Supplementary Material

Loss-of-function mutation survey revealed that genes with background-dependent fitness are rare and functionally related in yeast

Elodie Caudal¹, Anne Friedrich¹, Arthur Jallet¹, Marion Garin¹, Jing Hou¹,* and Joseph Schacherer¹,2,*

1. Université de Strasbourg, CNRS, GMGM UMR 7156, Strasbourg, France
2. Institut Universitaire de France (IUF)

* Corresponding authors
E-mail: jing.hou@unistra.fr (J.H.), schacherer@unistra.fr (J.S.)

This document includes:
Figure S1-6
Figure legends
Data File Descriptions
Figure S4

A. Respiration, oxidative phosphorylation, mitochondrial targeting
Transcription & chromatin organization
Nuclear-cytoplasmic transport

B. PMT1 GET2 PMT2
C. STI1 REC114 UBR2
D. SKT5
Figure S6

A. Genes Interaction degrees  

B. Co-annotation rate  

C. dN/dS (YN00)
**Supplemental figure legends**

**Figure S1** - (A) Number of reads (y-axis, log10 scale) vs. number of unique insertion sites (x-axis, log10 scale) across 107 diverse isolates. (B) Insertion preference comparison between the reference S288C and the other 106 selected isolates. Sequence motifs are on the x-axis and the percentage of reads with a given motif are presented as color coded bars. Error-bars correspond to the standard deviation across different isolates. (C) Insertion density comparison between essential and non-essential genes in S288C in the promoter region. Average insertion numbers in the -500 bp to +200 bp region relative to ATG are shown in 100 bp windows. (D) Insertion density comparison between essential and non-essential genes in S288C in the coding region (CDS). Average insertion numbers in the relative fractions of a given CDS are shown. (E) Insertion density comparison between essential and non-essential genes in S288C in the terminator region. Average insertion numbers in the -200 bp to +500 bp region relative to the stop-codon are shown in 100 bp windows.

**Figure S2** - (A) Predicted non-essential probabilities (y-axis) as a function of the number of insertions in the terminator region (300 bp after stop-codon). Non-essential genes are in blue and essential genes in red. (B) Predicted non-essential probabilities (y-axis) as a function of the number of insertions in a 10 kb region surrounding the CDS (5 kb before ATG and 5 kb after stop-codon). Non-essential genes are in blue and essential genes in red. (C) The number of strains retained as a function of cut-offs of the number of interpretable genes after removing low coverage regions (less than 50 insertions in the surrounding 10kb region and/or less than 3 insertions in the 300 bp terminator region). (D) Number of genes retained after imputation as a function of cut-offs of the number of interpretable genes after removing low coverage regions.

**Figure S3** - (A) Average non-essential probability or predicted fitness for every 10 successive essential genes along all 16 chromosomes for 52 strains that passed the coverage cut-offs. Strain-side clustering was based on predicted fitness for all genes. (B) Insertion profiles for gene related to galactose metabolism that are annotated as non-essential in S288C but detected as essential/sick in all or a fraction of the 39 strains in the final dataset. Chromosomal positions and gene orientations are schematically presented on the x-axis and insertion profiles for each strain are presented as black vertical bars. (C) Insertion profiles for essential genes predicted as non-essential in S288C. Shaded areas correspond to potential essential protein domains. (D) Insertion profiles for genes previously shown background-specific essentiality between S288C and Σ1278b. Domain-specific essentiality regions are indicated.

**Figure S4** - (A) Annotated network based on profile similarity as shown in Figure 4B. (B-D) Subnetworks with significant correlations independent from the large subnetwork involving respiration-related hits.

**Figure S5** - (A) Global correlation of predicted vs. observed fitness across 6 genes and 17 strain backgrounds. A total of 88 strain/gene combinations are obtained. Predicted fitness (y-axis) corresponds to the logistic probability based on insertion profiles and observed fitness (x-axis) corresponds to the colony size ratio between deletion mutant and wild type on YPD after tetrad
dissection. (B) Predicted fitness for genes annotated in respiration/mitochondrial targeting, transcription & chromatin organization and nuclear-cytoplasmic transport in the reference S288C with gene name annotations. Related to Figure 5B. (C) Rewiring diagrams for all 38 isolates relative to the reference S288C. Related to Figure 5C.

**Figure S6- Evolutionary features associated with genes with predicted background-dependent fitness.** (A) Genetic interaction degrees derived from the yeast global genetic interaction network\textsuperscript{37} for non-essential, class I background-dependent, class II background-dependent and essential gene categories. The number of genes annotated in each category are indicated. (B) Functional co-annotation rates\textsuperscript{37} for different gene categories. The co-annotation rate corresponds to the fraction of interaction partners that are annotated in the same biological process as the primary gene\textsuperscript{37}. (C) Mean non-synonymous vs. synonymous substitution rates (dN/dS) across 1,011 natural yeast isolates using the YN00 method\textsuperscript{26}. Comparisons between categories were performed using T-test, and significance levels are as indicated, with ns: P-value > 0.05, *: P-value < 0.05, **: P-value < 0.01, ***: P-value < 0.001 and ****: P-value << 0.0001.

**Supplemental datafiles**

**Dataset S1- Description of isolates used in this study.** This table includes the following columns:

- **Standardized name**
  Three letter code corresponding to the original isolates in Peter et al. 2018.
- **Isolation:**
  The ecological and/or geographical sources from which the strain was isolated as shown in Peter et al. 2018.
- **Eco_Origin**
  Idem. See Peter et al. 2018.
- **Geo_Origin**
  Idem. See Peter et al. 2018.
- **Continent**
  Idem. See Peter et al. 2018.
- **CLADES:**
  Idem. See Peter et al. 2018.
- **Strain_name**
  Common name for the isolates used
- **InsertionSites**
  The number of unique insertion sites in the isolate mutant pool.
- **Clean_reads**
  The total number of reads mapped to the isolate from the sequenced mutant pool.
- **Validation**
  Isolates used for functional validation by gene deletion.

**Dataset S2- Model construction and evaluations.** This table contains 4 tabs:
GenesInModel: 4600 ORFs and their essentiality annotations used to construct the logistic model. Insertion numbers and densities within coding sequence and surrounding regions are included. Insertion numbers calculated from S288C insertion profile. This tab includes the following columns:

**ORF**
Systematic name for the ORF studied.

**CHR**
Chromosome correspondence for the ORF.

**START**
Chromosomal location of start of the ORF, regardless of the direction.

**END**
Chromosomal location of end of the ORF, regardless of the direction.

**ORI**
The direction of the ORF, 1 corresponds to sense (5' to 3'), -1 corresponds to anti-sense (3' to 5')

**SIZE**
The size of the ORF in bp.

**N_zone10k**
Number of insertions 5 kb before ATG and 5 kb after stop codon.

**N_prom100**
Number of insertions in the -100 bp region before ATG.

**N_prom300**
Number of insertions in the -300 bp region before ATG.

**N_atg200**
Number of insertions in the -100 bp region before ATG and +100 bp after ATG.

**N_term100**
Number of insertions in the +100 bp region after stop codon.

**N_term300**
Number of insertions in the +300 bp region after stop codon.

**N_stop200**
Number of insertions in the -100 bp region before stop codon and +100 bp region after stop codon.

**N_cds**
Number of insertions in the CDS

**Strain**
Strain name

**Tot_insertion**
Total insertion sites for the strain

**Ess**
Annotated gene essentiality in the reference S288C

**N_zone10k_norm**
Density of insertions 5 kb before ATG and 5 kb after stop codon.

**N_prom100_norm**

Density of insertions in the -100 bp region before ATG.

\[ N_{\text{prom}300\_norm} \]
Density of insertions in the -300 bp region before ATG.

\[ N_{\text{atg}200\_norm} \]
Density of insertions in the -100 bp region before ATG and +100 bp after ATG.

\[ N_{\text{term}100\_norm} \]
Density of insertions in the +100 bp region after stop codon.

\[ N_{\text{term}300\_norm} \]
Density of insertions in the +300 bp region after stop codon.

\[ N_{\text{stop}200\_norm} \]
Density of insertions in the -100 bp region before stop codon and +100 bp region after stop.

\[ N_{\text{cds}\_norm} \]
Density of insertions in the CDS

ModelSummary: Features included in the logistic model and their coefficient.

CrossValidation: Summary of the cross-validation results.

CMStat: Confusion matrix, prediction accuracy and precision for essential/non-essential labels.

**Dataset S3- Raw and final dataset with predicted fitness.** This table contains 3 tabs:

Raw_data_pred: All raw predicted fitness based on the logistic model for 107 isolates. This tab includes all the columns described in TableS2, tab “GenesinModel”, plus the following columns:

- logOdds
  The log of odds for the predicted probability according to the logistic model.
- Prob
  The predicted probability of being non-essential according to the logistic model.
- window_10k
  The range of insertion numbers found in the 5 kb prior to the ORF and 5 kb after the ORF.

Pred_final_39: Predicted fitness for 39 isolates included in the final dataset. Raw, imputed and quantile normalized predictions are shown. This tab includes all the columns in the previous tab, plus the following two columns:

- Prob_imputed
  Imputed probability value using K-nearest neighbours (KNN).
- Prob_qt
  Quantile normalized probability value after KNN imputation.
Score_final_39: Differential fitness score by comparing the predicted fitness in a given isolate to S288C. This tab includes columns in the previous tabs, plus the following columns:

**FY**
Quantile normalized probability value after KNN imputation for the corresponding ORF in S288C.

**Score**
Quantile normalized probability value after KNN imputation for and given ORF in a given strain subtracted by the corresponding probability in S288C.

**Type**
Classification of ORF/Strain according to the score.

**Dataset S4- Background-dependent fitness variation genes identified in this study.** This table contains 5 tabs:

**Z-statistics:** Z-statistics for each of the 632 hits, including the number of genetic backgrounds impacted for each hit.

**Hits_SAFE_annotation:** Annotations for each hit into the 16 functional neighbourhoods according to SAFE \(^{37}\). This tab includes columns described in TableS3, plus the following columns:

- **GO**
  GO term associated with the gene.
- **external_gene_name**
  Common name for the ORF.
- **Network.Region.Name**
  Functional neighbourhoods according to SAFE annotations.
- **Enriched.GO.names**
  GO terms associated with the SAFE functional neighbourhood.

**Validation:** Phenotyping results for functional validation using gene deletion. This tab includes the following columns:

- **Gene**
  Gene name.
- **Strain**
  Strain name.
- **Group**
  Enriched biological function.
- **Condition**
  Growth condition for deletion and wild type strains.
- **Ratio_cond**
  Fitness based on the colony size ratio between mutant and wild type for a given condition.
- **Ratio_cond_sd**
Standard deviation based on 4 replicates.

*Ratio_dissection*
Fitness based on colony size ratio between mutant and wild type after tetrad dissection on YPD.

*Prob_qt*
Quantile normalized probability value after KNN imputation.

*Type*
Classification of ORF/Strain according to the score.

Enrichment_global: Enrichment for all hits across 16 functional neighbourhoods. This tab includes the following columns:

*SAFE*
Functional neighbourhood names

*OR*
Odds ratio of the one-sided Fisher’s exact test

*pval*
P-value of the one-sided Fisher’s exact test

*n_in_hits*
Number of genes with background-dependent fitness that belongs to the SAFE neighbourhood across all strains.

*n_hits*
Total number of genes with background-dependent fitness.

*n_go*
Number of genes belongs to the SAFE neighbourhood in the background.

*n_bk*
Number of genes in the background.

*ID*
Type of the hits analyzed.

Enrichment_Strain: Enrichment for hits in a given genetic background across 16 functional neighbourhoods. This tab includes the following columns in addition to ones described in the previous tab:

*n_common*
Number of genes with background-dependent fitness that belongs to the SAFE neighbourhood in a given strains.

*n*
Total number of genes with background-dependent fitness in a given strains.

*Strain*
Name of the strain.

**Dataset S5- Genetic interaction degree and dN/dS values and gene classifications.** This table contains the following columns:

*ORF*
Gene systematic name.
| **Gene** | Gene common name. |
|----------|--------------------|
| **Negative** | Number of negative genetic interactions, obtained from Costanzo et al. 2016. |
| **Positive** | Number of positive genetic interactions, obtained from Costanzo et al. 2016. |
| **Combined** | Total number of genetic interactions, obtained from Costanzo et al. 2016. |
| **Co-annotation.rate** | Percentage of interactions that belongs to the same SAFE functional neighbourhood, obtained from Costanzo et al. 2016. |
| **YN00_mean** | Mean dN/dS value calculated from the 1,011 yeast genomes. Data from Peter et al. 2018. |
| **ID** | Gene categories. |
| **type** | Type of the hits if the category corresponds to genes with background-dependent fitness phenotypes. |