nature genetics

Gene expression divergence in yeast is coupled to evolution of DNA-encoded nucleosome organization

Yair Field[1,5], Yvonne Fondufe-Mittendorf[2,5], Irene K Moore[2], Piotr Mieczkowski[3], Noam Kaplan[1], Yaniv Lubling[1], Jason D Lieb[3], Jonathan Widom[2] & Eran Segal[1,4]

Eukaryotic transcription occurs within a chromatin environment, whose organization has an important regulatory function and is partly encoded in *cis* by the DNA sequence itself. Here, we examine whether evolutionary changes in gene expression are linked to changes in the DNA-encoded nucleosome organization of promoters. We find that in aerobic yeast species, where cellular respiration genes are active under typical growth conditions, the promoter sequences of these genes encode a relatively open (nucleosome-depleted) chromatin organization. This nucleosome-depleted organization requires only DNA sequence information, is independent of any cofactors and of transcription, and is a general property of growth-related genes. In contrast, in anaerobic yeast species, where cellular respiration genes are relatively inactive under typical growth conditions, respiration gene promoters encode relatively closed (nucleosome-occupied) chromatin organizations. Our results suggest a previously unidentified genetic mechanism underlying phenotypic diversity, consisting of DNA sequence changes that directly alter the DNA-encoded nucleosome organization of promoters.

Changes in transcriptional regulation are important for generating phenotypic diversity among species, but the mechanisms underlying these regulatory changes are not well understood. Consistent with the centrality of transcription factors to transcriptional control, some phenotypic changes have been associated with changes in the binding-site content of promoters[1] or with changes in the targets bound by transcription factors[2]. However, modulation of other processes key to transcriptional regulation may also lead to phenotypic diversity. Recent studies that measured nucleosome occupancy genome-wide have revealed strong associations between chromatin organization and gene expression[3–9], and other studies have shown that the organization of nucleosomes is partly encoded in the genome through the sequence preferences of nucleosomes[3,10–13]. However, the relationship between evolutionary changes in DNA-encoded nucleosome organization and expression divergence has not been examined.

Here, we study the relationship between gene expression and the DNA-encoded nucleosome organization of promoters across two yeast species, the budding yeast *Saccharomyces cerevisiae* and the human pathogen *Candida albicans*, for which large compendia of gene expression data are available. These species show several phenotypic differences. Most notably, in high glucose, *C. albicans* grows by respiration and correspondingly activates transcription of genes required for the TCA cycle and oxidative phosphorylation, whereas *S. cerevisiae* grows primarily by fermentation and correspondingly reduces transcription of respiration genes. We henceforth term the respiratory growth 'aerobic' and the fermentative growth 'anaerobic'. Our approach consists of three steps. First, we quantify the extent to which the gene expression patterns of biologically meaningful gene sets are conserved across the two species. Next, we examine the DNA-encoded nucleosome organization of these gene sets using both a computational model and *in vitro* reconstitutions of nucleosome on purified DNA from each species. Finally, we test whether orthologous gene sets with divergent expression patterns between the two species show corresponding changes in their DNA-encoded nucleosome organization.

RESULTS

Expression changes linked to aerobic versus anaerobic growth

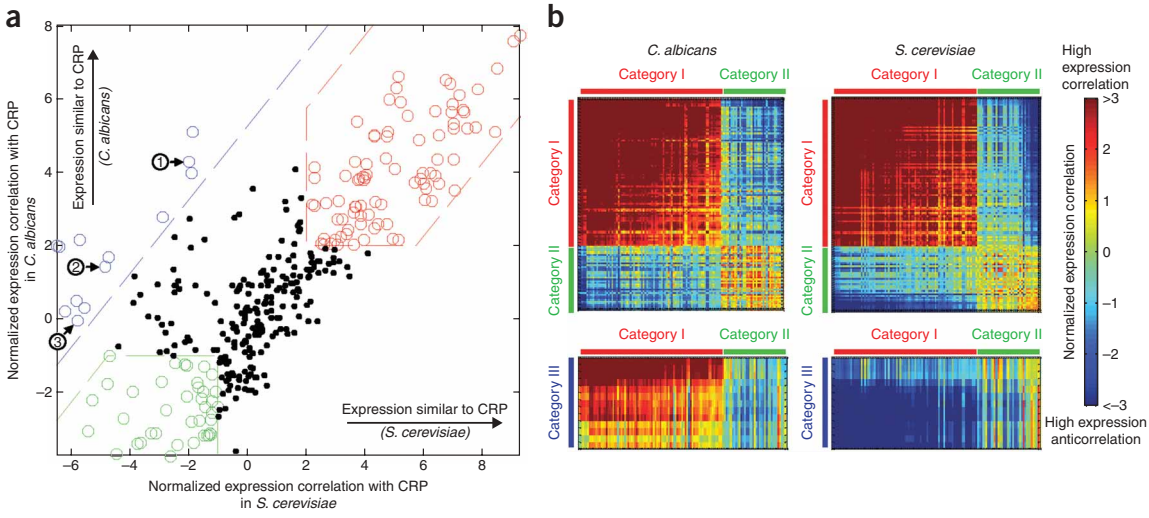
We downloaded two large collections of microarray-based gene expression data from ~1,000 and ~200 different cellular states and environmental conditions in *S. cerevisiae* and *C. albicans*, respectively, compiled in ref. 1. To compare the expression patterns of orthologous genes, we downloaded a yeast orthology map[14] and quantified the degree to which the co-expression relationships of a gene in one species are similar to the co-expression relationships of its orthologous counterpart in the other species. Such an approach has been successfully used to compare expression patterns across distant species[15,16].

[1]Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot, 76100, Israel. [2]Department of Biochemistry, Molecular Biology, and Cell Biology, Northwestern University, 2153 Sheridan Road, Evanston, Illinois 60208, USA. [3]Department of Biology, Carolina Center for Genome Sciences, and Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, USA. [4]Department of Molecular Cell Biology, Weizmann Institute of Science, Rehovot, 76100, Israel. [5]These authors contributed equally to this work. Correspondence should be addressed to E.S. (eran@weizmann.ac.il) or J.W. (j-widom@northwestern.edu).

Received 14 August 2008; accepted 12 January 2009; published online 1 March 2009; doi:10.1038/ng.324

To obtain insights at the level of biological processes, we used biologically meaningful gene sets from Gene Ontology[17] as the basic units of analysis[18], and restricted ourselves to only the 796 gene sets (of the total 2,152 gene sets) in which the average normalized correlation between all pairs of its member genes was above 0.5 in both expression compendia (Methods). We anchored our analysis around the cytosolic ribosomal protein (CRP)-encoding genes, as these genes have coherent expression patterns across many conditions[19] and their expression shows strong associations with cellular growth[20]. For each of the above 796 gene sets, we then computed, separately in each species, the average normalized correlation between the expression of every gene within the set and the expression of every one of the CRP genes (Methods). In this measure, gene sets that are active under typical growth conditions will have a high (positive) normalized correlation to the CRP genes, whereas gene sets that are inactive under typical growth conditions will have a low (negative) normalized correlation to the CRP genes.

Comparing these expression correlations for every gene set between the two species, we found three main categories of gene sets (**Fig. 1** and **Supplementary Table 1** online). The first ('category I') consists of 89 gene sets (totaling 1,333 and 1,321 genes in *S. cerevisiae* and *C. albicans*, respectively) whose expression in both species is highly correlated to that of the CRP genes. Many gene sets in this category are indeed related to cellular growth, including the CRP genes themselves (by construction), and genes involved in amino acid biosynthesis pathways and RNA processing. The second category ('category II') consists of 40 gene sets (447 and 448 genes in *S. cerevisiae* and *C. albicans*, respectively) whose expression in both species shows a strong anti-correlation with the expression of the CRP genes. This category includes many gene sets that are activated only in specific cellular states, and gene sets that are induced in response to environmental stress conditions[19], such as proteasome-, autophagy- and mating-related genes. The third category ('category III') consists of 13 gene sets (157 and 164 genes in *S. cerevisiae* and *C. albicans*, respectively) whose transcriptional program diverged between the two species, such that the correlation between the expression of their member genes and the expression of the CRP genes is much higher in *C. albicans* than in *S. cerevisiae*. This category includes gene sets related to cellular respiration and mitochondrial functions, such as the TCA cycle, oxidative phosphorylation and mitochondrial ribosomal genes. The



c

Category I (red)	Category II (green)	Category III (blue)
89 conserved CRP-like expression programs (total 1321/1333 genes in *C. albicans*/*S. cerevisiae*), including:	40 conserved anti-CRP-like expression programs (total 448/447 genes in *C. albicans*/*S. cerevisiae*), including:	13 diverged expression programs, with higher expression correlation to CRP genes in *C. albicans* (total 164 genes) than in *S. cerevisiae* (total 157 genes), including:
• cytosolic ribosome (CRP)	• autophagy	• mitochondrial ribosome (MRP)
• rRNA modification	• energy reserve metabolic process	① 57/56 genes in *C. albicans*/*S. cerevisiae*
• RNA processing	• mitotic recombination	• electron transport
• translation	• mating-type switching	② 18/18 genes in *C. albicans*/*S. cerevisiae*
• DNA-directed RNA pol I complex	• antioxidant activity	• TCA cycle intermediate metabolic process
• nucleobase biosynthetic process	• proteasome complex	③ 25/17 genes in *C. albicans*/*S. cerevisiae*
• steroid biosynthetic process	• condensed chromosome	
• snoRNA metabolic process	• kinetochore	

Figure 1 *S. cerevisiae* and *C. albicans* show large-scale changes in the transcriptional programs of cellular respiration and mitochondrial function genes. (**a**) For gene sets from Gene Ontology, shown is the average normalized Pearson correlation (see Methods) between the expression of their member genes and the expression of the cytosolic ribosomal protein (CRP) genes, computed separately for the expression compendia of *C. albicans* (*y* axis) and *S. cerevisiae* (*x* axis). Only gene sets that show coherent co-expression in both species are shown, where we define coherently co-expressed gene sets as those in which the average normalized correlation between its member genes is above 0.5. From these gene set expression correlation measures, we define three categories of gene sets (category I, red, high correlation with CRP genes in both species; category II, green, anticorrelation with CRP genes in both species; and category III, blue, higher correlation with CRP genes in *C. albicans* than in *S. cerevisiae*). Three gene sets from category III are numbered for reference in other figures. (**b**) Shown are the normalized expression correlations, computed as in **a**, between every pair of gene sets from all three categories defined in **a**. Note the expression divergence of gene sets from category III (blue), which show strong expression correlation with category I gene sets from *C. albicans*, but strong anticorrelation with category I gene sets from *S. cerevisiae*. (**c**) A subset of the list of gene sets in each of the categories defined in **a**, along with the number of gene sets and number of genes in each category. More details are given for the three numbered categories from **a**. See **Supplementary Table 1** for the full list.

expression divergence of a subset of these genes was reported previously[1], and reflects the difference in respiratory versus fermentative growth preferences between the species.

DNA-encoded nucleosome organization and expression changes

To study the relationship between transcriptional programs and chromatin organization, we examined DNA-encoded nucleosome organization over the promoter regions of the gene sets in each of the three categories, both experimentally and using a computational model of the nucleosome sequence preferences[21]. These sequence preferences are represented by a probability distribution over nucleosome-length sequences, estimated from a large set of fully sequenced *in vivo* nucleosomes from *S. cerevisiae*. The model uses this distribution to compute the probability that each base pair in the genome is covered by a nucleosome, in an assumed equilibrium between all competing nucleosome configurations. Using a cross-validation scheme, this sequence-based model was shown to be highly predictive of the experimentally measured nucleosome organization, suggesting that nucleosome organization is partly encoded in *cis* by the DNA, and that we can reliably use this model to examine the DNA-encoded nucleosome organization (see ref. 21 for an overview and evaluation of this model).

We used this model to compute the occupancy over the nucleosome-depleted region of every promoter in each of the two species. We focused on the 200 bp upstream of the translation start site, as *in vivo* measurements of nucleosome occupancy showed that promoters show a stereotyped depleted region of length ~100–150 bp within the 200 bp upstream of the translation start site[3–6,9]. We defined the occupancy over this promoter nucleosome-depleted region (henceforth termed 'PNDR') as the lowest average nucleosome occupancy across any 100-bp region in the 200 bp upstream of the translation start site. Other parameter choices that we tested for the region (in the range of 100–150 bp for the width of the least-occupied region and 200–400 bp for the overall length of the upstream region) gave equivalent results. Thus, when the PNDR score of each gene is computed by the model, it represents a predicted measure of the degree to which the gene's promoter encodes an open (nucleosome-depleted) or closed (nucleosome-occupied) nucleosome organization.

To test whether the DNA sequences of promoters from a given gene set encode a relatively open or closed nucleosome organization, we compared, separately for each species, the PNDR scores of the gene set's promoters to the PNDR scores of all other promoters. Specifically, we ranked all promoters by their PNDR score and measured the relative ranking of the gene set's promoters using a normalized Mann-Whitney rank statistic, which is equal to the area under the curve[22] (AUC) when plotting the fraction of the gene set's promoters above a given PNDR score versus the fraction of all other promoters above that PNDR score, for all possible PNDR values (**Fig. 2a**). In this measure, a gene set with a relatively closed promoter organization, in which every promoter has a PNDR score above that of every other promoter, will receive an AUC score of 1. A gene set in which every promoter has a PNDR score below that of every other promoter will receive an AUC score of 0 (relatively open nucleosome organization), and a gene set composed of randomly selected promoters set will receive, on average, an AUC score of 0.5.

For each gene set from the three categories above, defined solely by their expression profiles, we then compared the predicted PNDR AUC value in *S. cerevisiae* to the AUC in *C. albicans* (**Fig. 2b,c**). Notably, in both species, the growth-related gene sets of category I, whose expression profile in both species is highly correlated to that of the CRP genes, have AUC scores significantly lower than all other gene sets ($P < 10^{-13}$ and $P < 10^{-9}$ in Student's *t*-test for *S. cerevisiae* and *C. albicans*, respectively), indicating that their promoters encode relatively open nucleosome architectures. Conversely, the condition-specific gene sets of category II, in which expression is anti-correlated to that of the CRP genes in both species, have AUC scores significantly higher than all other gene sets ($P < 10^{-5}$ and $P < 10^{-18}$), indicating that their promoters encode relatively closed nucleosome architectures. These results suggest that both *S. cerevisiae* and *C. albicans* preserve a system-level relationship between transcriptional programs and DNA-encoded nucleosome organizations, whereby promoters of growth-related genes encode relatively open nucleosome organizations, whereas promoters of condition-specific genes encode relatively closed nucleosome organizations.

In contrast to the largely conserved nucleosome organization of gene sets from the first two categories, the aerobic cellular respiration gene sets of category III show many changes between the two species in the DNA-encoded nucleosome organization over their promoters. In *C. albicans*, aerobic respiration gene sets (category III) have AUC values significantly lower than all other gene sets ($P < 0.005$ in Student's *t*-test) and thus their promoters encode relatively open nucleosome organizations, whereas in *S. cerevisiae*, these aerobic respiration gene sets have AUC values significantly above all other gene sets ($P < 10^{-6}$) and thus their promoters encode relatively closed chromatin architectures. Notably, these changes in the DNA-encoded nucleosome organization are coupled to the expression divergence that the category III gene sets show between the two species, in a manner that may facilitate the transcriptional program of each species. Category III gene sets, which have higher expression correlation to growth-related genes in *C. albicans* than in *S. cerevisiae*, encode a relatively open nucleosome organization in *C. albicans*, in accordance with the trend observed for growth-related gene sets (category I), and a relatively closed nucleosome organization in *S. cerevisiae*, in accordance with the trend observed for gene sets whose expression is anticorrelated to growth-related genes (category II).

These results demonstrate that a global relationship between transcriptional programs and the DNA-encoded nucleosome organizations is conserved across two yeast species, even in the presence of expression divergence. Our results thus suggest a conserved design principle of transcriptional regulation in yeast, whereby the default repression of condition-specific genes (such as the aerobic respiration genes of *S. cerevisiae*) is facilitated by the relatively closed nucleosome organization encoded over their promoters. In conditions where activation of these genes is required, this repression is actively alleviated, presumably by the combined action of transcription factors and chromatin remodeling complexes. In contrast, for growth-related genes most commonly used by the organism (such as the aerobic respiration genes of *C. albicans*), the repression by nucleosomes is by default alleviated through the encoding of relatively open nucleosome organizations over their promoters. We note that although this global trend is strong in our analysis, it clearly does not apply to every growth-related or condition-specific gene set or individual gene within them, as some of these gene sets show moderate AUC values.

The same behavior was also evident when we created a single gene set for each of the three categories, consisting of all the genes from the gene sets of that category. In both species, when we plotted the average nucleosome occupancy predicted by the model across the promoters, we found stronger predicted nucleosome depletion (lower PNDR score) in category I promoters relative to category II promoters ($P < 10^{-6}$ and $P < 10^{-9}$ in Student's *t*-test for *S. cerevisiae* and

C. albicans, respectively). However, the promoters of genes involved in cellular respiration from category III differ in their average nucleosome occupancy between the two species, such that in *S. cerevisiae*, they encode the most closed nucleosome organization of all three categories (**Fig. 3a**), but in *C. albicans*, they encode the most open nucleosome organization of all three categories (**Fig. 3b**). Indeed, the category III promoters have a significantly more closed nucleosome organization in *S. cerevisiae* than in *C. albicans*, as the average difference in the predicted PNDR score between the two species is significantly higher than the difference obtained when randomly choosing the same number of promoters ($P < 10^{-4}$, 10,000 permutation tests).

Validation by assembly of nucleosomes on naked genomic DNA

As a direct experimental validation of the model predictions, we purified chicken erythrocyte histone octamers and assembled them on purified genomic DNA from both *S. cerevisiae* and *C. albicans* by salt gradient dialysis[23]. We then isolated mononucleosomes by standard micrococcal nuclease digestion, and used parallel sequencing to determine nucleosome positions. In each species, we carried out two completely independent experiments, and mapped ~10 million reconstituted nucleosomes. The resulting data provide a genome-wide map in each species, in which nucleosome positions are governed only by the intrinsic sequence preferences of nucleosomes[24]. For each

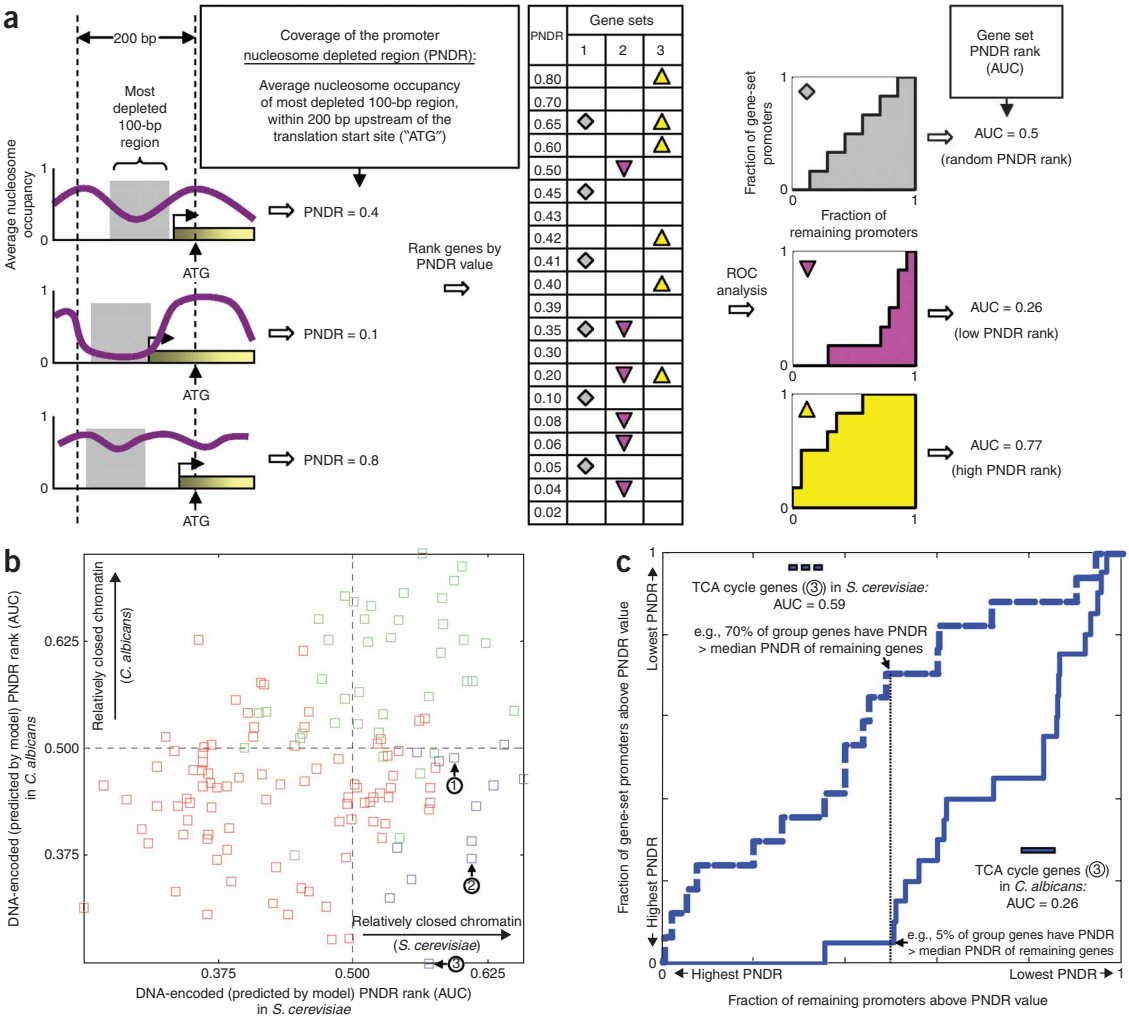


Figure 2 The expression divergence of cellular respiration genes is accompanied by changes in the DNA-encoded nucleosome organization of their promoters. (**a**) A toy example illustrating the rank statistic used to assess whether the DNA-encoded nucleosome organization of promoters of a given gene set encodes a relatively open or relatively closed organization. For each gene, we use the model of the nucleosome sequence preferences[21] to compute the DNA-encoded nucleosome coverage over the nucleosome-depleted region of its promoter (left, termed PNDR) and rank all genes by these PNDR scores (middle table; values were chosen arbitrarily for illustration). The rank statistic of each gene set is then obtained by computing the area under the curve (AUC) in a graph that plots the fraction of promoters from the gene set (*y* axis) that are above a certain PNDR score versus the fraction of all other promoters above that PNDR score, for all possible PNDR values (three plots on right). Plots for three sample gene sets are shown: a gene set with a random AUC rank of 0.5 (top, gray); a gene set whose genes have relatively open nucleosome organizations and thus relatively low PNDR scores (middle, pink); and a gene set whose genes have relatively closed nucleosome organizations and thus relatively high PNDR scores (bottom, yellow). ROC, receiver operating characteristic. (**b**) For every gene set from the three categories defined in **Figure 1a**, shown are its PNDR rank statistic, computed as explained in **a**, in both *C. albicans* (*y* axis) and *S. cerevisiae*. Gene sets from each category are colored as in **Figure 1a**. The three numbered gene sets from category III in **Figure 1a** are numbered here as well. (**c**) Example of an AUC plot for one of the gene sets from **b** (the TCA cycle gene set), in both *S. cerevisiae* (dashed line) and *C. albicans* (full line). The promoters of the gene set in this example have relatively high PNDR scores in *S. cerevisiae* and thus encode relatively closed nucleosome organizations, whereas in *C. albicans*, they have relatively low PNDR scores and thus encode relatively open nucleosome organizations. For example, only ~5% of the promoters in this gene set have higher PNDR scores than the PNDR score that is exceeded by ~50% of the promoters in *C. albicans*.

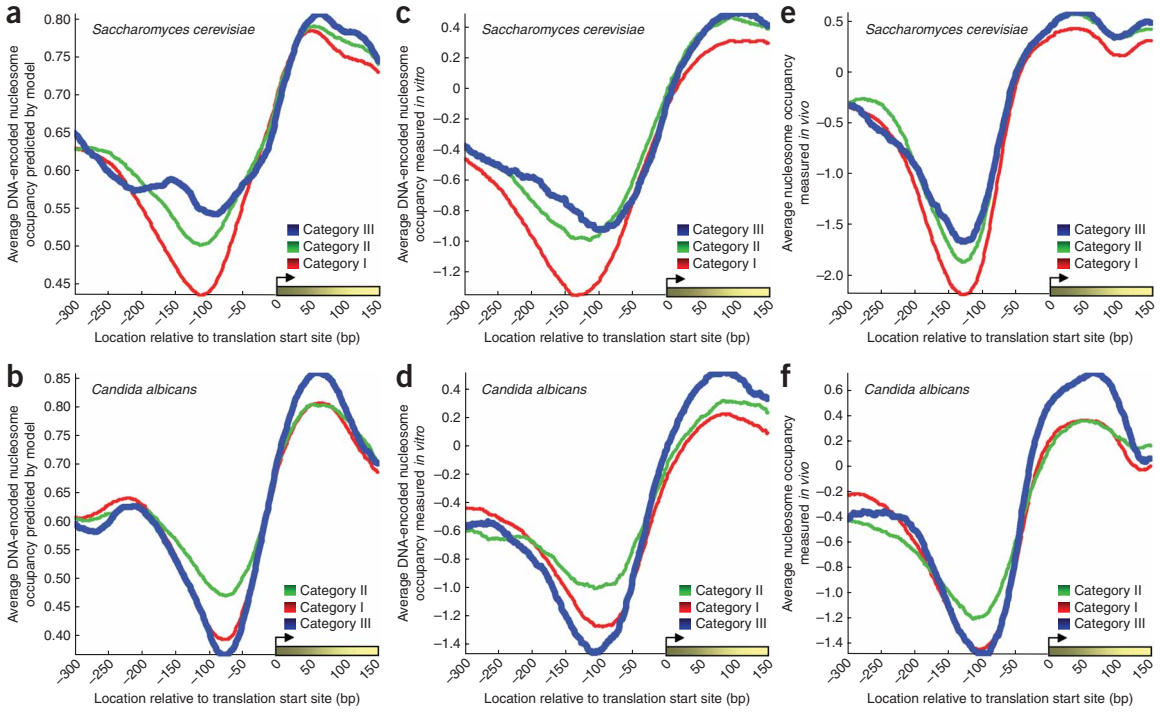


Figure 3 The DNA-encoded nucleosome organization of promoters driving genes involved in cellular respiration has diverged between *S. cerevisiae* and *C. albicans*. (**a**) For each of the three categories defined in **Figure 1a**, we created a single gene set per category that consists of all genes from all gene sets of that category. Shown is the average nucleosome occupancy in *S. cerevisiae*, predicted by our sequence-based model of nucleosome sequence preferences, across this unionized gene set per category. Average occupancy profiles are shown relative to the translation start site of the corresponding genes (we used translation start sites because transcription start sites are not well-annotated genome-wide for *C. albicans*). (**b**) Same as **a**, but for *C. albicans*. (**c**) Same as **a**, but using the *in vitro* map of nucleosome occupancy that we measured in *S. cerevisiae*. (**d**) Same as **c**, but for *C. albicans*. (**e**) Same as **a**, but using the *in vivo* map of nucleosome positions that we measured in *S. cerevisiae*. (**f**) Same as **e**, but for *C. albicans*.

map, we calculated the average nucleosome occupancy at every base pair as the log-ratio of the number of reads that cover that base pair and the median number of reads per base pair across the genome. The independent replicates of each species were in excellent agreement, so we averaged the replicates within each species to create two *in vitro* nucleosome occupancy maps, one in *S. cerevisiae* and one in *C. albicans*.

As a first validation, we compared the PNDR scores predicted by the model for each promoter, on which our above AUC analyses are based, to the PNDR scores computed from the *in vitro* maps. We found that these scores are in good agreement, with an overall correlation of 0.76 and 0.72 between the model PNDR scores and the PNDR scores computed from the *in vitro* maps in *S. cerevisiae* and *C. albicans*, respectively. Moreover, we found a correlation of 0.70 between the model-predicted and data-measured divergence in PNDR scores per promoter between *S. cerevisiae* and *C. albicans* (**Supplementary Fig. 1** online). We next examined the average nucleosome occupancy measured by these *in vitro* maps across the promoters of each of our three categories, and found that they are highly similar to those predicted by the model (**Fig. 3c,d**). As predicted by the model, the *in vitro* maps show stronger nucleosome depletion in category I promoters relative to category II promoters ($P < 10^{-4}$ and $P < 10^{-6}$ in Student's *t*-test for *S. cerevisiae* and *C. albicans*, respectively). The occupancy profiles of aerobic respiration promoters (category III) in the *in vitro* maps also agree with the model predictions, showing the most closed and open nucleosome organization of all three categories in *S. cerevisiae* and in *C. albicans*, respectively (**Fig. 3c,d**), and a significantly more closed nucleosome organization in *S. cerevisiae* than in *C. albicans* ($P < 10^{-4}$ in 10,000 permutation tests). A model constructed from the *S. cerevisiae in vitro* data[24] yielded equivalent predictions to those of the *in vivo*–based model that we use here (**Supplementary Fig. 2** online). Thus, in accordance with the model predictions, these *in vitro* nucleosome occupancy maps demonstrate that evolutionary changes in the DNA sequence of aerobic respiration gene promoters contributed to the divergence of nucleosome organization at these promoters in *S. cerevisiae* and *C. albicans*.

In vivo and DNA-encoded nucleosome organization are similar

Next, we tested whether the *in vivo* nucleosome organization of promoters from each of the above three categories is similar to their DNA-encoded organization, as predicted by the model and measured by the *in vitro* maps. To this end, we isolated mononucleosomes from both *S. cerevisiae* and *C. albicans* each cultured in their own 'normal' growth conditions (Methods), and used parallel sequencing to obtain genome-wide maps of *in vivo* nucleosome positions. The maps consist of ~10 million individual nucleosome reads in each species. We carried out two completely independent experiments in each species, calculated the average nucleosome occupancy at every base pair, and averaged the highly similar replicates within each species to create two *in vivo* nucleosome occupancy maps, one in *S. cerevisiae* and one in *C. albicans*.

We subjected these *in vivo* maps to the same tests that we had done for the *in vitro* maps and found that, indeed, the nucleosome organization of promoters *in vivo* is highly similar to the DNA-encoded nucleosome organization predicted by the model and measured by the *in vitro* maps[24]. As expected, the agreement between

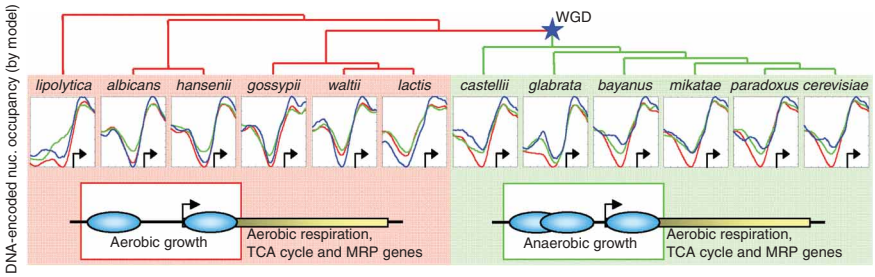


Figure 4 The emergence of anaerobic yeast species coincides with an evolutionary change in the DNA-encoded nucleosome organization of cellular respiration gene promoters. Shown is the average nucleosome occupancy, predicted by our model of nucleosome sequence preferences, across all genes from each of the three categories of gene sets defined in **Figure 1a**, for each of 12 different yeast species whose genomic sequence is available. Average nucleosome occupancy profiles are shown relative to the translation start site of the corresponding genes in each species, color-coded as in **Figure 3**. Yeast species are organized according to their phylogenetic tree, aerobic and anaerobic yeast species are indicated by pink (left) and green (right) boxes, respectively, and the point in evolution where the apparent whole-genome duplication event (WGD) has occurred is indicated (blue star). Note that in all of the aerobic yeast species, promoters of category III genes encode a relatively more open nucleosome organization compared to promoters of category II genes, whereas in all of the anaerobic yeast species, the situation is reversed, and promoters of category III genes encode a relatively more closed nucleosome organization compared to promoters of category II genes. MRP, mitochondrial robosomal proteins.

the model predictions and the *in vitro* maps is higher than the agreement between the model predictions and the *in vivo* maps. The PNDR scores predicted by the model for each promoter and those computed from the *in vivo* maps in each species are nonetheless in good agreement, with a correlation of 0.62 and 0.63 in *S. cerevisiae* and *C. albicans*, respectively. Similarly, there is good agreement (correlation = 0.60) between the model-predicted and data-measured difference in PNDR scores per promoter between the two species (**Supplementary Fig. 3** online). The *S. cerevisiae* model-predicted PNDR scores are also in agreement with other published nucleosome occupancy maps (**Supplementary Table 2** online). In accordance with the model predictions and the *in vitro* maps, the *in vivo* maps also reveal stronger nucleosome depletion in category I promoters relative to category II promoters in both species (**Fig. 3e**,**f**; $P < 10^{-3}$ and $P < 10^{-4}$ in Student's *t*-test in *S. cerevisiae* and *C. albicans*, respectively). Similarly, the *in vivo* maps indicate that promoters of genes involved in aerobic respiration (category III) have the most closed and open nucleosome organization of all three categories in *S. cerevisiae* and in *C. albicans*, respectively (**Fig. 3e**,**f**), and that they have a significantly more closed nucleosome organization in *S. cerevisiae* than in *C. albicans* ($P < 0.01$ in 10,000 permutation tests).

DNA-encoded nucleosome organization across 12 yeast species

To obtain a broader evolutionary perspective, we used our model to examine the DNA-encoded nucleosome organization of promoters in ten additional yeast species. Notably, we found that the relation between the DNA-encoded nucleosome organization of promoters from category I and category II is conserved across all of the yeast species that we examined. In all species, promoters of the growth-related genes from category I encode relatively open chromatin organizations, whereas promoters of condition-specific genes from category II encode relatively closed chromatin organizations. In contrast, the nucleosome organization of the promoters of respiration-related genes from category III has diverged in evolution, exactly at the point in which the yeast species that we examined show phenotypic divergence between aerobic and anaerobic growth. Specifically, the promoters of these genes are predicted to encode relatively open nucleosome organizations in all of the aerobic yeast species, and

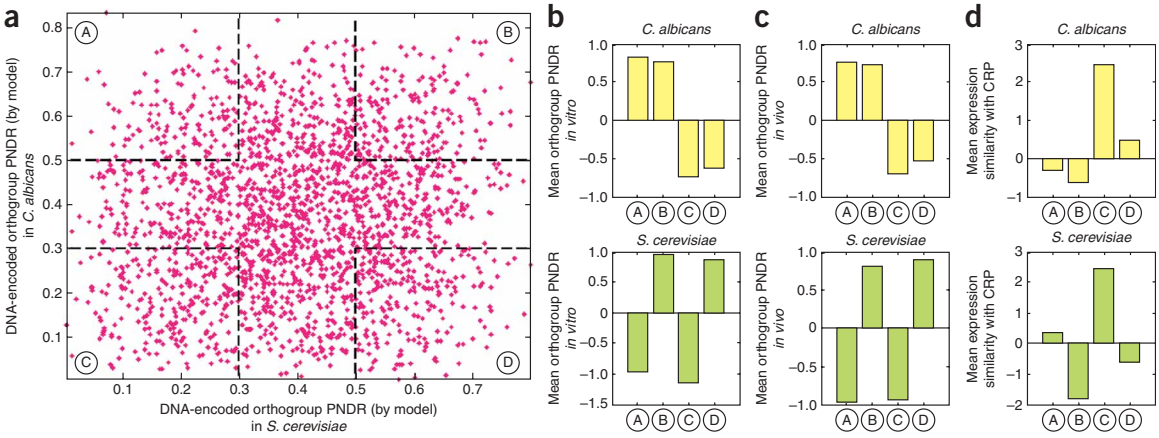


Figure 5 A global relationship between evolutionary changes in the DNA-encoded nucleosome organization and evolutionary changes in expression. (**a**) Shown is the PNDR score predicted by our model of nucleosome sequence preferences, for each ortholog from *C. albicans* (*y* axis) and *S. cerevisiae* (*x* axis). For orthologs that contain more than one gene in one of the species, the average PNDR score is shown. We used this plot to define four groups ('A', 'B', 'C' and 'D'), for all combinations of high (above 0.5) and low (below 0.3) PNDR scores in the two species. (**b**) For each of the four groups defined in **a**, shown are the PNDR scores computed from the *in vitro* nucleosome occupancy map in *C. albicans* (top) and *S. cerevisiae* (bottom). PNDR scores are shown as the difference between the actual PNDR score of the gene group and the average PNDR score of all genes in each respective species. For all groups, high and low PNDR scores predicted by the model in each species have correspondingly high and low PNDR scores in the *in vitro* maps, respectively. (**c**) Same as **b**, when computing the PNDR scores using the *in vivo* nucleosome occupancy maps in *C. albicans* (top) and *S. cerevisiae* (bottom). Here too, for all groups, the model predictions are upheld in the *in vivo* nucleosome maps in each species. (**d**) For each of the four groups from **a**, shown is the average normalized correlation between the expression of the genes from the group and the expression of the CRP genes in *C. albicans* (top) and *S. cerevisiae* (bottom).

relatively closed nucleosome organizations in all of the anaerobic yeast species (**Fig. 4**). Thus, our results demonstrate that a major phenotypic change across yeast species, namely the emergence of anaerobic yeast species, was accompanied by evolution of DNA-encoded nucleosome organization in a large number of aerobic respiration gene promoters. Notably, this phenotypic divergence coincides with a whole-genome duplication event in the evolutionary history of yeast, such that the six anaerobic yeast species descend from a post-genome duplication ancestor. Several studies have pointed out other unique evolutionary changes that coincide with this whole-genome duplication event[14,25].

Co-evolution of expression and nucleosome organization

Finally, we used the model to obtain a genome-wide view of changes in the DNA-encoded nucleosome organization across all of the genes that are conserved between *S. cerevisiae* and *C. albicans* (**Fig. 5a**). From this view, we extracted four extreme groups of genes on the basis of whether they had relatively open or closed organizations in *S. cerevisiae* and *C. albicans*, as determined by their PNDR scores. In all cases, high or low PNDR scores predicted by the model indeed have significantly high or low PNDR scores in both the *in vitro* and *in vivo* nucleosome occupancy maps (**Fig. 5b,c**). Examining the expression profiles of every group in each of the two species, we found a notable global trend, in that the two groups ('B' and 'C') whose DNA-encoded nucleosome organization is conserved between the two species also show conservation of their transcriptional programs, whereas the two groups ('A' and 'D') whose DNA-encoded organization diverged between the two species also show divergence in their transcriptional programs (**Fig. 5d**). For example, group A genes, which have a relatively closed (high PNDR scores) and open (low PNDR scores) chromatin organization in *C. albicans* and *S. cerevisiae*, respectively, show negative expression correlation to the CRP genes in *C. albicans* and positive expression correlation to the CRP genes in *S. cerevisiae*.

These results reinforce the trend that we observed for our three categories of gene sets, in that many individual genes whose DNA-encoded nucleosome organization has diverged between the two species also show divergence in their transcriptional programs. Moreover, in all cases, the direction of change in the encoded nucleosome organization is opposite to the direction of change in expression, such that changes that result in relatively more open and more closed nucleosome organizations are accompanied by higher and lower expression correlation to the CRP genes, respectively. Notably, our gene set–level analysis (**Fig. 1**) did not identify gene sets that show the expression divergence of group A genes, further highlighting the utility of this analysis at the level of individual genes. Although these are the global trends observed in each group, many individual genes within each group behave differently.

DISCUSSION

Our results suggest that yeast species exhibit a simple relationship between transcriptional programs and nucleosome organizations encoded in promoter sequences. Promoters of genes that are required for the typical mode of growth tend to encode relatively open nucleosome organizations, whereas promoters of genes that are not part of the typical growth pathways of the organism (for example, condition-specific and stress-response genes) tend to encode relatively closed nucleosome organizations. Notably, this relationship continues to hold even after the divergence of yeast into species that grow aerobically through pathways that involve cellular respiration and mitochondrial genes, and species that grow anaerobically through pathways that do not involve these genes[26]. We propose that this large-scale change in the expression of respiration genes is achieved, at least in part, through DNA sequence changes that alter the DNA-encoded nucleosome organization in their promoters. We provide strong support for this proposed mechanism by showing that these changes in nucleosome organization are also seen in a reconstitution of nucleosomes on purified DNA from *S. cerevisiae* and *C. albicans*. Our results thus show one case in which a system-level reprogramming of the yeast transcriptional network is associated with, and presumably achieved, in part, through evolution of intrinsic nucleosome organization encoded in the DNA sequence of promoters. This evolutionary mechanism for genetic change may also account for other types of phenotypic diversity observed across eukaryotic species.

METHODS

Parallel sequencing of *in vivo* nucleosome maps in *S. cerevisiae* and *C. albicans*. We extracted mononucleosomes from log-phase yeast cells using standard methods. *Saccharomyces cerevisiae* and *Candida albicans* mononucleosomes were prepared separately, and two independent replicates were taken from each species. The DNA was extracted, and protected fragments of length ~147 bp were cloned and sequenced on an Illumina GA II instrument.

Creation of *in vitro* nucleosome maps in *S. cerevisiae* and *C. albicans*. We purified *S. cerevisiae* genomic DNA from strain YLC8 [*MAT(a) ura3(Δ) leu2(Δ) his3(Δ) met15(Δ)*], and *C. albicans* DNA from strain SC5314, using standard methods. For both species, additional steps were taken to remove contaminating RNA. After recovery by ethanol precipitation, DNA was resuspended in TE buffer (10 mM Tris pH 8.0, 1 mM EDTA), and the DNA concentration was determined by agarose gel electrophoresis using ethidium stain, followed by comparison to mass standards using quantitative fluorometry. The sample was then subjected to RNase A digestion at 50 °C overnight, using 100 μg of RNase A for every 10 μg of DNA, followed by ethanol precipitation of the DNA. After resuspension of the pellet in TE, the genomic DNA was sheared twice each through a 25-gauge needle and then a 27.5-gauge needle. The entire mixture was then electrophoresed on a 20 × 20 cm, 1% agarose, 1 × TAE gel at 100 V for 6–8 h. We cut out the genomic DNA band, and then re-electrophoresed the agarose slab containing the DNA inside a dialysis bag, with occasional UV-light monitoring, to elute the DNA.

Histone octamer (HO) was purified from chicken erythrocytes using salt extraction and hydroxyapatite column chromatography, as previously described[27]. We reconstituted genomic DNA into nucleosomes under selective pressure for nucleosome-favoring sequences by salt gradient dialysis[23]. For *S. cerevisiae*, the reconstitution reaction used 40 μg HO + 100 μg DNA in a 200 μl volume. The resulting nucleosomes were biochemically isolated by micrococcal nuclease (MNase) digestion, in two independent experiments, using 6×10^{-3} or (separately) 6×10^{-4} units MNase (Sigma) per 10 μg competitively reconstituted DNA, in 10 mM Tris pH 8.0, 1 mM $CaCl_2$, for 5 min at 37 °C. For *C. albicans*, reconstitution reactions used 37 μg HO + 93 μg DNA in a 200 μl volume. Nucleosomes from two independent reconstitutions were digested (separately) with 6×10^{-3} units MNase per 10 μg competitively reconstituted DNA, for 5 min at 37 °C. After digestion, DNA was extracted, and protected fragments of length ~147 bp were isolated by PAGE, extracted from the gel. Samples were independently subjected to Illumina sequencing. A detailed comparison between the *in vitro* and *in vivo* nucleosome maps of *S. cerevisiae* has been presented previously[24].

Mapping and postprocessing of parallel sequencing reads. To map the reads resulting from the above sequencing experiments in each species we used NCBI BLAST[28] requiring 32 matches and allowing at most 1 gap. To estimate the mean DNA fragment length in each experiment, we superimposed the nucleosome reads of one strand and examined the distribution of nucleosome reads of the opposite strand. As expected, this distribution shows a strong peak at ~140–170 bp for all experiments, with slight variations between experiments. We used the maximum of the peak as an estimation of the mean DNA fragment length and extended all nucleosome reads to this length. We defined

repetitive regions as regions that were matched by a read that mapped to more than one place in the genome. We excluded repetitive regions and their 150-bp vicinity from our analyses. To obtain genomic nucleosome occupancy tracks, we summed for each position all reads covering it. We excluded base pairs covered by more than ten times the median genomic base-pair coverage (typically less than 1% of all base pairs). Finally, we normalized each track by the median base-pair coverage.

Datasets. The genome sequence and gene and chromosome annotations of the yeast species examined in this study were obtained from a recent compilation[14]. The member genes of Gene Ontology gene sets from *S. cerevisiae* were downloaded from the Gene Ontology repository[17]. The member genes of the same gene sets in other yeast species were defined as the orthologs of the original gene set from *S. cerevisiae*, using a recent orthology map across 17 yeast species[14]. For all pairwise comparisons between *S. cerevisiae* and *C. albicans*, we restricted our analysis to only genes that have orthologs in the other species, resulting in 2,835 genes in *S. cerevisiae* and 2,823 genes in *C. albicans*, representing 2,225 orthogroups. Note that the number of genes in each species differs, because the orthology map includes many-to-many relationships (that is, some orthogroups include more than one gene from one or both species). For the analysis across multiple yeast species (**Fig. 4**) the genes in each species were restricted to only those that have orthologs in all of the other species. Expression compendia of ~1,000 and ~200 gene expression measurements in *S. cerevisiae* and *C. albicans*, respectively, were downloaded from a previous compilation[1].

Computing expression correlations between gene sets. As our input gene sets, we used all of the gene sets from Gene Ontology[17] that have at least ten orthologous genes in each species. We then computed a transcription-program similarity measure between each pair of gene sets, separately in each species, as the average (over nonidentical gene pairs) normalized Pearson correlation between expression profiles. The normalized Pearson correlation is the Pearson correlation after subtraction of the mean and division by the s.d. of the Pearson correlation between every pair of genes in that species. This standardization corrects for potential biases in the Pearson correlation that may arise owing to size differences between the expression compendia of each species. We used these normalized correlations to further restrict the input gene sets to use only those for which the average normalized correlation between all pairs of genes from the gene set was above 0.5.

Model predictions. The model used for predicting nucleosome organization based on the genomic sequence has a single parameter that represents the apparent nucleosome concentration. We set this nucleosome concentration parameter to 1 in *S. cerevisiae*, and for each other species we set it such that the average genome-wide predicted nucleosome occupancy in that species is equal to that predicted for *S. cerevisiae* using a concentration of 1. To reduce potential biases that may arise from these concentration parameters, we restricted our analyses to examining the relative, rather than the absolute, relations between the nucleosome organization of different gene sets.

URLs. Data, results and model predictions, http://genie.weizmann.ac.il/pubs/nucleosomes09.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

We acknowledge with gratitude the gift of strains, protocols and advice from J. Berman (University of Minnesota), and thank H. Kelkar (University of North Carolina) for help with the Illumina sequencing data and the members of our respective laboratories for discussions and comments on the manuscript. This work was supported by grants from the US National Institutes of Health to J.D.L., from the NIH to J.W., and from the European Research Council (ERC) and NIH to E.S. N.K. is a Clore scholar. E.S. is the incumbent of the Soretta and Henry Shapiro career development chair.

AUTHOR CONTRIBUTIONS

Y.F., I.K.M., J.D.L., J.W. and E.S. conceived and designed the experiments. Y.F.-M. and I.K.M. performed the experiments. P.M. performed the sequencing. Y.F., N.K., Y.L., J.D.L., J.W. and E.S. analyzed the data. Y.F., J.D.L., J.W. and E.S. wrote the paper.

Published online at http://www.nature.com/naturegenetics/
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/

1. Ihmels, J. *et al.* Rewiring of the yeast transcriptional network through the evolution of motif usage. *Science* **309**, 938–940 (2005).
2. Borneman, A.R. *et al.* Divergence of transcription factor binding sites across related yeast species. *Science* **317**, 815–819 (2007).
3. Lee, W. *et al.* A high-resolution atlas of nucleosome occupancy in yeast. *Nat. Genet.* **39**, 1235–1244 (2007).
4. Yuan, G.C. *et al.* Genome-scale identification of nucleosome positions in *S. cerevisiae*. *Science* **309**, 626–630 (2005).
5. Albert, I. *et al.* Translational and rotational settings of H2A.Z nucleosomes across the *Saccharomyces cerevisiae* genome. *Nature* **446**, 572–576 (2007).
6. Shivaswamy, S. *et al.* Dynamic remodeling of individual nucleosomes across a eukaryotic genome in response to transcriptional perturbation. *PLoS Biol.* **6**, e65 (2008).
7. Whitehouse, I., Rando, O.J., Delrow, J. & Tsukiyama, T. Chromatin remodelling at promoters suppresses antisense transcription. *Nature* **450**, 1031–1035 (2007).
8. Ozsolak, F., Song, J.S., Liu, X.S. & Fisher, D.E. High-throughput mapping of the chromatin structure of human promoters. *Nat. Biotechnol.* **25**, 244–248 (2007).
9. Mavrich, T.N. *et al.* A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome. *Genome Res* **18**, 1073–1083 (2008).
10. Segal, E. *et al.* A genomic code for nucleosome positioning. *Nature* **442**, 772–778 (2006).
11. Ioshikhes, I.P., Albert, I., Zanton, S.J. & Pugh, B.F. Nucleosome positions predicted through comparative genomics. *Nat. Genet.* **38**, 1210–1215 (2006).
12. Peckham, H.E. *et al.* Nucleosome positioning signals in genomic DNA. *Genome Res.* **17**, 1170–1177 (2007).
13. Yuan, G.C. & Liu, J.S. Genomic sequence is highly predictive of local nucleosome depletion. *PLoS Comput. Biol.* **4**, e13 (2008).
14. Wapinski, I., Pfeffer, A., Friedman, N. & Regev, A. Natural history and evolutionary principles of gene duplication in fungi. *Nature* **449**, 54–61 (2007).
15. Stuart, J.M., Segal, E., Koller, D. & Kim, S.K. A gene-coexpression network for global discovery of conserved genetic modules. *Science* **302**, 249–255 (2003).
16. Bergmann, S., Ihmels, J. & Barkai, N. Similarities and differences in genome-wide expression data of six organisms. *PLoS Biol.* **2**, E9 (2004).
17. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29 (2000).
18. Segal, E., Friedman, N., Koller, D. & Regev, A. A module map showing conditional activity of expression modules in cancer. *Nat. Genet.* **36**, 1090–1098 (2004).
19. Gasch, A.P. *et al.* Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell* **11**, 4241–4257 (2000).
20. Mager, W.H. & Planta, R.J. Coordinate expression of ribosomal protein genes in yeast as a function of cellular growth rate. *Mol. Cell. Biochem.* **104**, 181–187 (1991).
21. Field, Y. *et al.* Distinct modes of regulation by chromatin encoded through nucleosome positioning signals. *PLoS Comput. Biol.* **4**, e1000216 (2008).
22. Hanley, J.A. & McNeil, B.J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **143**, 29–36 (1982).
23. Thastrom, A., Bingham, L.M. & Widom, J. Nucleosomal locations of dominant DNA sequence motifs for histone-DNA interactions and nucleosome positioning. *J. Mol. Biol.* **338**, 695–709 (2004).
24. Kaplan, N. *et al.* The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature* advance online publication, doi:10.1038/nature07667 (17 December 2008).
25. Man, O. & Pilpel, Y. Differential translation efficiency of orthologous genes is involved in phenotypic divergence of yeast species. *Nat. Genet.* **39**, 415–421 (2007).
26. Piskur, J. & Langkjaer, R.B. Yeast genome sequencing: the power of comparative genomics. *Mol. Microbiol.* **53**, 381–389 (2004).
27. Feng, H.P., Scherl, D.S. & Widom, J. Lifetime of the histone octamer studied by continuous-flow quasielastic light scattering: test of a model for nucleosome transcription. *Biochemistry* **32**, 7824–7831 (1993).
28. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).