nature Vol 464|4 March 2010|doi:10.1038/nature08791

ARTICLES

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] & Antonis Rokas[4]

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 previously unrecognized type of intraspecific multi-locus genetic variation that has been maintained over a vast period. The galactose (*GAL*) utilization gene network of *Saccharomyces kudriavzevii*, a relative of brewer's yeast, 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.

Genetic variation fuels evolution by providing the diversity on which natural selection acts. In contrast with the more common directional and stabilizing forms of selection, balancing selection favours 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[1–9]. In some cases, alternative alleles of single loci have been maintained for millions of years and span species boundaries, for example at the *MHC* locus in mammals[2,3].

In contrast to conventional single-locus balanced polymorphisms, the possibility that a complex multi-locus gene network may be maintained in alternative 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 between genes[6,10–13]. Therefore, keeping co-adapted interacting alleles or gene complexes together is probably crucial 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], by means of reduced gene flow between populations at the early stages of speciation[16,17], and even through inbreeding, as suggested in some theoretical studies[18,19].

Here we report a previously unrecognized 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. Although 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 without speciation. Moreover, their interacting genes comprise alternative 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 catalyse 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 model for studying eukaryotic regulatory networks[22–24], evolutionary processes[25–28] and systems biology[29].

The key features of the *GAL* gene network are preserved among most members of the genus *Saccharomyces sensu stricto* (hereafter referred to as *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 use 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$^+$ strains) as their sole carbon source[30].

Because 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$^+$

[1]Department of Biochemistry and Molecular Genetics, University of Colorado School of Medicine, Aurora, Colorado 80045, USA. [2]Center for Genome Sciences, Department of Genetics, Washington University in St Louis, School of Medicine, St Louis, Missouri 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, Tennessee 37235, USA.

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 (Supplementary Table 1), and crosses between a Gal$^-$ Japanese strain and a Gal$^+$ Portuguese strain produced viable (82%) spores that showed normal gene segregation (Supplementary Fig. 1). 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 in other *Saccharomyces* yeasts[31,32]. To determine whether the Gal$^+$ strains of *S. kudriavzevii* had obtained any genes from other species, we generated millions of mostly 36-base-pair (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* (Supplementary Fig. 2 and Supplementary Data 1). 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 (Fig. 1a, Supplementary Figs 3a and 4, and Supplementary Table 2). 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 about 89% as old as the species (Supplementary Fig. 3). This finding is remarkable, given the lack of any apparent pre-zygotic 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 as a result of 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 whether 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$^{\mathrm{T}}$ (refs 25, 34), confidently determining about 80% of the orthologous bases in the genome of each strain (Supplementary Table 1). 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 (Fig. 1 and Supplementary Fig. 4). 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 (d_s) between the Japanese (IFO1802$^{\mathrm{T}}$) and Portuguese reference (ZP591) strains across all annotated genes is 0.021 (Supplementary Data 2), whereas divergence between 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 more than 0.001, which also seems to be typical for wild populations of the genus[32]. Although the discovery and genome sequencing of additional strains might help in addressing 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 greater sequence divergence between the Japanese and Portuguese reference strains ($d_s = 0.159$, $P < 10^{-5}$). In every case, the region of elevated divergence is centred on the *GAL* gene(s) ($d_s = 0.939$), and there is no functional connection between adjacent genes. Moreover, levels of divergence rapidly decline towards the genome-wide background average, sometimes in the middle of open reading frames (Fig. 2 and Supplementary Fig. 5). 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 fitter 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 (Supplementary Data 2). The most

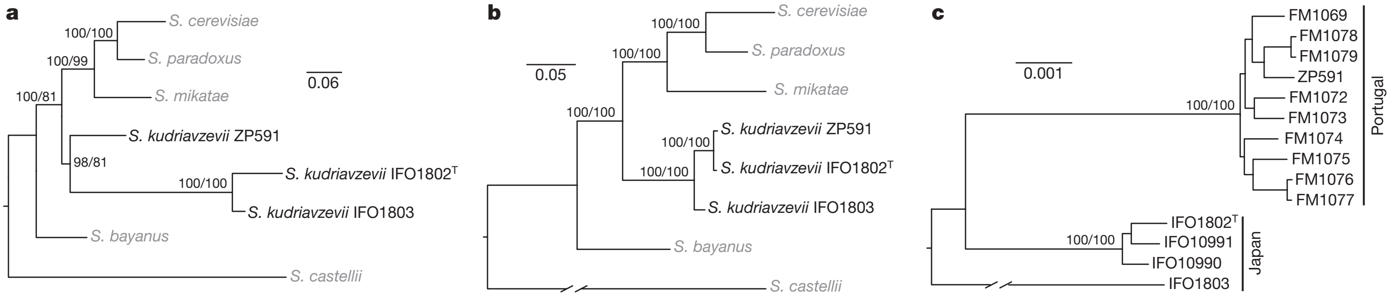

Figure 1 | The functional and non-functional *GAL* gene networks share a common ancestor within the *S. kudriavzevii* lineage. **a**, Phylogeny of functional *GAL* genes and pseudogenes. **b**, **c**, Genome-wide consensus phylogenies of *Saccharomyces* (**b**) and representative *S. kudriavzevii* strains (**c**). Support values are Bayesian posterior probabilities/maximum-likelihood (ML) bootstrap values. Scales show ML-estimated substitutions per site. Despite monophyly of the IFO1802$^{\mathrm{T}}$ 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**).

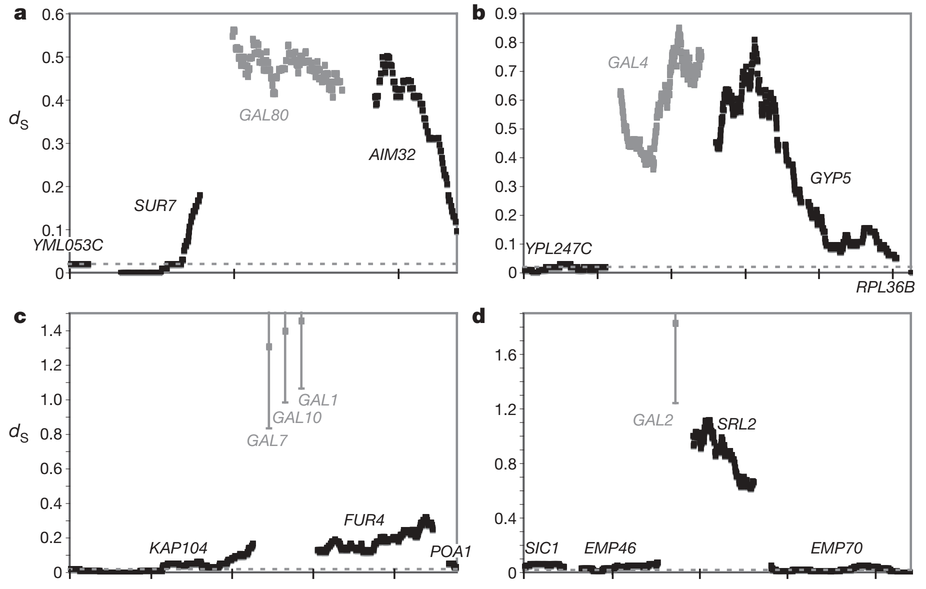

Figure 2 | The divergence of the *GAL* loci contrasts sharply with the rest of the genome. Sliding-window estimates (100 sites, step 1) of divergence at synonymous sites (d_s) for *GAL80* (**a**), *GAL4* (**b**), *GAL7*/*GAL10*/*GAL1* (**c**) and *GAL2* (**d**). *GAL* genes are shown in grey, *x*-axis ticks represent 1,000 aligned base pairs, and a dashed line shows the genome-wide average d_s of 0.021. Error bars show 95% binomial confidence intervals when few sites were available (Supplementary Table 2). Note that increase in d_s around the *GAL* genes strongly affects some linked sequences, whereas regions of sustained but moderate d_s elevation (for example *FUR4* and *GYP5*) provide evidence for continuing balancing selection.

divergent of these is the amino-acid permease *LYP1* with a highly elevated d_s 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 fivefold enrichment ($P < 10^{-5}$) suggests that gene flow has been selectively decreased 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 in which 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 five 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 *S. cerevisiae* does: 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 (that is, the *GAL* loci segregated independently; Supplementary Fig. 1).

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].

Last, although 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 and Supplementary Table 3)[22,23,27]. Indeed, *S. kudriavzevii gal80Δ* 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 smaller extent in glucose-repressing conditions,

Table 1 | Several genes involved in amino-acid metabolism and transport are significantly divergent and may be under balancing selection

Gene name		*P*	d_s	d_n	*S*	*N*	Description
Systematic	Common						
YNL268W	*LYP1*	$<10^{-10}$	0.3498	0.0097	342	840	Lysine permease
Skud1324.2	*SAM4* homologue	2.9×10^{-3}	0.2123	0.0054	34	185	*S*-Adenosylmethionine-homocysteine methyltransferase
YJR148W	*BAT2*	$<10^{-10}$	0.1420	0.0103	322	797	Branched-chain amino acid aminotransferase
YMR170C	*ALD2*	$<10^{-10}$	0.1141	0.0120	365	940	Cytoplasmic aldehyde dehydrogenase
YDR037W	*KRS1*	4.1×10^{-8}	0.0844	0.0028	413	1072	Lysyl-tRNA synthetase
YJL212C	*OPT1*	$<10^{-10}$	0.0793	0.0017	623	1762	Oligopeptide transporter
Skud2049.2	*CHA4* homologue	2.7×10^{-8}	0.0781	0.0169	497	1120	Transcription factor regulating amino acid catabolism
Skud1969.2	*SDL1* homologue	2.3×10^{-2}	0.0689	0.0082	268	743	L-Serine dehydratase
Skud2049.3	*CAR2* homologue	7.9×10^{-3}	0.0646	0.0073	347	964	L-Ornithine transaminase
YKR039W	*GAP1*	8.6×10^{-4}	0.0633	0.0064	428	974	General amino acid permease
YFL055W	*AGP3*	3.1×10^{-2}	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 (d_s) and non-synonymous (d_n) 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.

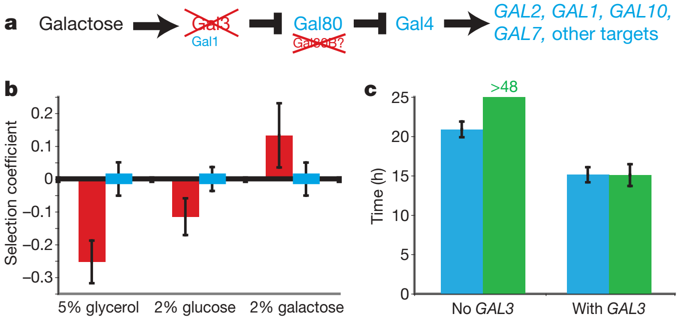

Figure 3 | Key roles of the Gal3 co-inducer and Gal80 co-repressor. **a**, Decreased but functional Portuguese *S. kudriavzevii* network (blue) and ancestral network (red). **b**, Malthusian selection coefficients acting on Portuguese *S. kudriavzevii gal80Δ* (red; $n = 12$) relative to isogenic *GAL80*$^+$ controls (blue; $n = 5$). The fitness defect in the absence of galactose would prevent the invasion of established Gal$^-$ populations by functional *GAL* alleles other than *GAL80*. **c**, Time to first doubling after transfer of stationary cultures to 2% galactose (blue, Portuguese *S. kudriavzevii* ($n = 8$); green, *S. cerevisiae* ($n = 3$)). The induction defect is rescued by *S. cerevisiae GAL3*, but *GAL3* is less important for induction in *S. kudriavzevii*. Results in **b** and **c** are shown as means ± s.d.

$P < 10^{-2}$; Fig. 3b). Thus, in genetic backgrounds that cause partial gene networks to be expressed constitutively (that is, *GAL4*$^+$ *gal80Δ* strains), functional *GAL* alleles (other than *GAL80*) would be strongly selected against, thereby imposing a moderate fitness cost when averaged across all genetic backgrounds that an allele might encounter in the absence of galactose. Because *GAL80* and linked sequences also show 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 crucial roles of *GAL80* and *GAL3* in the maintenance of distinct *GAL* states in *S. kudriavzevii* led us to consider other evolutionary changes that were likely to have 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 paralogue 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 paralogue, 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* markedly increased the speed of its response to galactose ($P < 10^{-3}$), rivalling the rapid response of wild-type *S. cerevisiae*. Although the naturally *gal3* Portuguese strains of *S. kudriavzevii* experience some delay in their response to galactose, they react much more quickly than *gal3Δ* mutants of *S. cerevisiae*[23] ($P < 10^{-2}$), suggesting that the Portuguese *S. kudriavzevii GAL* gene network contains compensatory mutations that partly mitigate the loss of *GAL3* and may have been crucial in 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 alternative states of six genes that must interact with specific alleles for optimal fitness. The numerous cases of long-term balancing selection[2,4,6–9], complex genetic interactions[6,10–13] 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 alternative states of a multi-locus gene network within a single species.

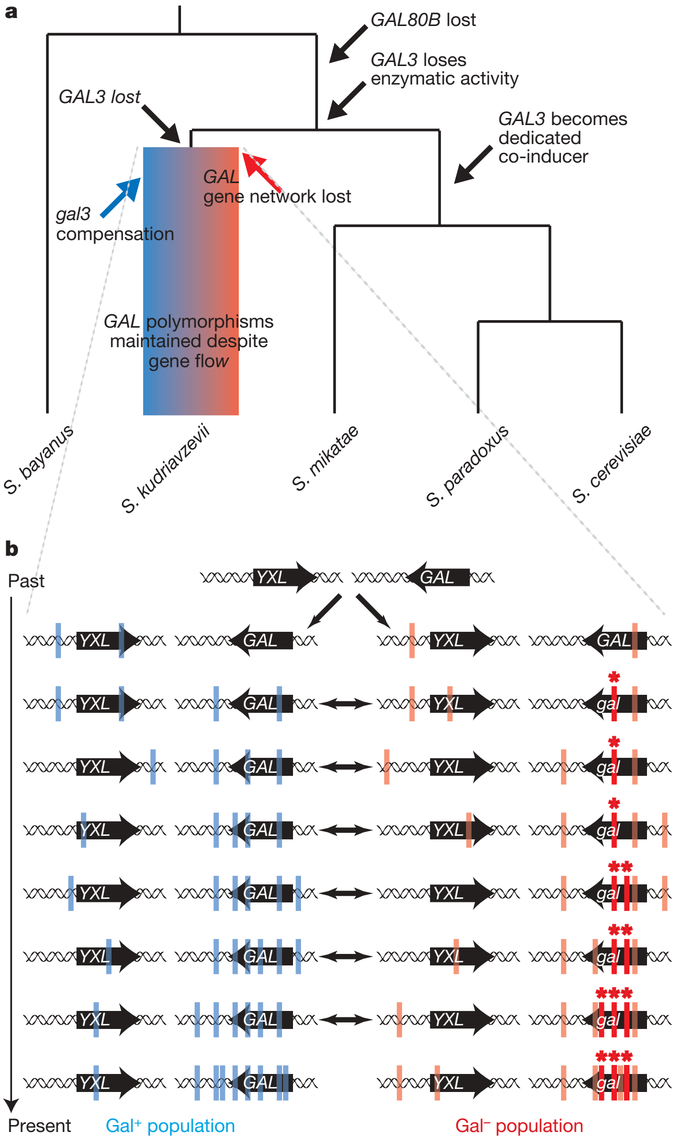

Figure 4 | Regulatory upheaval and the origin and maintenance of the *GAL* polymorphisms. **a**, Key changes in the *Saccharomyces GAL* gene network, including the origin of Gal$^-$ (red) and Gal$^+$ (blue) populations of *S. kudriavzevii* with some gene flow between them (gradient); the order of some events is uncertain. **b**, 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). Note that once inactivating mutations (dark red with asterisks) had formed *GAL* pseudogenes, gene flow was prevented within and decreased at linked sequences, resulting in the accumulation of lineage-specific variation at the *GAL* loci (*GAL*).

Genome-wide scans for cases of balancing selection either have 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 alternative 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 in the *GAL* gene network of *S. kudriavzevii*, and it is unknown whether these regions interact to contribute to phenotypic variation. Most BuGNPs are probably more quantitative (and are therefore harder to detect) than the one presented here, but we expect that gene networks could be maintained in alternative 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 alternative 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.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 19 May; accepted 15 December 2009.
Published online 17 February 2010.

1. Levene, H. Genetic equilibrium when more than one ecological niche is available. *Am. Nat.* **87,** 331–333 (1953).
2. Hughes, A. L. & Nei, M. Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature* **335,** 167–170 (1988).
3. Klein, J., Sato, A., Nagl, S. & O'hUigín, C. Molecular trans-species polymorphism. *Annu. Rev. Ecol. Syst.* **29,** 1–21 (1998).
4. Stahl, E. A., Dwyer, G., Mauricio, R., Kreitman, M. & Bergelson, J. Dynamics of disease resistance polymorphism at the *Rpm1* locus of *Arabidopsis. Nature* **400,** 667–671 (1999).
5. Colosimo, P. F. *et al.* Widespread parallel evolution in sticklebacks by repeated fixation of Ectodysplasin alleles. *Science* **307,** 1928–1933 (2005).
6. Kroymann, J. & Mitchell-Olds, T. Epistasis and balanced polymorphism influencing complex trait variation. *Nature* **435,** 95–98 (2005).
7. Charlesworth, D. Balancing selection and its effects on sequences in nearby genome regions. *PLoS Genet.* **2,** e64 (2006).
8. Mitchell-Olds, T., Willis, J. H. & Goldstein, D. B. Which evolutionary processes influence natural genetic variation for phenotypic traits? *Nature Rev. Genet.* **8,** 845–856 (2007).
9. Storz, J. F. *et al.* The molecular basis of high-altitude adaptation in deer mice. *PLoS Genet.* **3,** e45 (2007).
10. Hawthorne, D. J. & Via, S. Genetic linkage of ecological specialization and reproductive isolation in pea aphids. *Nature* **412,** 904–907 (2001).
11. Brem, R. B., Storey, J. D., Whittle, J. & Kruglyak, L. Genetic interactions between polymorphisms that affect gene expression in yeast. *Nature* **436,** 701–703 (2005).
12. Steiner, C. C., Weber, J. N. & Hoekstra, H. E. Adaptive variation in beach mice produced by two interacting pigmentation genes. *PLoS Biol.* **5,** e219 (2007).
13. Gerke, J., Lorenz, K. & Cohen, B. Genetic interactions between transcription factors cause natural variation in yeast. *Science* **323,** 498–501 (2009).
14. Navarro, A. & Barton, N. H. The effects of multilocus balancing selection on neutral variability. *Genetics* **161,** 849–863 (2002).
15. Dobzhansky, T. & Pavlovsky, O. Interracial hybridization and breakdown of coadapted gene complexes in *Drosophila paulistorum* and *Drosophila willistoni*. *Proc. Natl Acad. Sci. USA* **44,** 622–629 (1958).
16. Wu, C. I. & Ting, C. T. Genes and speciation. *Nature Rev. Genet.* **5,** 114–122 (2004).
17. Turner, T. L., Hahn, M. W. & Nuzhdin, S. V. Genomic islands of speciation in *Anopheles gambiae*. *PLoS Biol.* **3,** e285 (2005).
18. Wright, S. Genic and organismic selection. *Evolution* **34,** 825–843 (1980).
19. Neher, R. A. & Shraiman, B. I. Competition between recombination and epistasis can cause a transition from allele to genotype selection. *Proc. Natl Acad. Sci. USA* **106,** 6866–6871 (2009).
20. Clark, R. M. *et al.* Common sequence polymorphisms shaping genetic diversity in *Arabidopsis thaliana*. *Science* **317,** 338–342 (2007).
21. Coyne, J. A. & Orr, H. A. *Speciation* (Sinauer Associates, 2004).
22. Johnston, M. A model fungal gene regulatory mechanism: the GAL genes of *Saccharomyces cerevisiae. Microbiol. Rev.* **51,** 458–476 (1987).
23. Bhat, P. J. & Murthy, T. V. Transcriptional control of the GAL/MEL regulon of yeast *Saccharomyces cerevisiae*: mechanism of galactose-mediated signal transduction. *Mol. Microbiol.* **40,** 1059–1066 (2001).
24. Ptashne, M. & Gann, A. *Genes and Signals* (Cold Spring Harbor Laboratory Press, 2001).
25. Hittinger, C. T., Rokas, A. & Carroll, S. B. Parallel inactivation of multiple GAL pathway genes and ecological diversification in yeasts. *Proc. Natl Acad. Sci. USA* **101,** 14144–14149 (2004).
26. Martchenko, M., Levitin, A., Hogues, H., Nantel, A. & Whiteway, M. Transcriptional rewiring of fungal galactose-metabolism circuitry. *Curr. Biol.* **17,** 1007–1013 (2007).
27. MacLean, R. C. Pleiotropy and GAL pathway degeneration in yeast. *J. Evol. Biol.* **20,** 1333–1338 (2007).
28. Hittinger, C. T. & Carroll, S. B. Gene duplication and the adaptive evolution of a classic genetic switch. *Nature* **449,** 677–681 (2007).
29. Raj, A. & van Oudenaarden, A. Nature, nurture, or chance: stochastic gene expression and its consequences. *Cell* **135,** 216–226 (2008).
30. Sampaio, J. P. & 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.* **74,** 2144–2152 (2008).
31. Doniger, S. W. *et al.* A catalog of neutral and deleterious polymorphism in yeast. *PLoS Genet.* **4,** e1000183 (2008).
32. Liti, G. *et al.* Population genomics of domestic and wild yeasts. *Nature* **458,** 337–341 (2009).
33. Gonzalez, S. S., Barrio, E., Gafner, J. & Querol, A. Natural hybrids from *Saccharomyces cerevisiae*, *Saccharomyces bayanus* and *Saccharomyces kudriavzevii* in wine fermentations. *FEMS Yeast Res.* **6,** 1221–1234 (2006).
34. Cliften, P. *et al.* Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting. *Science* **301,** 71–76 (2003).
35. Bensasson, D., Zarowiecki, M., Burt, A. & Koufopanou, V. Rapid evolution of yeast centromeres in the absence of drive. *Genetics* **178,** 2161–2167 (2008).
36. Schacherer, J., Shapiro, J. A., Ruderfer, D. M. & Kruglyak, L. Comprehensive polymorphism survey elucidates population structure of *Saccharomyces cerevisiae. Nature* **458,** 342–345 (2009).
37. Tsai, I. J., Bensasson, D., Burt, A. & Koufopanou, V. Population genomics of the wild yeast *Saccharomyces paradoxus*: quantifying the life cycle. *Proc. Natl Acad. Sci. USA* **105,** 4957–4962 (2008).
38. Bubb, K. L. *et al.* Scan of human genome reveals no new loci under ancient balancing selection. *Genetics* **173,** 2165–2177 (2006).
39. Ossowski, S. *et al.* Sequencing of natural strains of *Arabidopsis thaliana* with short reads. *Genome Res.* **18,** 2024–2033 (2008).
40. Bentley, D. R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456,** 53–59 (2008).
41. Gibbons, J. G. *et al.* Benchmarking next-generation transcriptome sequencing for functional and evolutionary genomics. *Mol. Biol. Evol.* **26,** 2731–2744 (2009).
42. Smith, A. D., Xuan, Z. & Zhang, M. Q. Using quality scores and longer reads improves accuracy of Solexa read mapping. *BMC Bioinformatics* **9,** 128 (2008).
43. Zerbino, D. R. & Birney, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* **18,** 821–829 (2008).
44. Shimodaira, H. & Hasegawa, M. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Mol. Biol. Evol.* **16,** 1114–1116 (1999).

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

Acknowledgements We thank E. Barrio; B. L. Williams; the Centraalbureau voor Schimmelcultures, Netherlands and the National Institute of Technology and Evaluation Biological Resource Center, Japan for providing strains; J. Fay for critical reading of the manuscript; and the Howard Hughes Medical Institute (HHMI) and S. B. Carroll for supporting preliminary experiments when C.T.H. was a HHMI Predoctoral Fellow. This work was also supported by the James S. McDonnell Foundation (C.T.H., M.J.); the National Institutes of Health (C.T.H., 2T32HG00045; M.J., 5R01GM032540); the Fundação para a Ciência e a Tecnologia, Portugal (P.G. and J.P.S., PTDC/BIA-BDE/71734/2006); the Searle Scholars Program (A.R.); the National Science Foundation (A.R., DEB-0844968) and Vanderbilt University (A.R.). C.T.H. is the Maclyn McCarty Fellow of the Helen Hay Whitney Foundation.

Author Contributions C.T.H., P.G., J.P.S., M.J. and A.R. conceived and designed experiments; C.T.H., P.G. and J.D. performed experiments; C.T.H. sequenced and assembled genomes; C.T.H. and A.R. analysed data; C.T.H., M.J. and A.R. wrote the paper with advice and consent from all authors.

Author Information All short sequencing reads are deposited with Sequence Read Archive at the National Center for Biotechnology Information under accession number SRP001457 of SRA010159; the IFO1803 and ZP591 *GAL* sequences are deposited in GenBank under accession numbers GU299171–GU299178. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to M.J. (mark.johnston@ucdenver.edu).

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 (Supplementary Table 1). We attempted to create monosporic derivatives for all Portuguese strains by dissecting and isolating individual spores and allowing them to re-form 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 that 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, about 5 μg of gDNA was sonicated and ligated to Solexa sequencing adapters (Illumina) using either the manufacturer's kit or our custom protocol[41] (using gDNA instead of cDNA). Some libraries were prepared with multiplexing barcode adaptor pairs with an added 'T' overhang for ligation; reads were processed to sort by and remove the 'NT' barcodes and overhangs before analysis (Supplementary Table 1). Libraries were sequenced using the Solexa Genome Analyzer I or II and Pipeline Software (Illumina) in accordance with the manufacturer's instructions[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 (http://www.yeastgenome.org): *S. cerevisiae*[45], *S. paradoxus*[46], *Saccharomyces mikatae*[46], *S. kudriavzevii* IFO1802^T (ref. 34) and *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 by using RMAP v. 0.45 (ref. 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 artefacts. Complete data are presented in Supplementary Data 1, 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 (Supplementary Fig. 2). The Portuguese reference strain ZP591 was also analysed with one, two, three, four and five mismatches (Supplementary Data 1).

Reference-guided assemblies. Because the average short-read coverage of each genome was about tenfold, assembly of genome sequences required a reference genome. We started with the 3.4-fold coverage Sanger-sequenced genome of *S. kudriavzevii* IFO1802^T (ref. 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 and 9080 for precise boundaries). These alternative contigs overlap and are syntenic with the IFO1802^T contigs. Because we required all mapped reads to be unique, using these alternative 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 (ref. 42) with three mismatches and a quality filter of 5, which records the positions of all reads that can be uniquely mapped to the reference genome with three 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 five 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): first, one of the four possible bases must have $p_c < 10^{-5}$; second, there must be no other base with $p_c < 10^{-5}$; and third, 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 as a result of heterozygosity or as a result of 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 as being below 5×10^{-4} errors per base pair by reassembling 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 because the 3.4-fold coverage draft sequence has been estimated to contain at least 10^{-4} errors per base pair[34].

Prealigned ORF sequences were generated from the above reference-guided genome assemblies for each strain by using the ORF annotations from the reference genome. For genes wholly or partly included in the alternative 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. Although 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 (ref. 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 by 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 data set (Supplementary Fig. 4) was processed manually after alignment with DIALIGN, whereas the codon-based procedure was also used for all other phylogenetic data matrices.

Phylogenetics. Bayesian inference was conducted by using MRBAYES v. 3.1.2 (refs 48–50), assuming a GTR model of nucleotide substitution[51] and allowing for rate heterogeneity between 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 α 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 with PAUP* v. 4.0.b10 (ref. 52). The best-fit model of nucleotide evolution was estimated with MODELTEST v. 3.7 (ref. 53). Clade support was assessed by using 100 replicates of non-parametric bootstrap resampling. Agreement with the dominant genome-wide trees (Fig. 1b, c) was assessed for individual genes that contained ten or more parsimony informative sites by using the Shimodaira–Hasegawa test[44] as implemented in PAUP* v. 4.0.b10 (ref. 52).

Coalescence estimation. Rooted, time-measured phylogenies were inferred with BEAST v. 1.4.8 (ref. 54). Because the fungal and yeast fossil records are very poor and reliable fossil calibration points are unavailable, all branches were estimated in units of substitution per site. We assumed a GTR+GAMMA model of sequence evolution and the uncorrelated log-normal relaxed clock model. The Yule tree prior was chosen for the analysis of the *GAL* pseudogenes and functional genes (Supplementary Fig. 3a) and for the *Saccharomyces* yeast clade (Supplementary Fig. 3b). The coalescent (constant size) tree prior was used for the analysis of representative *S. kudriavzevii* strains (Supplementary Fig. 3c). Two to eight independent runs of 2,000,000 to 10,000,000 generations were performed 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 (more than 100) and verified visually 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 with the analysis of the *GAL* pseudogenes and functional genes (which contained 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 (d_s) and several other parameters were calculated by using both the ML-F3X4 and modified Nei-Gojobori[55] estimates implemented by CODEML of PAML v. 4.1 (ref. 56) for all annotated ORFs for which there were more than ten complete codons aligned. These two estimates produced similar genome-wide estimates of d_s, so we used only 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 and Supplementary Data 2). Unless otherwise stated, pairwise comparisons were between IFO1802^T (Gal$^-$ reference) and ZP591 (Gal$^+$ reference).

Because the *GAL* loci are significantly more divergent than the rest of the genome, we examined the extent of these highly divergent islands by graphing d_s along a sliding window. Position-based modified Nei–Gojobori estimates of d_s with a Jukes–Cantor correction were generated by using a one-site step and a 100-site window with DNASP v. 4.90.1 (ref. 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 by 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 (Supplementary Fig. 5).

Statistics. Experimental data were analysed with 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 and Supplementary Data 2) was assessed with the hypergeometric distribution on the pooled Gene Ontology (GO; 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* orthologue (when available) or homologue (for significant genes lacking a clear orthologue). 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 *hoΔ::natMX* or *hoΔ::kanMX* haploid progeny (Supplementary Table 4). Transformation of *S. kudriavzevii* was accomplished by using the standard lithium acetate protocol optimized for *S. cerevisiae*[61], except that incubations were performed at 22–23 °C and the heat shock was at 34 °C to accommodate the sensitivity of *S. kudriavzevii* to heat. FM1071-derived *ura3Δ* and *trp1Δ* auxotrophic strains were produced by replica-plating cells to 5-fluoroorotic acid or 5-fluoroanthranilic acid, 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 *hoΔ::kanMX* to *hoΔ::natMX* ($n = 5$; Supplementary Table 4). Each *hoΔ::kanMX gal80Δ* strain was competed against a single *hoΔ::natMX GAL80*$^+$ control strain lacking detectable defects, and each *hoΔ::natMX GAL80*$^+$ control strain was competed against the *hoΔ::kanMX GAL80*$^+$ progenitor of all strains. Each competition was conducted by mixing and co-culturing strains for about ten generations for 2 days in synthetic complete (SC) medium containing 2% glucose, performing colony counts of nat^r and kan^r colonies to establish starting frequencies, and inoculating fresh SC medium with various carbon sources at 1:1,000 dilution with aliquots from the same saturated medium[28]. Competitions were performed for 2 days in 2% glucose, for 4 days in 2% galactose, and for 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 described previously[28] (Fig. 3b).

To assess the effect of *ScerGAL3*$^+$ on the induction of *GAL* gene expression by galactose in a Gal$^+$ strain of *S. kudriavzevii*, we created a *trp1Δ::ScerGAL3*$^+$ and a control *trp1Δ* strain by targeting a *trp1Δ::ScerURA3*$^+$ strain of *S. kudriavzevii* with PCR products and selecting for 5-fluoroorotic acid resistance to replace *ScerURA3*$^+$ with the PCR products (Supplementary Table 4). 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 1 bp in a tract of 12 As 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 backcrossing 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 medium containing 2% raffinose. Induction of *GAL* gene expression was performed by inoculating the stationary phase cultures 1:20 in SC medium containing 2% galactose and measuring attenuance (D_{600}) values every 2 h 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 (Supplementary Table 4). These strains were sporulated, tetrads were dissected, and 297 monosporic segregants were recovered. In all, 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 (Supplementary Fig. 1), nor did the source of the parental mating types affect the offspring.

Triplicate *gal80Δ* and *GAL80*$^+$ green fluorescent protein-labelled strains of *S. cerevisiae* were constructed and competed against an otherwise identical blue fluorescent protein-labelled strain as described previously[28] (Supplementary Tables 3 and 4). Data reported are from a single experiment with quadruplicate biological replicates (from separate GFP colonies) of each genetically engineered strain, for a total of 12 replicates.

45. Goffeau, A. *et al.* Life with 6000 genes. *Science* 274, 546 563–567 (1996).
46. Kellis, M., Patterson, N., Endrizzi, M., Birren, B. & Lander, E. S. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423**, 241–254 (2003).
47. Morgenstern, B. DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment. *Bioinformatics* **15**, 211–218 (1999).
48. Huelsenbeck, J. P. & Ronquist, F. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **17**, 754–755 (2001).
49. Ronquist, F. & Huelsenbeck, J. P. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**, 1572–1574 (2003).
50. Altekar, G., Dwarkadas, S., Huelsenbeck, J. P. & Ronquist, F. Parallel Metropolis coupled Markov chain Monte Carlo for Bayesian phylogenetic inference. *Bioinformatics* **20**, 407–415 (2004).
51. Lanave, C., Preparata, G., Saccone, C. & Serio, G. A new method for calculating evolutionary substitution rates. *J. Mol. Evol.* **20**, 86–93 (1984).
52. Swofford, D. L. *PAUP*: Phylogenetic Analysis Using Parsimony (*and Other Methods)*. Version 4.0b10 edn (Sinauer, 2002).
53. Posada, D. & Crandall, K. A. MODELTEST: testing the model of DNA substitution. *Bioinformatics* **14**, 817–818 (1998).
54. Drummond, A. J. & Rambaut, A. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.* **7**, 214 (2007).
55. Nei, M. & Gojobori, T. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* **3**, 418–426 (1986).
56. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).
57. Rozas, J., Sanchez-DelBarrio, J. C., Messeguer, X. & Rozas, R. DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics* **19**, 2496–2497 (2003).
58. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genet.* **25**, 25–29 (2000).
59. Goldstein, A. L. & McCusker, J. H. Three new dominant drug resistance cassettes for gene disruption in *Saccharomyces cerevisiae*. *Yeast* **15**, 1541–1553 (1999).
60. Guldener, U., Heck, S., Fielder, T., Beinhauer, J. & Hegemann, J. H. A new efficient gene disruption cassette for repeated use in budding yeast. *Nucleic Acids Res.* **24**, 2519–2524 (1996).
61. Gietz, R. D. & Schiestl, R. H. Transforming yeast with DNA. *Methods Mol. Cell. Biol.* **5**, 255–269 (1995).