nature genetics

Predicting phenotypic variation in yeast from individual genome sequences

Rob Jelier[1], Jennifer I Semple[1], Rosa Garcia-Verdugo[1] & Ben Lehner[1,2]

A central challenge in genetics is to predict phenotypic variation from individual genome sequences. Here we construct and evaluate phenotypic predictions for 19 strains of *Saccharomyces cerevisiae*. We use conservation-based methods to predict the impact of protein-coding variation within genes on protein function. We then rank strains using a prediction score that measures the total sum of function-altering changes in different sets of genes reported to influence over 100 phenotypes in genome-wide loss-of-function screens. We evaluate our predictions by comparing them with the observed growth rate and efficiency of 15 strains tested across 20 conditions in quantitative experiments. The median predictive performance, as measured by ROC AUC, was 0.76, and predictions were more accurate when the genes reported to influence a trait were highly connected in a functional gene network.

A fundamental challenge in genetics is to predict differences in the phenotypes of individuals by using knowledge of their genetic variation. Rapid advances in sequencing technology have brought individual human whole-genome sequences within reach[1–4], and pilot projects to sequence individual genomes have been completed[5,6]. However, the possibility of predicting phenotypic variation from the genomic sequences of individuals is still largely unexplored[7]. Here we use the budding yeast *S. cerevisiae* as a model system to develop and assess a methodology for predicting phenotypic variation using genomic sequences. Budding yeast, which has complex phenotypes, provides many advantages for this type of study, including the diversity of systematic genetic and functional genomic data available that provide a rich overview of gene function[8]. Budding yeast can also be maintained in the laboratory as homozygotes for all alleles, avoiding the complication of heterozygosity[3,9]. Moreover, whole-genome sequences are available for many individual strains through the *S. cerevisiae* resequencing project[5]. Finally, large-scale experiments can be performed to evaluate the accuracy of predictions for many different phenotypes.

We developed a procedure for predicting phenotypic differences among individual *S. cerevisiae* strains and then evaluating these predictions, which consisted of three main steps (**Fig. 1a**). We first estimated for each gene in a strain the likelihood that protein function was altered as a result of sequence variations identified relative to a reference strain. Next, using gene sets derived from high-throughput reverse genetic screens, we estimated the total perturbation in the genes relevant for each phenotypic trait in each individual. This step allowed us to rank the strains according to their likelihood of being affected for each phenotype. Finally, we performed quantitative phenotyping experiments and compared the predicted rankings of strains to their observed phenotypic variation.

Partial genomic sequences are available for 38 *S. cerevisiae* strains, including the S288c reference strain[5]. Of these strains, 19 have at least 75% coverage at an error cutoff of one error per 10,000 bp, and we used these strains for our analysis. We estimated the effects of nonsynonymous SNPs (nsSNPs), premature stop codons, and insertions or deletions (indels) separately and then combined the estimates of their influence. Nonsynonymous variants were by far the most frequent, accounting for on average 94% of the analyzed variants (**Supplementary Table 1**). Numerous approaches have been developed to predict the effects of polymorphisms on protein function (see ref. 10 for an overview), with the general conclusion being that residue conservation is the best single predictor of effect[11]. We based our approach on the SIFT algorithm[12], which evaluates a multiple-sequence alignment of homologous proteins, adapting the algorithm to yeast by using a compiled yeast-specific test set (**Supplementary Table 2**). SIFT performed well on the yeast-specific test set (**Fig. 1**), although the coverage of the test set by SIFT, which depends on the availability of a multiple-sequence alignment, was only 73%. Both performance and coverage were augmented by improving the retrieval of orthologous sequences, which enhanced the underlying multiple-sequence alignments (**Fig. 1b**,**c** and Online Methods).

A similar test set does not exist for premature stop codons or indels, so we resorted to a set of heuristic rules to evaluate these variants (see Online Methods and **Supplementary Fig. 1**). For indels, we compared the occurrence rates in essential and nonessential genes, assuming the rate in essential genes to mostly reflect functionally neutral or falsely reported variations. Indeed, many indels were estimated to be falsely detected, a prediction that was confirmed by sequencing (8 out of 20 tested indels were not verified), whereas for missense polymorphisms, 50 out of 56 variants were confirmed (**Supplementary Table 3**). Using the score derived for each nsSNP, premature stop codon and indel,

[1]European Molecular Biology Laboratory—Centre for Genomic Regulation (EMBL-CRG) Systems Biology Research Unit, Centre for Genomic Regulation, Barcelona, Spain. [2]Institució Catalana de Recerca Estudis Avançats (ICREA), Centre for Genomic Regulation, Barcelona, Spain. Correspondence should be addressed to B.L. (ben.lehner@crg.eu).

Received 31 May; accepted 19 October; published online 13 November; doi:10.1038/ng.1007

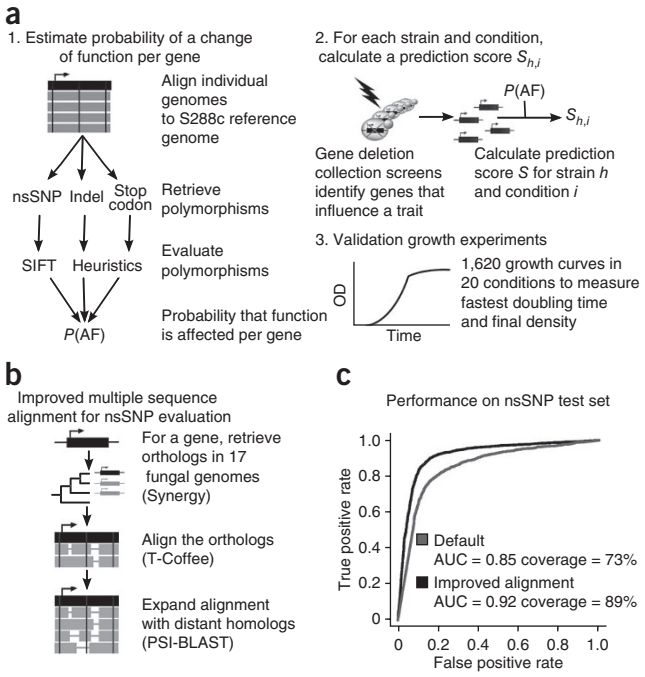

Figure 1 Genome-wide reverse genetic approach used to predict *S. cerevisiae* phenotypic variation from genomic sequences. (**a**) Overview of the procedure. First, polymorphisms are identified from high-coverage whole-genome sequences aligned to the S288c reference genome and are evaluated and combined to estimate for each gene whether its function had been altered (see Online Methods). Second, an *S* score is calculated for each individual strain, which predicts whether a given phenotype will be affected relative to the reference strain. The set of genes relevant for each condition is derived from genome-wide reverse genetic screens using the gene deletion collection. Third, phenotypic predictions are evaluated using quantitative phenotyping experiments. Growth experiments were performed under diverse environmental conditions or in the presence of small molecule inhibitors. For each phenotype, strains were classified according to deviations in either minimal doubling time or growth efficiency beyond a given threshold. The AUC from the ROC curve is used to characterize how well the strains with phenotypes are prioritized when sorted according to *S* score. (**b**) To evaluate the effect of nonsynonymous SNPs (resulting in amino acid alterations), the SIFT algorithm is used with a protein sequence alignment as input. Known fungal orthologs are identified, and a more sophisticated alignment algorithm is implemented before the retrieval of more distant homologs. (**c**) This substantially improved both coverage and prediction performance for a large reference set of polymorphisms with known functional consequences.

we first separately estimated the probability that the given mutation altered the function of a protein and then combined these probabilities naively to estimate the probability that each gene in the genome had an altered function (see Online Methods).

To associate gene variations with growth under a specific condition, we used data from genome-scale reverse genetic screens. The *S. cerevisiae* gene deletion collection[13] has been used to systematically identify sets of genes required for many different processes. A total of 177 gene sets for 115 distinct phenotypes were retrieved from the *Saccharomyces* Genome Database (SGD)[14], and we used data from these sets to predict whether strains were affected for each phenotype relative to the reference strain. To calculate a prediction score *S* for a strain *h* and a condition *i*, we combined the estimated change-of-function probabilities per gene, correcting for the overall sequence divergence of each strain by normalizing to the expected score per gene. The complete set of predictions is available in **Supplementary Table 4** and is illustrated in **Supplementary Figure 2**.

To assess the performance of our predictions, we conducted a total of 1,620 growth experiments using 15 strains across 20 conditions. We measured the maximum growth rate (doubling time) and the growth efficiency (yield) of each strain under each condition and compared them to the growth of the reference strain under the same condition (**Supplementary Tables 5** and **6**). A strain was considered defective for growth in a particular condition if its relative growth rate or efficiency deviated by more than 2 s.d. from growth under normal conditions. In three conditions, only one strain or none showed a relative growth

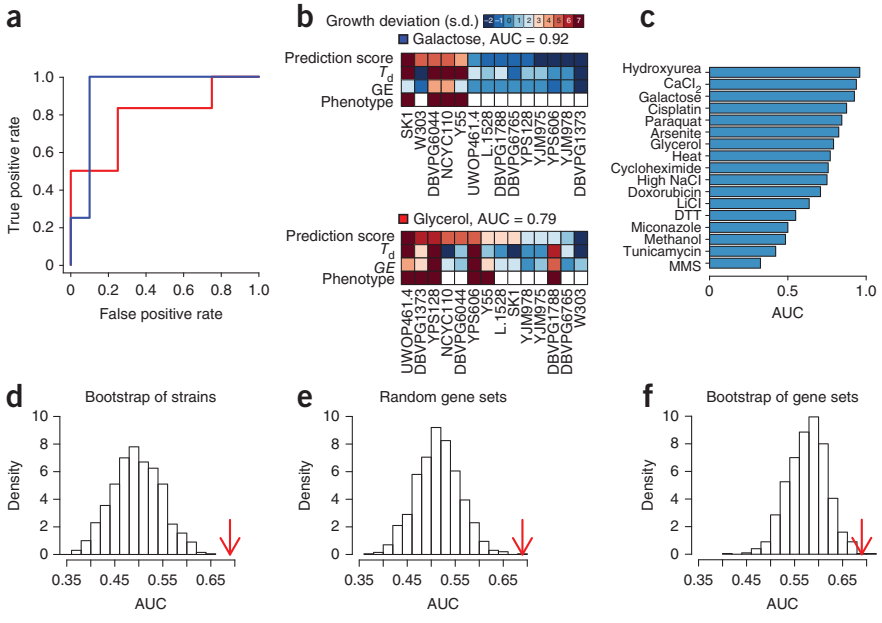

Figure 2 Testing predictive score performance for the identification of phenotypic variations in *S. cerevisiae* strains. (**a**) ROC curves were used to evaluate the prediction of growth phenotypes when strains were grown on the galactose (blue) and glycerol (red) alternative carbon sources. (**b**) Quantitative growth data for the 14 tested strains. Row 1, variation in the prediction score *S*; row 2, normalized deviation in doubling time (T_d) expressed in s.d.; row 3, normalized deviation in growth efficiency (GE) expressed in s.d.; row 4, the strains scored with a phenotype (deviation in either T_d or GE > 2 s.d.). (**c**) AUC performance for 17 conditions in which more than one strain was identified as having a growth defect. Random prediction gives an AUC of 0.5. (**d**–**f**) The significance of the overall AUC prediction is illustrated using three randomization experiments. The red arrow indicates the observed overall AUC. (**d**) A bootstrap of the strains ($P < 0.0001$). (**e**) Replacing the contents of the gene sets with genes randomly drawn from the set of genes represented by the haploid gene deletion collection ($P < 0.0001$). (**f**) A bootstrap of the gene sets. In this case, the mean of the distribution is shifted to >0.5 as a result of correlations between growth phenotypes ($P < 0.001$).

defect, and these conditions were not considered further. For every other condition, we sorted the strains according to their *S* prediction scores and evaluated how well these rankings predicted growth defects by determining the area under the receiver operating characteristic (ROC) curve (AUC)[15]. The AUC can be interpreted as the chance that a randomly selected strain with a phenotype is correctly distinguished from a randomly selected strain without a phenotype.

ROC curves for growth in galactose, which had a high-scoring AUC of 0.92, and glycerol, which had a reasonable AUC of 0.79, are shown in **Figure 2a**. The data on which the ROC curves are based are shown in **Figure 2b** (plots for the other conditions are available in **Supplementary Fig. 3**). Across all 17 conditions, the median AUC was 0.76 (**Fig. 2c**), and the overall AUC performance, calculated by combining the ranked strains across the conditions, was 0.69 ($P = 5.0 \times 10^{-7}$, Wilcoxon rank-sum test). Randomizing the matching of strains to phenotypes (**Fig. 2d**), genes to gene sets (**Fig. 2e**) or gene sets to conditions (**Fig. 2f**) confirmed that the predictions were highly specific ($P < 0.001$ in all cases).

For some phenotypes, multiple genome-wide screens have been performed, sometimes identifying gene sets that correlate only poorly between screens[16] and thereby resulting in very different predictions in our framework (**Supplementary Fig. 4**). The reliability of each gene set can be evaluated by quantifying the functional consistency of the set using an integrated gene network, such as YeastNet version 2.0 (refs. 17,18). To perform this analysis, we measured the extent to which genes within a set were connected to each other through predicted functional relationships relative to their connections to other genes. This comparison can be expressed as a network AUC for each gene set. Gene sets that were determined to be reliable according to the gene network (high network AUC) tended to show better predictive performance (Pearson's correlation between network AUC and prediction AUC = 0.5, $P = 0.0042$) (**Fig. 3a**). This correlation was observed across conditions as well as when comparing alternative gene sets for a particular condition. Thus, prediction performance was substantially influenced by the quality of the gene sets used to make the predictions. When alternative gene sets were available, we therefore used the set with the highest network AUC to make predictions.

The reliability of predictions also correlated with the magnitude of a phenotype. When phenotypes were defined by higher deviation thresholds, performance improved (**Fig. 3b** and **Supplementary Table 7**). For example, the median prediction AUC rose to 0.85 when a threshold of 6 s.d. was used. A separate observation was that

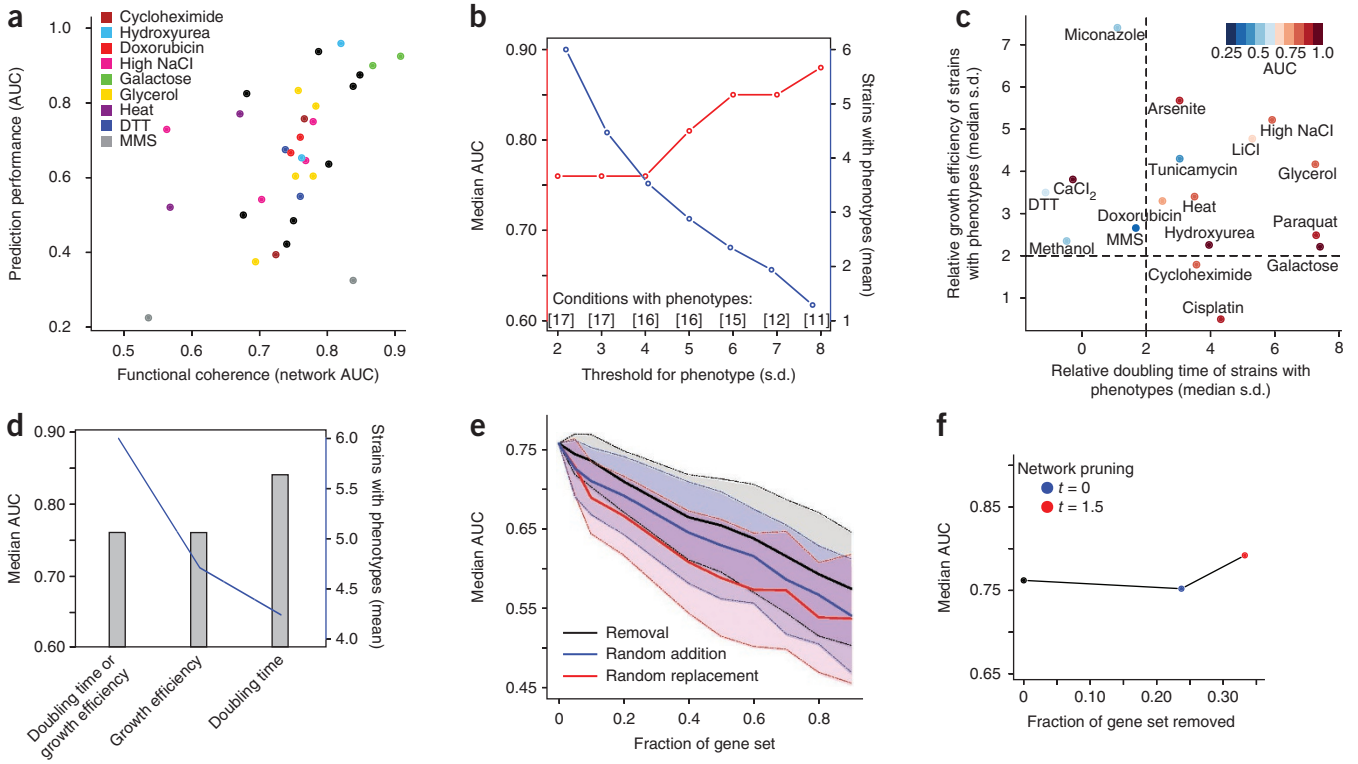

Figure 3 Influences on the classification performance of our predictive test score. (**a**) Performance, as measured by ROC AUC, of our prediction scores for a given condition and gene set correlates with the functional relationships within that gene set, as evaluated by examining the interactions in the YeastNet version 2 functional network[17,27] (Pearson's correlation = 0.5 between network AUC and ROC AUC from prediction, $P = 0.0042$). Conditions that matched several systematic gene deletion screens are color coded, and all gene sets are shown. (**b**) A higher threshold used to define strains with phenotypes improved predictive performance but reduced the number of strains with a phenotype and the number of conditions considered (conditions dropped out if no strain was considered to have a phenotype). The median prediction AUC is shown for each threshold between 2 and 8 s.d. (**c**) Median growth defects of strains with phenotypes for the different conditions. Conditions under which strains had larger T_d defects were better predicted. Predictions were worse for conditions in which most strains showed only GE phenotypes. The thresholds to define strains with phenotypes are indicated by the dotted lines. (**d**) T_d phenotypes were better predicted than GE phenotypes or phenotypes defined by either a T_d or GE defect. The mean number of strains with a phenotype per condition is also shown. The T_d bar represents 16 conditions, as high $CaCl_2$ concentration caused no T_d phenotypes. (**e**) Adding false positive genes to the gene sets had a more severe effect on predictive performance than removing genes (introducing false negatives). The lines indicate the mean overall AUC over 100 simulations, and the shaded areas show 1 s.d. The effect of randomly replacing genes (introducing false positives and false negatives) is also shown. (**f**) Through network-guided pruning, a large fraction of genes can be removed from each gene set without reducing predictive performance. Genes were removed if they did not interact (or interacted below a log-likelihood score *t*) with any other genes in the set in the YeastNet version 2 network. For $t = 1.5$, the result is significantly better than random removal ($P < 0.01$, overall AUC, 1,000 simulations).

growth-rate phenotypes could be more accurately predicted than final yields (**Fig. 3c**,**d**), possibly because the growth rate better matched the phenotype evaluated in most screens using the gene deletion collection. Indeed, considering growth rate alone, the median AUC increased to 0.84 at a 2-s.d. threshold (**Fig. 3d**).

To study the influence of false positive and false negative genes in a gene set assembled using a reverse genetic screen, we randomly removed or added genes to each set. The randomly added genes are not likely to be directly connected with the given phenotype and are therefore considered as false positives. Removing genes from the set involves genes that are likely to be relevant to the phenotype and therefore provides a simulation of relevant genes that were not retrieved by the screen or false negatives. Performance of our prediction score remained relatively robust and dropped off gradually when genes were randomly removed from each gene set (**Fig. 3e**); removing 10% of genes reduced the median AUC to 0.74, and removing 50% reduced it to 0.65. Only when ~70% of genes were removed were predictions no longer significant at the 1% level. In contrast, adding false positives to each of our gene sets had a stronger impact on performance (**Fig. 3e**), although including 10% false positives only reduced the median AUC to 0.71. These findings suggest that the defined gene sets that we were using include genes that do not substantially contribute to our phenotypic predictions. An integrated gene network provides one method to identify potential false positive genes in a set: genes without predicted functional connections to the other genes in a set may be considered less likely to represent genuine contributors to a phenotype. Indeed, removing genes that were unconnected (or weakly connected) within the network of each gene set substantially reduced the size of each gene set without affecting the overall performance of our prediction method (**Fig. 3f**). This approach of 'network-guided pruning' illustrates how background information on gene function can be used to refine the set of genes associated with a trait.

To further evaluate how variation within individual genes contributes to predictions, we measured covariance to quantify the agreement between the overall *S* score for the strains and the score of a single gene. To compare across conditions, we divided the covariance by the variance of the overall score (**Supplementary Table 8**). Under some conditions, a few genes were seen to have a larger effect on our prediction score, whereas for other conditions, a more even distribution of covariance scores was observed (**Fig. 4a**). To quantify the number of genes contributing to our prediction score 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. We determined the number of genes required at different cutoffs (**Fig. 4b**) and the fraction of the gene set needed to reach the cutoffs (**Fig. 4c**) (cumulative curves for all conditions are provided in **Supplementary Fig. 4**). Overall, the number of genes used to make predictions varied widely across conditions. For example, for growth with galactose, two genes were needed to reach 50% of the variance: *GAL3*, which had a stop codon in four strains, and *GAL2*, which had a nonsynonymous nucleotide transition (encoding p.Gly90Ser) in the W303 strain. To reach more than 90% of the variance, it was also necessary to consider *GAL4*, which had predicted deleterious mutations (causing p.Lys879Glu and p.Gly854Arg alterations) in two strains. In contrast, 59 of 374 genes were needed to reach 50% of the variance for strains growing in glycerol. The complexity that underlies predictions is therefore quite different across phenotypes, with between 1 and 59 genes required to reach 50% of the variance.

In summary, we have demonstrated here that it is possible to make accurate predictions about the phenotype of a *S. cerevisiae* strain by considering a set of genes relevant for that phenotype, as determined using data from previous reverse genetics screens, and by predicting the impact of genetic variations in the relevant genes on protein function. In this study, we considered only mutations in protein-coding regions and those predicted to cause loss-of-function alterations. Variation within regulatory regions and gain-of-function mutations are also expected to contribute to differences in phenotype, and incorporating the analysis of these into our approach could further improve predictions. However, a preliminary analysis suggests that a more comprehensive annotation of regulatory regions than that currently available for *S. cerevisiae* will likely be required for this purpose[19] (**Supplementary Note**). Further improvements in prediction could also derive from deeper sequencing and improved assembly, especially for the detection of insertions and deletions. Our analysis,

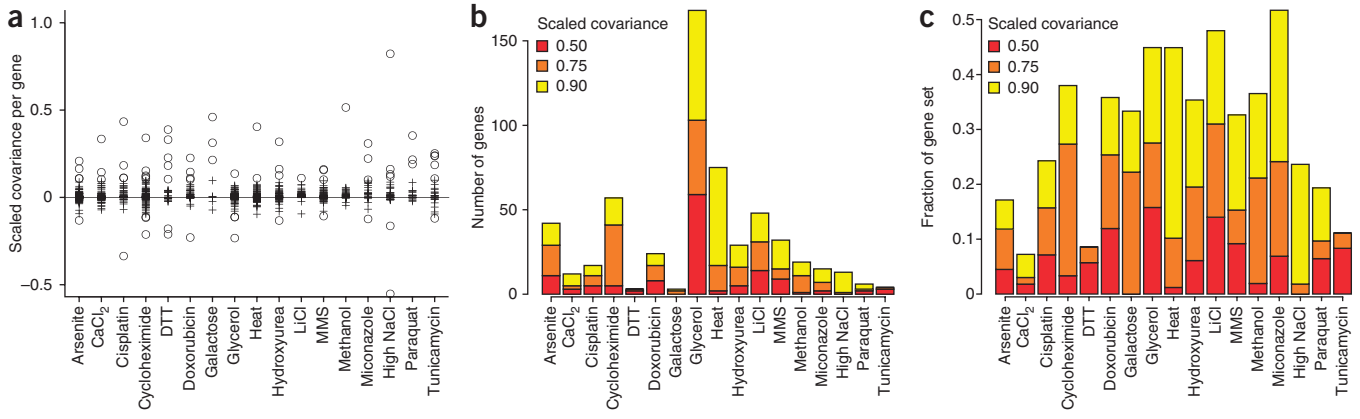

Figure 4 Per-gene contributions to overall *S* score variation across strains for each condition. For each condition, the covariance of the score per gene and *S* was divided by the variance of the *S* score as a proxy for the influence of each individual gene on the differences among strains. (**a**) Scaled covariance score per gene for each condition. Under some conditions, a few genes contributed considerably to the score differences between strains, whereas in other cases, individual genes contributed very little. In some cases, notably with high NaCl concentration, there were genes with negative covariance that were anti-correlated with the eventual scores per strain. Genes with a scaled covariance between –0.1 and 0.1 are indicated by crosses and other genes by open circles. (**b**) Evaluation of the number of genes required to explain the score differences between strains. The number of genes required to reach 50%, 75% and 90% of *S* variance is shown. Genes were added to the subset in the order of their absolute scaled covariance. Under some conditions, few genes were needed to reach 50% of the variance, although a large number of genes contributed to the overall scores. (**c**) The fraction of the total gene set for each condition required to achieve the three covariance levels.

using a network that reflects functional relationships between *S. cerevisiae* genes, showed that many of the gene sets retrieved from the SGD database are likely to be incomplete and to contain false positives, and additional reverse genetic screens or new methods to refine these gene sets would therefore be informative. Based on the analysis of systematic genetic interaction screens[20,21] and a few examples where interactions have been shown between quantitative trait loci[22,23], we suspect that considering nonadditive epistatic interactions will also be important for improving phenotypic predictions (**Supplementary Note**).

Importantly, genome-wide reverse genetic predictions in model organisms can be combined with extensive independent experimental validation. We therefore propose that further improvements in predictive performance may be best achieved through a competitive effort involving rounds of prediction and experimental evaluation by multiple groups, as is common in other fields of computational biology[24–26]. The challenge of making genome-wide reverse genetic predictions in model organisms should result in a deeper understanding of how to evaluate the effects of thousands of sequence variations on the phenotypes of an individual.

URLs. The analysis tool and our data sets are available at http://www.crg.eu/ben_lehner/datasets/.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturegenetics/.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

This work was funded by grants from the European Research Council (ERC), the Spanish Ministry of Science and Innovation (MCINN grant BFU2008-00365), the Catalan Agency for Management of University and Research Grants (AGAUR), ERASysBio+, the EMBO Young Investigator Program and the EMBL-CRG Systems Biology Program. R.J. was supported by a Juan de la Cierva Fellowship.

AUTHOR CONTRIBUTIONS

B.L. and R.J. designed the study, evaluated the results and wrote the manuscript. R.J. performed the analyses. J.I.S., R.G.-V. and R.J. designed and performed the growth experiments and sequence validation.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/naturegenetics/.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Bentley, D.R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53–59 (2008).
2. Kim, J.-I. *et al.* A highly annotated whole-genome sequence of a Korean individual. *Nature* **460**, 1011–1015 (2009).
3. Levy, S. *et al.* The diploid genome sequence of an individual human. *PLoS Biol.* **5**, e254 (2007).
4. Wheeler, D.A. *et al.* The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452**, 872–876 (2008).
5. Liti, G. *et al.* Population genomics of domestic and wild yeasts. *Nature* **458**, 337–341 (2009).
6. 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
7. Ng, P.C., Murray, S.S., Levy, S. & Venter, J.C. An agenda for personalized medicine. *Nature* **461**, 724–726 (2009).
8. Hillenmeyer, M.E. *et al.* The chemical genomic portrait of yeast: uncovering a phenotype for all genes. *Science* **320**, 362–365 (2008).
9. Chun, S. & Fay, J.C. Identification of deleterious mutations within three human genomes. *Genome Res.* **19**, 1553–1561 (2009).
10. Ng, P.C. & Henikoff, S. Predicting the effects of amino acid substitutions on protein function. *Annu. Rev. Genomics Hum. Genet.* **7**, 61–80 (2006).
11. Bromberg, Y., Yachdav, G. & Rost, B. SNAP predicts effect of mutations on protein function. *Bioinformatics* **24**, 2397–2398 (2008).
12. Ng, P.C. & Henikoff, S. Predicting deleterious amino acid substitutions. *Genome Res.* **11**, 863–874 (2001).
13. Giaever, G. *et al.* Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature* **418**, 387–391 (2002).
14. Engel, S.R. *et al.* *Saccharomyces* genome database provides mutant phenotype data. *Nucleic Acids Res.* **38**, D433–D436 (2010).
15. Hanley, J.A. & McNeil, B.J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **143**, 29–36 (1982).
16. Baryshnikova, A. *et al.* Quantitative analysis of fitness and genetic interactions in yeast on a genome scale. *Nat. Methods* **7**, 1017–1024 (2010).
17. Lee, I., Li, Z. & Marcotte, E.M. An improved, bias-reduced probabilistic functional gene network of baker's yeast, *Saccharomyces cerevisiae*. *PLoS ONE* **2**, e988 (2007).
18. Lee, I., Date, S.V., Adai, A.T. & Marcotte, E.M. A probabilistic functional network of yeast genes. *Science* **306**, 1555–1558 (2004).
19. Francesconi, M., Jelier, R. & Lehner, B. Integrated genome-scale prediction of detrimental mutations in transcription networks. *PLoS Genet.* **7**, e1002077 (2011).
20. Costanzo, M. *et al.* The genetic landscape of a cell. *Science* **327**, 425–431 (2010).
21. 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).
22. Gerke, J., Lorenz, K. & Cohen, B. Genetic interactions between transcription factors cause natural variation in yeast. *Science* **323**, 498–501 (2009).
23. Dowell, R.D. *et al.* Genotype to phenotype: a complex problem. *Science* **328**, 469 (2010).
24. Moult, J., Fidelis, K., Kryshtafovych, A., Rost, B. & Tramontano, A. Critical assessment of methods of protein structure prediction—Round VIII. *Proteins* **77** (suppl. 9), 1–4 (2009).
25. Leitner, F. *et al.* An overview of biocreative II.5. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **7**, 385–399 (2010).
26. Peña-Castillo, L. *et al.* A critical assessment of *Mus musculus* gene function prediction using integrated genomic evidence. *Genome Biol.* **9** (suppl. 1), S2 (2008).

ONLINE METHODS

Genome sequences. Genome sequences of 38 yeast strains were downloaded from the *Saccharomyces* Genome Resequencing Project[5]. Only the 19 strains with sequence coverage >75% and an imputed Phred quality score of Q40 were used in the analysis. Only SNPs and indels with an error rate of <0.001 were considered. The genomes were assembled relative to S288c, the strain used for the original sequencing project, such that observed variations are relative to this reference strain[5]. As the S288c strain was resequenced in this project, we were left with 18 strains for our analysis.

Construction of the SNP test set for *S. cerevisiae*. To evaluate the effects of sequence variations on protein function specifically in yeast, we constructed a data set composed of variations with characterized effects on protein function. The data set was created using the Protein Mutation Database (release March 2007)[28], SGD (March 2009)[14] and UniProt (March 2009)[29] databases. Retrieved variants were manually categorized as function changing or neutral according to the description of their effects. A total of 2,812 mutations were reported to be detrimental, compared to only 604 tolerated alterations. To increase the number of alterations without a functional effect, we added all variants in essential genes identified in our sequencing data. Because of the strong selective pressure on essential genes, we considered variation in essential genes to be neutral or close to neutral. The final test set contained 5,269 neutral variants. Excluding the variants within essential genes from the test set did not change the observed trends for the SIFT analysis. The compiled variants are provided in **Supplementary Table 2**.

Nonsynonymous SNPs. SNPs resulting in amino acid substitutions were evaluated with an adapted version of the SIFT algorithm[12]. SIFT evaluates a multiple-sequence alignment of homologous sequences and assesses the impact of amino acid substitutions by taking into consideration both the original and altered amino acid for the given residue in the homologs. As SIFT depends on multiple-sequence alignment, if a protein (or segment of a protein) is not covered by an alignment, the associated substitutions are not analyzed. We boosted SIFT performance on the test set by improving the input multiple-sequence alignments. This was achieved by retrieving previously identified orthologs of the *S. cerevisiae* genes from 17 fungal genomes[30] and by aligning them using T-Coffee[31] in the "accurate" mode.

The resulting alignments were used as input for a single PSI-BLAST run on the NCBI nonredundant database (downloaded February 2009) to retrieve more homologous sequences. We estimated the likelihood of a substitution with a functional effect based on the test set. We mapped the SIFT scores to a non-damaging rate, *P*(neutral), through a linear fit of the relationship between the –log-transformed SIFT scores and the proportion of substitutions with a phenotype in the test set. If a substitution did not retrieve a score, it was given a *P*(neutral) corresponding to the highest SIFT score.

Introduced stop codons. Premature stop codons occurred rarely, with only 112 instances identified. They showed a strong bias for the coding regions on the edges of genes. Of the nonsense mutations, 35% were located within nucleotides encoding the first or last 16 amino acids of the protein compared to the 5% of synonymous variants located in the same regions, which indicates a sevenfold over-representation. This suggests that the majority of these stop codons on gene edges are not damaging, and we set their non-damaging rate to 0.95 when within the region encoding the first or last 16 amino acids and to 0.01 otherwise. This choice, however, had little influence on predictions.

Insertions and deletions. To identify the indels most likely to have a functional effect, we studied their distribution. Genes with extreme numbers of indels (>20) were excluded (this removed 1,705 of the 4,329 indels). We used the program Repseek[32] to identify repeats within the genes. Indels are overrepresented in these repeat regions by more than 100-fold, and the predicted indels might be false. To obtain an indication of how often indels are not damaging or erroneous, we assumed that indels within essential genes have neither of these characteristics. We calculated the occurrence rate of indels (counted as units) per base and took the ratio of occurrences in essential and nonessential genes as an indication of the non-damaging rate, *P*(neutral), which is 0.87 in general, 1 in repeat regions (where all indels are non-damaging) and 0.64 in nonrepeat regions. The indels outside of repeat regions were divided into subclasses and, using the same logic as above, the non-damaging rate was estimated. Indels up to 15 bp in size causing frameshifts (58% of the total) had a non-damaging rate of 0.41, and in-frame indels had a rate of 0.6. Mid-sized indels (16–99 bp) had an estimated non-damaging rate of 0.49, and for large indels (>99 bp) the ratio of occurrence in essential to nonessential genes indicated that almost all indels were non-damaging.

Probability of affected function per gene. We defined the probability of a perturbed or altered function (AF) for a gene as a simple combination for all the variations in the gene (*k*) of the estimated probability that a variation does not cause a functional effect, *P*(neutral)

$$P(\mathrm{AF}) = 1 - \prod_{i=1}^{k} P_i(\mathrm{neutral})$$

Gene sets. To retrieve gene sets from genome-wide gene deletion screens, we used the SGD database (downloaded in September 2010)[14]. We filtered from the gene sets any gene annotated as dubious, silenced, merged or deleted. To avoid the inclusion of overly broad or incomplete gene sets, we only included gene sets with more than 5 genes but fewer than 500.

Calculating score per strain. A score for a strain *h* was based on the set of genes (of size *l*) selected as relevant for growth in the selected stress condition *i* as determined by a screen of the systematic gene deletion collection. For a given condition and gene set, the score *S* is given by

$$S_{h,i} = \sum_{j=1}^{l} \frac{1}{E_h} \cdot \log(1 - P_{h,j}(\mathrm{AF}))$$

which is analogous to combining the scores per gene. The value serves to correct for the evolutionary distances between strains. The evolutionary distance between strains correlates with a higher estimated rate at which gene function has altered, even though natural selection should prevent the actual rate from being too high. The expected score for a strain *h* over all *n* genes was calculated by

$$E_h = \frac{1}{n} \sum_{j=1}^{n} \log(1 - P_{h,j}(\mathrm{AF}))$$

Essential genes were not considered, as they were excluded from the systematic gene deletion screens.

Allele frequency of variations. Some of the variations occurred in many of the strains. If a variation is common it is less likely to be detrimental, as its spread should have been countered by natural selection. Also, in our case, if a variation is frequent, it might be that it is a variation specific to the reference strain. We chose to ignore any alleles that occurred in more than 80% of the considered strains. The effect of this variable on performance is shown in **Supplementary Figure 5**.

Growth experiments. Before the growth experiments, strains were grown in two consecutive pre-growth cultivations in synthetic complete medium (0.7% (w/v) yeast nitrogen base, 0.1% monosodium glutamic acid, 1% succinic acid, 2% glucose and 0.077% Complete Supplement Mixture (ForMedia), pH 5.8) at 30 °C. Growth experiments were performed in a 96-well Nunclon flat-bottom microtiter plate[33], and cells were incubated at 30 °C in a TECAN Infinite 200 plate reader with 120 µl of synthetic complete medium per well. An optical density measurement at 600 nm was made every 10 min to follow cell growth. The plates were shaken linearly every other minute for a 1-min interval. The start OD_{600} was in the linear range (~0.15). Measurements were taken for 48 h if the stationary phase had already been reached by this point or for 72 h in all other cases. Two strains, YJM789 and RM11-1a, were not available through the National Center of Yeast Cultures (NCYC), and two strains, UWOPS05.217.3 and UWOPS05.227.2, showed severe flocculation in our assays and were not included. The remaining 15 strains were grown in duplicate with two conditions

tested per plate; the outer row of wells was filled with sterile medium to minimize variation caused by evaporation. To obtain T_d, we determined the maximal slope for a linear fit over a 5-h period for data that were transformed as previously described[33]. GE was defined as the maximum OD measurement. T_d and GE were normalized to the growth of the S288c strain by determining the logarithmic strain coefficient (LSC)[5] with the equation

$$\mathrm{LSC} = \frac{1}{n}\sum_{i}^{n}\sum_{j}^{m}\log\left(\frac{\mathrm{reference}_i}{\mathrm{strain}_j}\right)$$

with n repeats of S288c and m repeats of a given strain in a certain condition. The LSC scores were corrected with values obtained under optimal conditions (2% glucose) and were based on at least two separate experiments. For growth on glycerol and galactose, the S288c strain showed impaired growth due to nonfunctional *HAP1* and *GAL2*, respectively[34]. Growth under these conditions was compared to that of the YJM978 strain, which grows well in these conditions, similarly to S288c in normal conditions.

Scoring phenotypes. We evaluated the performance of our predictions by sorting the strains according to their scores per phenotype. For every phenotype, the ranking was evaluated by taking the strains with impaired growth as positives and calculating the AUC for the ROC curve as a performance measure[15]. To produce a single overall score, the rank ordered lists of every phenotype were merged to form a single list for which an AUC could then be calculated. A P value for the AUC was estimated based on the relationship between the AUC and the Wilcoxon test statistic. The P value was confirmed by randomization experiments in which we calculated the overall AUC for the same conditions and strains.

Randomization experiments. The significance of the overall AUC was confirmed by three randomizations that took into account any remaining effects of sequence divergence between strains as well as correlations between strains and gene sets. These experiments included a bootstrap of the gene sets (gene sets were matched with conditions through sampling with replacement from the set of gene sets used for the final result), randomizations of the gene set contents while maintaining the size of the gene set (the content of the matched gene sets was replaced by randomly selecting genes from the set of nonessential genes, excluding dubious ORFs) and a bootstrap of the strains (the identity of strains was randomized through a sampling with replacement from the set of strains). For each of these strategies, 10,000 randomizations were performed.

Gene set evaluation using the YeastNet functional network. The functional coherence of gene sets was assessed using YeastNet version 2, which connects genes by a likelihood score of the probability that they act in a common biological process[17]. The online phenotype prediction interface was used to retrieve an AUC that reflected how strongly genes within a gene set share functional connections in comparison to the remaining genes. When pruning genes from a gene set, we removed all genes that did not share a functional edge in the network with any other gene in the gene set. We applied a minimum threshold to the confidence of a functional link (the log-likelihood score provided by YeastNet) for more stringent pruning.

27. McGary, K.L., Lee, I. & Marcotte, E.M. Broad network-based predictability of *Saccharomyces cerevisiae* gene loss-of-function phenotypes. *Genome Biol.* **8**, R258 (2007).
28. Kawabata, T., Ota, M. & Nishikawa, K. The protein mutant database. *Nucleic Acids Res.* **27**, 355–357 (1999).
29. UniProt Consortium. The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Res.* **38**, D142–D148 (2010).
30. Wapinski, I., Pfeffer, A., Friedman, N. & Regev, A. Natural history and evolutionary principles of gene duplication in fungi. *Nature* **449**, 54–61 (2007).
31. Notredame, C., Higgins, D.G. & Heringa, J. T-Coffee: a novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.* **302**, 205–217 (2000).
32. Achaz, G., Boyer, F., Rocha, E.P.C., Viari, A. & Coissac, E. Repseek, a tool to retrieve approximate repeats from large DNA sequences. *Bioinformatics* **23**, 119–121 (2007).
33. Toussaint, M. & Conconi, A. High-throughput and sensitive assay to measure yeast cell growth: a bench protocol for testing genotoxic agents. *Nat. Protoc.* **1**, 1922–1928 (2006).
34. Mortimer, R.K. & Johnston, J.R. Genealogy of principal strains of the yeast genetic stock center. *Genetics* **113**, 35–43 (1986).

doi:10.1038/ng.1007