An extensive analysis on the linear gene clustering in *Saccharomyces cerevisiae*

Athanasia Stavropoulou

Supervised by Christoforos Nikolaou
Computational Genomics group

MSc in Bioinformatics

Faculty of Medicine
UNIVERSITY OF CRETE
&
BIOMEDICAL SCIENCES RESEARCH CENTER
"Alexander Fleming"

February, 2022

A dissertation submitted in partial fulfilment of the requirements for the degree of M.Sc. in Bioinformatics.

Acknowledgements

It seems unreal that it has already been two and a half years since I decided to take the Bioinformatics Master's program at the University of Crete. I knew that these years would definitely not be a piece of cake, but now that I am nearing the end of this journey, I know it was worth the effort. Although SARS-CoV-2 made the process harder, I will remember these years with love and joy because of the people I met. The first year was quite intensive, with great teachers introducing us to the vast worlds of programming, linear algebra, biology and genomics. It would have certainly been a lot more difficult, had it not been for my fellow students with whom I built beautiful relationships in such a short period of time. Therefore, I want to thank all these great people for the times we had together and I hope that we will remain in touch in the years to come

In my second year, I and my partner in crime, Emilios Tassios, started our Master's thesis with the Computational Genomics Group in Athens. The change of city, my innate insecurity and the ongoing pandemic made things difficult at first, but knowing that I was in such a caring and friendly environment, somehting that I appreciated more than anything, helped me see things more brightly. Firstly I want to thank Dr. Christoforos Nikolaou, our wonderful mentor and supervisor, for his inspiring ideas in science, for supporting and for trusting us with those ideas, and for all the great conversations that we had, not only about science but also politics, psychology, music, art, history and pretty much every field possible (although I do not know anything about football). Many things could be said about my lab-partner and friend Emilios, a great guy with lots of interests... Pretty much all of us love him. I am thankful to him for all his support and for keeping up with my insecurity and pessimism, while at the same time instilling in me a bit of his own way of life and thinking. Last but not least, I want to thank all the amazing people in our group, Nikos, Michalis, Sofia, Antonis, Kleio, Jo and Gina for their constant support and to say that I feel lucky and grateful to be a part of this team. Finally, I want to thank my family for trusting and supporting me in every step I make, my dear friends and, of course, Stefanos who keeps supporting me from the first day that we met.

Athanasia 06/02/22

Abstract

Studies from the 2000s have reported on the non-randomness of gene distribution in the eukaryotic genome. Numerous studies have pointed out the existence of gene clustering in all major eukaryotic kingdoms, that is thought to enable the co-regulation of functionally-related and often co-expressed genes. In our current work we build on previous results of our lab (Nikolaou (2018); Tsochatzidou et al. (2017)), in order to study the existence of an underlying compartmentalized genomic organization in *saccharomyces cerevisiae*, which correlates the topological to the functional features of the genes. In order to achieve that, we have devised an algorithm that evaluates the existence of gene-clustering at the linear level and can be applied on any genomic categorization. This method works on the basis of a permutation test strategy, which assesses a) the enrichment of genes in specific chromosomes and b) the linear intergenic distances between consecutive input genes. Starting from a subset of genes of a specific type, the algorithm returns a set of delineated at coordinate level sub-clusters, which enclose genes positioned in greater proximity than expected by chance on a chromosome. We have applied this approach on a large variety of publicly available genomic categories, which include transcription factor gene-targets, gene ontology terms and other partitions related to evolutionary age, conservation level and transcriptional plasticity. We have detected clustering in almost every category that we examined. Seemingly, transcriptional regulation, expression variability and conservation level constrain the organization of the genes at the linear level. We have also found rare but interesting cases of sparse gene-positioning, regarding genes of a younger evolutionary age and genes considered integral components of the membrane. In a novel approach, our pipeline allows for the association of genomic properties through the assessment of overlapping sub-clusters. We have identified correlated patterns between clusters of genes with low conservation and high transcriptional plasticity, while also finding that clusters related to the membranes overlap with such regions, implying a positional bias of such functions towards the telomeres.

Contents

List of Figures

List of Tables

List of Abbreviations

TFs Transcription factors
WGD Whole genome duplicate
SSD Small scale duplicate
PFM Position frequency matrix
PSSM Position specific scoring matrix
SGD *Saccharomyces* Genome Database
SPB Spindle pole body
NE Nuclear envelope

1

Introduction

1.1 | *Saccharomyces cerevisiae*

1.1.1 | *Saccharomyces cerevisiae* as a model organism

Budding yeast *Saccharomyces cerevisiae* is an eukaryotic, single-cellular organism that is widely used as an experimental system. Its rapid growth and low living costs make its manipulation easier compared to the other animal models. The budding yeast genome was the first eukaryotic genome to be fully sequenced in 1996 (Goffeau et al. (1996)), aiding in the development of various functional genomic tools. Its genome consists of approximately 12Mb pairs and more than 6000 protein-coding genes (Goffeau et al. (1996)). One of yeast's major contributions in research is the gene-protein functional association through the construction of mutants (Botstein and Fink (2011)) leading to the creation of various kinds of libraries, e.g. deletion libraries (Scherens and Goffeau (2004)). The progress in studying yeast's genetics and molecular biology served as a motivation to introduce researchers in the development of many high-throughput technologies (Cho et al. (1998); DeRisi et al. (1997); Lashkari et al. (1997)) leading to the expansion of the yeast's and other model organisms' available genetic-data toolbox. Yeast and humans share a wide range of functional pathways (Kuchaiev and Pržulj (2011)) regarding the cell cycle (Hartwell (2002)), metabolism (Petranovic et al. (2010)) and other major pathways (Chen and Thorner (2007); De Virgilio and Loewith (2006)), making it also a suitable model to study human disorders (Petranovic and Nielsen (2008)).

1.1.2 | *Saccharomyces cerevisiae* nucleus and genome organization

The budding yeast is characterized by both unique and conserved eukaryotic traits. As mentioned before, the *Saccharomyces cerevisiae* genome is approximately 12Mb pairs long. It is or-

ganized in sixteen chromosomes which host approximately 6200 protein-coding genes. The positioning of genes on the chromosomes is quite compact as 70% of the genome is occupied by genes (Goffeau et al. (1996)), interrupted only by a low number of (approximately 250) introns (Barrass and Beggs (2003)). Although more complex eukaryotic genomes consist of repetitive elements, budding yeast chromosomes have little repetitive DNA (apart from the rDNA and telomeres) and potentially no satellite repeat DNA at the centromeres (Taddei and Gasser (2012)).

The basic principles governing the nuclear organization can be observed in all eukaryotes, from yeast to humans. During interphase, the budding yeast chromosomes achieve a very characteristic configuration in the nucleus (Duan et al. (2010)), which is thought to serve in the simplification of chromosomal intermingling (Pouokam et al. (2019)). This conformation is called Rabl and is characterized by a centromeric center adjacent to the spindle pole body (SPB) and extending chromosomal arms with telomeres anchored in the nuclear envelope (NE) (**Figure 1.1**). Transcription factors, nuclear pore proteins and chromatin remodellers play a crucial role in the spatial organization of the yeast genome (Brickner et al. (2019); Jo et al. (2021)). According to the Duan et al. (2010) model, chromosomes occupy distinct regions in the nucleus, with the smaller ones having higher inter-chromosomal contact frequencies, as they are cramped in a smaller part of the nucleus close to the centromeres.

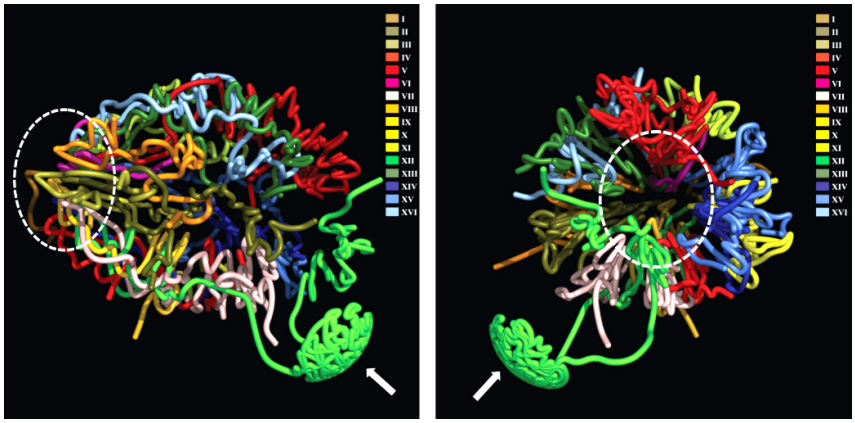
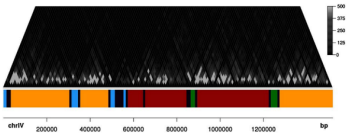
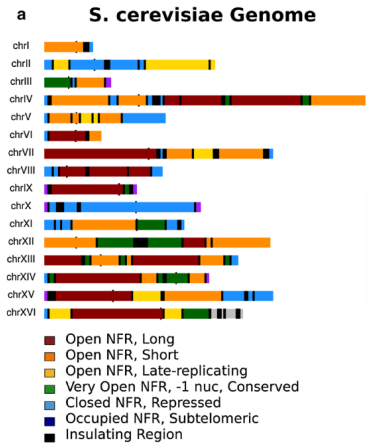


Figure 1.1: Two views representing the yeast genome from two different angles, as provided by Duan et al. (2010). This model was created by transforming the contact frequency of different regions into distance in space. Enclosed in the dashed lines is the centromeric center, while the arrow pinpoints the nucleolus formed by chromosome XII.

We know that the genome of the higher eukaryotes is not randomly positioned inside the nucleus, as there are functional compartments determined by specific enzymes and chromatin states (Lieberman-Aiden et al. (2009); Rao et al. (2014)), **Figure 1.3**). Similarly, a study in yeast revealed the existence of TAD-like domains, each characterized by distinct nucleosomal ar-



(a) TAD-like domains in yeast



(b) The segmentation of the genome on the basis of the TAD-like domains

Figure 1.2: Depicted in 1.2a are the TAD-like domains across the chromosome IV of the yeast as shown in Nikolaou (2018). Depicted at the top of 1.2a is the frequency of contacts in a 3C map using the data of Duan et al. (2010) while, shown at the bottom are the resulting TAD-like domains created on the basis of structural and functional properties. Black represents the insulating regions while the rest of the colours indicate different TAD classes. Shown in 1.2b is the distribution of the different TAD-like domains across the yeast chromosomes.

chitectures, transcriptional regulators and positional preferences, segmenting the genome into seven compartments, as shown in **Figure 1.2** (Nikolaou (2018)). Zooming out, we observe that the nuclear organization of the yeast genome seems to reflect distinct functional processes. The accumulation of the rDNA in the nucleolus opposite to the (Spindle Pole Body) SPB, is a site dedicated to ribosome biogenesis enriched in RNA-polI, serving as a most striking example of functional compartmentalization in the nucleus (Taddei and Gasser (2012)). Furthermore, telomeric regions near the periphery host repetitive DNA, while being unfavorable for RNA-pol II transcription, as they are enriched in silencing factors (Taddei and Gasser (2012)). Genomic compartmentalization seems to exist at multiple levels and it has inspired a number of studies trying to unveil the principles governing the organization of the genomes at both linear and three-dimensional levels.

1.2 | The non-random positioning of genes

1.2.1 | Gene Clustering

In the early 2000s, the science world was already talking about the non-randomness of the genes' distribution across the genome (Hurst et al. (2004)). Now it is known that gene positioning in either the linear or the spatial level affects the genes' regulation and thus their expression (Misteli (2004); Takizawa et al. (2008)). As already mentioned, the genome of higher eukayotes is organized in a non-random manner, segregated into open and closed chromatin and forming genome-wide compartments that occupy distinct regions in the nucleus (Lieberman-Aiden et al. (2009)) (**Figure 1.3**).

At the linear level, multiple studies reported cases of co-expressed genes clustering from yeast to humans. The first evidence came from Cho et al. (1998), who have shown, through mRNA level characterization, that 25% of genes with cell-cycle-dependent expression patterns were directly adjacent to genes induced in the same phase of the cell cycle (**Figure 1.4**). Clustering was also found in other species as well. Approximately 15% of *C. elegans* genes are contained in operon formations, transcribed in polycistronic pre-mRNAs, stretching between two and eight genes long (Blumenthal et al. (2002)). In *Arabidopsis thaliana* it was shown that neighbouring genes are co-expressed (Williams and Bowles (2004)). At a larger scale, in humans, it was shown that the housekeeping genes, highly expressed in a variety of tissues as defined by Serial Analysis of Gene Expression (SAGE) tags, show significantly smaller dispersion than expected by chance (Lercher et al. (2002)). Finally, Boutanaev et al. (2002) claim that 45% of genes expressed solely in the testes are organized in uninterrupted stretches of at least four genes. A looser definition of a cluster by permitting intervening genes of different expression patterns led to the identification of much larger clusters. Although clustering was found across many major species, it seems that there is a correlation between the physical cluster size and the complexity of the organisms, as cluster sizes range from a few kilobases in yeast, characterized by a compact genome, to several megabases in mammals (Hurst et al. (2004)).

Why are genes positionally clustered? One major hypothesis claims that the clustering of genes enables co-regulation, either on the small scale (e.g. bidirectional promoters) or on a broader scale (e.g. chromatin-mediated regulation). Various studies supporting this notion report that functionally related genes, participating in the same GO-terms (Tiirikka et al. (2014)) or in the same KEGG pathways (Lee and Sonnhammer (2003)), are clustered across many organisms. Nevertheless, the latter report that although there is a significant tendency for genes to cluster across all the species examined (human, worm, fly, A. thaliana and yeast), the fraction of pathways with significant chromosomal clustering was highly variable, ranging from 30% for *D. melanogaster* to 98% for yeast. Similarly, in yeast Teichmann and Veitia

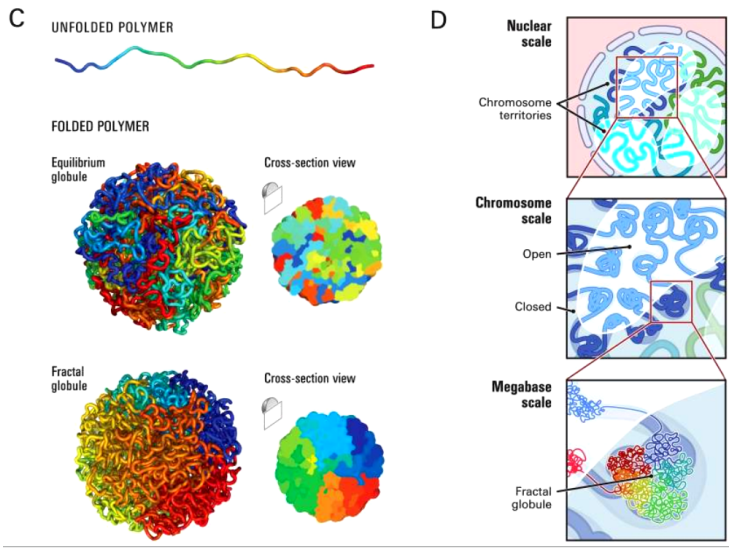


Figure 1.3: 'The chromatin packing in the nucleus is consistent with the behaviour of a fractal globule', as mentioned in Lieberman-Aiden et al. (2009). C) Top: an unfolded polymer chain, 4000 monomers. Depicted below are two folding models of the chromatin. The first refers to the equilibrium globule characterized by a highly entangled structure. The second refers to the fractal globule and is characterized by the formation of distinct monochromatic blocks as regions that are nearby at the linear level are brought together also at the 3D level. D) The genome architecture at three levels. Top: two distinct compartments reflecting the open and close chromatin with chromosomes occupying distinct territories, middle: individual chromosomes, bottom: at the scale of single megabases, the chromosome consists of a series of fractal globules

(2004) showed that genes, whose products participate in stable protein-protein interactions, are found to be strongly linked, which helps in their co-regulation and thus in the maintenance of the right stoichiometry. Additionally, Poyatos and Hurst (2006) have shown that proximal genes in a protein-protein interaction network in yeast tend to be positionally linked and often co-expressed. Janga et al. (2008) showed that most of the yeast transcription factors that they examined have positionally clustered gene-targets, implying that transcriptional regulation constrains the positioning of genes. A different work conducted on the human genome by Thévenin et al. (2014) has shown that functionally related gene groups are concentrated in specific chromosomes, while at the same time being positioned at smaller distances across the chromosomes. Finally, a work based on the 3D yeast model of Duan et al. (2010), claims that there is an enrichment of inter-chromosomal links connecting loci of genes with the same GO-term (Homouz and Kudlicki (2013)). In more complex genomes, increased co-expression has been shown to be a characteristic of specific chromosomal regions associated with Topologically Associated Domains (TADs) (Krefting et al. (2018)).

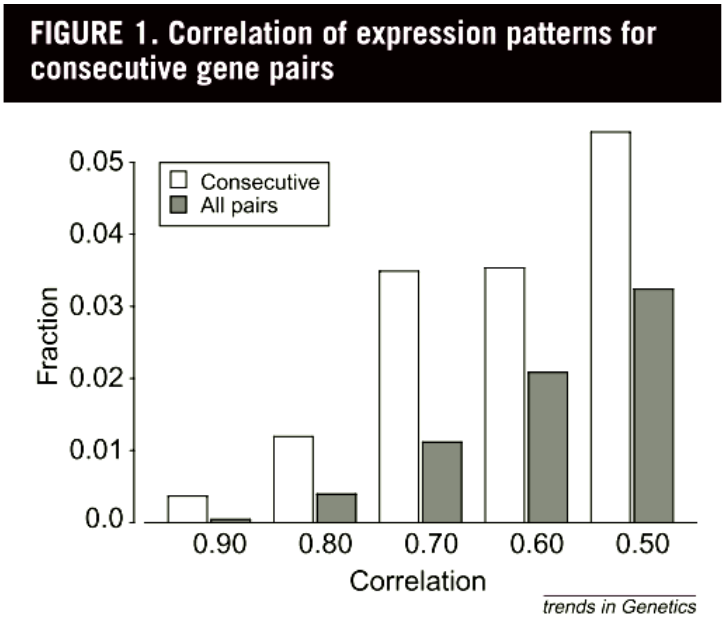
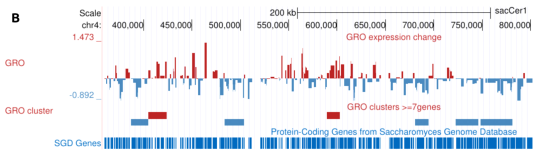
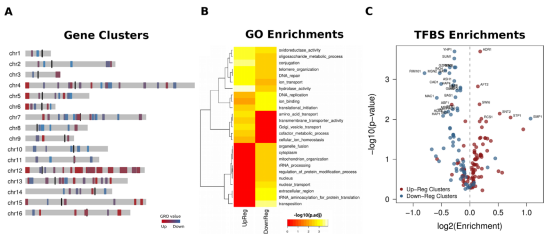


Figure 1.4: Depicted is a histogram, as shown in Kruglyak and Tang (2000), with a fraction of highly correlated expression patterns among consecutive gene-pairs (white) versus all gene pairs (gray). For example, the bars at 0.80 show that 1.2% of consecutive genes have expression patterns with a correlation between 0.80 and 0.90, compared to 0.40% of all gene pairs.

Although there is evidence that clustering is correlated to functionally related genes and co-expression, probably at both the linear and the 3D level, there are other studies supporting that this does not explain the full picture. Different studies in yeast claim that the drive behind the genes' clustering is the reduction of transcriptional noise, as there is evidence that they accumulate in constantly open chromatin "sinks" in which transcriptional bursting is minimized (Batada and Hurst (2007); Wang et al. (2011)). Supporting this notion, Kustatscher et al. (2017) note that the co-expression of functionally unrelated neighbouring genes may be a side effect of the selection for noise reduction. They claim that the genome "compensates" for such a co-expression by buffering the co-expressed genes' products at the protein level. Similarly, studies in the 3D level support that noise-reduction constrains the organization of the yeast genome at the 3D level as well (Singh et al. (2016)).



(a) Up- and Down-regulated gene clusters



(b) Topological and functional preferences of the gene clusters

Figure 1.5: Depicted in 1.5a are the up- and down-regulated gene clusters on chromosome 4, as shown in Tsochatzidou et al. (2017). The clusters were created by merging genes with similar (positive or negative) GRO values. Red indicates up-regulation, while blue indicates down-regulation. Depicted in A) is the distribution of the two clusters across the chromosomes while in B) is a heatmap representing the GO-terms enrichments across the two cluster types. Finally, depicted in C) is a volcano plot showing the significant enrichments of transcription factor binding sites (TFBS) across the two clusters types. Enrichments are shown as log2-based observed/expected ratios. Values >0 indicate enrichment and values <0 indicate depletion. P-values correspond to 1000 permutations for each transcriptional regulator.

1.2.2 | Gene clustering upon topoisomerase perturbation in yeast

The motivation for the current project came from a previous work of our lab, in which we have unveiled the existence of a compartmentalized organization of the yeast genome during stress. Tsochatzidou et al. (2017) worked on a genome-wide transcription run-on (GRO) experiment conducted shortly after the thermal inactivation of topoisomerase II which caused a transient accumulation of topological stress to the cell.

Upon these conditions they detected the emergence of 116 up- and down-regulated concordant gene clusters with more than seven genes each, which have strong positional and functional preferences, as depicted in **Figure 1.5**. These clusters tend to be co-expressed at levels which are higher than the ones expected by chance and are found to have opposing topological preferences, with the up-regulated clusters being positioned farther from the centromeres, while the down-regulated occupy regions near the centromeres. This segregation expands to the functional level as well, with up-regulated genes being mostly enriched in stress-related GO-terms and in more complex regulation patterns than the down-regulated genes, which were found to be mostly related to basic cellular functions characterized by a less complex regulation, depleted of TATA elements and transcription factor binding sites.

By analyzing more properties of gene-clusters, including the directionality of genes, the lengths of intergenic regions, their conservation level and even the frequency of contacts in the three-dimensional space, they have described a segregated genome architecture that resembles an "Urbanization process", as depicted in **Figure 1.6**. The genome near the centromeres,

which are enriched in down-regulated clusters, represents the "old city center", characterized by its ancient, tightly-positioned genes, related to basic cellular functions, while the "suburban genome" near the chromosomal outskirts, enriched in up-regulated clusters, is characterized by sparsely positioned, less conserved genes mostly related to stress response and secondary functions.

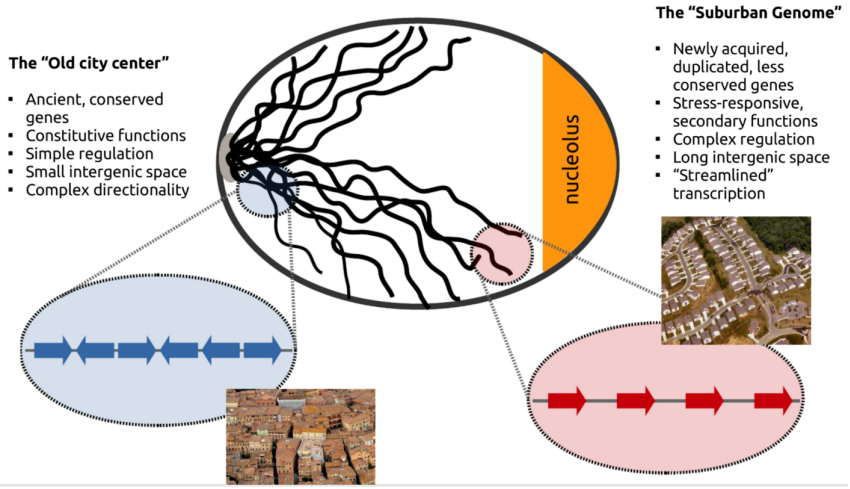


Figure 1.6: The topological and functional preferences of the co-regulated gene clusters reflect a compartmentalized genome architecture, as described in Tsochatzidou et al. (2017). Depicted here is a representation of the Genome Urbanization in the interphase nucleus acquiring the Rabl conformation. The centromeric 'old city center', enriched in down-regulated gene clusters under topoII deactivation, can be compared to a medieval town with crammed houses and narrow intertwined alleys as it consists of more conserved genes of complex directionality with tighter intergenic spaces. Telomeric regions resemble the 'suburban lanscape' of a modern city characterized by "younger" genes, sparsely positioned in a co-directional operon-like manner, enriched in the up-regulated clusters.

The emergence of co-regulated gene clusters which have distinct functional roles and occupy distinct genomic 'niches', implies the existence of a spatial preferences that may shape the organization of gene neighbourhoods, in such a way to enable the optimization of the cell's balance between homeostasis and stress response. Based on that specific idea, we currently wish to explore if such a notion applies widely to the *Saccharomyces cerevisiae* genome by studying the clustering of various gene categories and their overlaps across the chromosomes at a linear level.

1.3 | Aims and Objectives

Building on previous work of our lab (Tsochatzidou et al. (2017)), we currently study the topological and functional genomic compartmentalization of the *Saccharomyces cerevisiae* genome at the linear chromosomal level. Our main objective was to create an in-house algorithm that could be easily applied on any genomic categorization, aiming to identify cases of positional clustering across the yeast's genome and finally to delineate such regions at coordinate level. By applying this algorithm to many functional gene categorizations regarding transcriptional regulation, transcriptional variability, conservation level, gene origin and gene functionality we would gain insight in the extensive positional clustering of the yeast genome. The subsequent examination of the relationships between significantly overlapping sub-clusters would help us reveal underlying relationships between various clustered gene categories. Besides clustering, the algorithm would evaluate chromosomal preferences or avoidance tendencies of the input gene categories as well. The major advantage of this approach is that it is easily applicable in any genomic categorization, which in combination with the variety of available data on yeast, will give us a full picture of the level of clustering in its genome. The novel step is that, by delineating the positionally clustered genes into coordinate level, we are able to evaluate significant overlaps between all the categories examined, enabling us to study the principles governing the organization of the linear genome at a greater level.

2

Materials & Methods

2.1 | Information on the data used in this analysis

2.1.1 | Gene-coordinates at a linear level

An annotation file of the *Saccharomyces cerevisiae* genome sacCer2, (June 2008 assembly) as provided by the University of California Santa Cruz (http://genome.ucsc.edu) was used in this analysis. Mitochondrial genes were removed leaving a total of 7071 genes. Intergenic distances were calculated on the basis of transcript-coordinates of genes. Different coordinates were used only for the meta-assemblages analysis (mentioned below) taken from the paper itself and reported by Saccharomyces cerevisiae database (https://www.yeastgenome.org/, source: SGD_features.tab).

2.1.2 | Gene-coordinates at a three-dimensional level

The three-dimensional model of the *S. cerevisiae* genome was obtained by Duan et al. (2010). Michalis Georgoulopoulos during his master thesis resampled the aforementioned model at gene resolution by linearly interpolating the model's control points to approximate the center base pair of each gene (see full procedure here: https://github.com/mgeorgoulopoulos/ScerSeg). In more detail, the center base pair of each gene is found between two successive control points (assuming their uniform distribution on chromosomes) of the source model. A weighted average of these two points is assigned as the position of the gene in space. This dataset includes a total of 6496 genes.

2.1.3 | Transcription-factor binding

Three datasets containing transcription factor potential binding sites were used in this analysis. As described by Harbison et al. (2004), a genome-wide location analysis (chIP on chip) in 12 different environmental conditions with conservation criteria and previous knowledge were used to re-discover the binding specificities of 102 DNA-binding transcriptional regulators of *Saccharomyces cerevisiae* through the utilisation of six different motif discovery algorithms (**Figure 2.1**). These new specificities were then mapped on the yeast genome. Stringent binding and conservation criteria were used to create a transcriptional regulatory code. Variants of the map constructed with different binding and conservation thresholds were also made available. In this analysis we used a map of all potential interactions, thus constructed with lower-confidence information as provided by the University of California Santa Cruz or UCSC (Available on: `https://genome.ucsc.edu/cgi-bin/hgTables?hgsid=1198681503_0zToND2yJmhuG2Cf7zBOkiKBvJuG&hgta_doSchemaDb=sacCer2&hgta_doSchemaTable=transRegCode`). For each of these interacting regions the overlapping SacCer2 promoters were found (minus 300 and plus 100 base-pairs around the promoter). The final dataset contains 102 transcriptional regulators and 6026 gene promoters with which they have potential interactions (excluding the tRNA genes). 768 gene promoters were found with no available interactions ("no-signal" genes), possibly not overlapping with any regulator's motif or with any probe used in the chip. It is important to note that this map was created on the basis of regulators binding in multiple growth conditions.

MacIsaac et al. (2006) compiled a refined version of the above regulatory code by using two improved motif-discovery algorithms, thus enriching the existing interaction map. Again, different versions of the refined regulatory map were generated with loose or stringent criteria for binding and/or conservation available from: `http://fraenkel-nsf.csbi.mit.edu/improved_map/`. In the current analysis, two versions were used. The more relaxed version, with no conservation or binding criteria, contained potential interactions between 121 factors and 5693 gene promoters (without the tRNA genes). The second and more stringent version contained sites conserved across at least 3 out of 4 yeast *sensu stricto* species, and bound at p < 1E-3 in the location analysis of Harbison et al. (2004). This included 117 factors versus 1985 gene promoters. A factor is considered to have an interaction with a gene if there are bound instances for that factor in the intergenic region upstream a given gene.

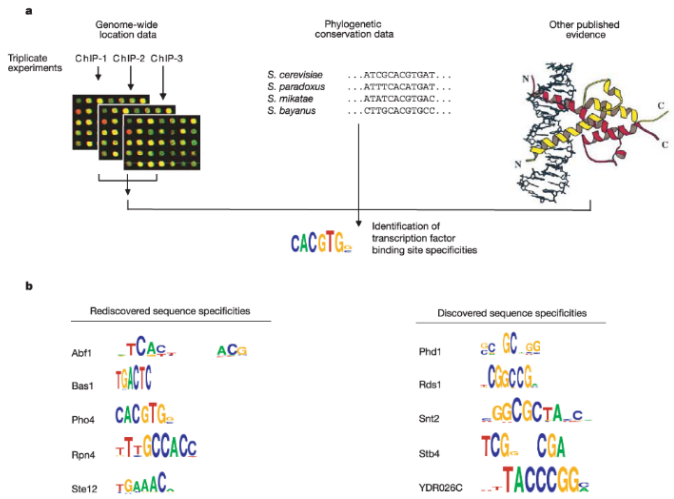


Figure 2.1: The procedure followed by Harbison et al. (2004) to discover the binding-site specificities for yeast transcriptional regulators. a, A combination of genome-wide location analysis (chIP on chip) to determine cis-regulatory sequences with phylogenetically conservation and previously published evidence was used to locate the potential binding-sites of many transcription factors in yeast. b, Some of the resulting factor specificities that were either rediscovered or newly discovered are shown. The total height of the column is proportional to the information content of the position, and the individual letters have a height proportional to the product of their frequency and the information content.

2.1.4 | Positional frequency matrices of transcription factors' binding motifs

The positional frequency matrices (PFM)s of *Saccharomyces cerevisiae* factors were obtained from the Jaspar database (Sandelin et al. (2004)), available at: `https://jaspar.genereg.net/download/data/2022/CORE/JASPAR2022_CORE_fungi_redundant_pfms_jaspar.txt`. Those were used to calculate the corresponding position specific scoring matrices (PSSM) with which the yeast genome was scanned in order to locate potential binding regions of transcription factors. Regions achieving the 66.6% of the max PSSM score per factor were further filtered by discarding those with scores lower than the 5% top scoring motifs per factor. Finally, from the aforementioned filtered regions, only those overlapping with the promoter regions used in Harbison et al. (2004) were kept. The information content for each factor was also calculated as the amount of uncertainty per motif position normalized by motif size. More specifically on the basis of the positional weight matrices, an entropy-based metric was calculated as the difference between the maximum entropy (expected nucleotide frequency across the genome) and the observed entropy (observed nucleotide frequency per motif position). This can be

mathematically explained as:

$$I = H_{max} - H_{observed}$$

where I is the information content and H is the entropy calculated as:

$$\Sigma_i^4 P[i] log(P[i])$$

with P[i] being the frequency of each of the four nucleotides in the specific motif position. The final information content is calculated as the sum of I across the motif's positions divided by the length of the motif.

2.1.5 | Genome-wide protein meta-assemblages

Protein meta-assemblages were obtained from Rossi et al. (2021). Rossi et al. (2021) performed ChIP-exo to map genome-wide binding in yeast with high accuracy. Each meta-assemblage describes a population-based consensus of composite target co-localization on the genome reflecting different regulatory patterns. As meta-assemblages are based on cell populations, their member targets tend to bind the same genomic locations, although not necessarily at the same time. In this analysis, 40 meta-assemblages were used containing 371 factors and encompassing a total of 6472 unique genes. A gene can be part of more than one meta-assemblage. Depicted in **Figure 2.2** are the functional categories describing most of the meta-assemblages used in this analysis, accompanied by the name of the factors that make up each meta-assemblage. The miscellaneous meta-assemblages refer to binding events that were either rare or highly isolated, thus do not represent a combination of co-localized factors on the genome. As these data are based on the sacCer3 assembly of the yeast genome and contain a number of non-coding elements we decided to do further analysis on the same gene pool as with the transcription factors before by using the sacCer2 assembly. As a result, the total number of gene-targets of the meta-assemblages was decreased to 6000 unique genes.

2.1.6 | Various gene classes

In all, 1091 essential (the rest are labeled as non_essential) genes of *S. cerevisiae* were obtained from a genome-scale functional profiling (Dow et al. (2002)). In addition, 1073 TATA (the rest are labeled as TATAless) genes were obtained from a concise data set, which was formed by taking into account the location and conservation of a TATA consensus in the gene's upstream region and the gene's sensitivity to TATA binding proteins (Basehoar et al. (2004)). 1018 small-scale duplicates (SSD) and 1092 whole-genome duplicates (WGD) were obtained from Fares et al. (2013) identified by performing all-against-all BLAST-searches using BLASTP. 373 genes

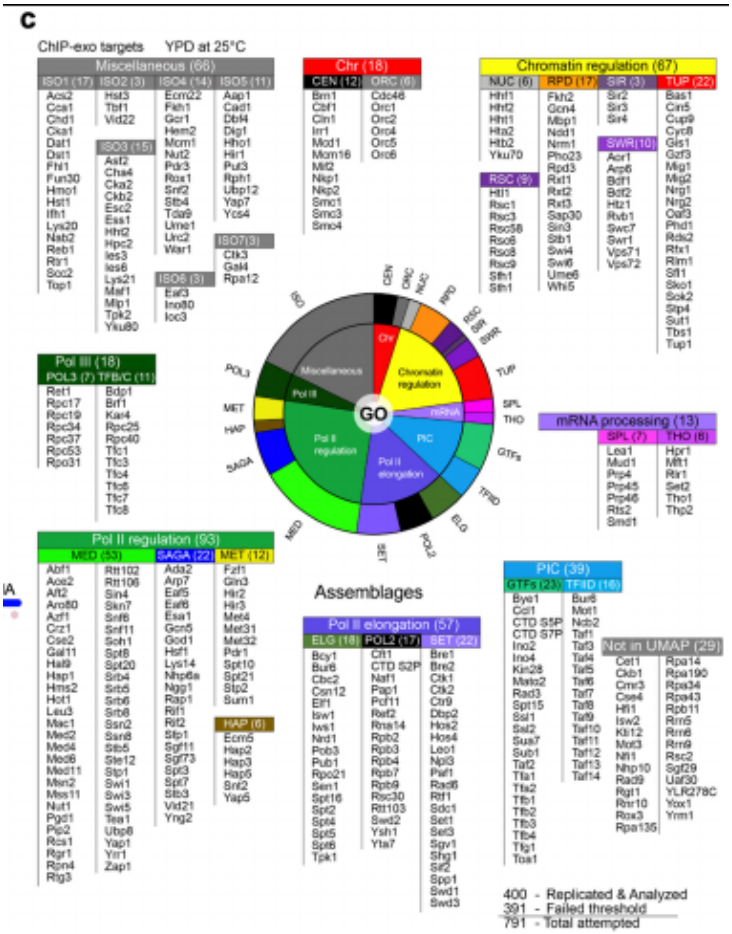


Figure 2.2: A pie-chart derived from Rossi et al. (2021) showing the general GO-terms related to most of the meta-assemblages (outer ring). Listed are also the common names of the assayed factors grouped on the basis of the related GO-terms and meta-assemblages.

expressed with high frequency (high_freq_epxr) and 495 genes expressed with low frequency (low_freq_expr) were obtained from the SPELL database (`https://spell.yeastgenome.org/`).

2.1.7 | Evolutionary origin of genes

The origin taxa of the *S. cerevisiae* genes were kindly provided by Nikos Vakirlis. These were determined by performing BLASTP searches on all the available RefSeq Fungi proteomes with an E-value cutoff of 1E-3. Finally, the origins of the *S. cerevisiae* genes were defined by tracing the most recent common ancestor of species that shared a homologue. **Table 2.1** depicts the distribution of genes across the evolutionary taxa. The taxa are ordered on the basis of decreasing evolutionary "age".

Origin	Number of genes
Fungi	4525
Ascomycota	295
Saccharomycetales	378
Saccharomycetaceae	226
Saccharomyces	101
Saccharomyces cerevisiae	563

Table 2.1: The number of genes across the evolutionary origin categories, ordered on decreasing "age".

2.1.8 | Gene-ontology terms

The *S. cerevisiae* genes with their corresponding gene-ontology terms were obtained from the *Saccharomyces* Genome Database or SGD (`http://www.yeast-genome.org`). The total number of genes was equal to 7127 and the GO-terms were equal to 5899. Genes can be matched to more than one GO-term. In this analysis only 1460 GO-terms with more than 5 genes each were used.

2.1.9 | Distances from Telomeres/Centromeres

Centromere and telomere coordinates were obtained from SGD (`http://www.yeast-genome.org`). Distances from the centromeres were scaled in the following way; the closest distance to the corresponding centromere was calculated for each gene and then divided over the size of the region spanning the centromere and the edge of the chromosomal arm in which the gene was lying. In this way, all distances were rescaled in a range from 0 (overlapping the centromere) to 1 (lying at the edge of the chromosomal arm). The same scaling was also performed for the distances from telomeres. Since telomeres are of very restricted size in the yeast genome, the edge of the corresponding chromosomal arm was used instead of the actual telomere coordinates.

2.1.10 | Transcriptional variability and Conservation score

We obtained normalized expression data from a compendium of 2400 experimental conditions from the SPELL database Hibbs et al. (2007). In order to assess expression variability, we calculated the standard deviation of gene expression levels for each gene across all conditions and then normalized it across genes with the use of a z-score. Sequence conservation was determined using phastCons scores for *S. cerevisiae* (Siepel et al. (2005)) as calculated on the basis of multiple genome alignments against six *Saccharomyces* species (*Saccharomyces paradoxus*,

Saccharomyces mikatae, *Saccharomyces kudriavzevii*, *Saccharomyces bayanus*, *Saccharomyces castelli*, *Saccharomyces kluyveri*)

2.2 | Linear-positional clustering

To test whether a group of genes (given as input) shows higher linear-positional clustering at a chromosome level than expected by chance, the observed average intergenic distances of consecutive genes, per chromosome, were compared to those obtained from 1000 permutation tests. In more detail, in each permutation, the order of genes in the dataframe was shuffled and the matching number of input genes per chromosome were kept. As an example, if the input consists of 30 genes positioned on chromosome I and 70 genes on chromosome IV, during each permutation test, 30 random genes of chromosome I and 70 of the chromosome IV are extracted to be further used. For each chromosome of the analysis a z-score and a p-value evaluating the significance of the difference between the observed and expected average distances is calculated. To achieve that, the average of intergenic distances between consecutive genes of the input is computed for each chromosome separately. From the corresponding random datasets, the randomly pulled genes are ordered according to their coordinates (consecutive) and the same property is calculated resulting in 1000 average intergenic distances for each chromosome of the analysis. In cases in which the intergenic distances per chromosome were more than 4 then distances higher than the 95% percentile were discarded, both for the observed and the random cases. This step minimizes the effect of outlier distances. Z-scores are finally calculated for each chromosome of the input, as the difference between the observed average of intergenic distances (χ) and the average of the 1000 random average intergenic distances (μ) divided by the standard deviation of the 1000 random averages (σ). Correspondingly, p-values are computed as the fraction of times the random average intergenic distances are equal to or more extreme than the observed average intergenic distance per chromosome. Significant cases of absolute z-scores higher than 1.96 with at least 5 genes per chromosome are kept in the end. Each step of the procedure is explained in greater detail below:

1. Working on the observed intergenic distances:
    - A dataframe containing the gene names and their corresponding chromosomes is inserted as input. The genes should be part of the distance matrices that were computed in a previous step.
    - Calculate and store the number of genes per chromosome (chromosomal profile). This information will be used to calculate a tendency of preference/avoidance for specific-chromosome positioning.

- Genes of the input per chromosome are sorted based on their linear order and then are divided into groups of consecutive two genes. For each pair of pseudo-consecutive genes (because not all genes are included in the input) the corresponding intergenic distance is traced back in the corresponding distance matrix. The top 5% of intergenic distances is discarded (only in cases with at least 5 genes) and the average is computed and stored in a dataframe accompanied by the corresponding chromosome.

2. Working on the expected intergenic distances
   - For each permutation (1000 in total) the order of all the yeast genes is shuffled. For the chromosomal preference test, the matching number of random genes as in the input is used to create a random chromosomal profile. The procedure following refers only to the linear-positional clustering test. For each chromosome a number of random genes matching the one of the original dataset is used. The procedure is then identical to the one mentioned above for the observed intergenic distances. As a result there are 1000 random average intergenic distances for each chromosome included in the analysis.
3. Evaluating the linear-clustering:
   - For every chromosome a z-score is calculated as the difference between the observed average intergenic distance and the average of the 1000 random average intergenic distances divided by the standard deviation of the 1000 random averages. P-values are calculated as the fraction of times the expected average distances are equal to or more extreme than the observed average intergenic distance per chromosome.
   - Results of absolute z-scores equal to or higher than the value 1.96 are labeled as statistically significant. Significant cases with less than 6 genes per chromosome are discarded.
4. Evaluating the chromosomal preference/avoidance
   - For each chromosome a p-value is calculated as the fraction of times the number of random genes per chromosome is equal to or more extreme than the observed number of genes per corresponding chromosome.
5. Making of Sub-clusters:
   - For each significant case (either negative or positive z-score) genes are divided into sub-clusters on the corresponding chromosomes on the basis of their intergenic distances. Each intergenic distance is subtracted from the average and divided by the

stadard deviation of all intergenic distances of the specific group of genes on the chromosome. Genes separated by a distance greater or equal to the one corresponding to 2 stardard deviations from the mean are split into separate sub-clusters otherwise merged into the same sub-cluster.

- A density property is calculated for each sub-cluster as the number of genes in the sub-cluster divided by the total number of genes in the chromosomal region with the same coordinates.

6. Final output:

   - A dataframe with all the resulting z-scores and p-values of the linear-positional clustering test.
   - A dataframe with only the significant z-scores of the linear-positional clustering test.
   - A dataframe containing the chromosomal preference/avoidance results.
   - A dataframe with the coordinates, genes and density of each sub-cluster.

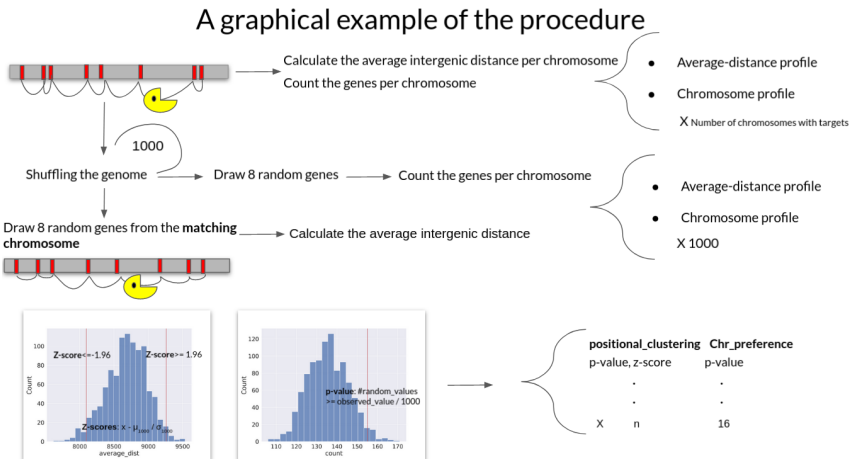


Figure 2.3: A graphical example of the procedure followed in order to evaluate the linear-clustering and the preference/avoidance at chromosomal level. Shown above is a toy example of 8 genes (red blocks) positioned on a single chromosome (gray large block). Their intergenic distances (curved lines) and their number per chromosome are the metrics computed against 1000 random permutations resulting in the calculation of p-values and z-scores. Distances higher than the top 5% are excluded (yellow PAC-MAN) only in cases in which the input genes are at least five.

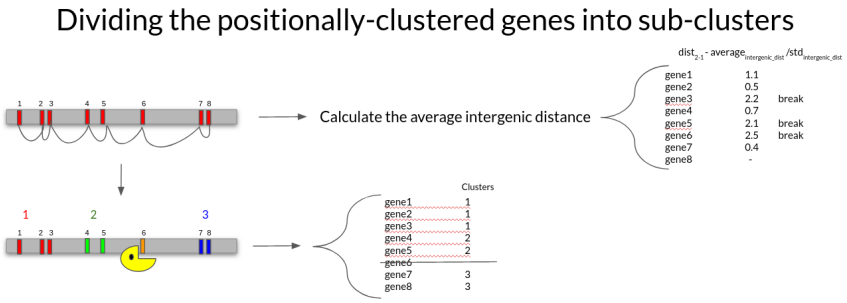


Figure 2.4: A graphical example of the procedure followed in order to divide the positionally-clustered genes in sub-clusters spanning across the chromosome. Genes are divided based on a distance threshold corresponding to 2Z otherwise merged in the same sub-cluster. Isolated genes on the chromosome are excluded (yellow PAC-MAN). Different colours indicate distinct sub-clusters.

2.2.1 | Computation of intergenic distances

To make the main algorithm time-efficient, a dictionary of sixteen matrices (one for each chromosome) was created containing the intergenic distances between all genes. The distance between two genes was calculated as the difference between the "start" coordinate of the second, in linear order, gene and the "end" coordinate of the first, in linear order, gene. The direction of transcription was not considered in this analysis. When two genes were overlapping their distance was set to zero. The resulting symmetric matrices with zeros in the diagonal were stored in a dictionary, saved as a PKL file and loaded in the main algorithm when needed.

2.2.2 | Estimation of Statistical Significance

To estimate the statistical significance of the properties described in this analysis, z-scores and p-values were computed against 1000 permutation tests. In each permutation test, there was a random selection of a gene subset which resulted in the creation of random distributions. In the case of chromosomal preference, the null hypothesis is that genes are equally distributed across chromosomes, while in terms of linear clustering the null hypothesis states that the intergenic distances will not differ significantly regardless of which genes are chosen. The p-values were computed as the fraction of values produced in N permutations that were, in each case, equal to or more extreme (greater or smaller) than the observed value. Z-scores were also calculated as the number of standard deviations (resulting from the random distribution) the observed average value was away from the average of N permutations.

2.2.3 | Chromosomal preference/avoidance

To test whether a group of genes (given as input) shows specific chromosomal preference (or avoidance) more often than expected by chance, the observed number of genes per chromosome were compared to the expected ones, resulting from 1000 permutation tests. More specifically, in each permutation, the order of genes was shuffled. Afterwards, the matching number of total genes as in the input were extracted from the shuffled dataset and their distribution across chromosomes was stored in a dataframe. Finally, p-values were computed, for each chromosome, as the fraction of times the observed number of genes was equal or more extreme (greater or smaller) than the one produced in N permutations of the original gene list.

2.3 | Overlap analysis between different sub-clusters

To test for significant overlaps between sub-clusters of positionally clustered genes of different gene categories, we used a procedure described in Andreadis et al. (2014). This algorithm computes a ratio by dividing the observed overlap of two clusters by their expected overlap as if they were two independent variables, using their coordinates. Statistical evaluation is accomplished by permutation tests, in which one of the two input's coordinates gets shuffled and the same procedure as before is followed. Finally, p-values are computed as the fraction of times the observed enrichment (or depletion) were more extreme (higher or lower) than the expected ones. In the current analysis, this approach was used to test the overlaps between the sub-clusters of all the combinations of the different categories yielding clustering results across chromosomes. Only significant overlaps (pvalue<=0.05) were further analyzed. It should be noted that, in order to get a manageable number of results, from the transcription factors datasets, only the Harbison et al. (2004)'s was used in this analysis. Networks were created with Cytoscape (`https://cytoscape.org/`).

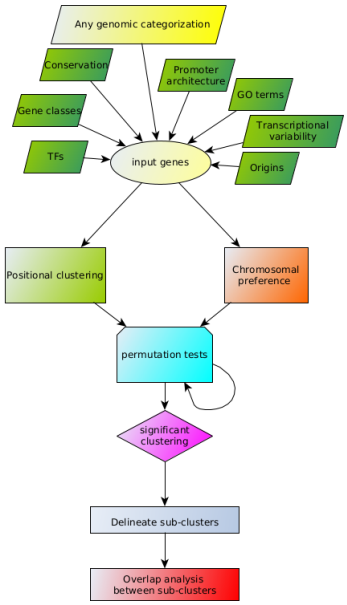


Figure 2.5: A flow chart depicting the order and the relationships of the most important procedures of this analysis. The green boxes indicate the data used so far as input to the algorithm.

2.4 | Polar coordinates and radial positioning of positionally clustered genes

To evaluate a potential preference/avoidance in the radial positioning of the positionally clustered genes, we divided the genome into three categories based on the genes' distance from a centromeric pole. According to Duan et al. (2010) three-dimensional model, the yeast genome acquires a Rabl configuration in the nucleus, in which the centromeres are accumulated in the spindle pole body of the nucleus with arms extending to the telomeres abutting the nuclear envelope. In our coordinate system the centromeric pole was used as a theoretical center on which the polar coordinates of genes were based. More specifically, 135 genes positioned near the centromeres (in the two-dimensional space) were chosen to represent this centromeric pole. These genes were chosen based on a distance threshold corresponding to the mean minus 1.65 standard deviations from the centromeres. Finally, the three-dimensional coordinates of this centromeric pole were defined as the average x, y and z of those 135 genes as provided by Michalis Georgoulopoulos (`https://github.com/mgeorgoulopoulos/ScerSeg`). The rest of the genes were divided into three categories based on their euclidean distance from the centromeric pole. For each chromosome separately, the genes corresponding to the bottom 25% distances were labeled as central, the genes corresponding to the top 25% were labeled as peripheral and the ones corresponding to the middle 50% as intermediate. The resulting radial

categories are shown in **Figure 2.6**.

Finally, an enrichment analysis was conducted for all the positionally clustered genes in each one of the three radial groups. This was accomplished by dividing the observed frequency of positionally clustered genes in each radial category by their expected frequency in the whole genome. Only genes included in the three-dimensional dataset provided by Michalis Georgoulopoulos were used in this analysis, so the number of genes used for the enrichment computations were slightly different from the actual number of positionally clustered genes. To statistically evaluate the resulting enrichments, permutation tests were conducted. In each permutation the positionally clustered genes were substituted with random ones coming from the same pool of genes while keeping their number per chromosome the same. For example, to test the preferential positioning of the 100 positionally clustered targets of a transcription factor X, 100 genes were randomly drawn from the initial subset of total targets in the corresponding dataset while keeping the number of genes per chromosome the same. Genes of chromosome XII were excluded from this analysis because of its special conformation forming the nucleolus.

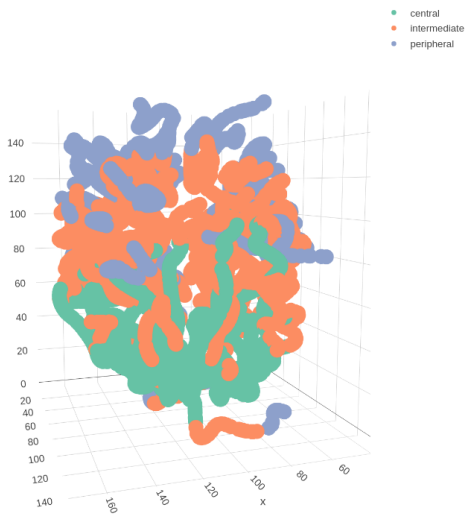
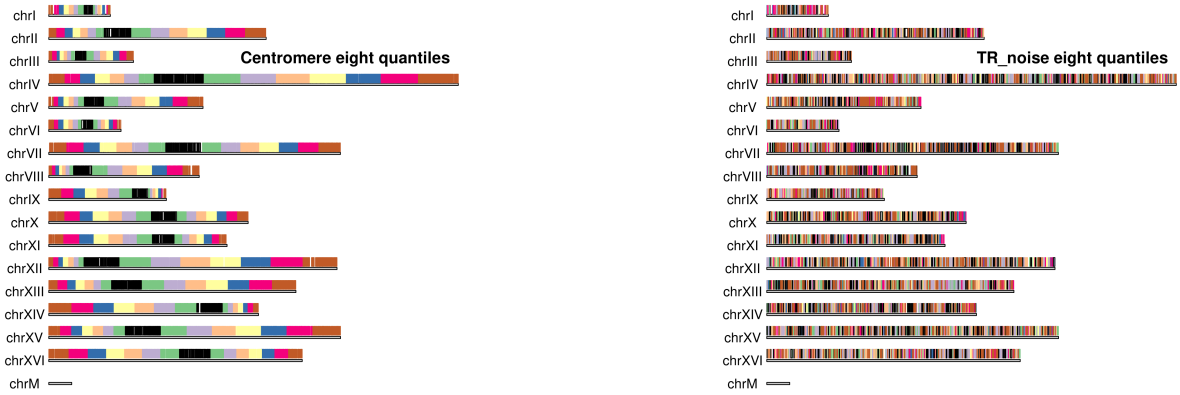


Figure 2.6: Genes in the three-dimensional space divided into three categories based on their euclidean distance from the centromeric pole (inside the green region). Genes near the centromeric pole are labeled as central (green), genes further from the centromeric pole are labeled as peripheral (blue) while the genes in the middle are labeled as intermediate (orange).



(a) Centromere quantiles

(b) TRnoise quantiles

Figure 2.7: Genes are divided into eight equisized groups (quantiles) based on the normalized distances from the centromeres 2.7a or on their transcriptional variability 2.7b. Each colour depicts a different quantile with black corresponding to genes positioned close to the centromeres or to genes with very low transcriptional variability and brown corresponding to the ones further from the centromeres or to the ones with the highest transcriptional variability.

2.5 | Enrichment analysis in various segments

To test the potential positional and functional tendencies of various gene categories, we divided the genome into eight proportionally sized groups (quantiles) based on the distribution of gene distances from the Telomeres/Centromeres (**Figure 2.7a**) or on the transcriptional variability distribution (**Figure 2.7b**). Each one of those eight quantiles in each case consisted of genes with low to high transcriptional variabilty or of genes positioned very close to the centromeres to the ones positioned very close to the chromosomal edges. These eight quantiles were used in an enrichment analysis of various gene categories in which the observed frequency of the gene category in the quantile is divided by the expected one in the genome. 1000 permutation tests were conducted for the statistical evaluation of the enrichments. A similar procedure was followed to compute enrichments of various gene categories in the chromosomes.

3

Results

3.1 | Compartmentalization of the yeast genome

In this section we describe some findings regarding the positional preferences of various gene categories. We have widely conducted enrichment analysis on segments of the genome made on the basis of centromere distances or transcriptional variability. This methodology is described in materials and methods and gives us insight into the yeast genome organization.

3.1.1 | Various gene categories in yeast show chromosomal preferences

By using a variety of available gene categorizations made on the basis of their regulation and their functional aspects, as described in materials and methods, we followed an enrichment analysis approach (materials and methods) in order to observe potential positional preferences. Starting this analysis from a chromosomal level, we have computed, for each chromosome, an enrichment as the observed frequency of a gene group over its expected frequency in the whole genome. Depicted in the **Figure 3.1** is a heatmap showing the hierarchical clustering of the resulting enrichment values of the various genomic categories (in the rows) across the sixteen chromosomes (in the columns).

Based on the resulting dendrogram at the top of the heatmap, chrI has the most distinct enrichment pattern among all chromosomes. Supported by statistical evidence, through permutation tests (results not shown), chrI is significantly enriched in TATA genes (enrichment approximates 2, pvalue<=0.01) and in the evolutionary origin category "*S. cerevisiae*" (enrichment approximates 1.4, pvalue=0.02) while is depleted in the Fungi (enrichment approximates 0.7, pvalue<=0.01) and in the essential genes (enrichment approximates 0.56, pvalue=0.01). Similar points can be made for other chromosomes as well. The chrVI owns the most extreme under-enrichment of the "*S. cerevisiae*" origin category. On the other hand there are chromo-

somes sharing very similar patterns across the gene categories, like chromosomes XIV and XVI or like chromosomes VII and X. The "*saccharomyces*" origin category, which is the category with the lowest sample size (approximates 100 genes), owns the most turbulent pattern across chromosomes, mostly found in specific chromosomes while depleted from the rest of them.

Keeping in mind that finite size effects and small gene set sizes are likely to affect every enrichment analysis, this simple approach provides us with insight into the genome organization across chromosomes. Based on the results above, chromosomes seem to have slightly different enrichment patterns between one another, across general gene categories, which is a process affected by evolutionary mechanisms and purely random processes. Nevertheless it is an interesting assumption that each chromosome is a distinct niche for different genomic categorizations potentially related to specific functionalities.

3.1.2 | Various gene categories have positional preferences in specific regions of the chromosomes

In order to study this time the positioning of the genomic categorizations used in the previous section in intrachromosomal level, we divided the chromosomes into eight equisized groups of genes (or else quantiles) on the basis of the genes' distances from the centromeres ending up with a chromosomal segmentation as depicted in **Figure 2.7a**. In this figure, black indicates the segments being very close to the centromeres while brown indicates genes positioned close to the telomeres. It is easily observed that this method of segmentation is not symmetric as it follows the positioning of the centromere on each chromosome.

By following the same approach as before, we found preferences of various gene categories to be positioned at specific distances around the centromeres. Some of the significant (evaluated through permutation tests) enrichments across different quantiles are shown in **Table 3.1**. It is observed that the only category found enriched near the centromeres are the essential genes while also being significantly depleted near the telomeres. TATA genes, which are primarily stress related genes (Basehoar et al. (2004); Huisinga and Pugh (2004)), are enriched near the telomeres, a finding that agrees with previous works in the field (Basehoar et al. (2004)). Similarly, the evolutionary "younger" genes of the *S. cerevisiae* origin category are significantly enriched in regions near the telomeres while the opposite is true for the "older" genes of the Fungi origin.

This genomic compartmentalization agrees with and is described in previous work of our lab (Tsochatzidou et al. (2017)). In addition, we found that small scale duplicates are significantly enriched in regions near the chromosomal edges while they are depleted near the centromeres. On the contrary, the whole genome duplicates are enriched in gene groups relatively closer to the centromeres while being depleted near the telomeres. Finally, genes that are ex-

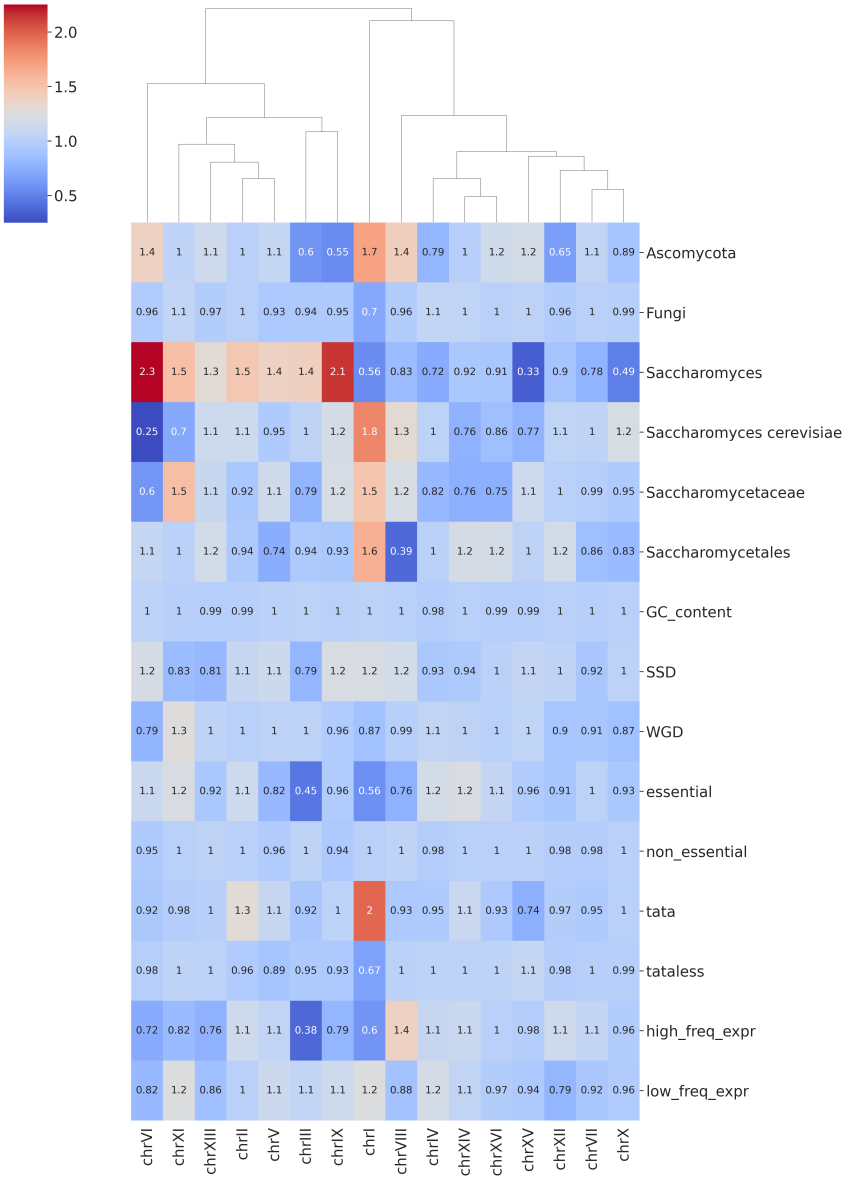


Figure 3.1: A heatmap that shows the enrichment values of various gene categories (rows), hierarchically clustered across the chromosomes (columns). Red indicates higher enrichment values than blue colour.

pressed in high frequency show a contradictory behaviour as although being enriched in regions closer to telomeres they are also depleted in a group of genes very close to the previous one but a "step" closer to the centromeres.

In the same manner, by dividing the genome into eight quantiles this time on the basis of the genes' trascriptional variability as shown in **Figure 2.7b**, we have found a tendency of genes

Gene category	quantile	enrichment value	pvalue
Essential	0	1.18	0.008
Essential	7	0.57	<0.001
TATA	6	0.84	0.02
TATA	7	1.25	<0.001
High freq expr	5	0.75	0.034
High freq expr	6	1.24	0.04
WGD	4	1.21	0.001
WGD	7	0.69	<0.001
SSD	2	0.81	0.005
SSD	7	1.39	<0.001
Fungi	7	0.85	0.022
S. cerevisiae	7	1.13	0.014

Table 3.1: The statistically significant enrichment values of various gene categories across segments (quantiles) positioned around the centromeres, at increasing distance. 0 indicates the group of genes positioned near the centromere while 7 indicates gene positioned near the telomeres.

of high transcriptional variability (seventh quantile) to be positioned closer to the telomeres than the centromeres (pvalue=0.001) while genes of lower variability tend to be depleted in the same regions (results not shown).

Based on the results above, specific group of genes characterized by different regulatory environments and different functionalities, seem to prefer distinct positioning in the chromosomes with some of them occupying the telomeric regions while others the centromeric. This compartmentalization reflects their correlation with specific functions and is governed by evolutionary mechanisms. We observe significant enrichments around the telomeres of gene groups reflecting stress responses, recent origin and more variable expression across different conditions. Essential genes occupy regions closer to the centromeres and avoid telomeric ones (Batada and Hurst (2007)) securing a more stable regulatory environment.

Finally, we observed a segregation in the positioning of the small scale duplicates (SSDs) and the whole genome duplicates (WGDs). SSDs seem to prefer to be positioned closer to the telomeres while the opposite is true for the WGDs. This segregation may reflect their functional differences as described by Fares et al. (2013). Based on their research , SSDs are more likely to form new functions (neo-functionalization) while WGDs are linked to the sub-functionalization of the ancestral functions. We assume that the telomeric environment is more likely to give birth to new functions as it is also enriched in younger genes.

3.1.3 | Transcription variability across various gene categories

Aiming to study the transcriptional variability across the same gene categories, we have segmented the genome into eight equisized groups (quantiles) on the basis of the transcriptional

variability of genes, as depicted in **Figure 2.7b**. Again black colour indicates low transcriptional variability while brown indicates high values of variability. These quantiles do not have profound positional preferences but as mentioned in the previous section we have detected a tendency of genes with high transcriptional variability to be closer to the telomeric regions. By conducting the same type of enrichment analysis, we computed the enrichment values of the various gene categories across the transcriptional variability quantiles and evaluated them through permutation tests. Depicted in the **Figure 3.2** are the enrichment values of the gene categories (rows) across the eight quantiles (columns) which are ordered in increasing transcriptional variability.

Firstly, we observed that extreme enrichments cannot be detected in the majority of the categories except for the genes with high expression frequency. This category is significantly enriched (enrichment approximates 2.4, pvalue<0.001) only in the quantile with the highest transcriptional variability while being depleted almost in the rest of the quantiles which ,in some cases, is also statistically significant (results not shown). This implies that genes with high expression frequency tend to have variable expression patterns across different conditions in yeast. This observation may be explained if one takes into account that highly expressed genes are qualified as such under only a certain subset of fast growth conditions.

We observe similar patterns between categories like the WGDs and the TATA genes. These show an increasing enrichment as the transcriptional variability increases, reaching an enrichment of around 1.3 and 1.4 respectively, being statistically significant in both cases (pvalue<0.001). A similar tendency can be observed for the SSDs although the enrichment gradient across quantiles is milder and there is no statistical significance. On the contrary, categories like the essential genes follow the opposite trend as they show higher enrichments in quantiles of low variability which decrease as the variability increases. Essential genes are significantly enriched, based on the permutation tests, in both the first and the second quantiles (enrichment approximates 1.2, pvalues=0.017 and 0.024).

Through the analysis of this whole section, we have gained knowledge on positional and functional preferences of some general gene categories of the yeast genome. In the next section, we hope to provide results that give us interesting insight in the organization of the yeast genome as we focus on the detailed studying of the potential clustering of genes in the linear level, covering a wide range of data regarding the transcriptional regulation, the conservation, the transcriptional variability, the origin of genes and their functionalities.

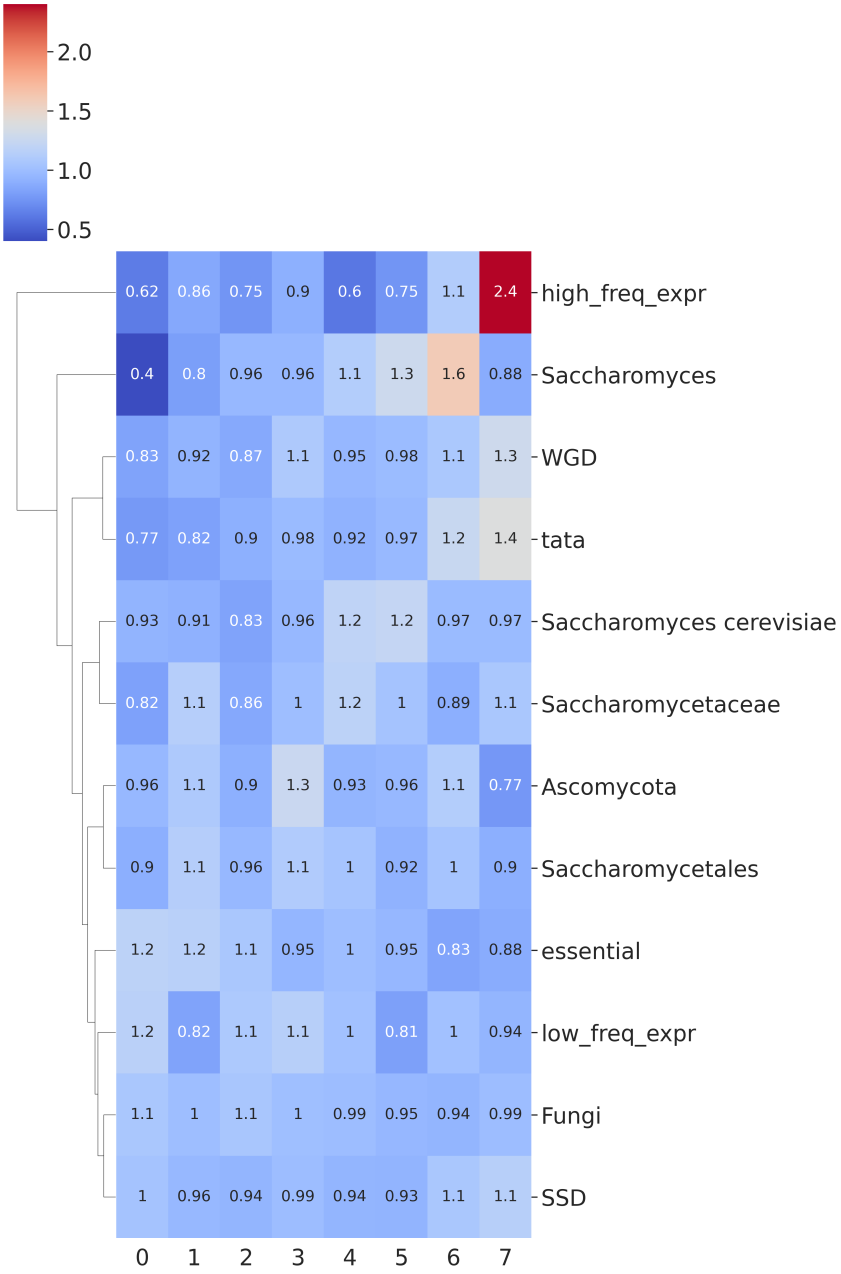


Figure 3.2: A heatmap that shows the enrichment values of various gene categories (rows) in the eight equisized transcriptional variability quantiles (columns), hierarchically clustered across the quantiles (dendrogram). Red indicates higher enrichment values than blue colour. Zero to seven indicates increasing transcriptional variability.

3.2 | Linear-positional clustering analysis

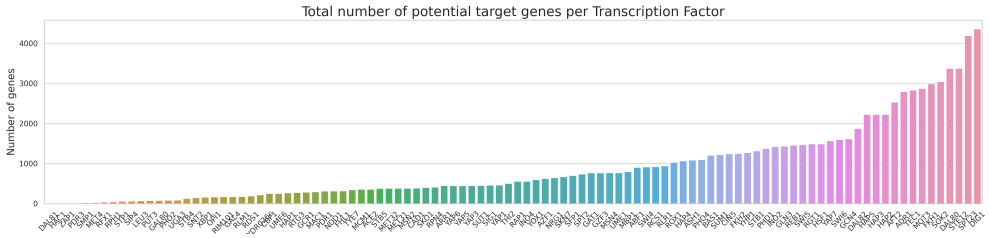
Based on previous studies of our lab (Tsochatzidou et al. (2017)), topoisomerase pertrurbation in yeast causes the emergence of either up- or down-regulated gene clusters. These clusters were shown to be co-expressed and to possess strong functional and positional preferences, organized in a compartmentalized manner that emerged as a response to stressing conditions. It is in our interest to unveil such cases of genome organization in the linear level by studying the clustering of many gene categorizations across the chromosomes. We hypothesize that if there is such an underlying organization in the yeast genome for many different gene groups, this would enable the cell to regulate each time, in response to different conditions, the appropriate "blocks" of clustered genes across the genome, in the same manner as in topoisomerase pertrurbation.

This section starts by describing the results of the linear-positional clustering analysis on various gene categorizations. As mentioned before, our main goal was to evaluate the potential clustering of various gene groups in the linear level which would give us insight on the principles governing the organization of the yeast genome. For that reason we came up with a method in which we evaluate the proximity of genes on the chromosomes by comparing the actual intergenic distances to randomized ones through permutation tests (materials and methods). The following subsections correspond to the analysis done on each one of the gene categorizations that we used. These include data on the transcriptional regulation, on the evolutionary origin, on the transcriptional variability, on the conservation and on the functionality of genes (see materials and methods).

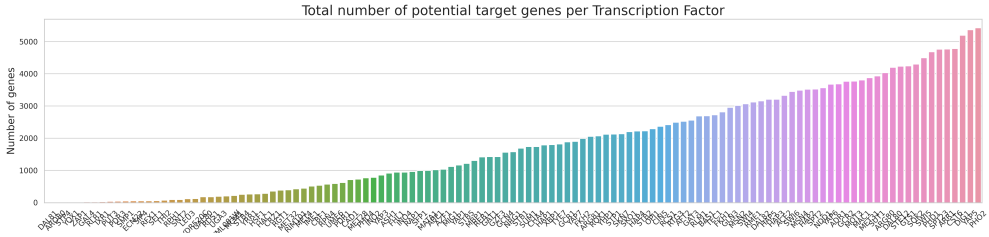
3.2.1 | Transcriptional regulation

This section provides analysis done on data regarding the transcriptional regulation in yeast. We were interested in studying the potential clustering of the gene-targets of various transcription factors. To achieve that, we used three datasets containing potential binding profiles of at least 102 trascription factors in yeast which are described thoroughly in materials and methods section (Harbison et al. (2004); MacIsaac et al. (2006)).

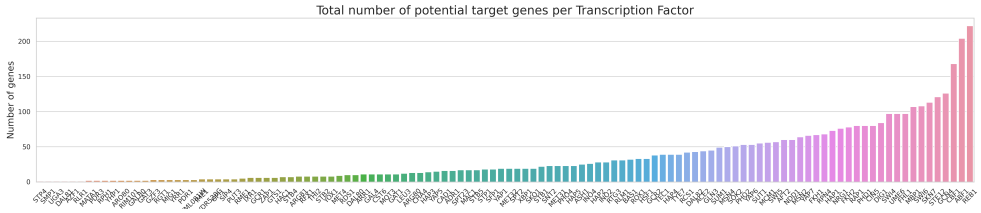
Figure 3.3 depicts the numbers of potential gene-targets for all the trasncription factors in the three different datasets used in this analysis. It is easily observed that the number of interactions in the first two datasets (**Figure 3.3a,3.3b**) are extremely high, with some transcription factors (e.g DIG1, SPT23) binding to more the the two thirds of the yeast genome. These two datasets contain interactions of both high and low confident levels which explains the high number of potential interactions. On the contrary, the third dataset (**Figure 3.3c**), which is based on the most stringent binding and conservation criteria, contains the fewer interactions



(a) Total gene-targets Harbison dataset



(b) Total gene-targets Macisaac dataset on loose-criteria



(c) Total gene-targets Macisaac dataset on stringent-criteria

Figure 3.3: The total number of potential gene-targets for each trascription factor corresponding to the three different datasets used in this analysis.

with REB1 reaching the maximum number being equal to 222. Thus, a threshold, limiting the number of total targets, was used to exclude cases of extreme binding that would probably falsely lead in extreme clustering. It was decided that transcription factors with more than 500 potential gene-targets were not furtherly analysed. The initial and final numbers of transcription factors per dataset are shown in **Table 3.2**

Dataset	Final number of TFs	Initial number of TFs
Harbison et al	50	102
Macisaac et al loose criteria	34	121
Macisaac et al stringent criteria	117	117

Table 3.2: The initial and final numbers of the transcription factors across the three datasets used in the clustering analysis after excluding those with more than 500 potential gene-targets.

Supplementary to the above, shown in the **Figure 3.4** is the normalized information content of all transcription factors' binding motifs (in y axis, PFMs provided by the Jaspar database:

materials and methods) against the number of potential interactions (x axis), based on the Harbison et al. (2004) dataset. We expect transcription factors with a low number of potential gene-targets to have a higher information content binding motif and the opposite for those having a high number of potential gene-targets. Indeed, the Spearman correlation (`https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.spearmanr.html`) between those two variables suggests a significant inverse relationship with a rho=-0.21 (p-value=0.045).

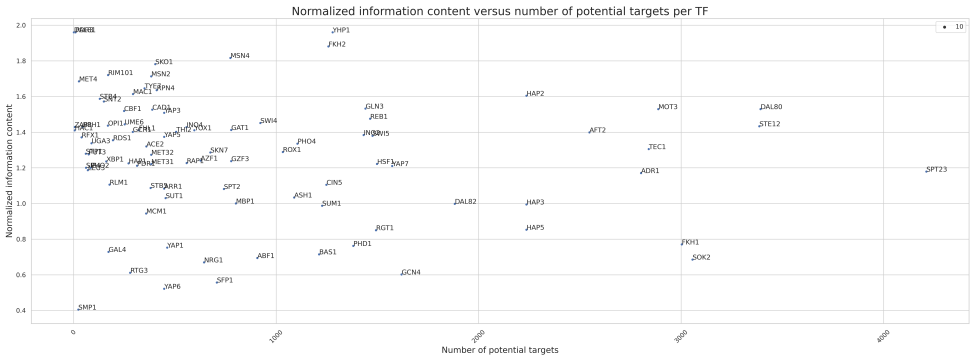


Figure 3.4: Depicted in the y axis is the normalized information content based on the binding motif (as provided by the Jaspar database) of each transcription factor versus the number of potential gene-targets (as provided by the Harbison et al. (2004) dataset) depicted in the x axis.

3.2.1.1 | Potential targets of transcription factors are linearly clustered on specific chromosomes

As mentioned in materials and methods, a z-score, evaluating the clustering of the input genes, is computed for each chromosome. A z-score lower than or equal to -1.96 indicates significant clustering on a specific chromosome. Depicted in **Figure 3.5** are the significant z-scores (in the cells of the heatmap) resulting from the clustering analysis done on the potential binding profiles of transcription factors (rows), coming from the three aforementioned datasets, against the chromosomes (columns).

As the three datasets consist of different interactions among transcription factors and genes, they yield different results. The first thing that we observed is that most of the transcription factors yielding results have clustered gene-targets in a restricted number of specific chromosomes. The transcription factors that correspond to extended clustering across chromosomes are expected to bind to a higher number of potential gene-targets. Depicted in **Table 3.3** are the percentages of TFs that yielded clustering results in at least one chromosome across the three datasets. This percentage is lower in the most stringent dataset because of the lower number of potential interactions between TFs and genes. Nevertheless, these results show that, on our

significance level, the 34% to 66% of the transcription factors that we used in each analysis own positionally clustered gene-targets in at least one chromosome.

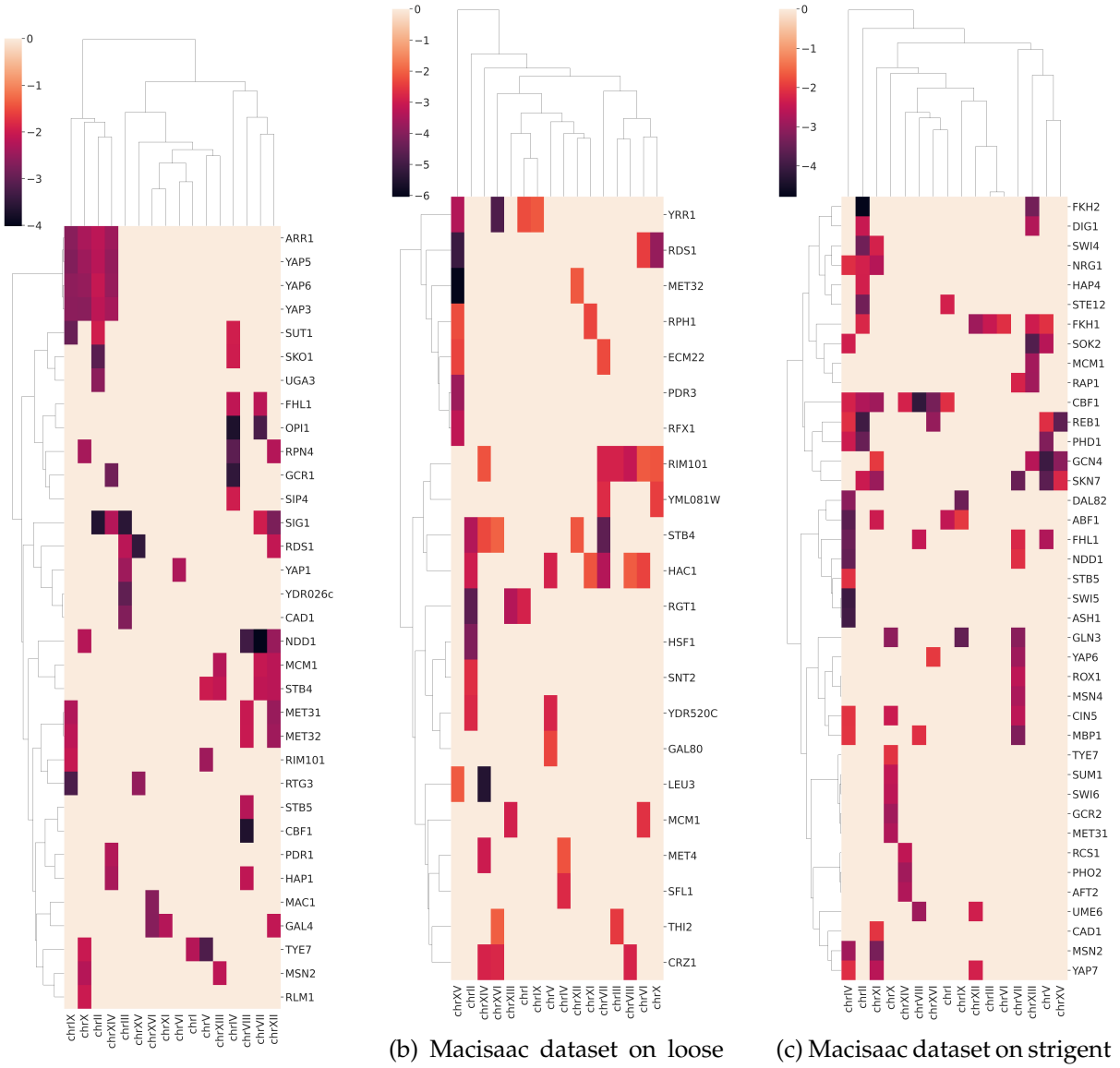
Dataset	Percentage of TFs with clustered targets in at least one chr
Harbison et al	66%
Macisaac et al loose criteria	64.7%
Macisaac et al stringent criteria	34.2%

Table 3.3: The percentage of transcription factors that yielded clustering results across the three datasets in at least one chromosome.

Some of the functional categories represented in the results are related to the regulation of the cell cycle (MBP1, SWI6, SWI4), to the regulation of the methionine biosynthesis (MET31-MET32), to stress responses (CAD1, MSN2), to the regulation of iron and copper (AFT2, MAC1) and to the regulation of mating and pseudohyphal growth (DIG1, STE12). Info on transcription factors were manually retrieved from the *Saccharomyces* Genome Database or SGD (https://www.yeastgenome.org/). Transcription factors yielding results in similar chromosomal patterns are not necessarily functionally correlated except for some cases in which transcription factors have almost identical binding patterns. Such examples are the YAP3-YAP5-YAP6-ARR1 group of TFs and MET31-MET32, both cases coming from the Harbison et al. (2004) dataset yielding a jaccard-index among their binding profiles higher than 95%. Cases of transcription factors with significantly overlapping clustering results are described in greater detail in the next chapters (see Overlap analysis). Only MCM1 yielded clustering results across the three datasets, with its targets clustered on chrXIII.

Depicted in **Table 3.4** are some of the most significant results (p-value<=0.005) of the chromosomal preference/avoidance results on the Harbison et al. (2004) dataset covering transcription factors that yielded clustering results in at least one chromosome. Fourteen transcription factors out of the 33 that yielded clustering results have significant chromosomal preference/avoidance tendencies. Based on those, the chromosomal preference do not coincide with the clustering results but it is possible that upon a less stringent statistical threshold we could detect mild correlations. On the contrary, such correlations can be easily detected in the results regarding the MacIsaac et al. (2006) stringent dataset (results not shown) but the sample size of genes-targets across chromosomes is significantly lower providing a less credible statistical analysis.

Based on the results so far, we detected many cases of various transcription factors whose potential gene-targets are positioned closer to each other than we would expect by chance. The transcription factors yielding such results represent many functional categories related to growth, cell-cycle and nutrient usage. Positional clustering is observed in specific chromosomes, depending on the factor, and do not seem to be highly correlated with the tendency of factors to prefer specific chromosomes over others, at least in one of the datasets that we



(a) Harbison dataset

(b) Macisaac dataset on loose criteria

(c) Macisaac dataset on strigent criteria

Figure 3.5: Depicted in these heatmaps are the resulting z-scores (in the cells of the heatmap) from the clustering analysis done on the potential targets of transcription factors across the three different datasets as described above. Shown in the rows are the transcription factors while the columns correspond to the chromosomes. Darker colours indicate a lower Z-score increasing in statistical significance.

TF	Chromosome	Preference/Avoidance	Observed	Expected	P-value
YAP1	chrXVI	preferred	54	35	0.002
YAP1	chrXII	avoided	27	41	0.005
YAP6	chrXV	avoided	40	23	<=0.001
RLM1	chrXII	avoided	5	15	<=0.001
ARR1	chrXV	avoided	23	40	0.001
SUT1	chrIII	avoided	4	12	0.001
CBF1	chrVII	avoided	10	22	0.003
CBF1	chrX	preferred	26	14	0.002
GCR1	chrVII	avoided	13	25	0.001
TYE7	chrV	preferred	29	17	0.004
MAC1	chrXI	preferred	26	15	0.005
SIG1	chrX	preferred	41	27	0.005
GAL4	chrXVI	preferred	25	13	0.001
UGA3	chrIII	preferred	8	2	0.005
MET31	chrXI	preferred	31	19	0.004
CAD1	chrXVI	preferred	46	30	0.004
CAD1	chrXII	avoided	22	34	0.004

Table 3.4: The chromosomal preference/avoidance results on Harbison et al. (2004) dataset. Here are shown only the results corresponding a p-value lower than or equal to 0.005 across TFs that yielded clustering results in at least one chromosome. If the expected number of gene-targets is higher than the observed one, then the chromosome is labeled as avoided.

checked based on a specific statistical threshold.

These results agree with previous work in the field done by Janga et al. (2008) in which they have shown that transcription factors have clustered gene-targets on specific chromosomes. They have also shown that transcription factors own chromosomal preferences and in more detail that some of them prefer to regulate targets on specific chromosomal regions. We have also conducted a similar analysis, although in the three dimensional level as described in (materials and methods). We tried to study potential tendencies of transcription factors to have positionally-clustered targets in specific chromosomal regions (central, intermediate, peripheral), through an enrichment analysis, but we did not get extensive results. Depicted in **Table 3.5** are some of those results corresponding to the positionally-clustered genes of the Harbison et al. (2004) dataset.

TF	Polar region	Enrichment	Number of genes	P-value
TYE7	peripheral	0.53	8	0.029
TYE7	intermediate	1.37	41	0.004
RIM101	central	2.39	9	0.005
SIG1	intermediate	1.18	65	0.019
SIG1	central	0.68	19	0.020
RTG3	peripheral	1.81	15	0.003
RTG3	central	0.12	1	0.002
MET31-32	intermediate	1.44	23	0.011
OPI1	peripheral	1.49	18	0.032

Table 3.5: The preference results of positionally-clustered gene-targets of transcription factors across three chromosomal regions (central, intermediate, peripheral). The significance level correspond to a p-value less than or equal to 0.05.

3.2.1.2 | Gene-targets of factors that co-localize on the genome are extensively clustered across chromosomes

Additionally to the above analysis on gene-targets of transcription factors in yeast, we have also used a recently published dataset published by Rossi et al. (2021). As described in material and methods, this dataset consists of forty clusters of factors created on the basis of their co-localization on the genome (except for the 'ISO' meta-assemblages). Rossi et al. (2021) refer to these categorizations as meta-assemblages and each one of those represent a different regulatory architecture on the genome. In the current analysis we have retrieved the gene-targets of each factor (detected by chip-exo sequencing) of the initial dataset and merged them on the basis of factors being part of the same meta-assemblage. Gene-targets may participate to more than one meta-assemblage. It should be highlighted that genes grouped in the same meta-assemblage are not necessarily targets of all factors participating in that meta-assemblage. Following the same method as with transcription factors, we excluded meta-assemblages with more than 800 total gene-targets resulting with twenty five meta-assemblages.

Dataset	Average number of gene-targets	Average number of significant z-scores
Harbison et al	312	2.42
Macisaac et al Stringent	75	2.3
Macisaac et al Loose	207	2.45
Rossi et al	455	7.13

Table 3.6: The average number of gene-targets and the average number of significant z-scores across chromosomes across the four datasets analyzed in the current section.

Depicted in the **Figure 3.6** are the results of positionally-clustered targets per meta-assemblage. The cells of this heatmap represent the z-scores evaluating the positional clustering of gene-targets per chromosome (rows) across the meta-assemblages (columns). The first observation that immediately stems from this figure, when compared to the previous results of the **Figure**

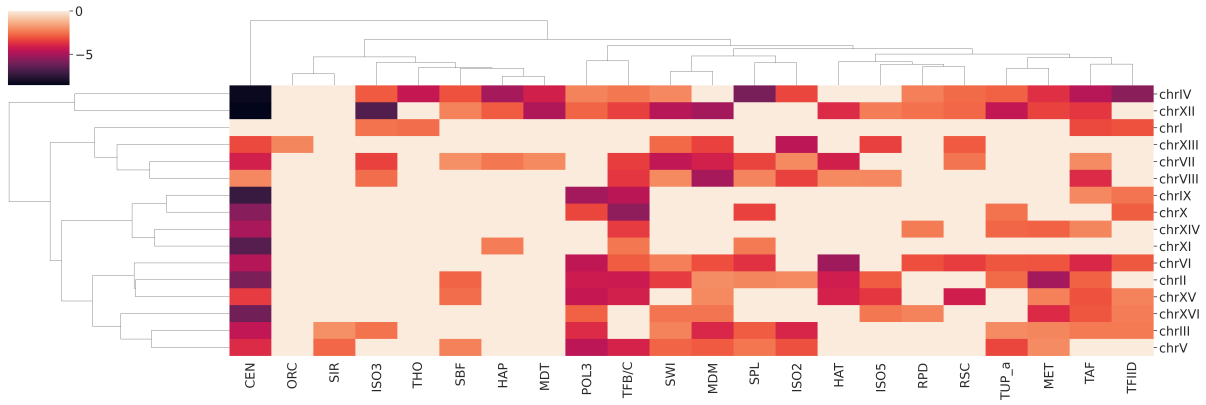


Figure 3.6: Depicted in this heatmap are the resulting z-scores (cells of the heatmap) from the linear clustering analysis done on targets of various factors that co-localize on the genome, derived from the meta-assemblages analysis by Rossi et al. (2021). Shown in the rows are the chromosomes while the columns correspond to distinct regulatory architecture categories (meta-assemblages). Darker colours indicate a lower z-score increased in statistical significance.

3.5, is that there is extensive clustering across many chromosomes for the majority of the meta-assemblages. To interpret these results we have to take into consideration the differences in sample sizes, as depicted in **Table 3.6**, between the datasets. Based on the information shown in the table, the dataset of meta-assemblages correspond to the higher average sample-size which may lead to an increase in the positional-clustering results compared to the other three datasets. Thus, it would be false to make a direct comparison between the different analyses. On the other hand, we cannot abolish the possibility that targets of a single transcription factor are less likely to be positionally clustered across many chromosomes while targets of factors that generally co-localize in the same genomic regions tend to cluster across many chromosomes.

Depicted in **Figure 3.7** is the correlation between the number of genes corresponding to each meta-assemblage versus the number of significant clustering cases across the chromosomes. The Pearson correlation between these two variables is equal to 0.48 and corresponds to a p-value equal to 0.023 indicating that the higher the number of genes the higher the resulting clustering results. Focusing on the figure we took an interest in the meta-assemblages that deviate from this correlation. The most striking case, which is the "CEN" meta-assemblage acts as a positive control for our methodology. This group consists of twelve factors responsible for proper chromosomal segregation during cell division (Rossi et al. (2021)). The gene-targets of these factors are expected to be enriched in regions around the centromeres which explains the extreme clustering across almost all the chromosomes and confirms that our algorithm efficiently detects clustering (**Figure 3.8**). On the other hand the "ISO" meta-assemblages are

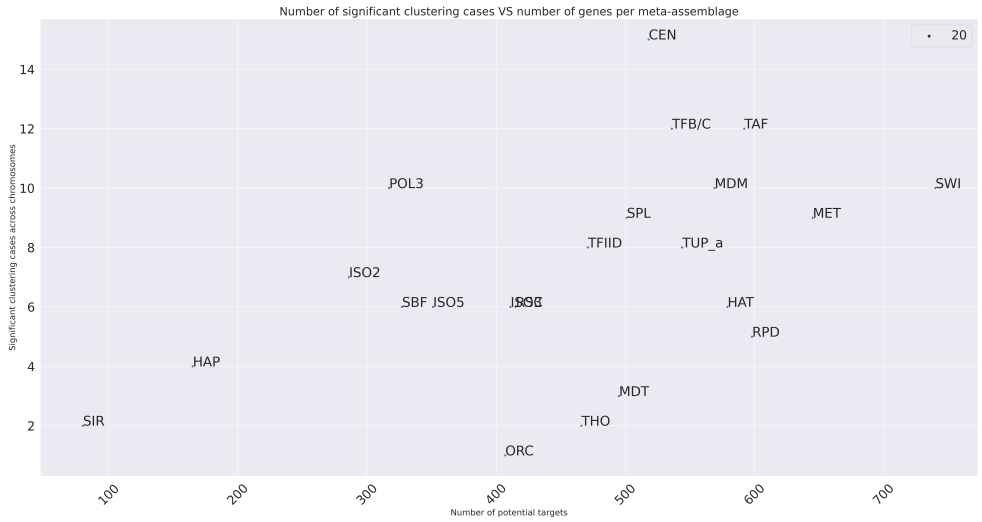


Figure 3.7: Depicted in this scatterplot is the relationship between the number of genes per meta-assemblage (x axis) against the resulting significant clustering cases across the chromosomes (y axis). The pearson correlation between these two variables is equal to 0.48 corresponding to a p-value equal to 0.023.

the groups not based on factor co-localization, acting as a negative control to our hypothesis. We observe that the clustering results that this group yields are not less than other groups with the similar number of genes implying that the absence of co-localization of factors does not necessarily correlate with lower clustering of gene-targets, unless there is another bias in that specific group as it consists of factors that mostly bind unique sites in the genome and thus did not cluster with other datasets. "ORC", "THO" and "MDT" yield the less extensive results related to their number of total genes. These categories represent factors related to the origin of replication complexes, to mRNA processing respectively and to the core mediator complex (Rossi et al. (2021)). Finally, "POL3" along with "TFBII/C" meta-assemblages consist of 18 factors regulators of tRNA transcription both yielding extensive clustering results across many chromosomes. Depicted in **Figure 2.2** are the functional categories (GO-terms) describing most of the meta-assemblages used in this analysis.

Based on the above analysis, although we detected extensive clustering across the meta-assemblages that we checked, we can not clearly attribute this clustering to the co-localization of factors partitioning the meta-assemblages. As these data are topologically biased regarding the co-localization of many factors on the genome they may also "hide", in some cases, a topological bias regarding their gene-targets as well, like the "CEN" case. Meta-assemblages are related to specific pathways and functions leading us to assume that their gene-targets as

well may be functionally related. This increases the chances of them being clustered across the genome based on previous studies as well (Lee and Sonnhammer (2003); Tiirikka et al. (2014)).

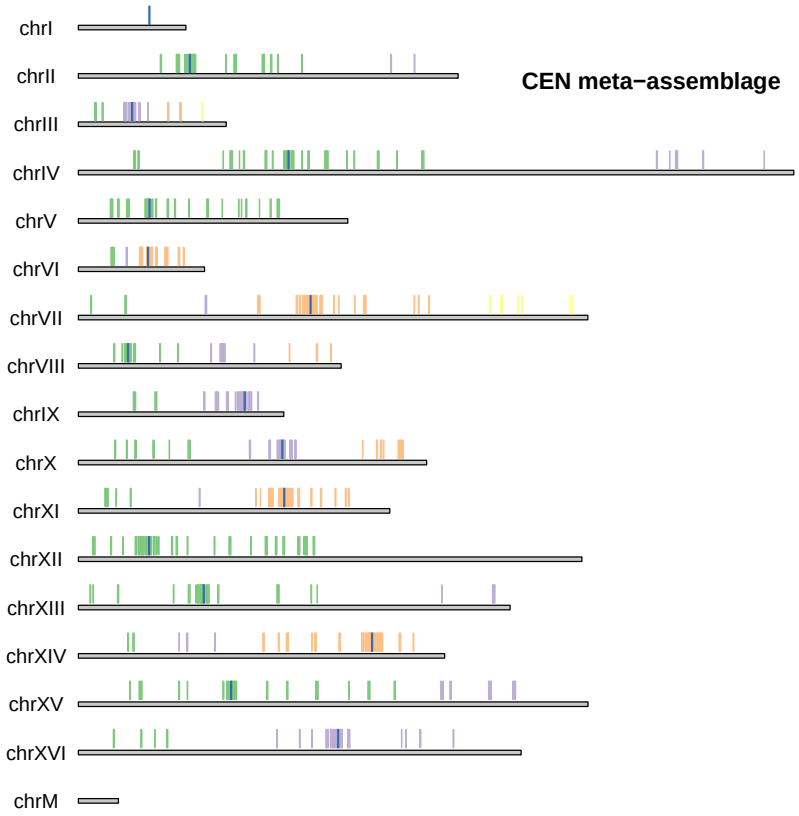


Figure 3.8: The resulting sub-clusters coming from the clustering analysis on the "CEN" meta-assemblage as provided by Rossi et al. (2021). The different colours indicate different sub-clusters. Those are clearly correlated with the centromere positioning as indicated in blue.

3.2.2 | General genomic categorizations

3.2.2.1 | Extensive clustering of essential yeast genes

Shown in the current section is the clustering analysis done on various gene categorizations described in material and methods. Depicted in **Figure 3.9** are the cases of significant clustering across chromosomes for these gene groups which are indicated below each sub-figure. As described in materials and methods, in the last step of our in-house algorithm, the positionally-clustered genes are divided into sub-clusters based on an intergenic distance threshold. These sub-clusters are indicated with different colours in the aforementioned figure.

(a) Essential (b) TATA (c) High freq expr

(d) Low freq expr (e) WGD (f) SSD

Figure 3.9: The resulting sub-clusters of positionally-clustered genes distributed across the sixteen chromosomes for various genomic categorizations. Depicted with the different colours are the genes divided into different sub-clusters while the label below each sub-figure indicates the corresponding gene group. Depicted only in 3.9c is a case of a significantly high z-score implying greater intergenic distances than expected by chance.

All these categories shown in this figure have positionally-clustered genes except for the genes with high expression frequency (high freq expr) which resulted with a rare but significantly high z-score, implying higher intergenic distances than expected by chance on chromosome seven. Such results are also found in other cases described in the next sections. The first observation is made on the essential genes which are clustered on the majority of chromosomes, a result supported by previous works in the field (Kamath et al. (2003); Pál and Hurst (2003)). Whole genome duplicates follow that trend while the rest of the categories show restricted results detected in specific chromosomes. Additionally, depicted in **Table 3.7** and **Figure 3.10** are the percentages of clustered genes per category and the density of the resulting sub-clusters calculated as the number of genes in the sub-cluster divided by the total number of genes enclosed in the same coordinates. This result suggests that functional categorizations tend to be more clustered than genes with common expression programs.

Gene class	Percentage of positionally-clustered genes
Essential	0.834
TATA	0.159
High freq expr	0.096
Low freq expr	0.214
WGD	0.488
SSD	0.205

Table 3.7: The percentage of the resulting positionally-clustered genes (except for high freq expr category) for each gene group given as input to the algorithm.

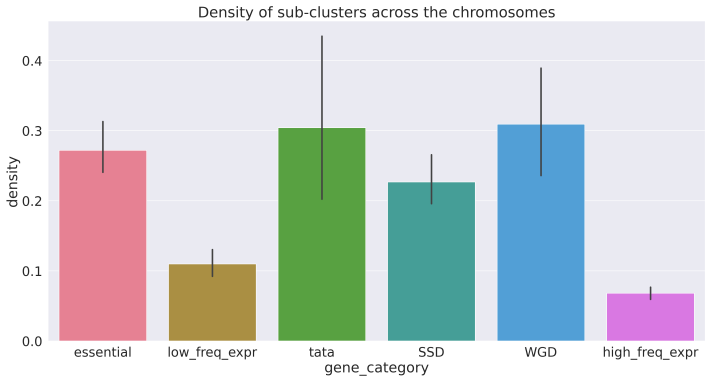


Figure 3.10: The density of the resulting sub-clusters coming from the algorithm when applied to the genomic categorization shown below each barplot.

Depicted in **Table 3.8** are the significant results of the preference/avoidance testing done on the same gene groups, as provided by our algorithm. Most of the significant results correspond to the essential genes containing both avoided and preferred chromosomes which match the enrichments shown in previous sections (**Figure 3.1**). Thus, positionally-clustered essential genes are observed also on avoided chromosomes (except for chrIII). We also observe that positionally-clustered essential genes on the other two "avoided" chromosomes (chrI and chrVIII) are less in number and thus are concentrated in a smaller part of the chromosome. TATA genes seem to prefer chrI and chrII which again match the results shown in **Figure 3.1** while the positionally-clustered results coincide with at least one preferred chromosome (chrII). Generally we observe that the preference/avoidance results do not predict the positionally-clustered cases across the chromosomes.

Finally, based on the analysis described in material and methods, we have computed enrichment values of the positionally-clustered genes for each category, in three segments of the genome (peripheral, central, intermediate) made on the basis of the genes' three-dimensional distances from a centromere pole as depicted in **Figure 2.6**. The only results supported by sta-

Gene class	Chromosome	Preference/Avoidance	Observed	Expected	P-value
Essential	chrI	avoided	11	19	0.02
Essential	chrIII	avoided	14	30	0.001
Essential	chrVIII	avoided	40	52	0.035
Essential	chrIV	preferred	161	136	0.011
Essential	chrXI	preferred	69	56	0.035
Essential	chrXIV	preferred	86	70	0.018
TATA	chrI	preferred	38	19	<=0.001
TATA	chrII	preferred	94	72	0.003
TATA	chrXV	avoided	72	97	0.002
SSD	chrXIII	avoided	64	78	0.032
WGD	chrXI	preferred	72	56	0.013

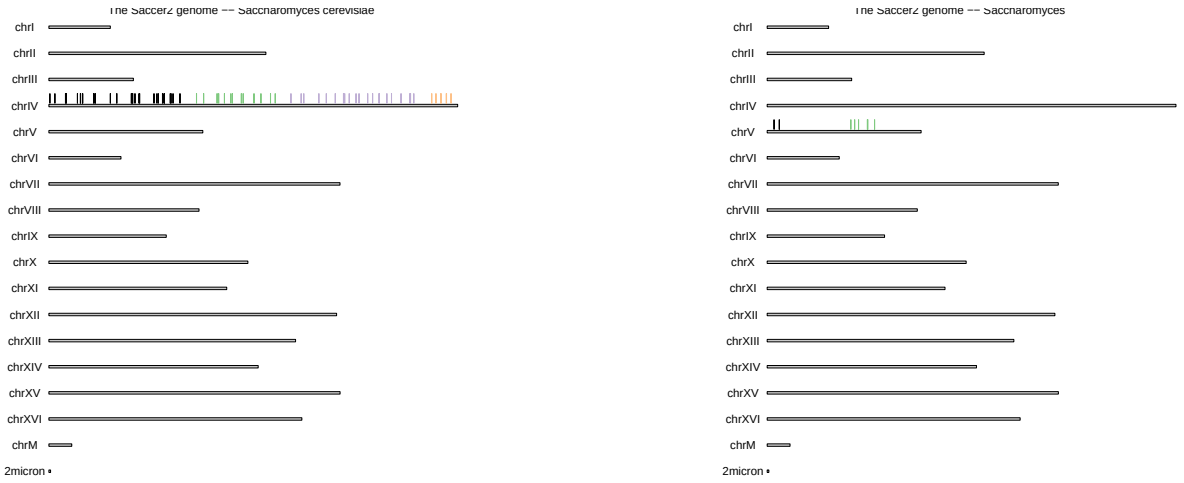
Table 3.8: Some significant preference/avoidance results as provided by our algorithm across the various genomic categorizations. P-value<=0.05

tistical significance was on the TATA genes. In more detail, positionally-clustered TATA genes are enriched in "central" chromosomal regions (enrichment=1.47, pvalue<=0.001) and depleted in the "peripheral" chromosomal regions (enrichment=0.58, pvalue<=0.001). These results suggest that clustered TATA genes are detected on two chromosomes (chrII, chrXIV) and seem to be mostly found closer to the centromeres which is contrary to the general preference of the TATA genes to be closer to telomeric regions, as described above.

3.2.2.2 | Genes of different evolutionary origins are not positionally-clustered across chromosomes

Moving on with the results, we have conducted the same analysis on genes grouped on the basis of their evolutionary origin, as described in material and methods. The abundant "Fungi" category, which contains more than 4000 genes, was excluded from this analysis. Depicted in **Figure 3.11** are the results coming from the clustering analysis conducted on the rest of the origin groups. None of the groups except for the *Saccharomyces* category, yield significantly negative z-scores implying no significant clustering for genes grouped on the basis of their evolutionary origin. The single clustering case of the *Saccharomyces* category, correspond to seven genes positioned on chromosome five (chrV) yielding a z-score equal to -2.11 (pvalue=0.018). On the contrary, the *Saccharomyces cerevisiae* category yielded a positive z-score, equal to 2.8 (pvalue<=0.001), implying a sparse positioning of the 70 genes of that category on chrIV.

Finally, depicted in **Table 3.9** are some of the most significant results regarding the preference/avoidance chromosomal tendencies of the origin groups. We observe that most of the significant results correspond to cases of chromosomal avoidance which may partly explain the absence of significant clustering across chromosomes.



(a) Saccharomyces cerevisiae (b) Saccharomyces

Figure 3.11: The results of the clustering analysis on the evolutionary origin categories. Positionally-clustered genes are detected only for the *Saccharomyces* category (on chrV). The *Saccharomyces category* yielded a significant positive z-score (equal to 2.8) on chrIV. The different colours indicate distinct sub-clusters.

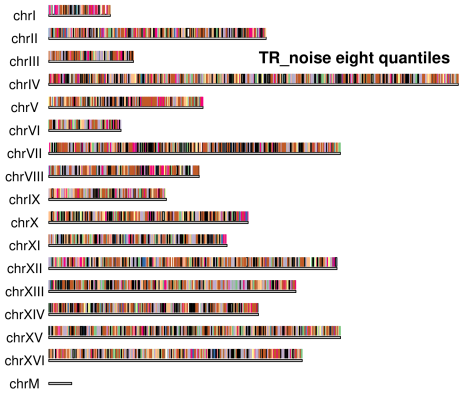
Origin	Chromosome	Preference/Avoidance	Observed	Expected	P-value
Fungi	chrI	avoided	57	81	<=0.001
Fungi	chrIV	preferred	602	565	0.005
Ascomycota	chrXII	avoided	17	26	0.027
Saccharomycetales	chrVIII	avoided	7	17	0.001
Saccharomyces cerevisiae	chrI	preferred	18	9	0.009
Saccharomyces cerevisiae	chrVI	avoided	3	11	0.002
Saccharomyces cerevisiae	chrXV	avoided	38	49	0.038

Table 3.9: Some of the most significant preference/avoidance results as provided by our algorithm across the origin categories. P-value<=0.05.

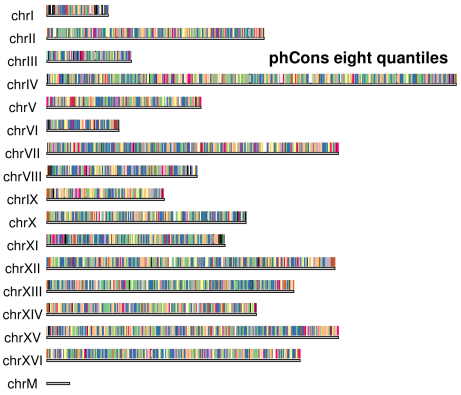
3.2.3 | Transcriptional variability and Conservation quantiles

3.2.3.1 | Gene groups of different transcriptional variability or conservation levels are positionally-clustered across specific chromosomes

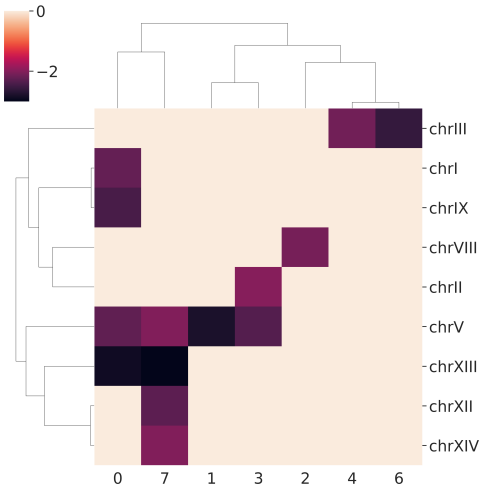
We have also performed analysis on data regarding the transcriptional variability and conservation of genes, as described in material and methods. To achieve this analysis, as mentioned before, we have divided the genes into eight equisized groups or quantiles on the basis of their transcription variability or conservation scores. We were interested in studying the positional clustering of gene groups varying in their conservation scores or transcriptional variability. Shown in the **Figures 3.12a, 3.12b** are the resulting transcriptional variability and conservation quantiles respectively, each one indicated with a different colour. Both transcriptional variability and conservation increases as the number-indicator of the quantile increases, ranging



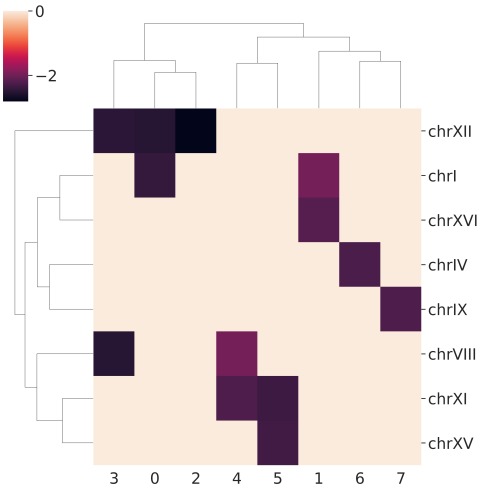
(a) Transciptional variability quantiles



(b) Conservation scores quantiles



(c) Clustering across Transcriptional variability quantiles

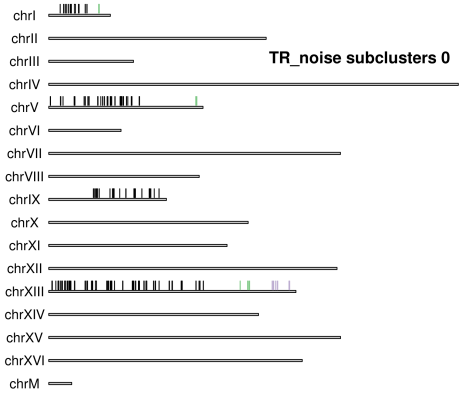


(d) Clustering across conservation quantiles

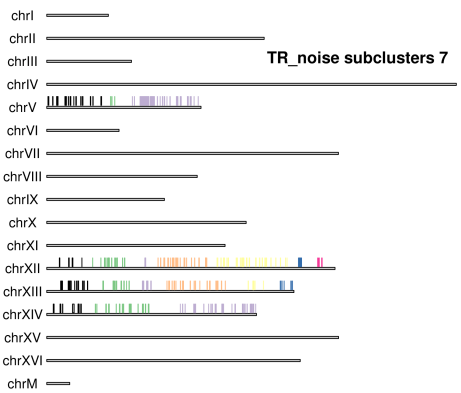
Figure 3.12: Depicted in 3.12a and 3.12b is the distribution of the transcriptional variability and conservation quantiles across the chromosomes, indicated with different colours. Shown in 3.12c and 3.12d are the significant z-scores coming from the clustering analysis conducted on the quantiles of both datasets. The increasing number in the x-axis indicate quantiles of increasing transcriptional variability (left) and conservation respectively (right). Depicted in the cells of these heatmaps are the z-scores (darker colours indicate lower z-scores) across the chromosomes (y-axis).

from zero to seven. Each transcriptional variability quantile comprised 833 genes while each conservation quantile consisted of approximately 778 genes.

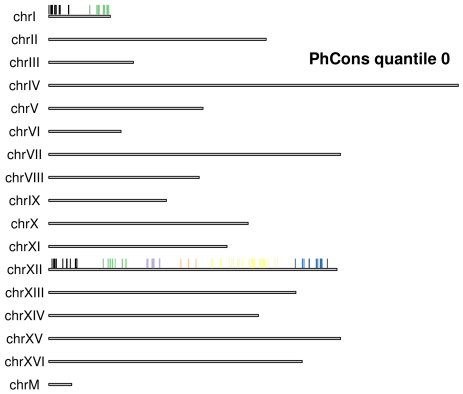
Depicted in the **Figures 3.12c, 3.12d** are the significant clustering results depicted as z-scores in a hierarchically clustered heatmap across the chromosomes (y axis) and the quantiles (x axis) for both datasets. In both cases the results are restricted in a small number of eight to nine chromosomes and the majority of the quantiles yield results in one to two chromosomes. Focusing



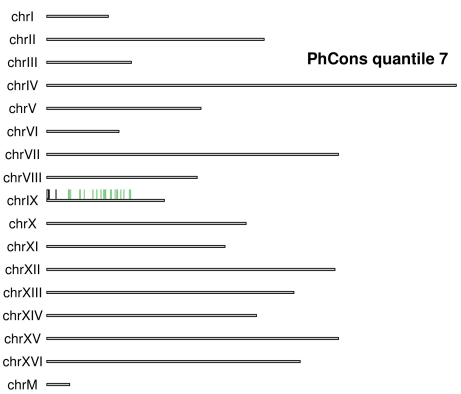
(a) Transcriptional variability - first quantile



(b) Transcriptional variability - eighth quantile



(c) Conservation - first quantile



(d) Conservation - eighth quantile

Figure 3.13: The distribution of the resulting sub-clusters of the positionally-clustered genes across the chromosomes for the first and the last transcriptional variability (up) and conservation quantiles (down). Each sub-cluster is indicated in different colours.

on the transcriptional variability results we observe that the most extensive results, across four chromosomes, correspond to the genes of the most extreme quantiles (zero and seven). We also observe that the clustering of those two quantiles, of the lowest and highest variability, coincide in the same two chromosomes (chrV, chrXII) and that most clustering results across the quantiles are accumulated in chrV. Accordingly, in the right, we observe that all conservation quantiles yielded clustering results across eight chromosomes. Clustering is more restricted in this case at most across two chromosomes. An overlap of clustering across chromosomes between consecutive quantiles is observed.

Depicted in **Figure 3.13** are the resulting sub-clusters of the clustering detected in the first and last quantile of both transcriptional variability and conservation data, as provided by our algorithm, indicated in distinct colours. We observe that genes with low transcriptional vari-

chr	observed_number	expected_number	pvalue	TRnoise-quantile
chrIX	22	31.131	0.044	0
chrV	30	42.477	0.014	0
chrVII	97	74.873	0.003	0
chrXIV	67	54.077	0.042	0
chrX	38	51.264	0.010	1
chrXV	89	75.245	0.050	1
chrIV	126	105.034	0.011	3
chrVII	55	74.614	0.006	4
chrXIII	48	64.262	0.011	4
chrIX	20	31.093	0.011	5
chrI	24	14.950	0.012	6
chrII	70	56.372	0.032	6
chrXI	33	43.337	0.050	6
chrV	64	42.287	0.001	7
chrXI	32	43.662	0.036	7

(a) Transcriptional variability - chromosomal preference/avoidance

quantile	polar_slice	count	enrichment	pvalue
0	peripheral	7	0.268325	0.000
0	intermediate	62	1.196347	0.014
0	central	35	1.341624	0.010
2	central	11	1.461732	0.046
3	central	25	1.384216	0.014
7	intermediate	88	1.139330	0.023
7	central	31	0.797308	0.031

(b) Transcriptional variability - polar preference/avoidance

chr	observed_number	expected_number	pvalue	conservation-quantiles
chrI	25	13.830	0.003	0
chrV	23	39.354	0.005	0
chrIV	119	97.798	0.009	1
chrVI	8	17.368	0.010	1
chrVIII	28	37.485	0.040	1
chrXIII	83	60.019	0.001	1
chrVI	10	17.195	0.035	2
chrX	63	47.981	0.011	2
chrXVI	81	59.722	0.002	2
chrVIII	28	37.976	0.047	3
chrI	7	14.161	0.028	4
chrVII	55	70.180	0.028	4
chrXIV	67	50.401	0.010	4
chrIV	80	97.501	0.019	5
chrIX	18	29.183	0.018	5
chrXI	53	40.347	0.022	5
chrII	64	52.579	0.045	6
chrV	51	38.898	0.022	7
chrXII	85	69.843	0.033	7

(c) Conservation - chromosomal preference/avoidance

quantile	polar_slice	count	enrichment	pvalue
0	peripheral	10	1.898353	0.003
0	central	0	0.000000	0.000
4	peripheral	12	0.629454	0.019
4	intermediate	49	1.293842	0.004
5	peripheral	39	1.274386	0.027
6	peripheral	19	0.728310	0.027
6	central	40	1.533285	0.001
7	intermediate	15	1.433411	0.037
7	central	0	0.000000	0.000

(d) Conservation - polar preference/avoidance

Figure 3.14: Four tables indicating the chromosomal preference/avoidance results (left) and the enrichment analysis across the polar segments of the genome (right) for both transcriptional variability and conservation quantiles. The significance level corresponds to a p-value equal to or less than 0.05. In figures 3.14a and 3.14c the first column corresponds to the chromosomes, the second and the third to the observed and expected number of genes per chromosome, the fourth to the p-values and the final column corresponds to the quantiles. In figures 3.14b and 3.14d the first column indicates the quantiles, the second the polar segments, the third the number of genes of the quantile positioned in that segment, the fourth the enrichment values and the last column corresponds to the p-values.

ability are not divided into many different sub-clusters implying that the distances between consecutive genes are close to the average intergenic distance of that gene group in each chromosome.

Finally, shown in **Figure 3.14** are the results of the chromosomal preference/avoidance tests (left) and the results on the enrichment analysis done across three polar segments of the yeast genome (right) for both the transcriptional variability and conservation quantiles. Based on **Figure 3.14a** we observe that the clustering results corresponding to the first transcriptional variability quantile (zero) are detected on two chromosomes that this group of genes generally avoids as the expected genes on those chromosomes (chrIX and chrV) are significantly more than the observed ones. On the other hand the last quantile (seven) seems to prefer the chrV on which its' genes are positionally-clustered as well. Although the two extreme quantiles (zero

and seven) both yield clustering results on chrV, the first avoids that chromosome while the second prefers it. We do not detect any other correlation between the preference/avoidance tests and the clustering results across the rest of the quantiles. Based on **Figure 3.14c** we observe extensive preference/avoidance results for the conservation quantiles, especially for the intermediate ones, but there is no correlation with the clustering results across chromosomes.

The enrichment analysis on polar segments is described in material and methods. In a nutshell, through this procedure we study the positional preferences of the resulting positionally-clustered genes across three segments of the genome (central, peripheral and intermediate). This is achieved by calculating enrichment values of the positionally-clustered genes across the three segments and by evaluating them trough permutation tests. The corresponding significant results of the current analysis are shown in **Figures 3.14b and 3.14d**. Based on 3.14b, positionally-clustered genes of the lower transcriptional variability quantiles are enriched in the intermediate (zero quantile) and central segments (zero,two,three quantiles) while the positionally-clustered genes of the highest trascriptional variability are enriched in the intermediate regions but depleted in the central ones. Similarly, the clustered genes of the lowest conservation (zero quantile) are enriched in the peripheral regions but are not found in the central ones, as shown in 3.14d. Generally, as the conservation increases the clustered genes are found mostly in the intermediate and central regions except for the fifth quantile whose clustered genes are enriched in the peripheral segments.

3.2.4 | Gene-ontology terms

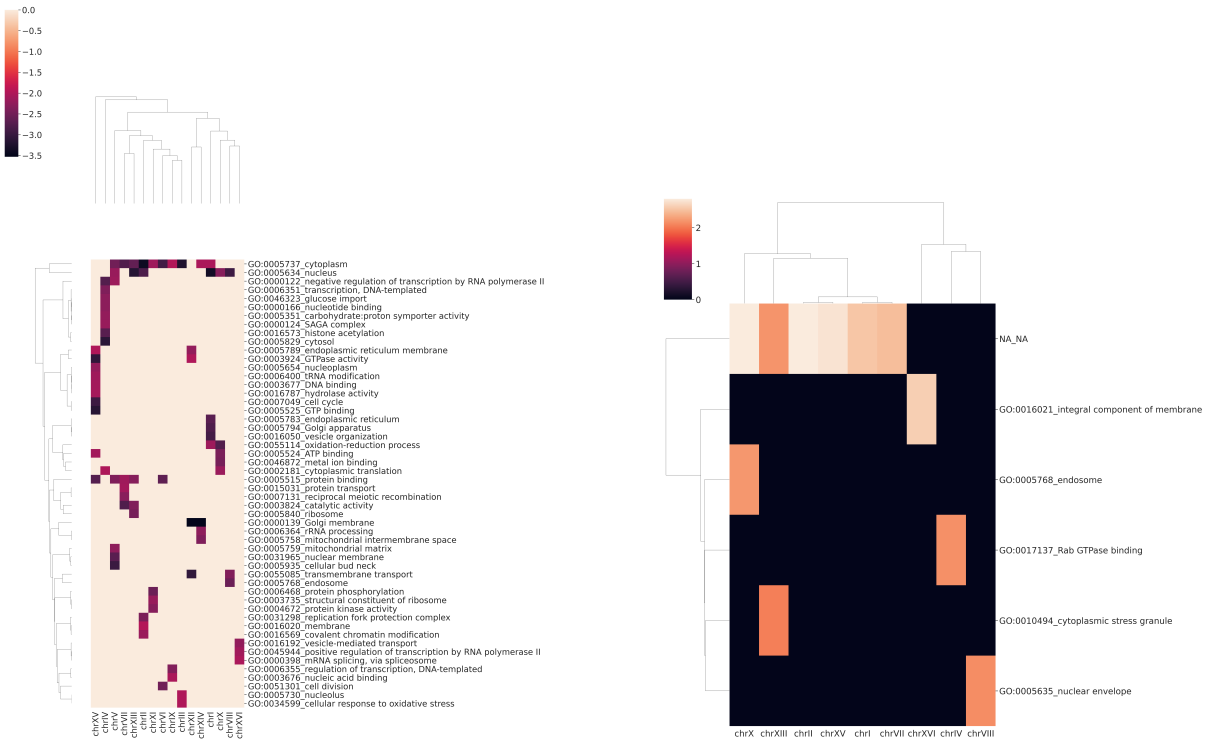
3.2.4.1 | Some Gene-ontology terms are positionally-clustered across specific chromosomes

The last gene categories that we studied in the current analysis are genes grouped on the basis of their corresponding GO-terms. Despite the high number of the GO-terms (5899), the corresponding genes are around 7000. The average number of genes embraced by a gene-ontology is equal to 8.05 genes, while the maximum number of genes in a gene-ontology is equal to 1305 and the minimum is equal to 1. Out of these 5899 terms only 1460 correspond to more than 5 genes, which are the GO-terms used in the analysis below. Based on **Table 3.10** only the 3.6% of the 1460 GO-terms yielded clustering results in at least one chromosome, which is a rather low percentage. This can be attributed to the low numbers of genes across the 1460 GO-terms. Supporting this, the computed average number of genes that correspond to the 53 GO-terms yielding clustering results is equal to 207.

Depicted in **Figure 3.15** are the resulting z-scores indicating clustering (left) or sparsity of genes (right) across chromosomes (columns) and GO-terms (rows). Shown are the z-scores

Number of significant z-scores	Number of GO-terms	Percentage over total GO-terms
80 < 0	53	3.6%
11 > 0	6	0.7%

Table 3.10: The number of significant z-scores (either positive or negative) across the chromosomes corresponding to a number of unique GO-terms. Shown in the third column is the percentage of GO-terms yielding results over the total number of GO-terms checked.

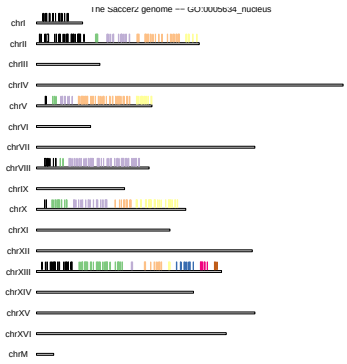


(a) Negative z-scores (b) Positive z-scores

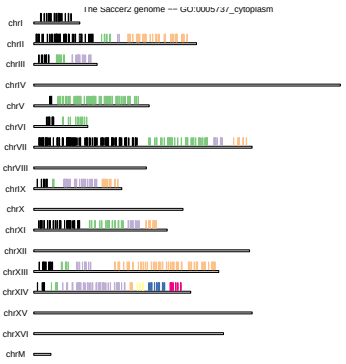
Figure 3.15: The resulting z-scores coming from the clustering analysis across GO-terms (rows) and chromosomes (columns). Shown in the left are the negative z-scores (<=-1.96) that indicate positional clustering while shown in the right are the positive z-scores (>=1.96) indicating higher than expected intergenic distances.

with an absolute value higher than or equal to the significance threshold 1.96. Focusing on the clustering cases shown in the left, we observe that most of the GO-terms yield results on a specific subset of chromosomes. The extensive clustering is detected mainly in abundant GO-terms like the "nucleus" and "cytoplasm" terms which both contain more than 1000 genes each. We also observe a tendency of clustered cases across GO-terms to be positioned on chrXV ad chrIV (10 clustering cases on each chromosome). These clustering findings are supported by previous work in the field (Tiirikka et al. (2014)).

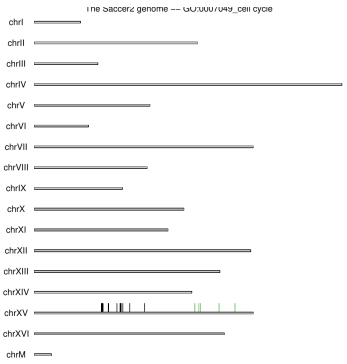
Strangely but interestingly enough, two GO-terms with more than 1000 genes each, yield



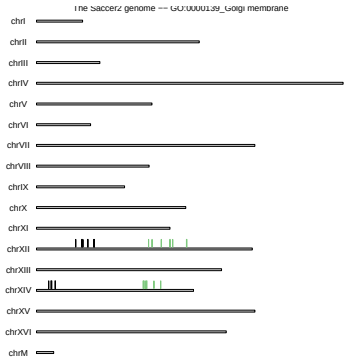
(a) "Nucleus" - Clustered



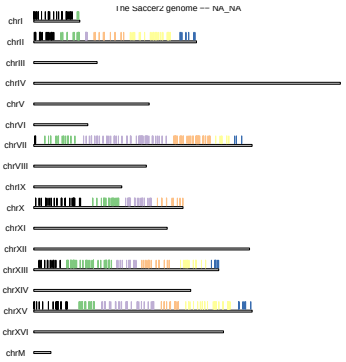
(b) "Cytoplasm" - Clustered



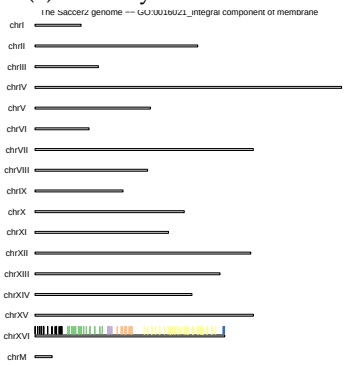
(c) "Cell-cycle" - Clustered



(d) "Golgi membrane" - Clustered



(e) NA_NA - Sparse positioning



(f) "Integral component of the membrane" - Sparse positioning

Figure 3.16: The sub-clusters coming from the clustering analysis done on genes grouped on the basis of the corresponding gene-ontology term. Indicated below each sub-figure is the corresponding GO-term along with whether the genes are positionally clustered (z-scores<=-1.96) or sparsely positioned (z-score>=1.96) across the chromosomes.

significant high z-scores indicating a more sparse positioning of genes than expected by chance (**figure 3.15b**). In more detail, the category NA_NA consists of 1239 genes that are not related to any gene ontology term and thus not related to a common functionality. Due to the high number of genes we would expect some clustering results or no significant results at all. On the contrary we observe an extensive sparse positioning of those genes across six chromosomes implying the existence of a positional bias.

In the same page, the GO-term "integral component of the membrane" which is the most abundant category with 1305 genes, yield a positive z-score (equal to 2.58) implying again a sparse positioning of those genes on chromosome chrXVI. It is noteworthy that, as mentioned before, the evolutionary gene category *S. cerevisiae* yields also a positive z-score on chrIV, implying a sparse positioning of 70 "new" genes on that chromosome. We also know (it is also

mentioned in supplementary),by doing a functional analysis on the origin categories, that the evolutionary "young" genes (*Saccharomyces* and *S. cerevisiae*) are enriched in the gene ontology term "integral component of the membrane" implying that those three categories share genes. Despite them yielding positive z-scores on different chromosomes, these results show a bias of genes related to integral components of the membrane to not cluster (or to be distantly positioned) across the genome with evolutionary newer genes following that tendency as they "acquire" such functionalities. Or else new genes may emerge from regions that tend to give birth to such functionalities and are found scattered across the genome. This notion is inspired and supported by ongoing work done by Aimilios Tassios and Nikos Vakirlis and by previous work in the field (Vakirlis et al. (2020)) in which they showed that emerging yeast orfs tend to form putative transmembrane domains potentially emerging from non-coding intergenic regions that upon hypothetical expression show a strong tendency for creating putative transmembrane domains. Finally, depicted in **Figure 3.16** are the sub-clusters of some GO-terms across the chromosomes as provided by our algorithm.

3.3 | Overlap analysis between sub-clusters of different gene-categories

In the final step of our methodology, as shown in **Figure 2.5**, an overlap analysis between all resulting sub-clusters is conducted. Aiming to identify significant overlap between the sub-clusters of different genomic categories we follow a procedure described in material and methods. Until now, by applying this algorithm on many genomic categorizations, we gained insight into the extensive positional clustering of the yeast genome. At this point, by examining the significantly overlapping sub-clusters, we hope to reveal underlying relationships between various gene categories that tend to cluster in the same regions. The combinations of different gene categories tested were approximately 19300 but only 1062 of those were statistical significant (pvalue<=0.05) corresponding to cases of significant overlap between the sub-clusters of categories. In order to examine the resulting relationships, we depicted the results through networks in which each node represents a gene category and each edge a significant overlap between two nodes. To have manageable networks, we divided the results based on the genomic categorizations while we also used different significant levels (either a pvalue<=0.05 or a pvalue<=0.01). In each sub-section below we elaborate on the resulting networks corresponding to each genomic category.

3.3.1 | General genomic categorizations

3.3.1.1 | Opposing tendencies between WGDs and SSDs

Depicted in **Figure 3.17** is the resulting network representing the significant overlaps (edges) between the sub-clusters of general genomic categorizations (blue nodes) as described above. Depicted in this figure are only the significant relationships observed between specific genomic categories (e.g WGD, SSD, TATA etc), which are represented as "central" nodes in the sub-networks, and all the other gene groups analyzed in the previous sections (e.g TFs, conservation, transcriptional variability etc.). These networks could be weighted based on the level of significance of each edge but in this study we analyze those relationships is a simpler manner.

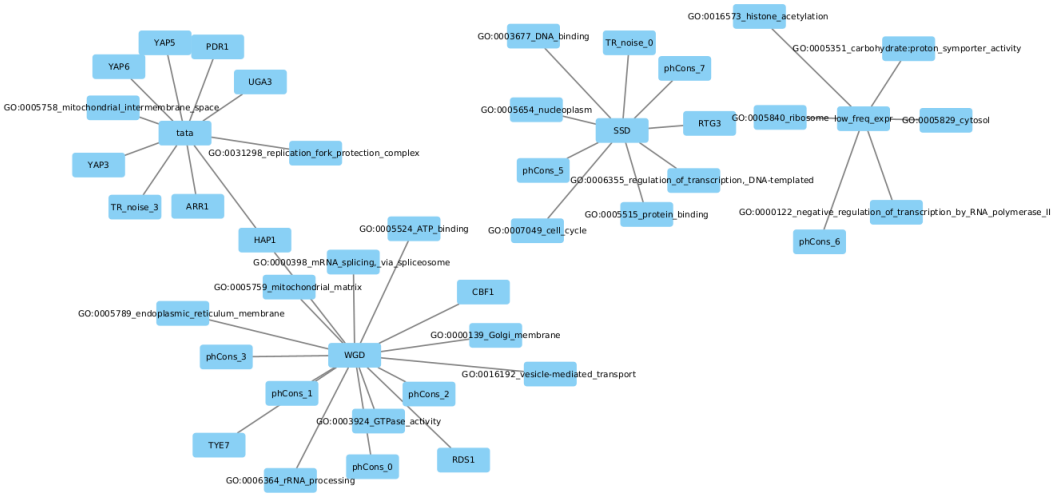


Figure 3.17: A network in which blue nodes represent the genomic categories and edges represent significant overlaps (pvalue<=0.05) between categories. The central nodes in this network are categories like "SSD", "WGD","TATA" etc. and form 45 significant overlaps with other categories.

The clustered TATA genes overlap with the clustered potential targets of factors like YAP3-5-6, ARR1 (activator of the basic leucine zipper (bZIP) family), PDR1 (pleiotropic drug response) and UGA3 (activator of GABA genes). TATA also overlap with two GO-terms, related to the mitochondrial intermembrane space and the replication fork complex, and with the clusters of an intermediate transcriptional variability quantile (Trnoise 3). Genes expressed in low frequency overlap mainly with some GO-terms related to histone acetylation or Ribosome and with highly conserved gene clusters (phCons 6). WGDs overlap with a plethora of categories. We observe that they overlap mainly with clusters of low conservation gene-groups (phCons 0-3), with GO-terms that are related to Golgi membrane, vesicle transport, rRNA Processing, mRNA splicing etc. and transcription factors like CBF1, RDS1 and TYE7. On the contrary, clustered SSDs overlap with clusters of genes that are generally highly conserved (phCons

5,7) and with clusters of genes characterized by the lowest transcriptional variability (TRnoise 0). They also overlap with some basic GO-terms related to DNA binding, protein binding, transcriptional regulation and cell cycle as well as with the RTG3 transcription factor. At this significance level we do not observe significant overlap related to the essential genes.

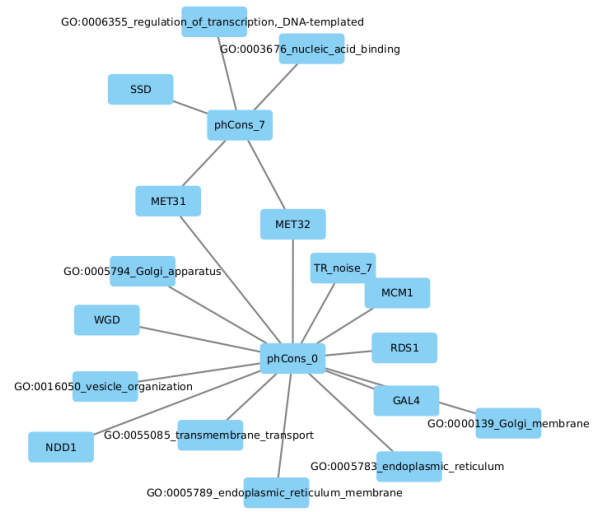
3.3.2 | Conservation and transcriptional variability quantiles

3.3.2.1 | Common tendencies between clusters of low conservation and high transcriptional variability

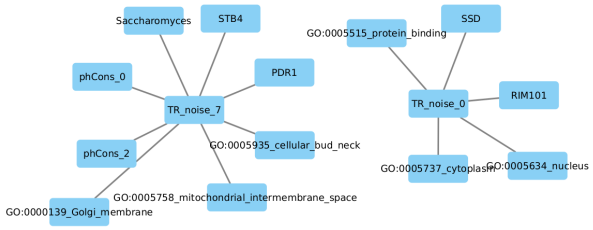
Depicted in **Figure 3.18** are the resulting networks regarding only the first and the last conservation (3.18a) and transcriptional variability (3.18b) quantiles (zero and seven quantiles). Based on **Figure 3.18a**, we observe a lot of GO-terms related to the membranes and the Golgi to significantly overlap with the low conservation clustered quantile, along with clusters of genes of the highest transcriptional variability and the WGDs. MET31 and MET32 (regulation of the methionine biosynthetic genes) overlap significantly with both the lowest and the highest conservation quantiles. Finally clusters of the highest conservation overlap with SSDs as already mentioned above, with MET31-32 and two GO-terms related to the regulation of transcription and nucleic acid binding.

Focusing on **Figure 3.18b** we observe that the clusters of the lowest transcriptional variability quantile overlap with SSDs and some basic GO-terms related to protein binding, nucleus and cytoplasm as well as with the factor RIM101 which is involved in the adaptation to alkaline conditions (as reported in SGD). On the other hand, the sub-clusters of the highest transcriptional variability overlap with gene clusters of GO-terms related to the Golgi membrane, the cellular bud neck and the mitochondrial intermembrane space as well as with low conservation gene clusters (phCons 0, 2), with the *Saccharomyces* origin category and the positionally clustered targets of PDR1 and STB4. PDR1 is involved in the pleiotropic drug response while STB4 is suggested to regulate the expression of genes encoding transporters (as reported in SGD).

In general, we observe common tendencies between the clusters of the lowest transcriptional variability and the highest conservation quantiles and between the lowest conservation and the highest transcriptional variability. The clustering of lowly conserved genes together with the clustering of membrane-related GO-terms reminds us of the relationship mentioned before, between the evolutionary younger genes and the potential transmembrane domains (mentioned above). The common clustering between low conservation and high transcriptional variability implies the existence of regions (or a single region) hosting clusters of evolutionary younger genes or very specialized genes present only in the yeast that intermingle with clusters of genes with highly variable expression patterns across different conditions.



(a) Conservation quantiles



(b) Transcriptional variability quantiles

Figure 3.18: Networks in which blue nodes represent the genomic categories and edges represent the significant overlaps (pvalue<=0.05) between categories. Shown in 3.18a, as central nodes, are the first and last conservation quantiles while in 3.18b are the first and last transcriptional variability quantiles.

3.3.3 | Gene-ontology terms

3.3.3.1 | Membrane related GO-terms overlap with less conserved and more transcriptionally variable clusters

Depicted in **Figure 3.19** are the resulting networks embracing the GO-terms' sub-clusters that significantly overlap (pvalue<=0.01) with other categories. Focusing on the wider network (up), we observe relationships that are pretty much already mentioned in the previous sections. Generally, GO-terms related to the membrane (e.g endoplasmic reticulum membrane, transmembrane transport, Golgi membrane) overlap with sub-clusters of lower conservation along with clusters of high transcriptional variability and the WGDs. On the other hand GO-terms related to basic processes like the cell cycle, DNA binding, tRNA modification, ribosome and

nucleoplasm overlap with clusters of more conserved genes. Unexpectedly the sub-clusters of RDS1, a factor involved in conferring resistance to cycloheximide (based on SGD), overlap with the GO-terms related to the cell cycle and GTPase activity.

Moving on to the networks in the middle, we observe a small network in the right embracing the GO-terms related to the nucleolus and the response to oxidative stress. Both the GO-terms overlap with regions of high transcriptional variability and the factor YDR026C or else NSI1 which is involved in the silencing of ribosomal DNA (based on SGD). Also the GO-term involved in response to stress overlaps with the targets of CAD1, a factor that is involved in stress responses, iron metabolism and pleiotropic drug resistance (based on SGD).

Based on the last small networks (down), GO-terms related to splicing, vesicle transport and endoplasmic reticulum overlap with lowly conserved gene clusters (phCons 1). The term cellular bud neck overlaps both with gene clusters of relatively low transcriptional variability (trNoise 1, 3) and with the highest variability quantile, as mentioned above (trNoise 7). The GO-term metal ion binding overlaps with the targets of both TYE7 and MSN2, with the first being involved in the activation of glycolytic genes and the latter being involved in various stress responses, based on SGD.

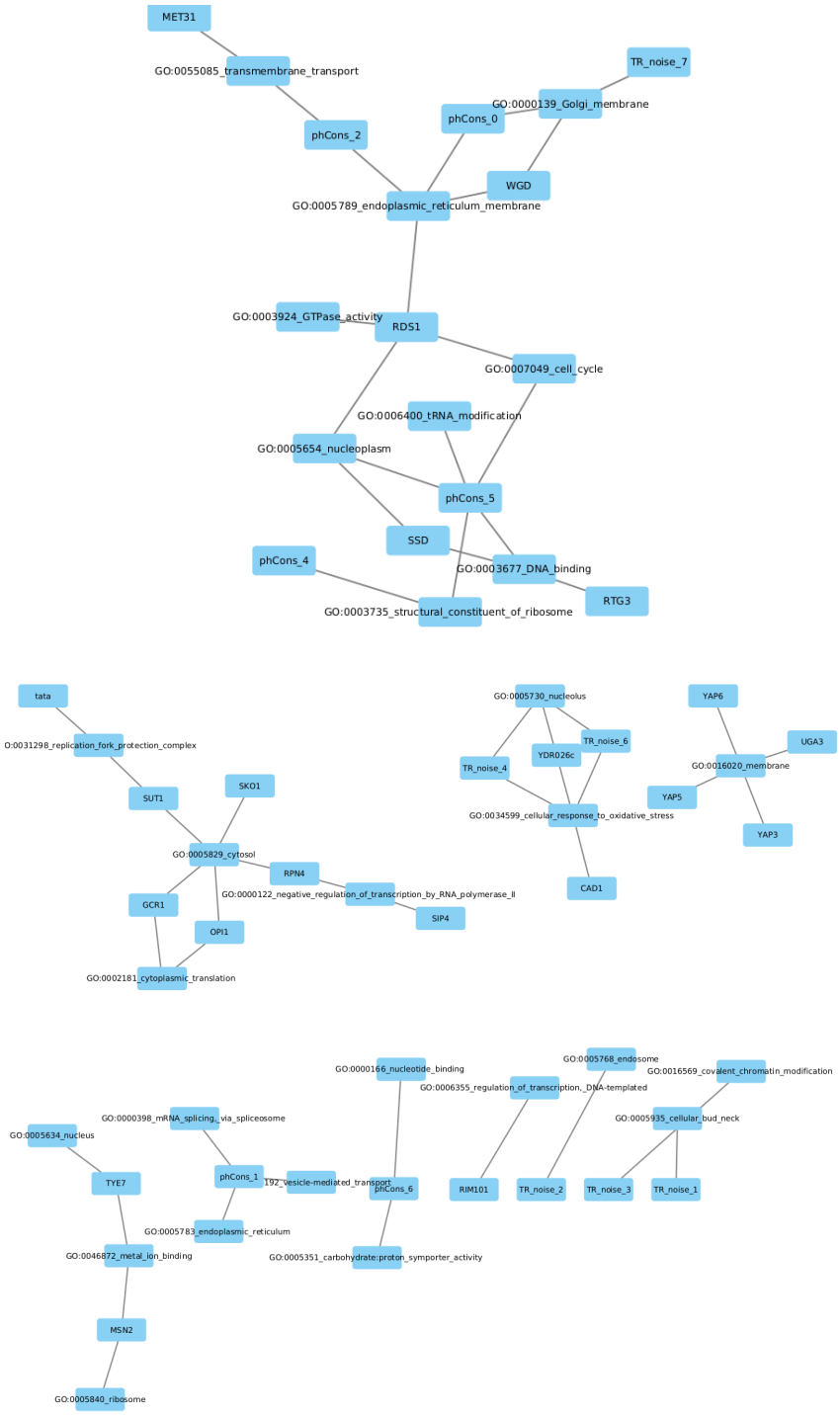


Figure 3.19: Networks in which blue nodes represent the genomic categories and edges represent the significant overlaps (pvalue<=0.01) between categories. The central nodes in this sub-networks are the GO-terms.

3.3.4 | Transcription factors

Finally, we study some of the relationships of transcription factors with the other categories. Depicted in **Figure 3.20** are the resulting networks representing the significant overlaps (pvalue<=0.01) of the transcription factors with the rest of the categories. Some of those relationships have already been described in the previous sections. Focusing on the first network (up) we observe that the clustered targets of the YAP5-6-3, ARR1 and PDR1 significantly overlap across the chromosomes which is expected as these factors have common potential targets. CAD1, which was mentioned before, overlaps with clusters of high transcriptional variability (trNoise 4, 6), with the GO-term "response to oxidative stress" and with the factor YAP1, which is also involved in the oxidative stress tolerance based on the SGD. Finally OPI1, a regulator of a variety of genes, significantly overlaps with many other categories. In more detail, OPI1's clustered targets overlap with the clustered targets of the GCR1, a factor generally involved in glycolysis , and the targets of FHL1, a regulator of ribosomal protein transcription. OPI1 overlaps also with highly conserved gene clusters (phCons 6).

Going on to the next sub-network (middle), we observe, as already mentioned, that MET31-32 along with NDD1 overlapping with gene clusters of low gene conservation. STB5 which is a factor involved in multidrug resistance and stress response overlaps with genes of intermediate conservation and low transcriptional variability. On the other hand, RDS1 overlaps with four GO-terms involved in the endoplasmic reticulum membrane, the cell cycle, the GTPase and the nucleoplasm along with the lowest conservation gene clusters. RDS1 is a factor with a restricted known role of conferring resistance to specific drugs.

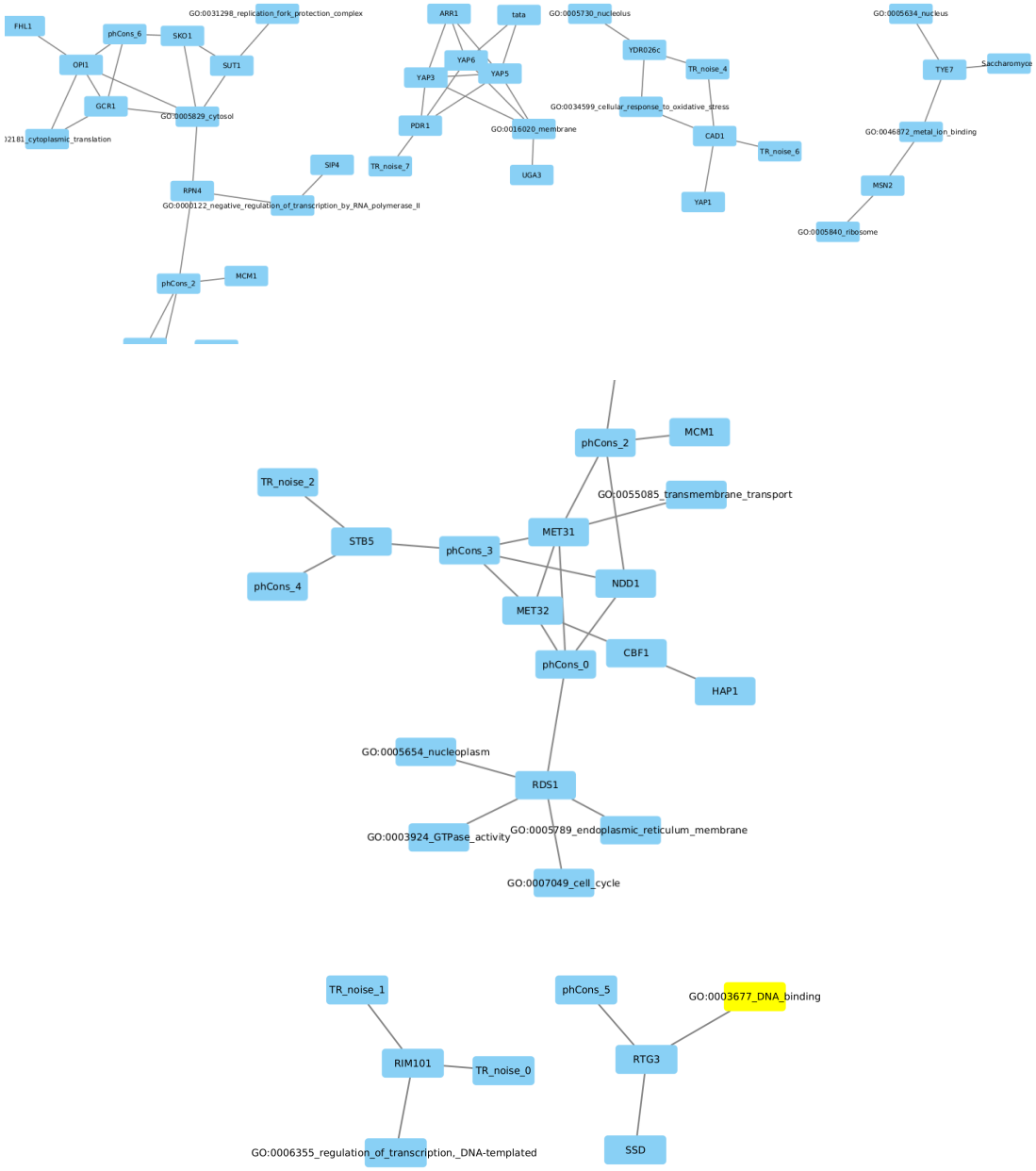


Figure 3.20: Networks in which blue nodes represent the genomic categories and edges represent the significant overlaps (pvalue<=0.01) between categories. The central nodes in this sub-networks are the transcription factors as provided by Harbison et al. (2004)

4

Supplementary results

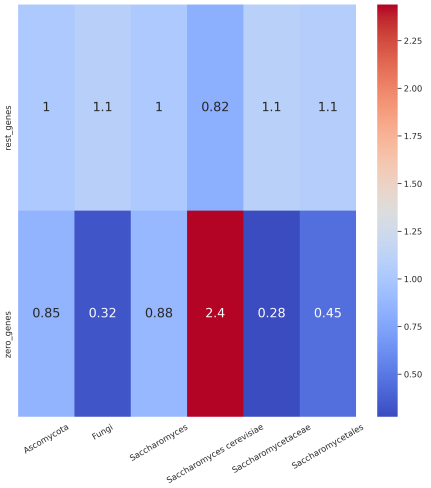
4.1 | "No-signal" genes in the Harbison et al dataset

4.1.1 | "No-signal" genes have specific positional and functional preferences potentially due to their high content in dubious elements

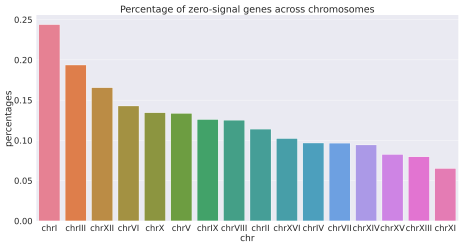
As described in materials and methods section one of the main datasets used in this analysis corresponds to a regulatory map which consists of the potential binding profiles of 102 transcription factors in yeast (Harbison et al. (2004)). 6026 gene promoter regions potentially interact with at least one transcription factor, based on this map. 768 gene promoters were found with no available interactions (no-signal genes), possibly not overlapping with any regulator's motif or with any probe used in the chip experiment.

Being curious about the 768 no-signal genes we studied more aspects about them, observing that they contain a high number of dubious ORFs, having both positional and functional preferences. Depicted in **Figure 4.1a** is a heatmap showing the enrichment of the no-signal genes and the rest of the genes in different evolutionary gene categories (materials and methods). Keeping in mind that the Fungi category indicates genes of the "oldest" origin and the "*Saccharomyces cerevisiae*" genes of the "younger" origin, it is easily observed that the no-signal genes are highly enriched in young genes which was also found to be statistically significant (pvalue<=0.01), through permutation tests. It is important to note that out of the 768 no-signal genes, we owned origin information for only the 377 genes possibly leaving out the dubius orfs. Out of those 377 no-signal genes, 150 genes are of the *S. cerevisiae* origin, significantly enriched (pvalue = 1.66E-026) in the GO-term "integral component of membrane" with 57 genes corresponding to the intersection, as provided by the gprofiler, (Raudvere et al. (2019)).

Based on the **Figure 4.1b** it is also observed that the no-signal genes are mostly found in the



(a) Antiquity of no-signal genes



(b) Percentage of no-signal genes across chromosomes

Figure 4.1: Depicted in 4.1a are the enrichment values of various evolutionary origin categories (as provided by Niko Vakirlis) across the no-signal genes and the rest of the genes used in the Harbison et al. (2004) dataset. Shown in 4.1b is the percentage of no-signal genes across the chromosomes.

smaller and less gene-rich chromosomes like chrI, chrIII and chrVI and in chrXII which is the one containing the rRNA genes also forming the nucleolus. Looking into the positional preferences of the no-signal genes with more detail, we found that they are positioned significantly closer to the telomeres (pvalue=3.83e-14) and further from the centromeres (pvalue=4.23e-22) while also being close to origin of replication sites (pvalue=4.60e-05) compared to the rest of the genes. We checked if this specific positioning is a characteristic of the evolutionary younger genes of the "*S. cerevisiae*" origin in which no-signal genes are enriched to, but we did not detect any significant positional tendency for this group of genes when we compared it to the other origin categories. Nevertheless, in the previous sections we note that a group of genes (quantile) positioned closer to telomeres are significantly enriched in younger genes. In the same manner, we compared both the phCons values and the transcriptional variability of the no-signal genes to the rest of the genes and we found that the no-signal genes have significantly lower phCons values (pvalue=3.51e-15) and higher transcriptional variability (pvalue=2.32e-08). All those statistical comparisons were completed by using the non parametric Mann–Whitney test (`https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.mannwhitneyu.html`).

Based on the analysis above, we observe that there is a special group of genes not included in the Harbison et al analysis. This may be due to the high percentage of dubious orfs or to

the young genes existing in this group which may not own consolidated transcriptional regulation mechanisms and thus are not included in the initial analysis. Nevertheless, these genes seem to have specific positional and functional characteristics. They are close to the telomeres, considered "recent" regarding their evolutionary age, thus being less conserved and they are more varied in their expression. Also, they are preferably positioned on specific chromosomes which are both smaller and less gene-rich. In the next subsection we tried to enlighten their potential interactions with trascription factors by scanning the genome based on the factors' position specific scoring matrices (PSSMs).

4.1.2 | "No-signal" genes are potentially regulated by a specific group of transcription factors

In order to determine if there are potential binding motifs of transcription factors overlapping with the promoters of the no-signal genes, we decided to create the positional specific scoring matrices (PSSM) of factors as provided by the Jaspar database (Sandelin et al. (2004)) and scan the whole yeast genome. This procedure is explained in more detail in materials and methods. In a nutshell, by setting thresholds on the basis of the maximum PSSM score per factor, we obtained regions highly matching the binding motifs of the transcription factors. Finally, out of those regions we kept only those overlapping with the promoter regions (as provided by Harbison et al) of the genes existing in the initial dataset. Depicted in **Figure 4.2** is the number of binding motifs overlapping with promoter regions of the no-signal genes (pink) and the rest of the genes (green). These numbers may include cases of binding motifs of a single factor found more than once in a promoter region. Based on our analysis, the number of potential regulators between the no-signal genes and the rest of the genes do not significantly differ.

To find out if there are prevailing factors regarding the regulation of the no-signal genes we conducted an enrichment analysis by which we obtained enrichment values of the various transcription factors' motifs, as provided by the Jaspar database, overlapping the no-signal genes' promoters or the rest of the genes' promoters. These results are depicted in the heatmap of the **Figure 4.3**. It is observed that there is a small fraction of transcription factors that their binding motifs tend to be mostly found around the promoters of no-signal genes. After verifying that these high enrichments are supported by an adequate number of promoter, we have also conducted statistical significance testing through random permutations (results not shown) yielding a high number of significant enrichments, although this number is expected to drop after multitest correction. Some of the transcription factors yielding a significantly high enrichment score in the no-signal genes are the RSC30, a component of the RSC chromatin remodelling complex, STB5, involved in oxidative stress response, FHL1, a transcriptional regulator of ribosomal proteins and finally STP3, a protein of unknown function. The informa-

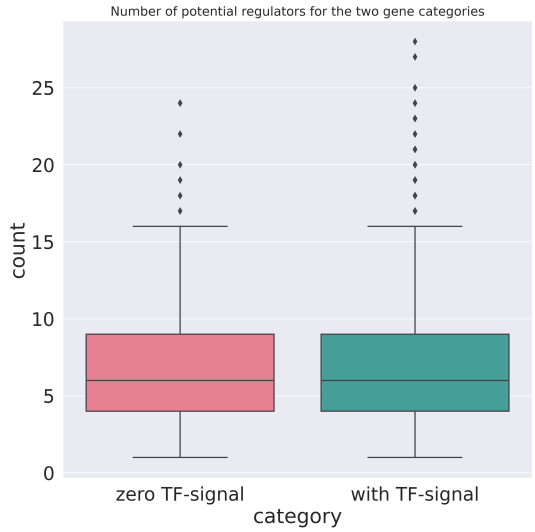


Figure 4.2: Depicted in these boxplots is the number of transcription factor motifs overlapping with the promoters of the no-signal genes (pink) and the rest of the genes (green) as provided by Harbison et al. (2004).

tion on those transcription factors were obtained from the Saccharomyces Cerevisiae Database (http://www.yeast-genome.org).

In conclusion, our analysis showed that the no-signal genes are potentially regulated by just as many transcription factors as the other genes, but it seems that there is a specific group of factors whose binding motifs are highly enriched in those genes. A potential explanation to this may be that the tendency of no-signal genes to be positioned close to telomeres may correlate with the binding of factors that "prefer" to regulate genes positioned near the telomeres. As described by Janga et al. (2008), transcription factors seem to have both chromosomal preferences and preferences on specific regions of the chromosomes.

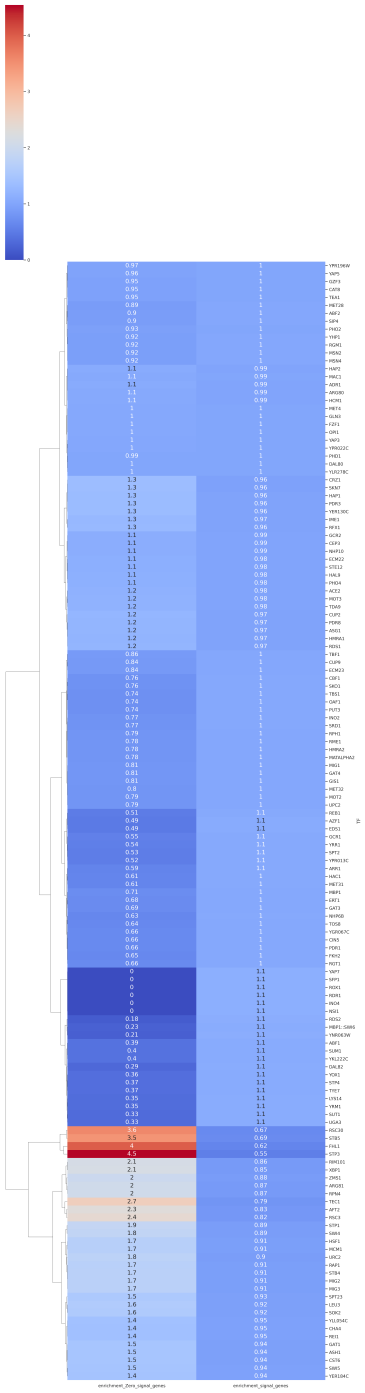


Figure 4.3: A heatmap that shows the enrichment values of transcription factors' motifs overlapping with the no-signal genes' promoters (left) and the rest of the genes' promoters (right). Red indicates higher enrichment values than blue colour.

5

Conclusions & Discussion

Our results from the initial enrichment analysis conducted at intra-chromosomal level indicate a topological and functional compartmentalization of the yeast genome that was in accordance with Tsochatzidou et al. (2017). Centromeric regions seem to be enriched in essential functions while, when moving closer to the telomeres, the 'landscape' changes. Telomeric regions are mostly enriched in evolutionary younger, TATA-regulated and transcriptionally variable genes, while under-enriched in essential functions and ancient genes. This matches the 'picture' described in Tsochatzidou et al. (2017), with telomeric regions being less conserved and mostly related to stress responses and secondary functions expected to be transcriptionally variable in different conditions. On the other hand, essential functions seem to strongly avoid telomeric ones, as also noted in Batada and Hurst (2007). On the chromosomal level, we observed that chromosomes have slightly different enrichment patterns across the same genomic categories, with chrI, the smallest chromosome, being the most striking example, implying that chromosomes may serve as distinct 'niches' for different gene categories. We also report on a topological segregation of the duplicate genes, as small scale duplicates (SSDs) are mostly found near the telomeres, while avoiding the regions closer to the centromeres, with whole genome duplicates (WGDs) having opposite tendencies. This division may reflect their distinct functional roles supported by the fact that SSDs are linked to neo-functionalization, while WGDs are linked to the sub-functionalization of ancestral functions (Fares et al. (2013)).

In the second part of our analysis, we have shown that the yeast genome is extensively clustered at the linear level. As described in Janga et al. (2008), we have also found that the potential targets of transcription factors are positionally clustered in specific chromosomes, although the chromosomal preferences of the factors do not correlate with the clustering. The most extensive clustering was reported on the analysis done on the meta-assemblages (Rossi et al. (2021)), each indicating a different regulatory architecture related to specific functions. Although this dataset may 'hide' positional biases, we assume that the co-localization of many

functionally-related factors in similar genomic regions reflect a co-regulation of common gene-targets that tend to be positioned closer to each other than compared to the gene-targets of a single transcription factor. Many studies claim that functionally-related, co-expressed genes are clustered in the genome (Cho et al. (1998); Lee and Sonnhammer (2003); Tiirikka et al. (2014)).

Essential genes are positionally clustered across almost all chromosomes, which agrees with previous works in the field (Batada and Hurst (2007); Pál and Hurst (2003)). This group of genes, which are potentially some of the most conserved genes, are found to be positionally clustered from yeast (essential) to humans (housekeeping) (Lercher et al. (2002)), implying that their clustering ensures their coordinated regulation. As described in Batada and Hurst (2007), essential genes in yeast may cluster in regions of open chromatin to minimize the effect of transcriptional bursting coming from the neighbouring genes. On the other hand, TATA genes are found to be clustered across only two chromosomes, which may reflect their positional preference near the less gene-dense telomeric regions (Basehoar et al. (2004)).

By testing the tendency of genes to cluster on the basis of transcriptional variability or conservation level, we have found that they are clustered across these gradients. Interestingly, we have detected more extensive clustering of genes with extreme transcriptional variability (either low or high) with them also overlapping significantly in chromosomes V and XIII. This implies that extreme values of transcriptional variability (either high or low) may constrain the positioning of genes across the chromosomes. Out of the 1460 gene ontology terms examined, only 3.6% were found to be positionally clustered across the chromosomes potentially due to the low number of genes participating in each term. These results partly agree with previous results (Lee and Sonnhammer (2003); Tiirikka et al. (2014)) in which they have shown that functionally-related genes (GO-terms, KEGG) are found clustered across many chromosomes.

Interestingly enough, we have also got some results indicating sparse positioning of genes (opposite of clustering). Based on our analysis, genes expressed in high frequency (high_freq_expr), genes of *Saccharomyces cerevisiae* origin and some GO-terms including the 'integral component of the membrane' were some of the cases found to be significantly sparsely positioned across some chromosomes. The most striking case is the extensive sparse positioning across six chromosomes of genes that cannot be categorized in any known GO-term. This implies that the genes of that category, potentially enriched in dubious orfs or pseudogenes, are positionally biased in such a way that they are found further from each other than expected by chance, although their number is not negligible. Nevertheless, the most interesting part is that two linked categories, the GO-term 'integral component of the membrane' and the evolutionary younger genes, enriched in the aforementioned GO-term, are not only positionally-clustered in any chromosome, but also tend to be sparsely positioned across the chromosomes. The above results seem to support and at the same time be inspired by the observations described in Vakirlis et al. (2020), in which it was claimed that emerging yeast orfs, potentially emerging

from non-coding intergenic regions, tend to form putative transmembrane domains.

The overlap analysis among the resulting sub-clusters revealed interesting relationships. We observed unexpected tendencies for the clustered SSDs and WGDs, with the latter overlapping with clusters with low gene conservation and with GO-terms related to Golgi-membrane, rRNA processing or mRNA splicing. On the contrary, the SSDs overlap with clusters of highly conserved genes, low transcriptional variability and with basic GO-terms, related to protein-binding and the cell cycle, which do not match the general tendency of SSDs to be positioned closer to the telomeres and to be linked to neo-functionalization (Fares et al. (2013)). Similarly, the clustered WGDs were expected to overlap with conserved categories and ancient functions. This implies that the clustered SSDs and WGDs may have opposite tendencies to the rest of the genes in each category, but also seem to maintain their opposing relationship. We generally observed a correlation between the overlaps found for clusters of conserved and low in transcriptional variability genes and between unconserved and highly variable gene clusters. Considering this as the genomic background, various GO-terms related to the membrane and the Golgi was found overlapping with the unconserved but highly variable clusters implying a tendency of those functions to be positioned closer to the telomeres.

Aiming to build on previous results of our lab (Nikolaou (2018); Tsochatzidou et al. (2017)), we decided to study the existence of an underlying genomic architecture in yeast that is founded on the topological and functional compartmentalization of genes. The current analysis conducted on various genomic categorizations, showed that the clustering of genes on the chromosomal level is commonly found in yeast, with some categories being more extensively clustered than others. The overlaps between all the resulting sub-clusters indicate that there may be genomic regions that play the role of 'control hubs' related to specific functionalities, while also being specifically positioned in the genome in a way that would be interesting to be depicted as an interactive map. Although improvements on the current methodology can always be made, these results confirm previous observations made on the yeast genome architecture and enrich our knowledge in the field. The secrets of the yeast genome's underlying architecture still remain to be studied.

References

Christos Andreadis, Christoforos Nikolaou, George S. Fragiadakis, Georgia Tsiliki, and Despina Alexandraki. Rad9 interacts with Aft1 to facilitate genome surveillance in fragile genomic sites under non-DNA damage-inducing conditions in S. cerevisiae. *Nucleic Acids Research*, 42(20):12650–12667, 10 2014. ISSN 0305-1048. doi: 10.1093/nar/gku915.

J.David Barrass and Jean D. Beggs. Splicing goes global. *Trends in Genetics*, 19(6):295–298, 2003. ISSN 0168-9525. doi: https://doi.org/10.1016/S0168-9525(03)00091-X.

Andrew D. Basehoar, Sara J. Zanton, and B. Franklin Pugh. Identification and distinct regulation of yeast TATA box-containing genes. *Cell*, 116(5):699–709, 2004. ISSN 00928674. doi: 10.1016/S0092-8674(04)00205-3.

Nizar N Batada and Laurence D Hurst. Evolution of chromosome organization driven by selection for reduced gene expression noise. *Nature genetics*, 39(8):945–949, 2007.

Thomas Blumenthal, Donald Evans, Christopher D. Link, Alessandro Guffanti, Daniel Lawson, Jean Thierry-Mieg, Danielle Thierry-Mieg, Wei Lu Chiu, Kyle Duke, Moni Kiraly, and Stuart K. Kim. A global analysis of Caenorhabditis elegans operons. *Nature*, 417(6891):851–854, 2002. ISSN 00280836. doi: 10.1038/nature00831.

David Botstein and Gerald R Fink. Yeast: An Experimental Organism for 21st Century Biology. *Genetics*, 189(3):695–704, 11 2011. ISSN 1943-2631. doi: 10.1534/genetics.111.130765.

Alexander M. Boutanaev, Alla I. Kalmykova, Yuri Y. Shevelyov, and Dmitry I. Nurminsky. Large clusters of co-expressed genes in the Drosophila genome. *Nature*, 420(6916):666–669, 2002. ISSN 00280836. doi: 10.1038/nature01216.

Donna Garvey Brickner, Carlo Randise-Hinchliff, Marine Lebrun Corbin, Julie Ming Liang, Stephanie Kim, Bethany Sump, Agustina D'Urso, Seo Hyun Kim, Atsushi Satomura, Heidi Schmit, Robert Coukos, Subin Hwang, Raven Watson, and Jason H. Brickner. The Role of Transcription Factors and Nuclear Pore Proteins in Controlling the Spatial Organization of the Yeast Genome. *Developmental Cell*, 49(6):936–947.e4, 2019. ISSN 18781551. doi: 10.1016/j.devcel.2019.05.023.

Raymond E Chen and Jeremy Thorner. Function and regulation in mapk signaling pathways: lessons learned from the yeast saccharomyces cerevisiae. *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research*, 1773(8):1311–1340, 2007.

Raymond J Cho, Michael J Campbell, Elizabeth A Winzeler, Lars Steinmetz, Andrew Conway, Lisa Wodicka, Tyra G Wolfsberg, Andrei E Gabrielian, David Landsman, David J Lockhart, et al. A genome-wide transcriptional analysis of the mitotic cell cycle. *Molecular cell*, 2(1):65–73, 1998.

Claudio De Virgilio and Robbie Loewith. The tor signalling network from yeast to man. *The international journal of biochemistry & cell biology*, 38(9):1476–1481, 2006.

Joseph L DeRisi, Vishwanath R Iyer, and Patrick O Brown. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, 278(5338):680–686, 1997.

Sally Dow, Ankuta Lucau-danila, Keith Anderson, Adam P Arkin, Anna Astromoff, Mohamed El Bakkoury, Rhonda Bangham, Rocio Benito, Sophie Brachat, Bruno Andre, Daniel F Jaramillo, Diane E Kelly, Steven L Kelly, and Peter Ko. Functional profiling of the Saccharomyces cerevisiae genome. *Nature*, (418):387–391, 2002.

Zhijun Duan, Mirela Andronescu, Kevin Schutz, Sean McIlwain, Yoo Jung Kim, Choli Lee, Jay Shendure, Stanley Fields, C Anthony Blau, and William S Noble. A three-dimensional model of the yeast genome. *Nature*, 465 (7296):363–367, 2010.

Mario A. Fares, Orla M. Keane, Christina Toft, Lorenzo Carretero-Paulet, and Gary W. Jones. The Roles of Whole-Genome and Small-Scale Duplications in the Functional Specialization of Saccharomyces cerevisiae Genes. *PLoS Genetics*, 9(1), 2013. ISSN 15537390. doi: 10.1371/journal.pgen.1003176.

A. Goffeau, B. G. Barrell, H. Bussey, R. W. Davis, B. Dujon, H. Feldmann, F. Galibert, J. D. Hoheisel, C. Jacq, M. Johnston, E. J. Louis, H. W. Mewes, Y. Murakami, P. Philippsen, H. Tettelin, and S. G. Oliver. Life with 6000 genes. *Science*, 274(5287):546–567, 1996. doi: 10.1126/science.274.5287.546.

Christopher T Harbison, D Benjamin Gordon, Tong Ihn Lee, Nicola J Rinaldi, Kenzie D Macisaac, Timothy W Danford, Nancy M Hannett, Jean-Bosco Tagne, David B Reynolds, Jane Yoo, et al. Transcriptional regulatory code of a eukaryotic genome. *Nature*, 431(7004):99–104, 2004.

Leland H Hartwell. Nobel lecture: Yeast and cancer. *Bioscience reports*, 22(3-4):373–394, 2002.

Matthew A. Hibbs, David C. Hess, Chad L. Myers, Curtis Huttenhower, Kai Li, and Olga G. Troyanskaya. Exploring the functional landscape of gene expression: Directed search of large microarray compendia. *Bioinformatics*, 23 (20):2692–2699, 2007. ISSN 13674803. doi: 10.1093/bioinformatics/btm403.

Dirar Homouz and Andrzej S. Kudlicki. The 3D Organization of the Yeast Genome Correlates with Co-Expression and Reflects Functional Relations between Genes. *PLoS ONE*, 8(1), 2013. ISSN 19326203. doi: 10.1371/journal.pone.0054699.

Kathryn L. Huisinga and B. Franklin Pugh. A genome-wide housekeeping role for tfiid and a highly regulated stress-related role for saga in saccharomyces cerevisiae. *Molecular Cell*, 13(4):573–585, 2004. ISSN 1097-2765. doi: https://doi.org/10.1016/S1097-2765(04)00087-5.

Laurence D. Hurst, Csaba Pál, and Martin J. Lercher. The evolutionary dynamics of eukaryotic gene order. *Nature Reviews Genetics*, 5(4):299–310, 2004. ISSN 14710056. doi: 10.1038/nrg1319.

Sarath Chandra Janga, Julio Collado-Vides, and M Madan Babu. Transcriptional regulation constrains the organization of genes on eukaryotic chromosomes. *Proceedings of the National Academy of Sciences*, 105(41):15761–15766, 2008.

Hyelim Jo, Taemook Kim, Yujin Chun, Inkyung Jung, and Daeyoup Lee. A compendium of chromatin contact maps reflecting regulation by chromatin remodelers in budding yeast. *Nature Communications*, 12(1):1–11, 2021. doi: 10.1038/s41467-021-26629-6.

Ravi S. Kamath, Andrew G. Fraser, Yan Dong, Gino Poulin, Richard Durbin, Monica Gotta, Alexander Kanapin, Nathalie Le Bot, Sergio Moreno, Marc Sohrmann, David P. Welchman, Peder Zipperien, and Julie Ahringer. Systematic functional analysis of the Caenorhabditis elegans genome using RNAi. *Nature*, 421(6920):231–237, 2003. ISSN 00280836. doi: 10.1038/nature01278.

Jan Krefting, Miguel A Andrade-Navarro, and Jonas Ibn-Salem. Evolutionary stability of topologically associating domains is associated with conserved gene regulation. *BMC biology*, 16(1):1–12, 2018.

S Kruglyak and H Tang. Regulation of adjacent yeast genes. *Trends in Genetics*, 16(3):109–111, 2000.

Oleksii Kuchaiev and Nataša Pržulj. Integrative network alignment reveals large regions of global network similarity in yeast and human. *Bioinformatics*, 27(10):1390–1396, 2011.

Georg Kustatscher, Piotr Grabowski, and Juri Rappsilber. Pervasive coexpression of spatially proximal genes is buffered at the protein level. *Molecular Systems Biology*, 13(8):937, 2017. ISSN 1744-4292. doi: 10.15252/msb.20177548.

Deval A Lashkari, Joseph L DeRisi, John H McCusker, Allen F Namath, Cristl Gentile, Seung Y Hwang, Patrick O Brown, and Ronald W Davis. Yeast microarrays for genome wide parallel genetic and gene expression analysis. *Proceedings of the National Academy of Sciences*, 94(24):13057–13062, 1997.

Jennifer M. Lee and Erik L.L. Sonnhammer. Genomic gene clustering analysis of pathways in eukaryotes. *Genome Research*, 13(5):875–882, 2003. ISSN 10889051. doi: 10.1101/gr.737703.

Martin J. Lercher, Araxi O. Urrutia, and Laurence D. Hurst. Clustering of housekeeping genes provides a unified model of gene order in the human genome. *Nature Genetics*, 31(2):180–183, 2002. ISSN 10614036. doi: 10.1038/ng887.

Erez Lieberman-Aiden, Nynke L. van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, Bryan R. Lajoie, Peter J. Sabo, Michael O. Dorschner, Richard Sandstrom, Bradley Bernstein, M. A. Bender, Mark Groudine, Andreas Gnirke, John Stamatoyannopoulos, Leonid A. Mirny, Eric S. Lander, and Job Dekker. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326(5950):289–293, 2009. doi: 10.1126/science.1181369.

Kenzie D. MacIsaac, Ting Wang, D. Benjamin Gordon, David K. Gifford, Gary D. Stormo, and Ernest Fraenkel. An improved map of conserved regulatory sites for Saccharomyces cerevisiae. *BMC Bioinformatics*, 7:1–14, 2006. ISSN 14712105. doi: 10.1186/1471-2105-7-113.

Tom Misteli. Spatial positioning: A new dimension in genome function. *Cell*, 119(2):153–156, 2004. ISSN 00928674. doi: 10.1016/j.cell.2004.09.035.

Christoforos Nikolaou. Invisible cities: segregated domains in the yeast genome with distinct structural and functional attributes. *Current Genetics*, 64(1):247–258, 2018. ISSN 14320983. doi: 10.1007/s00294-017-0731-6.

Csaba Pál and Laurence D. Hurst. Evidence for co-evolution of gene order and recombination rate. *Nature Genetics*, 33(3):392–395, 2003. ISSN 10614036. doi: 10.1038/ng1111.

Dina Petranovic and Jens Nielsen. Can yeast systems biology contribute to the understanding of human disease? *Trends in biotechnology*, 26(11):584–590, 2008.

Dina Petranovic, Keith Tyo, Goutham N Vemuri, and Jens Nielsen. Prospects of yeast systems biology for human health: integrating lipid, protein and energy metabolism. *FEMS yeast research*, 10(8):1046–1059, 2010.

Maxime Pouokam, Brian Cruz, Sean Burgess, Mark R. Segal, Mariel Vazquez, and Javier Arsuaga. The Rabl configuration limits topological entanglement of chromosomes in budding yeast. *Scientific Reports*, 9(1):1–10, 2019. ISSN 20452322. doi: 10.1038/s41598-019-42967-4.

Juan F. Poyatos and Laurence D. Hurst. Is optimal gene order impossible? *Trends in Genetics*, 22(8):420–423, 2006. ISSN 01689525. doi: 10.1016/j.tig.2006.06.003.

Suhas S.P. Rao, Miriam H. Huntley, Neva C. Durand, Elena K. Stamenova, Ivan D. Bochkov, James T. Robinson, Adrian L. Sanborn, Ido Machol, Arina D. Omer, Eric S. Lander, and Erez Lieberman Aiden. A 3d map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, 159(7):1665–1680, 2014. ISSN 0092-8674. doi: https://doi.org/10.1016/j.cell.2014.11.021.

Uku Raudvere, Liis Kolberg, Ivan Kuzmin, Tambet Arak, Priit Adler, Hedi Peterson, and Jaak Vilo. g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Research*, 47(W1):W191–W198, 05 2019. ISSN 0305-1048. doi: 10.1093/nar/gkz369.

Matthew J. Rossi, Prashant K. Kuntala, William K.M. Lai, Naomi Yamada, Nitika Badjatia, Chitvan Mittal, Guray Kuzu, Kylie Bocklund, Nina P. Farrell, Thomas R. Blanda, Joshua D. Mairose, Ann V. Basting, Katelyn S. Mistretta, David J. Rocco, Emily S. Perkinson, Gretta D. Kellogg, Shaun Mahony, and B. Franklin Pugh. A high-resolution protein architecture of the budding yeast genome. *Nature*, 2021. ISSN 14764687. doi: 10.1038/s41586-021-03314-8.

Albin Sandelin, Wynand Alkema, Pär Engström, Wyeth W Wasserman, and Boris Lenhard. Jaspar: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic acids research*, 32(suppl_1):D91–D94, 2004.

Bart Scherens and Andre Goffeau. The uses of genome-wide yeast mutant collections. *Genome biology*, 5(7):1–8, 2004.

Adam Siepel, Gill Bejerano, Jakob S. Pedersen, Angie S. Hinrichs, Minmei Hou, Kate Rosenbloom, Hiram Clawson, John Spieth, La Deana W. Hillier, Stephen Richards, George M. Weinstock, Richard K. Wilson, Richard A. Gibbs, W. James Kent, Webb Miller, and David Haussler. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research*, 15(8):1034–1050, 2005. ISSN 10889051. doi: 10.1101/gr.3715005.

Arashdeep Singh, Meenakshi Bagadia, and Kuljeet Singh Sandhu. Spatially coordinated replication and minimization of expression noise constrain three-dimensional organization of yeast genome. *DNA Research*, 23(2):155–169, 2016. ISSN 17561663. doi: 10.1093/dnares/dsw005.

Angela Taddei and Susan M Gasser. Structure and function in the budding yeast nucleus. *Genetics*, 192(1):107–29, sep 2012. ISSN 1943-2631. doi: 10.1534/genetics.112.140608.

Takumi Takizawa, Karen J. Meaburn, and Tom Misteli. The Meaning of Gene Positioning. *Cell*, 135(1):9–13, 2008. ISSN 00928674. doi: 10.1016/j.cell.2008.09.026.

Sarah Amalia Teichmann and Reiner Albert Veitia. Genes encoding subunits of stable complexes are clustered on the yeast chromosomes: An interpretation from a dosage balance perspective. *Genetics*, 167(4):2121–2125, 2004. ISSN 00166731. doi: 10.1534/genetics.103.024505.

Annelyse Thévenin, Liat Ein-Dor, Michal Ozery-Flato, and Ron Shamir. Functional gene groups are concentrated within chromosomes, among chromosomes and in the nuclear space of the human genome. *Nucleic Acids Research*, 42(15):9854–9861, 2014. ISSN 13624962. doi: 10.1093/nar/gku667.

Timo Tiirikka, Markku Siermala, and Mauno Vihinen. Clustering of gene ontology terms in genomes. *Gene*, 550(2): 155–164, 2014. ISSN 0378-1119. doi: https://doi.org/10.1016/j.gene.2014.06.060.

Maria Tsochatzidou, Maria Malliarou, Nikolas Papanikolaou, Joaquim Roca, and Christoforos Nikolaou. Genome urbanization: clusters of topologically co-regulated genes delineate functional compartments in the genome of Saccharomyces cerevisiae. *Nucleic Acids Research*, 45(10):5818–5828, 03 2017. ISSN 0305-1048. doi: 10.1093/nar/gkx198.

Nikolaos Vakirlis, Omer Acar, Brian Hsu, Nelson Castilho Coelho, S Branden Van Oss, Aaron Wacholder, Kate Medetgul-Ernar, Ray W Bowman, Cameron P Hines, John Iannotta, et al. De novo emergence of adaptive membrane proteins from thymine-rich genomic sequences. *Nature communications*, 11(1):1–18, 2020.

Guang Zhong Wang, Martin J. Lercher, and Laurence D. Hurst. Transcriptional coupling of neighboring genes and gene expression noise: Evidence that gene orientation and noncoding transcripts are modulators of noise. *Genome Biology and Evolution*, 3(1):320–331, 2011. ISSN 17596653. doi: 10.1093/gbe/evr025.

Elizabeth J.B. Williams and Dianna J. Bowles. Coexpression of neighboring genes in the genome of Arabidopsis thaliana. *Genome Research*, 14(6):1060–1067, 2004. ISSN 10889051. doi: 10.1101/gr.2131104.