Figures S1-S10

Available for download at http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.113.160291/-/DC1.

File S1

Supporting Methods

Supplementary Notes

Extended Methods

Genotyping

Sequencing reads from the ISA segregants, along with the parental strains SK1 and S96 (a haploid strain isogenic to S288c), were aligned to the S288c reference genome (build R63) using Novoalign (v2.07.06; http://www.novocraft.com/), allowing only unique alignments. Thereafter, GATK was used for realignment of the BAM files (Li et al. 2009), and subsequent SNP calling was performed using SAMtools (McKenna et al. 2010). The formula SAMtools applies for calling the genotype is modelled upon genotyping a population, and incorporates an allele frequency term. This is not applicable to our study, a cross between 2 parents, where the allele frequency at true SNP positions is 0.5. We thus used the genotype likelihood (PL stats generated by SAMtools) to infer the genotype. SNP positions, which correspond to a homozygous reference call in the S96 parent and a homozygous variant in the SK1 parent, were chosen first. From this set of SNPs, we selected SNPs where the calculated allele frequency was between 0.35 and 0.65 and the number of successfully genotyped segregants was more than 80%. This ensured that the genotypes segregated in a 1:1 manner as expected, and that duplicated or deleted regions were excluded. After generating the genotype matrix, using the R/qtl package, we further checked and removed switched alleles and markers not in linkage with their surrounding markers.

Bulk Segregant Analysis (BSA): Calculating allele frequencies

The allele frequency was calculated at each of the SNP positions used in ISA for all conditions, based on the ratio of base calls on different alleles from sequencing reads. The allele frequency was fitted using local polynomial regression assuming a binomial distribution. A bandwidth of 28kb validated with 5-fold cross-validation was used. Regions of interest were defined as intervals >30kb with fitted allele frequency >0.65 or <0.35. The peak position in the region of interest was defined according to the local maxima or minima in the region. Next, SNP positions were bootstrapped and the 95% confidence interval was determined for the peak position. To determine whether the allele frequency at the peak for a test condition was significant compared to the control (YPD 30°C, 100 generations), we first calculated the observed difference in allele frequency between the test condition and control. Then with each permutation, we randomly assigned reads in the region of interest to either test or control. The allele

frequency in both permutated datasets was fitted using local polynomial regression, and the difference in allele frequency calculated between test and control. This permutation was repeated 5000 times, and the p-value for the peak in the test condition was calculated as the probability of obtaining a value larger than the observed difference in the permutated dataset. After obtaining p-values for each peak in the test condition, they were corrected for multiple testing using the Benjamini-Hochberg method (Benjamini and Hochberg 1995).

Reciprocal Hemizygosity Scanning (RHS): Estimating allelic contributions

The fitness of each deletion strain was deduced from the signal intensity of the barcodes on the microarray. Each probe on the Genflex tag16k array (Affymetrix) is represented by five replicate features. Each tag was summarized by the $\log_2$-median intensity across all matching probes on the array. The $\log_2$ intensity distributions of the up and down tags (*i.e.*, the barcodes before and after the deletion cassette) on each microarray were shifted by a separate constant, so that all intensity distributions for growing strains had the same midpoint of the shortest interval containing half the data (a robust estimator of the mode of a distribution). Finally, the selection coefficient *s* (or relative growth rate of the strain in the pool) was estimated as the median across both up and down tags of the $\log_2$ fold change of normalized signal intensity between initial and final timepoints, divided by the pool generation number at the final timepoint. To control for pool construction effects, we focused on media-specific allelic effects, taking YPD as a control condition. For each gene in the genome, we modelled the selection coefficient $s_{i,j,k}$ of deleted allele *i* (0 for S96, 1 for SK1) in condition *j* (0 for YPD and 1 for the alternative condition) and pool *k* (0 for the first pool and 1 for the second pool) with the following linear model:

$$s_{i,j,k} = \beta_0 + \beta_a i + \beta_{a,c} ij + \beta_{0,k}(1-i)k + \beta_{1,k} ik + \varepsilon_{l,j,k.}$$

where β_0 is the intercept, β_a is the condition effect, β_c is the global allele effect, $\beta_{0,k}$ and $\beta_{1,k}$ are pool construction effects, and ε is a noise term. The terms of interest ($\beta_{a,c}$) were tested using a moderated t-test as implemented in the R limma package (Smyth et al. 2003). The moderated t-test robustly estimates the variance by following an empirical Bayes approach that effectively shrinks estimated sample variances towards a pooled estimate common to all strains. Obtaining robust estimates of the variance is important because of the small sample sizes. P-values were then corrected for multiple testing using Storey's false discovery rate approach (Storey and Tibshirani 2003).

Benjamini Y, Hochberg Y. 1995. Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. *J Roy Stat Soc B Met* **57**(1): 289-300.
Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**(16): 2078-2079.
McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M et al. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research* **20**(9): 1297-1303.
Smyth GK, Yang YH, Speed T. 2003. Statistical issues in cDNA microarray data analysis. *Methods in molecular biology* **224**: 111-136.
Storey JD, Tibshirani R. 2003. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America* **100**(16): 9440-9445.

File S2

RQTL data

Available for download at http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.113.160291/-/DC1.

Tables S1-S6

Available for download at http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.113.160291/-/DC1.

Table S1 Sequence of primers used for allele deletion or exchange, for *MUC1* repeat amplification, for the construction of the *ENA6* overexpression vector. The columns provide the name and the sequence of the primers, for following genes: *TAO3*, *AMN1*, *MUC1*, *SFL1*, *MKT1*, ENA6, *CRG1*
Meaning of abbreviations: del = ORF deletion, del_conf = deletion confirmation, swap = ORF amplification for allele exchange, swap_conf = allele exchange confirmation. If only del and conf primers are provided, conf primers also served as del_conf, swap, and swap_conf primers.

Table S2 Table of all the significant QTLs identified from ISA.
chr: chromosome of QTL
peak_pos: SNP position at which the peak is called
ci_start: start of 95% confidence interval for peak
ci_end: end of 95% confidence interval for peak
phenotype: the phenotype for which the QTL is called
LOD: LOD score of the peak as called from R/qtl

- Table of all the significant QTLs identified from BSA.
chr: chromosome of QTL
peak_pos: SNP position at which the peak is called
ci_start: start of 95% confidence interval for peak
ci_end: end of 95% confidence interval for peak
phenotype: the phenotype for which the QTL is called
peak_af: the allele frequency at peak_pos, 1 indicates 100% SK1 and 0 , 100% S288c padj: pvalue of the peak after adjustment with Benjamini-Hochberg

Table S3 Table with the estimated 95% confidence interval size for the QTL peaks of *CRG1* in cantharidin, *ENA* in NaCl and *MKT1* in YPE.

1. Samplesize: the number of segregants selected randomly each time, over 100 simulations
2. mean.ci_size: Mean size of the 95% confidence interval over 100 simulations
3. se.ci_size: standard error of the mean size of 95% confidence interval over 100 simulations
4. phe: Phenotype in which qtl is mapped.

Table S4 Table of heritability estimation
pheno : the phenotype
n.h.all : narrow sense heritability, all markers
b.h.all : broad sense heritability, all markers
n.h.qtl : narrow sense heritability, qtl markers (from Table S2 ISA)
b.h.qtl: broad sense heritability, qtl markers (from Table S2 ISA)

Table S5 List of markers used for testing interaction using Information Distance method. Distance values between markers.
For identity of markers in columns M1 and M2, refer to the list of markers
ID column is the Information Distance score
Left tail indicates the probability of observing in the permutated dataset, a score smaller than the Information Distance (ID) score
Right tail indicates the probability of observing in the permutated dataset, a score larger than the Information Distance (ID) score

Table S6 Summary from sequencing 50 RHS strains, including parental strains (raw RHS material), parental haploids for new RHS set, original haploid del strains (false positives), old RHS set (false positives), old RHS set (random strains), new RHS set (check if mutation is cured), newly reconstructed on hybrid. The columns indicate:
name = name of the strain incl. systematic ORF name
plate = internal plate specification
gene_del = name of deleted gene

S/K = strain background
wrong_gene_del = indicating if gene deletion is not correct
chr_aberration = kind of detected aneuploidy
comment = indicating possible origin of aneuploidy
point_mutation = number of possible point mutation high_confidence_point_mutation = number of mutation unique to this strain