624	176	17
Metallo-beta-lactamase superfamily	Pfam	753	6.39E-05	0.000270229	12	4
cgc4489_group_23	Exp.Cluster	95	5.30E-06	0.000575205	188	21
[cgc5767]:expression_class_SE_pi(83_min)	Exp.Cluster	164	3.47E-18	2.90E-05	101	16
WBPaper00025032:cluster_24	Exp.Cluster	172	1.15E-07	0.001010549	39	7
WBPaper00032165:differentially_expressed_with_age_medoid_1	Exp.Cluster	207	3.05E-17	0.005461796	184	18
WBPaper00025032:cluster_2	Exp.Cluster	248	4.79E-93	2.80E-08	168	28
[cgc5767]:expression_class_SE_pi(122_min)	Exp.Cluster	320	2.61E-08	0.002324185	75	10
[cgc5767]:expression_class_SE_pi(23_min)	Exp.Cluster	323	7.75E-75	0.004285331	81	10
WBPaper00027111:rde-3(r459)_upregulated	Exp.Cluster	361	8.89E-11	0.000643193	165	19
[cgc5767]:cluster_12	Exp.Cluster	363	6.49E-05	0.005305402	40	6
WBPaper00025032:cluster_14	Exp.Cluster	408	1.08E-05	0.001155034	23	5
WBPaper00025032:cluster_9	Exp.Cluster	441	3.82E-08	0.000304503	33	7
WBPaper00031477:Up_deps-1_vs_WT	Exp.Cluster	443	7.04E-05	3.23E-08	16	8
Group: Environmental genes						
structural molecule activity	GO	5198	8.31E-09	0.004613312	146	15
phosphate transport	GO	6817	1.62E-22	0.002211282	161	17
structural constituent of cuticle	GO	42302	8.55E-28	0.000421701	128	16
MSP (Major sperm protein) domain	Pfam	635	5.99E-18	0.002876323	57	8

haloacid dehalogenase-like hydrolase	Pfam	702	6.75E-05	0.008194329	34	5
Histone-like transcription factor (CBF/NF-Y) and archaeal histone	Pfam	808	2.02E-05	0.005117685	14	3
Nematode cuticle collagen N-terminal domain	Pfam	1484	8.13E-31	0.000462307	129	16
WBPaper00035892:KIM5_vs_OP50_Up	Exp.Cluster	58	6.38E-148	0.000568355	86	12
[cgc5976]:class_1	Exp.Cluster	99	3.31E-27	0.000406408	188	21
[cgc5767]:expression_class_ET_max(83_min)	Exp.Cluster	107	2.42E-67	0.006816769	179	17
WBPaper00006465:IV_ethanol_1ate_repression	Exp.Cluster	166	1.14E-05	0.001857926	64	9
WBPaper00028789:PA14_vs_OP50_downregulated_8hr	Exp.Cluster	177	5.32E-17	0.006816769	179	17
[cgc5767]:expression_class_SE_pi(186_min)	Exp.Cluster	191	4.02E-06	0.003928152	49	7
WBPaper00025032:cluster_5	Exp.Cluster	192	4.47E-05	0.009569486	105	11
WBPaper00028789:PA14_vs_OP50_downregulated_4hr	Exp.Cluster	195	1.39E-05	0.000568355	86	12
WBPaper00025099:NI_specific	Exp.Cluster	277	4.78E-06	0.000424978	63	10
WBPaper00027111:eri-1(mg366)_downregulated	Exp.Cluster	291	6.05E-11	5.26E-06	101	17
[cgc5767]:expression_class_SE_pi(66_min)	Exp.Cluster	335	1.59E-06	0.002613293	46	7
[cgc5767]:cluster_12	Exp.Cluster	363	6.49E-05	0.00461134	40	6
cgc4386_cluster_1_1	Exp.Cluster	375	2.17E-55	4.82E-11	100	24
WBPaper00028789:pmk-1_upregulated	Exp.Cluster	442	8.08E-19	0.004594249	61	8
WBPaper00027722:Microbacterium_nematophilum_upregulated	Exp.Cluster	447	4.41E-06	0.002876323	57	8
Group: Genotypic and Environmental (but not Interaction)						
ligand-dependent nuclear receptor activity	GO	4879	9.04E-32	2.93E-08	199	39
7 transmembrane receptor (rhodopsin family)	Pfam	1	1.06E-06	0.006431861	187	24
Ligand-binding domain of nuclear hormone receptor	Pfam	104	2.30E-32	9.04E-08	176	35
Zinc finger, C4 type (two domains)	Pfam	105	8.88E-43	3.90E-10	187	41
WBPaper00032948:MoltOssilate	Exp.Cluster	85	2.35E-17	0.006008072	186	24
WBPaper00025032:PAL-1_target_genes	Exp.Cluster	92	9.85E-45	0.003444899	141	20
WBPaper00027758:Group_IV	Exp.Cluster	93	2.78E-	0.0002781	77	15

			109	16		
[cgc5767]:expression_class_ET_max(83_min)	Exp.Cluster	107	2.42E-67	0.001760634	179	25
[cgc5767]:expression_class_SE_pi(83_min)	Exp.Cluster	164	3.47E-18	4.65E-08	101	25
WBPaper00025032:cluster_24	Exp.Cluster	172	1.15E-07	0.000685172	39	9
WBPaper00031850:mdt-15_RNAi_Down	Exp.Cluster	178	1.66E-05	0.001043688	137	21
WBPaper00025032:cluster_5	Exp.Cluster	192	4.47E-05	1.09E-07	105	25
[cgc5767]:cluster_7	Exp.Cluster	201	4.47E-20	0.001898587	59	11
WBPaper00032165:differentially_expressed_with_age_medoid_1	Exp.Cluster	207	3.05E-17	0.000107111	184	29
'WBPaper00025032:cluster_12	Exp.Cluster	218	3.80E-12	1.87E-05	76	17
[cgc5767]:cluster_5	Exp.Cluster	236	1.24E-25	4.22E-07	119	26
[cgc5767]:cluster_4	Exp.Cluster	242	1.39E-21	0.001840142	152	22
WBPaper00025032:cluster_2	Exp.Cluster	248	4.79E-93	4.85E-05	168	28
[cgc5767]:expression_class_SE_pi(122_min)	Exp.Cluster	320	2.61E-08	0.005448313	75	12
[cgc5767]:expression_class_SE_pi(23_min)	Exp.Cluster	323	7.75E-75	0.000507051	81	15
[cgc5767]:expression_class_SE_pi(66_min)	Exp.Cluster	335	1.59E-06	3.70E-05	46	12
WBPaper00031477:Up_deps-1_vs_WT	Exp.Cluster	443	7.04E-05	0.006326381	16	4
WBPaper00025032:cluster_61	Exp.Cluster	500	8.07E-06	6.81E-05	15	6

המותאמות לעבודה על תא בודד יהיה אפשרי לספק תמונה יותר מדויקת עבור תא בודד לעומת מיצוע על אוכלוסייה של תאים. סטטיסטיקות מתוחכמות יותר יוכלו לזהות באופן רגיש יותר העשרה-הדדית בין מדידות שונות ובכך לחסוך את הצורך בספים, אשר לרוב נקבעים באופן שרירותי, לקביעת סט גנים לבדיקה.

בפרויקט השני אנו נוקטים בגישה פרגמטית ומספקים תשתית אנליטית לחקר מבנה וארגון הכרומטין בגרעין תא, וכלים אלגוריתמיים לזיהוי קו-לוקליזציה של סמנים, כגון גנים, במיקומם המרחבי בגרעין. הנתונים עליהם אנו מסתמכים נובעים מניסוי המבוסס על שיטת 3C – Chromatin conformation capture. ניתוחים קודמים של מידע מניסויים אלו החלו לשפוך אור על מבנה הגנום הממוצע על אוכלוסיית שמרים מסונכרנת. נתונים אלה מספקים יכולת למדוד את הארכיטקטורה הגנומית בתלת-מימד, ובכך לענות על שאלות בסיסיות לגבי האופן שבו מבנה הגנום מקודד את הפנוטיפ. בעבודה זו אנו מפתחים ומפעילים כלים חישוביים אשר מניבים מודל של מבנה הגנום בשמר - *Saccharomyces cerevisae* באופן שמתערב בנתונים הגולמיים באופן מינימלי, ומזהים תבנית של קו-לוקליזציה של מטרות של גורמי שעתוק (Transcription Factors - TFs) באופן משמעותי מעבר לבקרה של הקו-לוקליזציה הידועה והצפויה מבחינת מיקום לאורך הגנום. על מנת לאפיין תבניות אלו ביתר דיוק אנו מבחינים כי גורמי השעתוק שמטרותיהם נוטים להתקבץ במרחב, בעצמם נוטים להיות בעלי רמות ביטוי גבוהות יותר, דבר המרמז על פעילות קונקרטית של אותם גורמי שעתוק והמטרות שלהם תחת תנאי סביבה דומים לאלה שבהם ניסוי דגימת המבנה בוצע. יתר על כן, הבחנה זו מרמזת על לוקליזציה מרחבית של גנים פעילים, מה שמספק ראיות לנכונותו של מודל מפעלי השעתוק (Transcription factories). שיטה זו חדשנית במספר מישורים – 1. השיטה מסתמכת על העשרה ולא על קורלציה ולכן רגישה יותר משיטות קודמות. 2. השיטה מונעת על ידי הנתונים ומצמצמת הטיות אשר מופיעות באופן נרחב במודלים רבים המסתמכים על הנחות קודמות. 3. השיטה משתמשת בנתוני ייחוס מתאימים אשר מבדילים באופן מדויק אפקטים הנובעים מתופעה ידועה אשר עלולים להטות את התוצאות. בחלק נוסף בעבודה אנו מתארים שיטה חישובית אשר מטרתה לחקור את מרחב הקונפורמציות והמבנים בגנום במטרה לקבץ מבנים דומים ולאפיין קבוצות אלו. בחלק זה של העבודה פותחו אלגוריתמים ייחודיים לזיהוי ואפיון תבניות עבור תתי-מבנים מרחביים במטרה לזהות ולמדוד דמיון מבני. שיטה זו אמנם לא הניבה תוצאות אשר פורשו מבחינה ביולוגית, אך מהווה כלי גנרי למחקרים מולטי-דיציפלינריים הדורשים שיטות לזיהוי תבניות וכריית נתונים. עם הופעת מבנים גנומיים נוספים באורגניזמים או סביבות שונות, שיטות אלו יהיו חשובות בניתוח והבנת הארגון המרחבי של סמנים בגנום, כגון גנים, אתרי תחילת שכפול ואתרים נוספים ומציאת קשרים ביניהם שבעבר לא היה ניתן עקב המודל הלינארי הגנומי המקובל.

לסיכום, אנו מזהים במחקר זה ראיות חשובות המקשרות בין הפלסטיות הגנומית ברמת המבנה המרחבי שלה והתוכן שלה ובין הביטוי של גנים פעילים. באופן זה אנו מדגישים את החשיבות בהסתכלות על תהליך הרגולציה של הביטוי הגנטי לא כמכונת טיורינג הפועלת על גנום חד-מימדי, אלא כעל אוסף תהליכים סטוכסטיים הפועלים במרחב דינאמי המווסת מבחינת הקונפורמציה שלו בהתאם לצרכי ופעילות התא. אנו מזהים כי חלק מהמידע המאופיין כהשפעות-*trans* הינו למעשה *cis* מבחינה מרחבית, ושהערך המוסף מהבנת שכבות הסיבוכיות הנוספות הללו דורש מחקרים נוספים. אנו מאמינים כי השיטות אשר דיסרטציה זו כוללת יספקו כר פורה למחקרים נוספים בתחום זה כאשר נתונים דומים יצוצו עם הקדמה הטכנולוגית ומהפכת הריצוף. כיוונים עתידיים בוודאי יכללו ניתוח קו-לוקליזציה לאלמנטים נוספים מלבד מטרות של גורמי שעתוק לזיהוי מחיצות פונקציונליות בגרעין. בנוסף, בוודאי יהיה מעניין לחקור כיצד משתנה מבנה הגנום בין אורגניזמים שונים או רקמות שונות, או אפילו תגובות לסביבות שונות. עם הופעת שיטות

תקציר

דיסרטציה זו טומנת בחובה שני פרויקטי מחקר נפרדים בעלי מטרה יחידה משותפת – חקר המנגנונים האחראיים לרגולציה גנטית. בביולוגיה, רגולציה גנטית הינו תחום מקיף אשר שואף לתאר את האינטראקציות המולקולריות בין גורמים תאיים אשר טווים את המכניזם שמוביל להגברת, או הפחתת, קצב השעתוק של גנים. שעתוק גנים – הידור של הקוד המקודד ב-DNA לכדי מכונה פעילה – החלבון, לבסוף משתף פעולה עם גורמים נוספים לקביעת הפנוטיפ של התא ושל האורגניזם כולו ולכן בעל חשיבות מכרעת בביולוגיה ובנגזרות פרקטיות מרפואה ועד ננו-טכנולוגיה. הקלות היחסית בה ניתן לקבוע את תוכן ה-DNA או ה-RNA בדגימה ביולוגית בעזרת טכנולוגיות הריצוף החדישות אשר צצות בשנים האחרונות ביצעה שינוי רדיקלי בהבנתנו את מנגנוני הרגולציה הגנטיים. מחד גיסא, יכולת זו הובילה לזיהוי שינויים ברמות ביטוי של גנים אורתולוגים, גנים מקבילים מבנית ופונקציונלית או בעלי דמיון רצפי, בזנים או מינים שונים ותחת שינויים בתנאי סביבה. ומאידך גיסא, התצורה הספציפית של הגנום אשר מכתיבה את המצב הרגולטורי של התא, מובהרת באופן הדרגתי בתנאי סביבה שונים ובמיקומים מסוימים בגנום.

מטרת הפרויקט הראשון הייתה לבחון השפעות שונות של סביבות שונות בה האורגניזם גדל על גמישות הביטוי הגנטי של צאצאיו דרך עדשת האבולוציה. בחלק זה במחקר זה אנו בוחנים את השפעת 5 תנאים שונים של מצע הגידול עבור 5 זנים של הנמטודה *C. elegans* אשר הסתעפו זה מזה והתפתחו באופן בלתי תלוי בסביבות גיאוגרפיות שונות. אנו דוגמים את רמות הביטוי של עוברי הנמטודה בשלב ארבע-תאי בו הביטוי הגנטי כולו הינו מתת-אם ותלוי באופן ישיר בתנאי הסביבה עמם התמודדה האם. על ידי שימוש בכלים סטטיסטיים איתנים אנו מראים שגנים בעלי שכבות רגולציה רבות - גנים אשר מאופיינים ב: אזורים אינטר-גנים ארוכים יותר על גבי ה-DNA, ריכוז מוטיבים ריצפיים הקשורים לרגולציה, ורמות ביטוי גבוהות, נוטים להיות גנים בעלי אינטראקציות גנוטיפ-סביבה. אינטראקציית גנוטיפ-סביבה מתרחשת כאשר שינוי של רמת הביטוי של גן מסוים אינה עקבית בשיעורה על פני סביבות שונות או בזנים שונים באופן שאינו אדטיבי, כלומר, ערך הביטוי של זן מסוים בסביבה מסוימת אינו מהווה קומבינציה לינארית של ערכי הביטוי של כל הזנים באותה סביבה וכל הסביבות בזן המסוים הנ"ל. לדוגמה, גן אשר ביטויו עולה בחשיפה לחום ביחס לסביבות האחרות בזן מסוים, אך ביטויו אחיד בשאר הזנים, הינו גן בעל אינטראקציית גנוטיפ-סביבה. מבחינה אינטואיטיבית אינטראקציות אלה מעניינות משום משהן מרמזות על קיומו של מנגנון הייחודי לזן מסוים להתמודדות עם סביבה ספציפית, ומאפשרות מחקר השוואתי למיצוי היסוד המולקולרי להבדל. בעזרת ריצוף חמשת הזנים האלו אנו מראים שגנים בעלי שונות ברמות הביטוי על פני הגנוטיפ מועשרים ב-SNP-ים (פולימורפיזם בנוקלאוטיד בודד) באזור הקודם לרצף המקודד ב-DNA, כמצופה. למרות זאת, גנים בעלי אינטראקציות גנוטיפ-סביבה אינם שונים באופן משמעותי סטטיסטית מבחינת כמות ה-SNP-ים שלהם מאשר שאר הגנים. יחד, תוצאות אלו מרמזות על כך שטיבם של גנים בעלי שכבות רגולציה רבות מוטה לאינטראקציות גנוטיפ-סביבה כתוצאה ישירה של שינויים באזורים המקדמים השכנים לפני הגן על גבי ה-DNA. תצפית זו עלולה לספק הבנות עמוקות יותר לגבי מוצאו של מגוון ביטויי הגנים יוצא הדופן אשר נצפה אפילו במינים קרובים מבחינה גנטית.

המחקר נעשה בהנחיית ד"ר זוהר יכיני בפקולטה למדעי המחשב ובשיתוף ד"ר איתי ינאי בפקולטה לביולוגיה.
אני מודה לטכניון – מכון טכנולוגי לישראל על התמיכה הכספית הנדיבה בהשלתמותי. כמו כן אני מודה ל-Merk על מענק ול-CGC על זנים אשר שימשו אותי במאמר - "A Genomic Bias for Genotype-environment Interactions in *C. elegans*".

שיטות חישוביות לניתוח ביטוי גנים באורגניזמי מודל

חיבור על מחקר

לשם מילוי חלקי של הדרישות לקבלת התואר
מגיסטר למדעים במדעי המחשב

שי בן-אלעזר

הוגש לסנט הטכניון – מכון טכנולוגי לישראל

חשון תשע"ב חיפה אוקטובר 2012