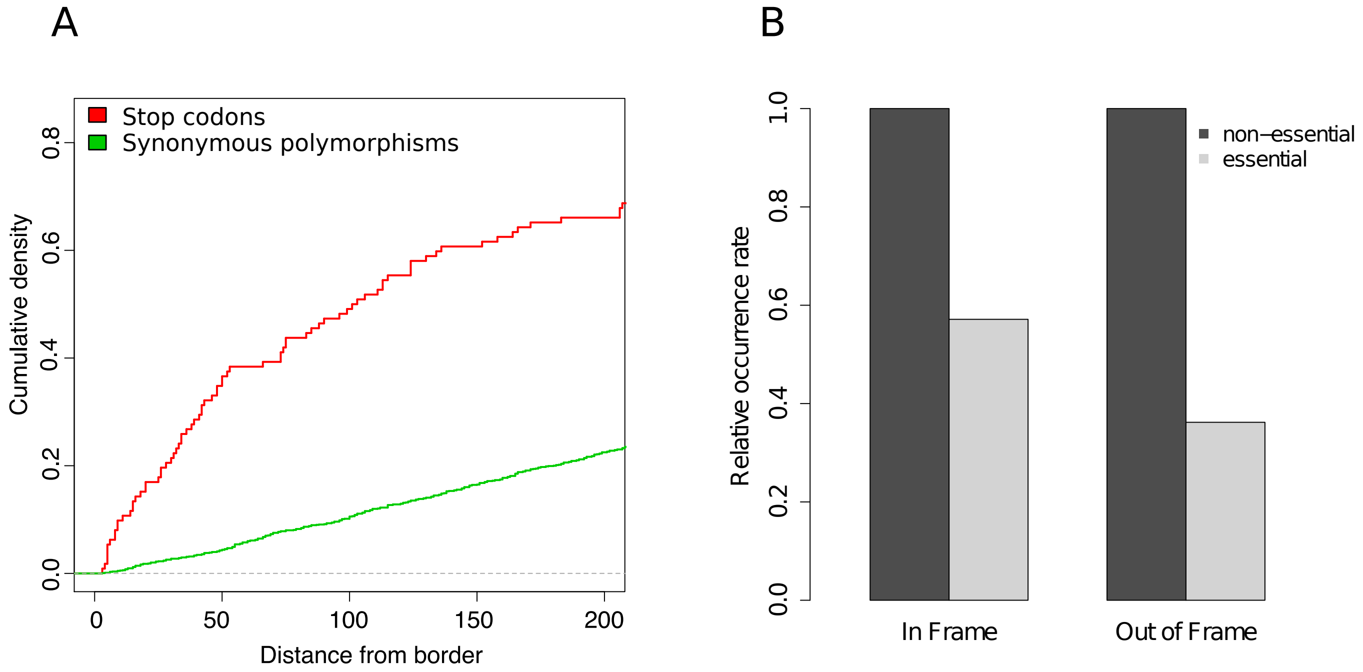
Predicting phenotypic variation in yeast from individual genome sequences

Rob Jelier, Jennifer I. Semple, Rosa Garcia-Verdugo, Ben Lehner

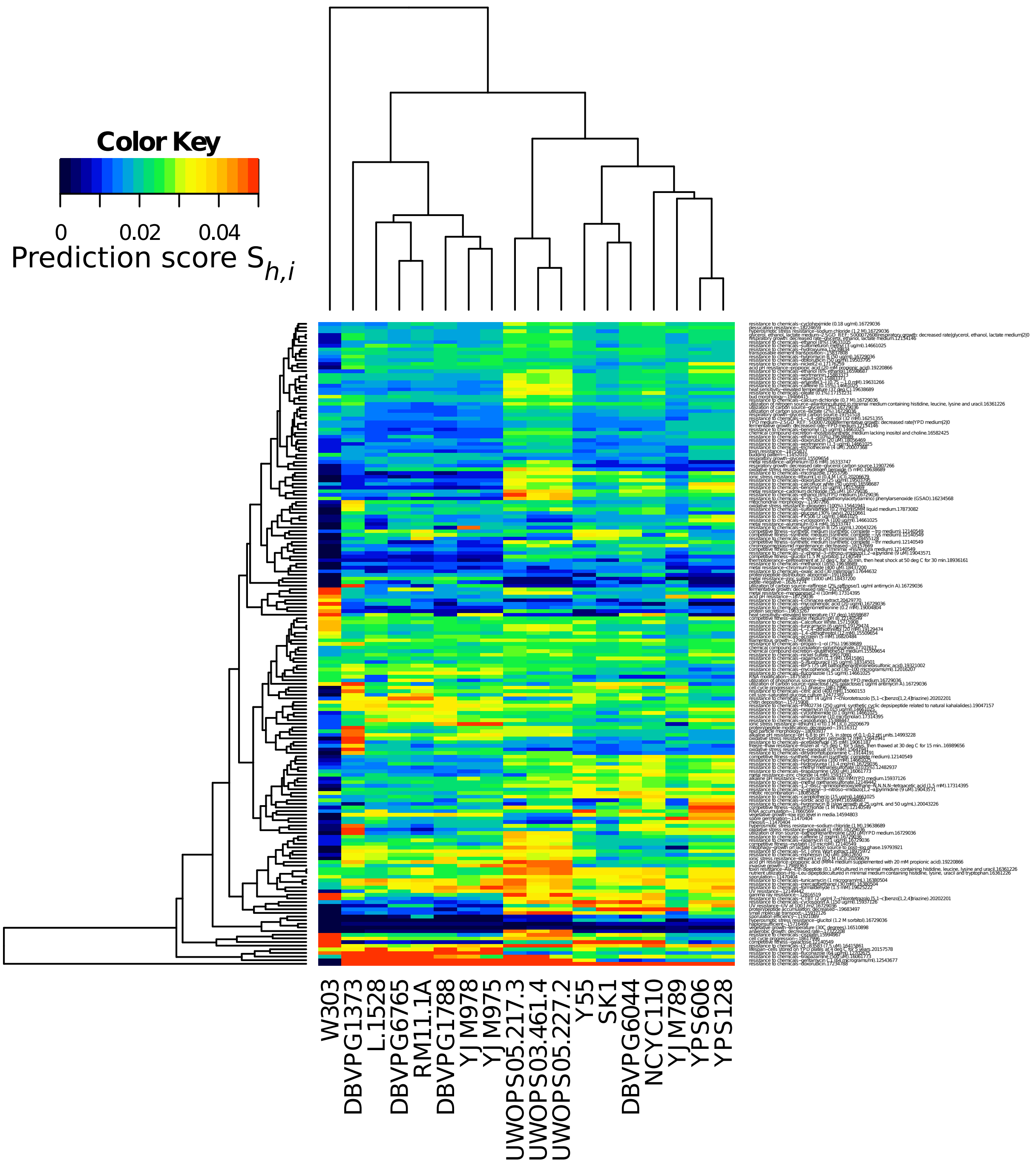
Supplementary information

Supplementary Figures

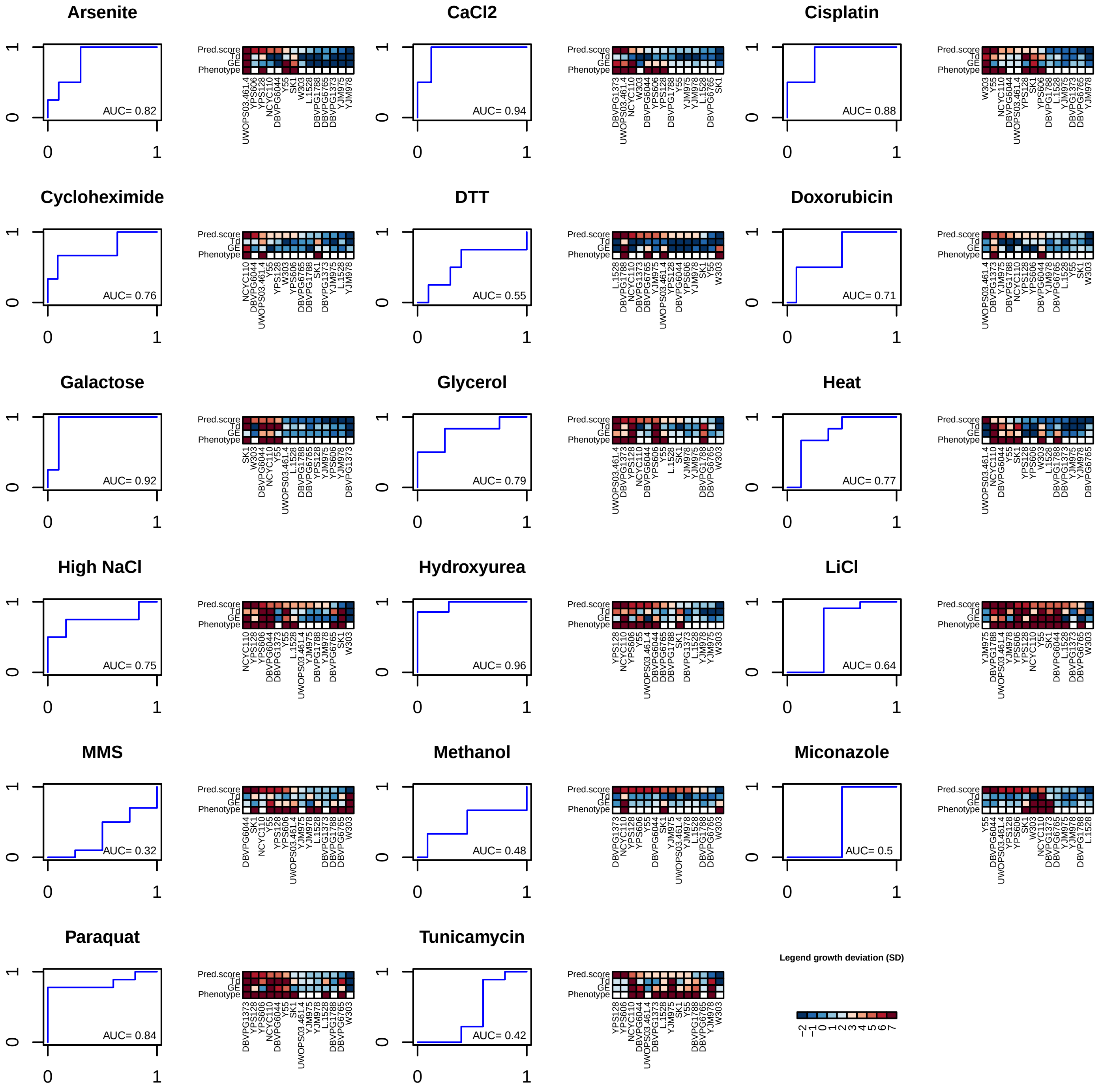
Supplementary figure 1: Heuristic rules for the evaluation of indels and stop codons. A. Stop codons have a marked over-representation in the start or end of genes when compared to synonymous polymorphisms. B. For indels we compare the occurrence rates of sub-classes of indels between essential and non-essential genes. Here the comparison is made for in frame and out of frame indels smaller than 16 base pairs.



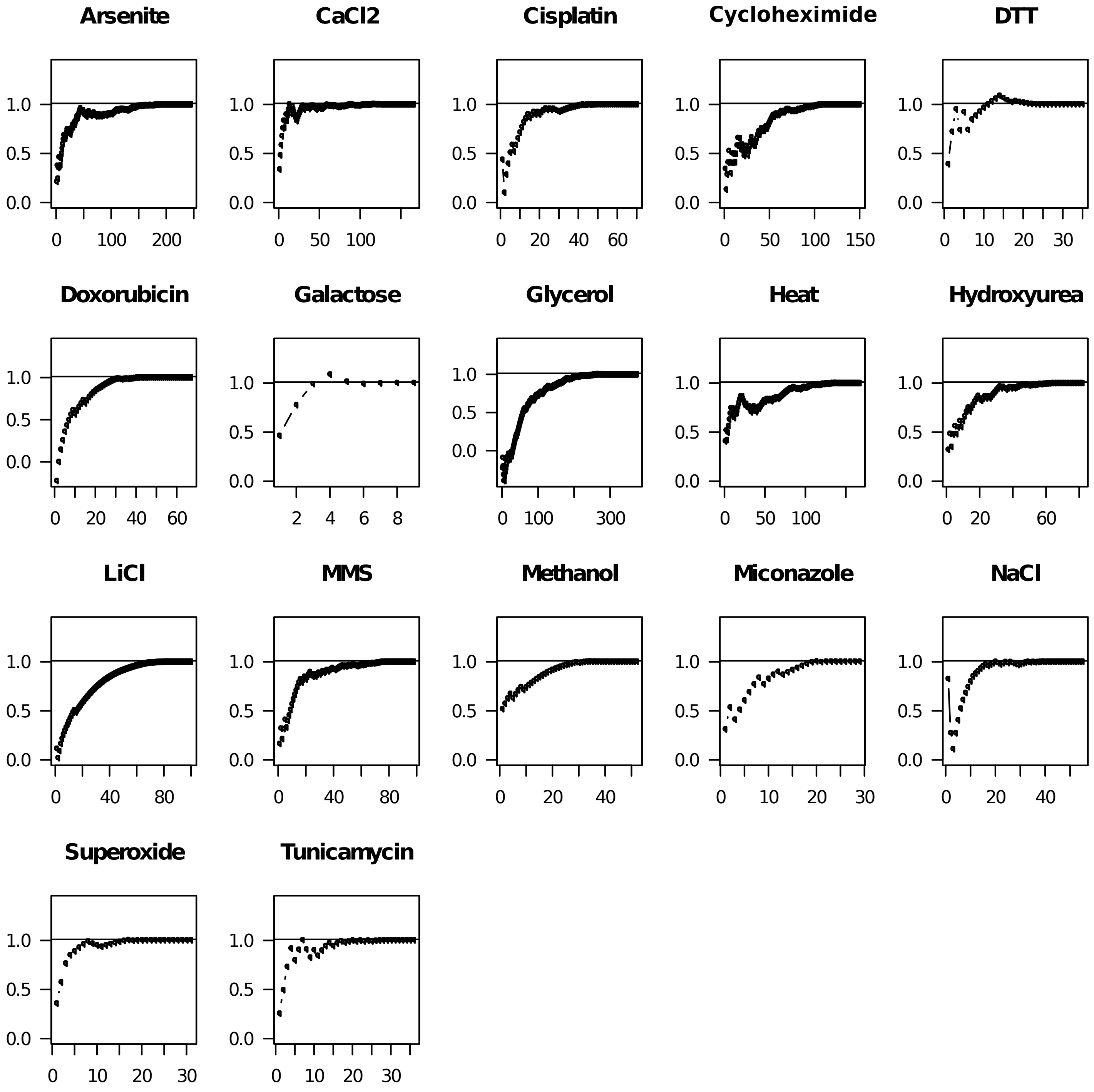
Supplementary figure 2: A heatmap with the $S_{h,i}$ prediction scores for all retrieved genome-wide gene deletion collection screens.



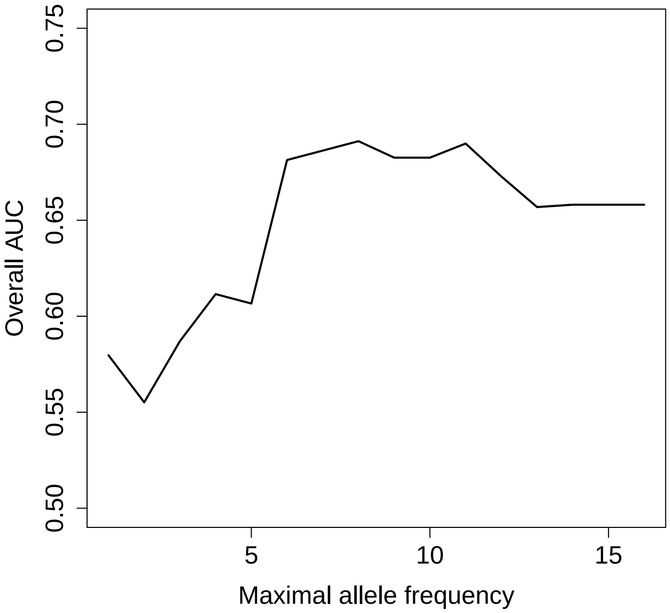
Supplementary figure 3: For all conditions, ROC curves and bars illustrating the underlying data in color-coded boxes for the 14 tested strains. The first bar shows the variation in the prediction scores $S_{h,i}$; the second bar shows doubling time (T_d) deviation, normalized and expressed in standard deviations (SD); the third bar normalized growth efficiency (GE) in SDs; the fourth bar indicates the strains scored with a phenotype (either T_d or GE deviation larger then 2 SD).



Supplementary figure 4: The covariance quantifies the agreement between the overall $S_{h,i}$ score for the strains and the score of a single gene. To compare the covariance across conditions we divide the covariance by the variance of the overall score for that condition. To quantify the number of genes relevant for predictions across strains, we sorted the genes according to their covariance, and counted the number of genes needed to reach a covariance level similar to the overall variance. This figure gives the development of the scaled covariance as genes are added to a running sum as it approximates the actual prediction score $S_{h,i}$.



Supplementary figure 5: Effect on the overall prediction AUC of ignoring polymorphisms that occur beyond the given frequency in our data set.



Supplementary tables

Supplementary table 1. Overview of variants in the 6609 genes of 18 high coverage strains when compared to the reference strain S288c. For each strain the total number of synonymous and non-synonymous SNPs, SNPs that introduce stop codons and indels are quantified across all genes. The last column shows the number of genes that have an estimated probability of altered function over 90%.

Strain	sSNP	nsSNP	Stop	Indel	P(LOF) $>$ 0.9
DBVPG1373	14805	7521	20	514	54
DBVPG1788	14774	7705	20	478	52
DBVPG6044	28530	14731	50	603	231
DBVPG6765	15971	8482	17	655	75
L.1528	14444	7429	17	481	51
NCYC110	25487	12661	34	484	154
RM11.1A	16349	8469	19	737	69
SK1	29543	15550	50	765	246
UWOPS03.461.4	24954	12518	30	460	179
UWOPS05.217.3	22641	11360	27	448	142
UWOPS05.227.2	25265	12766	31	487	182
W303	3084	1724	2	197	16
Y55	25777	13680	36	750	202
YJM789	19845	9469	27	691	64
YJM975	15088	7936	20	484	62
YJM978	15231	8055	20	475	64
YPS128	23550	10429	20	545	91
YPS606	23545	10503	21	554	92

Supplementary table 2 (.txt) Curated test set of gold standard positive (1) and gold standard negative (0) sequence variants in yeast proteins.

Supplementary table 3 (.xls). To assess the rate at which indels or potentially damaging nsSNPs are false reported, we PCR amplified and sequenced 20 indels and 56 SNPs. The file provides the details of each selected variant, the PCR primers for each variant, the expected polymorphism, and the sequencing result .

Supplementary table 4 (.txt). Phenotypic predictions ($S_{h,i}$ scores) for each strain based on each of 180 gene sets retrieved from genome-wide gene deletion collection screens.

Supplementary table 5 (.txt). The minimal doubling time and growth efficiency, normalized relative to the growth of S288c as the logarithmic strain coefficient (LSC), for all growth experiments used in this manuscript.

Supplementary table 6 (.txt). The mean minimal doubling time and growth efficiency LSC for every condition.

Supplementary table 7. The influence on performance of the threshold used to define a phenotype, the choice of response variable, the type of variation considered (var 1 = nsSNPs, var 2 = nsSNPs + indels, var 3 = nsSNPs+stops, and pruning the gene sets per condition. Gene set size (N), the Pubmed ID for the article from which the gene set was taken and the number of strains (Str) with a phenotype are provided.

			Default		SD3		SD4		SD5		SD6		SD7		SD8		rate Only		eff Only		rate ∧ eff		var 1	var 2	var 3	pruning thr=0		pruning thr=1.5	
Condition	PMID	N	Str	AUC	Str	AUC	Str	AUC	Str	AUC	Str	AUC	Str	AUC	Str	AUC	Str	AUC	Str	AUC	Str	AUC	AUC	AUC	AUC	N	AUC	N	AUC
Galactose	12140549	9	4	0.93	4	0.93	4	0.93	4	0.93	4	0.93	3	0.82	1	0.77	4	0.93	2	0.83	2	0.83	0.90	0.90	0.90	5	0.93	4	1.00
Miconazole 4ug/ml	17553796	29	4	0.50	3	0.45	3	0.45	3	0.45	3	0.45	3	0.45	1	0.46	1	0.46	4	0.50	1	0.46	0.28	0.55	0.28	15	0.68	12	0.85
Heat 40C	19638689	167	6	0.77	6	0.77	3	0.88	2	0.83	1	0.92	1	0.92	1	0.92	5	0.84	6	0.77	5	0.84	0.60	0.69	0.73	126	0.73	105	0.73
NaCl 1.25M	16729036	55	8	0.75	7	0.73	6	0.63	5	0.73	5	0.73	5	0.73	2	0.92	7	0.73	7	0.84	6	0.83	0.63	0.75	0.71	39	0.75	31	0.85
Paraquat 400 ug/ml	16729036	31	9	0.84	8	0.81	7	0.88	6	0.79	5	0.93	5	0.93	2	0.92	9	0.84	6	0.79	6	0.79	0.84	0.84	0.84	23	0.76	21	0.76
Tunicamycin 1.5ug/ml	16380504	36	9	0.42	7	0.45	5	0.47	4	0.55	3	0.45	2	0.67	2	0.67	6	0.35	9	0.42	6	0.35	0.42	0.47	0.40	26	0.42	24	0.44
MMS 0.01%	12482937	98	10	0.33	3	0.42	2	0.38	2	0.38	1	0.00	1	0.00	1	0.00	5	0.40	7	0.39	2	0.38	0.33	0.33	0.30	83	0.33	73	0.38
CaCl2 0.7M	16729036	166	6	0.94	3	1.00	3	1.00	2	0.96	1	0.85	0	-	0	-	0	-	6	0.94	0	-	0.75	0.88	0.88	140	0.79	129	0.90
LiCl 225mM	20206679	100	11	0.64	10	0.70	8	0.71	8	0.71	7	0.59	7	0.59	7	0.59	11	0.64	8	0.71	8	0.71	0.55	0.64	0.58	81	0.70	73	0.70
Arsenite 5mM	19631266	245	4	0.83	3	0.76	3	0.76	2	0.83	2	0.83	1	1.00	1	1.00	3	0.85	3	0.76	2	0.79	0.85	0.83	0.78	202	0.78	181	0.80
Hydroxyurea 15mg/ml	16729036	82	7	0.96	6	0.92	5	0.91	1	0.92	1	0.92	0	-	0	-	5	0.89	4	0.73	2	0.67	0.94	0.96	0.94	72	0.94	65	0.94
Methanol 15%	19638689	52	3	0.48	1	0.92	1	0.92	1	0.92	1	0.92	0	-	0	-	1	0.92	3	0.48	1	0.92	0.42	0.48	0.39	35	0.21	31	0.21
Glycerol	11907266	374	6	0.79	4	0.70	4	0.70	4	0.70	3	0.88	3	0.88	3	0.88	5	0.80	6	0.79	5	0.80	0.75	0.77	0.81	326	0.77	306	0.85
DTT 2.5mM	16380504	35	4	0.55	2	0.46	2	0.46	1	0.92	1	0.92	1	0.92	1	0.92	1	0.92	4	0.55	1	0.92	0.65	0.53	0.58	24	0.68	23	0.63
Cisplatin 150ug/ml	17093137	70	6	0.88	5	0.80	3	0.76	3	0.76	2	0.83	1	1.00	0	-	6	0.88	2	0.79	2	0.79	0.94	0.94	0.88	61	0.92	49	0.90
Cycloheximide 0.1ug/ml	16729036	150	3	0.76	3	0.76	1	1.00	1	1.00	0	-	0	-	0	-	2	0.58	1	1.00	0	-	0.76	0.76	0.79	133	0.70	111	0.67
Doxorubicin 20ug/ml	18056469	67	2	0.71	1	0.92	0	-	0	-	0	-	0	-	0	-	1	0.92	2	0.71	1	0.92	0.58	0.63	0.63	52	0.83	44	0.79
Overall				0.69		0.70		0.71		0.72		0.71		0.71		0.71		0.69		0.67		0.69	0.65	0.68	0.66		0.68		0.71
Median		70	6	0.76	4	0.76	3	0.76	2	0.81	2	0.85	1	0.85	1	0.88	5	0.84	4	0.76	2	0.79	0.65	0.75	0.73	61	0.75	49	0.79
Mean		104	6	0.71	4.47	0.74	3.53	0.74	2.88	0.77	2.35	0.75	1.94	0.74	1.29	0.73	4.24	0.75	4.71	0.71	2.94	0.73	0.66	0.70	0.67	84.88	0.70	75.41	0.73

Supplementary table 8 (.txt) Per gene and cumulative covariance over variance statistics.

Supplementary file (dataExplorer.rar).

A tool to explore the data that underlies our predictions. Expand the archive and read the Readme.txt for details. The program requires that a recent version of Java is installed.

Supplementary note

Alternative sources of gene sets

In the main text of our article we describe how gene sets from reverse genetics studies can be used to generate phenotypic predictions. However, we also explored the use of gene sets defined by gene expression profiling. We obtained expression data for five conditions that we had experimentally evaluated. As shown in the table below, overall performance is poor with a prediction AUC of 0.43. Three conditions predict poorly and one prediction predicts reasonably. Interestingly, the predictions for high salt are good. Investigating this further we found that the gene ENA1 has a strong influence on the prediction, and indeed a recent study suggested ENA1 harbours an important QTL for saline stress sensitivity[1]. Indeed ENA1 is a false negative in the gene set defined by the genome-wide reverse genetic screen because the deletion is not present in the deletion collection. Although using gene sets defined by expression changes may be useful in some cases, it is not yet clear to us how to best incorporate this data into the overall prediction model. Consistent with this, previous work has shown that the sets of genes that change expression in a particular condition and the set of genes that affect growth in that condition when mutated show little overlap *[2, 3, 4, 5, 6, 7]*.

Supplementary note table 1: Prediction results for five conditions where the gene sets were taken from mRNA profiling studies.

Condition	PMID	Set size	AUC
NaCl 1.25M	18753408	639	0.96
Tunicamycin 1.5ug/ml	10929718	129	0.2
MMS 0.01%	16709784	149	0.33
Arsenite 5mM	15575969	48	0.38
Cycloheximide 0.1ug/ml	10929718	68	0.82
Overall			0.43
Median			0.38

Assuming epistasis when calculating the prediction score

Systematic genetic interaction screens [8, 9] and a limited number of examples [10, 11] have highlighted the importance of epistasis in determining the phenotypic outcome of mutations when they are combined. We suspect that considering non-additive epistatic interactions between sequence variants will also be important to improve phenotypic predictions for individuals. As the specific interactions are not currently known or readily deducible, it is not possible to explicitly include them. However, we could make assumptions about the predominant type of epistasis, and observe the effects on predictions.

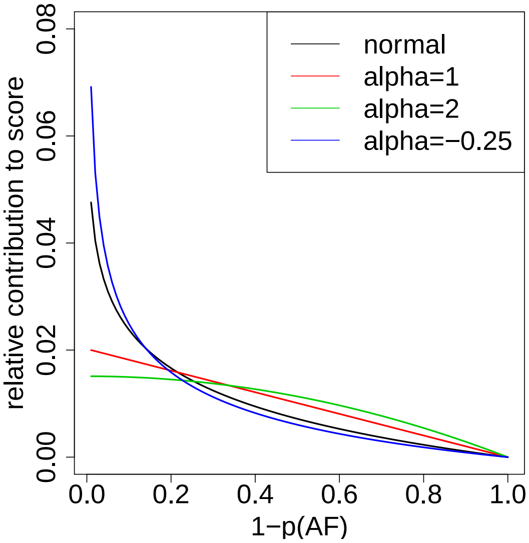
To examine the effects of assuming overall positive or negative epistasis we used the following formula for the prediction score:

$$S_{h,i} = \sum_{j=1}^{n} 1 - (1 - P_j(AF))^{\alpha}$$

Where $S_{h,i}$ is the prediction score for a strain h in a condition i, calculated over all genes j. The parameter alpha modifies the relative contributions of the estimated likelihoods of altered function of the genes $P_j(AF)$. To compensate for divergence between strains we normalize to the same expected score per gene set.

The figure below demonstrates the effect of varying alpha. The figure shows the relative contribution to $S_{h,i}$ of 100 genes with $P_j(AF)$ equally spaced between 0 and 0.99. The normal scheme takes the sum of the log transformed $1-P_j(AF)$ scores, this distribution of relative contributions is approximated when alpha approaches zero. If alpha is greater than zero, the relative contributions of different $P_j(AF)$ become more similar. In this way a large prediction score is more likely to reflect several genes with moderate scores rather than a single gene with a large score, which is similar to synergistic epistasis. If alpha is negative, fewer genes will have a significant influence on the overall score, which is similar to antagonistic epistasis. As an extreme example of antagonistic epistasis we also included a regime where for a gene set only the gene with the highest $P_j(AF)$ is considered.

Supplementary note figure 1: The relative contribution of genes to the prediction score for different weighting schemes. There are 100 genes with gene scores equidistantly from P(AF)=0 to P(AF)=0.99.



The results are presented in the following table. The most striking result is that using only the gene with the biggest p(AF) as a prediction score, predictive performance is very poor. The other methods differ only slightly, but it appears there is some evidence against assuming overall synergistic epistasis, yet no evidence for or against assuming antagonistic epistasis.

Supplementary note table 2: Performance for different schemes for calculating the prediction score.

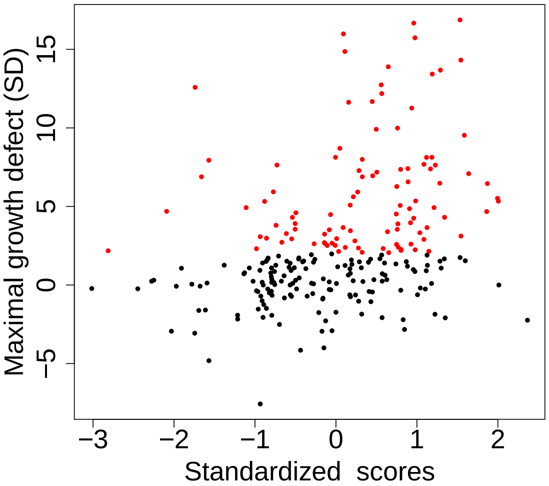
condition	normal	alpha=1	alpha=2	alpha=-0.25	highest P(AF)
Overall	0.69	0.68	0.66	0.69	0.55
Hydroxyurea 15mg/ml	0.96	0.94	0.88	0.98	0.8
CaCl2 0.7M	0.96	0.92	0.88	0.96	0.54
Cisplatin 150ug/ml	0.94	0.96	0.96	0.94	0.27
Galactose	0.9	0.9	0.9	0.9	0.9
Cykloheximide 0.1ug/ml	0.85	0.79	0.76	0.85	0.33
Arsenite 5mM	0.83	0.85	0.83	0.83	0.38
Superoxide (paraquat) 400 ug/ml	0.82	0.93	0.89	0.84	0.62
LiCl 225mM	0.76	0.85	0.88	0.73	0.3
Glycerol	0.71	0.73	0.73	0.73	0.54
Heat 40C	0.69	0.69	0.58	0.67	0.81
DTT 2.5mM	0.63	0.6	0.68	0.68	0.68
Doxorubicin 20ug/ml	0.63	0.67	0.67	0.5	0.42
NaCl 1.25M	0.56	0.46	0.33	0.65	0.56
Miconazole 4ug/ml	0.53	0.5	0.5	0.48	0.63
Methanol 15%	0.48	0.48	0.48	0.55	0.33
MMS 0.01%	0.43	0.4	0.4	0.4	0.5
Tunicamycin 1.5ug/ml	0.38	0.36	0.33	0.36	0.56

A linear model to relate the prediction scores to phenotypes

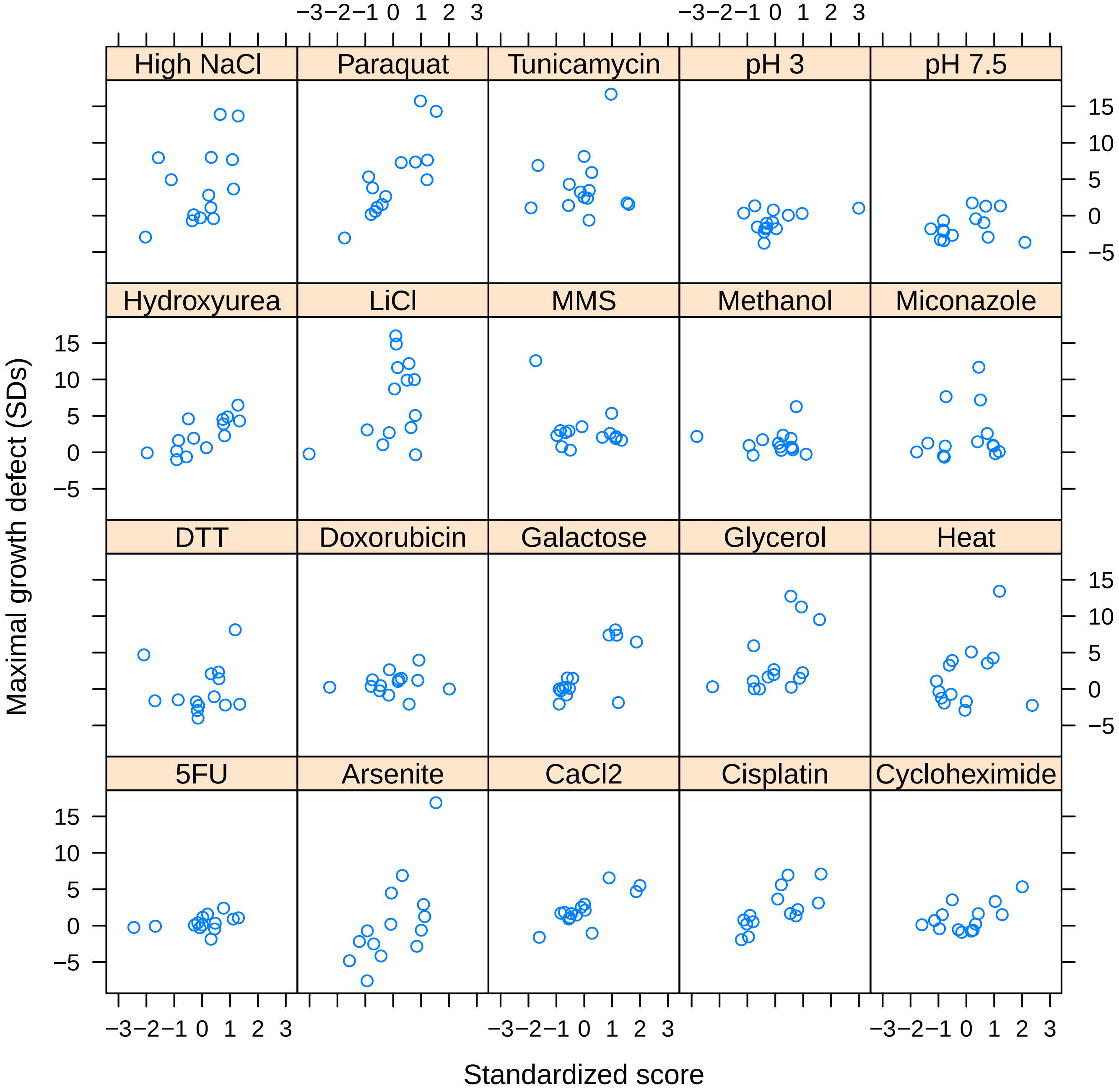
In the main text the output variable of the prediction process is the ranking of the strains based on their growth in a particular condition. We chose to use ranks because in several aspects the inputs to the predictions are not quantitative. For instance, a gene is either in the gene set or it is not, and no information is available on the actual size of the effect caused by the deletion of the gene. Additionally, the use of ranks makes the predictions directly comparable across conditions. This is relevant as the expected prediction score is not similar for every gene set. This is because some genes are much more prone to accumulate variations, e.g. because of repetitive sequences or the location on the chromosome, and these genes can be considerably more plentiful in some gene sets than in others. Further, using the threshold on the effect size is a simple and transparent way to combine the two response variables, i.e. growth rate and growth efficiency, requiring no further assumptions. The threshold was varied and predictions were shown to be better for larger effects. Here we illustrate that our conclusions do not depend on the chosen statistical methodology.

As an input to our analysis we used the $S_{h,i}$ prediction scores per strain over all conditions, and the deviations from normal for both growth rate and growth efficiency expressed in standard deviations. To combine the two response variables we simply took the largest deterioration of growth of the two for every condition and strain. As noted above, the $S_{h,i}$ prediction scores should not be compared directly between conditions. To make the comparisons the scores were standardized for every conditon. The analyses were performed in R, an environment for statistical computing [12]. We used the lme4 package [13] to fit a linear mixed effects model where the combined phenotype score is predicted by the prediction scores (as a fixed effect). To correct for any condition specific effects on the response variable, the conditions are fitted as fixed effects, also any strain specific effect is taken into account by fitting them as a random effect. An ANOVA comparison of the model with just the fits for the conditons and the strains to the model which included the prediction scores, demonstrates that the prediction scores improve the model with high statistical significance ($p = 8 \cdot 10^{-8}$) with both the AIC and BIC coefficients decreasing. The figure is a scatterplot that illustrates the overall relation between the standardized prediction scores and the combination of the two response variables. We also show the scores against the phenotypic variable for every condition separate.

Supplementary note figure 2: A scatterplot illustrating the relation between the standardized prediction score and the maximal growth defect across the conditions. All strains/condition pairs that exceed 2 SD are displayed in red.



Supplementary note figure 3: Scatter plots for every condition of the standardized sore vs. the maximal growth defect. Also the conditions that had less than 2 strains with a defect larger than 2 standard deviations, i.e. pH 3, pH 7.5 and 5-FU are shown.



Phenotype predictions through analysis of variation in transcription factor binding sites

Our predictions only relied on variations in protein coding sequences. However, variations in the regulatory regions of genes are also likely to contribute to phenotypes. Earlier, we evaluated numerous features of promoters and binding sites to predict if a variation in an experimentally-defined transcription factor binding site [14] is likely to be detrimental [15]. We fitted a generalized linear model to integrate the numerous features with conservation of the binding sites as an outcome variable. The assumption we use here is that sites that are conserved are more likely to cause a phenotypic change when disrupted. Our system predicts the likelihood of conservation for every transcription factor binding site with a variation. We treat these likelihoods the same as the $P(AF)$ as described in the main text to make prediction scores. The prediction results are shown in below. Overall very poor predictions are observed with an overall AUC near random. Correcting for the number of binding sites in a promoter does not change this conclusion. We think that the main reason for this result is that the annotation of important regulatory regions is likely to be very incomplete.

Supplementary note table 3: Prediction results for transcription factor binding sites. The fourth column represents the fraction of the gene set where a gene has a perturbation in the binding sites of its promoter in at least one strain.

Condition	Gene set size	AUC	Gene set fraction with variations
Galactose	9	0.9	0.11
Miconazole 4ug/ml	29	0.1	0.1
Heat 40C	167	0.46	0.18
NaCl 1.25M	55	0.52	0.16
Superoxide (paraquat) 400 ug/ml	31	0.16	0.1
Tunicamycin 1.5ug/ml	36	0.82	0.17
MMS 0.01%	98	0.55	0.17
CaCl2 0.7M	166	0.42	0.17
LiCl 225mM	100	0.58	0.12
Arsenite 5mM	245	0.25	0.17
Hydroxyurea 15mg/ml	82	0.29	0.21
Methanol 15%	52	0.67	0.23
Glycerol	374	0.54	0.15
DTT 2.5mM	35	0.33	0.17
Cisplatin 150ug/ml	70	0.59	0.17
Cykloheximide 0.1ug/ml	150	0.48	0.15
Doxorubicin 20ug/ml	67	0.58	0.19
Overall		0.48	0.16
Median	70	0.52	0.17

References

[1] Cubillos, F. A. *et al.* Assessing the complex architecture of polygenic traits in diverged yeast populations. *Mol Ecol* **20**, 1401–1413 (2011).

[2] Haugen, A. C. *et al.* Integrating phenotypic and expression profiles to map arsenic-response networks. *Genome Biol* **5**, R95 (2004).

[3] Deutschbauer, A. M., Williams, R. M., Chu, A. M. & Davis, R. W. Parallel phenotypic analysis of sporulation and postgermination growth in saccharomyces cerevisiae. *Proc Natl Acad Sci U S A* **99**, 15530–15535 (2002).

[4] Fry, R. C., Begley, T. J. & Samson, L. D. Genome-wide responses to dna-damaging agents. *Annu Rev Microbiol* **59**, 357–377 (2005).

[5] Birrell, G. W. *et al.* Transcriptional response of saccharomyces cerevisiae to dna-damaging agents does not identify the genes that protect against these agents. *Proc Natl Acad Sci U S A* **99**, 8778–8783 (2002).

[6] Smith, J. J. *et al.* Expression and functional profiling reveal distinct gene classes involved in fatty acid metabolism. *Mol Syst Biol* **2**, 2006.0009 (2006).

[7] Yeger-Lotem, E. *et al.* Bridging high-throughput genetic and transcriptional data reveals cellular responses to alpha-synuclein toxicity. *Nat Genet* **41**, 316–323 (2009).

[8] Costanzo, M. *et al.* The genetic landscape of a cell. *Science* **327**, 425–431 (2010).

[9] Lehner, B., Crombie, C., Tischler, J., Fortunato, A. & Fraser, A. G. Systematic mapping of genetic interactions in caenorhabditis elegans identifies common modifiers of diverse signaling pathways. *Nat Genet* **38**, 896–903 (2006).

[10] Gerke, J., Lorenz, K. & Cohen, B. Genetic interactions between transcription factors cause natural variation in yeast. *Science* **323**, 498–501 (2009).

[11] Dowell, R. D. *et al.* Genotype to phenotype: a complex problem. *Science* **328**, 469 (2010).

[12] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria (2010). ISBN 3-900051-07-0.

[13] Bates, D. & Sarkar, D. *lme4: Linear mixed-effects models using S4 classes* (2007).

[14] Harbison, C. T. *et al.* Transcriptional regulatory code of a eukaryotic genome. *Nature* **431**, 99–104 (2004).

[15] Francesconi, M., Jelier, R. & Lehner, B. Integrated genome-scale prediction of detrimental mutations in transcription networks. *PLoS Genet* **7**, e1002077 (2011).