nature genetics

Trans-acting regulatory variation in *Saccharomyces cerevisiae* and the role of transcription factors

Gaël Yvert[1,4], Rachel B Brem[1], Jacqueline Whittle[1], Joshua M Akey[1], Eric Foss[1], Erin N Smith[1,2], Rachel Mackelprang[1,3] & Leonid Kruglyak[1]

Natural genetic variation can cause significant differences in gene expression, but little is known about the polymorphisms that affect gene regulation. We analyzed regulatory variation in a cross between laboratory and wild strains of *Saccharomyces cerevisiae*. Clustering and linkage analysis defined groups of coregulated genes and the loci involved in their regulation. Most expression differences mapped to *trans*-acting loci. Positional cloning and functional assays showed that polymorphisms in *GPA1* and *AMN1* affect expression of genes involved in pheromone response and daughter cell separation, respectively. We also asked whether particular classes of genes were more likely to contain *trans*-regulatory polymorphisms. Notably, transcription factors showed no enrichment, and *trans*-regulatory variation seems to be broadly dispersed across classes of genes with different molecular functions.

Several studies have shown that natural genetic variation can cause significant differences in gene expression[1–7], suggesting that phenotypic diversity results not only from coding variation that affects protein function but also from regulatory variation that affects gene expression. Little is known about the mechanisms by which natural polymorphisms affect gene regulation, however. One can distinguish two classes of regulatory variations. First, a polymorphism can reside at the locus of the gene it regulates, affecting expression directly in *cis*, by altering functional motifs in the promoter region or the stability of the mRNA, or indirectly, by modifying the activity of the gene product and causing expression changes through feedback control. *cis*-regulatory variation has been studied by several approaches[3–5,8]. Alternatively, a polymorphism in one gene can affect expression of other genes. Such *trans*-regulatory variation is harder to study without prior information about regulators and their targets. The use of linkage analysis to map expression differences allows examination of both *cis*- and *trans*-acting regulatory variation. With this approach, we and others have shown that *trans*-acting loci seem to be responsible for most differences in gene expression between strains of yeast[2] and mice[7].

What types of genes contain *trans*-acting regulatory variation? A natural hypothesis is that polymorphisms in transcription factors affect expression levels of their targets, but regulatory variation could also reside elsewhere in regulatory networks. In this paper we used several approaches to address this question. We positionally cloned two *trans*-acting regulators and demonstrated their effects on expression of groups of genes with related functions. We also used a combination of clustering, linkage analysis and bioinformatics to examine the representation of classes of genes with different molecular functions among *trans*-acting loci.

RESULTS

Clustering of genes based on segregating variation

Correlation in expression among genes across a set of perturbations is now commonly used to define groups of genes with putative common function[9,10]. The perturbations used include different growth conditions[11], engineered mutations[12] and chronologic stages of a cellular process[13,14]. Because many differences in gene expression have been shown to exist between genetically diverse individuals or strains[1–4,6,7], the progeny of a mating between two genetically diverse parents provide a set of perturbations that reflects many different combinations of parental regulatory polymorphisms generated by random genetic segregation. Measuring transcript levels in the progeny thus allows simultaneous elucidation of the effects of many more factors affecting gene expression than is possible with conventional single-factor approaches[15,16].

To investigate the use of genetic variation for defining groups of co-regulated genes, we measured expression of all yeast genes in laboratory (BY) and wild (RM) strains of *S. cerevisiae* and in 86 segregants from a cross between them and used hierarchical clustering to group the genes by similarity in expression profile. To focus only on true coordinate regulation rather than chance correlation, we defined a statistically significant degree of correlation between genes by permutation testing and restricted further analysis to clusters in

[1]Howard Hughes Medical Institute and Division of Human Biology, Fred Hutchinson Cancer Research Center, Mailstop D4-100, 1100 Fairview Avenue North, Seattle, Washington 98103 USA. [2]Molecular and Cellular Biology Program and [3]Department of Genome Sciences, University of Washington, Seattle, Washington, USA. [4]Present address: Centre G. Durand, UMR-CNRS 5504, INSA, 135 Ave de Rangueil 31077 Toulouse cedex 04, France. Correspondence should be addressed to L.K. (leonid@fhcrc.org).

Published online 3 August 2003; doi:10.1038/ng1222

which all pairwise correlations exceeded 0.725. At this correlation, fewer than three clusters of two genes and no clusters of more than two genes are expected by chance. We found 593 clusters of at least two genes and 205 clusters of more than two genes in the actual data, comprising 1,861 genes. Because of the strict criterion used to define the clusters, a set of genes that are in fact coregulated could be split across several clusters. We observed that 76% of the clusters contained genes from different chromosomes, showing that most correlations were not merely due to coordinate regulation of genes in the same chromosomal region, for example, through shared *cis*-acting regulatory elements or chromatin effects. Annotations of genes in each cluster (downloaded from the Saccharomyces Genome Database) frequently showed their involvement in similar processes. We observed many correlations between well annotated genes and uncharacterized open reading frames (ORFs), allowing functional predictions (**Supplementary Table 1** online). Whereas most previous microarray studies have used specific treatments to group genes into clusters, our results reflect coregulation during standard exponential growth and therefore provide a useful exploratory set of genetic perturbations affecting different cellular processes.

Most expression differences map to *trans*-acting loci

Though successful in grouping genes with correlated expression, most expression profiling studies do not allow identification of the regulators responsible for coordinate expression of the clusters. In contrast, the use of segregating variation as the source of perturbations allows the regulators to be mapped through linkage analysis. We genotyped the segregants as previously described using oligonucleotide microarrays[2,17] and carried out linkage analysis with a map of 3,114 genetic markers. For each segregant, the mean expression value of the genes in a given cluster was treated as a quantitative phenotype. A total of 304 clusters containing 1,011 genes showed linkage to at least one position in the genome at $P < 3.4 \times 10^{-5}$, a significance level at which 10 clusters were expected to link by chance. At more stringent $P < 5 \times 10^{-7}$, 153 clusters containing 532 genes showed linkage (no clusters were expected to link by chance). Examples of clusters that link are shown in **Figure 1**. Because only one third of all genes fell into clusters, we also tested for linkage of expression levels of all genes individually. We mapped 2,294 expression levels at $P < 3.4 \times 10^{-5}$ (100 were expected to link by chance) and 992 expression levels at $P < 5 \times 10^{-7}$ (only one was expected to link by chance). For genes falling into clusters, the results obtained when mapping them as cluster members or individually were very similar. As expected, most linkages previously described[2] were confirmed in the new data

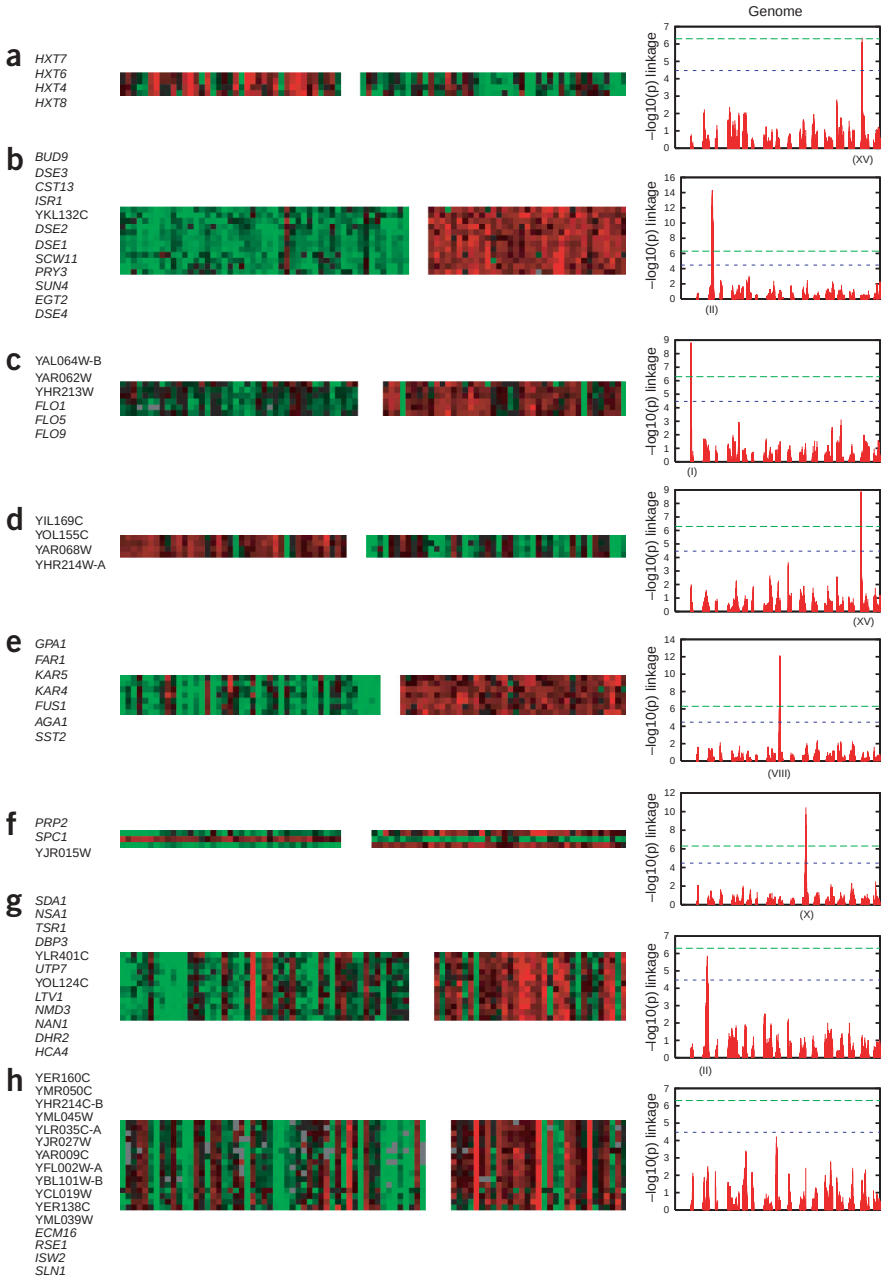

Figure 1 Genetic mapping of cluster regulators. Eight examples of clusters of genes with correlated expression are shown: the genes have a role in hexose transport (**a**), daughter-cell specificity (**b**), flocculation (**c**), unknown (**d**,**h**), pheromone response (**e**), RNA and signal peptide processing (**f**) and ribosomal RNA processing (**g**). For each cluster, rows represent the genes listed on the left and columns represent the segregants. The clusters are divided into two groups according to the genotypes of the segregants at the marker with highest linkage score (left, RM inheritance; right, BY inheritance). Over- and underexpression are indicated by red and green, respectively. The average relative change between the left and right blocks was 1.8 (**a**), 4.0 (**b**), 1.2 (**c**), 2.4 (**d**), 2.2 (**e**), 1.8 (**f**), 1.3 (**g**) and 1.5 (**h**). The linkage profile of expression of each cluster across the genome is shown on the far right; dashed lines correspond to 0.1% (green) and 4.3% (blue) false discovery rates. Chromosomes on which significant linkage was found are indicated in parentheses below the corresponding linkage peaks. The bottom cluster illustrates an example where no linkage was found despite a strong correlation in expression of many genes; the lack of linkage probably reflects polygenic control of expression of this cluster.

set, including all eight groups of genes linking to loci with widespread transcriptional effects. The larger cross allowed definition of five new such groups and addition of many genes to the previously defined groups, with a total of 1,265 expression levels falling into the 13 groups (**Supplementary Table 2** online).

Of the 2,294 genes whose expression linked somewhere in the genome, only 578 (25%) were located in the corresponding linked region (within 10 kb of a linked marker). Similarly, only 57 of the 304 clusters showing linkage contained a gene in the region of linkage. The fact that 75% of genes and 80% of clusters did not show self-linkage shows that most genetic variation in expression is due to *trans*-acting factors. Although our mapping resolution for each linkage is somewhat limited, we estimate that 100–200 distinct *trans*-acting loci affect the 1,716 expression levels that link in *trans*.

Missense mutation in *GPA1* affects pheromone response genes

One cluster of seven genes involved in pheromone response (**Fig. 1e**) linked to markers near *GPA1*, which encodes a G-protein α subunit coupled to pheromone receptors. We sequenced the *GPA1* allele from strain RM and found that it encoded two amino-acid changes, one of which (I469S) is non-conservative when compared with the BY strain sequence. This polymorphism is located in the C-terminal region of five amino acids known to be essential for binding of Gpa1 to the pheromone receptors Ste2 and Ste3 (refs. 18,19). Previous studies have shown that point mutations near position 469 (K467P and K468P) cause defects in mating and pheromone response[18], and mutants lacking the last 22 residues of Gpa1 have growth defects, suggesting constitutive activation of the pheromone response pathway[18]. Sequence alignment with other yeast species[20,21] showed that

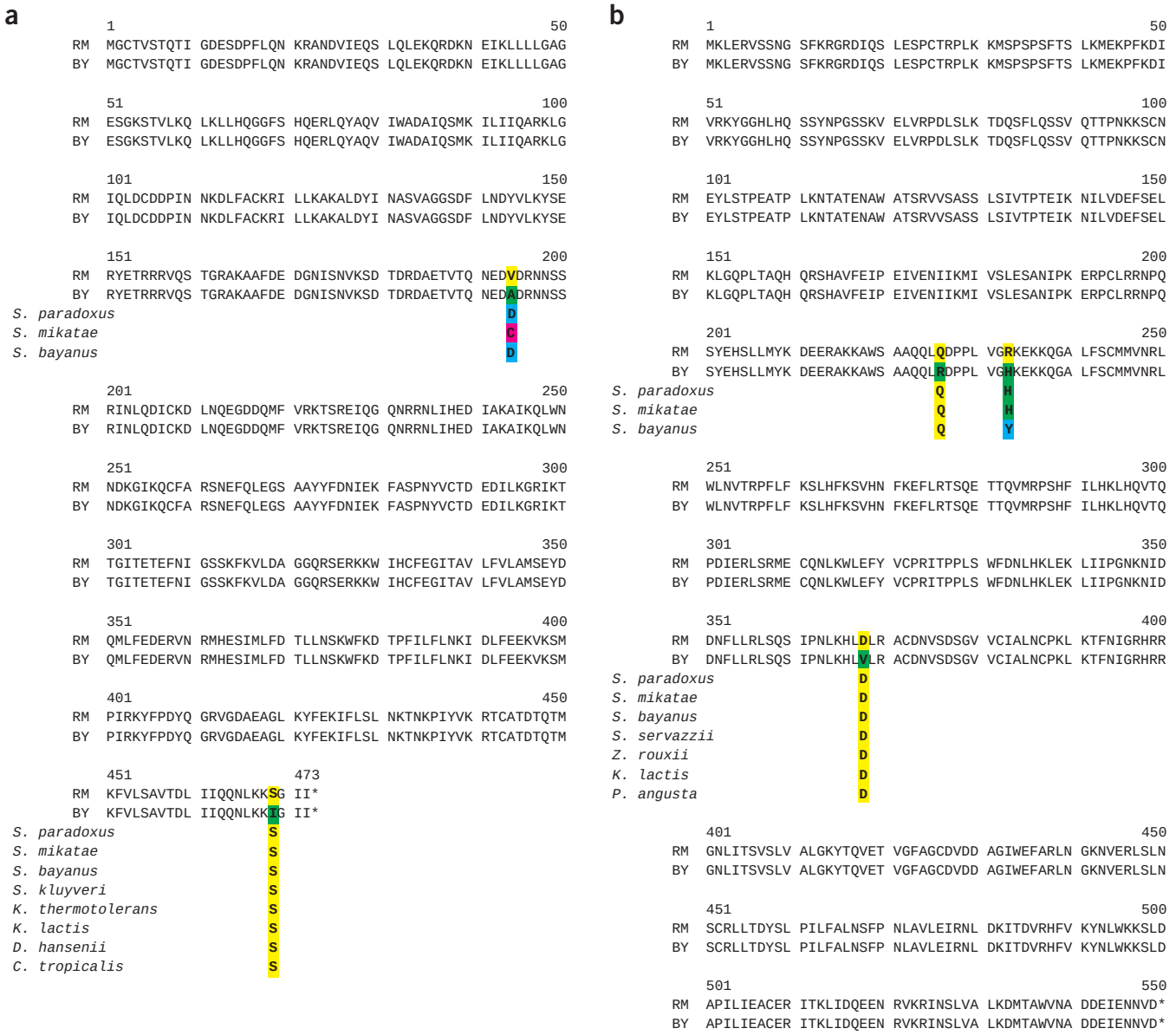

Figure 2 Coding sequence polymorphisms at *trans*-regulatory loci. (**a**) Polymorphisms in Gpa1. (**b**) Polymorphisms in Amn1. The amino-acid sequence in other yeast species is shown under the polymorphic residues when available[20,21]: *Saccharomyces paradoxus*, *Saccharomyces mikatae*, *Saccharomyces bayanus*, *Saccharomyces kluyveri*, *Kluyveromyces thermotolerans*, *Kluyveromyces lactis*, *Debaryomyces hansenii*, *Candida tropicalis*, *Saccharomyces servazzii*, *Zygosaccharomyces rouxii* and *Pichia angusta*.

Table 1 Expression of genes in cluster linking to *GPA1*

Gene	ORF	BY:RM[a]	Segregants[b]	Replacement[c]
GPA1	YHR005C	1.1	1.4	1.4
FAR1	YJL157C	1.1	1.7	1.8
KAR5	YMR065W	1.4	1.4	1.4
KAR4	YCL055W	1.5	1.6	1.6
FUS1	YCL027W	2.5	3.4	3.5
AGA1	YNR044W	6.2	3.7	3.8
SST2	YLR452C	2.4	2.2	1.9

[a]Ratio of expression in BY to expression in RM, averaged across 12 cultures. [b]Ratio of average expression in 40 segregants inheriting the BY allele of *GPA1* to average expression in 46 segregants inheriting the RM allele. [c]Ratio of expression in BY to expression in BY with the I469S substitution, averaged across two cultures.

Table 2 Expression of genes in cluster linking to *AMN1*

Gene	ORF	BY:RM[a]	Segregants[b]	RM *amn1*Δ:RM[c]
SCW11	YGL028C	18.0	12.2	16.2
DSE1	YER124C	21.4	18.0	32.9
DSE2	YHR143W	48.2	26.0	40.6
	YKL132C	7.7	3.2	1.1
ISR1	YPR106W	1.6	1.8	1.8
AMN1	YBR158W	2.2	2.2	0.08
PRY3	YJL078C	2.5	3.3	6.3
DSE4	YNR067C	2.6	2.8	2.9
EGT2	YNL327W	1.6	1.9	1.9
SUN4	YNL066W	4.5	3.3	4.3
DSE3	YOR264W	1.4	2.3	2.0
BUD9	YGR041W	1.8	2.1	2.3

[a]Ratio of expression in BY to expression in RM, averaged across 12 cultures. [b]Ratio of average expression in 45 segregants inheriting the BY allele of *AMN1* to average expression in 41 segregants inheriting the RM allele. [c]Ratio of expression in RM with the *AMN1* ORF deleted to expression in RM, averaged across two independent knock-out strains.

the serine is conserved in species as distant as *Candida tropicalis* (**Fig. 2a**). This observation suggested that Ser469 is important for Gpa1 function, and that the S469I polymorphism of the sequenced strain S288C (isogenic to BY) impairs Gpa1 function, perhaps by affecting binding to Ste2 and Ste3 pheromone receptors, leading to residual activation of the pathway in the absence of pheromones. To test this hypothesis, we replaced the isoleucine in BY with a serine and measured the expression of genes linking to *GPA1* relative to the original BY strain. As predicted, the expression of these pheromone response genes decreased to levels similar to those observed in segregants inheriting the RM allele (**Table 1**), confirming the functional role of Ser469 in pheromone signaling.

AMN1 affects gene expression and daughter cell separation

Another cluster of 12 genes, 8 of which were previously reported to be specifically expressed in daughter cells during budding[22], linked to a locus on chromosome II. Segregants that inherited this locus from the wild RM strain showed lower expression of these genes than did segregants that inherited the locus from the laboratory BY strain (**Fig. 1b**). We previously proposed that a polymorphism in *AMN1* (ref. 23), also called *CST13*, was responsible for variation in expression levels of these genes[2]. We developed additional markers in this region and narrowed the linkage to a region of 1.5 kb that included sequences upstream of *AMN1* and part of the *AMN1* ORF but did not include any other known ORFs. We found several polymorphisms in this region (**Fig. 2b**). Deletion of *AMN1* in RM caused expression of genes in the cluster to increase to the levels seen in BY (**Table 2**). In contrast, a strain isogenic to BY with a deletion of *AMN1* (ref. 24) showed no difference from BY in expression of daughter cell–specific genes (data not shown), suggesting that BY carries a loss-of-function mutation in *AMN1*.

We independently observed a phenotypic difference in daughter-cell separation between the parent strains, with RM but not BY showing clumpy growth (**Fig. 3a,b**). Because some of the daughter cell–specific genes act to effect daughter-cell separation after budding[22], and their expression is differentially regulated by the RM and BY alleles of *AMN1*, we predicted that clumpiness would map to *AMN1*. We scored 40 segregants for clumpiness and found linkage of clumpiness to *AMN1* with a lod score of 8, which confirmed the prediction. Furthermore, deletion of *AMN1* in RM eliminated clumpiness (**Fig. 3c**). A different wild strain also showed clumpy growth and low expression of the daughter cell–specific genes relative to BY (data not shown), further arguing that the BY laboratory strain carries a mutated copy of *AMN1* that leads to upregulation of the other daughter cell–specific genes and non-clumpy growth. Sequence comparison showed that two residues were conserved in other yeast species and RM but altered in BY (**Fig. 2b**); one of these, D368V, changes an acidic residue to a hydrophobic residue and might therefore be responsible for the functional differences between the two alleles.

Amn1 was recently shown to be a negative regulator of the mitotic exit network (MEN) in daughter cells[23]. MEN leads to activation of the transcription factor Ace2 (ref. 23), which is thought to drive the transcription of genes with daughter-specific expression[22]. Consistent with this model, we observe that deletion of *AMN1* in RM boosts expression of the daughter-specific genes. Notably, previous work on the function of *AMN1* was carried out in lab strain W303 (ref. 23), which carries the same allele of *AMN1*

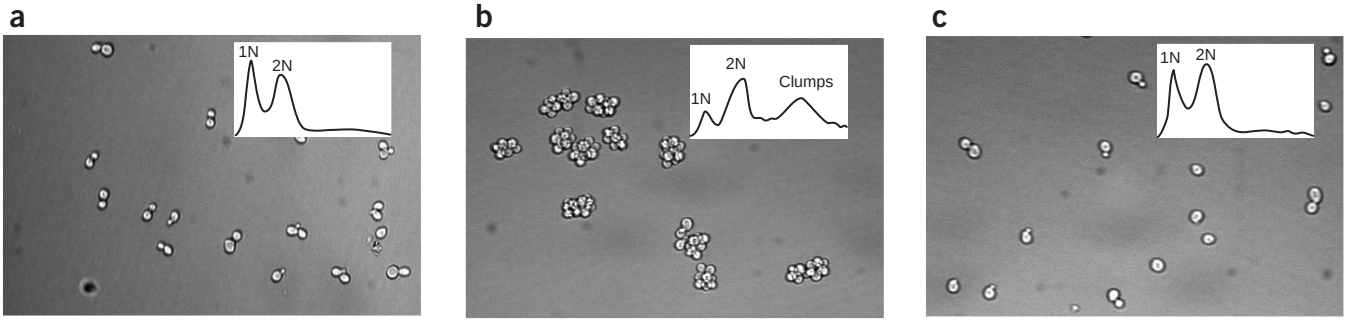

Figure 3 Effect of *AMN1* on clumpiness. Clumps of cells stained with DNA-binding dye generate fluorescence reflecting a particle (clump of cells) with DNA content greater than 2N. (**a**) BY. (**b**) RM. (**c**) RM *amn1*Δ.

as BY (data not shown). In our experiments, this allele behaves identically to a deletion with regard to expression of the daughter-specific genes and clumpiness, suggesting a possible separation of function between the activities of Amn1 in daughter-cell separation and cell cycle control.

Most *trans* variation does not map to transcription factors

Because transcription factors are natural candidates for harboring *trans*-acting regulatory variation, we investigated whether *trans*-acting loci preferentially mapped to transcription factors through several approaches. First, we used genome-wide studies of binding sites of 113 known or predicted transcription factors[25,26] to examine which transcription factors were probable regulators of expression clusters and whether these factors could explain the observed linkages. Thirty-one clusters showed statistically significant enrichment for targets of one or more factors (**Table 3**), with the number of transcription factors per cluster ranging from one to eight. As expected, Gcn4, Hap4, Ace2, Hir1 and Hir2, and Fhl1 and Rap1 had a high number of targets in clusters of genes involved in amino-acid synthesis[27], respiration[28], daughter-cell specificity[22], nucleosomes[29], and ribosome assembly[25], respectively. One might expect that regulatory polymorphisms responsible for the clusters would reside in the sequence of these transcription factors. In fact, although 15 of the 31 clusters showed significant linkage somewhere in the genome, only one linked to within 20 kb of a transcription factor whose targets were enriched in the cluster: three uncharacterized ORFs formed a cluster and were targets of Mcm1 and a predicted transcription factor encoded by YJL206C (**Table 3**), and this cluster linked to the region containing YJL206C.

This result suggested that most regulatory genetic variation does not reside in transcription factors. To confirm this, we considered all 992 genes showing highly significant linkage ($P < 5 \times 10^{-7}$). There were 807 instances of a transcription factor binding to the upstream region of a gene in this set, but we found only 7 cases (not different from the number expected by chance) in which a gene was a target of a transcription factor located at the locus of linkage. Therefore, few genetic differences in gene expression are due to polymorphisms in transcription factors that have been shown to bind to the genes of interest.

Because the genome-wide binding studies[25,26] excluded some transcription factors and, by their estimate, identified roughly two-thirds of the binding sites of the factors that were examined, some regulatory variation in transcription factors could have been missed by reliance on this data set. For example, 114 expression levels were linked to the gene encoding the transcription factor Hap1, which contains a functional polymorphism between the two parent strains[2], but this factor was not included in the genome-wide binding studies[25,26]. To examine more generally the extent to which regulatory variation resides in transcription factors, we considered a more complete list of 123 transcription factors[30] and asked whether *trans* linkages of expression levels were in general more likely to fall near transcription factors (without regard to their binding sites) than near other genes. The peaks of the 1,716 *trans* linkages (at $P < 3.4 \times 10^{-5}$) mapped to 394 separate markers, which correspond to 100–200 distinct *trans*-acting loci when linkage peaks near each other are grouped together. We found that 20% of the 123 transcription factors were located within 5 kb of a *trans*-linked marker, and 35% were within 10 kb, which is not different from what is expected by chance (22% within 5 kb and 37% within 10 kb). We therefore conclude that the set of *trans*-regulatory loci is not enriched in transcription factors.

Table 3 Clusters enriched for targets of specific transcription factors

Cluster annotation	Factor	Target genes/ total genes	*P* value[a]
Telomere maintenance	Gal4	8/33	$<1 \times 10^{-10}$
	Gat3	23/33	$<1 \times 10^{-10}$
	Msn4	10/33	$<1 \times 10^{-10}$
	Pdr1	13/33	$<1 \times 10^{-10}$
	Rgm1	10/33	$<1 \times 10^{-10}$
	Swi5	7/33	6.4×10^{-7}
	Rap1	8/33	9.0×10^{-6}
	Yap5	27/33	$<1 \times 10^{-10}$
COS genes	Gat3	5/10	2.9×10^{-8}
	Rgm1	5/10	5.0×10^{-10}
	Yap5	6/10	4.4×10^{-9}
RNA processing	Dig1	4/16	2.3×10^{-6}
Ribosomal proteins (10 clusters)	Fhl1	19/26	$<1 \times 10^{-10}$
	Rap1	13/26	$<1 \times 10^{-10}$
	Fhl1	9/13	$<1 \times 10^{-10}$
	Rap1	8/13	1.5×10^{-9}
	Fhl1	4/5	4.5×10^{-6}
	Rap1	4/5	5.7×10^{-6}
	Fhl1	5/6	1.7×10^{-7}
	Fhl1	4/4	9.2×10^{-7}
	Rap1	4/4	1.2×10^{-6}
	Fhl1	5/5	2.8×10^{-8}
	Rap1	4/5	5.7×10^{-6}
	Fhl1	9/11	$<1 \times 10^{-10}$
	Rap1	7/11	1.2×10^{-8}
	Fhl1	12/15	$<1 \times 10^{-10}$
	Rap1	7/15	2.1×10^{-7}
	Fhl1	4/5	4.5×10^{-6}
	Fhl1	6/8	2.3×10^{-8}
Daughter specificity	Ace2	7/12	$<1 \times 10^{-10}$
	Swi5	5/12	6.4×10^{-7}
Mating (2 clusters)	Mcm1	5/17	2.9×10^{-6}
	Gat3	4/7	3.9×10^{-7}
Leucine biosynthesis	Leu3	7/12	$<1 \times 10^{-10}$
Histones	Hir1	4/6	7.8×10^{-9}
	Hir2	4/6	1.9×10^{-9}
Unknown	Mcm1	3/3	2.8×10^{-6}
	YJL206C[b]	3/3	8.0×10^{-8}
Amino-acid biosynthesis	Gcn4	6/14	1.2×10^{-8}
ATP synthesis (4 clusters)	Hap4	7/9	$<1 \times 10^{-10}$
	Hap4	6/10	4.0×10^{-10}
	Hap4	3/4	5.3×10^{-6}
	Hap4	4/5	7.3×10^{-8}
Arginine biosynthesis	Arg80	3/4	8.0×10^{-8}
	Arg81	4/4	1.0×10^{-10}
	Gcn4	3/4	8.2×10^{-6}
Pheromone response (2 clusters)	Dig1	4/7	4.7×10^{-8}
	Ste12	4/7	2.2×10^{-7}
	Dig1	3/4	8.9×10^{-7}
	Ste12	3/4	2.8×10^{-6}
Unknown	Mcm1	3/3	2.8×10^{-6}
	YJL206C	3/3	8.0×10^{-8}
Glycine metabolism	Bas1	3/3	3.2×10^{-7}
FAS genes	Ino2	2/3	6.2×10^{-6}
Unknown	Hal9	2/2	6.5×10^{-6}

[a]Nominal *P* value reflecting deviation from binomial distribution. [b]Only case in which expression of the cluster links to the transcription factor.

Trans-acting variation distributed across molecular functions

To test whether the set of *trans*-regulatory loci is enriched for any class of genes, we next carried out the same analysis for a wide range of molecular functions obtained from the Gene Ontology database[31] (**Table 4**). None of the 31 molecular function categories analyzed showed statistically significant overrepresentation in the set of *trans*-regulatory loci. Only five categories (guanyl nucleotide binding, RNA binding, oxidoreductase, hydrolase and transcription cofactor) had an excess of 5 or more genes near *trans*-regulatory loci, no single category had an excess of more than 10, and, taken together, these excesses totaled 35 genes, which is considerably smaller than the estimated number of *trans*-regulatory loci (100–200). In addition, the same analysis was done with 55 Gene Ontology biological process categories, and again no evidence for enrichment was found (data not shown). These results strongly suggest that genetic variation is widely distributed across regulatory networks.

Table 4 Gene Ontology analysis

Gene Ontology molecular function	Number of genes[a]	Observed[b]	Expected[c]	χ^2 statistic[d]
Binding				
Cytoskeleton protein binding	59	20	21.1	0.06
ATP binding	160	58	56.3	0.05
Guanyl nucleotide binding	67	30	24.4	1.29
RNA binding	334	127	120.1	0.39
Chaperone	68	23	24.5	0.09
Enzyme				
Transferase				
Transferase, 1 carbon	63	22	23.1	0.05
Transferase, phosphorus	298	105	107.6	0.06
Transferase, glycosyl	87	33	30.2	0.27
Kinase	216	74	77.9	0.20
Nuclease	90	28	31.3	0.34
Ligase	112	38	39.7	0.07
Isomerase	43	13	14.4	0.14
Oxidoreductase	227	88	80.5	0.70
Hydrolase				
Hydrolase, acting on ester bonds	206	66	72.4	0.56
Hydrolase, acting on acid anhydrides	217	86	76.0	1.30
Hydrolase, acting on glycosyl bonds	59	18	14.1	1.05
Peptidase	103	39	37.1	0.10
Enzyme regulator	101	35	36.5	0.06
Signal transducer				
Receptor	31	12	11.0	0.09
Receptor signaling protein	78	30	27.9	0.15
Structural molecule	335	116	116.9	0.01
Transporter				
Protein transporter	39	8	14.1	2.61
Amino acid transporter	26	11	9.2	0.36
Carbohydrate transporter	31	6	10.1	1.65
Ion transporter	99	35	34.9	0.00
Carrier	164	59	56.5	0.11
Translational regulator	59	20	20.9	0.04
Triplet codon adaptor	300	101	109.1	0.60
Transcriptional regulator				
Transcription cofactor	35	16	10.6	2.78
Transcription factor	53	19	18.5	0.01
RNA polymerase II transcription factor	116	40	42.1	0.11

[a]Number of yeast genes in this Gene Ontology molecular function category. [b]Number of genes in category observed near *trans*-linked markers. [c]Number of genes in category expected to fall near *trans*-linked markers by chance. [d]χ^2 statistic for comparison of observed and expected numbers.

DISCUSSION

We found that more than one third of all yeast transcript levels segregate and show linkage to at least one locus in a cross between laboratory and wild strains of *S. cerevisiae*. Most of these linkages involve *trans*-acting factors (variations in genes affecting expression of other genes). We used positional cloning to identify *trans*-acting regulatory polymorphisms in two genes, *GPA1* and *AMN1*, and confirmed that these polymorphisms affect transcript levels of genes involved in pheromone response and daughter-cell separation, respectively, as well as the corresponding cellular phenotypes. It is notable that we were able to identify variation in these and other pathways in cells grown under a single set of conditions. Both polymorphisms involve missense changes in the laboratory strain relative to the wild strain and other yeast species, and both seem to impair the function of the protein products, perhaps reflecting adaptation to a new selective regime in the laboratory environment[32]. This observation raises the possibility that some conclusions drawn from studies in laboratory strains may not be generally applicable and underscores the importance of studying wild strains.

Segregating genetic variation was used to define clusters of significantly coregulated genes, with expression of half of the clusters showing linkage. We found a number of clusters that were significantly enriched in targets of particular transcription factors, explaining their coordinate regulation, but few genes linked to a transcription factor of which they were a target, and there was no tendency for linkages to fall near transcription factors. More generally, we found that no protein class defined by either molecular function or biological process in the Gene Ontology database was enriched for *trans*-regulatory variation, suggesting that such variation is broadly dispersed across regulatory networks.

Why did we find no enrichment for linkage of gene expression levels to transcription factors? One possibility is that transcription factors contain few functional polymorphisms owing to strong selective pressure against them, consistent with the recent observation of evolutionarily stable expression patterns of transcription factors in *Drosophila*[33]. A contrasting explanation is that expression of most genes is regulated by combinatorial binding of several transcription factors, and a polymorphism that affects the binding of one factor is usually insufficient to produce a detectable change in gene expression. Finally, transcriptional regulatory networks contain many components, and transcription factors may simply comprise only a small fraction of the target size for mutations that affect gene expression in *trans*. Transcript levels can be affected by polymorphisms in genes with roles in signal transduction, transcription, and mRNA processing, export and decay, as well as other regulatory and feedback processes in a cell. The three genes in which we have identified

natural *trans*-regulatory polymorphisms to date encode a transcription factor (Hap1; ref. 2), a G protein involved in receptor signaling (Gpa1) and a protein that seems to act through protein-protein interaction (Amn1; ref. 23). Identification of *trans*-regulatory variation in genes that would not have been considered likely to encode transcriptional regulators *a priori* underscores the power of the mapping approach for finding new nodes in regulatory networks.

METHODS

Strains, molecular biology and functional assays. Parent strains were BY4716, an S288C derivative (*MATα lys2Δ0*; ref. 34), and RM11-1a (*MATa leu2Δ0 ura3Δ0 ho::KAN*), a haploid derived from Bb32(3), a natural isolate collected by R. Mortimer[35]. Both strains were gifts from L. Hartwell (Fred Hutchinson Cancer Research Center, Seattle, Washington, USA). We amplified the 3′ end of *GPA1*, including 170 bp downstream of the stop codon, from the RM strain and fused it to the nourseothricin resistance gene of pAG25 (ref. 36). We transformed this construct into the BY strain as described[37]. After selection with 100 μg ml^{-1} cloneNAT, we tested colonies by PCR and sequenced them in duplicate for the presence of the polymorphism. We replaced the *AMN1* ORF with the gene for hygromycin-B resistance using standard PCR-based techniques[38,39]. We confirmed *amn1* deletion strains by PCR. Primer sequences are available on request. The source of the hygromycin-B resistance gene was pFVL100, a plasmid provided by F. van Leeuven (Fred Hutchinson Cancer Research Center, Seattle, Washington, USA). pFVL100 was created by subcloning a *Pac*I–*Bsm*I fragment containing the hygromycin-B resistance gene from pAG32 (ref. 36) into the vector backbone of pRS400 (ref. 34) cut with the same enzymes. We sonicated cultures at mid-log phase and then examined them for clumpiness both visually and by treating them with the fluorescent DNA-binding dye Sytox followed by flow cytometry, which was carried out as described previously[40].

Expression measurements and genotyping. We measured expression and carried out genotyping as described[2], except that we measured expression from cultures grown in 30 ml rather than 100 ml of medium. All array hybridizations were done in duplicate with the fluors reversed.

Gene clustering. We carried out hierarchical clustering of all genes in the data set (2 BY values, 2 RM values and 86 segregant values) using *cluster*[10], applying mean-centering and normalization of genes and arrays before absolute, centered, complete-linkage clustering. We estimated the significance of the correlation between genes empirically by running the clustering program on ten data sets of permuted data in which the 90 expression values for each gene were randomly redistributed. On average, only 2.5 clusters of two genes and no clusters of more than two genes were formed at $r^2 > 0.725$ in the permuted data.

Linkage analysis. We tested for linkage between an individual expression level and a genetic marker by separating the segregants into two groups according to marker genotype and comparing the expression levels between the groups by the Wilcoxon–Mann–Whitney test, with empirical false-discovery rates computed by permutation tests as described[2]. We used the same method to test linkage between a cluster of expression levels and a genetic marker, with the mean expression level of the genes in the cluster serving as the quantitative phenotype of each segregant.

Transcription factor analysis. We downloaded chromatin immunoprecipitation data[25,26] from the Young Lab website. We considered all interactions with a reported $P < 0.001$. For each gene cluster, we determined if the cluster was significantly enriched in targets of a specific factor by computing the probability p that the number of targets equals or exceeds the observed number by chance from the binomial distribution $B(n, f)$, where n is the cluster size and f is the fraction of all genes in the genome that are targets of the factor. Because 593 clusters and 113 factors were tested, no factor-cluster combination was expected to show $P < 10^{-5}$ by chance; we therefore reported the factor-cluster combinations meeting this significance level. In testing whether genes showing highly significant linkage were bound by transcription factors located at the region of linkage, we examined 807 gene–transcription factor pairs and for each pair asked whether the transcription factor fell within 20 kb of the linkage peak. In a genome of 12 Mb, about three false positives are expected. Finally, to test whether the set of *trans*-regulatory loci was enriched in transcription factors, we considered the list of all 123 transcription factors as a gene category in the analysis described below.

Gene Ontology analysis. We downloaded a complete list of yeast ORFs annotated with Gene Ontology terms from the Saccharomyces Genome Database. We mapped these terms to higher level parent terms using PERL scripts to query a locally installed Gene Ontology Molecular Function and Biological Process database. We generated categories by mapping Gene Ontology terms to either higher or lower level terms until each category contained at least 20 and at most 400 genes. We imposed these upper and lower bounds to provide enough genes in each category to conduct statistical tests but avoid saturating the genome with a particular term. For example, 1,720 yeast genes are associated with the Gene Ontology term 'enzyme', and one of them is likely to map near every *trans*-regulatory locus, which would make identifying enrichment impossible. To test whether a particular Gene Ontology category was overrepresented near *trans*-regulatory loci, we counted the observed number of genes in the category located within 10 kb of a *trans*-linked peak marker. Genes that mapped near multiple *trans*-linked markers were counted only once. The number of genes in a category expected to fall near *trans*-linked markers by chance was determined by randomly sampling an equal number of Gene Ontology–annotated genes without regard to category and counting the number of genes located within 10 kb of a *trans*-linked marker. Results from 1,000 random samples for each category were averaged. We then carried out a χ^2 test with 1 degree of freedom for each category to determine whether genes in that category fell near *trans*-linked markers more frequently than expected.

URLs. The Saccharomyces Genome Database is available at http://www.yeastgenome.org, the Gene Ontology Database at http://www.geneontology.org and the Young Lab website at http://staffa.wi.mit.edu/cgi-bin/young_public/navframe.cgi?s=17&f=downloaddata.

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

ACKNOWLEDGMENTS

We thank J. Delrow, D. Hare, J. Marty and the Fred Hutchinson Cancer Research Center microarray, oligo synthesis and sequencing facilities for technical assistance; M. Eberle and H. Childs for help in manuscript preparation; D. Gottschling and R. Young for discussions; and H. Coller for discussions and a critical reading of the manuscript. G.Y. and R.B. were supported by postdoctoral fellowships from the Human Frontier Science Program and the National Institutes of Health, respectively. This work was supported by the Howard Hughes Medical Institute, of which L.K. is an associate investigator. L.K. is a James S. McDonnell Centennial Fellow.

COMPETING INTERESTS STATEMENT

The authors declare that they have no competing financial interests.

Received February 11, 2003; accepted 26 June 2003.
Published online at http://www.nature.com/naturegenetics/

1. Oleksiak, M.F., Churchill, G.A. & Crawford, D.L. Variation in gene expression within and among natural populations. *Nat. Genet.* **32**, 261–266 (2002).
2. Brem, R.B., Yvert, G., Clinton, R. & Kruglyak, L. Genetic dissection of transcriptional regulation in budding yeast. *Science* **296**, 752–755 (2002).
3. Cowles, C.R., Hirschhorn, J.N., Altshuler, D. & Lander, E.S. Detection of regulatory variation in mouse genes. *Nat. Genet.* **32**, 432–437 (2002).
4. Yan, H., Yuan, W., Velculescu, V.E., Vogelstein, B. & Kinzler, K.W. Allelic variation in human gene expression. *Science* **297**, 1143 (2002).
5. Rockman, M.V. & Wray, G.A. Abundant raw material for *cis*-regulatory evolution in humans. *Mol. Biol. Evol.* **19**, 1991–2004 (2002).
6. Cheung, V.G. *et al.* Natural variation in human gene expression assessed in lymphoblastoid cells. *Nat. Genet.* **33**, 422–425 (2003).
7. Schadt, E.E. *et al.* Genetics of gene expression surveyed in maize, mouse and man. *Nature* **422**, 297–302 (2003).
8. Knight, J.C., Keating, B.J., Rockett, K.A. & Kwiatkowski, D.P. *In vivo* characterization of regulatory polymorphisms by allele-specific quantification of RNA polymerase loading. *Nat. Genet.* **33**, 469–475 (2003).
9. Wu, L.F. *et al.* Large-scale prediction of *Saccharomyces cerevisiae* gene function using overlapping transcriptional clusters. *Nat. Genet.* **31**, 255–265 (2002).
10. Eisen, M.B., Spellman, P.T., Brown, P.O. & Botstein, D. Cluster analysis and display

of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA* **95**, 14863–14868 (1998).
11. Gasch, A.P. *et al.* Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell* **11**, 4241–4257 (2000).
12. Hughes, T.R. *et al.* Functional discovery via a compendium of expression profiles. *Cell* **102**, 109–126 (2000).
13. Spellman, P.T. *et al.* Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell* **9**, 3273–3297 (1998).
14. Chu, S. *et al.* The transcriptional program of sporulation in budding yeast. *Science* **282**, 699–705 (1998).
15. Jansen, R.C. & Nap, J.P. Genetical genomics: the added value from segregation. *Trends Genet.* **17**, 388–391 (2001).
16. Jansen, R.C. Opinion: studying complex biological systems using multifactorial perturbation. *Nat. Rev. Genet.* **4**, 145–151 (2003).
17. Winzeler, E.A. *et al.* Direct allelic variation scanning of the yeast genome. *Science* **281**, 1194–1197 (1998).
18. Hirsch, J.P., Dietzel, C. & Kurjan, J. The carboxyl terminus of Scg1, the G alpha subunit involved in yeast mating, is implicated in interactions with the pheromone receptors. *Genes Dev.* **5**, 467–474 (1991).
19. Brown, A.J. *et al.* Functional coupling of mammalian receptors to the yeast mating pathway using novel yeast/mammalian G protein α-subunit chimeras. *Yeast* **16**, 11–22 (2000).
20. Kellis, M., Patterson, N., Endrizzi, M., Birren, B. & Lander, E.S. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423**, 241–254 (2003).
21. Souciet, J. *et al.* Genomic exploration of the hemiascomycetous yeasts: 1. A set of yeast species for molecular evolution studies. *FEBS Lett.* **487**, 3–12 (2000).
22. Colman-Lerner, A., Chin, T.E. & Brent, R. Yeast Cbk1 and Mob2 activate daughter-specific genetic programs to induce asymmetric cell fates. *Cell* **107**, 739–750 (2001).
23. Wang, Y., Shirogane, T., Liu, D., Harper, J.W. & Elledge, S.J. Exit from exit: resetting the cell cycle through Amn1 inhibition of G protein signaling. *Cell* **112**, 697–709 (2003).
24. Giaever, G. *et al.* Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature* **418**, 387–391 (2002).
25. Lee, T.I. *et al.* Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science* **298**, 799–804 (2002).
26. Simon, I. *et al.* Serial regulation of transcriptional regulators in the yeast cell cycle. *Cell* **106**, 697–708 (2001).
27. Jones, E.W., Pringle, J.R. & Broach, J.R. *The molecular and cellular biology of the yeast Saccharomyces* (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, 1992).
28. Forsburg, S.L. & Guarente, L. Identification and characterization of HAP4: a third component of the CCAAT-bound HAP2/HAP3 heteromer. *Genes Dev.* **3**, 1166–1178 (1989).
29. Spector, M.S., Raff, A., DeSilva, H., Lee, K. & Osley, M.A. Hir1p and Hir2p function as transcriptional corepressors to regulate histone gene transcription in the *Saccharomyces cerevisiae* cell cycle. *Mol. Cell. Biol.* **17**, 545–552 (1997).
30. Guelzim, N., Bottani, S., Bourgine, P. & Kepes, F. Topological and causal structure of the yeast transcriptional regulatory network. *Nat. Genet.* **31**, 60–63 (2002).
31. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29 (2000).
32. Olson, M.V. When less is more: gene loss as an engine of evolutionary change. *Am. J. Hum. Genet.* **64**, 18–23 (1999).
33. Rifkin, S.A., Kim, J. & White, K.P. Evolution of gene expression in the *Drosophila melanogaster* subgroup. *Nat. Genet.* **33**, 138–144 (2003).
34. Brachmann, C.B. *et al.* Designer deletion strains derived from *Saccharomyces cerevisiae* S288C: a useful set of strains and plasmids for PCR-mediated gene disruption and other applications. *Yeast* **14**, 115–132 (1998).
35. Mortimer, R.K., Romano, P., Suzzi, G. & Polsinelli, M. Genome renewal: a new phenomenon revealed from a genetic study of 43 strains of *Saccharomyces cerevisiae* derived from natural fermentation of grape musts. *Yeast* **10**, 1543–1552 (1994).
36. Goldstein, A.L. & McCusker, J.H. Three new dominant drug resistance cassettes for gene disruption in *Saccharomyces cerevisiae*. *Yeast* **15**, 1541–1553 (1999).
37. Gietz, R.D. & Woods, R.A. Transformation of yeast by lithium acetate/single-stranded carrier DNA/polyethylene glycol method. *Methods Enzymol.* **350**, 87–96 (2002).
38. Baudin, A., Ozier-Kalogeropoulos, O., Denouel, A., Lacroute, F. & Cullin, C. A simple and efficient method for direct gene deletion in *Saccharomyces cerevisiae*. *Nucleic Acids Res.* **21**, 3329–3330 (1993).
39. Lorenz, M.C. *et al.* Gene disruption with PCR products in *Saccharomyces cerevisiae*. *Gene* **158**, 113–117 (1995).
40. Foss, E.J. Tof1p regulates DNA damage responses during S phase in *Saccharomyces cerevisiae*. *Genetics* **157**, 567–577 (2001).