Genetic interactions between polymorphisms that affect gene expression in yeast

Rachel B. Brem[1]*, John D. Storey[2]*, Jacqueline Whittle[3] & Leonid Kruglyak[4]

Interactions between polymorphisms at different quantitative trait loci (QTLs) are thought to contribute to the genetics of many traits, and can markedly affect the power of genetic studies to detect QTLs[1]. Interacting loci have been identified in many organisms[1–5]. However, the prevalence of interactions[6–8], and the nucleotide changes underlying them[9,10], are largely unknown. Here we search for naturally occurring genetic interactions in a large set of quantitative phenotypes—the levels of all transcripts in a cross between two strains of *Saccharomyces cerevisiae*[7]. For each transcript, we searched for secondary loci interacting with primary QTLs detected by their individual effects. Such locus pairs were estimated to be involved in the inheritance of 57% of transcripts; statistically significant pairs were identified for 225 transcripts. Among these, 67% of secondary loci had individual effects too small to be significant in a genome-wide scan. Engineered polymorphisms in isogenic strains confirmed an interaction between the mating-type locus *MAT* and the pheromone response gene *GPA1*. Our results indicate that genetic interactions are widespread in the genetics of transcript levels, and that many QTLs will be missed by single-locus tests but can be detected by two-stage tests that allow for interactions.

Most heritable traits are affected by the inheritance of alleles at multiple loci, and the identification of these loci is a key challenge of modern genetic research. We recently showed that gene expression levels in yeast typically show multigenic inheritance and provide a good model for investigating the genetic basis of complex traits[7,11,12]. Here we use this system to examine the prevalence of genetic interactions in a large set of phenotypes. A genetic interaction between a pair of loci (sometimes termed epistasis) occurs when the effect of an allele at one locus changes as a function of the allele at the other. Previous biometric analyses have provided evidence for many interactions underlying transcript levels[6,7]. We sought to identify the loci involved in such interactions in a cross between a laboratory strain of yeast, BY, and a wild strain, RM. For this analysis we used previously described genotype and gene expression data from 112 segregants[7]. We first tested all possible pairs of loci for evidence of interaction underlying each transcript level. This strategy had little statistical power to map interactions, because of its requirement for a very large number of tests and the corresponding stringent correction of significance thresholds for multiple testing[13].

To improve power we used a two-stage search strategy. For each transcript, we first identified the 'primary' QTL with the strongest individual linkage. We then partitioned segregants on the basis of inheritance (either BY or RM) at the primary locus and tested each subgroup for further 'secondary' loci. We computed the joint significance of the strongest such secondary QTL with the primary locus by using a newly developed statistical method based on

estimation of the false discovery rate[13] (see Methods). This method provides an estimate of the overall fraction of transcripts for which both loci are involved in the genetics, as well as a probability for each transcript that both loci are true positives. This analysis indicated that the locus pairs identified for 57% of all transcripts were true positives, and it identified locus pairs for 225 transcripts at a false discovery rate of 5% (that is, fewer than 12 of these are expected to have either locus as a false positive).

Because the two-stage search takes account of inheritance at the primary locus in assessing the effect of the secondary locus, it can identify both locus pairs with interactions and locus pairs that act additively. We tested each of the 225 locus pairs for interacting effects by fitting a regression model relating inheritance at the linking loci to the transcript level. The model included an additive effect for each locus and a term representing the interaction between loci; the latter term was significant at $P < 0.05$ for 65% of transcripts. For comparison, we considered 547 transcripts for which two loci were identified as significant by their individual effects. Of these, only 13% had an interaction term significant at $P < 0.05$. A false discovery rate analysis[14] indicated that the interaction term represented a true positive for 91% of locus pairs mapped by the two-stage search, compared with 13% of pairs mapped independently.

We next sought to determine whether the secondary loci mapped by the two-stage search could have been detected by a single-locus linkage test. Under some interaction models, the individual effects of the loci are undetectable, whereas under others the loci retain an individual effect. We tested the individual effect of each secondary locus on the corresponding transcript and found that 87% of these were significant at $P < 0.05$, indicating that most secondary loci do have a residual individual effect. However, this significance threshold does not take into account the multiple tests performed in a genome scan. When we examined how many of the secondary loci were detectable at a false discovery rate of 5% in the context of a genome scan[7], we found that only 74 (33%) of the 225 were, indicating that 67% of secondary loci would have been missed without the two-stage search.

To identify pairs of interacting QTLs with effects on many transcripts, we constructed a histogram of the genetic positions of linking locus pairs, analogous to previous analyses with single loci[11,15]. Most of the pairs affected single transcripts, but a few affected multiple transcripts (Fig. 1). The largest number of transcripts linking to a single locus pair was 14, with four additional transcripts linking nearby (Fig. 1); because this group contained both YCR040W/*MATα1* and its silenced copy YCL066W/*HMLα1*, we eliminated the latter from further analysis, leaving 17 transcripts that linked to this locus pair. The primary QTL of the pair lay near *MAT*, which confers α or **a** mating type on a haploid yeast cell

[1]Program in Computational Biology, Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue N, M2-B876, Seattle, Washington 98109, USA. [2]Department of Biostatistics and Department of Genome Sciences, University of Washington, Campus Mail Stop 357232, Seattle, Washington 98195-7232, USA. [3]Infectious Disease Research Institute, 1124 Columbia Street, Suite 600, Seattle, Washington 98104, USA. [4]Lewis-Sigler Institute for Integrative Genomics and Department of Ecology and Evolutionary Biology, Carl Icahn Laboratory, Princeton University, Princeton, New Jersey 08544-1003, USA.
*These authors contributed equally to this work.

depending on the integration of genes at the locus. The secondary locus contained the gene encoding the G-protein subunit Gpa1, in which a single polymorphism in the BY parent strain, S469I, has previously been shown to affect the expression of pheromone response genes[12]. Of the 17 transcripts whose linkage to the *MAT* and *GPA1* loci was detected by the two-stage search, 11 were previously shown to be regulated by mating type (of which seven are also regulated by the pheromone response pathway); of the remaining six linking transcripts, two have a known function that does not involve mating, and four have no known function[16]. For 14 of the transcripts, the *MAT* and *GPA1* loci showed direct evidence for interaction in the regression model at $P < 0.05$; the independent effect of the *GPA1* locus exceeded the genome-wide significance threshold for single-locus linkage for only one of these 17 transcripts.

To test the hypothesis that variation at *MAT* and the S469I mutation in GPA1p interact genetically, we engineered isogenic yeast strains carrying each of the four possible combinations of alleles at the two loci. For each of these combinations we compared expression in the engineered strains with expression in segregants having the same allele combination (but having varying inheritance for the rest of the genome). For 7 of 11 transcripts previously known to be regulated by mating type, and 3 of 4 transcripts of unknown function, the pattern of effect of the *MAT*–*GPA1* genotype on expression was the same in engineered strains as in segregants (Fig. 2 and Supplementary Information), indicating that variation at *MAT* and *GPA1* is sufficient to explain transcript level differences across segregants. YOR090C/*PTC1* and YOR162C/*YRR1*, which have known functions unrelated to mating, did not show agreement between engineered strains and segregants (Supplementary Information), indicating that these effects might represent either false positives or QTLs linked, but unrelated, to *MAT* or *GPA1*.

In a wild-type haploid cell exposed to pheromone, *GPA1* activates the pheromone response pathway to prepare for mating[17,18]. We showed previously that, in the absence of endogenous pheromone, the S469I variant is associated with the upregulation of genes involved in mating-type-independent mating functions (for example cytoskeletal rearrangement and cell fusion)[12]. By contrast, most genes affected by the genetic interaction between *MAT* and *GPA1* perform mating-type-specific mating functions (for example pheromone detection and cell conjugation). Levels of α-specific transcripts in the **a** background have been measured at less than one copy per cell[19]. This indicates a possible molecular model for the interaction between *MAT* and *GPA1*: because of tight repression by *MAT*, other regulators have no impact on **a**-specific transcripts in the α background, and

vice versa (Fig. 2a–d and Supplementary Information). This situation is analogous to a polymorphism with effects on sex-limited traits in each of the two sexes.

GPA1 polymorphism affected the expression of *MATα1* and *MATα2* (Fig. 2e, f). This is in contrast to results from genome-wide studies[18,20] and a classic molecular biology report[21] showing no significant regulation of *MAT* genes by the pheromone response pathway. The discrepancy could indicate that the S469I variant of Gpa1p affects regulation directly at *MAT* through a different pathway from that of exogenous pheromone. Alternatively, *GPA1* might act indirectly through other regulators of *MAT*[22]. We speculate that expression changes due to polymorphism at *GPA1* might affect the regulatory activity of *MAT*-encoded proteins on their downstream targets. For several transcripts, polymorphism in *GPA1* seemed to show opposing effects in the two mating types (Fig. 2e, f and Supplementary Information); whether this represents a true result or an artefact is unclear.

Mapping of genetic factors that underlie polygenic physiological, medical and agricultural traits in outbred populations has met with limited success. It has been suggested that this might be mainly because the sizes of individual locus effects are modest, owing to genetic interactions between loci[1], but the prevalence of interactions has not been well characterized. Our results indicate that genetic interactions underlie the inheritance of roughly half of all transcript levels in yeast and that at least one member of an interacting locus pair typically has too small an individual effect to be identified on its own. We have shown that a two-stage search provides a useful strategy for the identification of such secondary loci. Because this
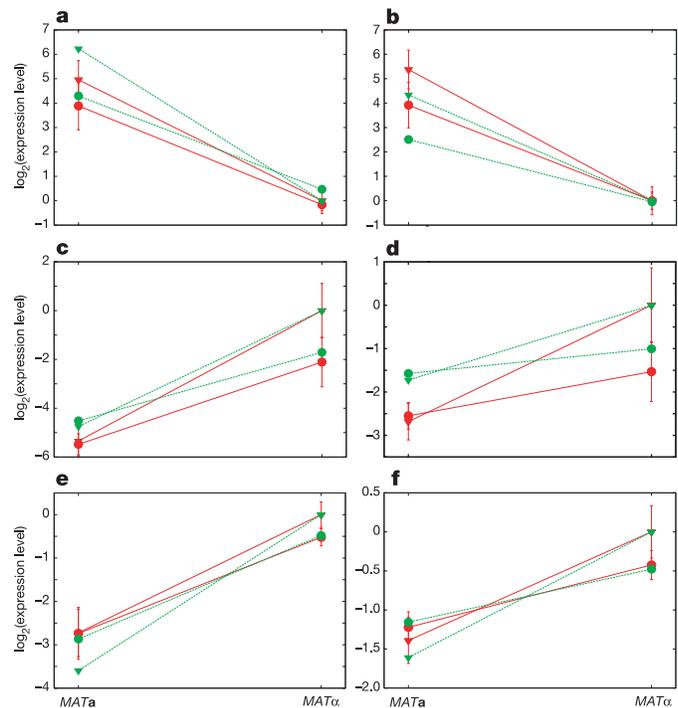
Figure 2 | Example transcripts showing genetic interaction between *MAT* and *GPA1*. Each panel represents one transcript: **a**, YFL026W/*STE2*; **b**, YIL015W/*BAR1*; **c**, YJR004C/*SAG1*; **d**, YGL090W/*LIF1*; **e**, YCR040W/*MATα1*; **f**, YCR039C/*MATα2*. Each red point represents the mean expression level over segregants with one genotype, normalized by the mean over segregants inheriting the BY allele at both loci; red error bars represent standard deviations among segregants with the respective genotype. Each green point represents the expression level in one engineered strain, normalized by the level in the engineered strain with the BY allele at both loci. Strain mating type is indicated on the *x* axis (*MAT***a** is the RM allele, *MAT*α is the BY allele). Triangles and circles represent strains carrying the BY and RM alleles of *GPA1*, respectively.
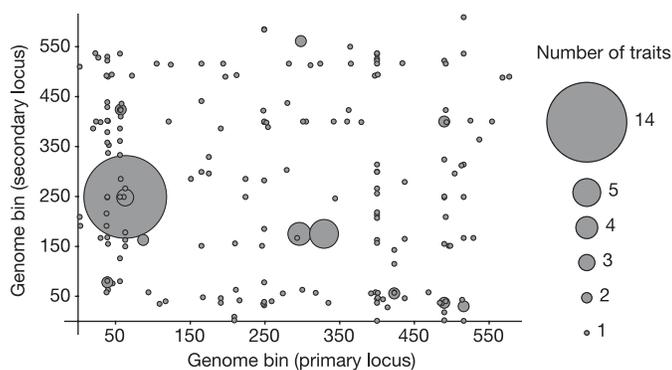

Figure 1 | Genome distribution of QTL pairs detected by the two-stage linkage search. On each axis the genome is divided into 611 bins of 20 kilobases each, shown in chromosomal order. The set of transcripts mapping to QTLs in each pair of bins in the two-stage analysis is represented as a circle, with the width proportional to the number of such co-linking transcripts (a scale is shown at the right); circles are centred on the corresponding bins. The largest circle represents the *MAT*–*GPA1* locus pair.

strategy relies on the identification of primary loci by their individual effects, the detection of interacting locus pairs in which both individual effects are small remains a challenge, so the overall prevalence of interactions may be even higher than estimated here. The discovery here of previously uncharacterized genetic effects in the well-studied yeast pheromone response network underscores the importance of interaction mapping in genetic analyses.

METHODS

Strains and expression measurements. Segregants, genotypes, expression measurements and single-locus linkage results were those of ref. 7. To find multiple independent linkages, for each transcript we identified the marker with the strongest linkage score on each chromosome; if two or more such chromosome peaks exceeded the genome-wide single-locus cut-off in ref. 7, we classified these as multiple independent QTLs. For the direct test of MAT–GPA1 interaction, the I469 and S469 alleles of GPA1 were each engineered into the S288c derivatives BY4709 (MATα URA3Δ0) and BY4724 (MATa LYS2Δ0 URA3Δ0) as in ref. 12; expression arrays were performed as in ref. 7, except that the reference sample was a 1:1 mixture of RNAs from the BY and RM strains. For two-stage mapping and interaction tests on the resulting loci, spatial loess normalization was performed on expression data using the genes in ref. 12. Missing genotype data were imputed with the use of a standard hidden Markov model algorithm implemented in R/qtl (ref. 23). Missing expression data were imputed by using a $K = 15$ nearest-neighbours method[24].

Two-locus mapping. A two-stage procedure was employed to identify pairs of linked loci for each expression trait. For each transcript and marker, a Wilcoxon rank-sum statistic was formed to quantify expression differences between the segregants grouped by inheritance at the locus. We identified the 'primary' QTL for each transcript as the locus with the most significant Wilcoxon rank-sum statistic. We then partitioned segregants on the basis of inheritance (either BY or RM) at the primary locus and similarly tested each subgroup for further 'secondary' loci. The locus with the highest statistic between either partition was chosen as the secondary locus. At both stages, the expression traits were randomly permuted (five times) and analogous maximal statistics were recalculated from this null distribution. These null statistics were pooled across transcripts, for a total of about 30,000 at each stage.

We considered a transcript to be a 'false discovery' if either the primary or the secondary locus was a false positive. Under this definition of a false discovery, it is not straightforward to calculate a P value for each transcript because the null distribution must account for an unknown mixture of three situations: both loci are false positives, the primary locus only is a false positive, or the secondary locus only is a false positive. We therefore employed a new statistical method to rank the expression traits for significance and calculate the false discovery rate for each significance cut-off[13]. In brief, we formed nonparametric empirical Bayes estimates of the posterior probability that the primary QTL is a true positive, and then the posterior probability that the secondary QTL is a true positive given that the primary locus is a true positive. The joint probability that both loci are true positives is equal to the product of these two probabilities. The observed and permutation-based null statistics for each stage were used to form conservative estimates of these probabilities (ref. 13). In the bayesian setting, the false discovery rate is exactly equal to the probability that a transcript is a false positive given that it is called significant[25]. Equivalently, it can be written as one minus the probability that the transcript is a true positive, given that it is called significant. The estimated joint linkage probabilities can be used to estimate this latter quantity directly, yielding an estimate of the false discovery rate for any chosen joint linkage probability cut-off. With a cut-off corresponding to a 5% false discovery rate, 225 expression traits were called significant for two-locus linkage with this method.

Interaction test. For each locus pair mapped in the two-stage model or mapped independently, we fitted the model

$$t = ax + by + cxy + d$$

over all segregants. Here t is $\log_2$ of the ratio of expression between the strain of interest and a reference sample, a, b, c, and d are parameters that are specific to the given transcript, x is inheritance at the first locus, and y is inheritance at the second locus. A standard F-test was used to test the null hypothesis that $c = 0$ by comparing the goodness-of-fit of the above full model with that of the purely additive model $t = ax + by + d$. Because these tests were performed on transcripts and locus pairs previously mapped by the two-stage or independent linkage calculations, there is a potential for the significance to be artificially inflated. However, because the model used in each linkage search is a restricted version of the purely additive model, the interaction term can in fact be tested on previously mapped locus pairs without incurring a bias. The program QVALUE

(faculty.washington.edu/~jstorey/qvalue/) was applied to these P values to estimate the total proportion of transcripts showing evidence for interaction.

1. Carlborg, O. & Haley, C. S. Epistasis: too often neglected in complex trait studies? Nature Rev. Genet. 5, 618–625 (2004).
2. Leamy, L. J., Workman, M. S., Routman, E. J. & Cheverud, J. M. An epistatic genetic basis for fluctuating asymmetry of tooth size and shape in mice. Heredity 94, 316–325 (2005).
3. Montooth, K. L., Marden, J. H. & Clark, A. G. Mapping determinants of variation in energy metabolism, respiration and flight in Drosophila. Genetics 165, 623–635 (2003).
4. Shook, D. R. & Johnson, T. E. Quantitative trait loci affecting survival and fertility-related traits in Caenorhabditis elegans show genotype–environment interactions, pleiotropy and epistasis. Genetics 153, 1233–1243 (1999).
5. Lynch, M. & Walsh, B. Genetics and Analysis of Quantitative Traits (Sinauer, Sunderland, Massachusetts, 1998).
6. Gibson, G. et al. Extensive sex-specific nonadditivity of gene expression in Drosophila melanogaster. Genetics 167, 1791–1799 (2004).
7. Brem, R. B. & Kruglyak, L. The landscape of genetic complexity across 5,700 gene expression traits in yeast. Proc. Natl Acad. Sci. USA 102, 1572–1577 (2005).
8. Kroymann, J. & Mitchell-Olds, T. Epistasis and balanced polymorphism influencing complex trait variation. Nature 435, 95–98 (2005).
9. Rawson, P. D. & Burton, R. S. Functional coadaptation between cytochrome c and cytochrome c oxidase within allopatric populations of a marine copepod. Proc. Natl Acad. Sci. USA 99, 12955–12958 (2002).
10. Caicedo, A. L., Stinchcombe, J. R., Olsen, K. M., Schmitt, J. & Purugganan, M. D. Epistatic interaction between Arabidopsis FRI and FLC flowering time genes generates a latitudinal cline in a life history trait. Proc. Natl Acad. Sci. USA 101, 15670–15675 (2004).
11. Brem, R. B., Yvert, G., Clinton, R. & Kruglyak, L. Genetic dissection of transcriptional regulation in budding yeast. Science 296, 752–755 (2002).
12. Yvert, G. et al. Trans-acting regulatory variation in Saccharomyces cerevisiae and the role of transcription factors. Nature Genet. 35, 57–64 (2003).
13. Storey, J. D., Akey, J. M. & Kruglyak, L. Multiple locus linkage analysis of genomewide expression in yeast. PLoS Biol. 3, e267 (2005).
14. Storey, J. D. A direct approach to false discovery rates. J. R. Statist. Soc. B 64, 479–498 (2002).
15. Schadt, E. E. et al. Genetics of gene expression surveyed in maize, mouse and man. Nature 422, 297–302 (2003).
16. Csank, C. et al. Three yeast proteome databases: YPD, PombePD, and CalPD (MycoPathPD). Methods Enzymol. 350, 347–373 (2002).
17. Bardwell, L., Cook, J. G., Inouye, C. J. & Thorner, J. Signal propagation and regulation in the mating pheromone response pathway of the yeast Saccharomyces cerevisiae. Dev. Biol. 166, 363–379 (1994).
18. Zeitlinger, J. et al. Program-specific distribution of a transcription factor dependent on partner transcription factor and MAPK signalling. Cell 113, 395–404 (2003).
19. Velculescu, V. E. et al. Characterization of the yeast transcriptome. Cell 88, 243–251 (1997).
20. Roberts, C. J. et al. Signaling and circuitry of multiple MAPK pathways revealed by a matrix of global gene expression profiles. Science 287, 873–880 (2000).
21. Hagen, D. C. & Sprague, G. F. Jr Induction of the yeast alpha-specific STE3 gene by the peptide pheromone a-factor. J. Mol. Biol. 178, 835–852 (1984).
22. Sudarsanam, P., Iyer, V. R., Brown, P. O. & Winston, F. Whole-genome expression analysis of snf/swi mutants of Saccharomyces cerevisiae. Proc. Natl Acad. Sci. USA 97, 3364–3369 (2000).
23. Broman, K. W., Wu, H., Sen, S. & Churchill, G. A. R/qtl: QTL mapping in experimental crosses. Bioinformatics 19, 889–890 (2003).
24. Troyanskaya, O. et al. Missing value estimation methods for DNA microarrays. Bioinformatics 17, 520–525 (2001).
25. Storey, J. D. The positive false discovery rate: A Bayesian interpretation and the q-value. Ann. Stat. 31, 2013–2035 (2003).