Published in final edited form as:

Nature. 2010 March 4; 464(7285): 54–58. doi:10.1038/nature08791.

Remarkably ancient balanced polymorphisms in a multi-locus gene network

Chris Todd Hittinger[1,2], **Paula Gonçalves**[3], **José Paulo Sampaio**[3], **Jim Dover**[1,2], **Mark Johnston**[1,2], and **Antonis Rokas**[4]

[1] Department of Biochemistry and Molecular Genetics, University of Colorado Denver Health Sciences Center, Aurora, CO 80045, USA

[2] Center for Genome Sciences, Department of Genetics, Washington University in St. Louis School of Medicine, St. Louis, MO 63108, USA

[3] Centro de Recursos Microbiológicos, Departamento de Ciências da Vida, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, 2829-516 Caparica, Portugal

[4] Department of Biological Sciences, Vanderbilt University, Nashville, TN 37235, USA

Abstract

Local adaptations within species are often governed by several interacting genes scattered throughout the genome. Single-locus models of selection cannot explain the maintenance of such complex variation because recombination separates co-adapted alleles. Here we report a novel type of intraspecific multi-locus genetic variation that has been maintained over a vast period of time. The galactose (*GAL*) utilization gene network of the brewer's yeast relative *Saccharomyces kudriavzevii* exists in two distinct states: a functional gene network in Portuguese strains and, in Japanese strains, a non-functional gene network of allelic pseudogenes. Genome sequencing of all available *S. kudriavzevii* strains revealed that none of the functional *GAL* genes were acquired from other species. Rather, these polymorphisms have been maintained for nearly the entire history of the species, despite more recent gene flow genome-wide. Experimental evidence suggests that inactivation of the *GAL3* and *GAL80* regulatory genes facilitated the origin and long-term maintenance of the two gene network states. This striking example of a balanced unlinked gene network polymorphism introduces a remarkable type of intraspecific variation that may be widespread.

Correspondence and requests for materials should be addressed to M. J.: mark.johnston@ucdenver.edu..

Author Contributions CTH, PG, JPS, MJ, and AR conceived and designed experiments; CTH, PG, and JD performed experiments; CTH sequenced and assembled genomes; CTH and AR analyzed data; CTH, MJ, and AR wrote the paper with advice and consent from all authors.

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Data Deposition All short sequencing reads have been deposited with NCBI's SRA as SRP001457 of SRA010159; GenBank accessions GU299171-GU299178 have also been deposited.

Reprints and permissions information is available at npg.nature.com/reprintsandpermissions.

Competing financial interests The authors declare no competing financial interests.

Genetic variation fuels evolution by providing the diversity upon which natural selection acts. In contrast with the more common directional and stabilizing forms of selection, balancing selection favors the maintenance of variation within a species. Overdominance, frequency-dependent selection, and local adaptations to heterogeneous ecological niches can all cause balancing selection, leaving the signature of unusually high levels of sequence divergence between segregating alleles[1],[2],[3],[4],[5],[6],[7],[8],[9]. In some cases, alternate alleles of single loci have been maintained for millions of years and span species boundaries, such as at the *MHC* locus in mammals[2],[3].

In contrast to conventional single-locus balanced polymorphisms, the possibility of a complex, multi-locus gene network being maintained in alternate states within a species has received little attention. However, quantitative genetic analyses have revealed numerous instances where ecologically relevant traits are sculpted from variation at multiple loci, often through epistatic interactions among genes[6],[10],[11],[12],[13]. Therefore, keeping co-adapted, interacting alleles or gene complexes together is likely key to optimal fitness. In principle, alternative allelic states of multiple genes under balancing selection could be maintained by tight linkage[6],[14], through chromosomal inversions[15], via reduced gene flow between populations at the early stages of speciation[16],[17], and even through inbreeding, as suggested by some theoretical treatments[18],[19].

Here we report a novel type of balanced polymorphism that we term a “balanced unlinked gene network polymorphism” (BuGNP), consisting of co-adapted alleles of several functionally related, unlinked genes with extremely elevated sequence divergence. The BuGNP we describe has persisted for nearly the entire history of the species, even as gene flow continued throughout the rest of the genome. While BuGNPs share some features with gene families under balancing selection[2],[4],[20] and with classical Dobzhansky-Muller incompatibilities between incipient species[16],[21], we show that BuGNPs can persist over vast periods of time without speciation. Moreover, their interacting genes comprise alternate network states that are most effective at performing a coordinated task when their allelic states are matched.

One species, two *GAL* gene network states

Galactose catabolism by the baker’s and brewer’s yeast, *Saccharomyces cerevisiae*, is governed by a network of seven interacting genes that regulate and effect the conversion of galactose into a glycolytic substrate to produce energy[22],[23]. This gene network encodes a transporter (Gal2), three enzymes that catalyze the conversion of galactose into glucose-6-phosphate (Gal1, Gal10, and Gal7), a transcriptional activator (Gal4), a co-repressor (Gal80), and a co-inducer (Gal3). This well-understood gene network is a paradigm for studying eukaryotic regulatory networks[22],[23],[24], evolutionary processes[25],[26],[27],[28], and systems biology[29].

The key features of the *GAL* gene network are preserved among most members of the *Saccharomyces sensu stricto* (hereafter, *Saccharomyces*) that includes *S. cerevisiae* and its close relative, *S. kudriavzevii*. Previous analyses of their genomes showed that the type strain of *S. kudriavzevii* and several more distantly related species have independently lost functional *GAL* genes, resulting in their inability to utilize galactose as a carbon source[25]. The genomes of the *S. kudriavzevii* type strain and of three other strains isolated in Japan possess heavily degenerated *GAL* pseudogenes that are syntenic with the functional *GAL* genes of the other *Saccharomyces* species. We were therefore surprised to find, in several Portuguese locations and substrates, 14 strains of *S. kudriavzevii* that can use galactose (Gal$^+$) as their sole carbon source[30].

Since the degree of sequence degeneration of the *GAL* pseudogenes suggests that their inactivation occurred soon after *S. kudriavzevii* diverged from *S. cerevisiae*[25], we first considered whether the Gal$^+$ strains thought to be *S. kudriavzevii* might instead be a different species, or a hybrid of two species. However, most of these strains readily produced viable spores (Table S1), and crosses between a Gal$^-$ Japanese strain and a Gal$^+$ Portuguese strain produced viable (82%) spores that exhibited normal gene segregation (Fig. S1). These results validate the initial characterization of the Portuguese strains as isolates of *S. kudriavzevii*[30] and show that they lack substantial intrinsic reproductive barriers that would isolate them from Japanese strains.

Genome sequencing has revealed rare introgression of genes among other *Saccharomyces* yeasts[31],[32]. To determine if the Gal$^+$ strains of *S. kudriavzevii* had obtained any genes from other species, we generated millions of mostly 36 bp DNA sequence reads from each of the 18 available strains of *S. kudriavzevii* and scanned those reads for evidence of introgression from other yeast species. We easily detected non-*S. kudriavzevii* sequences in a known *S. cerevisiae/S. kudriavzevii* hybrid used in winemaking[33] but found no evidence for introgression in any of the 18 wild isolates of *S. kudriavzevii* (Fig. S2, Database S1). We conclude that introgression had little, if any, impact on genome evolution of the Gal$^+$ strains.

To rule out the possibility of *GAL*-specific introgression, we obtained the sequences of all *GAL* genes and pseudogenes from all available strains of *S. kudriavzevii* and all previously sequenced genus members. Phylogenetic analyses are consistent with a monophyletic origin of the *S. kudriavzevii GAL* genes and pseudogenes and conclusively exclude introgression from any known lineage of yeast (Figs. 1a, S3a, S4, Table S2). Furthermore, we timed the split between the functional *GAL* genes and the pseudogenes within the *S. kudriavzevii* lineage by using a relaxed molecular clock approach. Our results indicate that the coalescence of the *GAL* pseudogenes and functional *GAL* genes occurred near the time of divergence of *S. kudriavzevii* as a distinct lineage. Specifically, the *GAL* pseudogenes are approximately 89% as old as the species (Fig. S3). This finding is remarkable given the lack of any apparent pre- or post-zygotic barrier to crosses between Gal$^+$ and Gal$^-$ *S. kudriavzevii* strains in the laboratory.

Ancient *GAL*, recent genome coalescence

A simple explanation for the extreme sequence divergence of the *GAL* genes would be that Gal$^+$ Portuguese strains and Gal$^-$ Japanese strains never had the opportunity to exchange genetic material in nature due to geographical or other extrinsic isolating mechanisms. This hypothesis predicts that all their genes should be highly divergent, like the *GAL* genes. Alternatively, the Gal$^+$ Portuguese strains and the Gal$^-$ Japanese strains might have exchanged genetic material throughout the rest of the genome, even as natural selection maintained two distinct *GAL* network states.

To determine if the *GAL* loci are representative of the divergence of Gal$^+$ Portuguese strains and their Gal$^-$ Japanese counterparts, we assembled draft genome sequences for each strain of *S. kudriavzevii* by mapping millions of short DNA sequence reads to the existing draft genome sequence of the Gal$^-$ Japanese type strain, IFO1802^T[25,34], confidently determining about 80% of the orthologous bases in the genome of each strain (Table S1). Sequence comparisons revealed that all strains share a recent common ancestor for nearly all genes, with coalescence at only 3% of the way back in the *S. kudriavzevii* lineage, except for the Gal$^-$ Japanese strain IFO1803. This highly divergent strain is an outgroup to all other strains of *S. kudriavzevii*, except at the *GAL* loci where it is monophyletic with the other Japanese strains (Figs. 1, S4). The difference in tree topology at the *GAL* loci suggests that this otherwise distinct Japanese lineage may have experienced a selective regime similar to the other three Japanese strains with respect to galactose.

The average genome-wide divergence of synonymous sites (*dS*) between the Japanese (IFO1802^T) and Portuguese reference (ZP591) strains across all annotated genes is 0.021 (Database S2), while divergence among all sites is 0.011. These values are slightly lower than the divergence between European and Far-Eastern populations of *Saccharomyces paradoxus*[32,35] and only slightly higher than the most divergent strains of *S. cerevisiae*[31,32,36]. Variation between strains within the Portuguese population ($N = 14$) and within the main Japanese population ($N = 3$) is usually above 0.001, which also appears to be typical for wild populations of the genus[32]. While the discovery and genome sequencing of additional strains might help address whether the *GAL* polymorphisms are fixed between populations, genome sequence data from the available strains suggests that gene flow between the Portuguese and main Japanese populations of *S. kudriavzevii* was recently extensive, or that they were founded from the same metapopulation.

Many genes adjacent to the *GAL* genes also have elevated sequence divergence between the Japanese and Portuguese reference strains ($dS = 0.159$, $P < 10^{-5}$). In every case, the region of elevated divergence is centered on the *GAL* gene(s) ($dS = 0.939$), and there is no functional connection between adjacent genes. Moreover, levels of divergence rapidly decline toward the genome-wide background average, sometimes in the middle of open reading frames (Figs. 2, S5). The persistence of *GAL* pseudogenes and linked polymorphisms in one population suggests that they are not simply rare deleterious alleles that have yet to be removed by purifying selection, as is likely for Gal$^-$ strains of *S. cerevisiae* with recent inactivating mutations in single *GAL* genes[32]. Instead, the striking localized peaks of extreme sequence divergence between populations are best explained by strong balancing selection on the *GAL* genes, which suggests that non-functional alleles are more fit in some genetic backgrounds and/or environmental conditions.

In addition to the *GAL* loci, 48 other genes are significantly more divergent than the genome average and are thus good candidates for genes under balancing selection (Database S2). The most divergent of these is the amino acid permease *LYP1* with a highly-elevated *dS* of 0.3498. In fact, nearly one-quarter (11/48) of these candidate genes are involved in amino acid metabolism and transport (Table 1). This highly significant 5-fold enrichment ($P < 10^{-5}$) suggests that gene flow has been selectively reduced for many genes involved in this process. The divergent alleles of these functionally related genes may constitute another BuGNP maintained within *S. kudriavzevii*.

Maintaining a gene network polymorphism

We know of no other case where two distinct states of a multi-locus gene network have been maintained for as long as the *GAL* BuGNP of *S. kudriavzevii*. In *S. cerevisiae*, the *GAL* genes are scattered over 5 different chromosomes, so their independent assortment in F_2 hybrid progeny would only rarely reconstitute a fully Gal^+ or Gal^- gene network in the absence of genome rearrangements or meiotic drive (1 in 32 for each state). We can imagine at least three evolutionary processes that may have facilitated the maintenance of the *GAL* BuGNP in *S. kudriavzevii*.

First, recovery of a complete gene network state is more likely in *S. kudriavzevii* because this species has one less gene in the network than does *S. cerevisiae*: it no longer relies on the Gal3 co-inducer, which is a highly degenerated pseudogene even in the Gal^+ Portuguese strains (Fig. 3a). Beyond this simplification of the gene network, laboratory crosses revealed no additional mechanism to increase the likelihood of recovering pure gene network states (*i.e.* the *GAL* loci segregated independently, Fig. S1).

Second, mating is not random because the extant *S. kudriavzevii* populations are highly structured, even though they share a recent common ancestor throughout most of their genome. The broad concordance of single-gene phylogenetic trees (Fig. 1) suggests that gene flow between populations is rare, relative to mating within populations. The infrequency of outcrossing and sexual reproduction in *Saccharomyces* yeasts[37] are both predicted to reinforce population structure and facilitate the maintenance of co-adapted alleles[19].

Finally, while natural selection would clearly limit the success of any pseudogene alleles invading a Gal^+ population in niches where galactose is a useful carbon source, we wondered whether some combinations of functional and pseudogene alleles might make cells unfit in environments that lack galactose. Specifically, the absence of the Gal80 co-repressor in a Gal^- population is expected to lead to constitutive, deleterious expression of partial *GAL* gene networks in strains containing invading functional alleles of *GAL4* and any *GAL* target genes (Fig. 3a, Table S3)[22],[23],[27]. Indeed, *S. kudriavzevii gal80*Δ mutants containing functional alleles of the rest of the *GAL* genes were at a significant disadvantage when grown without galactose, especially in non-glucose-repressing conditions ($P < 10^{-3}$; and to a lesser extent in glucose-repressing conditions, $P < 10^{-2}$; Fig. 3b). Thus, in genetic backgrounds that cause partial gene networks to be expressed constitutively (*i.e.* $GAL4^+$ *gal80*Δ strains), functional *GAL* alleles (other than *GAL80*) would be strongly selected against, thereby imposing a moderate fitness cost when averaged across all genetic backgrounds an allele might encounter in the absence of galactose. Since *GAL80* and linked sequences also exhibit ancient coalescence and the known *GAL80* pseudogene alleles are monophyletic, it is conceivable that functional *GAL80* (and perhaps other functional *GAL* genes) could also confer slight, conditional fitness costs by other unknown means in the Japanese population.

Origins of a gene network polymorphism

The seemingly critical roles of *GAL80* and *GAL3* in the maintenance of distinct *GAL* states in *S. kudriavzevii* leads us to consider other evolutionary changes that likely occurred in regulatory components near the time of the origin of the BuGNP (Fig. 4). The ancestral *Saccharomyces GAL* gene network presumably depended on the Gal80/Gal80b and Gal1/Gal3 paralog pairs for repression and induction, respectively (Fig. 3a)[25]. The only trace of *GAL80B* within the *S. cerevisiae/S. kudriavzevii* clade is a very small syntenic pseudogene fragment found in all strains of *S. kudriavzevii*, leaving Gal80 as the sole co-repressor. Gal3 function was also narrowed in the *S. cerevisiae/S. kudriavzevii* lineage by the loss of its galactokinase activity, once shared with its paralog, Gal1. We wondered whether the complete absence of functional *GAL3* in Portuguese strains of *S. kudriavzevii* impairs their ability to respond to galactose ($P < 10^{-2}$; Fig. 3c) and limits the utility of the functional network. Insertion of *S. cerevisiae GAL3* into the genome of a Portuguese strain of *S. kudriavzevii* dramatically sped its response to galactose ($P < 10^{-3}$), rivalling the rapid response of wild-type *S. cerevisiae*. Interestingly, although the naturally *gal3* Portuguese strains of *S. kudriavzevii* experience some delay in their response to galactose, they react much more quickly than *gal3*Δ mutants of *S. cerevisiae*[23] ($P < 10^{-2}$), suggesting that the Portuguese *S. kudriavzevii GAL* gene network contains compensatory mutations that partially mitigate the loss of *GAL3*, and which may have been critical to the evolution of the simplified gene network.

Balanced unlinked gene network polymorphisms

Regardless of the ecological and genetic circumstances that led to the retention of two very different states of the *GAL* gene network in *S. kudriavzevii*, this extreme example introduces a remarkable novel type of genetic variation in a sexual eukaryote: a BuGNP maintained as two alternate states of six genes that must interact with specific alleles for optimal fitness. The numerous cases of long-term balancing selection[2,4,6,7,8,9], complex genetic interactions[6,10,11,12,13], and theoretical considerations[18,19] all hint that BuGNPs might be important for explaining the evolution of complex traits, but we know of no other definitive examples of balancing selection acting to preserve alternates states of a multi-locus gene network within a single species.

Genome-wide scans for cases of balancing selection have either been negative or have identified only some of the features of BuGNPs. The paucity of balancing selection in humans[38], despite being the most thoroughly sampled of any species, suggests that maintenance of extreme multi-locus variation may require some or all of the features we have observed in *S. kudriavzevii*: broad geographic distribution, strong population structure, limited sexual reproduction, large effective population size, and interacting gene networks whose alternate states can create unfit genotypes when recombined. Indeed, population genomic studies of the selfing plant *Arabidopsis thaliana*[20,39] and of the *Anopheles gambiae* species complex[17] have revealed regions of unusually high divergence or reduced gene flow. However, the observed sequence divergence in these regions is much lower than that of the *GAL* gene network of *S. kudriavzevii*, and whether these regions interact to contribute to phenotypic variation is unknown. Most BuGNPs are probably more quantitative (and therefore harder to detect) than the one presented here, but we anticipate that gene networks could be maintained in alternate states in other broadly distributed species with similar life cycles.

METHODS SUMMARY

We generated millions of mostly 36 bp DNA sequence reads[40],[41] from each available strain of *S. kudriavzevii*, mapped[42] them to the genome sequence of the type strain (IFO1802^{T})[34] with some modified[25] and additional contigs, and confidently determined about 80% of the orthologous bases for each genome with a conservatively estimated error rate of less than 5 × 10^{-4}. For each highly divergent *GAL* locus, we generated alternate Portuguese-specific reference contigs from *de novo* short-read assemblies[43] and PCR-based Sanger-sequencing reads. Phylogenetic analyses, experimental manipulations of yeast, and other analyses were performed using standard procedures with modifications described in the Online Methods.

ONLINE METHODS

Preparation of libraries for sequencing genomic DNA

Genomic DNA (gDNA) was isolated from all 18 available strains of *S. kudriavzevii* (or their monosporic derivatives) and two hybrid strains (Table S1). We attempted to create monosporic derivatives for all Portuguese strains by dissecting and isolating individual spores and allowing them to reform homozygous diploid strains by selfing. Four strains sporulated poorly and had low spore viability. The genome sequences of three of them revealed the absence of any sequences corresponding to one of the mating types, suggesting the poor sporulation and sterility were caused by a loss of one of the *HML/HMR* loci and a homozygous *MAT* locus. For each strain, ~5 μg of gDNA was sonicated and ligated to Illumina’s Solexa® sequencing adapters using either the manufacturer’s kit or our custom protocol[41] (using gDNA instead of cDNA). Some libraries were prepared with multiplexing barcode adapter pairs with an added “T” overhang for ligation; reads were processed to sort by and remove the “NT” barcodes and overhangs prior to analysis (Table S1). Libraries were sequenced using Illumina’s Solexa® Genome Analyzer I or II and Pipeline Software according the manufacturer’s instructions (San Diego, CA)[40].

Screening sequence reads for evidence of hybridization and introgression

We screened sequence reads from each strain for evidence of hybridization and introgression by mapping reads to the genome sequences of all *Saccharomyces* species available from the *Saccharomyces* Genome Database (www.yeastgenome.org): *S. cerevisiae*[45], *S. paradoxus*[46], *Saccharomyces mikatae*[46], *S. kudriavzevii* IFO1802^{T}[34], *S. bayanus* var. *uvarum*[46]. To facilitate scans for chromosomal regions that might have been introgressed, and to limit false positives and annotation errors, we collected the sequences of all *Saccharomyces* open reading frames (ORFs) that were annotated as orthologous to single *S. cerevisiae* ORFs for which there was a single annotated copy in each species, providing a total of 2,805 annotated single copy ORFs per species. Reads were mapped to this library of 14,025 ORFs using RMAP v. 0.45[42] with no mismatches allowed, only retaining unambiguously mapped reads. We examined the proportion of reads mapped to *S. kudriavzevii* for each gene, and found that all introgression candidates were due to putative annotation or assembly errors (usually truncations of the annotated *S. kudriavzevii* ORF), had so few matches as to be uninformative, or were the result of other artifacts. Complete data are presented in Database S1, but imposing a stringent filter (requiring retained genes to have a normalized hit count that was no lower than one standard deviation below the mean) left no clear evidence of introgression, except for in a known[33] hybrid (W27) and another strain (CBS679) not recovered from the wild (Fig. S2). The Portuguese reference strain ZP591 was also analyzed with one, two, three, four, and five mismatches (Database S1).

Reference-guided assemblies

Since the average short-read coverage of each genome was about 10-fold, assembly of genome sequences required a reference genome. We started with the 3.4x Sanger-sequenced genome of *S. kudriavzevii* IFO1802^{T}[34], including several corrections, mostly to the *GAL* loci[25]. The extreme divergence between the sequences of the *GAL* loci of the Portuguese and Japanese strains made reference-guided assemblies impossible there. Instead, we created alternative Portuguese-specific contigs using VELVET[43] assemblies of ZP591 (the most thoroughly sequenced Portuguese strain) and Sanger-sequencing data obtained by a targeted PCR and sequencing strategy. The boundaries of these Portuguese-specific contigs were extended from the *GAL* loci until divergence returned to background levels (see contigs 9001, 9002, 9004, 9080 for precise boundaries). These alternate contigs overlap and are syntenic with the IFO1802^{T} contigs. Since we required all mapped reads to be unique, using these alternate contigs did not interfere with assemblies, but it required additional source-specific processing (see below). Complete functional *GAL1*, *GAL2*, *GAL4*, *GAL7*, *GAL10*, and *GAL80* genes were verified for ZP591 by both methods, and there were no discrepancies between VELVET and Sanger contigs assembled in these regions.

To create draft genome assemblies from short-read data for each strain, we first mapped all reads to the above reference genome using RMAPQ v. 0.45[42] with 3 mismatches and a quality filter of 5, which records the positions of all reads that can be uniquely mapped to the reference genome with 3 or fewer mismatches at high quality bases. This approach provided a good balance between coverage, variant detection, and error rates, but increasing the number of mismatches to 5 produced similar assemblies. Solexa® quality scores (Q) calculated by the Solexa® Pipeline Software are analogous to Phred scores, and the error probability (p) for a single base in a single read is defined as $p = 10^{-Q/10}/(1 + 10^{-Q/10})$. We produced assemblies by processing the BED-formatted RMAPQ output and the Q values for each position in each mapped read using custom Perl scripts to assign a cumulative error probability (p_c) to each base for each position. Specifically, p_c for each of the four possible bases for each position is calculated as the product of all mapped p values that support that base at that position. Positions that did not meet all of the following criteria were not called and were conservatively recorded in the assemblies as unknown bases (N): 1) one of the four possible bases must have $p_c < 10^{-5}$, 2) there must be no other base with $p_c < 10^{-5}$, and 3) the called base must have a p_c value that is at least 10^5-fold lower than the sum of all other possible bases.

This reference-guided assembly procedure requires that the support for a given base at a given position be strong and based on multiple quality reads and that there not be appreciable evidence for an alternative base at that position, either due to heterozygosity or due to the presence of nearly identical sequences elsewhere in the genome. This reference-guided assembly approach does not account for indels, but our downstream uses of the data (phylogenetics, divergence, and population genetics) would discard this data, even if it were available. We estimated our error rate at below 5×10^{-4} errors/bp by re-assembling the IFO1802^T genome from only new short-read data using the above procedure. This error estimate assumes that the reference sequence contains no errors, which is conservative since the 3.4x draft sequence has been estimated to contain at least 1×10^{-4} errors/bp[34].

Pre-aligned ORF sequences were generated from the above reference-guided genome assemblies for each strain using the ORF annotations from the reference genome. For genes wholly or partially included in the alternate Portuguese-specific contigs (*GAL7*, *GAL10*, *GAL1*, *GAL2*, *SRL2*, *GAL4*, *GYP5*, *GAL80*, *AIM32*, and *SUR7*), bases from the appropriate contigs were joined at their boundaries to create single complete ORFs for downstream analyses.

Alignments of *GAL* pseudogenes

While most of the genome was aligned during the assembly process, aligning the functional *GAL* genes to the heavily degenerated *GAL* pseudogenes required a separate procedure. First, we aligned the entire annotated pseudogene[25] to all the orthologous functional *Saccharomyces* genes, including those from ZP591, using DIALIGN v. 2.2.1[47]. This was performed both with and without Sanger-sequence data for the IFO1803 pseudogenes (which broadly agreed with the patchier assemblies from above). Alignments were trimmed in-frame by codon such that only complete codons significantly aligned by DIALIGN at all positions in all taxa were retained. Upstream and downstream boundaries were determined using BLASTX to compare the IFO1802^T pseudogene against the *S. cerevisiae* genome and trimming the alignment to the first or last identical amino acid, respectively. The full 83 strain dataset (Fig. S4) was processed manually following alignment with DIALIGN, while the codon-based procedure was also used for all other phylogenetic data matrices.

Phylogenetics

Bayesian inference was conducted using MRBAYES v. 3.1.2[48],[49],[50], assuming a GTR model of nucleotide substitution[51], and allowing for rate heterogeneity among sites by assuming that a certain proportion of sites were invariable and that the rates of the rest are determined according to the shape parameter alpha of the gamma distribution. Two independent analyses were run in parallel. Each analysis contained four chains (one cold and three incrementally heated), and trees were sampled every 1,000 generations. These analyses were run for 2,000,000 generations, by which time the average deviation of split frequencies was below 0.01. The trees and parameters sampled from the first 10% of generations from each of the two analyses were discarded as the burn-in.

ML analyses were performed using PAUP* v. 4.0.b10[52]. The best-fit model of nucleotide evolution was estimated by MODELTEST v. 3.7[53]. Clade support was assessed using 100 replicates of non-parametric bootstrap re-sampling. Agreement with the dominantgenome-wide trees (Fig. 1b, 1c) was assessed for individual genes that contained 10 or more parsimony informative sites using the Shimodaira-Hasegawa test[44] as implemented in PAUP* v. 4.0.b10[52].

Coalescence estimation

Rooted, time-measured phylogenies were inferred using BEAST v. 1.4.8[54]. Since the fungal and yeast fossil records are very poor and reliable fossil calibration points unavailable, all branches were estimated in units of substitution/site. We assumed a GTR+GAMMA model of sequence evolution and the uncorrelated lognormal relaxed clock model. The Yule tree prior was chosen for the analysis of the *GAL* pseudogenes and functional genes (Fig. S3a) and for the *Saccharomyces* yeast clade (Fig. S3b). The coalescent (constant size) tree prior was used for the analysis of representative *S. kudriavzevii* strains (Fig. S3c). Two to eight independent runs of 2,000,000 to 10,000,000 generations were run for each data matrix. The achievement of convergence was verified by examining the effective sample size of the likelihood and posterior probability parameters for each analysis (>100) and visually verified by inspection of the likelihood and posterior probability distributions across independent runs. The first 10% of sampled data points from each run was discarded as burn-in. To make the relaxed-clock phylogenies across all analyses comparable to the analysis of the *GAL* pseudogenes and functional genes (which contains 4,860 sites), other analyses were performed on 4,860 randomly selected orthologous sites.

To compare and search for highly divergent genes, the synonymous site divergence (*dS*) and several other parameters were calculated using both the ML-F3X4 and modified Nei-Gojobori[55] estimates implemented by CODEML of PAML v. 4.1[56] for all annotated ORFs where there were greater than 10 complete codons aligned. These two estimates produced similar genome-wide estimates of *dS*, so we only used the F3X4 estimate to test for statistical outliers based on a Poisson sampling distribution of inferred synonymous substitutions and a Bonferroni correction for multiple tests (Table 1, Database S2). Unless otherwise stated, pairwise comparisons were between IFO1802^{T} (Gal^{-} reference) and ZP591 (Gal^{+} reference).

Since the *GAL* loci are significantly more divergent than the rest of the genome, we examined the extent of these highly divergent islands by graphing *dS* along a sliding window. Position-based modified Nei-Gojobori estimates of *dS* with a Jukes-Cantor correction were generated using a one-site step and a 100-site window with DNASP v. 4.90.1[57] and are shown with an arbitrary intergenic spacer of 200 bp (Fig. 2). *GAL7*, *GAL10*, *GAL1*, and *GAL2* were collapsed into single points because of limited aligned data, and 95% confidence intervals were established using a binomial distribution of observed differences. A similar analysis of nucleotide divergence of all coding and non-coding regions of the *GAL* pseudogenes and functional genes also found levels significantly elevated over background (Fig. S5).

Statistics

Experimental data were analyzed using Wilcoxon rank sum tests implemented by MSTAT v. 5.01 (http://mcardle.oncology.wisc.edu/mstat). The enrichment of amino acid metabolism and transport among significantly divergent genes (Table 1, Database S2) was assessed using the hypergeometric distribution on the pooled Gene Ontology (http://www.geneontology.org)[58] biological process terms: amino acid metabolic process (GO: 0006520), oligopeptide transport (GO:0006857), and amino acid transport (GO:0006865). Annotations were from the *S. cerevisiae* ortholog (when available) or homolog (for significant genes lacking a clear ortholog). All *P* values are reported as one-tailed.

Genetic crosses and manipulations

Marked heterothallic haploids of *S. kudriavzevii* were created by replacing one copy of the coding sequence of *HO* of IFO1802^T or FM1071 (a monosporic derivative of ZP591) by transforming these strains with PCR-generated *natMX*[59] or *kanMX*[60] cassettes fused to about 70 bp of sequence upstream and downstream of *HO* and collecting stable *ho*Δ*::natMX* or *ho*Δ*::kanMX* haploid progeny (Table S4). Transformation of *S. kudriavzevii* was accomplished using the standard LiAc protocol optimized for *S. cerevisiae*[61], except incubations were carried out at room temperature (22-23°C), and the heatshock was at 34°C to accommodate the heat-sensitivity of *S. kudriavzevii*. FM1071-derived *ura3*Δ and *trp1*Δ auxotrophic strains were produced by replica-plating cells to 5-fluoroorotic acid (5FOA) or 5-fluoroanthranilic acid (5FAA), respectively, after transforming them with PCR products that introduced start to stop codon deletions and allowing them to recover on YPD plates. All gene deletions were verified by PCR.

To assess the effect of constitutive expression of functional *GAL* genes in a hybrid network, we precisely deleted *GAL80* from start codon to stop codon in several isogenic Portuguese strains (*N* = 12), as well as isogenic control strains whose drug resistance maker was changed from *ho*Δ*::kanMX* to *ho*Δ*::natMX* (*N* = 5, Table S4). Each *ho*Δ*::kanMX gal80*Δ strain was competed against a single *ho*Δ*::natMX GAL80*$^+$ control strain lacking detectable defects, and each *ho*Δ*::natMX GAL80*$^+$ control strain was competed against the *ho*Δ*::kanMX GAL80*$^+$ progenitor of all strains. Each competition was carried out by mixing and co-culturing strains for about 10 generations for 2 days in SC with 2% glucose, performing colony counts of nat^r and kan^r colonies to establish starting frequencies, and inoculating fresh SC media with various carbon sources 1:1000 with aliquots from the same saturated media[28]. Competitions were carried out for 2 days in 2% glucose, 4 days in 2% galactose, and 6 days in 5% glycerol, at which point colony counts of nat^r and kan^r strains established the ending frequencies, from which Malthusian selection coefficients (*m*) were calculated as previously described[28] (Fig. 3b).

To assess the effect of *ScerGAL3*$^+$ on galactose-induction of *GAL* gene expression in a Gal$^+$ strain of *S. kudriavzevii*, we created a *trp1*Δ*::ScerGAL3*$^+$ and a control *trp1*Δ strain by targeting a *trp1*Δ*::ScerURA3*$^+$ strain of *S. kudriavzevii* with PCR products and selecting for 5FOA-resistance to replace *ScerURA3*$^+$ with the PCR products (Table S4). The inserted *ScerGAL3*$^+$ gene included the coding sequence and the full upstream and downstream intergenic regions; it was free of errors, except for the deletion of one bp in a tract of 12 A’s downstream of the coding sequence. To remove any unintended mutations accumulated during strain construction, a panel of *trp1Δ::ScerGAL3*$^+$ and a panel of control *trp1Δ* strains were created by back-crossing them and collecting eight haploid backcross progeny for each genotype that were identical, except at the *ScerGAL3*$^+$ insertion site. These strains of *S. kudriavzevii*, and previously described strains of *S. cerevisiae*[28], were grown to saturation for 2 days in SC with 2% raffinose. Induction of *GAL* gene expression was carried out by inoculating the stationary phase cultures 1:20 in SC with 2% galactose and measuring OD_{600} values every 2 hours to determine the time to first doubling (Fig. 3c).

Two Portuguese/Japanese F_1 hybrid strains of *S. kudriavzevii* were constructed by crossing marked heterothallic haploids (Table S4). These strains were sporulated, tetrads were dissected, and 297 monosporic segregants were recovered. 96 F_2 segregants (48 from each parent) from fully viable tetrads were selected for genotyping. For each *GAL* locus that was functional in the Portuguese strains, we designed PCR primers and conditions that allowed us to distinguish between a functional gene and an orthologous pseudogene by their sizes, and we genotyped each strain at each locus (*GAL7/GAL10/GAL1*, *GAL2*, *GAL4*, and *GAL80*). No non-Mendelian segregation was detected (Fig. S1), nor did the source of the parental mating-types affect the offspring.

Triplicate *gal80Δ* and *GAL80*$^+$ GFP-labeled strains of *S. cerevisiae* were constructed and competed against an otherwise identical BFP-labeled strain as previously described[28] (Table S3, Table S4). Data reported is from a single experiment with quadruplicate biological replicates (from separate GFP colonies) of each genetically engineered strain, for a total of 12 replicates.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We thank Eladio Barrio, Barry L. Williams, The Netherland's CBS, and Japan's NITE-BRC for providing strains; Justin Fay for critical reading of the manuscript; and the Howard Hughes Medical Institute and Sean B. Carroll for supporting preliminary experiments when CTH was a HHMI Predoctoral Fellow. This work was also supported by the James S. McDonnell Foundation (CTH, MJ), NIH-NIGMS (MJ: 5R01GM032540), NIH-NHGRI (CTH: 2T32HG00045), Portugal's FCT (PG, JPS: PTDC/BIA-BDE/71734/2006), the Searle Scholars Program (AR), NSF (AR: DEB-0844968), and Vanderbilt University (AR). CTH is the Maclyn McCarty Fellow of the Helen Hay Whitney Foundation.

References

1. Levene H. Genetic equilibrium when more than one ecological niche is available. Am Nat. 1953; 87:331–333.
2. Hughes AL, Nei M. Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature. 1988; 335:167–170. [PubMed: 3412472]
3. Klein J, Sato A, Nagl S, O'hUigín C. Molecular trans-species polymorphism. Annu Rev Ecol Syst. 1998; 29:1–21.
4. Stahl EA, Dwyer G, Mauricio R, Kreitman M, Bergelson J. Dynamics of disease resistance polymorphism at the Rpm1 locus of Arabidopsis. Nature. 1999; 400:667–671. [PubMed: 10458161]

5. Colosimo PF, et al. Widespread parallel evolution in sticklebacks by repeated fixation of Ectodysplasin alleles. Science. 2005; 307:1928–1933. [PubMed: 15790847]
6. Kroymann J, Mitchell-Olds T. Epistasis and balanced polymorphism influencing complex trait variation. Nature. 2005; 435:95–98. [PubMed: 15875023]
7. Charlesworth D. Balancing selection and its effects on sequences in nearby genome regions. PLoS Genet. 2006; 2:e64. [PubMed: 16683038]
8. Mitchell-Olds T, Willis JH, Goldstein DB. Which evolutionary processes influence natural genetic variation for phenotypic traits? Nat Rev Genet. 2007; 8:845–856. [PubMed: 17943192]
9. Storz JF, et al. The molecular basis of high-altitude adaptation in deer mice. PLoS Genet. 2007; 3:e45. [PubMed: 17397259]
10. Hawthorne DJ, Via S. Genetic linkage of ecological specialization and reproductive isolation in pea aphids. Nature. 2001; 412:904–907. [PubMed: 11528477]
11. Brem RB, Storey JD, Whittle J, Kruglyak L. Genetic interactions between polymorphisms that affect gene expression in yeast. Nature. 2005; 436:701–703. [PubMed: 16079846]
12. Steiner CC, Weber JN, Hoekstra HE. Adaptive variation in beach mice produced by two interacting pigmentation genes. PLoS Biol. 2007; 5:e219. [PubMed: 17696646]
13. Gerke J, Lorenz K, Cohen B. Genetic interactions between transcription factors cause natural variation in yeast. Science. 2009; 323:498–501. [PubMed: 19164747]
14. Navarro A, Barton NH. The effects of multilocus balancing selection on neutral variability. Genetics. 2002; 161:849–863. [PubMed: 12072479]
15. Dobzhansky T, Pavlovsky O. Interracial Hybridization and Breakdown of Coadapted Gene Complexes in Drosophila Paulistorum and Drosophila Willistoni. Proc Natl Acad Sci U S A. 1958; 44:622–629. [PubMed: 16590252]
16. Wu CI, Ting CT. Genes and speciation. Nat Rev Genet. 2004; 5:114–122. [PubMed: 14735122]
17. Turner TL, Hahn MW, Nuzhdin SV. Genomic islands of speciation in Anopheles gambiae. PLoS Biol. 2005; 3:e285. [PubMed: 16076241]
18. Wright S. Genic and organismic selection. Evolution. 1980; 34:825–843.
19. Neher RA, Shraiman BI. Competition between recombination and epistasis can cause a transition from allele to genotype selection. Proc Natl Acad Sci U S A. 2009; 106:6866–6871. [PubMed: 19366665]
20. Clark RM, et al. Common sequence polymorphisms shaping genetic diversity in Arabidopsis thaliana. Science. 2007; 317:338–342. [PubMed: 17641193]
21. Coyne, JA.; Orr, HA. Speciation. Sinauer Associates, Inc.; Sunderland, MA: 2004.
22. Johnston M. A model fungal gene regulatory mechanism: the GAL genes of Saccharomyces cerevisiae. Microbiol Rev. 1987; 51:458–476. [PubMed: 2830478]
23. Bhat PJ, Murthy TV. Transcriptional control of the GAL/MEL regulon of yeast Saccharomyces cerevisiae: mechanism of galactose-mediated signal transduction. Mol Microbiol. 2001; 40:1059–1066. [PubMed: 11401712]
24. Ptashne, M.; Gann, A. Genes and Signals. Cold Spring Harbor Laboratory Press; Woodbury, NY: 2001.
25. Hittinger CT, Rokas A, Carroll SB. Parallel inactivation of multiple GAL pathway genes and ecological diversification in yeasts. Proc Natl Acad Sci U S A. 2004; 101:14144–14149. [PubMed: 15381776]
26. Martchenko M, Levitin A, Hogues H, Nantel A, Whiteway M. Transcriptional rewiring of fungal galactose-metabolism circuitry. Curr Biol. 2007; 17:1007–1013. [PubMed: 17540568]
27. MacLean RC. Pleiotropy and GAL pathway degeneration in yeast. J Evol Biol. 2007; 20:1333–1338. [PubMed: 17584228]
28. Hittinger CT, Carroll SB. Gene duplication and the adaptive evolution of a classic genetic switch. Nature. 2007; 449:677–681. [PubMed: 17928853]
29. Raj A, van Oudenaarden A. Nature, nurture, or chance: stochastic gene expression and its consequences. Cell. 2008; 135:216–226. [PubMed: 18957198]

30. Sampaio JP, Goncalves P. Natural populations of Saccharomyces kudriavzevii in Portugal are associated with oak bark and are sympatric with S. cerevisiae and S. paradoxus. Appl Environ Microbiol. 2008; 74:2144–2152. [PubMed: 18281431]

31. Doniger SW, et al. A catalog of neutral and deleterious polymorphism in yeast. PLoS Genet. 2008; 4:e1000183. [PubMed: 18769710]

32. Liti G, et al. Population genomics of domestic and wild yeasts. Nature. 2009; 458:337–341. [PubMed: 19212322]

33. Gonzalez SS, Barrio E, Gafner J, Querol A. Natural hybrids from Saccharomyces cerevisiae, Saccharomyces bayanus and Saccharomyces kudriavzevii in wine fermentations. FEMS Yeast Res. 2006; 6:1221–1234. [PubMed: 17156019]

34. Cliften P, et al. Finding functional features in Saccharomyces genomes by phylogenetic footprinting. Science. 2003; 301:71–76. [PubMed: 12775844]

35. Bensasson D, Zarowiecki M, Burt A, Koufopanou V. Rapid evolution of yeast centromeres in the absence of drive. Genetics. 2008; 178:2161–2167. [PubMed: 18430941]

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

37. Tsai IJ, Bensasson D, Burt A, Koufopanou V. Population genomics of the wild yeast Saccharomyces paradoxus: Quantifying the life cycle. Proc Natl Acad Sci U S A. 2008; 105:4957–4962. [PubMed: 18344325]

38. Bubb KL, et al. Scan of human genome reveals no new Loci under ancient balancing selection. Genetics. 2006; 173:2165–2177. [PubMed: 16751668]

39. Ossowski S, et al. Sequencing of natural strains of Arabidopsis thaliana with short reads. Genome Res. 2008; 18:2024–2033. [PubMed: 18818371]

40. Bentley DR, et al. Accurate whole human genome sequencing using reversible terminator chemistry. Nature. 2008; 456:53–59. [PubMed: 18987734]

41. Gibbons JG, et al. Benchmarking next-generation transcriptome sequencing for functional and evolutionary genomics. Mol Biol Evol. 2009; 26:2731–2744. [PubMed: 19706727]

42. Smith AD, Xuan Z, Zhang MQ. Using quality scores and longer reads improves accuracy of Solexa read mapping. BMC Bioinformatics. 2008; 9:128. [PubMed: 18307793]

43. Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 2008; 18:821–829. [PubMed: 18349386]

44. Shimodaira H, Hasegawa M. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Mol Biol Evol. 1999; 16:1114–1116.

45. Goffeau A, et al. Life with 6000 genes. Science. 1996; 274:546, 563–547. [PubMed: 8849441]

46. Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES. Sequencing and comparison of yeast species to identify genes and regulatory elements. Nature. 2003; 423:241–254. [PubMed: 12748633]

47. Morgenstern B. DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment. Bioinformatics. 1999; 15:211–218. [PubMed: 10222408]

48. Huelsenbeck JP, Ronquist F. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics. 2001; 17:754–755. [PubMed: 11524383]

49. Ronquist F, Huelsenbeck JP. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics. 2003; 19:1572–1574. [PubMed: 12912839]

50. Altekar G, Dwarkadas S, Huelsenbeck JP, Ronquist F. Parallel Metropolis coupled Markov chain Monte Carlo for Bayesian phylogenetic inference. Bioinformatics. 2004; 20:407–415. [PubMed: 14960467]

51. Lanave C, Preparata G, Saccone C, Serio G. A new method for calculating evolutionary substitution rates. J Mol Evol. 1984; 20:86–93. [PubMed: 6429346]

52. Swofford, DL. PAUP*: Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4.0b10. Sinauer; Sunderland, MA: 2002.

53. Posada D, Crandall KA. MODELTEST: testing the model of DNA substitution. Bioinformatics. 1998; 14:817–818. [PubMed: 9918953]

54. Drummond AJ, Rambaut A. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol Biol. 2007; 7:214. [PubMed: 17996036]

55. Nei M, Gojobori T. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol. 1986; 3:418–426. [PubMed: 3444411]

56. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 2007; 24:1586–1591. [PubMed: 17483113]

57. Rozas J, Sanchez-DelBarrio JC, Messeguer X, Rozas R. DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics. 2003; 19:2496–2497. [PubMed: 14668244]

58. Ashburner M, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000; 25:25–29. [PubMed: 10802651]

59. Goldstein AL, McCusker JH. Three new dominant drug resistance cassettes for gene disruption in Saccharomyces cerevisiae. Yeast. 1999; 15:1541–1553. [PubMed: 10514571]

60. Guldener U, Heck S, Fielder T, Beinhauer J, Hegemann JH. A new efficient gene disruption cassette for repeated use in budding yeast. Nucleic Acids Res. 1996; 24:2519–2524. [PubMed: 8692690]

61. Gietz RD, Schiestl RH. Transforming yeast with DNA. Methods Mol Cell Biol. 1995; 5:255–269.

Hittinger *et al.*
Black & White;
Two columns;
No relationship
between panels;
Would also be acceptable
as 1-column vertical;
chris.hittinger@
ucdenver.edu
Thanks!

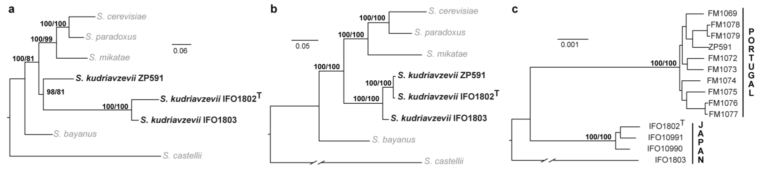


Figure 1. The functional and non-functional *GAL* gene networks share a common ancestor within the *S. kudriavzevii* lineage

Phylogeny of functional *GAL* genes and pseudogenes (**a**). Genome-wide consensus phylogeny of *Saccharomyces* (**b**) and representative *S. kudriavzevii* strains (**c**). Support values are Bayesian posterior probabilities/maximum likelihood (ML) bootstrap values. Scales show ML-estimated substitutions/site. Despite monophyly of the IFO1802^T and IFO1803 *GAL* pseudogenes (**a**), only two of 2,734 tested non-*GAL* genes (*SRL2* and *GYP5*, both *GAL*-adjacent) are monophyletic (Shimodaira-Hasegawa tests, $P < 0.05$)[44] (**b**). Only six of 1,642 genes (*YBR159W*, *YGL100W*, *YJR006W*, *YJR013W*, *YKL077W*, and *YPR071W*) reject[44] the monophyly of both the Portuguese and non-IFO1803 Japanese populations (**c**).

Hittinger *et al.*
Black & White;
One-and-a-half columns
(136 mm);
No relationship
between panels;
chris.hittinger@
ucdenver.edu
Thanks!

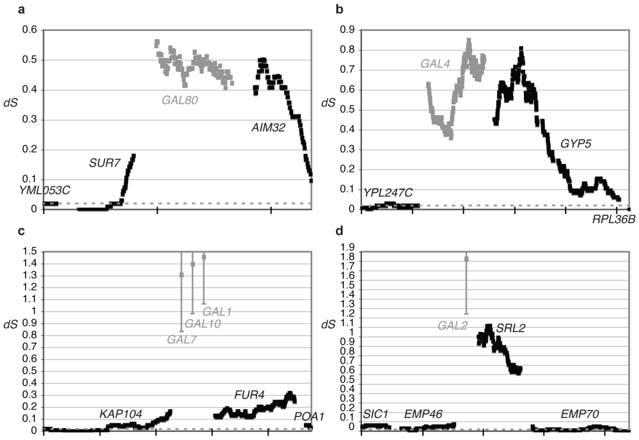


Figure 2. The divergence of the *GAL* loci sharply contrasts with the rest of the genome

Sliding window estimates (100 sites, step of one) of divergence at synonymous sites (*dS*) for: *GAL80* (**a**), *GAL4* (**b**), *GAL7/GAL10/GAL1* (**c**), and *GAL2* (**d**). *GAL* genes are gray, x-axis ticks represent 1,000 aligned bps, and a dashed line shows the genome-wide average *dS* of 0.021. Error bars are 95% binomial confidence intervals where few sites were available (Table S2). Note that elevation of *dS* around the *GAL* genes strongly affects some linked sequences, while regions of sustained but moderate *dS* elevation (*e.g. FUR4* and *GYP5*) provide evidence for ongoing balancing selection.

Hittinger *et al.*
Color;
One column;
No relationship
between panels;
chris.hittinger@
ucdenver.edu
Thanks!

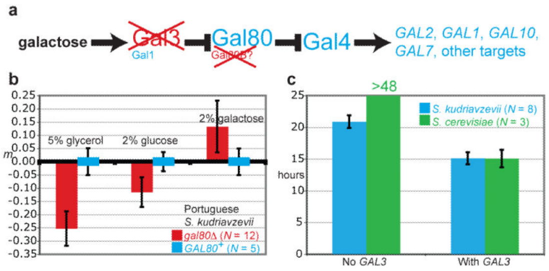
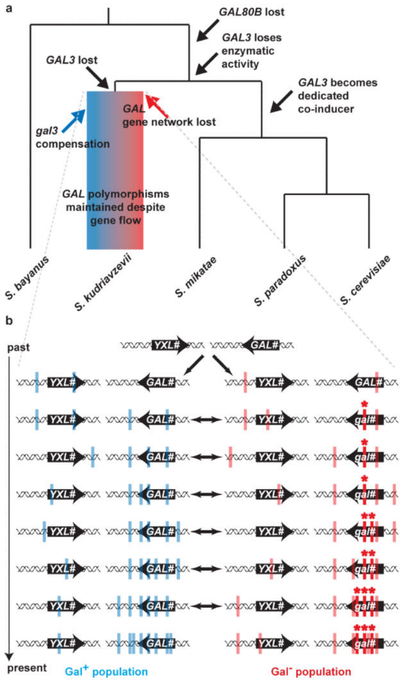


Figure 3. Key roles of the Gal3 co-inducer and Gal80 co-repressor

Reduced but functional Portuguese *S. kudriavzevii GAL* network (blue) and ancestral network (red) (**a**). Selection coefficients (m ± s.d.) acting on Portuguese *S. kudriavzevii gal80*Δ (red) relative to isogenic controls (blue) (**b**). The fitness defect in the absence of galactose would prevent invasion of established Gal^{-} populations by functional *GAL* alleles, other than *GAL80*. Hours (± s.d.) to first doubling after transferring stationary cultures to 2% galactose (blue, Portuguese *S. kudriavzevii*; green, *S. cerevisiae*) (**c**). Note that the induction defect is rescued by *S. cerevisiae GAL3* but that *GAL3* is less important for induction in *S. kudriavzevii*.



Hittinger *et al.*
Color;
One column;
panel a is connected
to b by the dashed lines;
chris.hittinger@
ucdenver.edu
Thanks!

Figure 4. Regulatory upheaval and the origin and maintenance of the *GAL* polymorphisms

Key changes in the *Saccharomyces GAL* gene network, including the origin of Gal$^-$ (red) and Gal$^+$ (blue) populations of *S. kudriavzevii* with some gene flow between them (gradient); the order of some events is uncertain (**a**). Model showing population-specific variation (light red or blue) arising as mutations and the elimination of variation from most of the genome (*YXL#*) during gene flow (two-headed arrows) (**b**). Note that once inactivating mutations (dark red with asterisks) formed *GAL* pseudogenes, gene flow was prevented within and reduced at linked sequences, resulting in the accumulation of lineage-specific variation at the *GAL* loci (*GAL#*).

Table 1

Several genes involved in amino acid metabolism and transport are significantly divergent and may be under balancing selection.

Systematic	Common	P	dS	dN	S	N	Description
YNL268W	*LYP1*	*P*<E-10	0.3498	0.0097	342	840	Lysine permease
Skud1324.2	*SAM4* homolog	2.9E-03	0.2123	0.0054	34	185	S-adenosylmethionine-homocysteine methyltransferase
YJR148W	*BAT2*	*P*<E-10	0.1420	0.0103	322	797	Branched-chain amino acid aminotransferase
YMR170C	*ALD2*	*P*<E-10	0.1141	0.0120	365	940	Cytoplasmic aldehyde dehydrogenase
YDR037W	*KRS1*	4.1E-08	0.0844	0.0028	413	1072	Lysyl-tRNA synthetase
YJL212C	*OPT1*	*P*<E-10	0.0793	0.0017	623	1762	Oligopeptide transporter
Skud2049.2	*CHA4* homolog	2.7E-08	0.0781	0.0169	497	1120	Transcription factor regulating amino acid catabolism
Skud1969.2	*SDL1* homolog	2.3E-02	0.0689	0.0082	268	743	L-serine dehydratase
Skud2049.3	*CAR2* homolog	7.9E-03	0.0646	0.0073	347	964	L-ornithine transaminase
YKR039W	*GAP1*	8.6E-04	0.0633	0.0064	428	974	General amino acid permease
YFL055W	*AGP3*	3.1E-02	0.0543	0.0043	489	1185	Low-affinity amino acid permease

Divergence of these genes is shown between the Japanese (IFO1802^{T}) and Portuguese (ZP591) reference strains at synonymous (*dS*) and non-synonymous (*dN*) sites, as well as the number of sites (*S* and *N*, respectively) and a Bonferroni-corrected *P* value calculated from the Poisson distribution of synonymous substitutions.