Supplementary Material

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

Elodie Caudal[1], Anne Friedrich[1], Arthur Jallet[1], Marion Garin[1], Jing Hou[1,*] and Joseph Schacherer[1,2,*]

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 S1

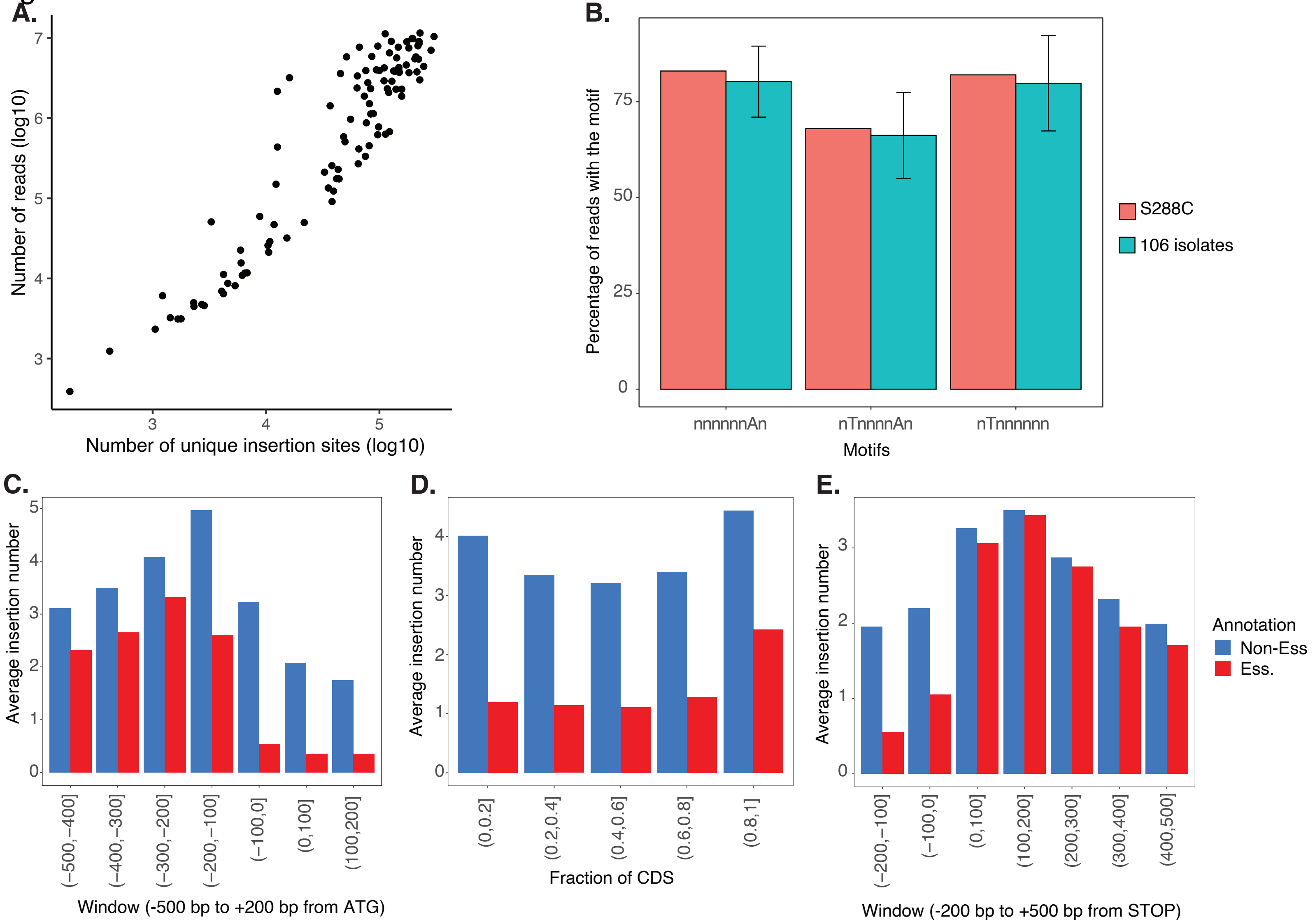

A.
Number of reads (log10)
7
6
5
4
3
3
4
5
Number of unique insertion sites (log10)
B.
Percentage of reads with the motif
75
50
25
0
nnnnnnAn
nTnnnnAn
nTnnnnnn
Motifs
S288C
106 isolates
C.
Average insertion number
5
4
3
2
1
0
(−500,−400]
(−400,−300]
(−300,−200]
(−200,−100]
(−100,0]
(0,100]
(100,200]
Window (-500 bp to +200 bp from ATG)
D.
Average insertion number
4
3
2
1
0
(0,0.2]
(0.2,0.4]
(0.4,0.6]
(0.6,0.8]
(0.8,1]
Fraction of CDS
E.
Average insertion number
3
2
1
0
(−200,−100]
(−100,0]
(0,100]
(100,200]
(200,300]
(300,400]
(400,500]
Window (-200 bp to +500 bp from STOP)
Annotation
Non-Ess
Ess.

Figure S2

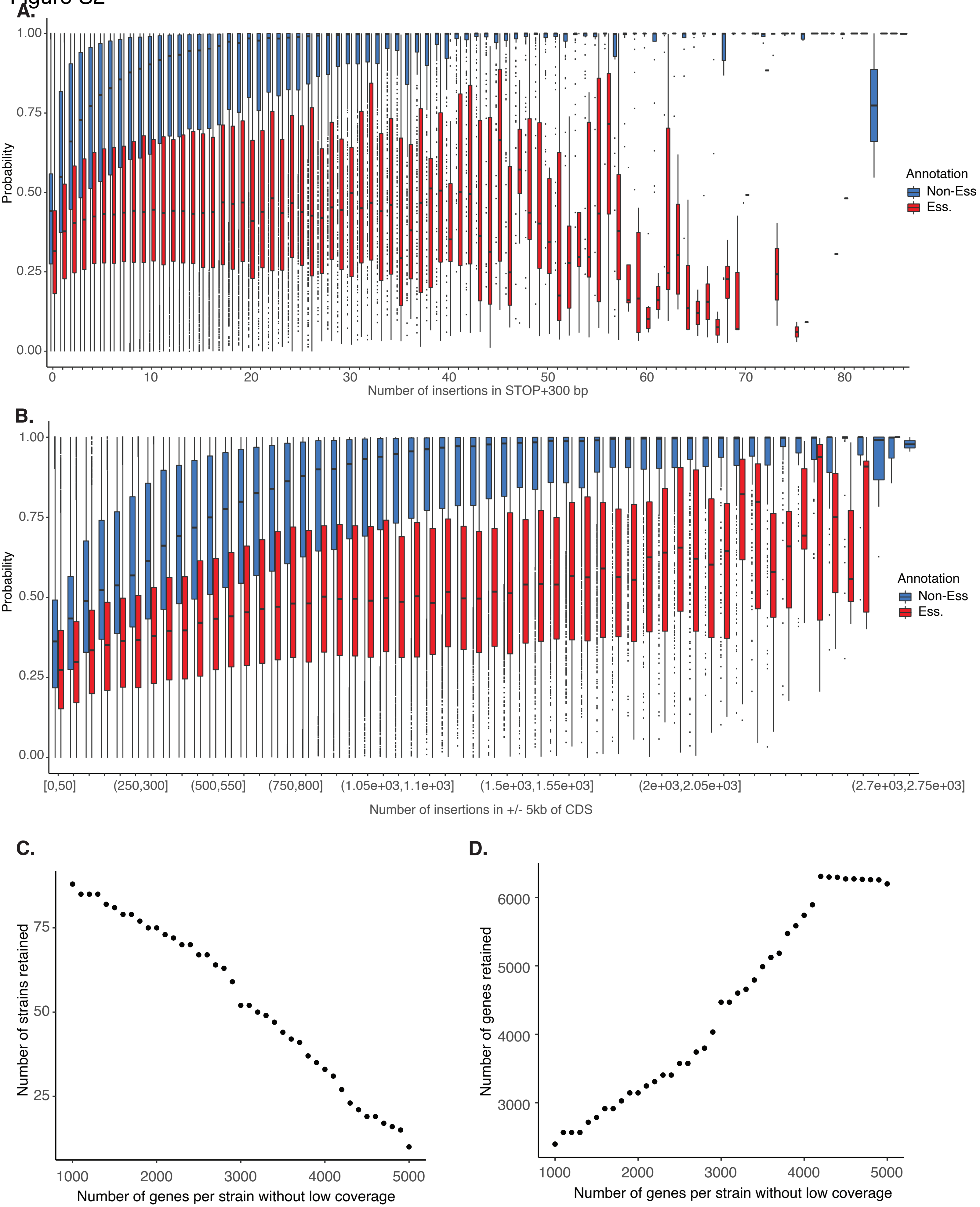

A.
Probability
1.00
0.75
0.50
0.25
0.00
0
10
20
30
40
50
60
70
80
Number of insertions in STOP+300 bp
Annotation
Non-Ess
Ess.
B.
Probability
1.00
0.75
0.50
0.25
0.00
[0,50]
(250,300]
(500,550]
(750,800]
(1.05e+03,1.1e+03]
(1.5e+03,1.55e+03]
(2e+03,2.05e+03]
(2.7e+03,2.75e+03]
Number of insertions in +/- 5kb of CDS
Annotation
Non-Ess
Ess.
C.
Number of strains retained
75
50
25
1000
2000
3000
4000
5000
Number of genes per strain without low coverage
D.
Number of genes retained
6000
5000
4000
3000
1000
2000
3000
4000
5000
Number of genes per strain without low coverage

Figure S3

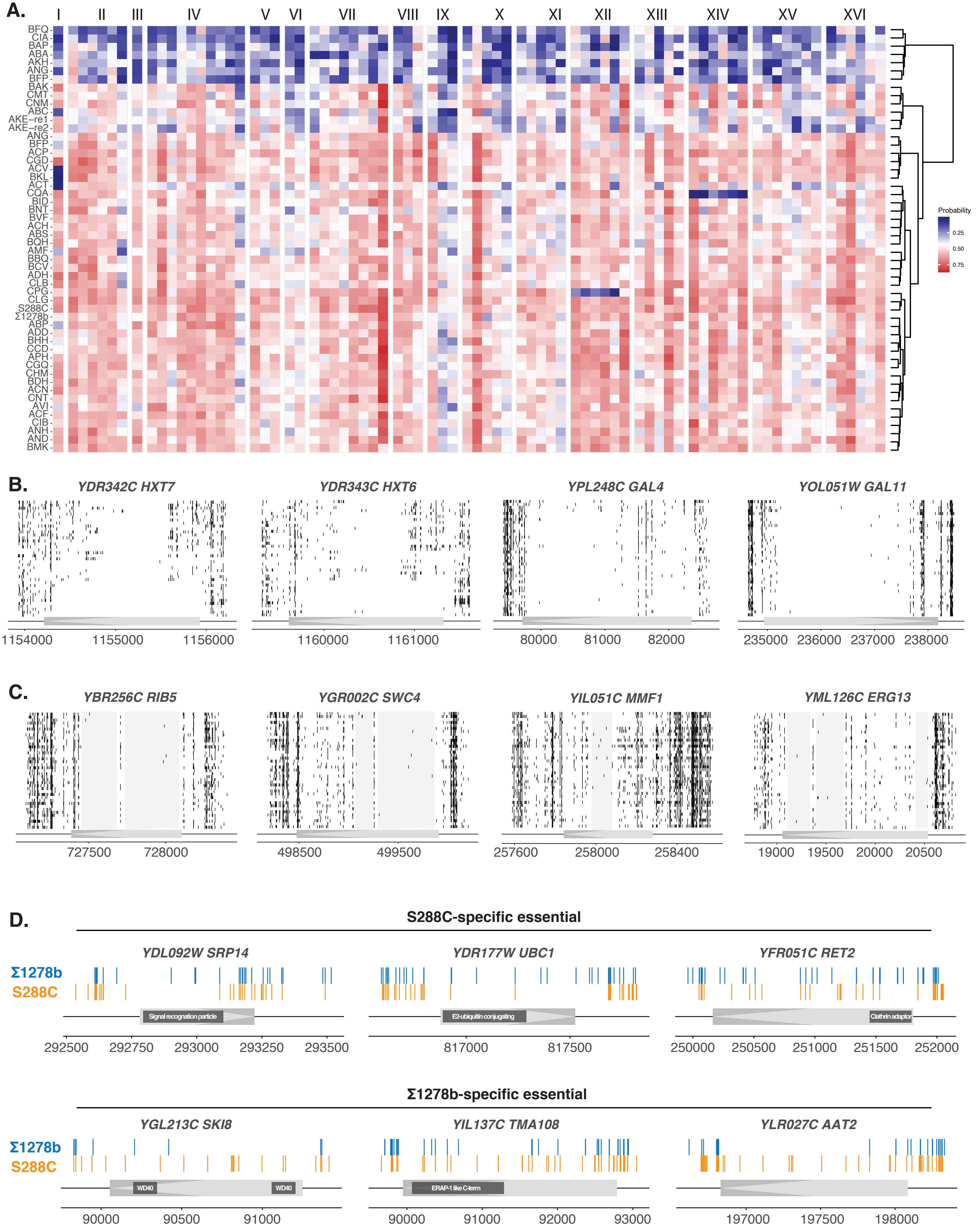

A.
I
II
III
IV
V
VI
VII
VIII
IX
X
XI
XII
XIII
XIV
XV
XVI
BFQ
CIA
BAP
ABA
AKH
ANG
BFP
BAK
CMT
CNM
ABC
AKE−re1
AKE−re2
ANG
BFP
ACP
CGD
ACV
BKL
ACT
CQA
BID
BNT
BVF
ACH
ABS
BQH
AMF
BBQ
BCV
ADH
CLB
CPG
CLG
S288C
Σ1278b
ABP
ADD
BHH
CCD
APH
CGQ
CHM
BDH
ACN
CNT
AVI
ACF
CIB
ANH
AND
BMK
Probability
0.25
0.50
0.75
B.
YDR342C HXT7
1154000
1155000
1156000
YDR343C HXT6
1160000
1161000
YPL248C GAL4
80000
81000
82000
YOL051W GAL11
235000
236000
237000
238000
C.
YBR256C RIB5
727500
728000
YGR002C SWC4
498500
499500
YIL051C MMF1
257600
258000
258400
YML126C ERG13
19000
19500
20000
20500
D.
S288C-specific essential
YDL092W SRP14
Σ1278b
S288C
Signal recognation particle
292500
292750
293000
293250
293500
YDR177W UBC1
E2-ubiquitin conjugating
817000
817500
YFR051C RET2
Clathrin adaptor
250000
250500
251000
251500
252000
Σ1278b-specific essential
YGL213C SKI8
Σ1278b
S288C
WD40
WD40
90000
90500
91000
YIL137C TMA108
ERAP-1 like C-term
90000
91000
92000
93000
YLR027C AAT2
197000
197500
198000

Figure S4

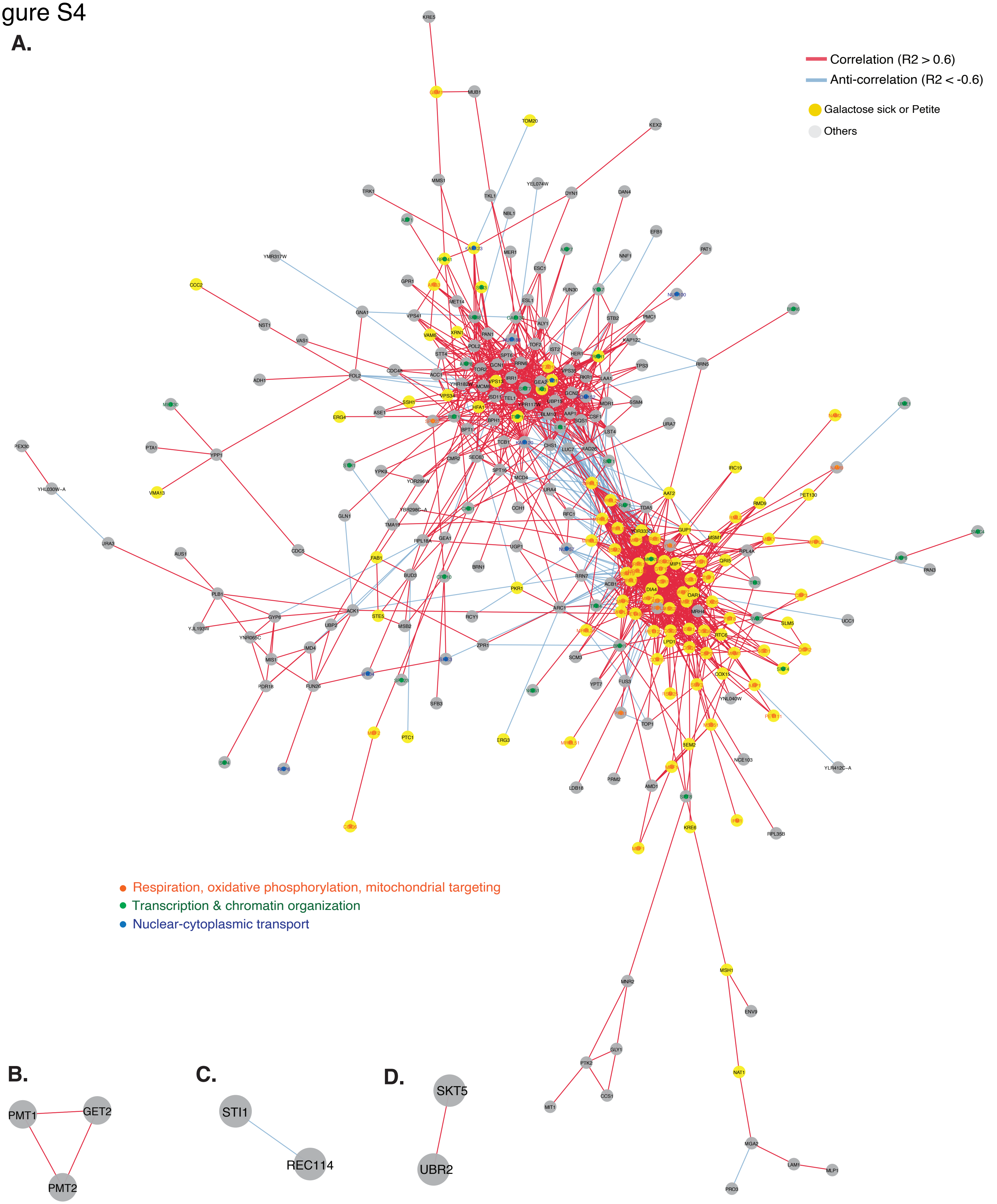

A.
Correlation (R2 > 0.6)
Anti-correlation (R2 < -0.6)
Galactose sick or Petite
Others
● Respiration, oxidative phosphorylation, mitochondrial targeting
● Transcription & chromatin organization
● Nuclear-cytoplasmic transport
KRE5
MUB1
TOM20
KEX2
MMS1
TRK1
TKL1
YEL074W
DYN1
DAN4
NBL1
EFB1
PAT1
YMR317W
MER1
NNF1
ESC1
CCC2
GPR1
FUN30
NUP100
MET14
ESL1
GNA1
VPS41
STB2
PMC1
NST1
VAM6
XRN1
PAN1
ALY1
VAS1
KAP122
TOF2
POL2
IST2
STT4
HER1
CDC48
GCN1
RRN6
TPS3
ADH1
FOL2
ACC1
VPS35
RKR1
LAA1
RRN5
VPS13
IRR1
GEA2
YHR182W
MCM6
VPS34
SSH1
GCN2
ISD11
TEL1
UBP15
MDR1
SSM4
YPR117W
ASE1
ERG4
HFA1
AAP1
BLM10
CSF1
BPH1
SQS1
URA7
NAM2
UME1
PEX30
PTA1
YPP1
LST4
BPT1
TCB1
CHS1
LUC7
RAD28
CMR2
SEC63
YPK9
IRC19
SPT16
YOR296W
MCD4
VMA13
YHL030W-A
URA4
AAT2
RMD9
PET130
YBR298C-A
CCH1
TDA5
RFC1
GLN1
TMA19
YOR333C
GUP1
MSM1
URA3
AUS1
CDC5
RPL18A
GEA1
UGP1
RPL4A
BRN1
MIP1
QRI5
PAN3
FAB1
BUD3
RRN7
PKR1
ACB1
DIA4
OAR1
PLB1
ACK1
ARC1
MRH4
UCC1
YJL193W
GYP6
STE5
RCY1
UBP2
MSB2
SLM5
YNR065C
LPD1
RTC6
IMD4
ZPR1
MIS1
SCM3
COX15
PDR1B
FUN26
YPT7
FUS3
YNL040W
SFB3
TOP1
PTC1
ERG3
BEM2
NCE103
YLR412C-A
LDB18
PRM2
AMD1
KRE6
RPL35B
MSH1
MNR2
ENV9
GLY1
PTK2
NAT1
CCS1
MIT1
MGA2
LAM1
MLP1
PRO3
B.
PMT1
GET2
PMT2
C.
STI1
REC114
D.
SKT5
UBR2

Figure S5

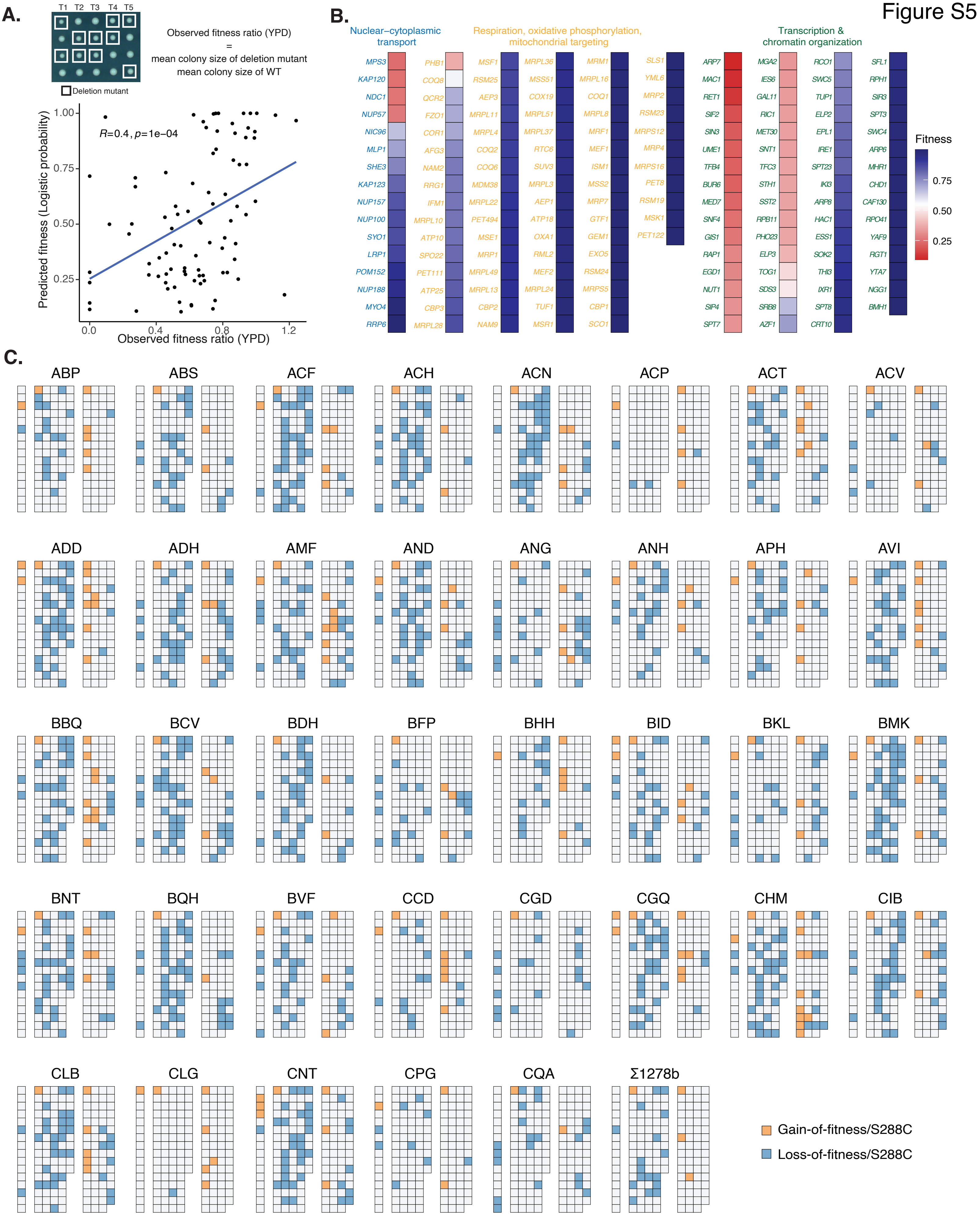
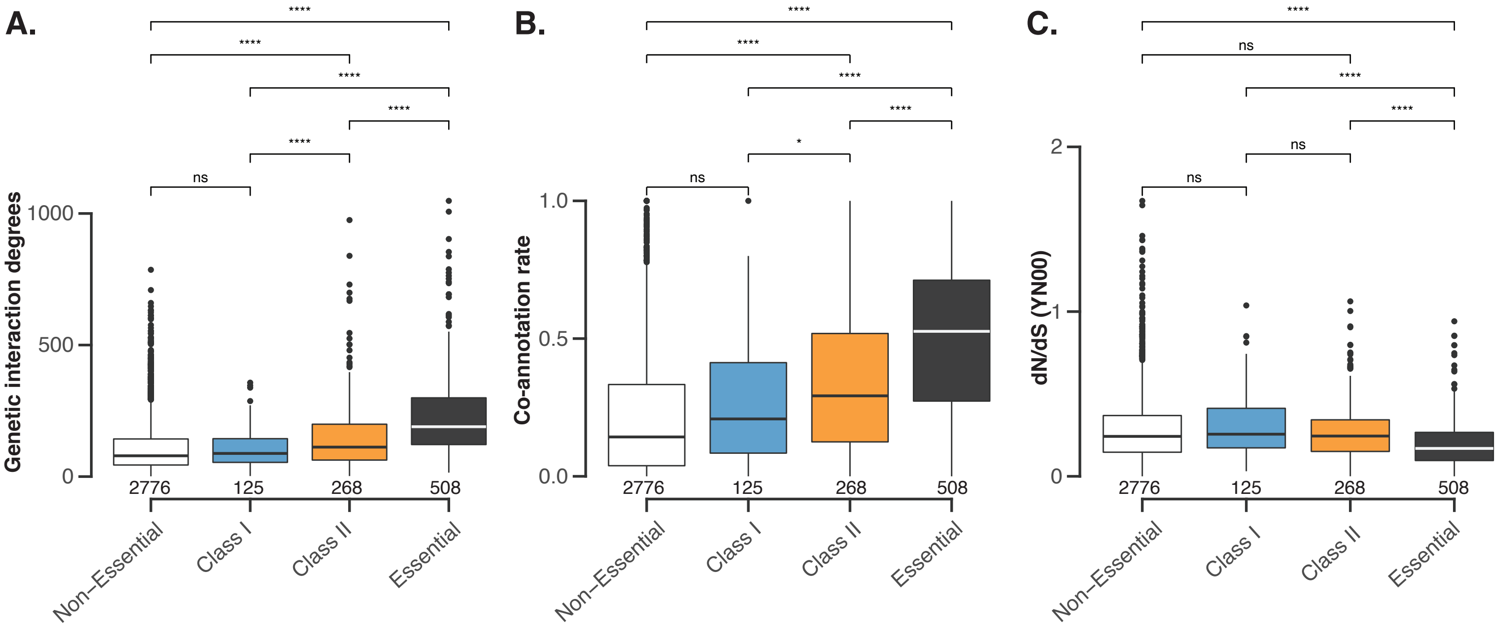

A.
T1 T2 T3 T4 T5
Deletion mutant
Observed fitness ratio (YPD)
=
mean colony size of deletion mutant
mean colony size of WT
R=0.4, p=1e−04
Predicted fitness (Logistic probability)
1.00
0.75
0.50
0.25
0.0
0.4
0.8
1.2
Observed fitness ratio (YPD)
B.
Nuclear−cytoplasmic transport
Respiration, oxidative phosphorylation, mitochondrial targeting
Transcription & chromatin organization
MPS3
KAP120
NDC1
NUP57
NIC96
MLP1
SHE3
KAP123
NUP157
NUP100
SYO1
LRP1
POM152
NUP188
MYO4
RRP6
PHB1
COQ8
QCR2
FZO1
COR1
AFG3
NAM2
RRG1
IFM1
MRPL10
ATP10
SPO22
PET111
ATP25
CBP3
MRPL28
MSF1
RSM25
AEP3
MRPL11
MRPL4
COQ2
COQ6
MDM38
MRPL22
PET494
MSE1
MRP1
MRPL49
MRPL13
CBP2
NAM9
MRPL36
MSS51
COX19
MRPL51
MRPL37
RTC6
SUV3
MRPL3
AEP1
ATP18
OXA1
RML2
MEF2
MRPL24
TUF1
MSR1
MRM1
MRPL16
COQ1
MRPL8
MRF1
MEF1
ISM1
MSS2
MRP7
GTF1
GEM1
EXO5
RSM24
MRPS5
CBP1
SCO1
SLS1
YML6
MRP2
RSM23
MRPS12
MRP4
MRPS16
PET8
RSM19
MSK1
PET122
ARP7
MAC1
RET1
SIF2
SIN3
UME1
TFB4
BUR6
MED7
SNF4
GIS1
RAP1
EGD1
NUT1
SIP4
SPT7
MGA2
IES6
GAL11
RIC1
MET30
SNT1
TFC3
STH1
SST2
RPB11
PHO23
ELP3
TOG1
SDS3
SRB8
AZF1
RCO1
SWC5
TUP1
ELP2
EPL1
IRE1
SPT23
IKI3
ARP8
HAC1
ESS1
SOK2
THI3
IXR1
SPT8
CRT10
SFL1
RPH1
SIR3
SPT3
SWC4
ARP6
MHR1
CHD1
CAF130
RPO41
YAF9
RGT1
YTA7
NGG1
BMH1
Fitness
0.75
0.50
0.25
C.
ABP
ABS
ACF
ACH
ACN
ACP
ACT
ACV
ADD
ADH
AMF
AND
ANG
ANH
APH
AVI
BBQ
BCV
BDH
BFP
BHH
BID
BKL
BMK
BNT
BQH
BVF
CCD
CGD
CGQ
CHM
CIB
CLB
CLG
CNT
CPG
CQA
Σ1278b
Gain-of-fitness/S288C
Loss-of-fitness/S288C

Figure S6


A.
Genetic interaction degrees
0
500
1000
2776
125
268
508
Non–Essential
Class I
Class II
Essential
ns
****
****
****
****
****
B.
Co-annotation rate
0.0
0.5
1.0
2776
125
268
508
Non–Essential
Class I
Class II
Essential
ns
*
****
****
****
****
C.
dN/dS (YN00)
0
1
2
2776
125
268
508
Non–Essential
Class I
Class II
Essential
ns
ns
****
****
ns
****

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[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[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[37]. (C) Mean non-synonymous *vs.* synonymous substitution rates (dN/dS) across 1,011 natural yeast isolates using the YN00 method[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_prom300_norm
Density of insertions in the -300 bp region before ATG.
N_atg200_norm
Density of insertions in the -100 bp region before ATG and +100 bp after ATG.
N_term100_norm
Density of insertions in the +100 bp region after stop codon.
N_term300_norm
Density of insertions in the +300 bp region after stop codon.
N_stop200_norm
Density of insertions in the -100 bp region before stop codon and +100 bp region after stop.
N_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.