



Supporting Online Material for

A Systems Approach to Mapping DNA Damage Response Pathways

Christopher T. Workman, H. Craig Mak, Scott McCuine, Jean-Bosco Tagne,
Maya Agarwal, Owen Ozier, Thomas J. Begley, Leona Samson, Trey Ideker*

*To whom correspondence should be addressed. E-mail: trey@bioeng.ucsd.edu

Published 19 May 2006, *Science* **312**, 1054 (2006)
DOI: 10.1126/science.1122088

The main PDF file includes the following:

Materials and Methods
Figs. S1 to S8
References

Other Supporting Online Material for this manuscript includes the following:
(available at www.sciencemag.org/cgi/content/full/312/5776/1122088/DC1)

Tables S1 to S9 as zipped archive 1122088data_files.zip

Supporting Online Material for:

A Systems Approach to Mapping DNA Damage Response Pathways

Christopher T. Workman^{1,*}, H. Craig Mak^{1,*}, Scott McCuine¹, Jean-Bosco Tagne², Maya Agarwal¹, Owen Ozier², Thomas J. Begley³, Leona D. Samson⁴, and Trey Ideker^{1†}

¹ University of California San Diego, La Jolla, CA 92093

² Whitehead Institute for Biomedical Research, Cambridge, MA 02139

³ University of Albany-State University of New York, Rensselaer, NY 12144

⁴ Massachusetts Institute of Technology, Cambridge, MA 02139

* These authors contributed equally to the manuscript.

† To whom correspondence should be addressed. E-mail: trey@bioeng.ucsd.edu

Methods

Strains and Media. Strains used in phenotyping and expression profiling were derived from the haploid BY4741. The parent strain was obtained from ATCC (Manassas, Virginia, USA), and all nonessential deletion strains constructed by the *Saccharomyces* Gene Deletion Project were obtained from Research Genetics (Huntsville, AL). Epitope-tagged strains (c-myc) used in construction of the physical regulatory network were derived from W303 and obtained from the laboratory of Dr. Richard A. Young at the Whitehead Institute (Cambridge, MA). Cells were cultured in standard yeast rich media (YPD) at 30°C except when noted.

Phenotypic Sensitivity Assays. Parent and all viable deletion strains harboring single knockouts of transcription factors (TFs) profiled in Lee *et al.* (1) were arrayed into 96-well plates containing YPD broth and grown to saturation. Settled cells in each well were re-suspended to ensure homogeneity, and 5µl from each well was spotted simultaneously onto YPD agar plates with the use of a Hydra liquid handling apparatus (Robbins Scientific). The process was repeated on additional YPD agar plates containing a range (0.01% to 0.1%) of methyl methanesulfonate concentrations (Sigma Chemical Company) to determine the optimum concentration to adequately detect sensitive knockout strains. MMS was added to cooled agar and used fresh within one day to ensure the stability of the agent. Spotted plates were grown for 60 h at 30°C and imaged using a Gel Doc 1000 (BioRad) running Quantity One software,

and all screens were performed in triplicate using fresh cultures. Images for YPD and YPD+0.025% MMS are shown in fig. S1. Eight TF knockout strains were found to be sensitive using this approach. This list was combined with the list of MMS-sensitive knockouts determined by Begley et al. (2) (score \geq 5) to obtain the 14 MMS-sensitive strains reported in the main text.

Whole Genome Expression Analysis. Gene expression experiments were processed in parallel using two distinct biological samples from each strain (colonies of similar size picked from fresh YPD + G418 agar plates) grown to saturation in YPD overnight at 30°C. The overnight culture was diluted 1:100 in a flask of 100ml fresh YPD and grown in a shaking incubator at 30°C (180 rpm) until the culture reached an OD₆₀₀ of 0.8 – 1.0. Every effort was made to closely match the final optical density of each of the samples. Once a culture reached the appropriate cell density, the sample was split in half, MMS added to a final concentration of 0.03% in the non-reference sample, and both cultures grown for an additional hour. Cells were harvested by centrifugation at 3000 rpm for five minutes at room temperature in a Legend RT centrifuge (Kendro Laboratory Products, Asheville, NC, USA). Cell pellets were immediately snap-frozen in liquid nitrogen to suspend gene expression (including any temperature stress response genes) and stored at –20°C prior to RNA extraction. Total RNA from each sample was isolated by hot acid phenol extraction and mRNA-purified via Poly(A)Pure kits (Ambion). Labeling of cDNA was performed in a dye-reversal scheme by direct incorporation using a CyScribe First-Strand cDNA Labeling Kit (Amersham Biosciences). Corresponding Cy-3 and Cy-5 labeled samples were co-hybridized to microarrays containing the Yeast Genome Oligo Set Version 1.1 (Qiagen).

Genome Wide Transcription Factor Binding Analysis. Samples were processed in parallel using three distinct biological replicates from each epitope-tagged strain and grown to saturation in YPD overnight at 30°C. The overnight culture was diluted 1:100 in a flask of 50 ml fresh YPD and grown in a shaking incubator at 30°C (180 rpm) until the culture reached an OD₆₀₀ of 0.8 – 1.0. MMS was added to a final concentration of 0.03%, and the culture was grown for an additional hour. Protein-DNA binding locations were assayed as previously described by Lee *et al.* (1) with corresponding IP-enriched and

unenriched samples co-hybridized to a single cDNA microarray containing all yeast intergenic sequences derived from PCR amplification.

Array Preparation and Hybridization. Microarrays were spotted on UltraGAPS II slides (Corning) using an OmniGrid 100 microarrayer (GeneMachines). After spotting, arrays were baked at 80°C for 2 hours, cross-linked at 300 mJ in a UV Stratalinker 2400 (Stratagene), and stored under vacuum. Hybridizations were conducted at 42°C for 15 hours using the Lucidea SlidePro automated hybridization machine (Amersham Biosciences), and arrays scanned using GenePix 4000A (Axon Instruments) or ScanArray Express scanners (Perkin-Elmer) at a 10.0 μm resolution.

Data Post-processing and Significance Assessment. Scanned images were processed using GenePix Pro 3.0 (Axon Instruments) or QuantArray (PerkinElmer) software to obtain raw Cy-3 and Cy-5 foreground (f_3, f_5) and background (b_3, b_5) intensity measurements for each spot on the array. Cy-3 and Cy-5 background intensities were smoothed using a 7x7 median spatial filter to obtain (b_3', b_5') which were then subtracted from foreground measurements to obtain the background-adjusted intensities: $f_3 - b_3'$ and $f_5 - b_5'$. Next, log ratios $L = \log_{10}[(f_5 - b_5') / (f_3 - b_3')]$ were corrected for cyanine-dye dependent bias by scaling the background-adjusted intensities using Qspline normalization (3). Corrected log ratios L' were spatially normalized by subtracting the median L value within a 9x9 window. The four replicate arrays for each gene expression experiment were processed using the VERA package (4) to estimate multiplicative and additive errors and to associate a p -value of differential expression with each gene. This approach was also applied for error modeling and significance assessment of the chIP-chip data. However, unlike for gene expression analysis, in which both increases and decreases in fluorescent intensity are of interest, DNA binding is indicated only for increases in intensity, representing increased promoter binding in the IP-enriched sample versus the IP-unenriched sample. Thus, significance of DNA binding must be assessed using a one-sided test. The VERA likelihood ratio test was therefore modified to use a one-sided statistic [force $\mu_x \geq \mu_y$ in the denominator of Eqn. 5 of reference (4)].

Overlap in Binding Between Conditions. For each transcription factor, p -values of binding from MMS treated (+MMS) and untreated (–MMS) experiments are integrated to select sets of genes bound in –MMS only, +MMS only, or in both conditions, as displayed in Fig. 2 of the main text. Typically, genes are selected in a single condition by including those with p -values below a certain threshold, such as $t=0.01$. However, given that two conditions c_1 and c_2 are under consideration, the dependency in the corresponding pair of p -values (p_1, p_2) measured for each gene can be exploited to achieve a more sensitive selection. Treating the pair of values as replicate measurements of binding, a compounded p -value of binding (in one, both, or neither condition) can be computed by multiplying p -values, i.e., genes for which $p_1 \cdot p_2 < t'$ are chosen as bound.

Based on this observation, we combine the one- and two-condition thresholding methods, associated with t and t' above, to partition the two-dimensional space of p -values into four regions as shown in fig. S2. Region (A) is defined by $(p_1 \leq t, p_2 > t'/t)$ and represents binding in c_1 only. Region (C) is symmetrically defined by $(p_1 > t'/t, p_2 \leq t)$ and represents binding in c_2 only. Region (B) is the area defined by $(p_1 \cdot p_2 \leq t', p_1 \leq t'/t, p_2 \leq t'/t)$ and represents binding in c_1 and c_2 . The remaining values of (p_1, p_2) not covered by regions (A, B, or C) represent lack of binding. Furthermore, for a given simple threshold t , the compound threshold t' is set so that regions (A, B, and C) are equal in area, which ensures that under the null hypothesis (i.e., no genes are bound in either condition) the expected numbers of genes which fall into each region are the same. This two-dimensional thresholding scheme is similar to one proposed by Zaykin *et al.* (5) for epidemiological studies.

Significance of expansion or contraction. Given overlap counts (A, B, and C in fig. S2) for each TF, we computed the significance of a possible change in genome-wide binding pattern between untreated and MMS-treated conditions. To construct a null model in which the TF binding profiles did not change across conditions, additional chIP-chip experiments were performed in untreated conditions for each of three transcription factors Gcn4, Crt1, and Rpn4. As with all other experiments, three replicate experiments per TF were normalized and processed using the VERA/SAM package. These data were

compared to the original chIP experiments for Gcn4, Crt1, and Rpn4 generated by the Young laboratory (1, 6), also in untreated conditions.

Each of these untreated (original Young lab) vs. untreated (new Ideker lab) comparisons was analyzed to compute the overlap in binding, exactly as described above for the untreated (original Young lab) vs. treated (new Ideker lab) binding profiles. Promoter counts falling into regions A, B, or C were aggregated across Gcn4, Crt1, and Rpn4 yielding A=347, B=459, C=239. These totals were used to derive a pooled estimate of $\frac{A+C}{2B} = 64\%$ for the proportion of the average number of promoters bound in exclusively one condition (i.e. one experiment in one lab) compared to the number bound in both conditions (i.e. both experiments) in the negative control.

This expected value was compared to the $\frac{A}{B}$ and $\frac{C}{B}$ proportions for the comparison between treatment conditions (absence or presence of MMS) for each of the 30 TFs in our study using a one-sided Fisher's Exact Test (FET). The p -value of contraction (P_C) or expansion (P_E) was defined as the FET significance that the first or second fraction, respectively, was higher than expected under the null model.

Motif Enrichment in Transcription Factor Binding Data. Transcription factor binding site motifs were defined as position-specific weight matrices (PWMs) of log-likelihood ratios (7). PWMs were compiled for 114 different individual TFs from Harbison et al. (6) and public data bases (8, 9). When more than one matrix was defined for the same TF, the PWM with the highest information content per position (relative entropy) was selected. Using the PWM scoring functionality of ANN-Spec (10), the score distribution for each motif was determined over all possible subsequences of the intergenic regions (as defined by the probes on the chIP-chip array) such that a score threshold could be selected to ensure a rate of $< 10^{-4}$ predicted sites per base pair.

ANN-Spec was used to discover potentially new motifs in promoter sets defined by the method described in “**Overlap in Binding Between Conditions**” ($t=10^{-3}$). This defined three non-overlapping sets for each TF: +MMS-only, -MMS-only, and BOTH. For each sequence set, 100 training runs were

performed for pattern widths 8 to 14 using ANN-Spec's discriminative mode against all intergenic regions. Training allowed for zero or more sites per alignment (so called 'ZOPS' occurrence model). Redundant or approximately redundant alignments were merged (correlation coefficient >0.8, over 6 or more contiguous alignment positions in either orientation) and the remaining alignments were filtered against the 114 known TF-motifs in a similar way. The remaining 222 alignments were converted to PWMs and represented our novel motif models (108 in -MMS-only, 71 in +MMS-only and 43 in BOTH).

For each TF and known and novel motif, PWM enrichments were calculated (hypergeometric p-value) in either the -MMS-only or +MMS-only non-overlapping binding sets to identify differentially enriched motifs. Differential enrichment was defined when binding p-values $<10^{-7}$ ($<10^{-3}$ after Bonferroni correction) were observed in one condition, and p-values $>10^{-2}$ were observed in the other condition.

Assessing overlaps in binding between TF pairs. Each of the sets of genes bound by a TF (in -MMS or +MMS conditions using a strict threshold of $p \leq 0.001$) was systematically compared to the sets of genes bound by each of the other TFs using the hypergeometric test. Significant overlaps between sets ($p \leq 0.01$ after Bonferroni correction) are displayed in Fig. 3 of the main text. TFs are linked by a green line if a significant number of genes were bound by both TFs in -MMS and (a possibly different set of genes were bound) in +MMS conditions. If two TFs only bound the same genes in either -MMS or +MMS, they are linked with a blue or orange line, respectively. Hierarchical clustering was performed using the ClustArray program (<http://www.cbs.dtu.dk/services/DNAarray/>).

Deletion Buffering Analysis. Expression profiles (+/-MMS treatment for wild type and 27 TF knockout strains) were analyzed to identify genes that were differentially expressed in the vast majority of profiles but did not change in expression in a particular TF knockout background. In these cases, the gene was said to be deletion-buffered by the TF in question and the TF was "epistatic" to this gene.

For each (gene, TF) combination, the probability of buffering was computed using a Bayesian score function, as follows. Let T_{ko} vs. T_{other} represent the event of true differential expression of the gene in the TF knockout vs. all other strains. Similarly, let p_{ko} and p_{other} represent the corresponding observed values for the gene, which in this case are p -values of differential expression. The value for p_{other} is computed as the combined p -value of differential expression over all profiles excluding the knockout under consideration, according to Fisher's rule of multiplication (11):

$$m = \prod_{\forall i \neq ko} p_i$$

$$p_{other} = m \sum_{i=0}^{n-1} \frac{(-\ln m)^i}{i!}$$

Since most genes are likely regulated (directly or indirectly) by only a fraction of the TFs encoded by the genome, they are expected to behave similarly between the wild type and most of the 27 knockout backgrounds. The combined p -value thus provides a measure of differential expression in + vs. – MMS conditions that is more robust than the single p -value obtained for a wild type experiment. Given measurements of p_{ko} and p_{other} , the probability of deletion-buffering is equivalent to the probability that the gene is truly differentially expressed in other experiments but not in the knockout in question. Using Bayes' rule:

$$\Pr(T_{ko} = 0, T_{other} = 1 | p_{ko}, p_{other}) = \frac{\Pr(p_{ko}, p_{other} | T_{ko} = 0, T_{other} = 1) \Pr(T_{ko} = 0, T_{other} = 1)}{\Pr(p_{ko}, p_{other})}$$

$$\equiv \frac{\Pr(p_{ko} | p_{other}, T_{ko} = 0, T_{other} = 1) \Pr(p_{other} | T_{ko} = 0, T_{other} = 1)}{\Pr(p_{ko}, p_{other})} \quad [\text{Because } \Pr(T_{ko}=0, T_{other}=1) \text{ is constant}]$$

$$\equiv \frac{\Pr(p_{ko} | T_{ko} = 0) \Pr(p_{other} | T_{other} = 1)}{\Pr(p_{ko}, p_{other})} \quad [\text{Given } T, p \text{ is conditionally independent of other variables}]$$

By identity, $\Pr(p_{ko} | T_{ko} = 0) = p_{ko}$. The remaining two terms are estimated empirically from the available frequency distributions. $\Pr(p_{ko}, p_{other})$ is estimated by a two-dimensional histogram of the observed values over all genes and experimental conditions in our study. $\Pr(p_{other} | T_{other} = 1)$ is estimated by a histogram of p_{other} restricted to “true” differentially-expressed genes, defined as the set of 255 genes for which the

median absolute expression change was greater than two-fold over at least two of seven time points (5, 15, 30, 45, 60, 90, 120 min.) monitored by Gasch *et al.* after MMS addition (12). Fig. S5 illustrates the probability of deletion-buffering for the factor Crt1 as a function of p_{ko} and p_{other} .

Assembly of Transcriptional Pathways From Integrated Data. Physical mechanisms of transcriptional regulation were modeled using an approach described previously (13). Briefly, we postulated that the regulatory effects of deleting a gene are propagated along paths of physical interactions (protein-protein and protein-DNA). We formalized the properties of these paths and interactions using a factor graph and found the most probable set of paths using the max-product algorithm (14). The raw data used in the modeling procedure included ChIP-chip interactions measured in the presence (this study) and absence (1) of MMS, protein-protein interactions from the Database of Interaction Proteins as of April 2004 (15), and the 341 deletion-buffering interactions found using the method described in “**Deletion Buffering Analysis**”.. For the 30 transcription factors in our study (see Fig. 1a), we used the method described in “**Overlap in Binding Between Conditions**” to identify genes bound in either +MMS only, -MMS only, or both conditions. For 75 additional factors from Lee *et al.* (1) for which we had data for the -MMS condition only, we used a strict p-value threshold of $p \leq 0.001$ to select significant promoter-binding interactions and classified these as -MMS only interactions.

Figure S8 shows all paths that directly connect TFs to deletion-buffered genes. In Fig. 5b of the main text, all paths that directly or indirectly (via one intermediate TF) connect TFs to deletion-buffered genes are shown. To generate integrated models for the *mbp1Δ* and *rtg1Δ* validation experiments (Fig. 5c), the same promoter-binding and protein-protein interaction network was used along with either 14 buffering interactions from the *mbp1Δ* experiment or 20 from the *rtg1Δ* experiment (buffering p-value ≤ 0.004).

References

1. T. I. Lee *et al.*, *Science* **298**, 799 (2002).
2. T. J. Begley, A. S. Rosenbach, T. Ideker, L. D. Samson, *Mol Cell* **16**, 117 (Oct 8, 2004).
3. C. Workman *et al.*, *Genome Biol* **3**, research0048 (Aug 30, 2002).
4. T. Ideker, V. Thorsson, A. Siegel, L. Hood, *Journal of Computational Biology* **7**, 805 (2000).
5. D. V. Zaykin, L. A. Zhivotovsky, P. H. Westfall, B. S. Weir, *Genet Epidemiol* **22**, 170 (Feb, 2002).
6. C. T. Harbison *et al.*, *Nature* **431**, 99 (Sep 2, 2004).
7. G. D. Stormo, *Bioinformatics* **16**, 16 (Jan, 2000).
8. E. Wingender *et al.*, *Nucleic Acids Res* **29**, 281 (Jan 1, 2001).
9. J. Zhu, M. Q. Zhang, *Bioinformatics* **15**, 607 (Jul-Aug, 1999).
10. C. T. Workman, G. D. Stormo, *Pac Symp Biocomput*, 467 (2000).
11. T. L. Bailey, M. Gribskov, *Bioinformatics* **14**, 48 (1998).
12. A. P. Gasch *et al.*, *Mol Biol Cell* **12**, 2987 (Oct, 2001).
13. C. H. Yeang, T. Ideker, T. Jaakkola, *J Comput Biol* **11**, 243 (2004).
14. F. Kschischang, B. Frey, H. Loeliger, *IEEE Transactions on Information Theory* **47**, 498 (2001).
15. I. Xenarios *et al.*, *Nucleic Acids Res* **30**, 303 (Jan 1, 2002).
16. S. A. Jelinsky, L. D. Samson, *Proc Natl Acad Sci U S A* **96**, 1486 (Feb 16, 1999).
17. A. P. Gasch *et al.*, *Mol Biol Cell* **11**, 4241 (Dec, 2000).
18. P. T. Spellman *et al.*, *Mol Biol Cell* **9**, 3273 (Dec, 1998).
19. S. A. Jelinsky, P. Estep, G. M. Church, L. D. Samson, *Mol Cell Biol* **20**, 8157 (2000).

Figure S1.

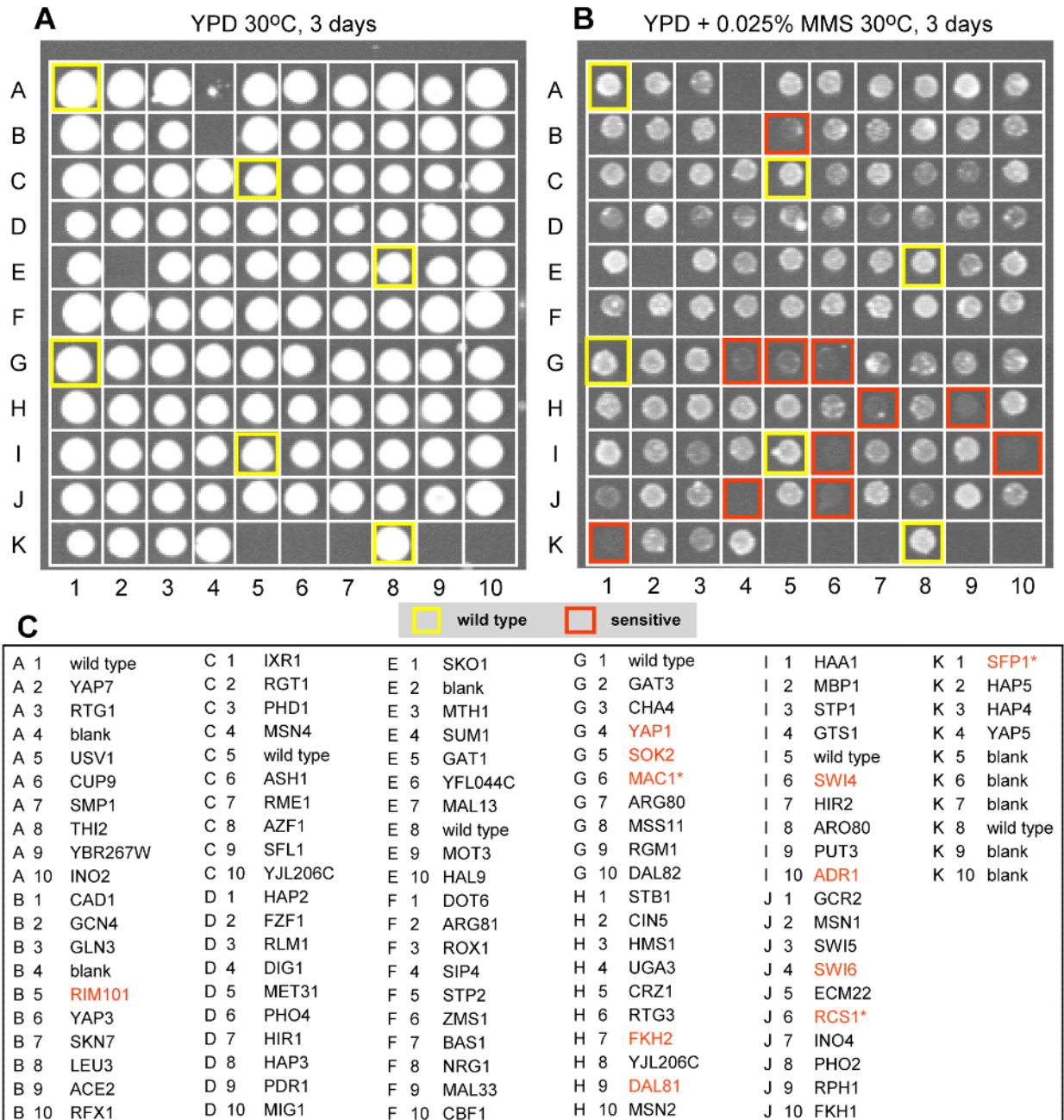


Figure S1: MMS sensitivity assay of transcription factor deletions strains. Light images of yeast colonies on agar plates are shown. Each position in the grid contains the colonies for TF-deletion or wild-type strains grown for 60 hours in YPD (**A**) and YPD +0.025% MMS (**B**). The key for the location of yeast strains is shown in (**C**). Sensitive mutants for this assay are indicated in red. Factors marked with ‘*’ were not found to be sensitive in subsequent trials.

Figure S2.

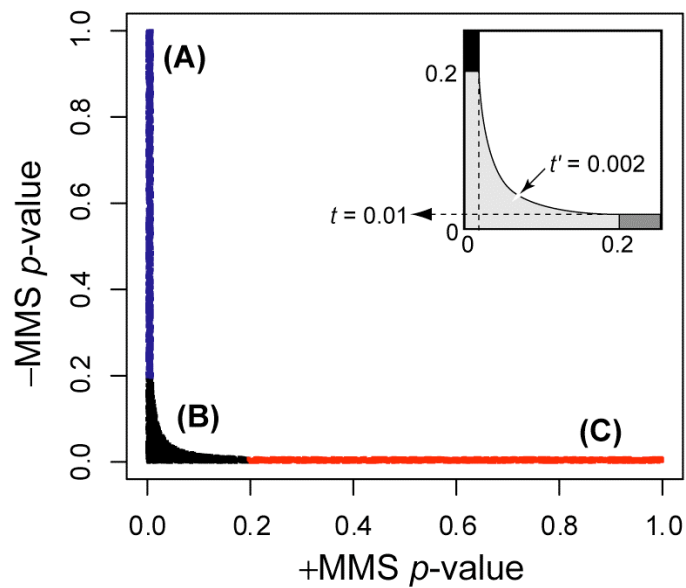


Figure S2: Scoring overlap in TF binding between conditions. For a given TF, each gene is assigned a pair of p -values for binding in the presence (+MMS) versus absence (-MMS) of MMS. Genes with p -value pairs that fall into the region (A), (B), or (C) are considered bound in +MMS only, both conditions, or -MMS only, respectively. The inset shows an enlargement of the plot for p -values < 0.25. The regions shown correspond to p -value thresholds of $t=0.01$ and $t'=0.002$. The thresholds used in the actual study were $t=0.001$ and $t'=1.5 \times 10^{-4}$.

Figure S3.

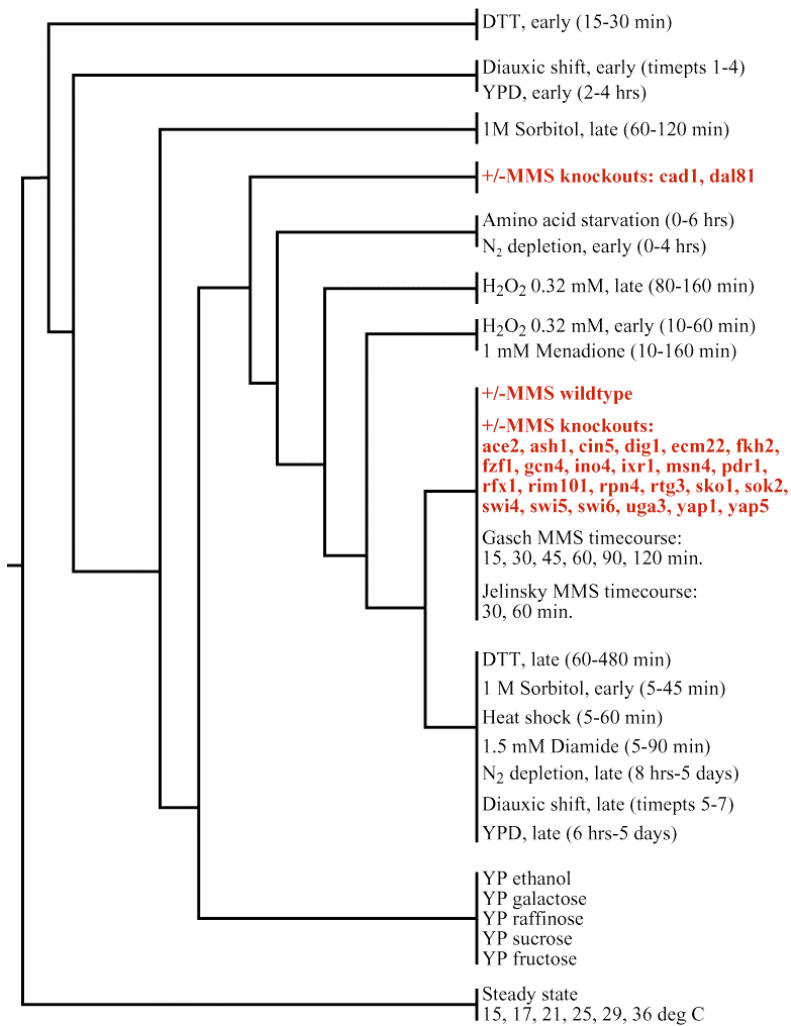


Figure S3: Clustering of MMS expression responses with other stresses. Expression profiles generated in this study (wild type and TF knockouts +/- 0.03% MMS) were clustered together with previous expression data for MMS [Gasch et al. (12) and Jelinsky et al. (16)] and other environmental stresses (17). Hierarchical clustering was performed using ClustArray (<http://www.cbs.dtu.dk/services/DNAarray/>) to construct a dendrogram on the conditions. The dendrogram shows that the +/- MMS wild type profile from this study is in relative agreement with previous ones. Moreover, all but two (*cad1Δ* and *dal81Δ*) of the +/- MMS knockout profiles are more similar to the MMS wild type response than to other stress responses.

Figure S4.

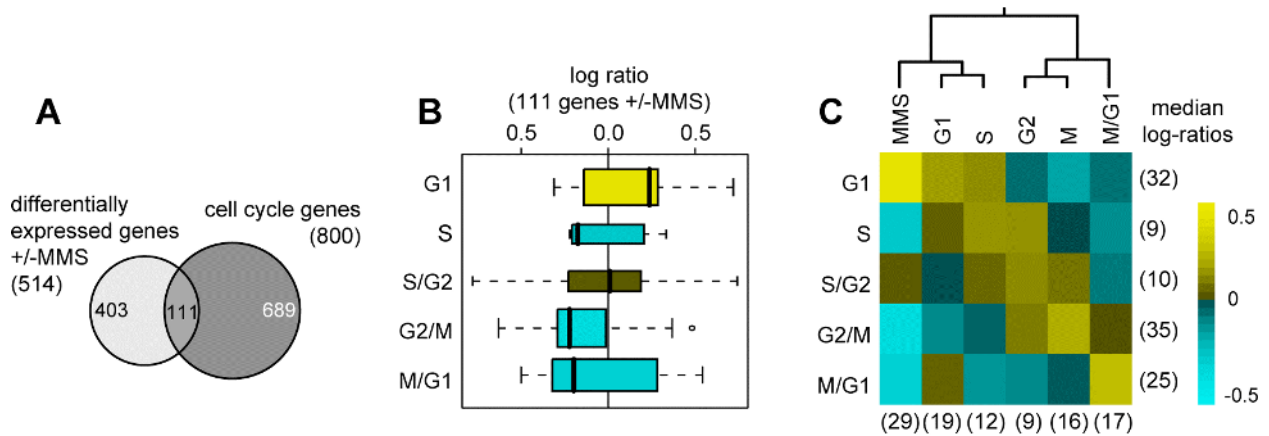


Figure S4: MMS responsive genes implicate G1/S checkpoint. (A) Approximately 22% of the MMS-responsive genes (111 of 514) have been shown to be temporally regulated during the cell cycle (18). (B) Distributions of the median log-ratios (over the 27 TF knockouts and wild-type) for genes found to be up-regulated in G1, S, S/G2, G2/M, M/G1. Boxplot colors correspond to the color scale in C. (C) Clustering these MMS responsive genes with representative profiles from cell cycle phases G1, S, G2, M and M/G1 (18) shows that the MMS profile is most similar to that of late G1 and S phases. Further evidence of G1 and S phase arrest can be found in the expression responses of *ACE2* and *SWI5*, cell cycle regulators expressed in M/G1 phase (18). Both factors are down regulated in response to MMS (Fig. 2a in the main text), and have contracting binding profiles (Fig. 2b). Moreover, the Ace2 and Swi5 TFs only show significant binding overlap in -MMS (Fig. 3). Together, these findings confirm that exposure to MMS causes cells to progress more slowly through G1/S and S phase (19).

Figure S5.

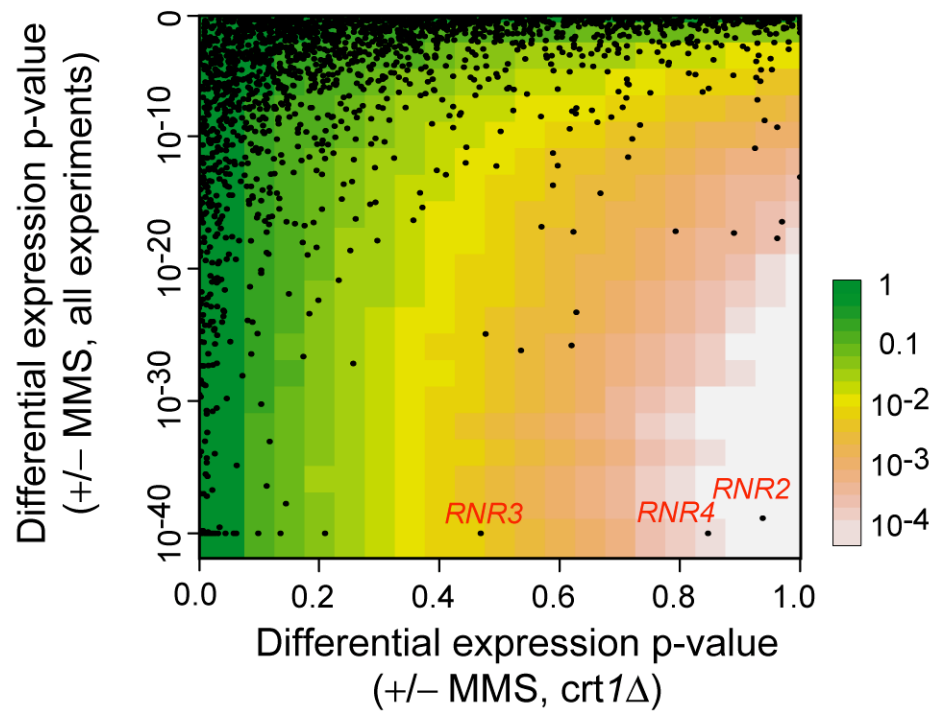


Figure S5: *Crt1* deletion-buffering p -values (colors ranging from green to white) are computed for each gene (black points) based on their p -values of differential expression for *crt1Δ* (x-axis) versus all experiments (y-axis).

Figure S6.

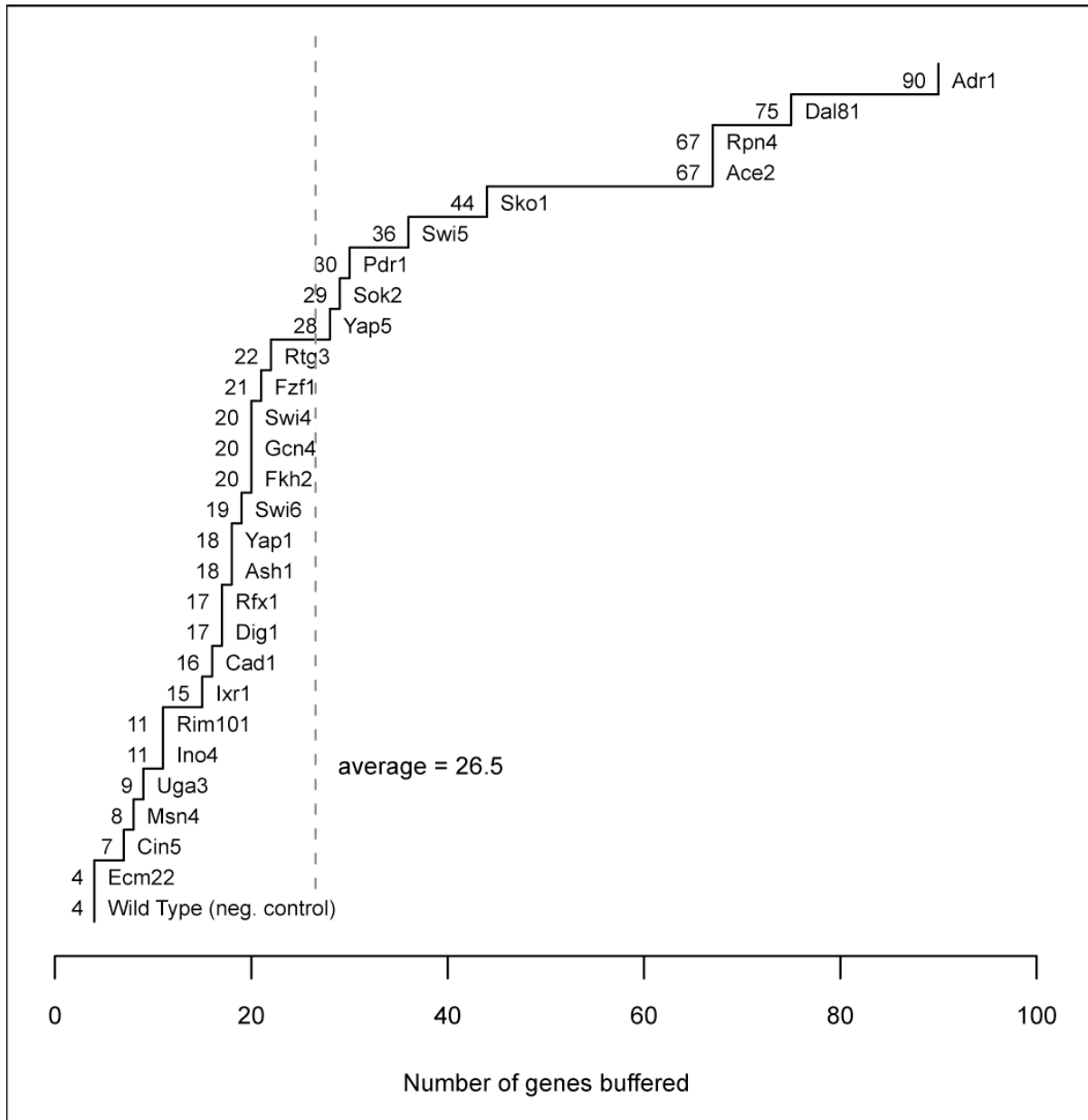


Figure S6: Number of genes buffered by each transcription factor deletion (buffering p-value <0.005). Because more than one transcription factor may buffer each gene, the number of buffered genes (341) is less than the total number of buffering interactions observed (757). As indicated by the dashed grey line, the average number of genes buffered by each TF is 26.5.

Figure S7.

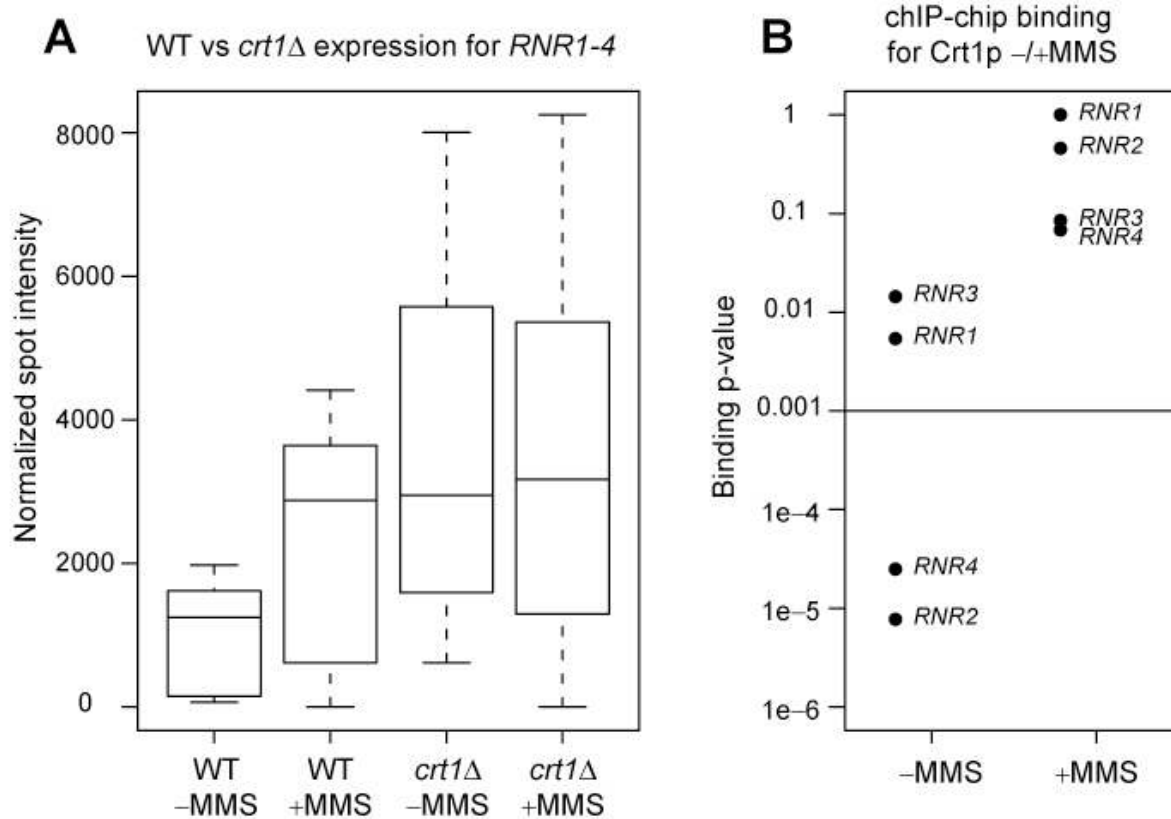


Figure S7: Correspondence between expression and regulation of the *RNR* genes by Crt1p. (A) Distributions of normalized spot intensities (cy3 and cy5 channels) for the four *RNR* genes in *crt1*Δ + vs. -MMS over four replicate microarrays (16 spot intensities for each distribution). The wild-type expression levels show an induction of these genes after exposure to MMS while the *crt1*Δ experiment shows constitutive expression for both + and -MMS. **(B)** The binding data show strong evidence for Crt1p binding upstream of *RNR2,4* (and marginal evidence for *RNR1,3*) in nominal growth conditions while binding evidence suggests a lack of Crt1p binding for all *RNR* genes after exposure to MMS. These data support the induction (by derepression) observed in the first two distributions in (A).

Figure S8.

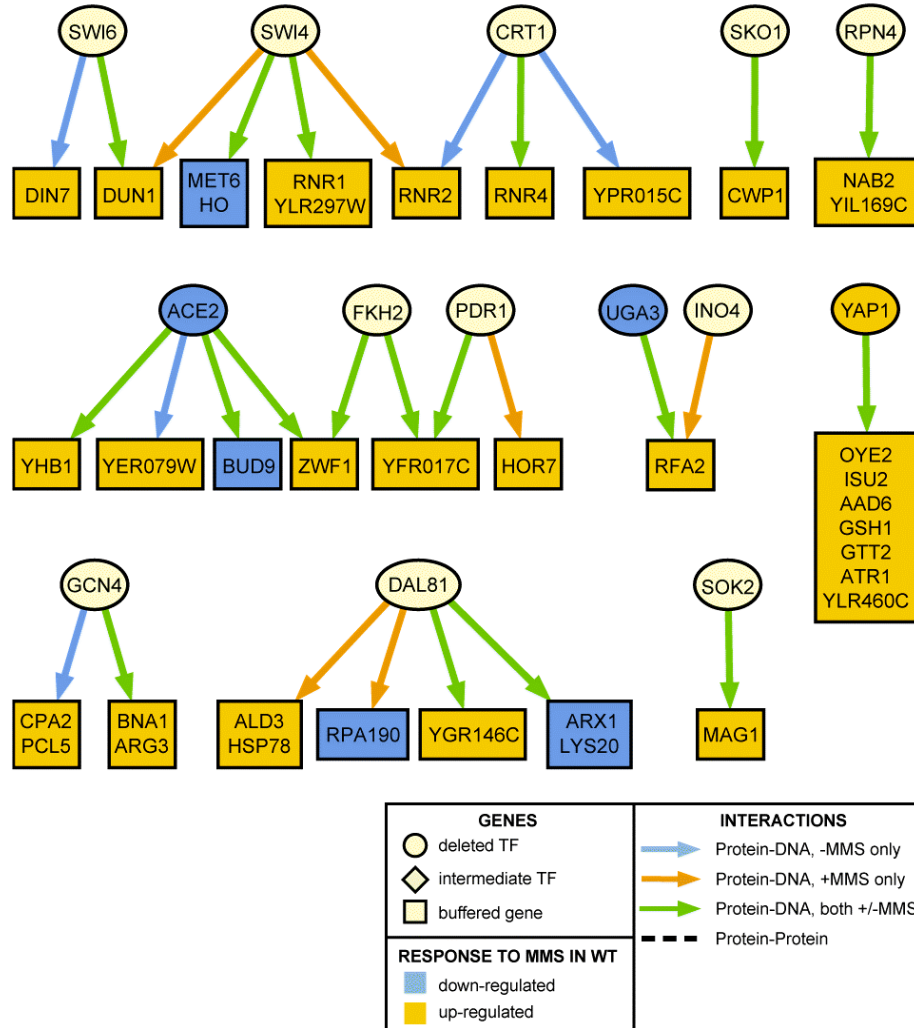


Figure S8: Direct binding interactions supported by deletion-buffering effects. Circles or squares represent deleted TFs or target genes, respectively. Protein-DNA interactions between TFs and the gene promoters they bind are represented as arrows colored blue (bound in -MMS only), orange (+MMS only), or green (both + and -MMS). Genes are colored orange or blue, representing induction or repression in response to MMS. Each binding interaction shown is supported by a deletion-buffering effect. For example, deletion of the transcription factor *SWI6* results in lack of expression of *DIN7* and *DUNI* in response to MMS. In total, 42 binding interactions covering 37 distinct target genes are supported.