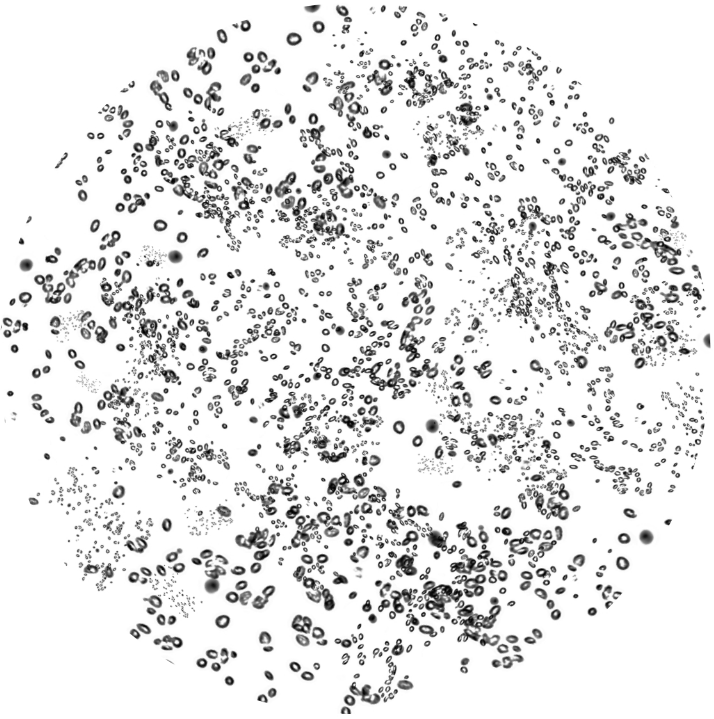

A Study of Sequence and Structural Constraints in the Bidirectional Promoters of *S. cerevisiae*

Papazogonopoulos Dionysios _ M.Sc. Thesis

ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ

M.Sc. Thesis

A Study of Sequence and Structural Constraints in the Bidirectional Promoters of *S. cerevisiae*

Papazogonopoulos Dionysios

University of Crete
School of Sciences and Engineering
Department of Biology
Joint Graduate Programme in Molecular Biology and Biomedicine (MBB)

Supervising professor: Christoforos Nikolaou
Heraklion, October 2016

Ευχαριστίες

Καθώς γράφω τις τελευταίες αυτές γραμμές πριν ολοκληρωθεί η πτυχιακή μου εργασία, και μαζί το μεταπτυχιακό πρόγραμμα, νιώθω την ανάγκη να ευχαριστήσω όλα εκείνα τα άτομα που κατά την διάρκεια των δύο αυτών χρόνων ήταν αρωγοί στην προσπάθειά μου. Η βοήθεια και συμπαράστασή τους ήταν αναγκαία στο να τα καταφέρω και να φτάσω ως το τέλος.

Πρώτα απ' όλους θα ήθελα να ευχαριστήσω τον καθηγητή και επιβλέποντα μου, κ. Χριστόφορο Νικολάου που μου έδωσε την ευκαιρία να εργαστώ στην ομάδα του, και που όχι μόνο μου παρείχε απαραίτητες και χρήσιμες συμβουλές σε στιγμές άγχους και αγωνίας, αλλά ήταν καταλυτικός παράγοντας στην ήρεμη και ομαλή πορεία της πτυχιακής. Τον ευχαριστώ βαθιά για όλα.

Θέλω να εκφράσω, επίσης, την ευγνωμοσύνη μου και στους δύο ακόμα καθηγητές που απαρτίζουν την τριμελή επιτροπή μου, κ Χαράλαμπο Σπηλιανάκη και κ. Ιωάννη Ηλιόπουλο. Η συνεισφορά τους δεν περιορίζεται μόνο στα πλαίσια των δύο αυτών χρόνων που διήρκεσε το μεταπτυχιακό πρόγραμμα, αλλά εκτείνεται σε ολόκληρο το διάστημα των σπουδών μου, από την πρώτη μέρα ως προπτυχιακός φοιτητής που πάλευε να κατανοήσει τις βασικές έννοιες της βιολογίας και της βιοχημείας, μέσα από το μάθημα του κ. Σπηλιανάκη, έως και την μετέπειτα “στροφή” του ενδιαφέροντός μου στην βιοπληροφορική μέσα από την καθοδήγηση του κ. Ηλιόπουλου. Νιώθω τυχερός που είχα την ευκαιρία να είμαι μαθητής και των τριών και ελπίζω η καλή αυτή συνεργασία να μείνει ως παρακαταθήκη για το μέλλον.

Η ηρεμία, το ευχάριστο κλίμα και η καλή συνεργασία στο περιβάλλον του εργαστηρίου είναι απαραίτητες συνθήκες για την αποδοτική και παραγωγική εργασία. Γι' αυτό οφείλω να ευχαριστήσω και όλα τα πρότερα και νυν μέλη της ομάδας “Υπολογιστικής Γονιδιωματικής”, την Μαρία Τσοχατζίδου, τον Σάββα Παραγκαμιάν, τον Βενιαμίν (Μπεν) Παπαδάκη, την Κλειώ-Μαρία Βέρρου, την Ιωάννα Πιτσιδιανάκη, το Γιώργο Κολιοπάνο, τον Ανδρέα Αγγελόπουλο, την Άννα Καραβαγγέλη, το Θεόφιλο Χαλκιαδάκη, τη Μαρία Μαλλιαρού, τον Άκη Λινάρδο, τον Αντώνη Κλωνιζάκη, την Δήμητρα Βλαχοκώστα, αλλά και τον πολύ καλό μου φίλο και συνοδοιπόρο των φοιτητικών μου χρόνων Αντώνη Παπαδάκη. Είμαι ιδιαίτερα χαρούμενος που η συνύπαρξη με όλους σας στον χώρο του εργαστηρίου δεν περιορίστηκε στο επίπεδο της τυπικής συνεργασίας, αλλά απέκτησε χαρακτήρα φιλίας.

Καθ' όλη την διάρκεια των φοιτητικών μου χρόνων είχα, επίσης, την ευκαιρία να διαμορφώσω δυνατές και διαχρονικές φιλίες που λειτούργησαν ως σταθεροποιητικός μηχανισμός στην προσπάθεια μου. Ευχαριστώ, λοιπόν, τους καλούς μου φίλους Γιώργο Γελαδάκη, Ηλία Τζελέπη, Μαριέτα Ρήγα, Γιάννη Ευαγγελάκο, Νίκο Σκεντέρη και Ανθή Συμεωνίδου. Ακόμη, ένα μεγάλο ευχαριστώ αξίζει στην αγαπημένη μου Δήμητρα Τάσσου, που με ανεχόταν και με στήριζε καθημερινά.

Τέλος, οφείλω ένα μεγάλο ευχαριστώ στους γονείς μου, την μητέρα μου, τον θείο και θεία μου και την πολυαγαπημένη γιαγιά μου που, παρ' όλες τις δικές τους προσωπικές δυσκολίες, με εμπιστεύτηκαν και μου παρείχαν, όλα αυτά τα χρόνια, ό,τι ήταν δυνατό για να πραγματοποιήσω τα θέλω και τις επιθυμίες μου. Χωρίς εσάς, κυριολεκτικά, δεν θα είχα φτάσει ως εδώ.

Table of Contents

Abstract

Bidirectional gene pairs, also known as head-to-head gene pairs, are defined as two genes on different DNA strands with adjacent 5'-ends that are transcribed divergently. Their transcription is often coordinated in order to achieve their biological role. The region between a bidirectional gene pair is designated as a putative bidirectional promoter, which is suggested to coordinately regulate the expression of the pair. Given that divergently transcribed genes constitute quite a substantial percentage of the total genes in a wide range of organisms, many studies have already focused on the characteristics of bidirectional promoters. In this study, we focused our work on bidirectional promoters of *S. cerevisiae* from the point of view of sequence and structural constraints. We performed various types of analyses that included data for the phylogenetic conservation of sequences, the robustness of symmetry of DNA curvature, as a mark of structural information, data from nucleosome maps to quantify nucleosomal occupancy and, lastly, data of single nucleotide polymorphisms (SNPs). With the use of all that, we have tried to discover and describe the inherent sequential and structural elements that distinguish bidirectional intergenic regions and promoters from the intergenic regions of other possible gene pair organizations, in order to draw conclusions on the underlying mechanisms and constraints that define the regulation and diversification of bi-directional transcription in a simple eukaryotic genome.

Περίληψη

Ένα ζεύγος γονιδίων χαρακτηρίζεται ως “αμφίδρομο” ή “head-to-head” όταν τα δύο γονίδια βρίσκονται σε διαφορετική αλυσίδα του DNA, τα 5' -άκρα τους είναι σε κοντινή απόσταση και η μεταγραφή τους γίνεται με αμφίδρομο τρόπο. Συχνά, υπάρχει συντονισμένη ρύθμιση στην μεταγραφή των δύο γονιδίων ενός αμφίδρομου ζεύγους, έτσι ώστε αυτά να επιτελέσουν με βέλτιστο τρόπο τον βιολογικό τους ρόλο. Η περιοχή μεταξύ των δύο γονιδίων του αμφίδρομου ζεύγους αναφέρεται συχνά ως ο πιθανός “αμφίδρομος υποκινητής”, που είναι υπεύθυνος για τον συντονισμό και την ρύθμιση της έκφρασης του ζεύγους γονιδίων. Τα αμφίδρομα ζεύγη γονιδίων αποτελούν ένα αξιοσέβαστο ποσοστό των συνολικών γονιδίων σε ένα μεγάλο εύρος οργανισμών, όπως για παράδειγμα η ζύμη και ο άνθρωπος, γι' αυτό και πολλές μελέτες έχουν επικεντρωθεί στην μελέτη των χαρακτηριστικών των αμφίδρομων υποκινητών τους. Σε αυτή την εργασία, επικεντρωθήκαμε στην μελέτη των αμφίδρομων υποκινητών του *S. cerevisiae*, με στόχο να βρούμε περιορισμούς στην αλληλουχία και την δομή τους. Πραγματοποιήσαμε αναλύσεις που περιλαμβάνουν την χρήση διαφόρων ειδών δεδομένων, όπως δεδομένα φυλογενετικής συντήρησης των αλληλουχιών, δεδομένα δομικής πληροφορίας που σχετίζονται με την συμμετρία της καμπυλότητας του DNA, δεδομένα από νουκλεοσωμικούς χάρτες για την ποσοτικοποίηση της ύπαρξης των νουκλεοσωμάτων στις υπό εξέταση περιοχές και, τέλος, δεδομένα πολυμορφισμών (SNPs). Με την χρήση όλων αυτών, προσπαθήσαμε να ανακαλύψουμε και να περιγράψουμε τα εγγενή εκείνα στοιχεία στην αλληλουχία και στη δομή που ξεχωρίζουν τους αμφίδρομους υποκινητές από τις περιοχές μεταξύ των γονιδίων των υπόλοιπων πιθανών διευθετήσεων, έτσι ώστε να βγάλουμε συμπεράσματα για τους βαθύτερους μηχανισμούς και περιορισμούς που καθορίζουν την ρύθμιση και τους ποικίλους φαινοτύπους της αμφίδρομης γονιδιακής έκφρασης ενός απλού ευκαρυωτικού οργανισμού.

1.Introduction

1.1 Bidirectional genes

Bidirectional or divergent genes pairs are defined as pairs of adjacent or overlapping Open Reading Frames (ORFs) that are located on opposite strands of DNA and are transcribed in opposing directions (Fig. 1). The region between the Transcription Start Sites (TSSs) of these two ORFs (intergenic space), which usually contains the regulatory elements for transcription, is often termed as the putative "bidirectional promoter" (Li et al. 2006a). Such promoters have been described in a plethora of organisms, ranging from yeast to human (Neil et al. 2009; Core et al. 2008; Seila et al. 2008; Preker et al. 2008; Xu et al. 2009; Wang et al. 2009; Koyanagi et al. 2005; Trinklein et al. 2004) and, owing to this abundance, they have been the focus of many studies. As Pol II, the molecular player shouldered with the duty of transcribing the majority of transcriptionally active regions, binds on such promoters, it can transcribe DNA in both directions as it is directionally unbiased. Divergent transcription can result in the production of either two mRNAs (head-to-head genes) or a single mRNA and a corresponding upstream non-coding RNA (ncRNA) (Xu et al. 2009; Neil et al. 2009; Core et al. 2008; Seila et al. 2008; Preker et al. 2008). The number of bidirectional or "head-to-head" genes as a percentage of the total genome depends on the size of the genome. Therefore, while in humans only 11% of the genes appear to have such organization, in the much smaller and more compact yeast genome this ratio rises to half of the mRNA coding genes (Xu et al. 2009).

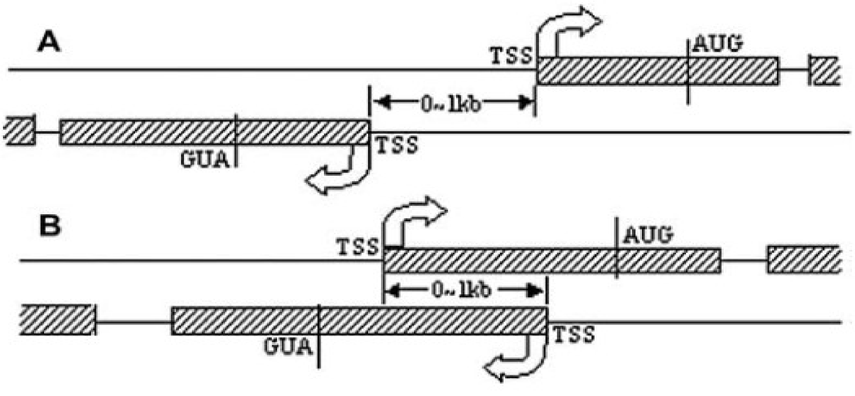

Figure 1: Schematic illustration of non-overlapping (A) and overlapping (B) head-to-head gene pairs (Li et al. 2006a)

1.2 Biological significance of bidirectional gene pairs

1.2.1 Regulation of expression

The bidirectional organization of genes helps in the finer coordination of their transcription and regulation. Through shared cis-regulatory sequences and/or the localization in the same chromatin domain genes become more effective in achieving their biological role. Such a synchronization of expression is needed for several reasons. Known examples of genes that profit from this kind of organization include the complex-forming histone genes H2A and H2B (Trappe et al. 1999) and the collagen genes COL4A1 and COL4A2 (Burbelo et al. 1988). In the case of histone genes, concerted expression serves to maintain a stoichiometric relationship (Albig et al. 1997; Ahn & Gruen 1999; Maxson et al. 1983). In other examples, correlated expression helps with certain cellular processes or responses. More specifically, the mouse genes RanBP1 and Htf9-c are co-regulated through different time-points during cell cycle progression (Guarguaglini et al. 1997), the human PSENEN and U2AF1L4 are concertedly regulated for their possible involvement in the regulation of T-cell activity (Didych et al. 2013), while the, also human, genes HSP60 and HSP10 are coordinated to respond to a heat shock signal (Hansen et al. 2003). Lastly, there are also examples of negative correlation between bidirectional gene pairs, such as the mouse thymidine kinase gene and kynurenine formamidase gene (Schuettengruber et al. 2003).

1.2.2 Similarity in function

The coordination of expression described above increases the potential of gene pairs to perform similar functions. Functional analyses of bidirectional gene pairs based on gene ontology (GO) demonstrated a tendency of these pairs to co-function (Li et al. 2006b), especially in functional categories such as metabolism, chromosome organization and DNA packaging, anion transport, nucleic acid binding, catalytic activity, intracellular and organelle components, protein complex and collagen type IV. A stronger correlated expression and evolutionary conservation was shown for bidirectional genes of similar functions (Chen et al. 2010). Therefore, bidirectional gene pairs seem to be coordinately expressed for their common functions.

1.3 Bidirectional promoters

As mentioned previously, the region between the two TSSs of bidirectional genes is considered to be the location of promoters and regulatory elements with bidirectional activities. Using deletions in these areas (bidirectional promoters), complemented with transcriptional activity assays, it has been shown that most of the promoters examined contain a shared region necessary for full promoter activity in both directions (Trinklein et al. 2004). Thus, given their importance, they have been the object of studies concerning their architectural characteristics and qualities, some of which are described below.

1.3.1 Length

In humans, a bidirectional promoter's length -that is, for regions that are non-overlapping- is found to be less than 1kb (Adachi & Lieber 2002). Koyanagi et al found a minor peak in the distribution of lengths (< 1kb), although they were not able to explain it (Koyanagi et al. 2013). In other studies, the length of a bidirectional promoter and the degree of expression correlation have not been significantly correlated (Chen et al. 2010; Trinklein et al. 2004). In yeast, the promoter's length distribution has revealed two subpopulations centered at 290bp and 771bp, respectively, with the typical distance being 490bp (Pelechano et al. 2006). In a study of the evolution of yeast's intergenic regions Sugino R. and Innan H. showed that newly created divergent pairs have longer intergenic regions than conserved divergent gene pairs, something which they attributed to the slowest rate of shrinkage of the new pairs and the selective pressures exercised to prevent potential deleterious co-expression of genes (Sugino & Innan 2012).

1.3.2 Core promoter elements and TF binding motifs

Most bidirectional promoters are characterized by the lack of TATA boxes and an enrichment in CpG islands (Adachi & Lieber 2002; Trinklein et al. 2004), a fact that is suggestive of enrichments in constitutive, highly expressed genes. The co-regulation of a bidirectional gene pair is achieved by the synergic cooperation of TFs that bind to their promoters, such as GABP/NRF2 and YY1 regulate the human PREPL and C2ORF34 gene pair (Huang & Chang 2009). A small set of motifs, such as GABPA, MYC, E2F1, E2F4, NRF-1, CCAAT, YY1, ACTACAnnTCC is found to be overrepresented in bidirectional promoters (Lin et al. 2007; Liu et al. 2011).

1.3.3 Chromatin structure

Chromatin structure plays an important role in dictating the initiation of transcription in both directions. As has been previously described, Pol II promoters appear to accomodate nucleosome-free regions (NFRs) (Yuan et al. 2005a; Albert et al. 2007; Mito et al. 2005; Mavrich et al. 2008; Ozsolak et al. 2007). This lack of nucleosomes promotes the binding of transcriptional complexes and the unwinding of DNA strands. Nucleosome assembly in these areas is disfavored by characteristic sequential elements, while allowing the recruitment of transcription factors (Guertin & Lis 2013). Specifically in yeast, poly(A-T) tracts, which are found frequently in promoter sequences, is a mean to this end (Sekinger et al. 2005; Segal, Fondufe-Mittendorf, Chen, Thåström, Field, Irene K. Moore, et al. 2006; Kaplan, Irene K. Moore, et al. 2009). It has also been suggested that the competition between transcription factors and nucleosomes for DNA binding results in the creation of NFRs (Zhang et al. 2009; Floer et al. 2010; Bai et al. 2011; Ozonov & van Nimwegen 2013). Thus, in bidirectional promoters, two distinct preinitiation complexes can be harbored on the two DNA strands and drive sense and anti-sense expression (Rhee & Pugh 2012), either independently or coordinately (Xu et al. 2009; Murray et al. 2012). Finally, in addition to lower nucleosome occupancy, bidirectional promoters in yeast have been found to be flanked by strongly positioned nucleosomes, which probably results in lower gene expression variability of the corresponding bidirectional gene pair (Woo & Li 2011).

1.4 Robustness of SymCurv

Nucleosomes are a vivid part of DNA structure and directly affect chromatin organization. Nnucleosomal organization plays a key role in significant molecular processes, such as the regulation of transcription (Guenther et al. 2007; Mellor 2006) and DNA replication (Eaton et al. 2010; Yin et al. 2009). Many studies have tried to experimentally address the question of nucleosome positioning in yeast and produced results that include the coordinates of nucleosomes (Yuan et al. 2005b; Shivaswamy et al. 2008; Lee et al. 2007; Wal & Pugh 2012). Other studies were focused on determining the sequence properties of nucleosomal DNA (Stein et al. 2009; Caserta et al. 2009; Ioshikhes et al. 2006; Kaplan, Irene K Moore, et al. 2009; Ogawa et al. 2010; Kaplan, Irene K Moore, et al. 2009; Peckham et al. 2007; Segal, Fondufe-Mittendorf, Chen, Thåström, Field, Irene K Moore, et al. 2006) as a means to a better prediction of the positions of nucleosomes, but their conclusions were insufficient in providing a concise framework that explains the dynamics of the process of nucleosome positioning. They suggested that predictions are possible only for a subset of nucleosomes and that the existence of sequence constraints in nucleosomal DNA could not be strongly supported. In their study, Nikolaou et al, explored the concept of existence of sequence constraints that drive the positioning of 'consistent' nucleosomes -that is, nucleosomes that appear to have stable positions across different studies (Nikolaou et al. 2010). Given that nucleosomes affect chromatin structure and conformation, they hypothesized that the DNA structure may reflect these constraints better. For their work they used the Symmetry of DNA Curvature (Tilgner et al. 2009), which is a measure of structural information, and extended the concept by calculating the robustness of SymCurv values. They reasoned that the existence of structural constraints within a sequence will produce quantifiable changes in the structural level even if small changes in the level of sequence occurred, producing SymCurv values of high variance (low robustness).

Robustness of SymCurv values is a good indication of structural information that is not necessarily reflected in the level of sequence and, in this work, we wanted to apply this measure in order to determine differences between divergent (bidirectional) and the other forms of gene pair organization.

2. Results

2.1 Length of intergenic regions

In order to explore any possible difference in the length of the intergenic region between bidirectional gene pairs and the other gene organizations, we calculated the distances of the two TSSs for every gene pair. To do this, we categorized all *S. cerevisiae* genes (as described in section 3.1) in three major categories: tandem (minus-minus (-/-), plus-plus (+/+)), convergent (plus-minus (+/-)) and divergent (minus-plus (-/+)). As shown in Fig. 2, we observed a wider distribution of minus-plus -that is divergent gene pairs- compared to the other sub-categories, with slightly elevated intergenic lengths. Minus-minus and plus-plus gene groups (which correspond to tandem organization) appear to have essentially similar distributions, while plus-minus genes (convergent) have overall lower intergenic distances.

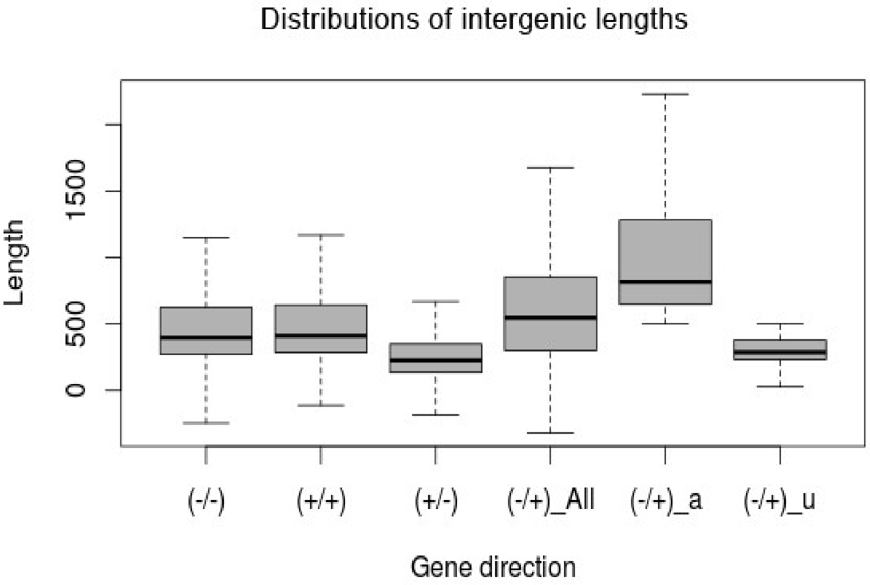

Figure 2: Box-plots depicting the distributions of the size of the intergenic regions in the four gene organizations. The length was measured in bp. The last two boxes (-/+)_a and (-/+)_u correspond to the two subgroups of divergent genes (>500 bp and <500 bp accordingly).

2.2 Conservation and robustness of SymCurv

In order to explore further the differences between bidirectional intergenic regions (minus-plus) and the intergenic regions of the other types of gene organization we incorporated data on the robustness of Symmetry of Curvature (SymCurv) and directly compared them with data of sequence conservation. In this way, it would be possible to determine any constraints in the structural level of the bidirectional regions, that may not be apparent at the one-dimensional level. For this purpose, we calculated the mean robustness and conservation scores for each intergenic region. Being in a different scale, the two different types of scores were normalized and plotted in R. We have also calculated conservation scores for genes' regions (Supplementary Fig. 1 & 2)

Figures 3 and 4 correspond to the conservation and robustness scores in the four different groups. Each group appeared to have its own trend in both conservation and robustness. Intergenic regions of divergent gene pairs have a steady rise in conservation until around the middle of total region followed by a steady decline right before the TSS of the sequential gene. In terms of robustness they present two local peaks, one around bin 35 and one around bin 50, while in the regions before and after the appear to have lower robustness values. The peaks near bins 0 and 100 in both conservation and robustness and in all four groups are possibly observed because of the proximity to genes (TSSs), which are naturally more conserved in terms of sequence and contain less information on the structural level (robustness).

To directly compare the conservation and robustness patterns we created four distinct plots corresponding to the four groups and two additional plots analyzing in a higher resolution these values (Fig. 5). For these last two plots, we used the two subgroups of divergent (minus-plus) gene pairs produced by a division of the initial list based on the intergenic distance (>500bp or <500bp).

All four gene orientations appeared to have their own conservation and robustness patterns and we found no regions that do contain structural information (low robustness) and, at the same time, are depleted of sequential information (low conservation). The two subgroups of divergent gene pairs gave more schematic results. Smaller intergenic regions (<500bp) have a clear negative correlation of the two scores, with conservation peaking at around the middle of the region where robustness has its lowest scores. This pattern, high sequence significance (conservation) and low robustness (existence of structural constraints) signifies the existence of regulatory elements, possibly the bidirectional promoter. This pattern was also observed in the subgroup with bigger intergenic regions (>500bp), but, this time, it was placed right before the TSSs of each gene. This could suggest the existence of two different promoters, one for each gene. In the middle of the region, the pattern is reversed, showing the lack of an apparent biological function of this subregion.

2.3 Conservation, robustness and nucleosome occupancy around TSS

We next wanted to compare and characterize the regions around the TSS of the different gene orientations in terms of conservation, robustness and nucleosome occupancy. To do this, firstly we redefined the categories to tandem (which includes both minus-minus and plus-plus), divergent (minus-plus) and convergent (plus-minus). Using the coordinations of TSSs of genes in tandem and in divergent formation we created 1000bp-long regions centered on TSS (+/- 500bp) of each gene. Using the new coordinations we calculated the conservation and robustness scores for these regions and the nucleosome occupancy. In terms of conservation, as we expected, we observed an interval of lower conservation (0-40 bins), which is the intergenic segment, followed by a decline right before the TSS (50) and a sharp rise once we enter the gene territory (Fig. 6). The two classes (tandem, divergent) showed the same trend in conservation values along the 1000bps, with the interesting exception of the segment around bin 40, where we see slightly more elevated conservation values. In terms of robustness, values of divergent gene pairs are constantly higher than those derived from gene pairs in tandem, until around bin 85 where we see a shift for a small region (Fig. 7). As was the case with conservation values, robustness is lower in the intergenic region (bins 0-40), followed by a steep increase around the TSS (bin number 50) and, from then and on, scores are higher.

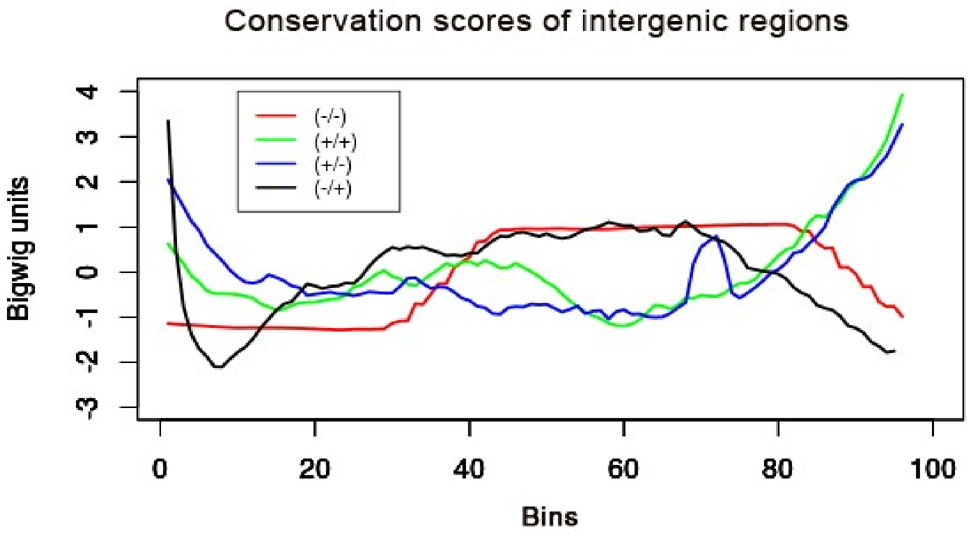

Figure 3: Mean conservation scores of intergenic regions in the four categories of gene organization depicted as continuous signal. Each colored line represents a different group as show in the legend. Conservation scores were normalized as z scores (y axis), while the intergenic regions were divided into 100 bins (x axis).

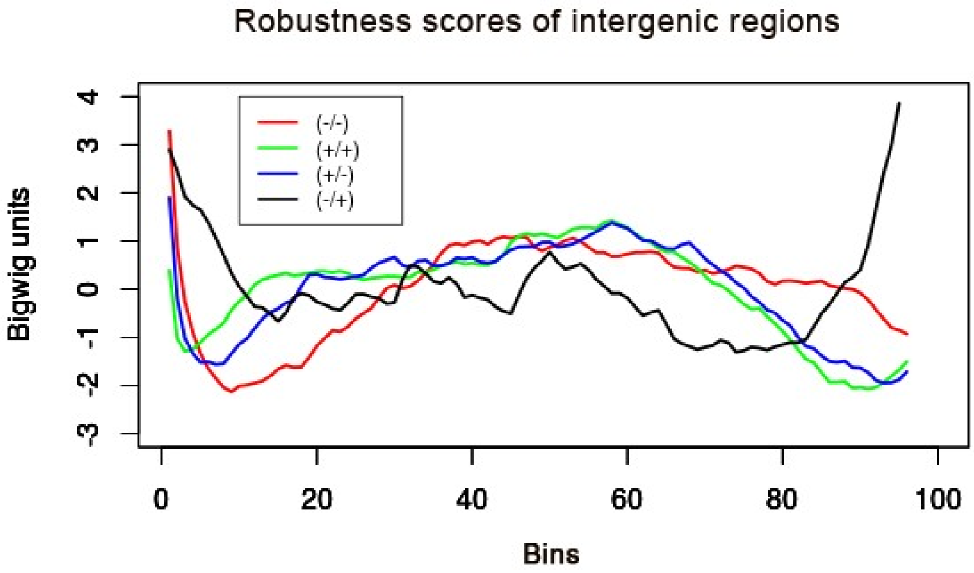

Figure 4: Mean robustness scores of intergenic regions in the four categories of gene organization depicted as continuous signal. Each colored line represents a different group as show in the legend. Conservation scores were normalized as z scores (y axis), while the intergenic regions were divided into 100 bins (x axis).

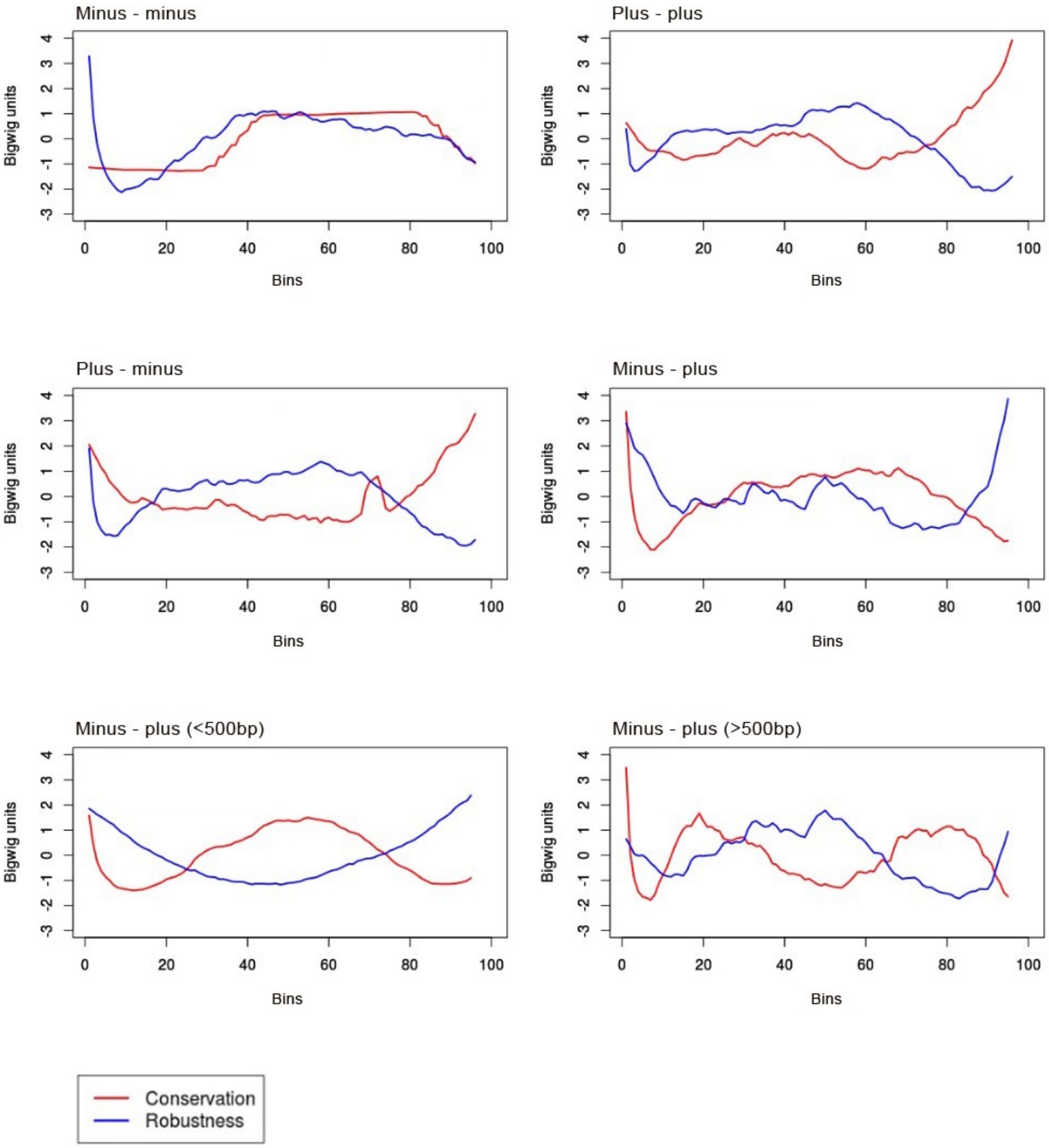

Figure 5: Plots showing the mean conservation (red) *and robustness* (blue) *scores (y axes) in each bin (x axes) as a continuous signal. Each of the first four plots corresponds to a different group of gene pair orientation. The last two plots show the two different subgroups of divergent gene pairs that contain regions smaller or bigger than 500bp.*

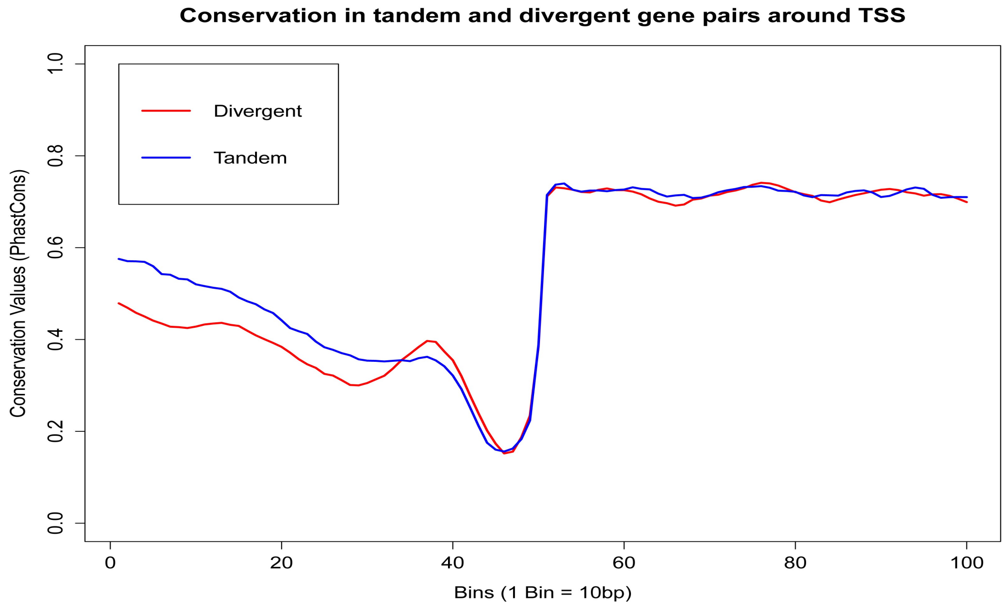

Figure 6: Mean conservation scores in tandem and divergent gene pairs around TSS. The figure shows the mean conservation values of a 1000bp-long regions in divergent gene pairs (red) and in tandem gene pairs (blue). In the x axis we have the number of bins used to calculate the conservation values (y axis) using the bigWigSummary utility. TSS is positioned in the middle of x axis (bin 50)

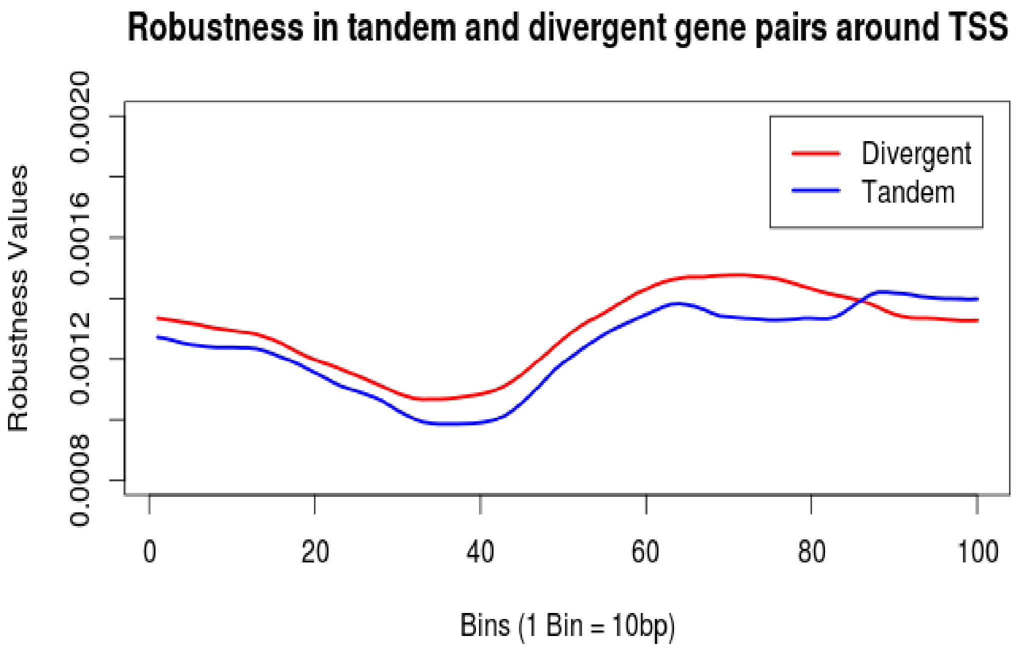

Figure 7: Mean robustness scores in tandem and divergent gene pairs around TSS. The figure shows the mean robustness values of a 1000bp-long regions in divergent gene pairs (red) and in tandem gene pairs (blue). In the x axis we have the number of bins used to calculate the robustness scores (y axis) using the bigWigSummary utility. TSS is positioned in the middle of x axis (50).

Finally, nucleosome occupancy scores appear to have almost the same height and trend before and after bin 40 in both divergent and tandem gene pairs (Fig. 8). However, in the case of divergent gene pairs we observe a deeper cleft around bin 40. This seems to be the only substantial difference between these two classes of gene pairs.

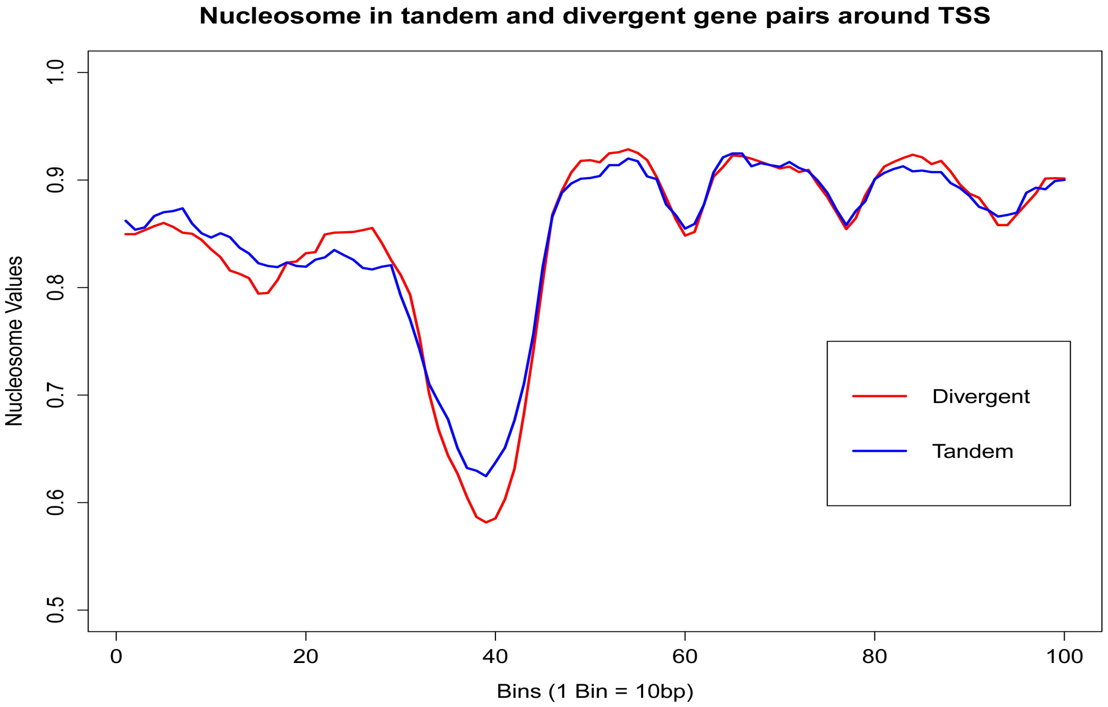

Figure 8: Mean nucleosome occupancy scores in tandem and divergent gene pairs around TSS. The figure shows the mean conservation values of a 1000bp-long regions in divergent gene pairs (red) and in tandem gene pairs (blue). In the x axis we have the number of bins used, where every bin unit corresponds to 10bp in the real genome. The nucleosome scores (y axis) were generated by the bigWigSummary utility. TSS is positioned in the middle of x axis (50).

For a direct side-to-side comparison of the conservation and nucleosome occupancy scores around TSS, we plotted the data again as shown in figures 9 & 10.

2.4 SNPs density

In order to extend our analysis we went on to incorporate data of sequence polymorphism of the *S. cerevisiae* genome. We used the data from (Schacherer et al. 2007) and produced SNPs density scores again for each of the four gene pair categories of gene and intergenic regions. We observed no significant difference between the distributions of SNPs density in genes and some minor differences in the distributions of the intergenic regions. More specifically, we noticed that the majority of the values in the distribution of convergent (plus-minus) intergenic regions are lower and the majority of the divergent values appear to have a smaller intequantile range (Fig. 11 & 12).

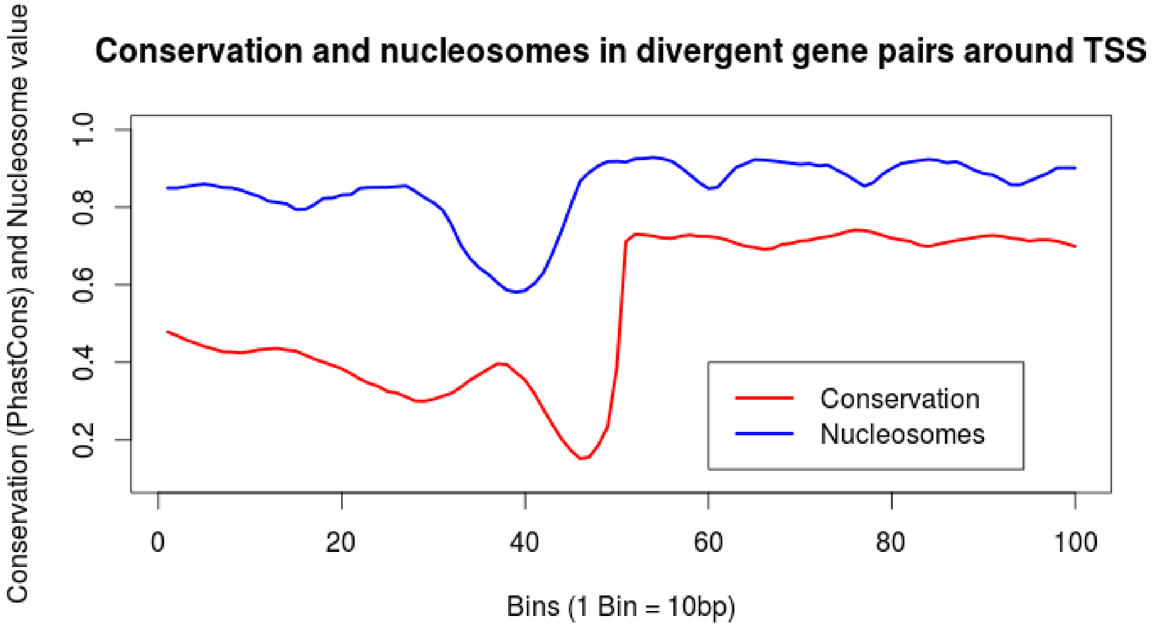

Figure 9: Mean conservation and nucleosome occupancy scores in divergent gene pairs around TSS. The figure shows the mean conservation scores of a 1000bp-long regions (red) and the mean nucleosome occupancy scores (blue). In the x axis we have the number of bins used, where every bin corresponds to 10bp in the real genome. The conservation and nucleosome scores (y axis) were generated by the bigWigSummary utility. TSS is positioned in the middle of x axis (50).

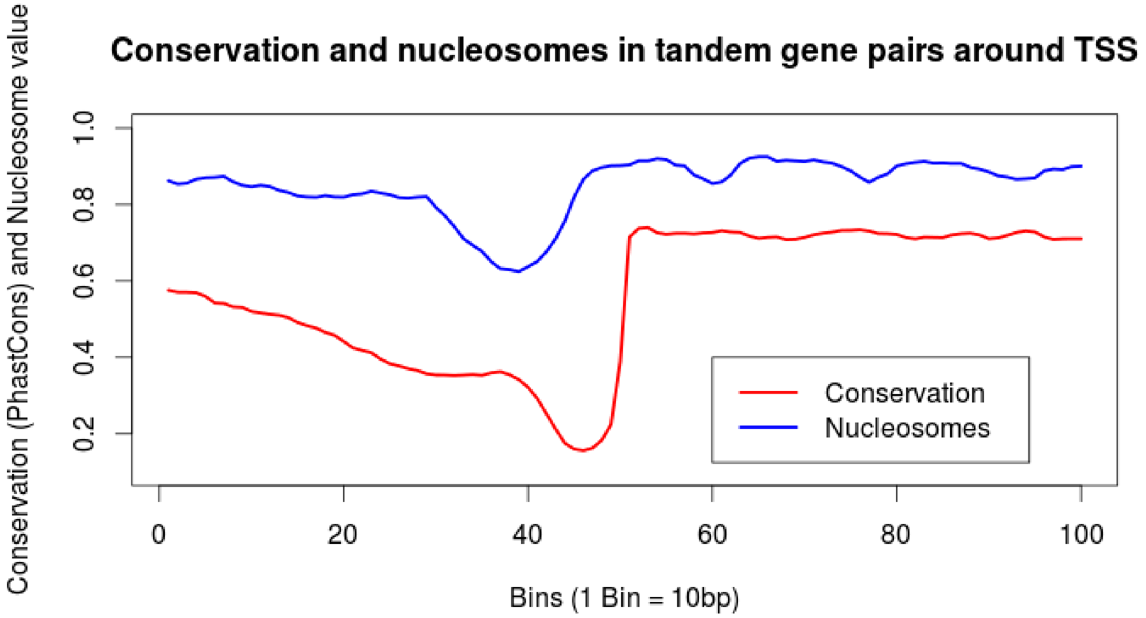

Figure 10: Mean conservation and nucleosome occupancy scores in tandem gene pairs around TSS. The figure shows the mean conservation scores of a 1000bp-long regions (red) and the mean nucleosome occupancy scores (blue). In the x axis we have the number of bins used, where every bin corresponds to 10bp in the real genome. The conservation and nucleosome scores (y axis) were generated by the bigWigSummary utility. TSS is positioned in the middle of x axis (50).

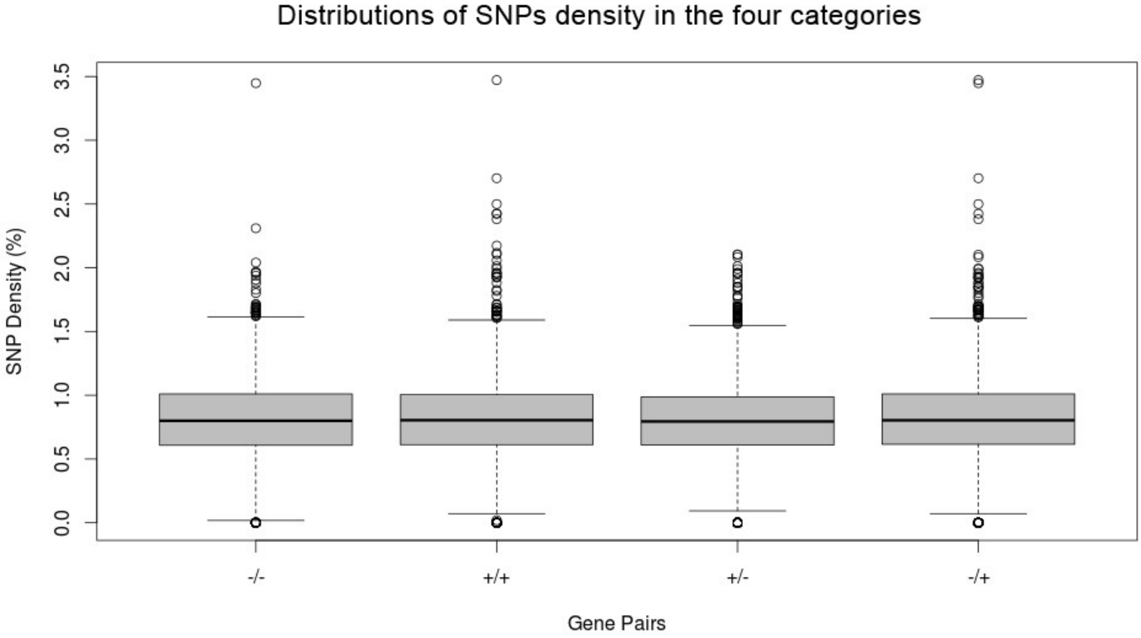

Figure 11: Gene SNPs density distributions of the four gene orientations. Box-plots show the SNPs density of the genes (y axis) in every gene orientation (x axis).

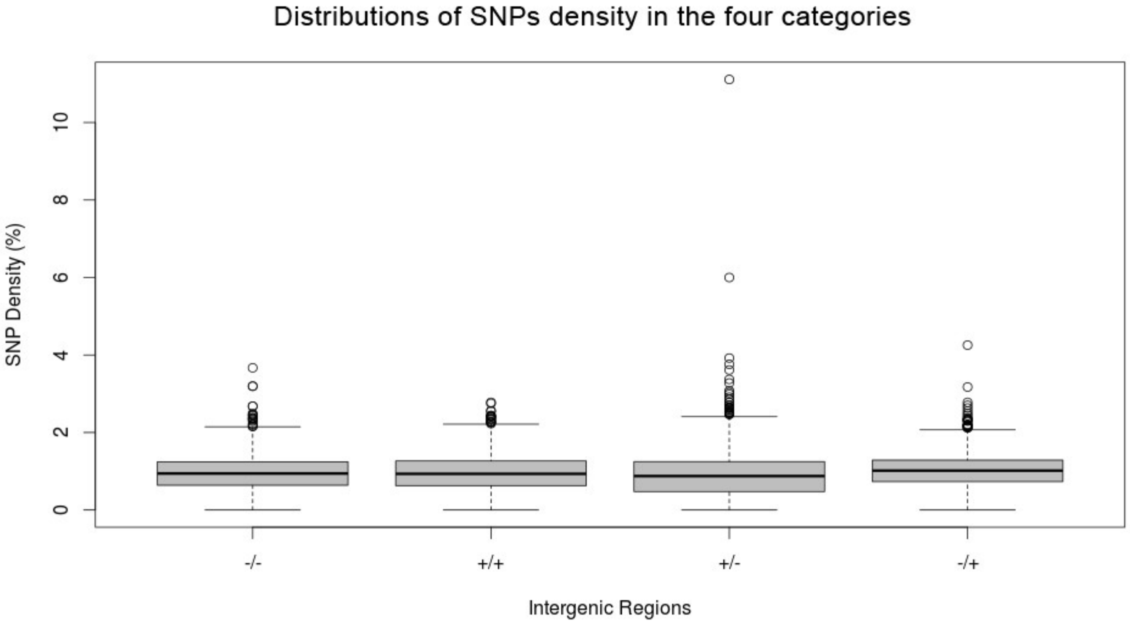

Figure 12: Intergenic SNPs density distributions of the four gene orientations. Box-plots show the SNPs density of the intergenic regions (y axis) in every gene orientation (x axis).

3. Materials and methods

3.1 Genome data and gene pair determination

S. cerevisiae gene coordinates were obtained from the UCSC track “SGD Genes” and assembly “Oct. 2003 (SGD/saccer1)(UCSC n.d.). A perl programm processed these data to produce groups of gene pairs. This algorithm used each and every gene of the initial gene list and paired it with the next in line. Every pair was then appointed to one of the categories, according to the genes' direction. For example, if one gene is on minus strand and the consecutive gene is on plus strand, this pair was categorized as divergent. Divergent genes were divided further into two subgroups, based on the length of the intergenic region (>500bp or <500bp). The coordinates of the TSSs were used to determine the length of the intergenic region between gene pairs.

3.2 Conservation scores

Conservation data for each genomic position was derived from the UCSC track “Yeast (*S. cerevisiae*) Genome (saccer1)” (UCSC n.d.). Conservation scores are produced by a program called “PhastCons” by the alignment of 7 yeast species (*S. cerevisiae, S. paradoxus, S. mikatae, S. kudriavzevii, S. bayanus, S. castelli, S. kluyveri*) and range from 0 (no conservation) to 1 (total conservation). Mean conservation scores were obtained for coding and non-coding positions, which includes every gene pair category described (tandem, divergent, convergent).

3.2.1 PhastCons

PhastCons (Siepel et al. 2005) is a program created to identify evolutionary conserved elements by aligning sequences from multiple species, given a phylogenetic tree. In its core lies a Phylogenetic Hidden Markov Model (phylo-HMM), a statistical model of sequence evolution. In its analysis, PhastCons, considers *n* number of species, their relative phylogeny and uses statistical models of nucleotide substitution between pairs of bases (e.g. a higher frequency of transitions than transversions). It is a part of the PHAST package and is written in C.

3.2.2 bigWigSummary

Conservation scores are encoded in a bigWig format file, which is a format used for the visualization of dense and continuous data. In order to assign each region or position a score we used the bigWigSummary utility, provided by UCSC (UCSC n.d.). bigWigSummary can be used either directly as a terminal command or be a part of a larger program. To produce the scores, bigWigSummary, requires as input the following: a bigWig file (.bw), a description of the value to be produced (e.g. min, max, mean), the coordinates of the region to be scored (chromosome, starting coordinate, ending coordinate) and the number of parts (bins) this region will be divided. So, for example, if we want to score the genomic region with coordinates chr1 1000 1200 and achieve a resolution of 10 bps, we have to set the number of bins to 20.

3.3 Robustness of SymCurv

For calculating the Robustness of SymCurv scores for the querying regions and positions we used the data from (Nikolaou et al. 2010). Robustness scores are also encoded in a bigWig format and were extracted with the same procedure as conservation scores (described above).

Even if it was not a part of the work of this study, the methodology for the calculation of SymCurv and Robustness values is described below in great detail, as we strongly believe it is a an important step for a better understanding of the reader.

3.3.1 Calculation of SymCurv values

The symmetry of curvature of DNA sequences was calculated by applying symmetry constraints on a DNA sequence's predicted curvature. Given a sequence on which curvature values are computed for each trinucleotide step, a symmetric pattern around a given nucleotide position n would imply similar values of curvature at equal distances from this position on either direction. That is, the value of curvature at position *n-1*, $Curv_{n-1}$ should be similar to the value at position *n+1*, $Curv_{n+1}$, and this should hold for all pairs of positions at distance *i* from *n* for *i = 1, ..., m*, where *m* is a suitably adjusted parameter. At each such distance, we can compute the absolute difference between the corresponding curvature values $d_i = |Curv_{n-i} - Curv_{n+i}|$. The lower this value, the higher the symmetry within the given distance *i* from *n*.

The symmetry of the curvature of the sequence centered at position *n* on a window of length m was defined as the inverse sum of the distances over all values from *1* to *m*:

$$S_{sym}(m) = \sum_{i=0}^{m} \frac{1}{d_i} \qquad (1)$$

The inversion in the S_{sym} formula is performed to maximize the symmetry score; the more symmetric the values on either side of position n, the closer the sum of distances will approach zero and thus larger the symmetry value S_{sym} will become. As values of d_i are generally in the range of two orders of magnitude, the inversion in equation 1 is used (rather than, for example, the simpler use of its negative value) to increase the dynamic range of S_{sym} values and thus 'spread' the S_{sym} value range to better capture differences between sequences.

Based on the above definition of Symmetry, the symmetry of curvature is calculated as follows. Given a genomic sequence, the method proceeds by first calculating the curvature values and subsequently applying the symmetry constraints on the resulting curvature data.

First, DNA curvature values of the given sequence are calculated using BENDS [46], extended with the use of trinucleotide parameters as described previously [47,48]. The output of this step is an array in which a curvature value is attributed to every nucleotide, calculated through a window of length of 30 bp centered on each nucleotide and sliding 1 bp at a time. According to this scheme, each trinucleotide and its reverse complement (for example, TAA/TTA) are equivalent in terms of structural parameters (roll, tilt and twist angles). It is thus an easy matter to apply the calculation to both the sequence under examination and its reverse complementary. This makes sense from the physical point of view as the nucleosome forming potential of a given DNA sequence is expected to be strand-independent.

Secondly, nucleosomal sequences have been reported [34,36,39,40] to be flexible around their central region, where local distortions are relaxed. Thus, the region of the pseudodyad axis may be expected to produce lower curvature values, separating two parts of overall higher curvature. A local curvature minimum is thus set as a prerequisite for a given site to be considered as a possible dyad axis, and the calculation of S_{sym} is only to take place in sites fulfilling this condition. Thus, the curvature values array is scanned for local minima. For positions that fulfill the above criterion, a local minimum score is calculated according to the formula:

$$S_{min}(n)=\frac{1}{(Curv_{(n-1)}-Curv_{(n+1)})+(Curv_{(n+1)}-Curv_{(n)})} \qquad (2)$$

if $Curv_{(n-1)} > Curv_{(n)}$ and $Curv_{(n-1)} > Curv_{(n)}$, while S_{min} (n)= 0, otherwise $Curv_n$ is the curvature value at position n on the genomic sequence. The inversion in the S_{min} formula is performed to selectively increase the scores for mild local minima, as the local decrease in curvature on the dyad axis region is expected to be a smooth, minor decrease rather than an acute one.

Thirdly, the *SymCurv* symmetry score at every local minimum site is calculated using equation 1. The length parameter *m* was set to 25, based on the combined size of the pseudodyad axis and the immediate flanking regions. The calculation is thus conducted over a window of 50 nucleotides, which corresponds to five DNA double-helical pitches. The overall score of the symmetry of curvature, SymCurv, is calculated as the product of the two scores.

$$SymCurv_{(n,m)}=S_{(min(n))}S_{(sym(n,m))} \quad , \quad (3)$$

where m = 25.

It should be noted here that use of *m* values in the range of (*m* = 15 to 35), corresponding to three to seven helical turns yields similar results. Therefore, the usage of *m* = 25 (~five helical turns) was chosen as it is closer to the known size of the dyad axis region [49] and its immediate flanking sequences, which have been shown to be contributing the most to histone binding [50,51].

SymCurv thus assigns a value for each nucleotide. Given a region of size *L* nucleotides, we may calculate an overall SymCurv value for the genomic segment as the average over all nucleotides.

$$SymCurv(L)=\sum_{i}^{L} SymCurv\frac{(i)}{L} \qquad (4)$$

3.3.2 Robustness of SymCurv

Given a DNA sequence of length *L*, its average SymCurv value is initially calculated as described above. We then produce the complete set of DNA sequences, which differ from the original by one nucleotide, by mutating all individual positions but keeping the rest of the sequence intact. For a sequence of length *L*, there are *3L* one-nucleotide mutants or neighbors. We then calculate the average SymCurv values for all neighbors and defined the distance within them as the variance of the four values for each nucleotide position. Thus for nucleotide *i* the distance *D(i)* is:

$$D(i) = var(SymCurv[A], SymCurv[G], SymCurv[C], SymCurv[T])$$

where *SymCurv[X]* is the value of SymCurv at the *i*-th nucleotide for the neighbor bearing nucleotide *X* at that specific position. The overall distance for the complete sequence D_{seq} is then calculated as the average over all *L* positions:

$$D_{seq} = \sum_{i}^{L} D\frac{(i)}{L}$$

As high variance is a measure of variability, which is inversely related to robustness, we may define robustness *(R)* as the negative logarithm of the above distance:

$$R_{seq} = -\log(D_{seq})$$

The logarithm is used here to decrease the dynamic range of *R* purely for practical reasons, as overall distances *(D)* exhibit a range of values over several orders of magnitude.

A property such as R_{seq} is used as a measure of the variance of the SymCurv values between the one-nucleotide neighbors. It represents the tendency of a given sequence to radically alter its structural properties (as measured by SymCurv) given a single mutation anywhere within it. In this sense, robust sequences will tend to have low (strongly negative) values of *D*. By contrast, sequences under strong structural constraints will tend to have increased variance, as even single nucleotide mutations may bring about notable changes in the structural profile, and their robustness will therefore be decreased.

3.4 Nucleosomes

As the set of nucleosome coordinations we used the data from (Wal & Pugh 2012). Their map of nucleosome positions was produced by with the use of Micrococal Nuclease (Mnase) digestion followed by chromatin immunoprecipitation and facilitated library construction for deep sequencing. The set was obtained as a file in BED (Browser Extensible Data) format. For the analysis of nucleosome occupancy around TSS, we wanted to transform the data to a continuous signal like the one produced by conservation and robustness data. To do this, we first created a bedgraph file, where the presence of a nucleosome in a region was scored as 1 and the absence as 0. Then, using the wigToBigWig utility from UCSC we created a bigWig format file that was used to obtain the final nucleosomal values for the same regions used in conservation and robustness analyses around TSS.

3.5 SNPs

SNPs data used were obtained in directly from the authors of (Schacherer et al. 2007) and were processed so they can be easily parsed. SNPs coordinates were intersected with gene and intergenic regions, so every single SNP will be appointed to a specific region. For each of the gene and intergenic regions, we calculated the SNP density as follows:

Density = (total number of SNPs in a region / length of the region) * 100

3.6 Programming languages and environments

Data parsing and manipulation were performed with the use of the Perl programming language. Statistical analysis and plotting of the data were performed using the R language and its integrated development environment RStudio.

3.6.1 Perl

Perl is a family of high-level, general-purpose, interpreted, dynamic programming languages. Perl was originally developed by Larry Wall in 1987 as a general-purpose Unix scripting language to make report processing easier. Since then, it has undergone many changes and revisions. The Perl languages borrow features from other programming languages including C, shell script (sh), AWK, and sed. They provide powerful text processing facilities without the arbitrary data-length limits of many contemporary Unix command-line tools, facilitating easy manipulation of text files. Perl 5 gained widespread popularity in the late 1990s as a CGI scripting language, in part due to its unsurpassed regular expression and string parsing abilities. In addition to CGI, Perl 5 is used for graphics programming, system administration, network programming, finance,bioinformatics, and other applications.

3.6.2 R and Rstudio

R is a programming language and environment mainly used for statistical computing and graphics (R-project n.d.). R has its roots in the S language and environment which was developed at Bell Laboratories (formerly AT&T, now Lucent Technologies) by John Chambers and colleagues. R can be considered as a different implementation of S. In contrast with S, R is available as Free Software under the terms of the Free Software Foundation’s GNU General Public License in source code form. It compiles and runs on a wide variety of UNIX platforms and similar systems (including FreeBSD and Linux), Windows and MacOS. R provides a wide variety of statistical (linear and nonlinear modeling, classical statistical tests, time-series analysis, classification, clustering) and graphical techniques, and is highly extensible. Because of all that and its usability and flexibility, R is widely used in the field of bioinformatics, for the analysis and visualization of biological data and the construction of biological models. The R bioinformatics community is highly active and constantly developing and many R packages have been created for the specific use with bioinformatics data.

4. Discussion

Bidirectional gene pairs and promoters have been found to be an integral part of the genome in a wide range of organisms and species. Although many studies have focused on the characterization of these genomic elements, especially in the human genome, detailed information about the existence of constraints in their sequences and structures has not yet been described. Trying to address this problem we performed a number of analyses that were targeted to 1) the documentation of any possible differences between bidirectional genes and promoters and the other gene pair orientations and 2) a description of their unique inherent properties.

We have started our analysis by comparing the length of the intergenic sequences between the four gene organizations and found that divergent genes have a unique wider distribution of their lengths, with the sizes of the intergenic regions being slightly elevated. We proceeded using data concerning the phylogenetic conservation of these intergenic sequences and the robustness of the symmetry of DNA curvature and tried to draw conclusions on the existence of constraints in the three-dimensional structure of these regions, rather than in the linearity of their sequence. We have visualized and directly compared the conservation and robustness scores that characterize the intergenic regions in all four gene pair orientations, in the form of a continuous signal, and documented their trademark patterns.

We have deepened our comparison by focusing on the regions around the TSS of genes in tandem and divergent orientation by creating 1000bp-long regions centered on the TSSs for which we recalculated the conservation and robustness scores, as well as by integrating data from nucleosomal maps in order to produce nucleosome occupancy scores for these regions. We observed that the conservation signal follows the same trend in the two gene orientations compared, but with an interesting exception right before the TSS mark, where the conservation scores of divergent genes are higher. Regions produced based on the TSSs of divergent genes appear to have constantly higher robustness values, except after the TSS mark, when we enter the gene territory and the pattern is reversed. Nucleosome occupancy signal appears to have one major difference between the two groups right before the TSS mark, where we see that in divergent gene pairs, the pre-TSS region is more nucleosome-depleted with lower nucleosome occupancy scores.

Lastly, we made use of annotated SNPs data of the *S. cerevisiae* genome and calculated the SNPs density of both coding and intergenic regions in each of the four gene pair categories. Although coding regions were found to have similar distributions, some minor differences were found in the distributions of intergenic regions. More specifically, convergent gene pairs appear to have mostly lower SNPs density than the other three, while divergent gene pairs showed a smaller intequantile range.

This work provides some first conclusions about the distinguishing elements in the form of sequence and structural constraints of bidirectional promoters in *S. cerevisiae*, upon which future studies can be based on to extend further our knowledge around such promoters and gene pairs. The types of data used in this work can be exploited in various analyses that will complement the results presented here and fill in the missing details creating a more substantial and robust framework. Such analyses can also integrate epigenetic data, such as histone and DNA modifications, for a deeper characterization of the distinguishing elements of bidirectional promoters. Gene expression data can also be used, in order to firstly determine the percentage of gene pairs that are actually co-expressed and, secondly, focus on specific pairs of divergent formation to delineate the mechanisms of their co-regulation.

Another field of study is the role of bidirectional genes in the wider context of gene order architecture and evolution. Thus, an evolutionary approach for the understanding of the creation and maintenance of bidirectional promoters will be instructive. *S. cerevisiae* has undergone a whole genome duplication (WGD) followed by an extensive reorganization process of gene order (Wolfe & Shields 1997; Kellis et al. 2004; Dietrich et al. 2004), during which new bidirectional genes emerged (Sugino & Innan 2012). This unique evolutionary history establishes *S. cerevisiae* as a very good model for evolutionary studies that will focus on the existence, if any, and the type of natural selection forces (positive, negative or neutral) that take place and shape the characteristics of both the intergenic non-coding sequences of bidirectional promoters and the regulation of genes and coding regions of the corresponding bidirectional pairs.

All things considered, this study, by no means covers the full spectrum of analyses that will bring out the full potential of bidirectional gene pairs and promoters as systems of study of the regulatory mechanism , the functional aspects and the evolution of the genome architecture. However, we hope that our limited results will encourage future works that will enrich our basic knowledge concerning the role of bidirectional gene organization in the context of fundamental biological processes.

5. References

Adachi, N. & Lieber, M.R., 2002. Bidirectional gene organization: A common architectural feature of the human genome. *Cell*, 109(7), pp.807–809. Available at: http://www.ncbi.nlm.nih.gov/pubmed/12110178 [Accessed October 20, 2016].

Ahn, J. & Gruen, J.R., 1999. The genomic organization of the histone clusters on human 6p21.3. *Mammalian Genome*, 10(7), pp.768–770. Available at: http://www.ncbi.nlm.nih.gov/pubmed/10384058 [Accessed October 20, 2016].

Albert, I. et al., 2007. Translational and rotational settings of H2A.Z nucleosomes across the Saccharomyces cerevisiae genome. *Nature*, 446(7135), pp.572–6. Available at: http://www.nature.com/doifinder/10.1038/nature05632 [Accessed October 19, 2016].

Albig, W. et al., 1997. Human histone gene organization: nonregular arrangement within a large cluster. *Genomics*, 40(2), pp.314–322. Available at: http://www.ncbi.nlm.nih.gov/pubmed/9119399 [Accessed October 20, 2016].

Bai, L., Ondracka, A. & Cross, F.R., 2011. Multiple Sequence-Specific Factors Generate the Nucleosome-Depleted Region on CLN2 Promoter. *Molecular Cell*, 42(4), pp.465–476. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21596311 [Accessed October 20, 2016].

Burbelo, P.D., Martin, G.R. & Yamada, Y., 1988. Alpha 1(IV) and alpha 2(IV) collagen genes are regulated by a bidirectional promoter and a shared enhancer. *Proceedings of the National Academy of Sciences of the United States of America*, 85(24), pp.9679–82. Available at: http://www.ncbi.nlm.nih.gov/pubmed/3200851 [Accessed October 20, 2016].

Caserta, M. et al., 2009. A translational signature for nucleosome positioning in vivo. *Nucleic Acids Research*, 37(16), pp.5309–5321. Available at: http://www.ncbi.nlm.nih.gov/pubmed/19596807 [Accessed October 23, 2016].

Chen, Y.-Q. et al., 2010. Sorting out inherent features of head-to-head gene pairs by evolutionary conservation. *BMC bioinformatics*, 11 Suppl 1(Suppl 11), p.S16. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21172051 [Accessed October 20, 2016].

Core, L.J., Waterfall, J.J. & Lis, J.T., 2008. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science (New York, N.Y.)*, 322(5909), pp.1845–8. Available at: http://www.ncbi.nlm.nih.gov/pubmed/19056941\nhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2833333 [Accessed October 19, 2016].

Didych, D.A. et al., 2013. Human PSENEN and U2AF1L4 genes are concertedly regulated by a genuine bidirectional promoter. *Gene*, 515(1), pp.34–41. Available at: http://www.ncbi.nlm.nih.gov/pubmed/23246698 [Accessed October 20, 2016].

Eaton, M.L. et al., 2010. Conserved nucleosome positioning defines replication origins. *Genes and Development*, 24(8), pp.748–753. Available at: http://www.ncbi.nlm.nih.gov/pubmed/20351051 [Accessed October 23, 2016].

Floer, M. et al., 2010. A RSC/nucleosome complex determines chromatin architecture and facilitates activator binding. *Cell*, 141(3), pp.407–418. Available at: http://www.ncbi.nlm.nih.gov/pubmed/20434983 [Accessed October 20, 2016].

Guarguaglini, G. et al., 1997. Expression of the murine RanBP1 and Htf9-c genes is regulated from a shared bidirectional promoter during cell cycle progression. *The Biochemical journal*, 325 (Pt 1(Pt 1), pp.277–286. Available at: http://www.ncbi.nlm.nih.gov/pubmed/9224656 [Accessed October 20, 2016].

Guenther, M.G. et al., 2007. A Chromatin Landmark and Transcription Initiation at Most Promoters in Human Cells. *Cell*, 130(1), pp.77–88. Available at: http://www.ncbi.nlm.nih.gov/pubmed/17632057 [Accessed October 23, 2016].

Guertin, M.J. & Lis, J.T., 2013. Mechanisms by which transcription factors gain access to target sequence elements in chromatin. *Current Opinion in Genetics and Development*, 23(2), pp.116–123. Available at: http://www.ncbi.nlm.nih.gov/pubmed/23266217 [Accessed October 20, 2016].

Hansen, J.J. et al., 2003. Genomic structure of the human mitochondrial chaperonin genes: HSP60 and HSP10 are localised head to head on chromosome 2 separated by a bidirectional promoter. *Human Genetics*, 112(1), pp.71–77. Available at: http://www.ncbi.nlm.nih.gov/pubmed/12483302 [Accessed October 20, 2016].

Huang, C.-C. & Chang, W.-S.W., 2009. Cooperation between NRF-2 and YY-1 transcription factors is essential for triggering the expression of the PREPL-C2ORF34 bidirectional gene pair. *BMC molecular biology*, 10, p.67. Available at: http://www.ncbi.nlm.nih.gov/pubmed/19575798 [Accessed October 21, 2016].

Ioshikhes, I.P. et al., 2006. Nucleosome positions predicted through comparative genomics. *Nature genetics*, 38(10), pp.1210–1215. Available at: http://www.ncbi.nlm.nih.gov/pubmed/16964265 [Accessed October 23, 2016].

Kaplan, N., Moore, I.K., et al., 2009. The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature*, 458(7236), pp.362–366. Available at: http://www.nature.com/doifinder/10.1038/nature07667 [Accessed October 20, 2016].

Kaplan, N., Moore, I.K., et al., 2009. The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature*, 458(7236), pp.362–366. Available at: http://www.ncbi.nlm.nih.gov/pubmed/19092803 [Accessed October 23, 2016].

Koyanagi, K.O. et al., 2005. Comparative genomics of bidirectional gene pairs and its implications for the evolution of a transcriptional regulation system. *Gene*, 353(2), pp.169–176.

Lee, W. et al., 2007. A high-resolution atlas of nucleosome occupancy in yeast. *Nature genetics*, 39(10), pp.1235–44. Available at: http://www.ncbi.nlm.nih.gov/pubmed/17873876.

Li, Y.-Y. et al., 2006a. Systematic analysis of head-to-head gene organization: evolutionary conservation and potential biological relevance. *PLoS computational biology*, 2(7), p.e74. Available at: http://www.ncbi.nlm.nih.gov/pubmed/16839196 [Accessed October 19, 2016].

Li, Y.-Y. et al., 2006b. Systematic analysis of head-to-head gene organization: evolutionary conservation and potential biological relevance. *PLoS computational biology*, 2(7), p.e74. Available at: http://www.ncbi.nlm.nih.gov/pubmed/16839196 [Accessed October 20, 2016].

Lin, J.M. et al., 2007. Transcription factor binding and modified histones in human bidirectional promoters. *Genome research*, 17(6), pp.818–27. Available at: http://www.ncbi.nlm.nih.gov/pubmed/17568000 [Accessed October 21, 2016].

Liu, B., Chen, J. & Shen, B., 2011. Genome-wide analysis of the transcription factor binding preference of human bi-directional promoters and functional annotation of related gene pairs. *BMC systems biology*, 5 Suppl 1(Suppl 1), pp.S2–S2. Available at: http://bmcsystbiol.biomedcentral.com/articles/10.1186/1752-0509-5-S1-S2 [Accessed October 21, 2016].

Mavrich, T.N. et al., 2008. Nucleosome organization in the Drosophila genome. *Nature*, 453(7193), pp.358–362. Available at: http://www.ncbi.nlm.nih.gov/pubmed/18408708 [Accessed October 19, 2016].

Maxson, R. et al., 1983. Expression and organization of histone genes. *Annual review of genetics*, 17(1), pp.239–277. Available at: http://www.annualreviews.org/doi/10.1146/annurev.ge.17.120183.001323 [Accessed October 20, 2016].

Mellor, J., 2006. Dynamic nucleosomes and gene transcription. *Trends in Genetics*, 22(6), pp.320–329. Available at: http://www.ncbi.nlm.nih.gov/pubmed/16631276 [Accessed October 23, 2016].

Mito, Y., Henikoff, J.G. & Henikoff, S., 2005. Genome-scale profiling of histone H3.3 replacement patterns. *Nature genetics*, 37(10), pp.1090–1097. Available at: http://www.nature.com/doifinder/10.1038/ng1637 [Accessed October 19, 2016].

Murray, S.C. et al., 2012. A pre-initiation complex at the 3???-end of genes drives antisense transcription independent of divergent sense transcription. *Nucleic Acids Research*, 40(6), pp.2432–2444. Available at: http://www.ncbi.nlm.nih.gov/pubmed/22123739 [Accessed October 20, 2016].

Neil, H. et al., 2009. Widespread bidirectional promoters are the major source of cryptic transcripts in yeast. *Nature*, 457(7232), pp.1038–42. Available at: http://www.nature.com/doifinder/10.1038/nature07747 [Accessed October 19, 2016].

Nikolaou, C. et al., 2010. Structural constraints revealed in consistent nucleosome positions in the genome of S. cerevisiae. *Epigenetics & Chromatin*, 3(1), p.20. Available at: http://www.epigeneticsandchromatin.com/content/3/1/20.

Ogawa, R. et al., 2010. Computational prediction of nucleosome positioning by calculating the relative fragment frequency index of nucleosomal sequences. *FEBS Letters*, 584(8), pp.1498–1502. Available at: http://www.ncbi.nlm.nih.gov/pubmed/20206172 [Accessed October 23, 2016].

Ozonov, E.A. & van Nimwegen, E., 2013. Nucleosome Free Regions in Yeast Promoters Result from Competitive Binding of Transcription Factors That Interact with Chromatin Modifiers I. Ioshikhes, ed. *PLoS Computational Biology*, 9(8), p.e1003181. Available at: http://dx.plos.org/10.1371/journal.pcbi.1003181 [Accessed October 20, 2016].

Ozsolak, F. et al., 2007. High-throughput mapping of the chromatin structure of human promoters. *Nature biotechnology*, 25(2), pp.244–248. Available at: http://www.nature.com/doifinder/10.1038/nbt1279 [Accessed October 15, 2016].

Peckham, H.E. et al., 2007. Nucleosome positioning signals in genomic DNA. *Genome Research*, 17(8), pp.1170–1177. Available at: http://www.ncbi.nlm.nih.gov/pubmed/17620451 [Accessed October 23, 2016].

Preker, P. et al., 2008. RNA exosome depletion reveals transcription upstream of active human promoters. *Science (New York, N.Y.)*, 322(5909), pp.1851–4. Available at: http://www.sciencemag.org/cgi/doi/10.1126/science.1164096 [Accessed October 19, 2016].

Rhee, H.S. & Pugh, B.F., 2012. Genome-wide structure and organization of eukaryotic pre-initiation complexes. *Nature*, 483(7389), pp.295–301. Available at: http://www.ncbi.nlm.nih.gov/pubmed/22258509 [Accessed October 20, 2016].

Schacherer, J. et al., 2007. Genome-wide analysis of nucleotide-level variation in commonly used Saccharomyces cerevisiae strains J. Fay, ed. *PLoS ONE*, 2(3), p.e322. Available at: http://dx.plos.org/10.1371/journal.pone.0000322 [Accessed October 23, 2016].

Schuettengruber, B. et al., 2003. Alternate activation of two divergently transcribed mouse genes from a bidirectional promoter is linked to changes in histone modification. *Journal of Biological Chemistry*, 278(3), pp.1784–1793. Available at: http://www.ncbi.nlm.nih.gov/pubmed/12411446 [Accessed October 20, 2016].

Segal, E., Fondufe-Mittendorf, Y., Chen, L., Thåström, A., Field, Y., Moore, I.K., et al., 2006. A genomic code for nucleosome positioning. *Nature*, 442(August), pp.772–8. Available at: http://www.nature.com/doifinder/10.1038/nature04979 [Accessed October 20, 2016].

Segal, E., Fondufe-Mittendorf, Y., Chen, L., Thåström, A., Field, Y., Moore, I.K., et al., 2006. A genomic code for nucleosome positioning. *Nature*, 442(August), pp.772–8. Available at: http://www.ncbi.nlm.nih.gov/pubmed/16862119 [Accessed October 23, 2016].

Seila, A.C. et al., 2008. Divergent transcription from active promoters. *Science (New York, NY)*, 322(5909), pp.1849–1851. Available at: http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=19056940&retmode=ref&cmd=prlinks\npapers3://publication/doi/10.1126/science.1162253 [Accessed October 19, 2016].

Sekinger, E.A., Moqtaderi, Z. & Struhl, K., 2005. Intrinsic histone-DNA interactions and low nucleosome density are important for preferential accessibility of promoter regions in yeast. *Molecular Cell*, 18(6), pp.735–748. Available at: http://www.ncbi.nlm.nih.gov/pubmed/15949447 [Accessed October 20, 2016].

Shivaswamy, S. et al., 2008. Dynamic remodeling of individual nucleosomes across a eukaryotic genome in response to transcriptional perturbation. *PLoS Biology*, 6(3), pp.0618–0630. Available at: http://www.ncbi.nlm.nih.gov/pubmed/18351804 [Accessed October 23, 2016].

Siepel, A. et al., 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome research*, 15(8), pp.1034–50. Available at: http://www.ncbi.nlm.nih.gov/pubmed/16024819 [Accessed October 21, 2016].

Stein, A., Takasuka, T.E. & Collings, C.K., 2009. Are nucleosome positions in vivo primarily determined by histone-DNA sequence preferences? *Nucleic Acids Research*, 38(3), pp.709–719. Available at: http://www.ncbi.nlm.nih.gov/pubmed/19934265 [Accessed October 23, 2016].

Tilgner, H. et al., 2009. Nucleosome positioning as a determinant of exon recognition. *Nat Struct Mol Biol*, 16(9), pp.996–1001. Available at: http://www.nature.com/doifinder/10.1038/nsmb.1658 [Accessed October 23, 2016].

Trappe, R., Doenecke, D. & Albig, W., 1999. The expression of human H2A-H2B histone gene pairs is regulated by multiple sequence elements in their joint promoters. *Biochimica et Biophysica Acta - Gene Structure and Expression*, 1446(3), pp.341–351.

Trinklein, N.T. et al., 2004. An abundance of bidirectional promoters in the human genome. *Genome Research*, 14(1), pp.62–66. Available at: http://www.ncbi.nlm.nih.gov/pubmed/14707170 [Accessed October 19, 2016].

UCSC, Conservation data. Available at: http://hgdownload.soe.ucsc.edu/goldenPath/sacCer3/phastCons7way/.

UCSC, UCSC Genome Browser: bigWig Track Format. Available at: https://genome.ucsc.edu/goldenpath/help/bigWig.html.

UCSC, UCSC Table browser. *Online resource*. Available at: https://genome-euro.ucsc.edu/cgi-bin/hgTables?hgsid=218530754_LPVrgO9NzdIj4f0Y9jdL7mgArat5&clade=other&org=&db=sacCer1&hgta_group=genes&hgta_track=sgdGene&hgta_table=sgdGene&hgta_regionType=genome&position=&hgta_outputType=primaryTable&hgta_outFileName=.

Wal, M. & Pugh, B.F., 2012. Genome-wide mapping of nucleosome positions in yeast using high-resolution MNase ChIP-Seq. *Methods in Enzymology*, 513, pp.233–250. Available at: http://www.ncbi.nlm.nih.gov/pubmed/22929772 [Accessed October 22, 2016].

Wang, Q. et al., 2009. Searching for bidirectional promoters in Arabidopsis thaliana. *BMC bioinformatics*, 10 Suppl 1(Suppl 1). Available at: http://www.biomedcentral.com/1471-2105/10/S1/S29 [Accessed October 19, 2016].

Woo, Y.H. & Li, W.-H., 2011. Gene clustering pattern, promoter architecture, and gene expression stability in eukaryotic genomes. *Proceedings of the National Academy of Sciences of the United States of America*, 108(8), pp.3306–11. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21300882 [Accessed October 21, 2016].

Xu, Z. et al., 2009. Bidirectional promoters generate pervasive transcription in yeast. *Nature*, 457(7232), pp.1033–7. Available at: http://www.nature.com/doifinder/10.1038/nature07728 [Accessed October 19, 2016].

Yin, S. et al., 2009. The impact of nucleosome positioning on the organization of replication origins in eukaryotes. *Biochemical and Biophysical Research Communications*, 385(3), pp.363–368. Available at: http://www.ncbi.nlm.nih.gov/pubmed/19463783 [Accessed October 23, 2016].

Yuan, G.-C. et al., 2005a. Genome-scale identification of nucleosome positions in S. cerevisiae. *Science (New York, N.Y.)*, 309(5734), pp.626–30. Available at: http://www.ncbi.nlm.nih.gov/pubmed/15961632 [Accessed October 19, 2016].

Yuan, G.-C. et al., 2005b. Genome-scale identification of nucleosome positions in S. cerevisiae. *Science (New York, N.Y.)*, 309(5734), pp.626–30. Available at: http://www.ncbi.nlm.nih.gov/pubmed/15961632 [Accessed October 15, 2016].

Zhang, Y. et al., 2009. Intrinsic histone-DNA interactions are not the major determinant of nucleosome positions in vivo. *Nature structural & molecular biology*, 16(July), pp.1–7. Available at: http://www.nature.com/doifinder/10.1038/nsmb.1636 [Accessed October 20, 2016].

6. Supplementary Figures

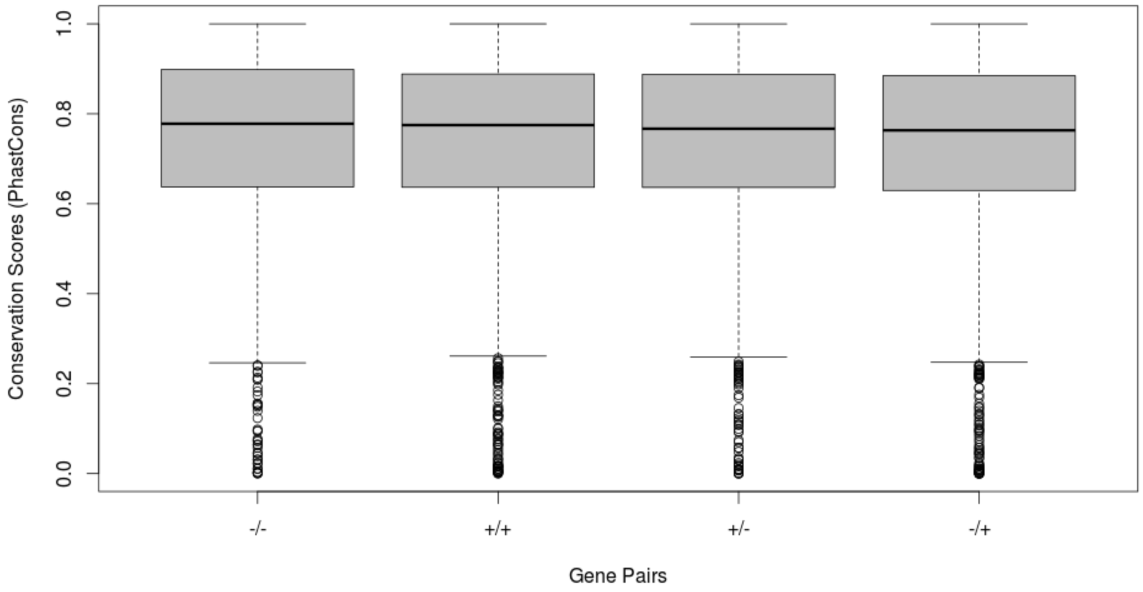

Figure 1: Distributions of PhastCons conservation scores (y axis) of genes in each orientation (x axis). As expected, regions of genes are highly conserved and appear to have no difference in their distributions of PhastCons scores.

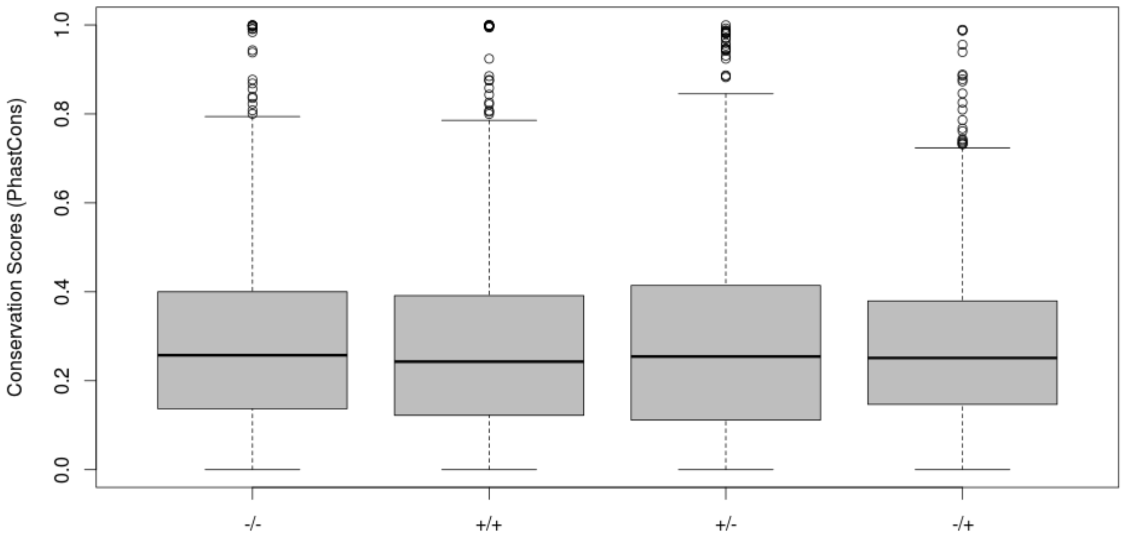

FIgure 2: Distributions of PhastCons conservation scores (y axis) of intergenic regions in each gene pair orientation (x axis).

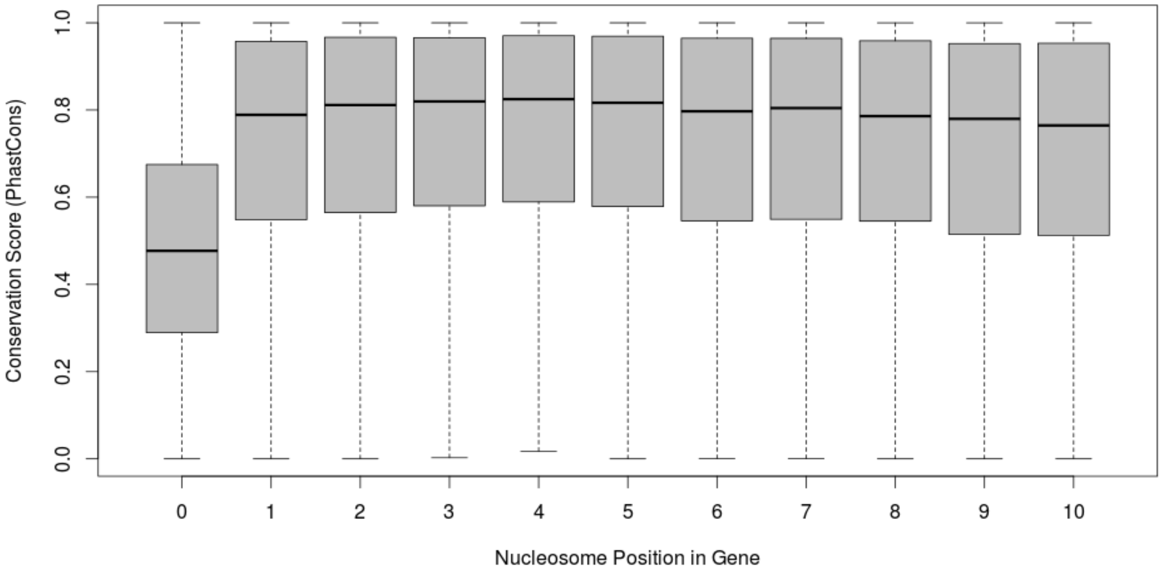

Figure 3: Distributions of PhastCons mean conservation scores (y axis) of the nucleosomal regions within genes. Using the nucleosomal coordinates we positioned each nucleosome in the linear scale according to its proximity to TSS. As zero (0) we appointed any nucleosome that was found to be overlapping with the TSS.

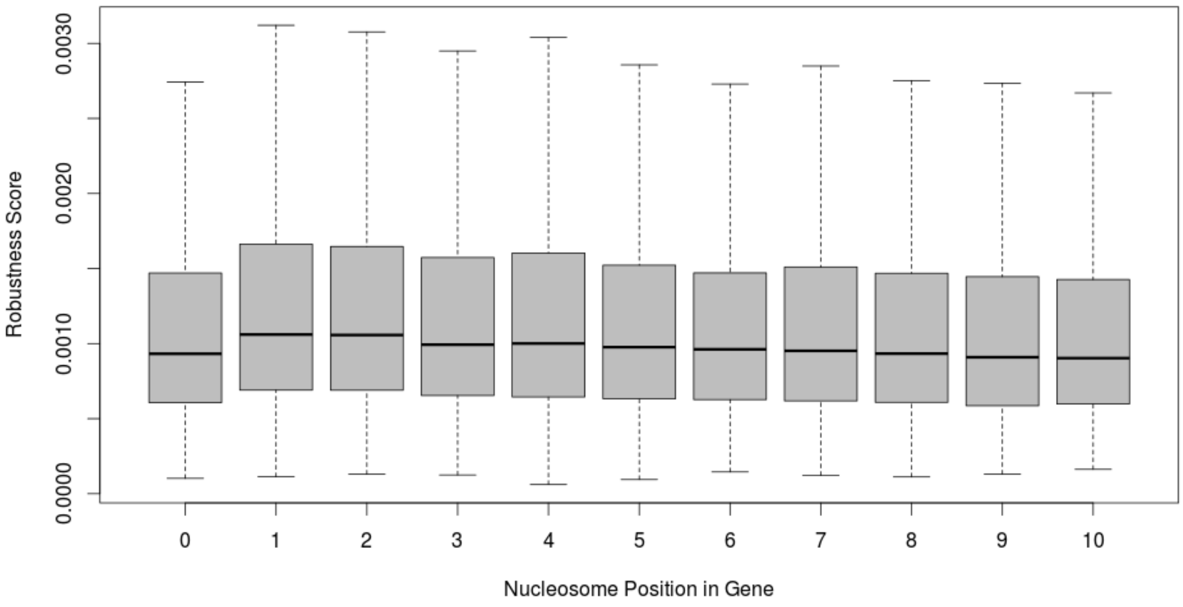

Figure 4: Distributions of PhastCons mean robustness scores (y axis) of the nucleosomal regions within genes. Using the nucleosomal coordinates we positioned each nucleosome in the linear scale according to its proximity to TSS. As zero (0) we appointed any nucleosome that was found to be overlapping with the TSS.

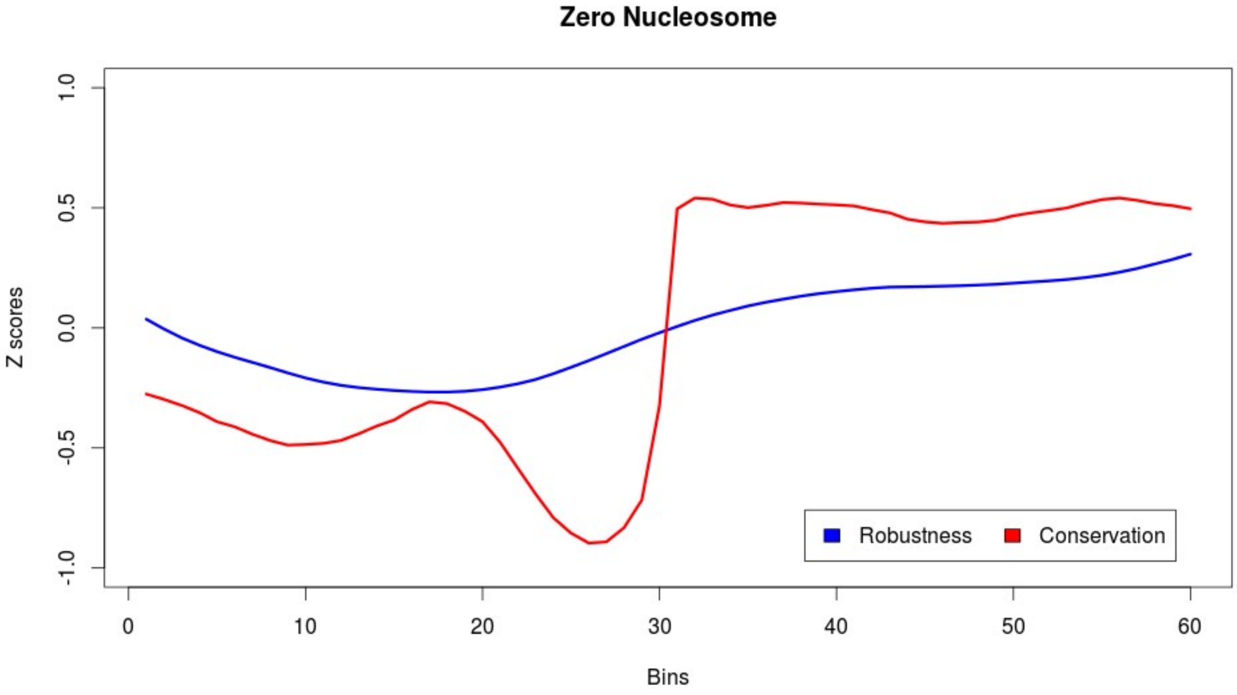

Figure 5: Mean PhastCons conservation and robustness scores around zero nucleosome. Using the nucleosomal coordinates we positioned each nucleosome in the linear scale according to its proximity to TSS. As zero (0) we appointed any nucleosome that was found to be overlapping with the TSS. The figure shows the mean conservation (red) and robustness (blue) values of a 600bp-long region around zero nucleosome. In the x axis we have the number of bins used for the division of the region, where every bin unit corresponds to 10bp in the real genome. The nucleosome scores (y axis) were generated by the bigWigSummary utility. Zero nucleosome is positioned in the middle of x axis (50).

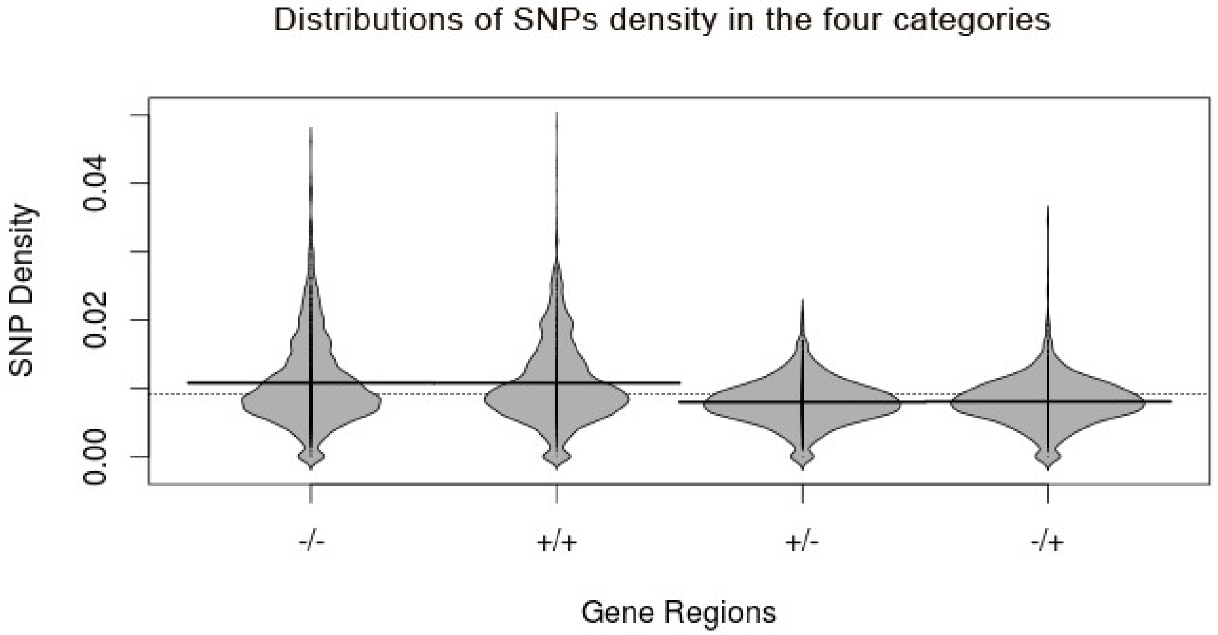

Figure 6: Gene SNPs density distributions of the four gene orientations. Bean-plots show the SNPs density of the genes (y axis) in every gene orientation (x axis).

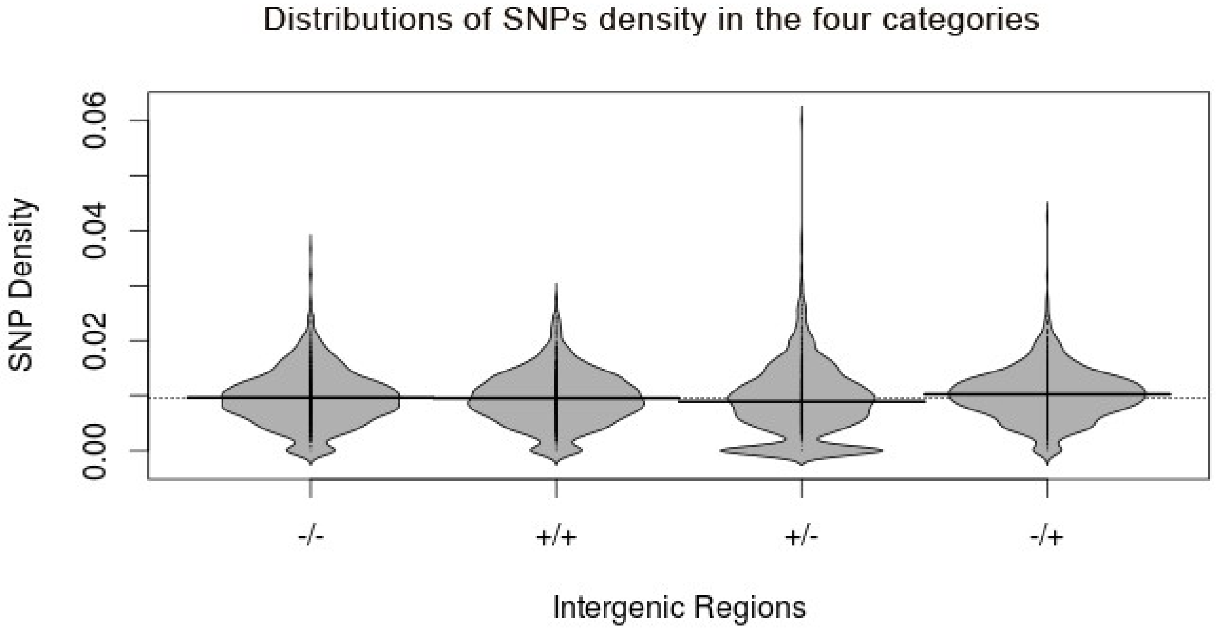

Figure 7: Intergenic SNPs density distributions of the four gene orientations. Bean-plots show the SNPs density of the intergenic regions (y axis) in every gene orientation (x axis).