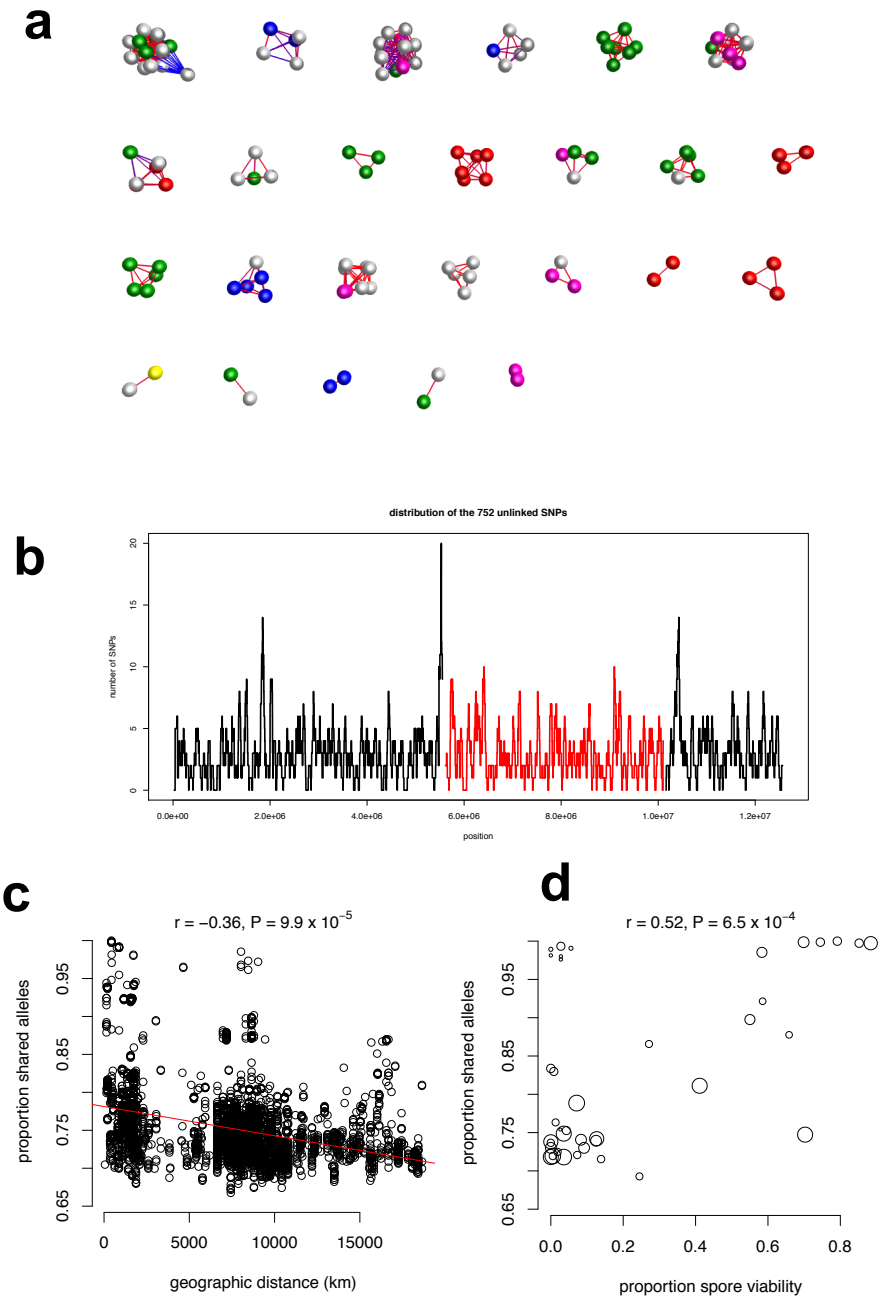
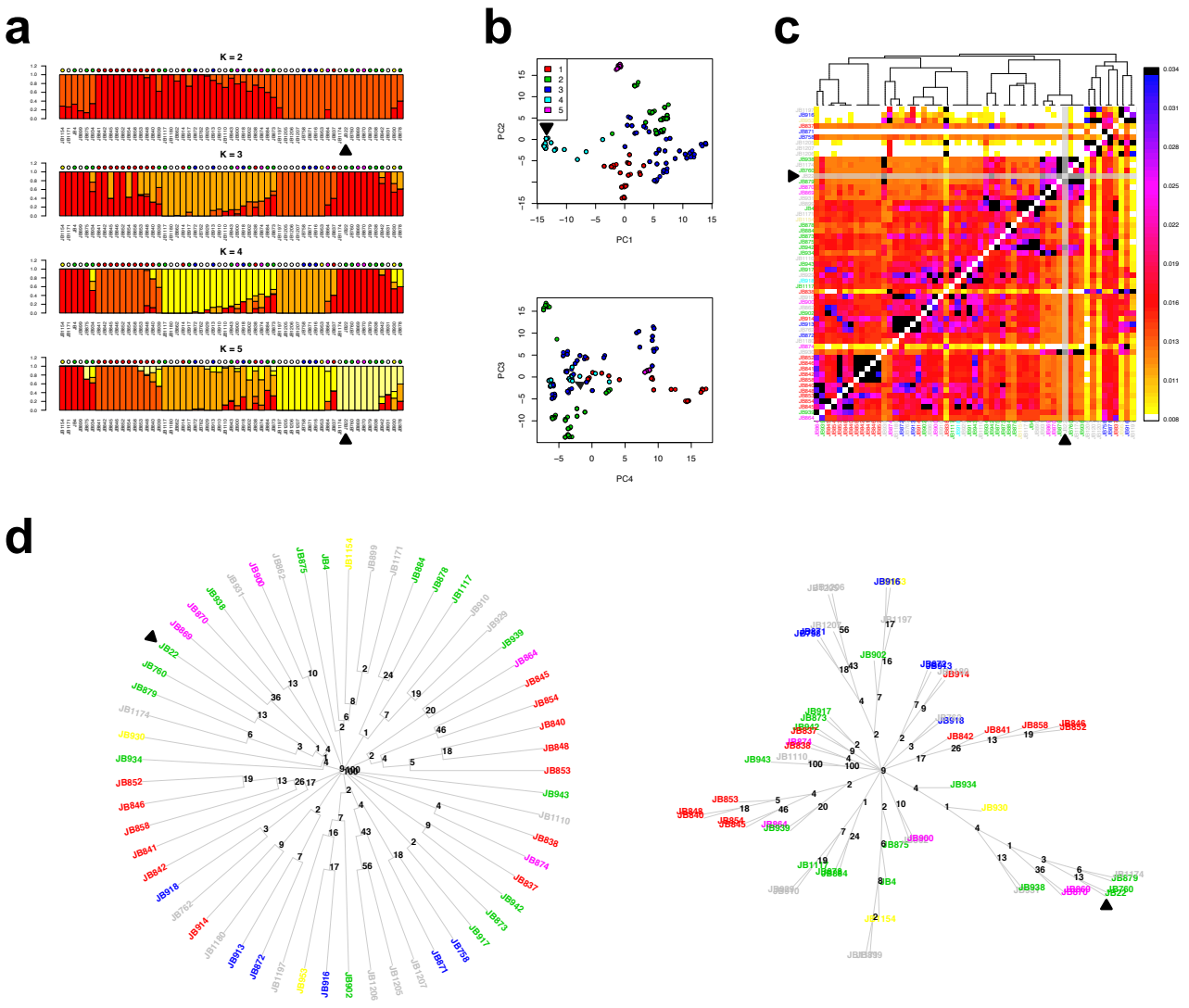


Supplementary Figure 1

Clonal clusters and isolation by distance.

a, For all strains, we calculated the number of allelic differences using all SNPs. Pairs with <150 SNPs were considered near-clonal, and these pairs were clustered using Markov clustering. Strains (spheres) are colored according to the continent that they were isolated from, with grey spheres indicating unknown locations. Coloring as in Fig. 1A; red (Americas), pink (Africa), green (Europe), blue (Asia), yellow (Australia). **b,** The 752 unlinked SNPs used for descriptions of population structure are evenly distributed across the genome. For each 50kb window of the genome, with 1kb step-size, we show the number of SNPs from the 752 unlinked set. Chromosome 1 and

3 are in black, and chromosome 2 is in red. We note a slight bias to the right side of chromosome 1, which contains 20 of these 752 SNPs. **c**, Genetic distance is correlated with geographic distance. For each pairwise comparison of the 161 strains, we calculated the proportion of shared alleles from the 752 unlinked SNPs ('drift distance') and the great circle distance (distance around the globe) between the locations that strains were collected from. A Mantel test with 10,000 re-samplings showed that these two matrices were anti-correlated ($r = -0.36$, $P = 9.9 \times 10^{-5}$). This correlation is also present when we use only the 57 non-clonal strains ($r = -0.28$, $P = 9.9 \times 10^{-5}$) **d**, Genetic distance is correlated with spore viability. For 43 crosses, we recorded the spore survival by tetrad analysis. Spore survival was correlated with the proportion of shared alleles from the 752 unlinked SNPs (Pearson's product-moment correlation, $r = 0.51$, $P = 6.4 \times 10^{-4}$). Some strains do not produce many viable spores even when mated to themselves (low self-cross viability). The plot represents this by scaling each circle size to the lowest self-cross viability of the parents, showing that all low viability outliers (top left of plot) have at least one low self-viability parent. When excluding crosses with the lowest self-cross viability < 0.3, the correlation between spore viability and genetic distance is stronger (Pearson $r = 0.76$, $P = 1.2 \times 10^{-7}$).

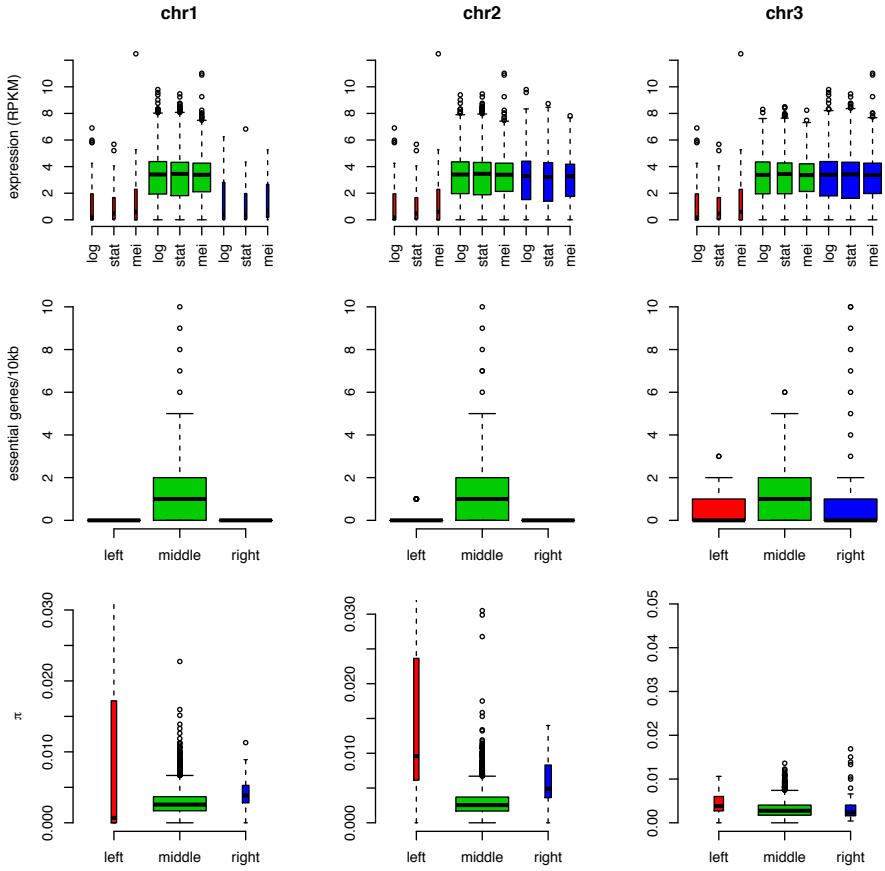


Supplementary Figure 2

Population structure and relatedness between strains.

In each panel Leupold's 972 reference strain (JB22) is indicated by a black triangle.
a, *Admixture* results. Each bar represents the proportion of SNPs assigned to each of the 2-5 populations, with the strain name below the bar. Geographic locations of the strains are shown in colored dots above the bar; Australia (yellow), Europe (green), Americas (red), Africa (pink), Asia (blue). **b**, Principal components plot colored by admixture clusters. Principal component coordinates as described for Fig. 1, using 752 unlinked SNPs. Strains (filled dots) are colored according to their *Admixture* cluster with K=5. As in Fig. 1, the 57 non-clonal strains are indicated with thick black borders. **c**, *fineSTRUCTURE* analysis of shared haplotypes. The heatmap depicts the proportion of the genome for which each strain in the columns shares most recent common ancestry with each other strain (i.e. relative to all other strains) in the rows, as inferred by *ChromoPainter* (note therefore values in each column add up to 1.0)[3]. Strains are colored along the axes by their geographic sampling location, as above. The row and column of Leupold's 972 reference strain (JB22) is indicated with grey shading. The tree at top shows the hierarchical merging of each strain based on genetic similarity, as inferred by *fineSTRUCTURE*[3,9]. This tree was inferred by first taking the sample configuration with the highest posterior probability among 100 posterior samples taken every 10,000 iterations from a Markov-Chain-Monte-Carlo (MCMC) run following 1 million "burn-in" iterations, next performing an additional 100,000 hill-climbing steps to find a solution with higher posterior probability, and then constructing the tree by the stepwise merging of clusters as described in Lawsen et al.[3,10]. Strains connected by a horizontal row at the bottom of the tree are inferred by *fineSTRUCTURE* to form a genetically homogeneous cluster. **d**, Majority consensus trees of the 57 non-clonal

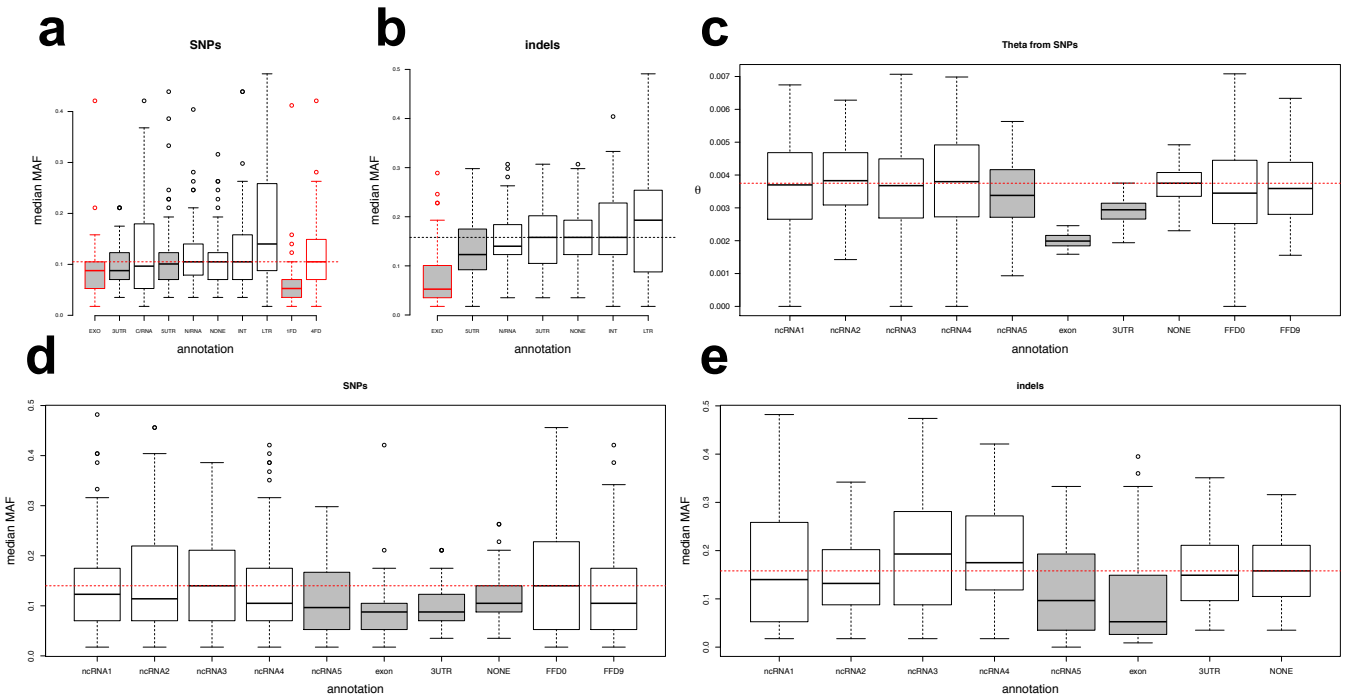
strains. A consensus tree generated from 100 trees, each estimated from a window of 1/100th of the genome. Branch values show the percentage of windows that support each clade, with strain names colored according to their geographic origin as for Fig. 1. The two trees have identical topology, branch lengths are adjusted to give a radial presentation in the left tree, and all branch lengths are equal in the right tree. The historic recombination of these strains is illustrated by the fact that all but one of the internal clades have less than 56% support. To generate this tree, we divided the genome into 100 non-overlapping windows, and produced alignments for all of the four-fold degenerate sites from each window (~10,000 sites each). We estimated the best tree for each window using the GTRAMMA model in RaXML[11,12], and calculated the consensus tree using the CONSENSE function from Phylip (http://evolution.genetics.washington.edu/phylip.html), using Majority rule (extended).



Supplementary Figure 3

The terminal 100 kb of all chromosomes contains excess diversity and unusual properties.

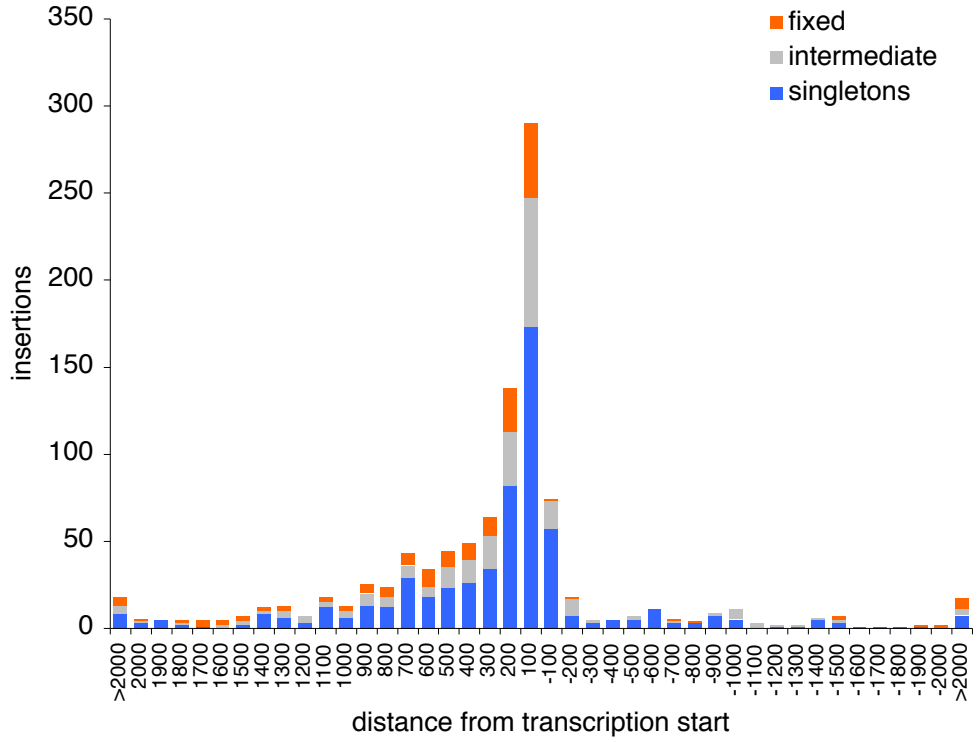
The columns of the 9 panels show the 3 chromosomes. The rows show the expression levels of protein-coding genes (top), the number of essential genes (middle), and the diversity (π; bottom). **Expression panels (top)** show the range of expression levels (in reads/kb/million reads, RPKM) for genes during exponential growth (log), stationary phase (stat) and meiotic differentiation (mei) (S.A., unpublished data). For each chromosome, we show the expression levels for the left 100 kb of the chromosome in red, the right 100 kb of the chromosome in blue, and all other genes in green. Box widths are proportional to the number of genes. We note that, in general, genes in chromosome ends are expressed at lower levels under all conditions tested. **Essential gene panels (middle)** show the number of essential genes/10kb window, with box fill colors as above. Essential genes are defined as those annotated with the Fission Yeast Phenotype Ontology id (FYPO:0000049 - inviable) in PomBase (www.pombase.org). **Diversity panels (bottom)** show the distribution of average pairwise similarity (π) for the 10 kb windows in left, middle and right regions of each chromosome. Chromosome ends have higher diversity, indicating less purifying selection. Not shown: the ends of chromosomes contain an excess of common LTR insertions (present in at least half of the non-redundant 57 strains, per 10 kb window of the genome). Windows within 100 kb of the chromosome ends had significantly more common insertions (ends mean = 0.74 transposons/window, internal regions mean = 0.15 transposons/window, Mann-Whitney test $P = 4.8 \times 10^{-11}$).



Supplementary Figure 4

Differences in diversity in various genome annotations.

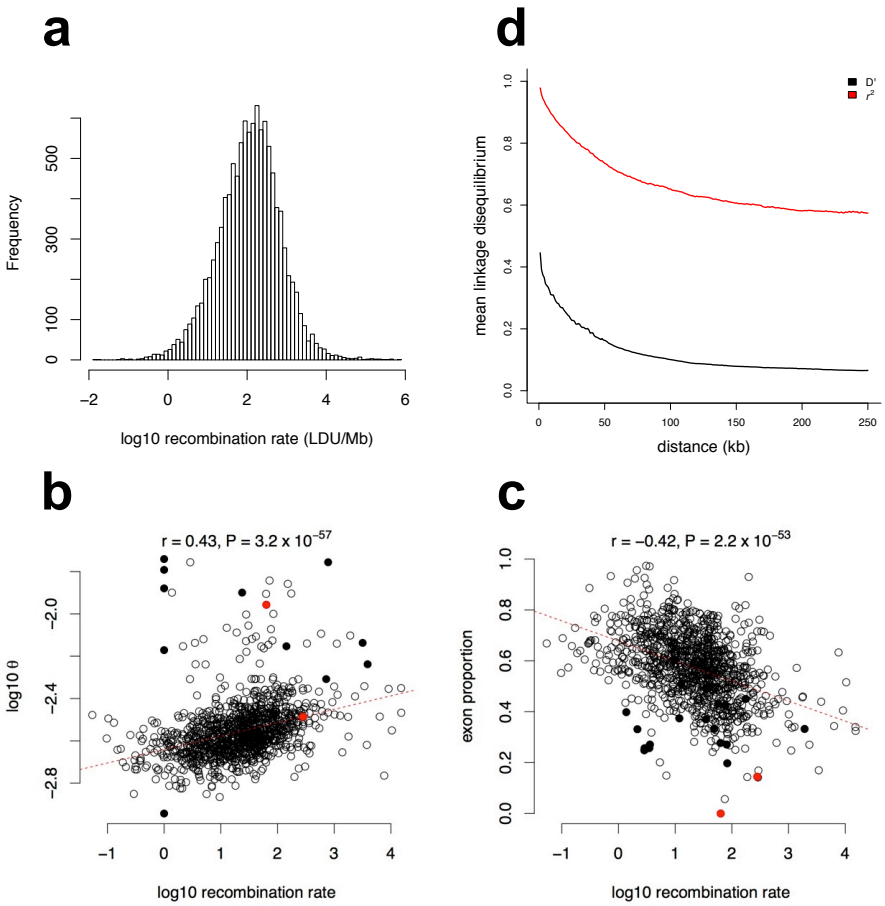
a, SNP median minor allele frequency. Median minor allele frequency calculated with SNPs from 100 windows of the genome, using sites specific to one annotation. Colors are as in Fig. 3b. C/RNA indicates canonical RNAs (rRNA, tRNA, snoRNA, snRNA). One-sided Mann-Whitney test P-values vs the FFD site neutral proxy were: $3.4x10^{-13}$ (exon), $4.4x10^{-3}$ (3' UTRs), 0.97 (canonical ncRNAs), 0.013 (5' UTRs), $P = 1$ (lncRNAs), 0.0078 (nun-annotated regions), 0.55 (introns), $3.7x10^{-3}$ (LTRs, which have higher median MAF), $7.3x10^{-16}$ (one-fold degenerate sites). This supports the conclusion from theta that exons, UTRs, but not lncRNAs have been subject to purifying selection. **b, Indel median minor allele frequency.** Median minor allele frequency calculated with indels from 100 windows of the genome, using sites specific to one annotation. Colors are as Fig. 3B One-sided Mann-Whitney test P values vs the neutral proxy of un-annotated sites were; $1.5x10^{-7}$ (exons), $2.8x10^{-3}$ (5' UTRs), 0.5 (lncRNAs), 0.077 (3' UTRs), 0.42 (introns), 0.66 (transposon LTRs). Here, exons and 5'UTRs show evidence for constraint, but 3' UTRs or lncRNAs do not. **c, Diversity (θ) in lncRNA expression fractions.** θ, calculated using SNPs, from left to right; 5 expression-fractions of non-canonical lncRNAs (ncRNA1 to lncRNA5, with lncRNA5 including the top 20% most highly expressed lncRNAs), exonic sites, 3'-UTRs, un-annotated regions, four-fold degenerate sites from low expression genes (FFD0, lowest 10%), four-fold degenerate sites from high expression genes (FFD9, highest 10%). In this analysis, we use un-annotated regions as a neutral proxy, and the red horizontal line shows the median value for these sites. Annotations that show significantly lower diversity than the neutral proxy are shaded grey, one-sided Mann-Whitney test P values are: $2.7x10^{-3}$ (ncRNA5), $6.9x10^{-29}$ (exons), $1.2x10^{-19}$ (3'-UTRs). **d, SNP median MAF in lncRNA expression fractions.** Median minor allele frequency of SNPs, with annotation classes as above. In this analysis, we use four-fold degenerate sites from low expression genes as a neutral proxy, and the red horizontal line shows the median value for these sites. Annotations that show significantly lower diversity than the neutral proxy are shaded grey, one-sided Mann-Whitney test P-values are: 0.012 (ncRNA5), $1.3x10^{-5}$ (exons), $2.3x10^{-5}$ (3'-UTRs), 0.026 (un-annotated regions). **e, Indel median MAF in lncRNA expression fractions.** Median minor allele frequency of indels, with annotation classes as above. In this analysis we use un-annotated regions as a neutral proxy, and the red horizontal line shows the median value for these sites. Annotations that have significantly lower diversity than un-annotated regions are shaded grey, one-sided Mann-Whitney test P-values are: $7.0x10^{-3}$ (ncRNA5), $8.7x10^{-11}$ (exons).



Supplementary Figure 5

A sharp peak of LTR insertions within 500 nt regions upstream of transcription start sites.

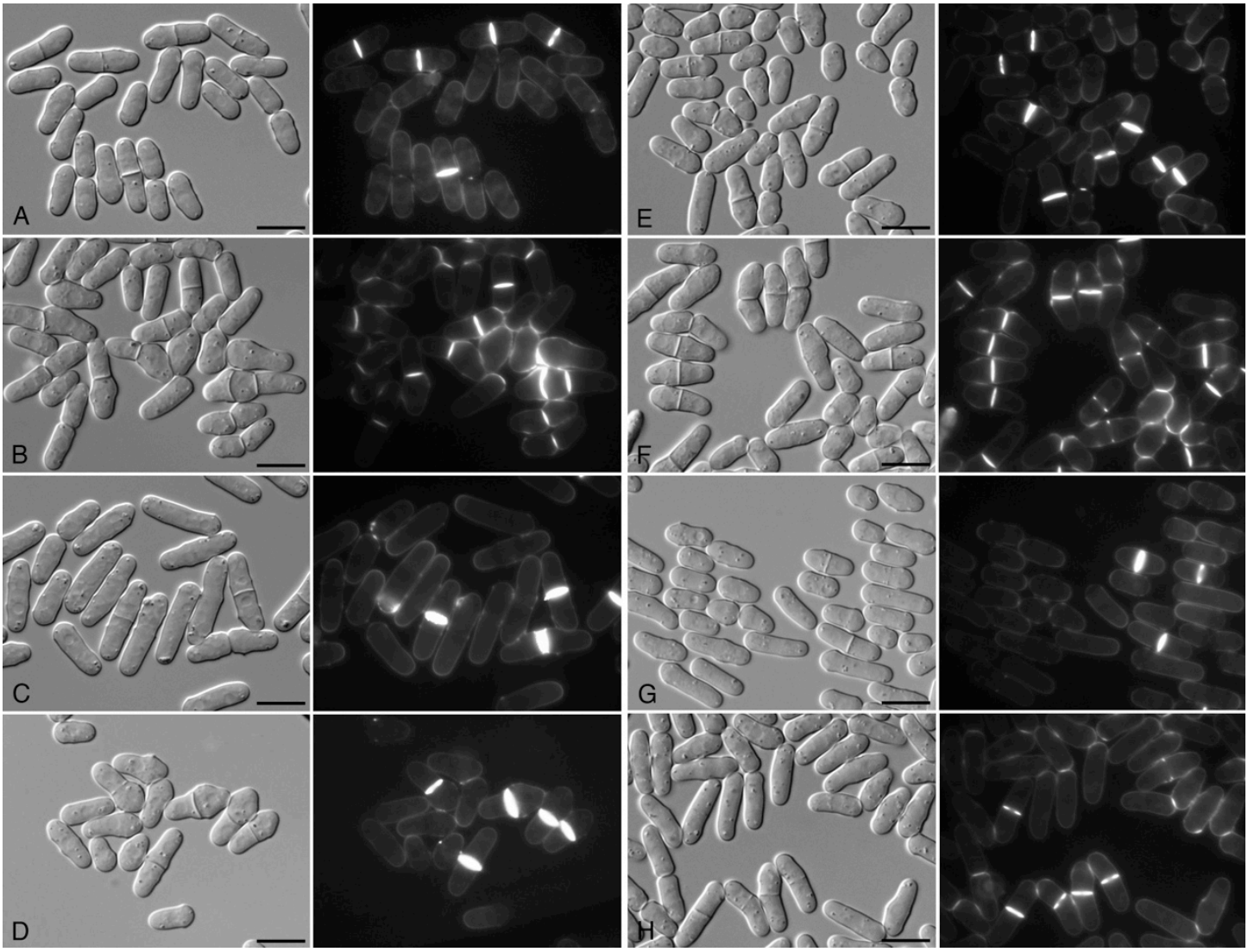
Histogram of LTR insertions in 100 bp bins around transcription start sites (TSS) of protein-coding genes. Positive and negative x-values denote regions up- and down-stream of TSS, respectively. The number of insertions is shown for 'fixed' insertions (present in all 57 strains), 'singletons' (present in single strain only), and 'intermediate' (all other insertions).



Supplementary Figure 6

Recombination rate and linkage decay.

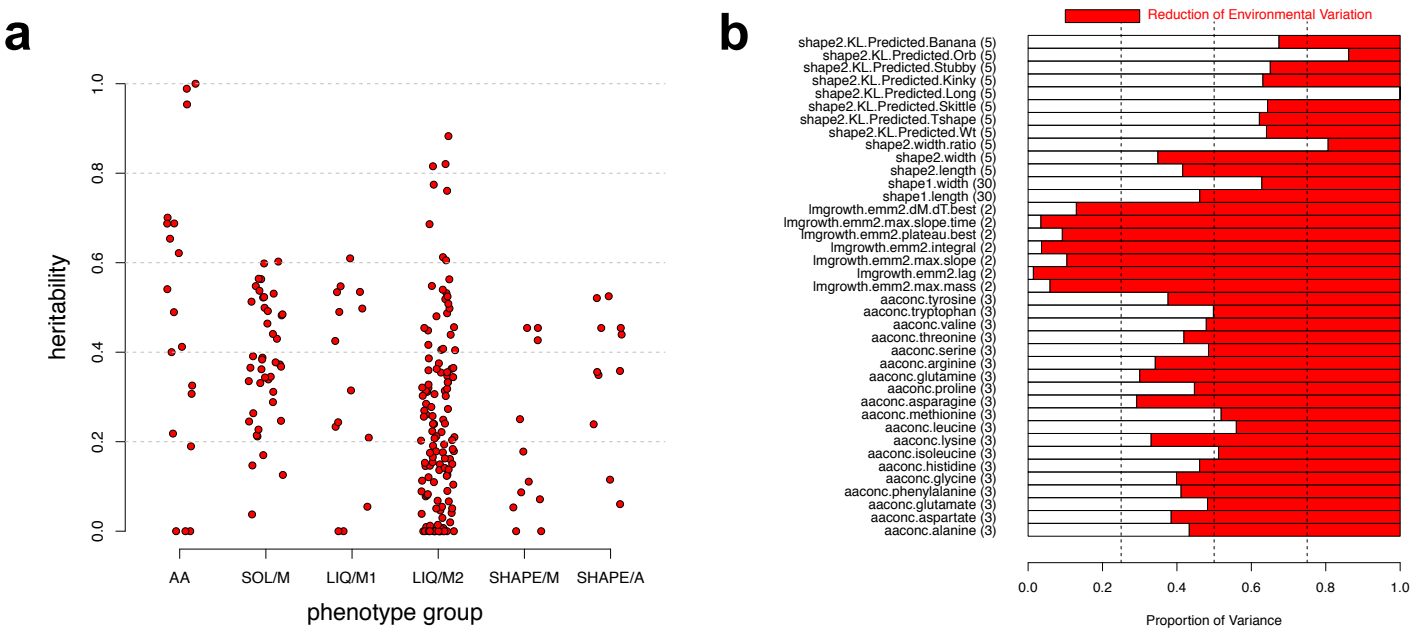
a, Recombination rate is log-normally distributed. For each SNP, we calculated the recombination rate in linkage disequilibrium units/Mb. The plot shows the distribution of non-zero rates in a $\log_{10}$ scale. **b**, Recombination rate is correlated with diversity. Filled red and black circles indicate centromeric and telomeric regions, respectively, as in Fig. 3C. Diversity (Watterson's θ), calculated as in Fig. 3C (in 10 kb genomic windows) is correlated with average recombination rate (LDU/Mb) (Spearman rank correlation $\rho = 0.43$, $P = 2.2 \times 10^{-57}$). **c**, Diversity is calculated as above. The recombination rate is negatively correlated with exon density (the proportion of each 10 kb window that is annotated as exon (Spearman $\rho = -0.42$, $P = 2.2 \times 10^{-53}$). **d**, Linkage disequilibrium (LD) declines to 50% of its value within 21 kb. Using SNPs with minor allele frequencies >0.05, we calculated the D' and r^2 measures of linkage disequilibrium for all pairs of SNPs up to 250 kb apart (see Methods). We show the mean D' and r^2 for all pairwise comparisons within each 1 kb window of distance.



Supplementary Figure 7

Microscopy images of selected strains.

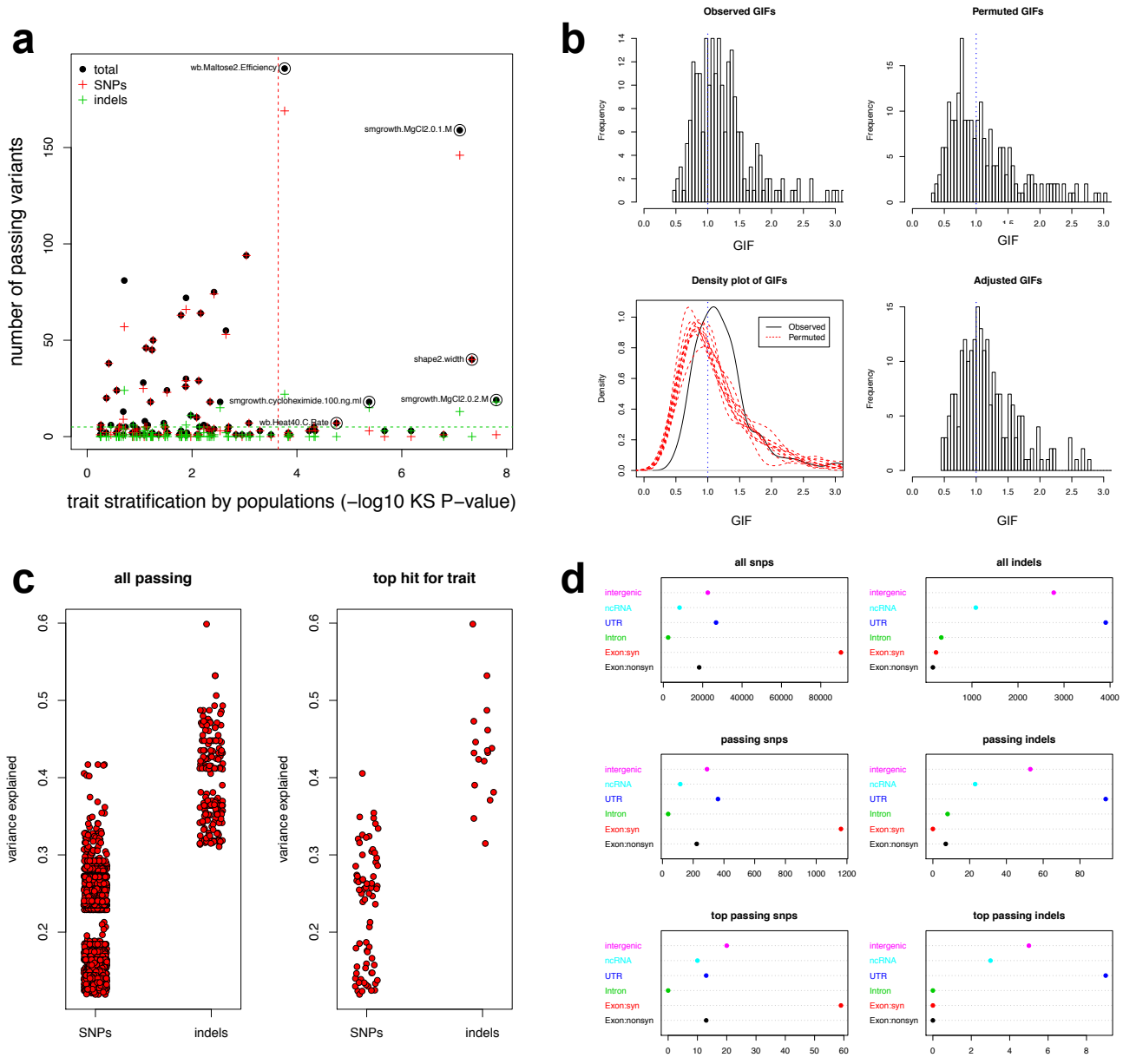
All strain descriptions (long, misshapen) are in comparison to Leupold's 972 reference strain. Left, DIC micrographs; right, calcofluor-stained cells, fluorescence microscopy (calcofluor stains the cell wall and division septum). Strains from top are: A) Leupold's 972 reference strain, B) JB762 which has branched, multi-septated and pear shaped cells, C) JB1207 which has long cells, D) JB1117 where cells are weakly misshapen/pear-shaped and slightly curved, E) JB939 which has misshapen cells, F) JB914 which is near-filamentous on solid media, bright calcofluor staining between cells shows that cells that have undergone cell division remain attached at the septum, G) JB930 which has short cells, H) JB1116 which contains "banana" shaped (curved) cells.



Supplementary Figure 8

Trait heritability and the value of repeat trait measurements.

a, Traits collected using all methods are heritable. Here we show heritability estimates according to the method of data collection. All methods are sufficiently accurate to detect some heritable traits. Data collection types from left are: AA) amino acid concentrations determined by mass spectrometry, SOL/M) colony size on various solid media, LIQ/M1) growth parameters in liquid YES rich media and EMM2 minimal media, LIQ/M2) growth parameters in various liquid media from Brown et al (2011), SHAPE/M) manually-defined shape parameters, SHAPE/A) automated definitions of shape parameters. **b,** Repeat measurements reduce non-genetic sources of variation (experimental noise/environmental variation). This plot shows the proportion of variation removed for each phenotype due to repeats, calculated as the adjusted r^2 from regressing the 179 individual phenotypic values on the factor clonal ID. For example, for the trait "Predicted Banana", for each clone, we recorded average phenotypic values across 5 samples, which removed approximately 30% of phenotypic variation. Repeated measurements for clones can substantially increase power to detect causal variants; for example, suppose we can remove 50% of variation through repeated measurements, then the proportion of variance explained by each variant effectively doubles (a variant which explains X% of total variation will explain 2X% of the variance which remains).

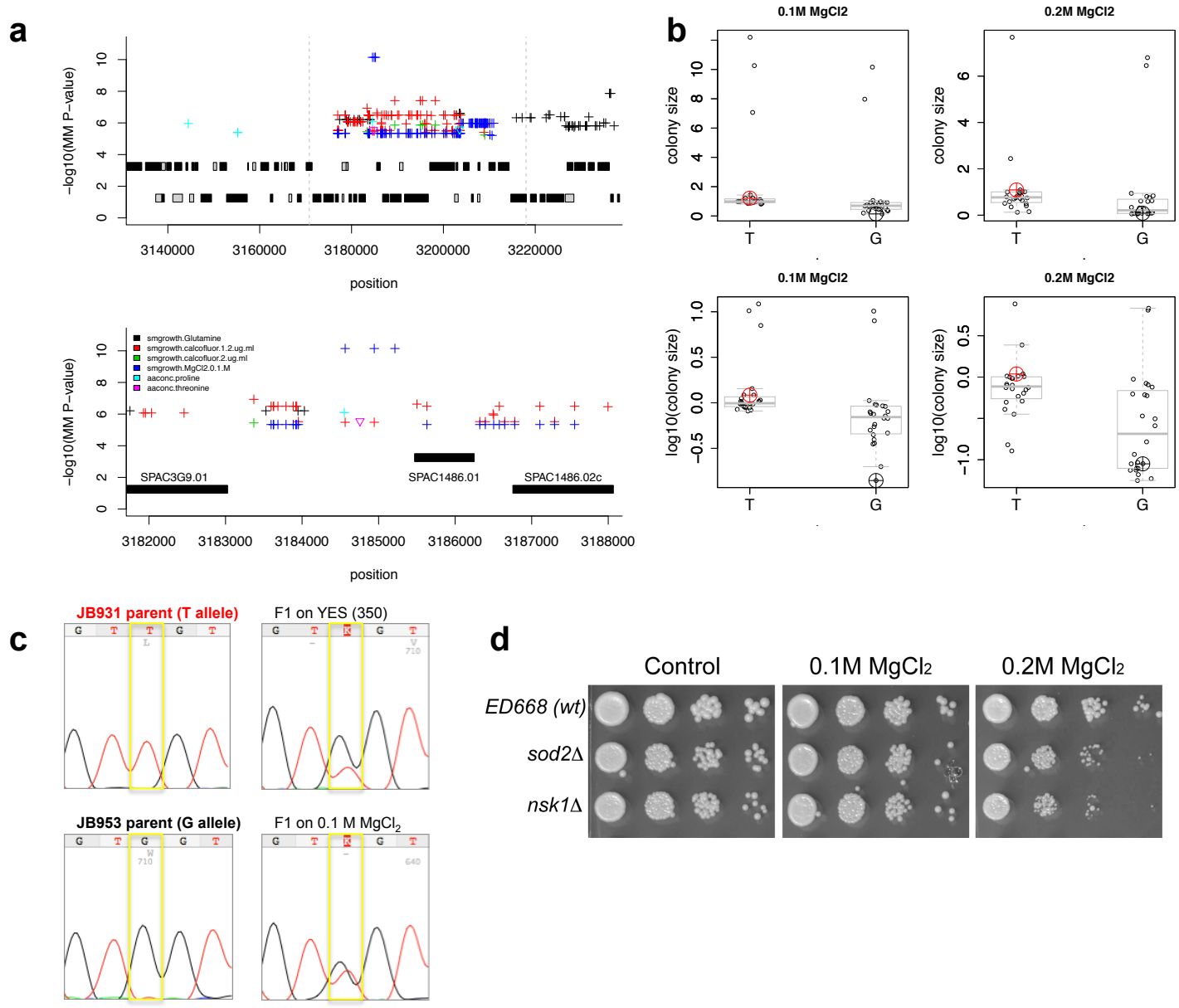


Supplementary Figure 9

Analysis of GWAS results.

a, To examine whether the mixed model GWAS controlled for population structure, we compared the degree of population stratification of each trait to the number of variants that passed the P-value threshold. To calculate the degree of population stratification, we divided the strains into 5 groups (defined by *Admixture*), and used a Kolmogorov-Smirnov test to determine whether the trait was significantly different between these 5 groups, using the log(P-value) as metric. This metric is not significantly correlated with the number of passing variants (Spearman rank correlation P >0.05). Circles show the number of all variants that are significant in the GWAS, red crosses indicate the number of passing SNPs and green crosses the number of passing indels. Traits that we might evaluate with caution because they are significantly population-stratified and have many passing variants are indicated with a black circle. The red vertical line shows the Bonferroni-corrected P-value threshold for the Kolmogorov-Smirnov tests, the green vertical line shows the median number of passing variants. **b,** Genomic inflation factors (GIFs). The GIF is the observed median P-value divided by the median expected P-value. Under a null model of no associations and unlinked variants, the expectation if for the GIF to be 1. We show the distribution of GIFs from the 223 traits (top left), GIFs from permuted data (top right), density plot of observed GIFs vs 10 sets of permutations (each one per trait) (bottom left), and the distribution of adjusted GIFs (observed median P-value/median P-value from

permuted data)(bottom right). Although the distribution of observed GIFs is slightly skewed to values larger than 1, adjusted GIFs (observed median/median from permuted data) is close to 1. **c,** Associated indels tend to explain a greater proportion of trait variance. For all variants associated with a trait (left) and for the most significant variant associated with the 89 traits (right) we show the estimated variance explained by the trait. **d**, Annotations of variants used for the GWAS analysis (top panel), all variants passing the P-value threshold (middle panel), and the most significant variant from each of the 89 traits (top hits) (lower panel). The annotations from top are intergenic regions (un-annotated as any other of the categories below), long non-coding RNAs (ncRNA), 5' and 3' untranslated regions (UTR), synonymous sites in exons (Exon:syn), and non-synonymous sites in exons (Exon:nonsyn). Indels that are multiples of 3 nucleotides are categorized as Exon:syn, all others are categorized as Exon:nonsyn. Chi-squared tests showed no significant difference between SNPs in any three groups, or indels in any three groups, including no bias towards non-synonymous SNPs.



Supplementary Figure 10

The GWAS hotspot on chromosome 1.

The 10kb region that contains the largest number of significant associations in the mixed model, and also the passing variant with the lowest P-value, is on chromosome 1 (Fig. 4b). Here we show: **a,** The passing variants in this 10kb window (top panel), with the window indicated by vertical grey lines, and the local neighborhood of three genes (lower panel). In both panels protein-coding genes are show below variants as black rectangles and ncRNAs as grey rectangles, with forward strand genes above reverse strand genes. The most significant variants are three SNPs between *nsk1* (SPAC3G9.01) and *sod2* (SPAC1486.01). These variants are in perfect LD, and are associated with growth in solid media with 0.1 M $MgCl_2$. *nsk1* is a reverse strand gene (transcribed from right to left), *sod2* is a forward strand gene, so these variants are in the promoter region of both genes. **b,** Shows the distribution of values for growth in solid media with 0.1 M $MgCl_2$ (left), categorizing strains by the genotype of one of these three variants (chromosome 1, position 3185213). The top right panel shows the trait values for the non-clonal 57 strains in 0.2 M $MgCl_2$. Because some strains are clear outliers, we show the trait on a log scale in the two lower plots. The box and whisker plots overlaid show the median and interquartile ranges of trait values. The red and black crosshairs show the trait values for the two parents used in the cross (panel c, below), JB931, which has the T allele (red) and JB953, which has the G allele (black). **c,** PCR and ABI capillary sequencing of the parents and F1 of a cross between two strains with the two genotypes at chromosome 1, position 3185213 (JB931 x JB953). The left panel shows the parents, and the right

panels show pools of F1 segregants grown on YES rich media without $MgCl_2$ (top) or in YES rich media with 0.1 M $MgCl_2$ (below). The segregating allele is indicated with a yellow box. The T allele is enriched relative to the G allele on $MgCl_2$, as expected from the trait values in part (b). The increase in signal from the favored allele is likely due to either increased colony size of segregants with the favored T allele (expected from the association) and/or increased survival of segregants with the favored T allele. Pools contained at least 35 colonies. **d,** Spot assays of serial ten-fold dilutions of *sod2* and *nsk1* deletion strains on control rich media (YES), and rich media with 0.1 $MgCl_2$ or 0.2 M $MgCl_2$. Both deletion strains show less dense growth on 0.2 M $MgCl_2$, consistent with these genes affecting sensitivity to this stress. Deletion strains are from the Bioneer Version 2.0 deletion collection, and ED668 is the corresponding wild-type strain (genotype *h+ ade6- M216 ura4-D18 leu1-32*).

Click inside this box and insert figure. For best results, use Insert menu to select a saved file; do not paste images. Source images must be in RBG color profile, at a resolution of 150—300 dpi. Optimize panel arrangement to a 2:3 height-to-width ratio; maximum online display is 600h x 900w pixels. Reduce empty space between panels and around image. Keep each image to a single page. Delete these instructions before inserting the image.

Supplementary Figure 11

Insert figure title here by deleting or overwriting this text; keep title to a single sentence.

Insert figure caption here by deleting or overwriting this text; captions may run to a second page if necessary.

Click inside this box and insert figure. For best results, use Insert menu to select a saved file; do not paste images. Source images must be in RBG color profile, at a resolution of 150—300 dpi. Optimize panel arrangement to a 2:3 height-to-width ratio; maximum online display is 600h x 900w pixels. Reduce empty space between panels and around image. Keep each image to a single page. Delete these instructions before inserting the image.

Supplementary Figure 12

Insert figure title here by deleting or overwriting this text; keep title to a single sentence.

Insert figure caption here by deleting or overwriting this text; captions may run to a second page if necessary.

Click inside this box and insert figure. For best results, use Insert menu to select a saved file; do not paste images. Source images must be in RBG color profile, at a resolution of 150—300 dpi. Optimize panel arrangement to a 2:3 height-to-width ratio; maximum online display is 600h x 900w pixels. Reduce empty space between panels and around image. Keep each image to a single page. Delete these instructions before inserting the image.

Supplementary Figure 13

Insert figure title here by deleting or overwriting this text; keep title to a single sentence.

Insert figure caption here by deleting or overwriting this text; captions may run to a second page if necessary.

Click inside this box and insert figure. For best results, use Insert menu to select a saved file; do not paste images. Source images must be in RBG color profile, at a resolution of 150—300 dpi. Optimize panel arrangement to a 2:3 height-to-width ratio; maximum online display is 600h x 900w pixels. Reduce empty space between panels and around image. Keep each image to a single page. Delete these instructions before inserting the image.

Supplementary Figure 14

Insert figure title here by deleting or overwriting this text; keep title to a single sentence.

Insert figure caption here by deleting or overwriting this text; captions may run to a second page if necessary.

Click inside this box and insert figure. For best results, use Insert menu to select a saved file; do not paste images. Source images must be in RBG color profile, at a resolution of 150—300 dpi. Optimize panel arrangement to a 2:3 height-to-width ratio; maximum online display is 600h x 900w pixels. Reduce empty space between panels and around image. Keep each image to a single page. Delete these instructions before inserting the image.

Supplementary Figure 15

Insert figure title here by deleting or overwriting this text; keep title to a single sentence.

Insert figure caption here by deleting or overwriting this text; captions may run to a second page if necessary.