A genomic model of condition-specific nucleosome behaviour explains transcriptional activity in yeast

Supplementary Material

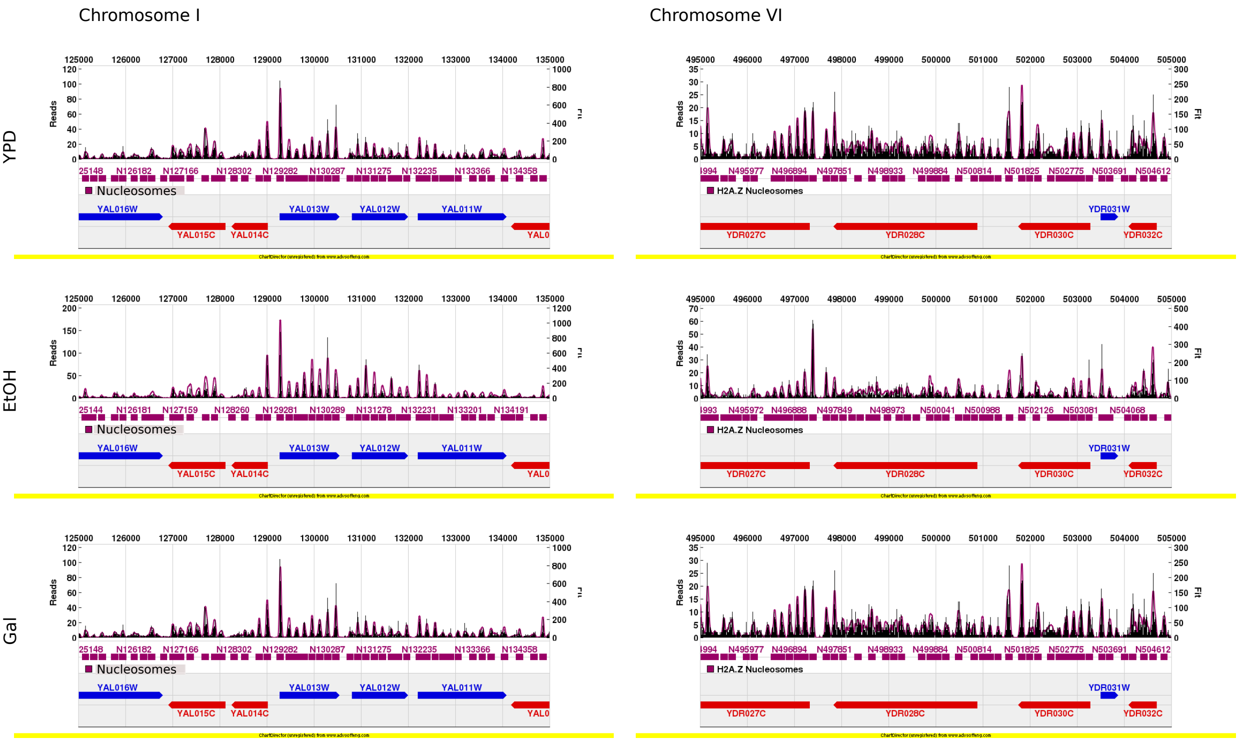
Judith B Zaugg[1] and Nicholas M Luscombe[1,2]

[1] EMBL-European Bioinformatics Institute, Wellcome Trust Genome Campus, Cambridge CB10 1SD, UK
[2] Genome Biology Unit, EMBL Heidelberg, Meyerhofstrasse 1, Heidelberg D-69117, Germany

[§]Corresponding authors:
Judith B Zaugg: zaugg@ebi.ac.uk; +44 (0)1223 492572
Nicholas M Luscombe: luscombe@ebi.ac.uk; +44 (0)1223 492572

Contents

Figure S1: Nucleosome prediction output by GeneTrack

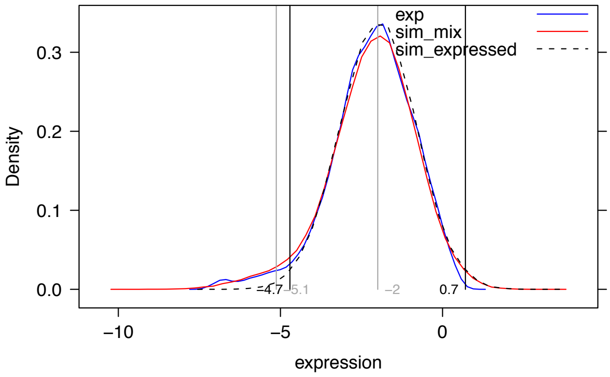


To identify individual nucleosome-binding positions, we applied the GeneTrack algorithm (Albert et al. 2008) to the MNase-Seq data. Figure S1 shows the raw MNase-Seq data, Gaussian-smoothing and peak-calling outputted by GeneTrack for two different 10kb stretches of the yeast genome in the YPD, EtOH and Gal conditions. The reads (black) are smoothed by a Gaussian filter (pink line) with fitting window of 75bp and standard deviation of 15bp. Individual nucleosome positions (horizontal pink bars) were identified using an exclusion zone of 147bp.

GeneTrack applies Gaussian-smoothing to each genomic position: ie, each base position is represented by a normal distribution with peak height equal to the number of reads and standard deviation equal to the fitting tolerance supplied by the user (Albert et al. 2008). Gaussian-distribution values are summed at each base position and joined together to produce a continuous line. Individual nucleosomes are identified by searching for non-overlapping peaks according to a user-supplied exclusion zone. These peaks correspond to the most likely positions of nucleosomes and peak heights correspond to the estimated number of MNase-Seq reads present at the particular position

Figure S2: Definition of expression states

The ON and OFF gene states were calculated using gene expression data from each cellular condition. Normalised expression values from replicate experiments were averaged: the density plot below displays the distribution of



expression values of genes in YPD (blue line). A normal mixture model consisting of two normal distributions was fitted to the data (red line); the larger distribution corresponds to expressed gene values (black dotted line).

We distinguish between ON and OFF genes using a threshold of 1% FDR in the distribution of expressed genes (vertical grey line).

To ensure the robustness of results, we repeated the analyses presented in the paper using an additional threshold of 5% FDR. The differences in nucleosome-binding properties are similar to those reported using a 1% FDR cut-off: +1 nucleosomes bind more strongly in ON genes than OFF, whereas the -1 nucleosomes are stronger in the OFF state for promoters in the Closed configuration. The p-values are given for Wilcoxon's rank sum test and are even more significant than for the 1% FDR; this is most likely due to the larger sample size of OFF genes.

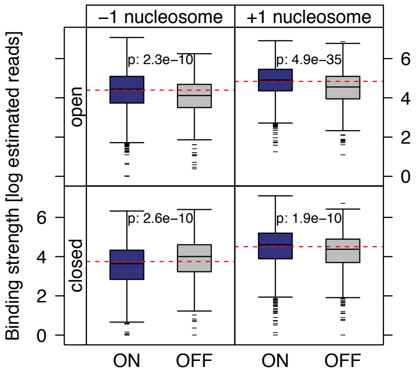


Figure S3: Definition of expression states using RNA-Seq data
To confirm the robustness of results further, we repeated the analyses using expression data obtained by RNA-Seq (Nagalakshmi et al. 2008). Note that these data are available only for YPD. Nagalakshmi and colleagues used the median read count for the last 30bp of each transcript as a measure of expression level (Nagalakshmi et al. 2008). The density plot of log2(read count) values below shows a similar distribution to that obtained from microarray data: there is a bimodal distribution comprising populations of expressed genes and unexpressed ones, which we used as the natural threshold between ON and OFF genes (red line). Since the expression value was determined at the 3'end of genes, we excluded those transcripts that differed by more than 200bp in the position of their 3'end between the two studies in the analysis below. The differences in nucleosome binding properties are comparable to what was reported for microarray data.

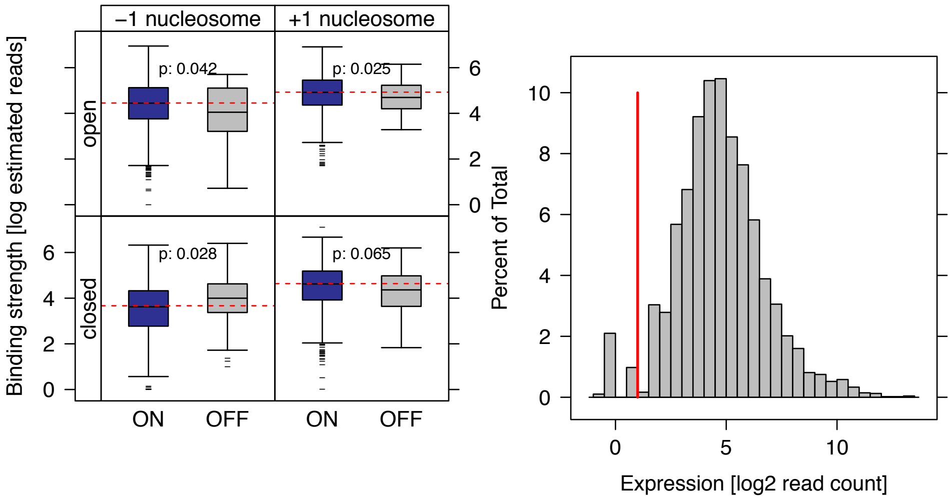


Figure S4: Testing the model using alternative nucleosome datasets
A comparison of the Kaplan (Kaplan et al. 2009) and Mavrich (Mavrich et al. 2008) datasets demonstrates that many promoters share very similar NFR sizes (red line). There is a minor proportion of promoters that have much larger NFRs in the Mavrich dataset (blue lines): however because the correlation between the two datasets is still good for these promoters, we speculate that data for -1 nucleosome has been lost in the Mavrich dataset.

This is consistent with findings from a study by Weiner and colleagues, who showed that the -1 nucleosome is especially unstable, and that extended exposure to MNase digest leads to their loss (Weiner et al. 2010). The fact that the Mavrich dataset contains ~10 times fewer reads than the Kaplan dataset (1.2M v 12.5M) also suggests that nucleosomes have been lost in the former dataset. Nonetheless, given the good correlation in NFR sizes and nucleosome-binding positions between the Kaplan and Mavrich datasets, we suggest that our observations are robust.

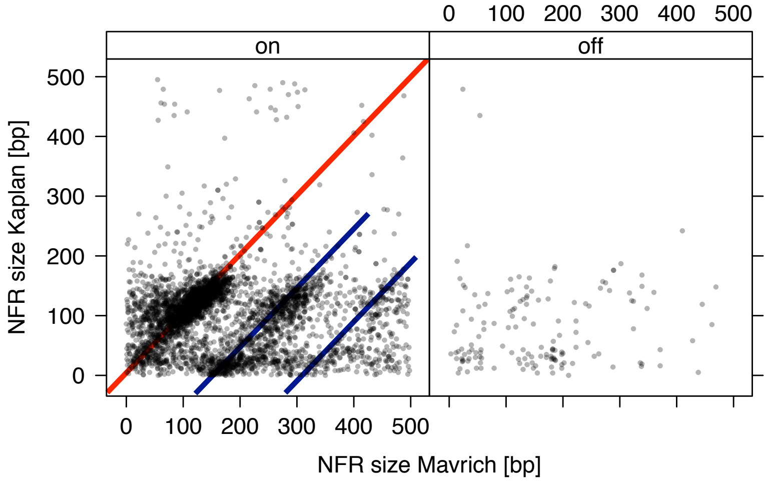
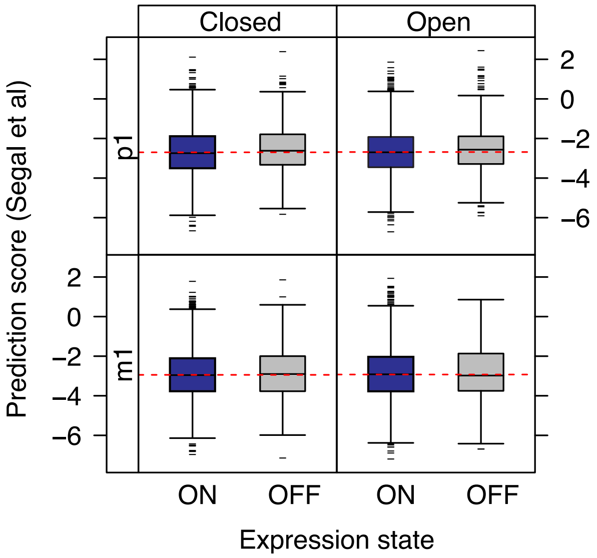


Figure S5A: Properties of predicted nucleosomes based on DNA sequence
To test whether nucleosome positions are determined by the underlying genomic sequence, we checked whether the four-state model holds true for predicted nucleosomes. We calculated scores of nucleosome-binding based on two models: the nucleosome-DNA interaction model by Kaplan and Segal (Segal et al. 2006; Kaplan et al. 2009) and the DNA sequence based hidden Markov Model by Xi (Xi et al. 2010). Both algorithms calculate probability scores for the presence of a nucleosome at a given genomic location, given the underlying nucleotide sequence. We applied both algorithms to the yeast genome, and then we separated the promoters into the same four classes as in Figure 1C (based on YPD).

We compared the distributions of nucleosome probability scores at the +1 and -1 positions across the four promoter states. For the Segal prediction method, we found no significant difference between ON and OFF as well as between Open and Closed promoters (left; Wilcoxon Rank Sum test). For the Xi predictions, we found no differences among the +1 nucleosomes, whereas the -1 nucleosome positions had higher probability scores among OFF genes compared with OFF (right). That the Xi prediction method performs better for OFF genes appears to support the notion that nucleosomes in ON genes are positioned by trans-acting factors. However, we note that neither of the algorithms reveal a similar pattern as the experimentally measured nucleosome positions (Figure 1C).



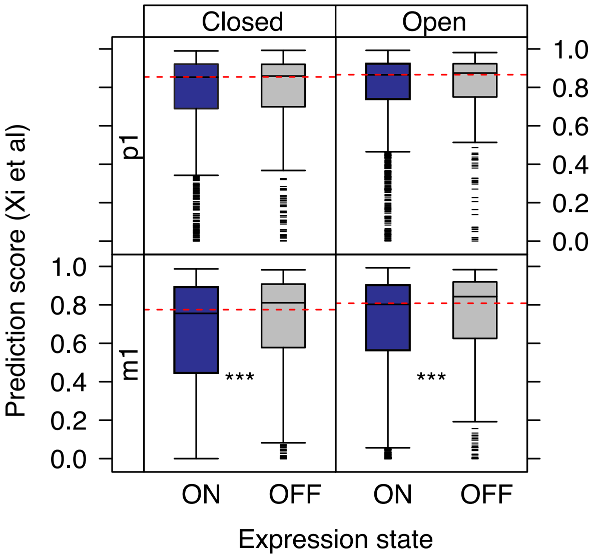


Figure S5B: Examples of predicted nucleosomes

To examine the above results in greater detail, we compared the positions of experimentally measured nucleosomes with the two prediction methods at individual promoters. Below, we show one representative example for each promoter state. We find that the predictions from both Kaplan (red) and Xi (blue) algorithms differ substantially from the nucleosome positions obtained from the data (grey boxes).

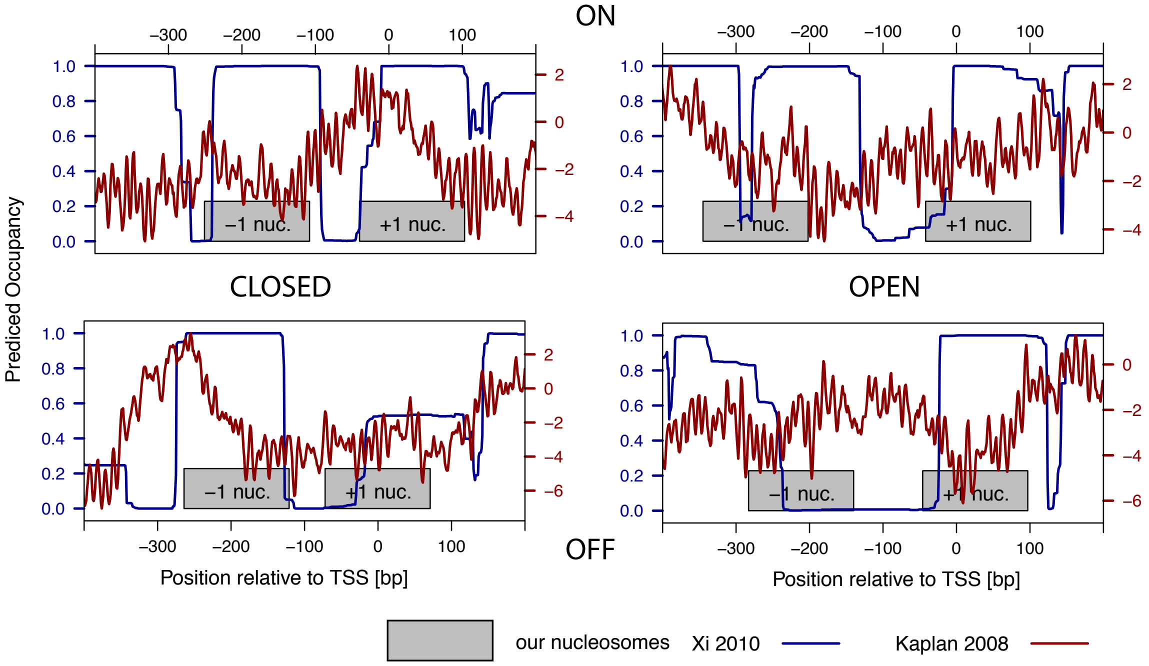


Figure S6: Effect of sample sizes on nucleosome positioning

Figure 1 in the main manuscript shows the nucleosome-binding distributions for all promoters in the respective categories. Since there are many more ON genes than OFF, we tested whether the distributions are affected by sample size by randomly sampling the same numbers of ON and OFF promoters in the Open and Closed states. The plots below show the distributions of the +1 and -1 nucleosomes in the Open/Closed and ON/OFF states relative to the TSS. The +1 nucleosome is well-defined in all ON promoters, and the -1 nucleosome is well-positioned in the Open state compared with Closed. Therefore, the positioning of +1 and -1 nucleosomes is not an artefact of samples sizes.

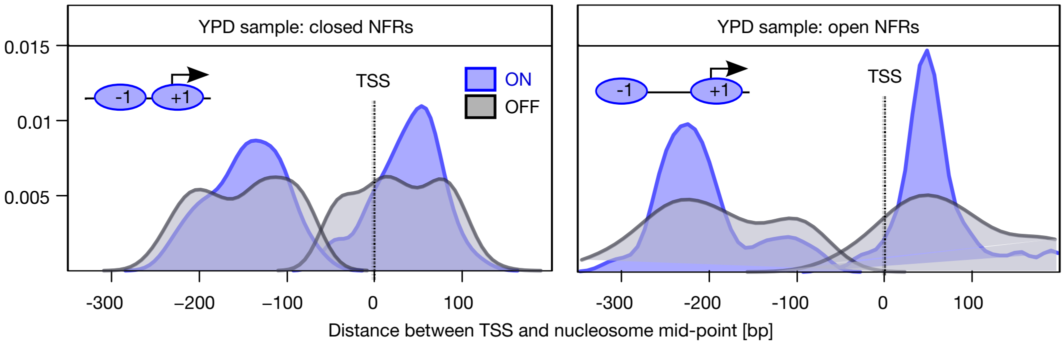


Figure S7: Expression states correlates with nucleosome occupancy

Previous studies showed that average nucleosome-occupancies along the gene body are lower for expressed genes compared with unexpressed ones. In Figure S7, we double-checked that this observation is true for the Kaplan dataset. Standardised nucleosome occupancies across entire genes are shown for those that are expressed (ON, blue) and not expressed (OFF, grey). ON genes have significantly lower occupancy than OFF gnes in all growth conditions (YPD, EtOH, and Gal from left to right). Wilcoxon rank sum test, p-value < 0.001 (***).

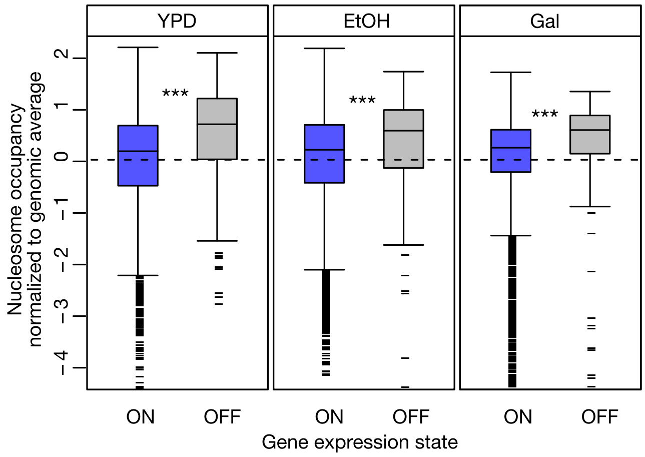
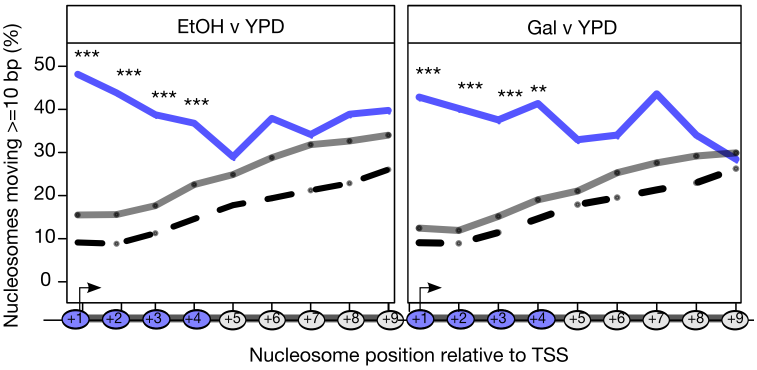


Figure S8: Nucleosome movement starts at 5‘-end of genes and propagates to the 3’-end.

For each nucleosome in the reference YPD1 dataset, we identified the equivalent nucleosome in the EtOH, Gal and YPD2 datasets as the ones that occupy overlapping genomic positions; we excluded non-overlapping nucleosomes since they cannot be distinguished from eviction events or missing data. If more than one nucleosome in the condition under comparison overlapped with a reference nucleosome in YPD1, we retained the one displaying greater overlap (numbers shown in SOM Table 2). We used the distance between the centres of equivalent nucleosomes as a measure of movement between cellular conditions. The distances measured in the YPD1 v YPD2 comparison were used as a control.

The method gives a very conservative estimate of nucleosome movement, as it assumes that nucleosomes do not shift more than the length of DNA wrapped around them (ie, 147bp). Since 98% of reference nucleosomes in YPD1 have a counterpart in each condition, the method captures most of the movement in the data (Table S3).



The plot shows the proportion of nucleosomes at each position along the gene (numbered +1 to +9) that shifts at least 10bp between two conditions for YPD1 v EtOH (left panel) and YPD1 v Gal (right panel). Nucleosomes in gene that don't change expression state (grey solid lines) behave similarly to the YPD1 v YPD2 control (black dashed lines): there is least movement at the 5'-end and a slight increase towards the 3'-end of the gene. Nucleosomes in genes that change expression state between the two growth conditions display much greater movement (blue solid lines): movement is greatest at the 5'-end and steadily decreases towards the 3'-end. Significantly larger proportions of nucleosomes in switching genes move in the first four positions compared with non-switching genes. Fisher test, *p*-value < 0.01(**) and < 0.0001(***).

Table S1

Numbers of predicted nucleosome positions in each growth condition. Note YPD1 is used throughout the analysis (See methods).

	# nucleosomes	# +1	# -1
YPD1	53,657	5,120	3,851
YPD2	53,290	5,118	3,942
EtOH	53,637	5,122	3,985
Gal	53,271	5,119	3,928

Table S2

Enrichment of poly(dA:dT) of different sizes in open NFR promoters.

length of track	adjusted p-value	ratio	significant
4	1.9E-01	1.56	
5	1.4E-03	1.43	**
6	3.7E-03	1.27	**
7	9.5E-04	1.31	***
8	1.0E-02	1.27	**
9	4.8E-03	1.36	**
10	2.8E-02	1.32	*
11	1.4E-01	1.25	

Promoters were scanned for motifs of poly(dA:dT) of size 4-11. Association with Open and Closed promoters was tested using Fisher's exact test. The enrichment for poly(dA:dT) in Open promoters is significant but very small with Odd ratios between 1.2 and 1.5

Table S3

Numbers of nucleosome pairs for YPD1 v EtOH, YPD1 v Gal and YPD1 v YPD2.

	Total		**Intergenic**		**Genic**	
	#pairs	**% YPD1 nucleosomes**	**# pairs**	**% YPD1 nucleosomes**	**# pairs**	**% YPD1 nucleosomes**
YPD1 v EtOH	52,653	98.1	2,958	100	49,695	98.0
YPD1 v Gal	52,481	97.8	2,914	100	49,567	97.7
YPD1 v YPD2	52,504	97.9	2,972	100	49,532	97.7

Table S4

GO enrichment analysis - performed using GO-profiler (Reimand et al. 2007) - of genes whose NFR configurations (Open/Closed) change between growth in EtOH and YPD.

Open in EtOH, closed in YPD

p-value	GO cat	process	description
6.25E-05	GO:0010035	BP	response to inorganic substance
3.48E-05	GO:0000302	BP	response to reactive oxygen species
2.01E-05	GO:0042743	BP	hydrogen peroxide metabolic process
2.01E-05	GO:0070301	BP	cellular response to hydrogen peroxide
2.01E-05	GO:0042744	BP	hydrogen peroxide catabolic process
7.90E-05	GO:0004096	MF	catalase activity
2.83E-05	**KEGG:04146**	**ke**	**Peroxisome**

Open in YPD, closed in EtOH

p-value	GO cat	process	description
2.49E-05	GO:0000056	BP	ribosomal small subunit export from nucleus
1.81E-05	GO:0006084	BP	acetyl-CoA metabolic process
6.34E-05	GO:0046356	BP	acetyl-CoA catabolic process
6.34E-05	GO:0006099	BP	tricarboxylic acid cycle
3.22E-07	GO:0006793	BP	phosphorus metabolic process
3.22E-07	GO:0006796	BP	phosphate metabolic process
3.74E-07	GO:0016310	BP	phosphorylation
6.34E-05	GO:0055114	BP	oxidation reduction
3.57E-08	GO:0006091	BP	generation of precursor metabolites and energy
9.40E-06	GO:0015980	BP	energy derivation by oxidation of organic compounds
7.20E-06	GO:0045333	BP	cellular respiration
6.34E-05	GO:0022900	BP	electron transport chain
6.34E-05	GO:0022904	BP	respiratory electron transport chain
1.11E-07	GO:0006119	BP	oxidative phosphorylation
6.34E-05	GO:0042773	BP	ATP synthesis coupled electron transport
4.86E-05	GO:0045269	CC	proton-transporting ATP synthase, central stalk
4.86E-05	GO:0005756	CC	mitochondrial proton-transporting ATP synthase, central stalk
4.17E-06	GO:0045259	CC	proton-transporting ATP synthase complex
4.12E-05	GO:0005753	CC	mitochondrial proton-transporting ATP synthase complex
4.12E-05	GO:0022890	MF	inorganic cation transmembrane transporter activity
1.08E-03	**KEGG:00020**	**ke**	**Citrate cycle (TCA cycle)**
1.28E-03	**KEGG:00190**	**ke**	**Oxidative phosphorylation**
1.57E-04	REAC:504351	re	Citric acid cycle (TCA cycle)

Data S1

The dataset describes the properties (expression, promoter nucleosomes, categories) of the ORFs and contains the following information:

1: ORF; 2: gene name; 3: chromosome; 4: strand; 5: start position; 6: end position; 7: expression level YPD; 8: expression level EtOH; 9: expression level Gal; 10: expression state YPD; 11: expresion state EtOH; 12: expression state Gal; 13: switching class EtOH v YPD; 14: switching class Gal v YPD; 15: NFR configuration YPD; 16: NFR configuration EtOH; 17: NFR configuration Gal; 18: +1 nucleosome id YPD1; 19: +1 nucleosome id YPD2; 20: +1 nucleosome id EtOH; 21: +1 nucleosome id Gal; 22: -1 nucleosome id YPD1; 23: -1 nucleosome id YPD2; 24: -1 nucleosome id EtOH; 25: -1 nucleosome id Gal

Data S2-S5

Datasets S2 to S5 describe the nucleosome data sets for YPD1, YPD2, EtOH and Gal respectively. Each file contains the following information:

1: nucleosome id; 2: chromosome; 3: start position; 4: binding strength; 5: binding focus

Data S6:

The dataset describes the relations between nucleosomes in different data sets. The file contains the following information:

1: id; 2: chromosome; 3: start position YPD1 nucleosome; 4: nucleosome id YPD1; 5: nucleosome id EtOH; 6: nucleosome id Gal; 7: nucleosome id in YPD2; 8: distance EtOH-YPD1; 9: distance Gal -YPD1; 10: distance YPD2-YPD1

Bibliography

Albert I, Wachi S, Jiang C, Pugh BF. 2008. GeneTrack--a genomic data processing and visualization framework. *Bioinformatics* **24**(10): 1305-1306.

Kaplan N, Moore IK, Fondufe-Mittendorf Y, Gossett AJ, Tillo D, Field Y, LeProust EM, Hughes TR, Lieb JD, Widom J et al. 2009. The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature* **458**(7236): 362-366.

Mavrich TN, Ioshikhes IP, Venters BJ, Jiang C, Tomsho LP, Qi J, Schuster SC, Albert I, Pugh BF. 2008. A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome. *Genome Res* **18**(7): 1073-1083.

Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M. 2008. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320**(5881): 1344-1349.

Reimand J, Kull M, Peterson H, Hansen J, Vilo J. 2007. g:Profiler--a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic Acids Res* **35**(Web Server issue): W193--200.

Segal E, Fondufe-Mittendorf Y, Chen L, Thåström A, Field Y, Moore IK, Wang J-PZ, Widom J. 2006. A genomic code for nucleosome positioning. *Nature* **442**(7104): 772-778.

Weiner A, Hughes A, Yassour M, Rando OJ, Friedman N. 2010. High-resolution nucleosome mapping reveals transcription-dependent promoter packaging. *Genome research* **20**(1): 90-100.

Xi L, Fondufe-Mittendorf Y, Xia L, Flatow J, Widom J, Wang J-P. 2010. Predicting nucleosome positioning using a duration Hidden Markov Model. *BMC bioinformatics* **11**: 346.