Published in final edited form as:
Nature. 2010 April 22; 464(7292): 1187–1191. doi:10.1038/nature08934.

Genetic Analysis of Variation in Transcription Factor Binding in Yeast

Wei Zheng[1,*], **Hongyu Zhao**[2,3], **Eugenio Mancera**[4], **Lars M. Steinmetz**[4], and **Michael Snyder**[1,5]

[1]Department of Molecular, Cellular and Developmental Biology, Yale University, New Haven, CT 06520.

[2]Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT 06520.

[3]Department of Epidemiology and Public Health, Yale University School of Medicine, New Haven, CT 06520.

[4]Genome Biology Unit, European Molecular Biology Laboratory, Heidelberg, Germany.

[5]Department of Genetics, Stanford University School of Medicine, Stanford, CA 94305.

Abstract

Variation in transcriptional regulation is thought to be a major cause of phenotypic diversity1,2. Although widespread differences in gene expression among individuals of a species have been observed3-8, few studies have examined the variability of transcription factor (TF) binding, and thus the extent and underlying genetic basis of TF binding diversity is largely unknown. In this study, we mapped differences in transcription binding among individuals and elucidated the genetic basis of such variation on a genome-wide scale. Whole-genome Ste12 binding profiles were determined using ChIP Seq in pheromone-treated cells of 43 segregants of a cross between two highly diverged yeast strains and their parental lines. We identified extensive Ste12 binding variation among individuals and mapped underlying *cis*- and *trans*- acting loci responsible for such variation. We showed that the majority of TF binding variation is *cis*-linked and that many variations are associated with polymorphisms residing in the binding motifs of Ste12 as well as those of several proposed Ste12 cofactors. We also identified two trans factors, *AMN1* and *FLO8*, that modulate Ste12 binding to promoters of more than 10 genes under α-factor treatment. Neither of these two genes was known to regulate Ste12 previously, and we suggest that they may be mediators of gene activity and phenotypic diversity. Ste12 binding strongly correlates with gene expression for more than 200 genes indicating that binding variation is functional. Many of the variable bound genes are involved in cell wall organization and biogenesis. Overall these studies

Correspondence should be addressed to M.S. (mpsnyder@stanford.edu)..

[*]Present address: Biostatics Resources, Keck Laboratory, Yale University, New Haven, CT 06520.

Author contribution statement. WZ and MS designed the study, EM and LS provided yeast strains, WZ performed experiments, WZ and HZ analysed the data, EM, LMS and MS provided helpful suggestions in data analysis, and all authors co-wrote the paper.

The authors have no competing interests.

Raw data are in the GEO database (http://www.ncbi.nlm.nih.gov/geo/) under accession number GSE19636.

identified genetic regulators of molecular diversity among individuals and provide novel insights into mechanisms of gene regulation.

Many, and possibly the majority, of phenotypic differences among individuals are likely to be elicited by alterations in gene expression and the underlying transcriptional regulation1,2. Differences in gene expression have been observed among individuals in a variety of species, including yeast3-8. These differences may be mediated at a variety of levels, including transcription, mRNA processing and mRNA stability, and the contribution of each level is unknown. Recently, we and others discovered a surprisingly high frequency of TF binding site variation among closely related species9,10, but the extent to which this occurs in individuals is unknown.

To investigate the extent of diversity in transcription factor binding among individual strains of yeast, we used ChIP-Seq to generate high-resolution Ste12 binding profiles for two distinct yeast strains, S96 (isogenic to S288c) and HS959 (isogenic to YJM789), and 43 *MATa* genotyped segregants of these lines. For the parent strains, the profiles were generated in the presence and absence of the mating pheromone α-factor9, and for the segregants only α-factor treatment conditions were examined. At least two biological replicates were performed for each strain along with one input DNA control (see Methods).

From vegetatively grown cells, we found 536 and 846 binding regions in S96 and HS959 cells, respectively (Fig. 1b; Tables S2, S3). After treatment with α factor, Ste12 bound many more regions using the same threshold (1179 in S96 and 1112 in HS959, Fig. 1b, Tables S4, S5), consistent with the role of Ste12 as a major transcriptional activator in pheromone response11. Assignment of binding regions revealed that many of the gene targets were identical in vegetative and pheromone treated cells (Fig. S1); many of these common targets are involved in conjugation, and may have an additional role during vegetative growth.

Comparison of the signal tracks revealed extensive differences among the different yeast strains and their progeny. We quantitatively identified the variable binding regions across all 43 segregants and the two parental strains under pheromone-treatment conditions by calculating a Normalized Difference (NormDiff) Score and defined the genomic positions within the top 2.5% of binding variance across all 45 strains as variable (see Methods). Adjacent variable regions were grouped into 930 genomic “traits”. Inspection of the signal tracks revealed that the majority (~ 78%) of the variable binding regions exhibit Mendelian segregation in the progeny (Fig. 1a). Interestingly, a large number of transgression effects were observed in which the segregants either obtained or lost binding relative to the parents (Fig. 1a).

To identify the underlying QTLs for each binding variation trait, we used the genotype information from Mancera et al. which contained 53,460 markers12 (see Methods). The neighbouring markers with the same genotype distribution across strains were grouped, resulting in 2,592 nonredundant marker sets of 1,132 bp median size. We performed single marker regression with the 930 binding traits and the 2,592 markers and found significant marker-trait associations for 195 traits at a local false discovery rate (FDR) of 0.01 (see Methods). As shown in Figs. 2a, 166 traits were associated with *cis*-markers (defined as markers on the same chromosome as the associated traits); 157 of these *cis*-traits were within 10kb of the associated marker. 35 traits were associated with *trans*-markers (defined as markers on different chromosome from associated traits); the trans markers each affected between one and 11 loci (Fig. 2b). Six traits were affected by both *cis*- and *trans*- markers. The effect for the *cis* markers is much stronger than for the trans markers (Fig. 2c). Thus, the vast majority of significantly associated markers (85%) lay *cis* to the variable binding site, indicating that most binding differences are mediated by adjacent regulatory elements. Changing the linkage threshold yielded similar results (Fig. S4).

We next examined the underlying genetic basis of the variable binding loci that were affected in *cis*. Searching for position weight matrix (PWM) matches of the Ste12 motif in the genomic regions of the 166 significant *cis*-variable binding traits of the two parental strains revealed that 102 contained at least one Ste12 motif. 36 of these Ste12 motifs in 31 variable binding regions were affected by polymorphisms: 14 contained SNPs (Fig. 3 and Fig. S5) and, the majority (22) were disrupted by Indels. For the majority of SNPs (11 out of 14), the poorer match to the Ste12 consensus motif corresponds to a lower Ste12 binding signal. Two examples are shown in Fig. 3a (see Fig. S5 for two exceptions). In contrast to that observed for the variable regions, we found 177 Ste12 motifs in the non-variable Ste12 binding regions and only 2 out of these 177 motifs contained a SNP or Indel (Pearson's Chi-square test, $P < 0.001$, Table S6).

To examine whether there are *cis*-mutations in the binding sequences of other TFs affecting Ste12 binding, we searched PWM matches of all known TF recognition consensus sequences13 in all Ste12 binding regions with a particular focus on the variable binding regions. Excluding Ste12, 38 TF consensus sequences were over-represented in Ste12 bound regions (Table S6, Bonferroni corrected $P < 0.05$; see Methods). Of the 38 TF recognition sequences, 219 motifs contained at least one polymorphism in the 166 *cis*- variable Ste12 binding regions. In contrast, only 175 contain motifs with polymorphisms in more than 1,000 non-variable Ste12 binding regions (Pearson's Chi-Square test, P <0.001). In total, of 166 *cis*-traits, 72 traits had polymorphic motifs: 10 (6.0%) with an altered Ste12 motif, 41 (24.7%) with altered motifs of other TFs, and 21 (12.7%) with altered motifs of both Ste12 and other TFs. We presume that the *cis*-variable regions which lack polymorphisms are due to differences in other TF binding regions, chromatin effects, or perhaps affect other genes which in turn affect TF binding through feedback loops or post-transcriptional mechanisms14.

We next analyzed whether polymorphisms in a specific TF motif correlated with Ste12 binding in the variable regions. The presence of polymorphisms in six TF motifs correlated positively or negatively with Ste12 binding (Table S6). For example, for Yhp1 which associates with Mcm115 (a Ste12 associated protein16), and Yap5, whose deletion causes a defect in haploid invasive growth17 (Fig. 3 b, c), 18 variations (4 SNPs and 14 Indels) and 27 variations (12 SNPs and 15 Indels) lie in their respective binding motifs. In both cases the poorer consensus motifs match associated with reduced Ste12 binding (Binomial test, P = 0.0096 and 0.0038 respectively, Table S6). These results suggest that Yhp1 and Yap5, along with four other factors, function to facilitate Ste12 binding.

In our linkage analysis, we found 194 *trans*-markers significantly linked to binding variation traits on different chromosomes (Fig. 2b). We used an independent hierarchical clustering method that has greater statistical power6 to also identify *trans*-regulators (see Methods). The 930 quantitative traits were first grouped into 121 mutually exclusive clusters of high significance. 85 of the clusters contained three or more traits (the largest cluster had 46 traits) and 70% of the clusters contain traits on different chromosomes. We next tested for association between these cluster-level traits and the genotype markers. At an FDR of 0.01 we found significant linkages for 51 clusters (P-value $< 2 \times 10^{-5}$), including 16 clusters with traits on different chromosomes. Six distinct genotype markers from 7 trans clusters were in common with those identified by the linkage analysis (Figs. 2a and S6; see methods) and are potential *trans*-QTLs that affect Ste12 binding.

Four of the common trans clusters with Ste12 binding variation contained distinct classes of genes. Two clusters shown in Fig. S6a and g exhibit Ste12 binding variation in the promoters of *DSE1*, *DSE2*, *SCW11*, *CTS1*, and *WTM1* which are enriched for the GO categories: “cytokinesis” (corrected $P << 0.001$) and “cell wall organization and biogenesis” (corrected $P = 0.045$). The cluster in Fig. S6e includes binding variation in promoters of several *FLO* genes (e.g. *FLO1*, *FLO11),* suggesting that the QTL regulates pathways involved in flocculation18 and/or cell-cell interaction19. The fourth cluster (Fig. S6f) includes 9 binding variations on chromosomes II, IV, VI, VII, XI, XIV and XV, with the highest association to a marker on chromosome II. The chromosome II region contains several interesting genes including chromatin remodelling factors. These results indicate that not only can *trans* loci be identified, but they also affect distinct classes of genes.

To identify the *trans*-acting genes responsible for the Ste12 binding differences, we disrupted 12 candidate genes in each of the parental strains. The genes were located primarily in three regions containing the *trans*-QTLs shown in Fig. S6a/g, e, and f. For two regions we found a single gene, *AMN1* and *FLO8*, respectively, whose disruption recapitulated the Ste12 binding pattern of the other parent. *AMN1* deletion increased Ste12 binding whereas loss of *FLO8* diminished Ste12 binding, indicating that these are negative and positive regulators of Ste12 binding, respectively (Fig. 4). Interestingly, whereas Flo8 affects highly localized Ste12 peaks, Amn1 affects broadly bound Ste12 peaks, suggesting that Amn1 may operate over larger genomic regions.

AMN1 was previously shown to affect vegetative gene expression in a different set of yeast crosses6,7. Our data therefore indicate that *AMN1* has a role in regulating transcription under several conditions. *FLO8* is often disrupted in laboratory strains20, involved in pseudohyphal growth18,21, and was shown previously to have a role in Ste12/Tec1 binding at one promoter22. However, its binding motif is not enriched near Ste12 binding sites and it was not known to have role in either Ste12 function or in mating. Our data indicate that *FLO8* is directly or indirectly important for Ste12 binding at a number of loci during α-factor treatment. Flo8 has been reported to localize to specific loci, including the promoters of *FLO1*, *FLO11*, and *CIN5*23, consistent with a direct regulatory role at these genes. Thus, our study revealed two novel loci affecting Ste12 binding during the pheromone response and yielded several other candidate regions.

To determine if the variation in Ste12 binding is functional, we examined the correlation between binding and gene expression at the Ste12 target genes. Gene expression was measured using DNA microarrays for all 45 strains grown under the same conditions as the ChIP-Seq experiments. Many traits exhibit a positive correlation between binding and expression (permutation test, $P < 0.001$; Fig. 5), consistent with Ste12's role as a transcriptional activator. Interestingly, however, a number of traits were negatively correlated with gene expression (Fig. 5), suggesting that Ste12 may also function as a repressor, which had not been reported previously. Overall, 222 unique target genes (28%) had an absolute correlation coefficient for gene expression and Ste12 binding intensity higher than 0.335; 171 were positively correlated and 51 negatively correlated (Fig. 5). Although the results are highly significant ($P < 0.001$ from 1,000 permutations), we presume that the absolute correlation is not higher because of the presence of additional binding sites for Ste12 and/or other factors, or additional mechanism of gene regulation. The positively correlated genes were enriched in the GO category “cell wall organization and biogenesis” ($P < 0.001$).

We next grouped the patterns of binding variation and expression variation at individual genes and found 14 clusters of variable Ste12 binding regions that exhibited strongly correlated or anti-correlated gene expression (Fig.5a, Fig. S8). 12 of 14 clusters include *cis*-binding variations on the same chromosome. For the *trans*-cluster affected by *AMN1*, all Ste12 binding variations are positively correlated with expression of their target genes, suggesting that *AMN1* regulates gene expression through Ste12 binding to these genes. Another *trans*-cluster with binding traits on chromosome VI and X also positively correlated with gene expression. Notably, these traits correspond to a translocation between chromosomes VI and X in HS959, and are actually in *cis* for segregants inheriting the HS959 genotype (Fig. S8), thereby suggesting that coordinated expression patterns can disperse in strain diversification.

Interestingly, we often observed combinations of both positively and negatively correlated binding-expression patterns in the same cluster (Fig. 5a, second and third cluster; Fig. S8). These results suggest that Ste12 binding may act as an activator on some gene promoters and a repressor on others. Alternatively, the differential expression may be due to an indirect effect of Ste12 binding through an activator or repressor of the target genes.

Although the *trans*-cluster affected by *FLO8* was not among the most highly correlated for binding and expression, most binding traits in the *FLO8* cluster are correlated with the expression of downstream gene targets. Specifically, four binding regions positively correlate with the expression of three adjacent genes, *FLO1* (two regions; r = 0.554 and 0.577), *FLO11* (r = 0.574), and *CIN5* (r = 0.490) and one region negatively correlates with the expression level of its neighbouring genes *YPS6* and *GTT1* (r = −0.306 and −0.340 respectively). Furthermore, although not directly downstream, all binding traits in this cluster are highly correlated with the expression level of two filamentous growth genes, *HMS1* (r = 0.600, 0.615, 0.650, 0.725, 0.715 respectively) and *TMN3* (r = −0.500, −0.591, −0.599, −0.526, −0.643 respectively). Thus, *FLO8*-influenced Ste12 binding appears to be functional in regulating gene expression. The overall correlation of Flo8-mediated binding and gene expression may not have been revealed initially because some genes are influenced positively and others negatively.

Overall, our studies have revealed that TF binding variation in different yeast strains is wide-spread and that the majority of the strongest differences can be explained by genetic variants acting in *cis*. The *cis* events are often mediated by variation in consensus sequences required for binding of Ste12 or other co-associated TFs. Through genetic analysis on Ste12 binding, we were able to pinpoint potential causative DNA polymorphisms in TF binding motifs, identify novel association between TFs, and identify novel functional roles of Ste12 in activation and repression of different genes.

Our results concerning Flo8 and the covariance of motifs of other TFs such as Yhp1 and Yap5 with Ste12 binding, suggest that many factors not previously known to be involved in mating may have direct roles in the binding of key regulators such as Ste12 at specific loci. This result is consistent with the analysis of TF binding during the salt response where many different combinations of factors were present at different gene promoters24. Our results suggest that although major transcriptional regulators can control the expression of many inducible genes, the particular combination of factors at specific loci are ultimately responsible for individual gene expression. Thus, gene expression is regulated both globally and locally by different factors. Analysis of both *cis*- and *trans*-acting factors in different individuals of the same species can reveal a large number of new regulators of gene expression, even in well-characterized processes such as yeast mating.

The observation that *AMN1* and *FLO8* vary in different laboratory strains6,7,18,20 and affect several cellular processes such as vegetative growth and mating raises the possibility that these genes determine not one, but several phenotypic traits. Each of these genes, along with many other QTLs identified in our study, affect the expression of genes involved in the response of cells to the environment, such as constituents of the cell wall and flocculation genes. Thus, we speculate that *AMN1* and *FLO8* and other QTLs are important for controlling the response to different environmental conditions and phenotypic diversity.

Methods Summary

We performed over 200 ChIP-Seq experiments to determine the Ste12 binding profiles in 43 *MATa* segregants and their isogenic *MATa* parent strains upon α-factor treatment1; the parental strains were also analyzed without pheromone treatment. Ste12 bound peaks were identified using MACS2. To identify the variable binding regions across different strains, we calculated a normalized difference score (NormDiff) and then the variance of the NormDiff scores at each genomic position for all 45 strains, and the top 2.5% variable positions (variance > 10) were chosen. 930 highly variable regions were obtained, and the sum of NormDiff scores in each region was treated as quantitative traits. We used these 930 variable binding traits and 2592 merged genotype markers to conduct linkage analysis and identify *cis*- and *trans*- factors by single marker regression. For *cis*-factors, motif analysis was performed to search for enriched TF motifs in variable Ste12 binding regions, for co-occurrence of SNP/Indel containing motifs and differences in Ste12 binding. To identify trans-factors, we used two methods: a) the linkage analysis of individual traits and b) hierarchical clustering of all 930 variable binding traits and linkage analysis with cluster-level traits. The trans-associated markers passing significance threshold (FDR = 0.01) in both methods were considered further. Six genomic regions around most significant trans-associated markers were examined in detail: 12 candidate genes in these regions were deleted in both parent strains, and ChIP-Seq experiments were used to examine Ste12 binding profiles in the mutant strains. To determine whether Ste12 binding variation affects gene expression, microarray experiments were performed and compared with the Ste12 binding data.

Methods

Strains and Plasmids

Parental yeast strains are S96 (isogenic to *S288c*) and YJM789. YHS959 is an isogenic *MATa* derivative of YJM789. *MATa* segregants obtained from S96/YJM789 cross3 were selected by two criteria: 1) a relatively even distribution of breakpoints over the genome so that loci are as dissociated as possible; 2) a relatively even distribution of *STE12* polymorphisms over the segregants. All strains were genotyped previously with high-resolution tiling array at 55958 SNP and Indel markers3. *STE12* loci were tagged with c-myc at the 3′ end using one-step gene replacement with plasmid pFA6a (9X-myc, *KanMX4*). Successful tagging for all strains was confirmed by colony PCR. To identify causative genes in 3 trans-QTLs, 12 candidate genes AMN1, ICS2, SLI15, FLO8, GLE2, CTF19, HIR1, LDB7, PDR3, SCT1, ALK2, SLA1 were deleted from each parental strain by homologous recombination.

Pheromone treatment of yeast culture

Yeast strains were grown in 500ml YPAD media to mid-log phase (O.D.$_{600}$ = 0.6), and then 0.5 ml 5 mM alpha factor (solution in 0.1M sodium acetate pH 5.2, GenScript) was added. Cultures were incubated for 30 minutes at 30°C with vigorous shaking. The optimal pheromone concentration and incubation time were determined by measuring the expression levels of four known *STE12* target genes (*AFR1*, *FIG1*, *FUS2*, *PCL2*) using PCR.

ChIP-sequencing

Pheromone-treated yeast cultures were cross-linked with 1% formaldehyde for 15min at room temperature. Cells were harvested using filtration, and washed twice with water. Chromatin immunoprecipitation (ChIP) were performed as described in 4, using 160 uL of 50% anti-myc EZ-view affinity gel (Sigma). At least two independent biological replicates grown at different times were completed for each strain. Before adding antibody, 10% volume of the sonicated cell lysate was used to prepare input DNA from each sample that was not subjected to immunoprecipiation. Illumna sequencing libraries were made from ChIP and input DNAs and sequencing using multiplexing barcoded adapters as described in 1. The background signal for the input DNA in our experiments was quite low. The number of uniquely mapped sequencing reads determined for each segregant is summarized in Table S1.

In total, over 200 ChIP-Seq experiments were performed. For each strain, an average of 3.4 and 1.8 million uniquely mapped reads of 30 bp were obtained for ChIP and Input sequencing, respectively (Supplementary Table 1). At this sequencing depth, we expect to identify over 95% of the TF binding sites with more than 2-fold enrichment of sequencing tags1. Approximately 61.6% of the sequencing reads uniquely mapped to the S288c reference genome.

Identifying Ste12 binding peaks and target genes in parental strains

To quantify the binding differences between the two parent strains, we first identified the Ste12 binding regions in parental strains by comparing each ChIP dataset against its corresponding control data using the peak scoring algorithm MACS2. We scored Ste12 binding peaks in parental strains S96 and YJM789 under vegetative growth; S96 and HS959 under pheromone treatment. The parameters used are: effective genome size = 1.20e+07; band width = 15; model fold = 4; pvalue cutoff = 1.00e-05. Output from MACS and the raw sequencing data are available in the GEO database.

We assigned binding regions to the nearest genes and found statistically significant overlap between the gene sets identified in the two parental strains: 302 (39.27%) and 616 (53.94%) overlap in vegetative growth and pheromone-treated cells, respectively (hypergeometric P < 0.001 in both cases). Within a strain 311 (31.19%) and 376 (32.84%) of the vegetative gene targets overlapped with targets in the pheromone treated condition for S96 and HS959, respectively (Supplementary Fig.1). GO analysis revealed that many of the shared genes for both vegetative and pheromone treated cells are enriched in conjugation genes (Fig. 1c and Fig. S1)

Identification of quantitative Ste12 binding differences

After scaling ChIP and Input data, at every genomic position, we calculated a normalized difference (NormDiff) score to represent the scale-adjusted, background extracted ChIP signal. Detailed steps are as follows.

We first used Eland to uniquely mapped sequencing reads and generate signal maps, using a similar strategy as 5 and 6. Let x1, … , xn be genomic positions covering an entire chromosome. We observe A (xi) and B (xi) be the signal at position xi, i = 1, … , n for our experimental signal map and the control signal map. Assume the model $A(x_i) \sim Poisson(f + g)$, $B(x_i) \sim Poisson(cg)$ where f (xi) is the signal caused by TFBS, g(xi) is the non-random background signal, potentially caused open chromatin status, or other systematic bias. From Poisson model assumption, we have the following properties: $E(A - B/c) = $ f; $Var(A - B/c) = f + g + g/c$. Our purpose is to evaluate f (xi).

To estimate scaling factor c, we calculated a series of ratios $r_i = B(x_i)/A(x_i)$ for positions x_1, … , x_n, $i = 1, …, n$. At each position, the ratio r_i is an estimate for the scaling factor c. However if the position is in a TFBS, r_i will be an underestimate for c. For sequence-specific transcription factors, there are usually less than half of the genomic regions enriched for a ChIP signal; therefore we use the median of ratios to estimate c. In case $A(x_i) = 0$, we flag the position by assigning a large number to r_i. We denote $\sigma = \sqrt{f + g + g/c}$ and define normalized difference score (NormDiff) as $Z\left(x_i\right)=\frac{A\left(x_i\right)-B\left(x_i\right)/c}{\sigma}$, Under $H_0 : f(x_i) = 0$, we have $Z(x_i) \sim N(0,1)$. Inspired by the dynamic parameter selection in 2, we estimate the unknown standard deviation σ by average signal densities from local or global ranges.

$$\widehat{\sigma}=\max_{w\in\{1,10,[all]\}}\sqrt{\bar{A}_w+\frac{\bar{B}_w}{c^2}}$$

where $\bar{A}_w$ and $\bar{B}_w$ are average ChIP and Input signal densities per window, from $2w+1$ positions centered at position x_i. The default settings we use are w = 10 bp, w = 10 bp, and all positions across the chromosome. Comparing to a naïve estimate of σ that only uses information at position x_i, this dynamic estimate $\widehat{\sigma}$ is robust to both very low signal densities and very high signal densities caused by random fluctuations in a small local region.

To summarize NormDiff score into quantitative traits, we first calculated $Z(x_i)$ for each strain at each genomic position and the variance of $Z(x_i)$ across 45 strains, denoted as $\text{var}(Z(x_i))$. Then we defined x_j as a highly variable position when $\text{var}(Z(x_j))$ is one of the largest 2.5% among all the variances $\{\text{var}(Z(x_i)), i=1, \ldots, n\}$. Finally we merged all selected x_j to highly variable regions at least 20 bp apart from each other and at least 20 bp in length, and used the sum of $Z(x_j)$ in each highly variable region as a quantitative trait. Selection of high variance regions implies that in some of the strains, the ChIP signal is highly different from the background, therefore they correspond to “peaks” caused by Ste12 binding instead of small increases over the background.

84% of the 930 variable traits overlap with binding regions in at least one parental strain, and 51% of binding regions in at least one parental strains overlap with the 930 variable traits.

Investigation of transgressive inheritance

We define transgressive score as the ratio between standard deviation of NormDiff scores among 43 segregants and absolute difference between NormDiff scores of two parental strains. Biologically, if transgressive score is low, the phenotypes are likely to follow Mendelian inheritance, otherwise they exhibit transgressive inheritance. We calculated the transgressive score for each of the 930 variable binding traits, and found that 78% binding variation traits have transgressive score less than 3, and thus exhibit Mendelian inheritance.

Linkage analysis

To test for linkage between an individual quantitative trait and a genetic marker, we performed simple linear regression to compare the quantitative trait differences between genotype groups. The genotype information was obtained from Mancera et al.3. In their original study, 55958 markers were genotyped. We removed 2498 markers with less than 5 occurrences of either parental genotype. A major chromosome region filtered by this step is chrIII:177396..210630, which due to our selection of *MATa* segregants for pheromone treatment, caused a biased representation of the S96 parental genotype in markers near the

MATa locus. The genotype of the markers was evenly distributed across the 43 segregants: 50.7% inherited the S96 allele, and 49.3% inherited the HS959 allele.

Neighboring markers tend to segregate together due to strong linkage. These markers are redundant in QTL analysis and were consolidated into 2592 nonredundant marker sets spanning 9.15 Mb in physical distance, and 6995 cM in genetic distance. To compute the empirical false discovery rate, we randomly shuffled the columns of quantitative traits for 45 strains (therefore maintaining the correlation structure among different traits), and repeated the linkage analysis. To test for linkage between a cluster of quantitative traits and a genetic marker, we first computed the average trait for the cluster (trait with negative correlations to other traits in the same cluster was subtracted from others), and applied the same approach on linkage analysis.

Motif Analysis

We used TAMO package7 to search for 124 position weight matrices (PWM) of fungi transcription factors8 in sequences of variable Ste12 binding regions using a best match cutoff of 0.9 fold. To test whether the motif occurrence frequency is different between two sets of non-overlapping genomic regions, we compared motif PWM occurrence frequency of the set of interest and the control set, and then calculated Binomial p-value with edge effects taken into consideration7.

Trait clustering

We performed hierarchical clustering on 930 quantitative traits with Cluster3.0 9, using absolute centered correlation and complete linkage. The correlation threshold to divide clusters was determined using similar strategy as in Yvert et al.10. The highest pair-wise correlation between traits is 0.98, whereas the value obtained from a 100 permutation average is only 0.73. For clusters with three or more traits, we defined the cluster-wise correlation value as the smallest pair-wise correlation value in the cluster. The highest cluster-wise correlation of the 3-trait cluster is 0.97, while the maximal value obtained from 100 permutations is only 0.53. We therefore used 0.53 as cut-off (where we expect 0 false positive clusters with more than two traits), and obtained 121 mutually exclusive clusters.

Microarray experiment and data analysis

Microarray experiments were performed for 43 segregants and 2 parental strains in pheromone treated cells, and also for 2 parental strains and 1 segregant (SEG8) without pheromone treatment. Ambion Ribopure-yeast kit was used to extract total RNA from 10 ml yeast cultures. RNA integrity was examined by Agilent Bioanalyzer, and all samples had 260/280 ratios ~2.2, and 260/230 ratios >2.0. mRNA purification, cDNA synthesis and hybridization to Agilent Yeast Gene Expression 8x15K Array steps were performed at Microarray Core Facility at Cornell University.

To check alpha factor specific pattern of differential gene expression, quantile normalization was performed across 48 samples, data were log-transformed and results subjected to Principal Component Analysis (PCA). Alpha factor treated and untreated groups were clearly separated along the direction of first principal component. Gene set enrichment analysis (GSEA)11 was performed, and exhibited significant concordance between our lists of differentially expressed genes and Ste12 binding target genes.

To determine the correlation between gene expression level and Ste12 binding strength, we first updated 6233 annotated ORFs on the array with the newest SGD annotation, and obtained 6217 unique ORF targets. Pearson's correlation of log-transformed gene expression data and NormDiff scores of nearest target gene across 45 strains were calculated. The histogram is compared with background histogram calculated from 1,000 permutations of correlation between Ste12 binding and expression of a randomly chosen gene (Fig. 5).

Investigation of potential artefact

To ensure that the Ste12 binding variation observed from the ChIP-Seq data are not due to experimental artefacts, we performed several types of analyses. We first examined the reproducibility between biological replicates. We divided the entire nuclear genome into 500bp bins, counted the number of sequencing reads in each bin for each strain, and calculated pair-wise Pearson's correlation between replicates and different strains. The correlations between two biological replicates of the same strain are significantly higher than those obtained from two different strains (Fig. S2b, median correlation between replicate is 0.96 and that between different strains is 0.89; Wilcoxon rank-sum P < 0.001). We also repeated this analysis on 930 variable binding regions, and reached similar results (Fig. S2c, median correlation between replicates is 0.98 and that between different strains is 0.95; Wilcoxon rank-sum P < 0.001). This indicates that the variations we observed are mostly related to the differences in genetic background.

Second, we examined the effect of alignment bias. Since the ChIP-Seq data were aligned against the S288c reference genome, which is a perfect match to one parental strain (S96), but different from the other parental strain (YJM789), it is possible that some sequencing reads were not aligned. We expect this effect to be small because the alignment algorithm (Eland) allows up to two mismatches for each 30 bp read and YJM789 only has an average of 5 single nucleotide polymorphisms (SNPs) per kb and 0.5 insertion/deletions (Indels) with respect to the S288C genome12. To estimate the mappability due to polymorphisms between S288c and YJM789, we simulated the case that a 30bp read is randomly drawn from the genome, and used location information of 55958 SNP and Indel markers to determine whether the 30bp region contains more than one Indel or more than two SNPs, and thus not mappable. We randomly drew 1,000,000 reads, and recorded the percentage of non-mappable reads. Through simulation, we estimated that the probability of a read unable to match due to the presence of a polymorphism is as low as 2.05%. Nonetheless, we also aligned sequencing reads from each individual strain against the genomic DNA sequence of YJM78912. We then counted the number of reads that mapped to the 930 variable regions in S288c and their homologous regions in YJM789. The pairwise Pearson's correlation between different aligning schemes for the same strain were significantly higher than inter-strain correlations (Fig. S2d, median correlation between mapping schemes is 0.98 and that between different strains is 0.93; Wilcoxon rank-sum $P < 0.001$). This result indicates that the mappability of the reads with YJM789 genotype is not a major source of TF binding variation.

Finally, for each of the 930 variable binding regions, we inspected whether the two alignment schemes yielded a similar variation pattern across all 45 strains by calculating the correlations between different mapping schemes for each region across the 45 strains. In this scenario, the correlation between two mapping schemes of the ChIP sample should deviate from 1.0 due to the composite effect of Ste12 binding differences across strains and mappability issue caused by polymorphisms between S288c and YJM789; whereas the correlations between two mapping schemes of the Input sample deviate from 1.0 only due to the mappability issue. Therefore by comparing the correlation histograms of ChIP and Input samples, we can estimate to what extent mappability issue confounds the binding differences we are interested in. The correlations from Input samples are extremely high, (median r = 0.95), whereas the correlation from ChIP samples of different strains are lower but still reasonably high (median r = 0.70). Thus, we conclude that few binding sites differences are likely to be due to read alignment issues. Importantly, the significantly associated *cis*- and *trans*- traits exhibit high correlations from Input samples (median r = 0.93 for *cis*- traits and 0.94 for *trans*- traits), suggesting the robustness of our results.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We thank J. Gagneur for comments on data analysis. We thank the Cornell Microarray Facility for helping with the gene expression experiments. Research was funded by NIH grants to MS, HZ and LMS.

References

1. King MC, Wilson AC. Evolution at two levels in humans and chimpanzees. Science. 1975; 188(4184):107. [PubMed: 1090005]
2. Pennisi E. Searching for the genome's second code. Science. 2004; 306(5696):632. [PubMed: 15499005]
3. Morley M, et al. Genetic analysis of genome-wide variation in human gene expression. Nature. 2004; 430(7001):743. [PubMed: 15269782]
4. Schadt EE, et al. Genetics of gene expression surveyed in maize, mouse and man. Nature. 2003; 422(6929):297. [PubMed: 12646919]
5. Brem RB, Yvert G, Clinton R, Kruglyak L. Genetic dissection of transcriptional regulation in budding yeast. Science. 2002; 296(5568):752. [PubMed: 11923494]
6. Yvert G, et al. Trans-acting regulatory variation in Saccharomyces cerevisiae and the role of transcription factors. Nat Genet. 2003; 35(1):57. [PubMed: 12897782]
7. Ronald J, Brem RB, Whittle J, Kruglyak L. Local regulatory variation in Saccharomyces cerevisiae. PLoS Genet. 2005; 1(2):e25. [PubMed: 16121257]
8. Li Y, et al. Mapping determinants of gene expression plasticity by genetical genomics in C. elegans. PLoS Genet. 2006; 2(12):e222. [PubMed: 17196041]
9. Borneman AR, et al. Divergence of transcription factor binding sites across related yeast species. Science. 2007; 317(5839):815. [PubMed: 17690298]
10. Wilson MD, et al. Species-specific transcription in mice carrying human chromosome 21. Science. 2008; 322(5900):434. [PubMed: 18787134]
11. Roberts CJ, et al. Signaling and circuitry of multiple MAPK pathways revealed by a matrix of global gene expression profiles. Science. 2000; 287(5454):873. [PubMed: 10657304]

12. Mancera E, et al. High-resolution mapping of meiotic crossovers and non-crossovers in yeast. Nature. 2008; 454(7203):479. [PubMed: 18615017]
13. MacIsaac KD, et al. An improved map of conserved regulatory sites for Saccharomyces cerevisiae. BMC Bioinformatics. 2006; 7:113. [PubMed: 16522208]
14. Rockman MV, Kruglyak L. Genetics of global gene expression. Nat Rev Genet. 2006; 7(11):862. [PubMed: 17047685]
15. Pramila T, et al. Conserved homeodomain proteins interact with MADS box protein Mcm1 to restrict ECB-dependent transcription to the M/G1 phase of the cell cycle. Genes Dev. 2002; 16(23):3034. [PubMed: 12464633]
16. Hwang-Shum JJ, et al. Relative contributions of MCM1 and STE12 to transcriptional activation of a- and alpha-specific genes from Saccharomyces cerevisiae. Mol Gen Genet. 1991; 227(2):197. [PubMed: 1905781]
17. Jin R, Dobry CJ, McCown PJ, Kumar A. Large-scale analysis of yeast filamentous growth by systematic gene disruption and overexpression. Mol Biol Cell. 2008; 19(1):284. [PubMed: 17989363]
18. Kobayashi O, Suda H, Ohtani T, Sone H. Molecular cloning and analysis of the dominant flocculation gene FLO8 from Saccharomyces cerevisiae. Mol Gen Genet. 1996; 251(6):707. [PubMed: 8757402]
19. Smukalla S, et al. FLO1 is a variable green beard gene that drives biofilm-like cooperation in budding yeast. Cell. 2008; 135(4):726. [PubMed: 19013280]
20. Liu H, Styles CA, Fink GR. Saccharomyces cerevisiae S288C has a mutation in FLO8, a gene required for filamentous growth. Genetics. 1996; 144(3):967. [PubMed: 8913742]
21. Gimeno CJ, Fink GR. Induction of pseudohyphal growth by overexpression of PHD1, a Saccharomyces cerevisiae gene related to transcriptional regulators of fungal development. Mol Cell Biol. 1994; 14(3):2100. [PubMed: 8114741]
22. Rupp S, et al. MAP kinase and cAMP filamentation signaling pathways converge on the unusually large promoter of the yeast FLO11 gene. EMBO J. 1999; 18(5):1257. [PubMed: 10064592]
23. Borneman AR, et al. Target hub proteins serve as master regulators of development in yeast. Genes Dev. 2006; 20(4):435. [PubMed: 16449570]
24. Ni L, et al. Dynamic and complex transcription factor binding during an inducible response in yeast. Genes Dev. 2009; 23(11):1351. [PubMed: 19487574]

References

25. Lefrancois P, et al. Efficient yeast ChIP-Seq using multiplex short-read DNA sequencing. BMC Genomics. 2009; 10(1):37. [PubMed: 19159457]
26. Zhang Y, et al. Model-based analysis of ChIP-Seq (MACS). Genome Biol. 2008; 9(9):R137. [PubMed: 18798982]
27. Mancera E, et al. High-resolution mapping of meiotic crossovers and non-crossovers in yeast. Nature. 2008; 454(7203):479. [PubMed: 18615017]
28. Aparicio O, Geisberg JV, Struhl K. Chromatin immunoprecipitation for determining the association of proteins with specific genomic sequences in vivo. Curr Protoc Cell Biol. 2004 **Chapter 17**, Unit 17 7.
29. Robertson G, et al. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. Nat Methods. 2007; 4(8):651. [PubMed: 17558387]
30. Rozowsky J, et al. PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. Nat Biotechnol. 2009; 27(1):66. [PubMed: 19122651]
31. Gordon DB, Nekludova L, McCallum S, Fraenkel E. TAMO: a flexible, object-oriented framework for analyzing transcriptional regulation using DNA-sequence motifs. Bioinformatics. 2005; 21(14):3164. [PubMed: 15905282]
32. MacIsaac KD, et al. An improved map of conserved regulatory sites for Saccharomyces cerevisiae. BMC Bioinformatics. 2006; 7:113. [PubMed: 16522208]

33. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A. 1998; 95(25):14863. [PubMed: 9843981]

34. Yvert G, et al. Trans-acting regulatory variation in Saccharomyces cerevisiae and the role of transcription factors. Nat Genet. 2003; 35(1):57. [PubMed: 12897782]

35. Subramanian A, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005; 102(43):15545. [PubMed: 16199517]

36. Wei W, et al. Genome sequencing and comparative analysis of Saccharomyces cerevisiae strain YJM789. Proc Natl Acad Sci U S A. 2007; 104(31):12825. [PubMed: 17652520]

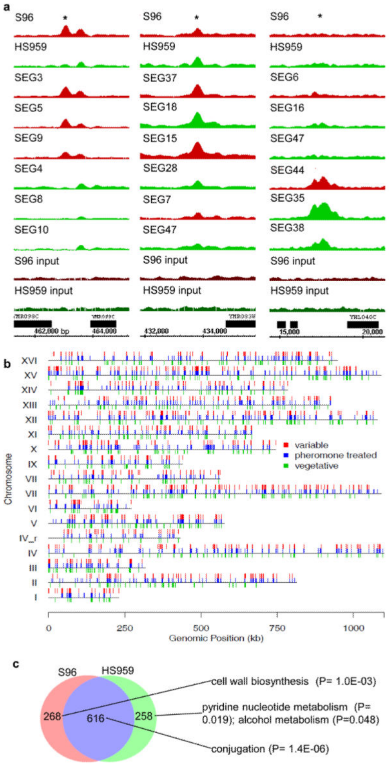

Figure 1. Extensive Ste12 binding variations among S288c × JM789 derivatives

a, ChIP-Seq signal tracks showing Ste12 binding sites that segregate in a Mendelian (Left) or transgressive fashion (Middle and Right). The color indicates genotype background in the depicted regions: red (S96), green (HS959). Asterisks indicate peaks of interest. **b,** Overall genomic positions of Ste12 binding regions in S96 or HS959 under vegetative growth condition (green), pheromone treatment (blue), and most variable binding regions across segregants (red). **c,** Overlap of target genes (from 6644 annotated genes) between parent strains with pheromone treatment. Enriched GO categories are listed.

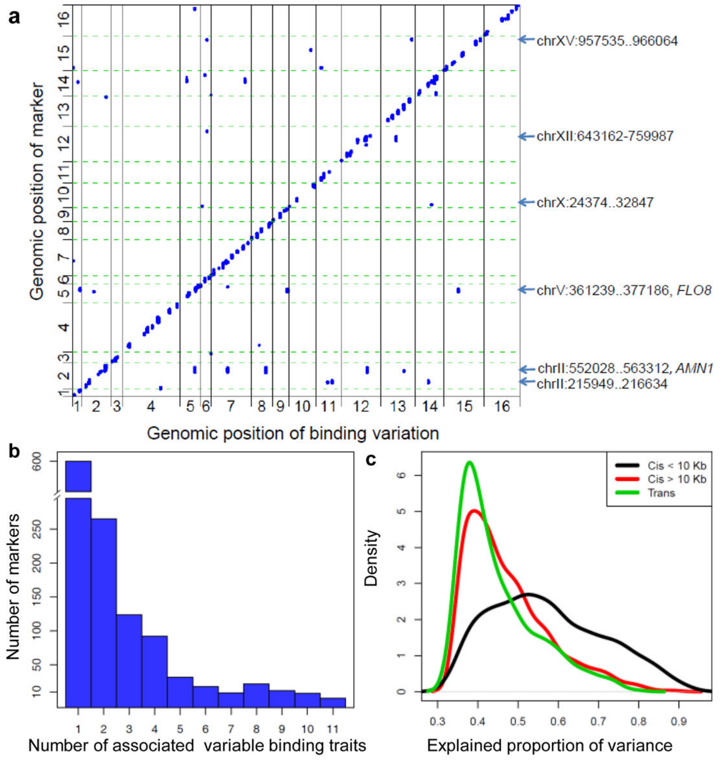

Figure 2. Whole-genome linkage analysis of variable Ste12 binding traits

a, Chromosomal position of significantly associated binding traits (X axis) relative to markers (Y axis) passing threshold (FDR = 0.01). See Fig. S4 for results with different thresholds. Trans-QTL regions validated by the clustering method are shown on the right. Two experimentally validated causative genes (*AMN1* and *FLO8*) underlying the trans-QTLs are also shown. **b,** Histogram of markers significantly associated with multiple binding traits. **c,** Effect size (explained variance in quantitative traits) of cis-QTLs < 10kb to the trait (black), cis-QTLs > 10kb to the trait (red), and trans-QTLs (green).

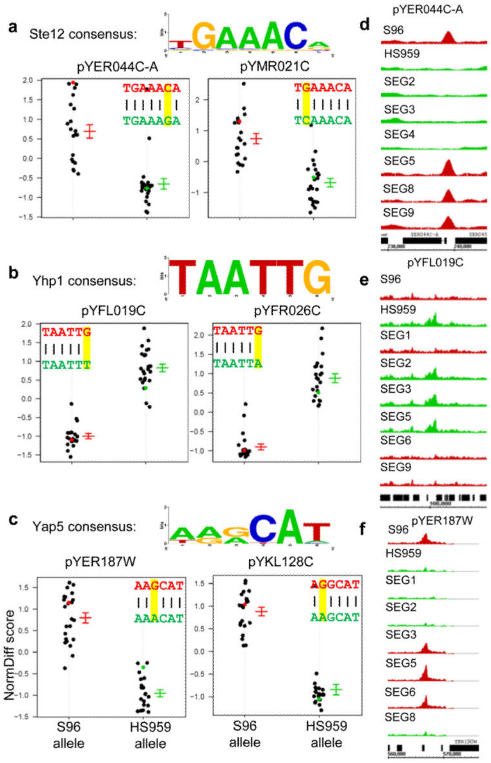

Figure 3. Motif analysis of cis-variable binding regions

a-c, Ste12 binding signals (NormDiff scores) against the genotypes of Ste12 motif, Yhp1 motif and Yap5 motif in two target loci. Each dot plot shows strains with S96 inheritance (red dot), strains with HS959 inheritance (green dot). The mean and standard error of each group, and Ste12 consensus sequence are also shown (Red, S96; green, HS959). **d-f,** ChIP-Seq signal tracks of the depicted regions in dot plots a-c. The color in each track indicates genotype background in the depicted regions: red (S96), green (HS959). Additional examples are in Fig. S5.

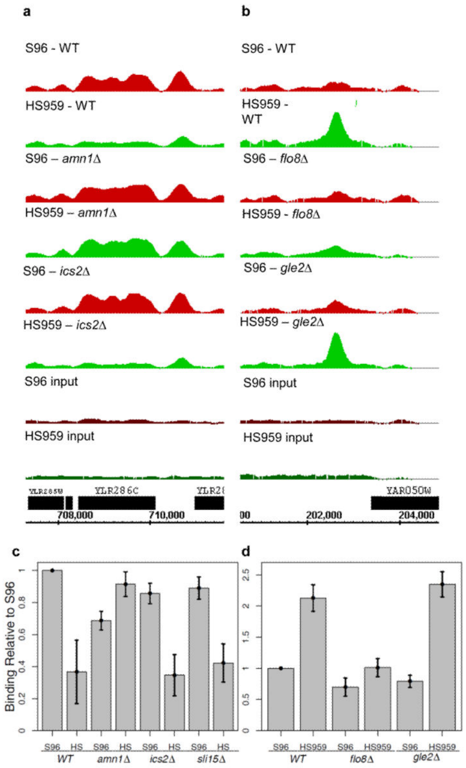

Figure 4. Validation of two causative quantitative trait genes

a, ChIP-Seq signal tracks showing Ste12 binding of wild-type, *amn1*Δ, and *ics2*Δ parent strains, corresponding to variable binding trait B12S708456E710089. **b,** ChIP-Seq signal tracks showing Ste12 binding of wild-type, *flo8*Δ, and *gle2*Δ parent strains, corresponding to variable binding traits B1S202520E202741. *ics2*Δ and *gle2*Δ strains have no effect on binding. **c, d,** Comparison of Ste12 binding across 10 and 5 associated variable binding regions in WT and knockout strains rspectively. Error bar represents s.e.m. Additional results are in Fig. S7.

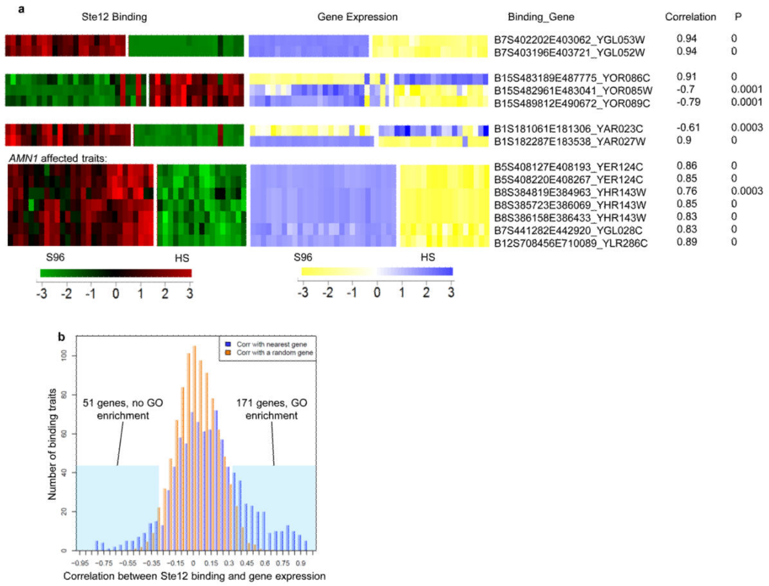

Figure 5. Ste12 binding significantly correlates with downstream gene expression

a, Examples of high correlation between binding variation and gene expression variation. Columns in the heat maps are ordered by the genotype of markers with highest association to the Ste12 binding traits. 10 additional clusters are shown in Fig. S8. **b,** Histogram (blue) and background histogram (yellow) of correlation coefficients between Ste12 binding and expression of nearest gene. Regions with absolute Pearson's correlation |r| > 0.335 are shaded.