Published in final edited form as:
Mol Ecol. 2011 October ; 20(19): 4085–4097. doi:10.1111/j.1365-294X.2011.05225.x.

GENOME-WIDE ASSOCIATION ANALYSIS OF CLINICAL VERSUS NON-CLINICAL ORIGIN PROVIDES INSIGHTS INTO *SACCHAROMYCES CEREVISIAE* PATHOGENESIS

Ludo A. H. Muller[1,*], **Joseph E. Lucas**[2], **D. Ryan Georgianna**[1], and **John H. McCusker**[1]
[1]Department of Molecular Genetics and Microbiology, Duke University Medical Center, Durham, NC 27710, USA

[2]Institute for Genome Sciences and Policy, Duke University, Durham, NC 27708, USA

Abstract

Because domesticated *Saccharomyces cerevisiae* strains have been used to produce fermented food and beverages for centuries without apparent health implications, *S. cerevisiae* has always been considered a Generally Recognized As Safe (GRAS) microorganism. However, the number of reported mucosal and systemic *S. cerevisiae* infections in the human population has increased and fatal infections have occured even in relatively healthy individuals. In order to gain insight into the pathogenesis of *S. cerevisiae* and improve our understanding of the emergence of fungal pathogens, we performed a population-based genome-wide environmental association analysis of clinical versus non-clinical origin in *S. cerevisiae*. Using tiling array-based, high density genotypes of 44 clinical and 44 non-clinical *S. cerevisiae* strains from diverse geographical origins and source substrates, we identified several genetic loci associated with clinical background in *S. cerevisiae*. Associated polymorphisms within the coding sequences of *VRP1*, *KIC1*, *SBE22* and *PDR5*, and the 5′ upstream region of *YGR146C* indicate the importance of pseudohyphal formation, robust cell wall maintenance and cellular detoxification for *S. cerevisiae* pathogenesis, and constitute good candidates for follow-up verification of virulence and virulence-related factors underlying the pathogenicity of *S. cerevisiae*.

Keywords

Clinical origin; genome-wide association analysis; environmental association; pathogenesis; *Saccharomyces cerevisiae*; tiling array

INTRODUCTION

Domesticated strains of the hemiascomycetous yeast *Saccharomyces cerevisiae* have been employed for centuries in the production of fermented food and beverages, which has been considered safe practice as *S. cerevisiae* was generally recognized as being non-pathogenic. However, an increase in the number of reported mucosal and systemic infections caused by *S. cerevisiae* warranted a reevaluation and this species is now regarded as an opportunistic pathogen of low virulence (de Hoog, 1996; Enache-Angoulvant & Hennequin, 2005).

Corresponding author: John H. McCusker, Department of Molecular Genetics and Microbiology, Duke University Medical Center, box 3020, Durham, NC 27710, USA, Telephone: +1 919 681 6778, Fax: +1 919 684 8735, mccus001@mc.duke.edu.
[*] Current address: Institut fu□r Biologie - Botanik, Freie Universita□t Berlin, Altensteinstraße 6, 14195, Berlin, Germany

DATA ACCESSIBILITY
Microarray data: ArrayExpress accession E-MTAB-627

Clinical syndromes caused by *S. cerevisiae* include pneumonia, peritonitis, vaginitis and fungemia (Muñoz *et al*., 2005), and although *S. cerevisiae* infections often involve immunocompromised or critically ill patients, fatal infections also occur in relatively healthy individuals (Smith *et al*., 2002). *S. cerevisiae* is commonly present as part of the microflora in the gastrointestinal tract, the respiratory tract and the vagina (Kwon-Chung & Bennett, 1992; Salonen *et al*., 2000), and infection is believed to occur through translocation of ingested microorganisms from the oral or enteral mucosa and contamination of catheter insertion sites (Hennequin *et al*., 2000; Muñoz *et al*., 2005). The sources of emerging *S. cerevisiae* pathogens are largely unclear, but some have shown to be food and drink related (de Llanos *et al*., 2004; de Llanos *et al*., 2006a), while others are nosocomial. A notable example of nosocomial infections are those caused by *S. cerevisiae* var. *boulardii*, which is used as a probiotic to treat antibiotic-associated diarrhoea and *Clostridium difficile* infections (Guslandi, 2006), and which has caused fungemia in various cases (Bassetti *et al*., 1998; Muñoz *et al*., 2005). Nevertheless, the genetic diversity among pathogenic *S. cerevisiae* strains is large and a single origin is unlikely (Muller & McCusker, 2009).

Although *S. cerevisiae* infections are opportunistic in nature and host factors are believed to be more important for the establishment of infection than yeast characteristics (Klingberg *et al*., 2008), virulence does vary greatly among *S. cerevisiae* strains and is expected to have an effect on the extent of invasiveness (Clemons *et al*., 1994). Both virulence factors, which interact directly with the host, and fitness attributes that enhance the ability to thrive in the host contribute to fungal pathogenicity, and most knowledge regarding virulence and virulence-related factors is available for *Candida albicans*, the major fungal pathogen of humans. Recognized virulence factors of *C. albicans* include morphogenesis (the transition between yeast cells and filamentous growth; Sundstrom, 2006), secretion of degradative enzymes such as aspartyl proteinases and phospholipases (Hube & Naglik, 2002), adhesin-promoted binding to host tissue (Hoyer *et al*., 2008) and invasin-enhanced uptake by host cells (Phan *et al*., 2007). Suspected virulence factors in *S. cerevisiae* are pseudohyphal growth and high levels of phospholipase activity, while the ability to grow at high temperatures has been proposed as an important virulence-related fitness attribute (McCusker *et al*., 1994; de Llanos *et al*., 2006b). However, the intrinsic properties that determine pathogenesis in *S. cerevisiae* remain poorly understood.

Here, we report on the first population-based genome-wide environmental association analysis of clinical versus non-clinical background in *S. cerevisiae*. Using 135,771 single feature polymorphisms (SFPs) in independent case-control samples of 44 clinical and 44 non-clinical *S. cerevisiae* strains from various geographical origins and source substrates, we identified several polymorphisms, consituting both coding and regulatory variation, associated with clinical origin in *S. cerevisiae*. Our results support the importance of yeast-hypha morphogenesis for *S. cerevisiae* pathogenicity, in addition to suggesting the involvement of cell wall maintenance and cellular detoxification.

MATERIAL AND METHODS

Yeast strains

Segregants of 44 clinical and 44 non-clinical *S. cerevisiae* isolates of diverse geographical backgrounds and source substrates, including the well known lab strain *S. cerevisiae* S288c, were used in this study (see Table S1). Segregants were obtained through sporulation of *S. cerevisiae* isolates, dissection of the spore tetrads and germination of the ascospores using standard methods described by Sherman (1991). All strains were kept in glycerol (15% v/v) at −80 °C, or for shortterm storage on 1% yeast extract, 2% peptone and 2% dextrose (YPD) agar medium at 4 °C. All strains have been deposited in, and should be requested from, the Phaff Yeast Culture Collection (http://www.phaffcollection.org/).

DNA extraction and microarray hybridization

Total DNA was extracted from yeast cultures grown overnight in 50 mL YPD medium using the Qiagen Genomic-tip 100/G (Qiagen) according to the manufacturer's instructions. Ten micrograms of purified DNA were digested with 1 U of DNase I (New England Biolabs) for 2 min at 37 °C in 1× DNase I Reaction buffer (New England Biolabs) to obtain fragments of 25-50 bp. DNase I was heat inactivated at 95 °C for 20 min, and the digestion products were analyzed on a 2% agarose gel. The fragmented DNA was end-labeled by incubation at 37 °C for 1 h with 20 U of terminal deoxynucleotidyl transferase (New England Biolabs) and 1 nmol of biotin-11-ddATP (Perkin Elmer) in 1× NEBuffer 4 (New England Biolabs), and the labeling reaction was terminated by heat inactivation of the terminal transferase at 75 °C for 25 min. The target DNA was hybridized onto GeneChip® *S. cerevisiae* Tiling 1.0R Arrays (Affymetrix) following standard Affymetrix protocols for DNA hybridization, washing and staining (see Gresham *et al*., 2006), and the arrays were scanned using the Affymetrix scanner at 0.7 μm resolution. The hybridization intensities of the 9 central pixels were determined, and an average intensity at each oligonucleotide feature was computed using the GeneChip® Operating Software (Affymetrix).

Data analysis

The raw hybridization intensity data of the perfect match (PM) oligonucleotide features were extracted from the individual ".cel" files, background-corrected with the RMA algorithm (Irizarry *et al*., 2003), quantile-normalized across microarrays (Bolstad *et al*., 2003) and $\log_2$-transformed using the AROMA.AFFYMETRIX v1.1.0 package (Bengtsson *et al*., 2008) in R v2.9.0 (R Development Core Team, 2007). Single feature polymorphisms (SFPs) were identified based on the bimodality of their respective transformed hybridization intensity distributions. Modalities of the hybridization intensity distributions were estimated and individual observations were assigned to genotype classes using a normal mixture modeling and model-based clustering approach, as implemented in the MCLUST v3 package for R (Fraley & Raftery, 2006). Model-based clustering was performed according to the normal mixture model with the highest Bayesian Information Criterion-value (BIC; see e.g. Fraley & Raftery, 2002) among the equal variance-models with up to two mixture components. Analysis was restricted to PM features with a unique match in the *S. cerevisiae* S288c reference genome (Goffeau *et al*., 1996) and in order to reduce the number of false genotype calls, data of predicted polymorphic PM features were only withheld for further analysis if the mean transformed hybridization intensities of the different genotype classes differed by at least 4 and the rare genotype class was comprised of at least 5 yeast strains (minor allele frequency (MAF) ≥ 5.6%). In all subsequent analyses, the two recognized genotype classes were considered "S288c-like" (high hybridization intensity class) and "non S288c-like" (low hybridization intensity class).

Pairwise linkage disequilibria (LD) between the SFPs were measured as r^2 using PLINK v1.07 (Purcell *et al*., 2007) and heatmaps were generated with the LDHEATMAP package (Shin *et al*., 2006) in R v2.9.0. Genotype data from a subset of 5,000 SFPs that appear in linkage equilibrium were selected (PLINK parameters: pairwise $r^2 \leq 0.5$, window size = 50 loci, step size = 5 loci) and used to assess the genetic relationships among the different yeast strains. Classical multidimensional scaling was performed using the CMDSCALE routine in R v2.9.0 with a genetic distance matrix based on the identity-by-state (IBS) estimates between all pairs of yeast strains. In addition, genetic structure was evaluated using the Bayesian approach implemented in STRUCTURE v2.3.2 (Pritchard *et al*., 2000; Falush *et al*., 2003; Hubisz *et al*., 2009) to take the clinical/non-clinical origin of the samples into account. The LOCPRIOR model (Hubisz *et al*., 2009) with admixture and correlated allele frequencies was used and three independent runs for each value of *K* (the number of clusters) between 1 and 5 were performed, with 100,000 iterations after a burn-in period of

50,000 iterations. Using ARLEQUIN v3.5.1.2 (Excoffier & Lischer, 2010), analysis of molecular variance (AMOVA; Excoffier *et al*., 1992) was performed in order to estimate the proportions of the total genetic variance between and within the groups of clinical and nonclinical yeast strains.

Genome-wide association mapping was performed using the non-parametric Fisher exact test and the parametric linear mixed model implemented in the method of Efficient Mixed-Model Association Expedited (EMMAX; Kang *et al*., 2010) to identify genetic loci associated with clinical background in *S. cerevisiae*. Fisher exact tests of association between background and single markers were carried out in R v2.9.0. A log quantile-quantile (Q-Q) plot was generated by plotting the negative logarithm (base 10) of the observed *P*-values against $-\log_{10}(r/(n+1))$, with r equal to the rank of the $n = 135{,}771$ *P*-values, and inflation of the observed *P*-values was measured as the slope of the linear regression between the log-transformed quantiles. The method of Efficient Mixed-Model Association (EMMA; Kang *et al*., 2008) as implemented in EMMAX was used to account for patterns of genetic relatedness between the yeast strains. In the applied model, a random genetic term with a variance-covariance matrix proportional to the kinship matrix based on pairwise IBS estimates was included together with the SFPs being tested as fixed terms. Inflation of the observed *P*-values was estimated as described for the Fisher exact tests. Genome-wide significance was assessed using a nominal 5% significance treshold with Bonferroni correction for 135,771 tests ($P\text{-value} \leq 0.05 / 135{,}771 = 3.68 \times 10^{-7}$).

Verification of DNA polymorphisms

Underlying DNA polymorphisms of SFPs 47,217, 58,321, 79,325, 79,327, 79,345, 79,346, 94,444, 94,445, 94,446, 94,447, 94,468, 94,469, 120,072, 120,073, 120,120 and 120,121, which were found to be associated with clinical origin in *S. cerevisiae* (see Results), were verified using DNA sequences of *S. cerevisiae* strains S288c, YJM789 (YJM145), YJM975 and RM11-1a (YJM1293) retrieved from the NCBI database (National Center for Biotechnology Service; http://www.ncbi.nlm.nih.gov/). SFPs 65,184, 65,185, 65,186 and 65,201 were verified using *S. cerevisiae* S288c DNA sequences and sequences obtained by polymerase chain reaction (PCR) amplification and subsequent Sanger sequencing of the corresponding genetic loci in *S. cerevisiae* strain YJM1199. Oligonucleotide primers were designed based on the *S. cerevisiae* S288c reference sequence using PRIMER3 v0.4.0 software and PCRs were performed in total volumes of 20 μL containing 0.5 mM of the forward and reverse primers (see Table S2), 1U Taq DNA polymerase, 10mM Tris-HCl, pH 9, 15mM $MgCl_2$, 50mM KCl, 200 mM of all four dNTPs and approximately 2 ng of genomic DNA template. The following PCR temperature profile was used: initial denaturation at 95 °C for 2 min; 30 cycles of 95 °C for 20 s, 55 °C for 20 s, 72 °C for 1 min; final extension at 72 °C for 10 min. Direct sequencing of the PCR products was performed using the ABI BigDye® Terminator reaction mixture and an ABI PRISM® 3730 DNA Analyzer (Applied Biosystems) following the manufacturer's instructions.

RESULTS

Genomic DNA extracted from segregants of 44 clinical and 44 non-clinical *S. cerevisiae* isolates was hybridized onto separate GeneChip® *S. cerevisiae* Tiling 1.0R Arrays, which carry 2,470,820 PM oligonucleotide features with a unique match in the *S. cerevisiae* S288c reference genome. Combined, the unique microarray features cover approximately 93% of the *S. cerevisiae* S288c nuclear genome. After background correction and normalization, a total of 135,771 oligonucleotide features (5.5%) were found to cover polymorphic loci in the *S. cerevisiae* nuclear genome based on the bimodality of the distributions of their $\log_2$-transformed hybridization intensity values and a minor allele frequency cut-off value of 5.6%. The fraction of polymorphic oligonucleotide features per chromosome varied between

4.1% and 11.5% (average of 6.3%), with relatively high frequencies of SFPs for the three smallest chromosomes (chromosome 1, 11.5%, chromosome 3, 10.8%, and chromosome 6, 11.4%; see Table 1) due to the occurrence of large sequence polymorphisms. The distance between adjacent polymorphic features (calculated using the position of the middle nucleotide of each oligonucleotide feature in the *S. cerevisiae* S288c reference genome) varied between 4 bp and 16,018 bp, with an average of 89 bp (see Figure S1 for the genomic distribution of SFPs).

Based on the hybridization intensity, two allele classes were recognized for each SFP: “S288c-like” (high hybridization intensity class) and “non S288c-like” (low hybridization intensity class). Per SFP marker, the frequency of the “S288c-like” allele class varied between 5.7% and 94.3% (average of 75.7%) across all yeast strains (see Figure S2a). Per strain, the frequency of the “S288c-like” allele class varied between 59.4% and 99.9% (average of 75.7%) across all SFPs (see Figure S2b), with the highest frequency observed for strain S288c. Using the hybridization intensity data of strain YJM145 and the genome sequence of its haploid isoform YJM789 (Wei *et al*., 2007), the frequency of correctly called genotypes was estimated to be 94.6%. Of the 94,125 SFP markers covering loci with an “S288c-like” allele in YJM145, 94,061 (99.9%) allowed correct genotype calling, while 34,403 of the 41,646 SFP markers covering loci with “non S288c-like” alleles (82.6%) provided correct genotype calls.

The results of the classical multidimensional scaling ordination are presented in Figure 1 as a plane projection of the two most informative axes accounting for the genetic relatedness among the segregants of the clinical and non-clinical *S. cerevisiae* isolates. Combined, the first two factors accounted for 31% of the total variation in the sample. Although strongly differentiated clusters of yeast strains are not apparent, there seems to be some genetic differentiation between the groups of clinical and non-clinical *S. cerevisiae* strains. This was confirmed by the analysis of molecular variance, which indicated a limited (2.7%) but significant (*P*-value < 0.05) proportion of the total variance to be due to differences between these two groups of yeast strains (see Table 2). However, the bayesian analysis of genetic structure, as implemented in STRUCTURE v2.3.2, provided no evidence for clustering according to clinical/non-clinical origin and seemed to support the absence of a clear genetic structure among the yeast strains, assigning relatively few strains to single clusters and indicating most strains to be of mixed ancestry (results not shown).

Figure 2 provides an overview of the results of the genome-wide association analysis using Fisher exact tests (Figure 2a) and EMMAX (Figure 2b). Magnifications of the genomic regions surrounding the SFPs giving the strongest association signals (see Table 3), including heatmaps of pairwise linkage disequilibria, are presented in Figure 3. Before correcting the *P*-values for inflation due to population structure, Fisher exact tests and EMMAX indicated 20 SFPs to reach genomewide significance. Twelve of these corresponded to sharply defined peaks within single genes and eight are located in intergenic regions. SFP 47,217 (*P*-value = 3.16×10^{-7}, OR = ∞) is located in a small LD block of four partially overlapping SFPs on chromosome 6 in between open reading frames (ORFs) *YFR012W* and *YFR012W-A* (see Figure 3a). SFP 58,321 (*P*-value = 1.24×10^{-7}, OR = 6.04) on chromosome 7 overlaps with the heat shock transcription factor Hsf1p binding site between *YGR146C-A* and *YGR147C* (see Figure 3b). The partially overlapping SFPs 65,184, 65,185 and 65,186 (*P*-values ≤ 4.04×10^{-8}, ORs = ∞) are located within *YHR102W* (*KIC1*) on chromosome 8 and appear in LD with SFP 65,201 (*P*-value = 1.15×10^{-7}, OR = ∞) located within *YHR103W* (*SBE22*; see Figure 3c). SFPs 79,325 and 79,327 (*P*-values = 5.01 × 10^{-8}, ORs = 0.00), together with SFPs 79,345 and 79,346 (*P*-values = 1.36×10^{-12}, ORs = 0.00), are located within a 1.6 kbp LD block containing *YJR153W* (*PGU1*) on chromosome 10 (see Figure 3d). The LD block of SFPs 94,444, 94,445, 94,446 and 94,447

(*P*-values = 2.50×10^{-8}, ORs = 7.71) is located within *YLR337C* (*VRP1*) on chromosome 12 and is in LD with SFPs 94,468 and 94,469 (*P*-values = 2.50×10^{-8}, ORs = 7.71) located between *YLR340W* (*RPP0*) and *YLR341W* (*SPO77*; see Figure 3e). Finally, on chromosome 15, SFPs 120,072 and 120,073 (*P*-values = 3.16×10^{-7}, ORs = ∞) are located within a 2.2 kbp LD block partially overlapping *YOR153W* (*PDR5*) and SFPs 120,120 and 120,121 (*P*-values = 3.55×10^{-7}, ORs = 15.30) are located within *YOR154W* (*SLP1*; see Figure 3f).

However, the log Q-Q plots of the observed *P*-values (see Figure 4) indicated systematic deviations of the observed *P*-value distributions from their null expectations under a model of no association and an excess of significant associations is expected based on the estimations of the inflation factors for the *P*-values generated by the Fisher exact tests ($\lambda_{\text{fisher}} = 1.67$; see Figure 4a) and by EMMAX ($\lambda_{\text{EMMAX}} = 1.15$; see Figure 4b). After correcting the observed *P*-values using the estimated inflation factors ($P_{\text{corrected}} = P_{\text{observed}}^{1/\lambda}$), EMMAX identified significant association at a genome-wide level of SFPs 79,345 and 79,346 (*P*-values = 5.13×10^{-11}) and SFPs 94,444, 94,445, 94,446, 94,447, 94,468 and 94,469 (*P*-values = 2.56×10^{-7}) with clinical origin in *S. cerevisiae* (see Figures 2 and 3, and Table 3).

The pseudoheritability, the additive component of heritability as estimated with the kinship matrix by EMMAX, of the clinical background in *S. cerevisiae* is 99%. Including the 20 SFPs that reach genome-wide significance before correction of the *P*-values for inflation due to population structure in the AMOVA, these SFPs explain 0.5% of the total genetic variance and 4.7% of the variance due to differences between clinical and non-clinical *S. cerevisiae* strains (results not shown). Among the 20 significantly associated SFPs, linkage disequilibrium was mostly restricted to the SFPs located on the same chromosome (see Table 4).

All reported SFPs associated with clinical origin in *S. cerevisiae* cover at least one single nucleotide polymorphism (SNP) or sequence polymorphism spanning two or three nucleotides (see Figure S3). The majority of the SNPs were found to be transitions and the polymorphisms located within coding sequences were mostly synonymous polymorphisms. Nevertheless, SFPs were found to cover non-synonymous nucleotide polymorphisms in the coding sequences of *YLR337C* (*VRP1*; D159I and I160V; see Figure S3f) and *YOR153W* (*PDR5*; T1444A and D1447G; see Figure S3h).

DISCUSSION

Treating an environmental variable as a phenotype and applying similar methods as those developed for mapping phenotypes is a potentially powerful approach to identify loci under selection, and can provide valuable insights into the genetic and ecological bases of natural selection (Coop *et al*., 2010). Unusually strong correlations between allele frequencies and environmental variables may indicate the corresponding loci, assuming they affect relevant phenotypes, to be under selection by those environmental factors or correlated ecological variables. Although this strategy often fails at identifying causal selection pressures, as many environmental and ecological factors covary, it has the advantage that the involvement of specific phenotypes is not assumed *a priori* and thus allows for a more exploratory approach. In the presented population-based case-control study, we performed a genome-wide environmental association analysis of clinical versus non-clinical origin in *Saccharomyces cerevisiae*. Assuming the distribution of genetic variants related to pathogenicity in *S. cerevisiae* differs between clinical and non-clinical strains, we aimed at identifying significant correlations between allele frequencies and clinical origin to gain insight into the virulence factors and virulence-related traits that contribute to *S. cerevisiae* pathogenesis.

Genotypes of 44 clinical and 44 non-clinical *S. cerevisiae* strains were obtained using the high-density Affymetrix GeneChip® *S. cerevisiae* Tiling 1.0R Array and relatively constant levels of diversity, detected as single microarray feature polymorphisms, were revealed across the genome. Genetic structure in our sample of yeast strains was largely independent of geographical origin. This is in agreement with our earlier work, which included 63 of the 88 *S. cerevisiae* isolates used in the present study and indicated high levels of gene flow between geographical locations (Muller & McCusker, 2009), as well as with studies by Liti *et al*. (2009) and Schacherer *et al*. (2009), which revealed relatively low differentiation among worldwide *S. cerevisiae* isolates and suggested ecological rather than geographical differentiation. Consistent with ecological differentiation and as reported by Malgoire *et al*. (2005) and Klingberg *et al*. (2008), we found limited but significant genetic differentiation between the clinical and the non-clinical *S. cerevisiae* strains. Nevertheless, patterns of genetic relatedness among the *S. cerevisiae* strains in our sample are complex and confounding of the results of the association analysis due to this population structure could be assumed from the deviations of the *P*-value distributions from their null expectations. Although the mixed-model approach implemented in EMMAX, which takes genetic relatedness among samples into account, reduced the number of associations of low-to-moderate significance in comparison with the Fisher exact tests, significant inflation of *P*-values remained and warranted correction.

Even if the limited numbers of clinical and non-clinical strains resulted in low statistical power, which may have been further reduced by the possible ability of some of the non-clinical strains to establish infection when given the opportunity (see e.g. Wheeler *et al*., 2003), we were able to identify several common variants associated with a clinical origin in *S. cerevisiae*. Some of the identified polymorphisms indicate genes that consitute known virulence factors, most notably yeast-pseudohyphal morphogenesis, or fitness attributes that have been shown to contribute to fungal pathogenicity, such as a heat stress response and cellular detoxification, to be involved in *S. cerevisiae* pathogenesis.

After applying the correction for confounding due to population structure, EMMAX identified three polymorphisms to be associated with clinical background in *S. cerevisiae* at a genome-wide significance level and two of these suggested single genes to be involved: *PGU1* and *VRP1*. The *PGU1* gene encodes for a secreted endo-polygalacturonase, a cell wall-degrading pectinase that is induced during invasive growth by the Kss1p mitogen-activated protein kinase (MAPK) signaling pathway (Madhani *et al*., 1999; Wong Sak Hoi & Dumas, 2010). Expression of *PGU1* is regulated by the Ste12p transcription factor, which also targets *FLO11*. Although *FLO11* encodes a cell surface-protein that promotes yeast flocculation and pseudohyphal growth directly, *PGU1* is not required for morphogenesis but may play a role in processes related to filamentous growth (Gancedo, 2001; Gognies *et al*., 2006). Alternatively, the significant association of *PGU1* with clinical origin may indicate plant associated *S. cerevisiae* strains to be an important source of human infections as pectinases are proven virulence factors of plant pathogens (Reignault *et al*., 2008).

VRP1, on the other hand, which encodes an actin-binding protein essential for correct polarization of actin patches during cell growth, has been shown to be directly involved in pseudohyphal morphogenesis in *S. cerevisiae* and *Candida albicans* (Wu & Jiang, 2005; Borth *et al*., 2010). Noteworthy thereby is the physical interaction between Vrp1p and End3p (Tarassov *et al*., 2008), a protein involved in cytoskeletal organization and cell wall morphogenesis known to affect high temperature growth in *S. cerevisiae* (Sinha *et al*., 2008).

The functional relevance of the third polymorphism indicated to be associated with clinical origin in *S. cerevisiae* by EMMAX, found in the intergenic region between *RPP0* and

SPO77, is less clear. While *SPO77* is involved in spore wall formation during sporulation (Coluccio *et al*., 2004), *RPP0* apparently affects invasive growth (Jin *et al*., 2008). However, the identified locus is not linked to any other polymorphism covering the *RPP0* ORF. Instead, it is tightly linked to the *VRP1* polymorphism, rendering significance of *RPP0* questionable.

Applying a correction for confounding due to population structure using the inflation factor resulted in several associations identified by Fisher exact tests and EMMAX becoming nonsignificant at the genome-wide level. However, considering the fact that Q-Q plots can overestimate the impact of population structure (see e.g. Voorman *et al*., 2011), the large differences in minor allele frequency between the clinical and the non-clinical yeast strains and the functions of the identified genes, we do believe there is real information in these results. Among these associations are polymorphisms within the genes *KIC1*, *SBE22*, *PDR5* and *SLP1*, and polymorphisms in the intergenic regions between *YGR146C-A* and *NAT2*, and between *YFR012W* and *YFR012W-A*. *KIC1* encodes a protein kinase that is part of the RAM (Regulation of Ace2p activity and cellular Morphogenesis) signaling pathway and plays a role in maintaining cell wall integrity (Vink *et al*., 2002). Mutations in *KIC1* have been shown to cause cell separation defects and loss of polarity in *S. cerevisiae* (Nelson *et al*., 2003), and deletion of *KIC1* in *C. albicans* resulted in defective mouse infectivity and hyphal formation (Noble *et al*., 2010). *SBE22* is involved in the transport of cell wall components from the Golgi apparatus to the cell surface and deletion of *SBE22* results in the formation of an abnormal cell wall structure with a reduced mannoprotein layer (Santos & Snyder, 2000). Amongst others, Sbe22p is involved in the transport of Crh2p (Rodriguez-Peña *et al*., 2002), a chitin transglycosylase encoded by a member of the *CRH* family which has been shown to be critical for virulence of *C. albicans* in an animal model of systemic infection (Pardini *et al*., 2006). *PDR5* encodes for an ATP-binding cassette (ABC) transporter, which is involved in the pleiotropic drug response and mediates resistance to fungicides and other xenobiotic compounds (Ernst *et al*., 2005).

While *SLP1* encodes for an integral membrane protein of unknown function, its relevance for *S. cerevisiae* pathogenesis is unclear. This is also the case for the polymorphism located in the 5′ upstream region of *YFR013W* (*IOC3*), in between the uncharacterized ORFs *YFR012W* and *YFR012W-A*. However, the polymorphism identified in between *YGR146C-A* and *NAT2* flanks a binding site for the heat shock transcription factor Hsf1p located in the 5′ upstream region of *YGR146C* (*ECL1*). Hsf1p, which activates transcription of at least 59 genes upon heat shock (Yamamoto *et al*., 2005), is believed to regulate stress-induced cell wall remodeling (Imazu & Sakurai, 2005) and has been shown to contribute significantly to the virulence of *C. albicans* (Nicholls *et al*., 2011). Hsf1p binds to the promotor of *YGR146C* and although the function of Ygr146cp is unclear, transcription of *YGR146C* is induced upon transient cell wall damage (Imazu & Sakurai, 2005).

In all cases, large differences in allele frequency distributions were observed between the clinical and the non-clinical yeast groups, with the “non S288c-like” allele class usually being more prevalent among the clinical yeast strains. The “S288c-like” allele class was only found to be more common for the *PGU1* associated polymorphism and this may reflect its indirect clinical relevance as the non-clinical *S. cerevisiae* S288c, which is avirulent in mice (Clemons *et al*., 1994), has about 90% of its genome derived from a strain which was originally isolated from rotten fruit (Mortimer & Johnston, 1986). Single nucleotide polymorphisms, both in coding and in regulatory non-coding regions, most often underlied the observed microarray feature polymorphisms, and both synonymous and non-synonymous clinical SNP variants were found in coding regions. Whether the identified polymorphisms also constitute functional variation and are not just in linkage disequilibrium with the unidentified causative alleles will have to be determined in experimental studies.

Nevertheless, our results suggest that both coding and regulatory variation are important for pathogenicity in *S. cerevisiae*. In addition, the restriction of pairwise linkage disequilibrium to those polymorphisms located in the same chromosomal region may suggest that a single clinical genetic background is absent and is consistent with the opportunistic nature of *S. cerevisiae* pathogenesis (Klingberg *et al*., 2008; Muller & McCusker, 2009).

In summary, we were able to identify different polymorphisms associated with clinical background in *S. cerevisiae* by performing a population-based genome-wide environmental association analysis with relatively few clinical and non-clinical *S. cerevisiae* isolates and high density, tiling array-based genotypes. Polymorphisms within the coding sequences of *VRP1, KIC1*, *SBE22* and *PDR5*, and the 5′ upstream region of *YGR146C* found to be associated with a clinical origin support the relevance of pseudohyphal formation, robust cell wall maintenance and cellular detoxification for *S. cerevisiae* pathogenesis, and we propose these genes as good candidates for follow-up studies that aim to reveal the virulence and virulence-related factors important for the pathogenicity of *S. cerevisiae*.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The work was supported by the National Institutes of Health grants GM070541 and GM081690 awarded to JHM. The authors wish to thank F. Dietrich and P. Magwene for helpful comments and suggestions, and E. Winzeler, K. Clemons, R. Barton, J. Bakken, L. Hartwell, W. Schell, J. Haber, P. Sniegowski, D. Stevens, K. Clemons, C. Kurtzman and C. Kaufman for generously providing strains and/or isolates.

REFERENCES

Bassetti S, Frei R, Zimmerli W. Fungemia with *Saccharomyces cerevisiae* after treatment with *Saccharomyces boulardii*. American Journal of Medicine. 1998; 105:71–72. [PubMed: 9688023]

Bengtsson, H.; Simpson, K.; Bullard, J.; Hansen, K. Technical Report #745. Department of Statistics, University of California; Berkeley, USA: 2008. Aroma.affymetrix: a generic framework in R for analyzing small to very large Affymetrix data sets in bounded memory.

Bolstad BM, Irizarry RA, Astrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on bias and variance. Bioinformatics. 2003; 19:185–193. [PubMed: 12538238]

Borth N, Walther A, Reijnst P, Jorde S, Schaub Y, Wendland J. *Candida albicans* Vrp1 is required for polarized morphogenesis and interacts with Wal1 and Myo5. Microbiology. 2010; 156:2962–2969. [PubMed: 20656786]

Clemons KV, McCusker JH, Davis RW, Stevens DA. Comparative pathogenesis of clinical and nonclinical isolates of *Saccharomyces cerevisiae*. Journal of Infectious Diseases. 1994; 169:859–67. [PubMed: 8133102]

Coop G, Witonsky D, Di Rienzo A, Pritchard JK. Using environmental correlations to identify loci underlying local adaptation. Genetics. 2010; 185:1411–1423. [PubMed: 20516501]

Enache-Angoulvant A, Hennequin C. Invasive *Saccharomyces* infection: a comprehensive review. Clinical Infectious Diseases. 2005; 41:1559–1568. [PubMed: 16267727]

Ernst R, Klemm R, Schmitt L, Kuchler K. Yeast ATP-binding cassette transporters: cellular cleaning pumps. Methods in Enzymology. 2005; 400:460–484. [PubMed: 16399365]

Excoffier L, Lischer HEL. Arlequin suite ver 3.5: A new series of programs to perform population genetics analyses under Linux and Windows. Molecular Ecology Resources. 2010; 10:564–567. [PubMed: 21565059]

Excoffier L, Smouse PE, Quattro JM. Analysis of molecular variance inferred from metric distances among DNA haplotypes: Application to human mitochondrial DNA restriction data. Genetics. 1992; 131:479–491. [PubMed: 1644282]

Falush D, Stephens M, Pritchard JK. Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. Genetics. 2003; 164:1567–1587. [PubMed: 12930761]

Fraley C, Raftery AE. Model-based clustering, discriminant analysis and density estimation. Journal of the American Statistical Association. 2002; 97:611–631.

Fraley, C.; Raftery, AE. Technical Report #504. Department of Statistics, University of Washington; Seattle, USA: 2006. MCLUST version 3 for R: normal mixture modelling and model-based clustering.

Gancedo JM. Control of pseudohyphae formation in *Saccharomyces cerevisiae*. FEMS Microbiology Reviews. 2001; 25:107–123. [PubMed: 11152942]

Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M, Louis EJ, Mewes HW, Murakami Y, Philippsen P, Tettelin H, Oliver SG. Life with 6000 genes. Science. 1996; 274(546):563–567.

Gognies S, Barka EA, Gainvors-Claisse A, Belarbi A. Interactions between yeasts and grapevines: filamentous growth, endopolygalacturonase and phytopathogenicity of colonizing yeasts. Microbial Ecology. 2006; 51:109–116. [PubMed: 16408245]

Gresham D, Ruderfer DM, Pratt SC, Schacherer J, Dunham M, Botstein D, Kruglyak L. Genome-wide detection of polymorphisms at nucleotide resolution with a single DNA microarray. Science. 2006; 311:1932–1936. [PubMed: 16527929]

Guslandi M. Are probiotics effective for treating *Clostridium difficile* disease and antibiotic-associated diarrhea? Nature Clinical Practice Gastroenterology and Hepatology. 2006; 3:606–607.

de Hoog GS. Risk assessment of fungi reported from humans and animals. Mycoses. 1996; 39:407–417. [PubMed: 9144996]

Hennequin C, Kauffmann-Lacroix C, Jobert A, Viard JP, Ricour C, Jacquemin JL, Berche P. Possible role of catheters in *Saccharomyces boulardii* fungemia. European Journal of Clinical Microbiology & Infectious Diseases. 2000; 19:16–20.

Hoyer LL, Green CB, Oh SH, Zhao X. Discovering the secrets of the *Candida albicans* agglutinin-like sequence (ALS) gene family - a sticky pursuit. Medical Mycology. 2008; 46:1–15. [PubMed: 17852717]

Hube, B.; Naglik, J. Extracellular hydrolases. In: Calderone, RA., editor. Candida and Candidiasis. ASM Press; Washington, DC: 2002. p. 107-122.

Hubisz MJ, Falush D, Stephens M, Pritchard JK. Inferring weak population structure with the assistance of sample group information. Molecular Ecology Resources. 2009; 9:1322–1332. [PubMed: 21564903]

Imazu H, Sakurai H. *Saccharomyces cerevisiae* heat shock transcription factor regulates cell wall remodeling in response to heat shock. Eukaryotic Cell. 2005; 4:1050–1056. [PubMed: 15947197]

Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP. Summaries of Affymetrix GeneChip probe level data. Nucleic Acids Research. 2003; 31:e15. [PubMed: 12582260]

Jin R, Dobry CJ, McCown PJ, Kumar A. Large-scale analysis of yeast filamentous growth by systematic gene disruption and overexpression. Molecular Biology of the Cell. 2008; 19:284–296. [PubMed: 17989363]

Kang HM, Zaitlen NA, Wade CM, Kirby A, Heckerman D, Daly MJ, Eskin E. Efficient control of population structure in model organism association mapping. Genetics. 2008; 178:1709–1723. [PubMed: 18385116]

Kang HM, Sul JH, Service SK, Zaitlen NA, Kong SY, Freimer NB, Sabatti C, Eskin E. Variance component model to account for sample structure in genome-wide association studies. Nature Genetics. 2010; 42:348–354. [PubMed: 20208533]

Klingberg TD, Lesnik U, Arneborg N, Raspor P, Jespersen L. Comparison of *Saccharomyces cerevisiae* strains of clinical and nonclinical origin by molecular typing and determination of putative virulence traits. FEMS Yeast Research. 2008; 8:631–640. [PubMed: 18355272]

Kwon-Chung, K.; Bennett, JE. *Saccharomyces* Meyen Ex Reess. In: Kwon-Chung, JJ.; Bennet, JE., editors. Medical mycology. Lea and Febiger; Philadelphia, PA: 1992. p. 772-773.

de Llanos R, Querol A, Planes AM, Fernandez-Espinar MT. Molecular characterization of clinical *Saccharomyces cerevisiae* isolates and their association with non-clinical strains. Systematic and Applied Microbiology. 2004; 27:427–435. [PubMed: 15368848]

de Llanos R, Querol A, Peman J, Gobernado M, Fernandez-Espinar MT. Food and probiotic strains from the *Saccharomyces cerevisiae* species as a possible origin of human systemic infections. International Journal of Food Microbiology. 2006a; 110:286–290. [PubMed: 16782223]

de Llanos R, Fernández-Espinar MT, Querol A. A comparison of clinical and food *Saccharomyces cerevisiae* isolates on the basis of potential virulence factors. Antonie Van Leeuwenhoek. 2006b; 90:221–231. [PubMed: 16871421]

Liti G, Carter DM, Moses AM, Warringer J, Parts L, James SA, Davey RP, Roberts IN, Burt A, Koufopanou V, Tsai IJ, Bergman CM, Bensasson D, O'Kelly MJ, van Oudenaarden A, Barton DB, Bailes E, Nguyen AN, Jones M, Quail MA, Goodhead I, Sims S, Smith F, Blomberg A, Durbin R, Louis EJ. Population genomics of domestic and wild yeasts. Nature. 2009; 458:337–341. [PubMed: 19212322]

Madhani HD, Galitski T, Lander ES, Fink GR. Effectors of a developmental mitogen-activated protein kinase cascade revealed by expression signatures of signaling mutants. Proceedings of the National Academy of Sciences of the United States of America. 1999; 96:12530–12535. [PubMed: 10535956]

Malgoire JY, Bertout S, Renaud F, Bastide JM, Mallie M. Typing of *Saccharomyces cerevisiae* clinical strains by using microsatellite sequence polymorphism. Journal of Clinical Microbiology. 2005; 43:1133–1137. [PubMed: 15750073]

McCusker JH, Clemons KV, Stevens DA, Davis RW. *Saccharomyces cerevisiae* virulence phenotype as determined with CD-1 mice is associated with the ability to grow at 42 degrees C and form pseudohyphae. Infection and Immunity. 1994; 62:5447–5455. [PubMed: 7960125]

Mortimer RK, Johnston JR. Genealogy of principal strains of the yeast genetic stock center. Genetics. 1986; 113:35–43. [PubMed: 3519363]

Muller LAH, McCusker JH. Microsatellite analysis of genetic diversity among clinical and nonclinical *Saccharomyces cerevisiae* isolates suggests heterozygote advantage in clinical environments. Molecular Ecology. 2009; 18:2779–86. [PubMed: 19457175]

Muñoz P, Bouza E, Cuenca-Estrella M, Eiros JM, Pérez MJ, Sánchez-Somolinos M, Rincón C, Hortal J, Peláez T. *Saccharomyces cerevisiae* fungemia: an emerging infectious disease. Clinical Infectious Diseases. 2005; 40:1625–1634. [PubMed: 15889360]

Nelson B, Kurischko C, Horecka J, Mody M, Nair P, Pratt L, Zougman A, McBroom LD, Hughes TR, Boone C, Luca FC. RAM: a conserved signaling network that regulates Ace2p transcriptional activity and polarized morphogenesis. Molecular Biology of the Cell. 2003; 14:3782–3803. [PubMed: 12972564]

Nicholls S, MacCallum DM, Kaffarnik FA, Selway L, Peck SC, Brown AJ. Activation of the heat shock transcription factor Hsf1 is essential for the full virulence of the fungal pathogen *Candida albicans*. Fungal Genetics and Biology. 2011; 48:297–305. [PubMed: 20817114]

Noble SM, French S, Kohn LA, Chen V, Johnson AD. Systematic screens of a *Candida albicans* homozygous deletion library decouple morphogenetic switching and pathogenicity. Nature Genetics. 2010; 42:590–598. [PubMed: 20543849]

Pardini G, De Groot PW, Coste AT, Karababa M, Klis FM, de Koster CG, Sanglard D. The *CRH* family coding for cell wall glycosylphosphatidylinositol proteins with a predicted transglycosidase domain affects cell wall organization and virulence of *Candida albicans*. Journal of Biological Chemistry. 2006; 281:40399–40411. [PubMed: 17074760]

Phan QT, Myers CL, Fu Y, Sheppard DC, Yeaman MR, Welch WH, Ibrahim AS, Edwards JE Jr, Filler SG. Als3 is a *Candida albicans* invasin that binds to cadherins and induces endocytosis by host cells. PLoS Biology. 2007; 5:e64. [PubMed: 17311474]

Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics. 2000; 155:945–959. [PubMed: 10835412]

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC. PLINK: a toolset for whole-genome association and population-based linkage analysis. American Journal of Human Genetics. 2007; 81:559–575. [PubMed: 17701901]

R Development Core Team. R: a language and environment for statistical computing. R Foundation for Statistical Computing; Vienna: 2007. ISBN 3-900051-07-0

Reignault P, Valette-Collet O, Boccara M. The importance of fungal pectinolytic enzymes in plant invasion, host adaptability and symptom type. European Journal of Plant Pathology. 2008; 120:1–11.

Rodriguez-Peña JM, Rodriguez C, Alvarez A, Nombela C, Arroyo J. Mechanisms for targeting of the *Saccharomyces cerevisiae* GPI-anchored cell wall protein Crh2p to polarised growth sites. Journal of Cell Science. 2002; 115:2549–2558. [PubMed: 12045225]

Salonen JH, Richardson MD, Gallacher K, Issakainen J, Helenius H, Lehtonen OP, Nikoskelainen J. Fungal colonization of haematological patients receiving cytotoxic chemotherapy: emergence of azole-resistant *Saccharomyces cerevisiae*. Journal of Hospital Infection. 2000; 45:293–301. [PubMed: 10973747]

Santos B, Snyder M. Sbe2p and sbe22p, two homologous Golgi proteins involved in yeast cell wall formation. Molecular Biology of the Cell. 2000; 11:435–452. [PubMed: 10679005]

Schacherer J, Shapiro JA, Ruderfer DM, Kruglyak L. Comprehensive polymorphism survey elucidates population structure of *Saccharomyces cerevisiae*. Nature. 2009; 458:342–345. [PubMed: 19212320]

Sherman, F. Getting started with yeast. In: Guthrie, C.; Fink, GR., editors. Methods in Enzymology. Academic Press; New York, NY: 1991. p. 3-21.

Shin JH, Blay S, McNeney B, Graham J. LDheatmap: An R function for graphical display of pairwise linkage disequilibria between single nucleotide polymorphisms. Journal of Statistical Software. 2006; 16 Code Snippet 3.

Sinha H, David L, Pascon RC, Clauder-Münster S, Krishnakumar S, Nguyen M, Shi G, Dean J, Davis RW, Oefner PJ, McCusker JH, Steinmetz LM. Sequential elimination of major-effect contributors identifies additional quantitative trait loci conditioning high-temperature growth in yeast. Genetics. 2008; 180:1661–1670. [PubMed: 18780730]

Smith D, Metzgar D, Wills C, Fierer J. Fatal *Saccharomyces cerevisiae* aortic graft infection. Journal of Clinical Microbiology. 2002; 40:2691–2692. [PubMed: 12089311]

Sundstrom, P. *Candida albicans* hypha formation and virulence. In: Heitman, J.; Filler, SG.; Edwards, JE.; Mitchell, AP., editors. Molecular Principles of Fungal Pathogenesis. ASM Press; Washington, DC: 2006. p. 45-47.

Tarassov K, Messier V, Landry CR, Radinovic S, Serna Molina MM, Shames I, Malitskaya Y, Vogel J, Bussey H, Michnick SW. An in vivo map of the yeast protein interactome. Science. 2008; 320:1465–1470. [PubMed: 18467557]

Vink E, Vossen JH, Ram AF, van den Ende H, Brekelmans S, de Nobel H, Klis FM. The protein kinase Kic1 affects 1,6-beta-glucan levels in the cell wall of *Saccharomyces cerevisiae*. Microbiology. 2002; 148:4035–4048. [PubMed: 12480907]

Voorman A, Lumley T, McKnight B, Rice K. Behavior of QQ-plots and genomic control in studies of gene-environment interaction. PLoS One. 2011; 6:e19416. [PubMed: 21589913]

Wei W, McCusker JH, Hyman RW, Jones T, Ning Y, Cao Z, Gu Z, Bruno D, Miranda M, Nguyen M, Wilhelmy J, Komp C, Tamse R, Wang X, Jia P, Luedi P, Oefner PJ, David L, Dietrich FS, Li Y, Davis RW, Steinmetz LM. Genome sequencing and comparative analysis of *Saccharomyces cerevisiae* strain YJM789. Proceedings of the National Academy of Sciences of the United States of America. 2007; 104:12825–12830. [PubMed: 17652520]

Wheeler RT, Kupiec M, Magnelli P, Abeijon C, Fink GR. A *Saccharomyces cerevisiae* mutant with increased virulence. Proceedings of the National Academy of Sciences of the United States of America. 2003; 100:2766–2770. [PubMed: 12589024]

Wong Sak Hoi J, Dumas B. Ste12 and Ste12-like proteins, fungal transcription factors regulating development and pathogenicity. Eukaryotic Cell. 2010; 9:480–485. [PubMed: 20139240]

Wu X, Jiang YW. Genetic/genomic evidence for a key role of polarized endocytosis in filamentous differentiation of *S. cerevisiae*. Yeast. 2005; 22:1143–1153. [PubMed: 16240455]

Yamamoto A, Mizukami Y, Sakurai H. Identification of a novel class of target genes and a novel type of binding sequence of heat shock transcription factor in *Saccharomyces cerevisiae*. Journal of Biological Chemistry. 2005; 280:11911–11919. [PubMed: 15647283]

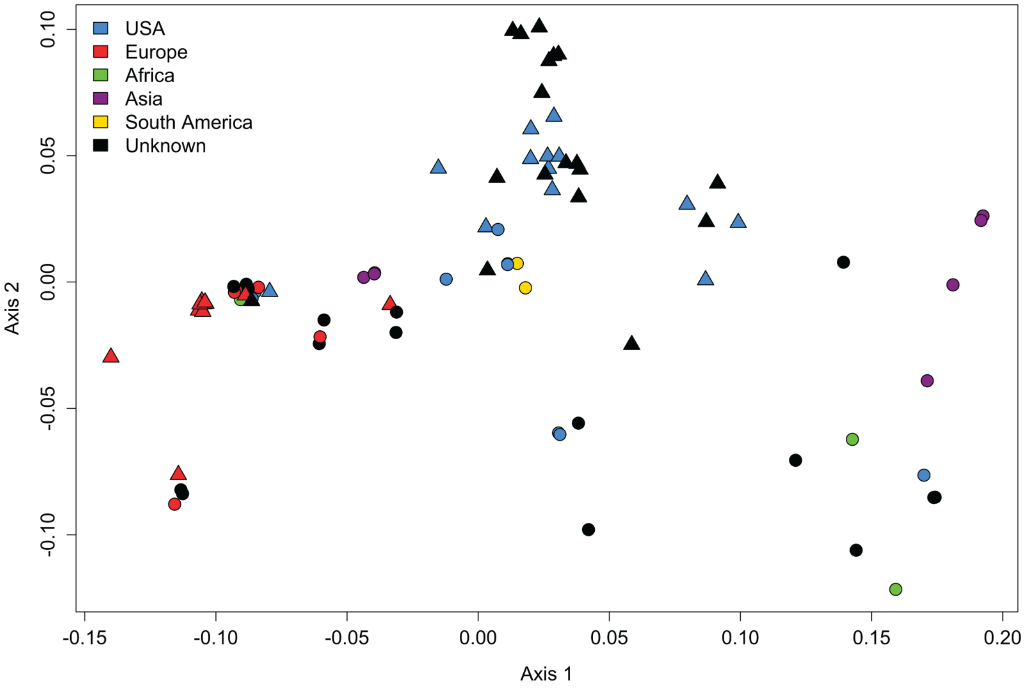


Figure 1.

Results of a classical multidimensional scaling analysis performed using the cmdscale routine in R v2.9.0 (R Development Core Team, 2007) with a genetic distance matrix based on the pairwise identity-by-state (IBS) estimates, calculated using genotype data from 5,000 GeneChip® *S. cerevisiae* Tiling 1.0R single feature polymorphisms in approximate linkage equilibrium, among the segregants of 88 clinical and non-clinical *Saccharomyces cerevisiae* isolates; filled circles, nonclinical *S. cerevisiae* strains; filled triangles, clinical *S. cerevisiae* strains.

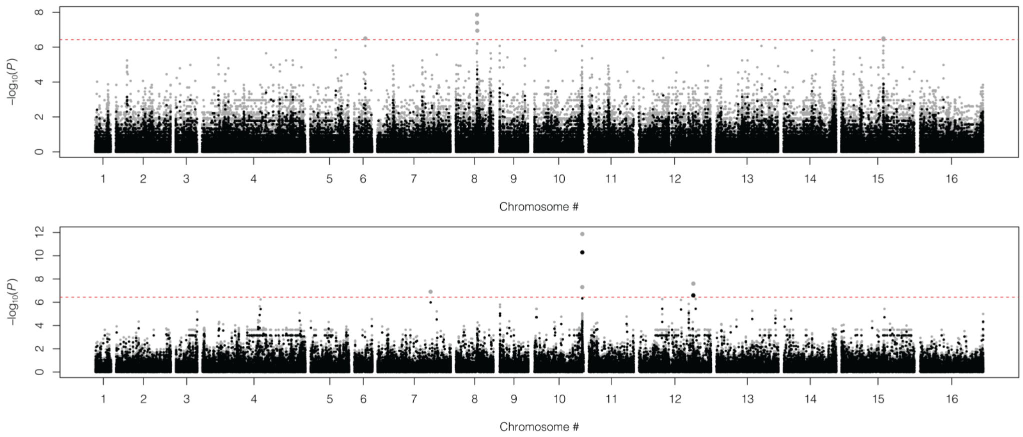
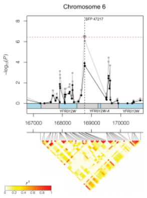
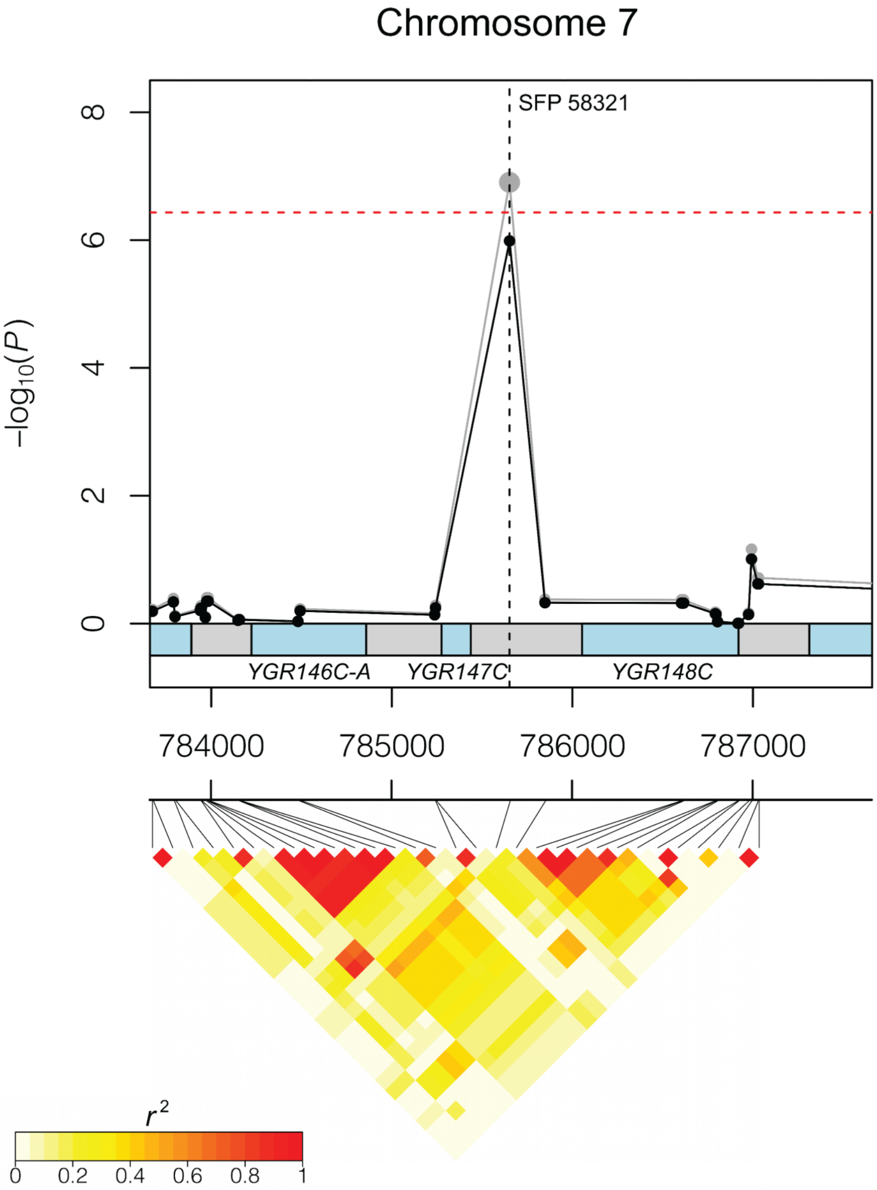
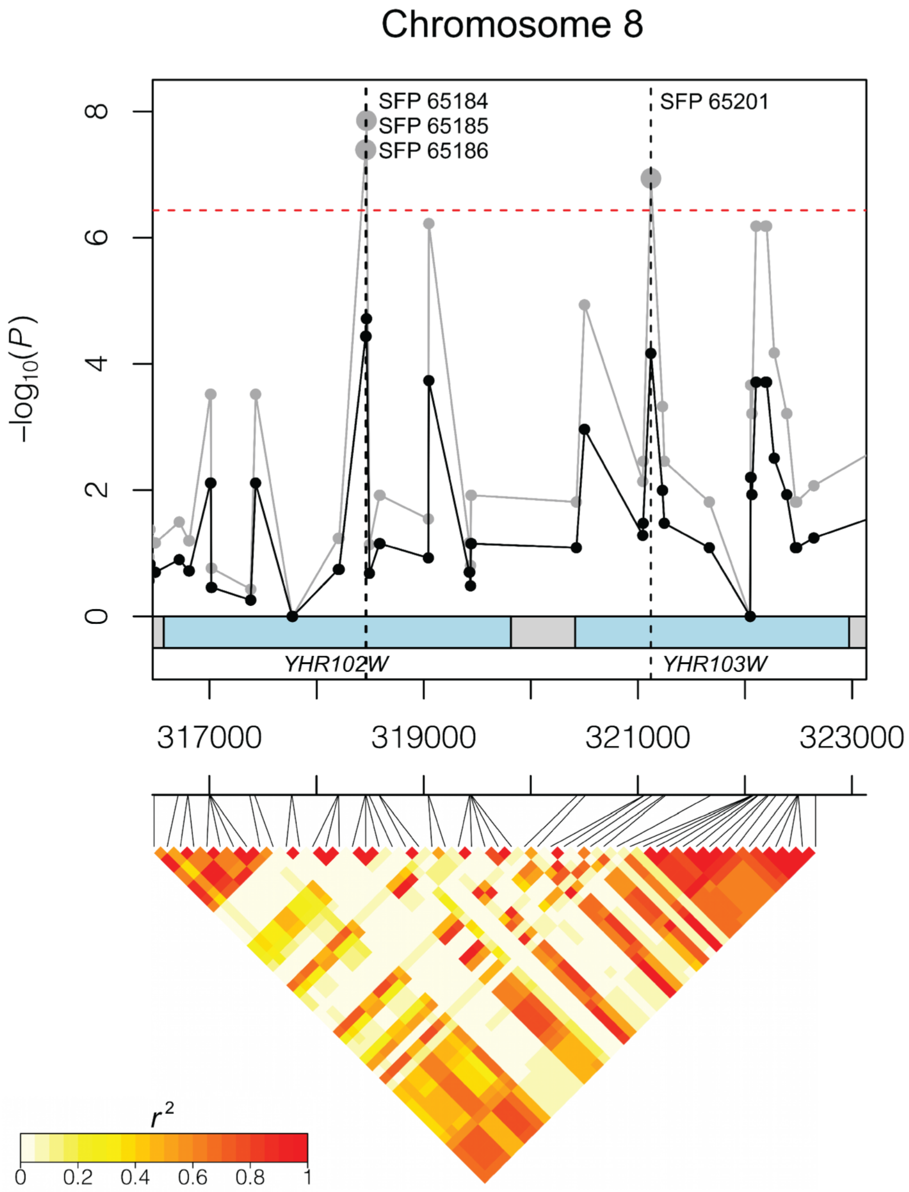


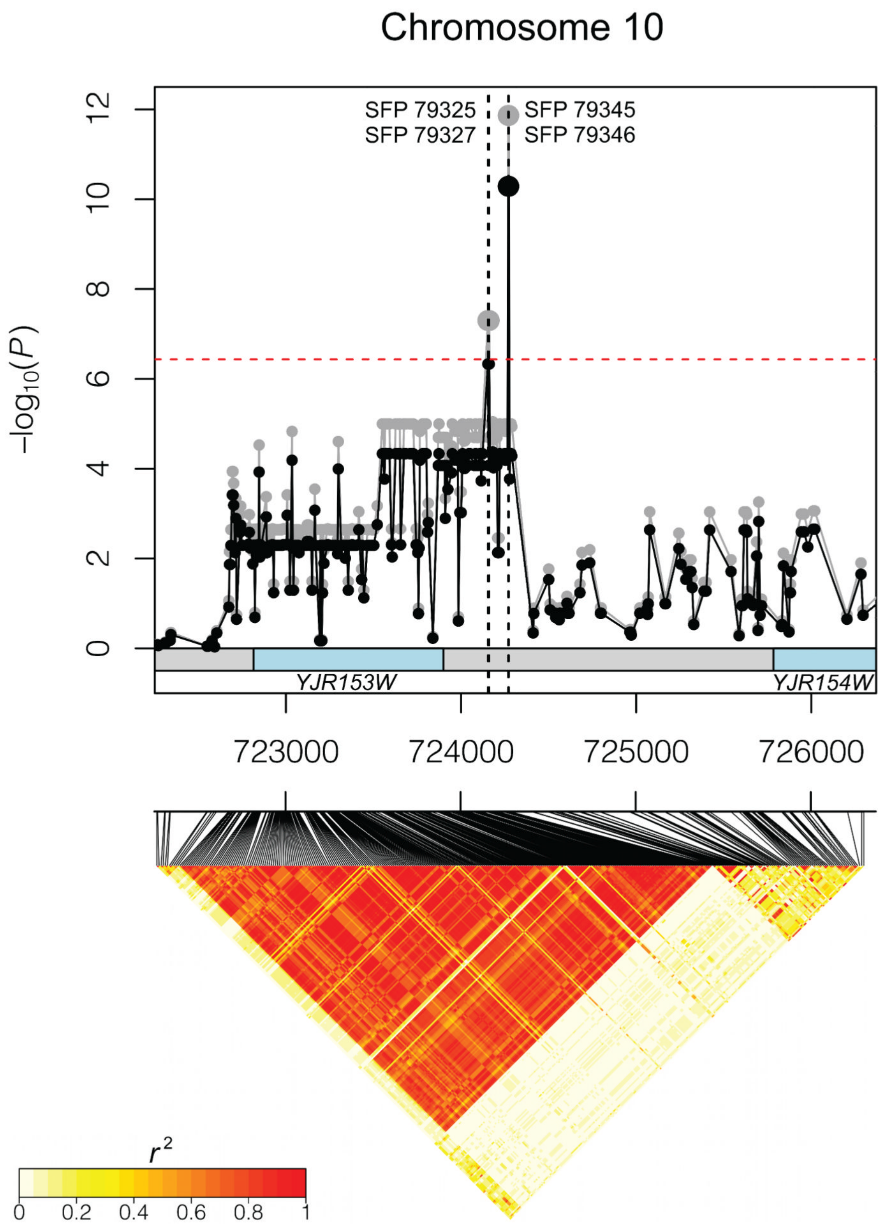
Figure 2.

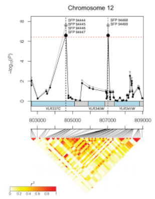
Manhattan plots illustrating the results of a genome-wide association analysis using Fisher exact tests (a) and Efficient Mixed-Model Association Expedited (EMMAX) (b) of clinical origin in *Saccharomyces cerevisiae*, with the x-axis representing genomic position and the y-axis indicating $-\log_{10}(P)$ before (grey) and after (black) correction for confounding due to population structure by division with the inflation factor λ ($\lambda_{\mathrm{Fisher}} = 1.67$ and $\lambda_{\mathrm{EMMAX}} = 1.15$); the dashed line corresponds to a nominal 5% significance threshold with Bonferroni correction for 135,771 tests and bold circles indicate a genome-wide significant association (P-value $\leq 3.68 \times 10^{-7}$).


Chromosome 6
SFP 47217
$-\log_{10}(P)$
0
2
4
6
8
YFR012W
YFR012W-A
YFR013W
167000
168000
169000
170000
r^2
0
0.2
0.4
0.6
0.8
1


Chromosome 7
SFP 58321
$-\log_{10}(P)$
0
2
4
6
8
YGR146C-A
YGR147C
YGR148C
784000
785000
786000
787000
r^2
0
0.2
0.4
0.6
0.8
1


Chromosome 8
SFP 65184
SFP 65185
SFP 65186
SFP 65201
$-\log_{10}(P)$
0
2
4
6
8
YHR102W
YHR103W
317000
319000
321000
323000
r^2
0
0.2
0.4
0.6
0.8
1


Chromosome 10
SFP 79325
SFP 79327
SFP 79345
SFP 79346
$-\log_{10}(P)$
0
2
4
6
8
10
12
YJR153W
YJR154W
723000
724000
725000
726000
r^2
0
0.2
0.4
0.6
0.8
1


Chromosome 12

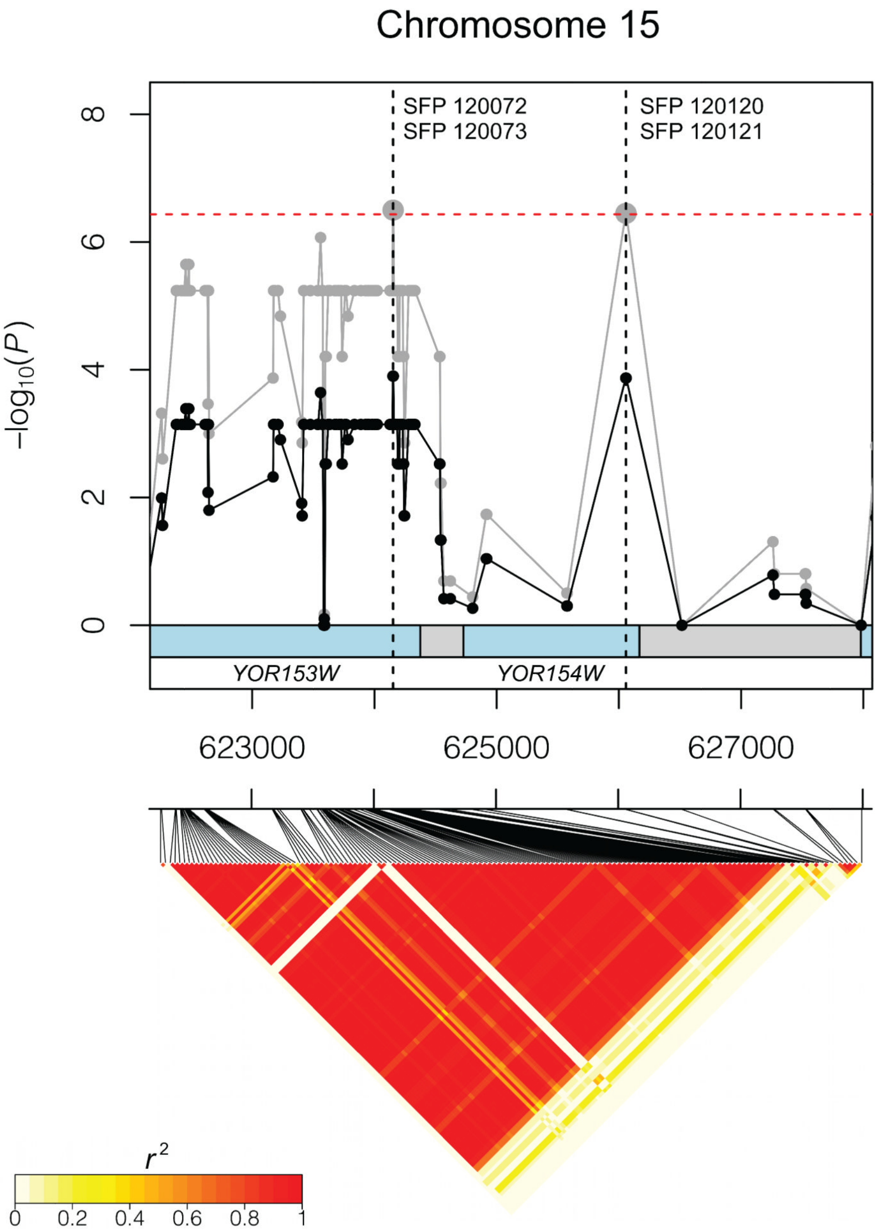


Figure 3.

Magnifications of the genomic regions surrounding single feature polymorphisms (SFPs) that were found to be associated with clinical origin in *Saccharomyces cerevisiae* at a genome-wide significant level according to Fisher exact tests (a, c and f) and Efficient Mixed-Model Association Expedited (b, d and e); the x-axis indicates genomic position in basepair positions and the y-axis represents $-\log_{10}(P)$ before (grey) and after (black) correction for confounding due to population structure by division with the inflation factor λ; the red dashed line corresponds to a nominal 5% significance threshold with Bonferroni correction for 135,771 tests and bold circles indicate a genome-wide significant association (P-value $\leq 3.68 \times 10^{-7}$); the black dashed lines indicate the positions of the SFPs; heatmaps illustrate pairwise linkage disequilibria (LD) between the SFPs, measured as r^2, and are shaded with a white-to-red gradient indicating low to high LD values.

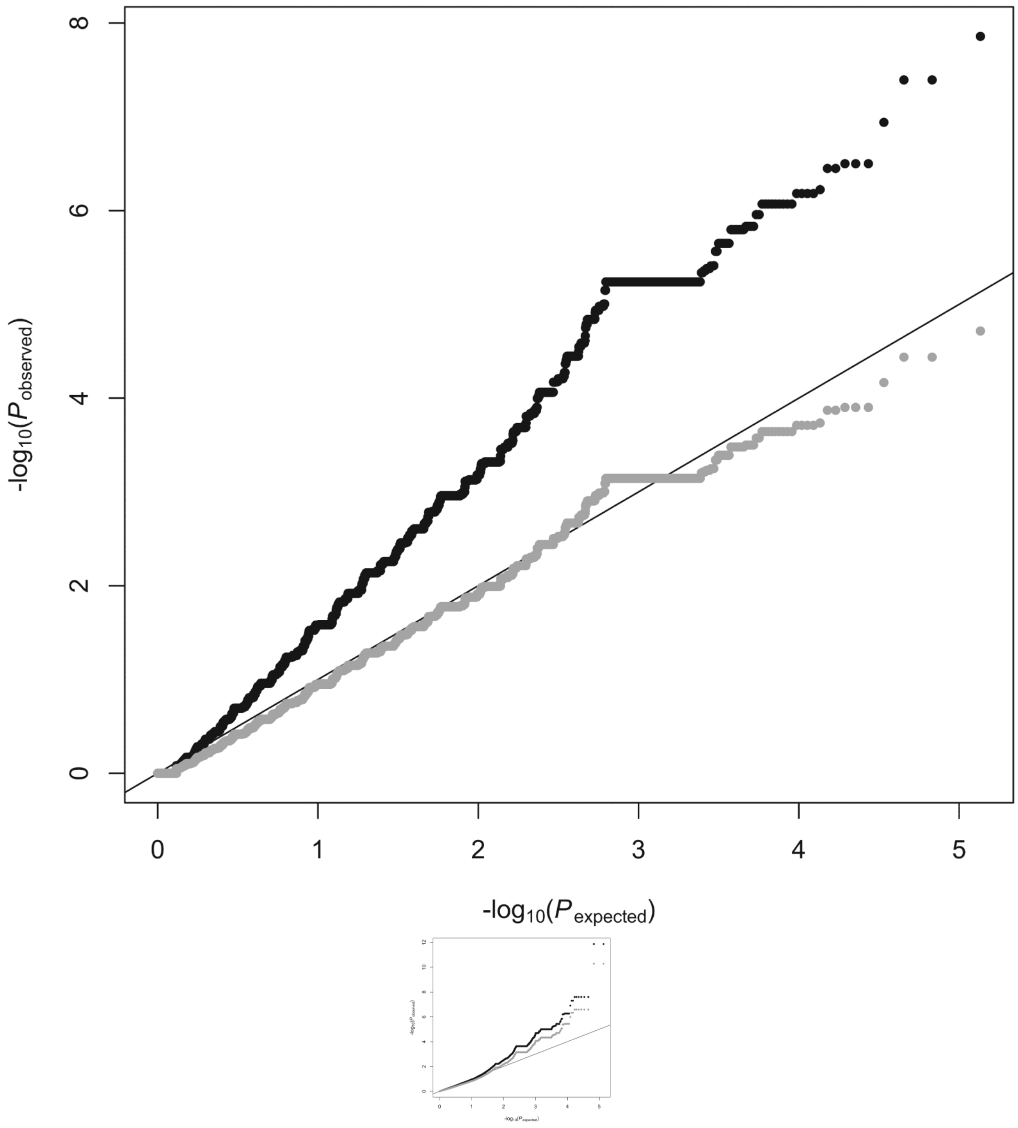


Figure 4.

Log quantile-quantile plots of observed *P*-values generated by Fisher exact tests (a) and Efficient Mixed-Model Association Expedited (EMMAX) (b) versus expected *P*-values under the null hypothesis of no association before (black) and after (grey) correction for confounding due to population stratification by division with the inflation factor λ (λ_{Fisher} = 1.67 and λ_{emmax} = 1.15).

Table 1

Summary of perfect match (PM) oligonucleotide features with unique occurrences in the *Saccharomyces cerevisiae* S288c reference genome and bimodal hybridization intensity distributions based on the hybridization of genomic DNA from the segregants of 88 *S. cerevisiae* isolates onto separate Affymetrix GeneChip® *S. cerevisiae* Tiling 1.0R Arrays; indicated coverage is based on the *S. cerevisiae* S288c reference genome sequence.

			Single feature polymorphisms	
Chromosome #	**# Unique features**	**% Coverage**	**#**	**%**
1	41436	83	4785	11.5
2	170119	94	7624	4.5
3	63927	92	6874	10.8
4	312429	92	17022	5.4
5	118808	93	7578	6.4
6	54768	92	6220	11.4
7	225318	94	11420	5.1
8	111848	90	6047	5.4
9	88944	92	4896	5.5
10	151020	91	7192	4.8
11	143000	97	7920	5.5
12	214231	90	9340	4.4
13	192873	94	7955	4.1
14	161499	93	7800	4.8
15	226315	94	13673	6.0
16	194285	93	9425	4.9
Total	2470820	93	135771	5.5

Table 2

Results of analysis of molecular variance (AMOVA; Excoffier *et al.*, 1992) with genotype data from 5,000 single feature polymorphism markers in 88 *Saccharomyces cerevisiae* strains grouped according to their clinical or non-clinical origin; D.f., degrees of freedom.

Source of variation	D.f.	Sum of squares	Variance component	Percentage of total	*P*-value
Among groups	1	1363.34	17.07	2.71	0.00218
Within groups	86	52657.32	612.29	97.29	
Total	87	54020.66	629.36		

Table 3

Overview of single feature polymorphisms (SFPs) that reach genome-wide significance in an analysis of environmental association with clinical origin in *Saccharomyces cerevisiae* using Fisher exact tests and Efficient Mixed-Model Association Expedited (EMMAX); genome positions refer to the position of the middle nucleotide of each oligonucleotide feature in the February 2011 assembly of the *S. cerevisiae* S288c reference genome; *P*-values that surpass genome-wide significance before and after correction for confounding due to population stratification using the inflation factor λ (λ_{Fisher} = 1.67 and λ_{EMMAX} = 1.15) are indicated in italics; MAF, minor allele frequency; MA, minor allele; S, "S288c-like" allele class; NS, "non S288c-like" allele class; OR, odds ratio; ORF, open reading frame

			MAF (MA)			*P*-value		
SFP #	**Chr #**	**Position**	**Clinical**	**Non-clinical**	**OR (95% CI)**	**Uncorrected**	**Corrected**	**ORF (Gene)**
A. Fisher exact test								
47,217	6	168,779	0.43 (NS)	0.00 (NS)	∞ (7.21 - ∞)	*3.16×10^{-7}*	1.26×10^{-4}	*YFR012W/YFR012W-A* *
65,184	8	318,468	0.48 (NS)	0.00 (NS)	∞ (8.68 - ∞)	*4.04×10^{-8}*	3.65×10^{-5}	*YHR102W (KIC1)*
65,185	8	318,472	0.48 (NS)	0.00 (NS)	∞ (8.68 - ∞)	*4.04×10^{-8}*	3.65×10^{-5}	*YHR102W (KIC1)*
65,186	8	318,476	0.50 (NS)	0.00 (NS)	∞ (9.50 - ∞)	*1.39×10^{-8}*	1.92×10^{-5}	*YHR102W (KIC1)*
65,201	8	321,132	0.46 (NS)	0.00 (NS)	∞ (7.92 - ∞)	*1.15×10^{-7}*	6.82×10^{-5}	*YHR103W (SBE22)*
120,072	15	624,164	0.43 (NS)	0.00 (NS)	∞ (7.21 - ∞)	*3.16×10^{-7}*	1.26×10^{-4}	*YOR153W (PDR5)*
120,073	15	624,168	0.43 (NS)	0.00 (NS)	∞ (7.21 - ∞)	*3.16×10^{-7}*	1.26×10^{-4}	*YOR153W (PDR5)*
120,120	15	626,069	0.39 (S)	0.09 (NS)	15.30 (4.42 - 69.44)	*3.55×10^{-7}*	1.35×10^{-4}	*YOR154W (SLP1)*
120,121	15	626,073	0.39 (S)	0.09 (NS)	15.30 (4.42 - 69.44)	*3.55×10^{-7}*	1.35×10^{-4}	*YOR154W (SLP1)*
B. EMMAX								
58,321	7	785,660	0.39 (S)	0.20 (NS)	6.04 (2.18 - 18.09)	*1.24×10^{-7}*	1.03×10^{-6}	*YGR146C-A/YGR147C (NAT2)* *
79,325	10	724,251	0.00 (NS)	0.34 (NS)	0.00 (0.00 - 0.21)	*5.01×10^{-8}*	4.67×10^{-7}	*YJR153W (PGU1)/YJR154W* *
79,327	10	724,259	0.00 (NS)	0.34 (NS)	0.00 (0.00 - 0.21)	*5.01×10^{-8}*	4.67×10^{-7}	*YJR153W (PGU1)/YJR154W* *
79,345	10	724,367	0.00 (NS)	0.41 (NS)	0.00 (0.00 - 0.15)	*1.36×10^{-12}*	*5.13×10^{-11}*	*YJR153W (PGU1)/YJR154W* *
79,346	10	724,371	0.00 (NS)	0.41 (NS)	0.00 (0.00 - 0.15)	*1.36×10^{-12}*	*5.13×10^{-11}*	*YJR153W (PGU1)/YJR154W* *
94,444	12	804,623	0.27 (NS)	0.05 (NS)	7.71 (1.55 - 75.64)	*2.50×10^{-8}*	*2.56×10^{-7}*	*YLR337C (VRP1)*
94,445	12	804,627	0.27 (NS)	0.05 (NS)	7.71 (1.55 - 75.64)	*2.50×10^{-8}*	*2.56×10^{-7}*	*YLR337C (VRP1)*
94,446	12	804,631	0.27 (NS)	0.05 (NS)	7.71 (1.55 - 75.64)	*2.50×10^{-8}*	*2.56×10^{-7}*	*YLR337C (VRP1)*
94,447	12	804,635	0.27 (NS)	0.05 (NS)	7.71 (1.55 - 75.64)	*2.50×10^{-8}*	*2.56×10^{-7}*	*YLR337C (VRP1)*
94,468	12	807,034	0.27 (NS)	0.05 (NS)	7.71 (1.55 - 75.64)	*2.50×10^{-8}*	*2.56×10^{-7}*	*YLR340W (RPP0)/YLR341W (SPO77)* *

			MAF (MA)			*P*-value		
SFP #	**Chr #**	**Position**	**Clinical**	**Non-clinical**	**OR (95% CI)**	**Uncorrected**	**Corrected**	**ORF (Gene)**
94,469	12	807,038	0.27 (NS)	0.05 (NS)	7.71 (1.55 - 75.64)	2.50×10^{-8}	2.56×10^{-7}	*YLR340W (RPP0)/YLR341W (SPO77)* *

* intergenic SFP.

Table 4

Pairwise linkage disequilibria (LD) between SFPs significantly associated with clinical background in *S. cerevisiae* (significance assessed before correction of *P*-values for inflation due to population structure), measured as r^2 using PLINK v1.07.

SFP #	**47217**	**58321**	**65184**	**65185**	**65186**	**65201**	**79325**	**79327**	**79345**	**79346**	**94444**	**94445**	**94446**	**94447**	**94468**	**94469**	**120072**	**120073**	**120120**
58321	0.00																		
65184	0.38	0.00																	
65185	0.38	0.00	1.00																
65186	0.43	0.00	0.94	0.94															
65201	0.41	0.00	0.94	0.94	0.88														
79325	0.06	0.00	0.06	0.06	0.07	0.06													
79327	0.06	0.00	0.06	0.06	0.07	0.06	1.00												
79345	0.07	0.00	0.08	0.08	0.09	0.08	0.67	0.67											
79346	0.07	0.00	0.08	0.08	0.09	0.08	0.67	0.67	1.00										
94444	0.02	0.16	0.06	0.06	0.03	0.06	0.04	0.04	0.05	0.05									
94445	0.02	0.16	0.06	0.06	0.03	0.06	0.04	0.04	0.05	0.05	1.00								
94446	0.02	0.16	0.06	0.06	0.03	0.06	0.04	0.04	0.05	0.05	1.00	1.00							
94447	0.02	0.16	0.06	0.06	0.03	0.06	0.04	0.04	0.05	0.05	1.00	1.00	1.00						
94468	0.02	0.16	0.06	0.06	0.03	0.06	0.04	0.04	0.05	0.05	1.00	1.00	1.00	1.00					
94469	0.02	0.16	0.06	0.06	0.03	0.06	0.04	0.04	0.05	0.05	1.00	1.00	1.00	1.00	1.00				
120072	0.44	0.03	0.38	0.38	0.35	0.41	0.06	0.06	0.07	0.07	0.05	0.05	0.05	0.05	0.05	0.05			
120073	0.44	0.03	0.38	0.38	0.35	0.41	0.06	0.06	0.07	0.07	0.05	0.05	0.05	0.05	0.05	0.05	1.00		
120120	0.06	0.20	0.10	0.10	0.12	0.11	0.02	0.02	0.04	0.04	0.16	0.16	0.16	0.16	0.16	0.16	0.18	0.18	
120121	0.06	0.20	0.10	0.10	0.12	0.11	0.02	0.02	0.04	0.04	0.16	0.16	0.16	0.16	0.16	0.16	0.18	0.18	1.00