A consensus *S. cerevisiae* metabolic model Yeast8 and its ecosystem for comprehensively probing cellular metabolism

Hongzhong Lu1,5, Feiran Li1,5, Benjamín J. Sánchez1, Zhengming Zhu1,2, Gang Li1, Iván Domenzain1, Simonas Marcišauskas1, Petre Mihail Anton1, Dimitra Lappa1, Christian Lieven3, Moritz Emanuel Beber3, Nikolaus Sonnenschein3, Eduard J. Kerkhoven1 & Jens Nielsen1,3,4

Genome-scale metabolic models (GEMs) represent extensive knowledgebases that provide a platform for model simulations and integrative analysis of omics data. This study introduces Yeast8 and an associated ecosystem of models that represent a comprehensive computational resource for performing simulations of the metabolism of *Saccharomyces cerevisiae*—an important model organism and widely used cell-factory. Yeast8 tracks community development with version control, setting a standard for how GEMs can be continuously updated in a simple and reproducible way. We use Yeast8 to develop the derived models panYeast8 and coreYeast8, which in turn enable the reconstruction of GEMs for 1,011 different yeast strains. Through integration with enzyme constraints (ecYeast8) and protein 3D structures (proYeast8DB), Yeast8 further facilitates the exploration of yeast metabolism at a multi-scale level, enabling prediction of how single nucleotide variations translate to phenotypic traits.

---

1 Department of Biology and Biological Engineering, Chalmers University of Technology, Kemivägen 10, SE-412 96 Gothenburg, Sweden. 2 School of Biotechnology, Jiangnan University, 1800 Lihu Road, 214122 Wuxi, Jiangsu, China. 3 The Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, DK-2800 Kgs Lyngby, Denmark. 4 BiInnovation Institute, Ole Maaløes Vej 3, DK2200 Copenhagen N, Denmark. 5 These authors contributed equally: Hongzhong Lu, Feiran Li. Correspondence and requests for materials should be addressed to J.N. (email: nielsenj@chalmers.se)

There are amendments to this paper
In the era of big data, computational models are instrumental for turning different sources of data into valuable knowledge for e.g. biomedical\textsuperscript{1,2} or industrial use\textsuperscript{3}. As a bottom-up systems biology tool, genome scale metabolic models (GEMs) connect genes, proteins and reactions, enabling metabolic and phenotypic predictions based on specified constraints\textsuperscript{4,5}. The quality and scope of GEMs have improved as demonstrated by the models for human\textsuperscript{6}, yeast\textsuperscript{7}, and E. coli\textsuperscript{8}. With their thousands of reactions, genes and proteins, GEMs also represent valuable organism-specific databases. Therefore, it is important to keep track of changes as new knowledge is added to GEMs, in order to make model development recorded, repeatable, free and open to the community. Approaches to record updates of a GEM exist for this purpose\textsuperscript{9,10}, albeit with some limitations on their simplicity and flexibility.

\textit{S. cerevisiae} is a widely used cell factory\textsuperscript{11,12} and is extensively used as a model organism in basic biological and medical research\textsuperscript{13,14}. Recently, the emergence of technologies, such as CRISPR\textsuperscript{15} and single cell omics data generation\textsuperscript{16}, have accelerated the development in systems biology. Consistent with strong research interests in yeast, the relevant GEMs have also undergone numerous rounds of curation since the first published version in 2003\textsuperscript{17}. These GEMs have contributed significantly to systems biology studies of yeast including their use as platforms for multi-omics integration\textsuperscript{18,19}, and use for in silico strain design\textsuperscript{20,21}. However, the hitherto latest version, Yeast8\textsuperscript{22}, with only 909 genes, falling behind the latest genome annotation, presents a bottleneck for the use of yeast GEMs as a scaffold for integrating omics datasets.

Metabolism is complex and regulated at several different levels\textsuperscript{23,24}. Traditional GEMs only consist of reactions and their related gene and protein identifiers, and therefore cannot accurately predict cellular phenotypes under varied environmental conditions other than nutritional conditions, e.g. simulating the impact of temperature on growth rates\textsuperscript{25}. Recently, enzyme constraints\textsuperscript{26} and protein 3D structures\textsuperscript{27} have been integrated into GEMs, thereby expanding their scope of application and laying the foundation for whole cell modelling. GEMs constrained with \( k_{\text{cat}} \) values and enzyme abundances have been able to directly integrate proteomics data and correctly predict cellular phenotypes under conditions of stress\textsuperscript{26}. GEMs with protein 3D structures connect the structure-related parameters and genetic variation\textsuperscript{27} with cellular metabolism\textsuperscript{6}, thus enlarging the prediction scope of GEMs. However, it remains challenging to directly predict cellular metabolism based on changes in protein sequences with current GEMs. Advanced functional mutation cluster analysis constrained with protein 3D structures is therefore needed to integrate knowledge on protein structures into GEMs.

This study presents Yeast8, the latest release of the consensus GEM of \textit{S. cerevisiae}\textsuperscript{22,28–30}. We also introduce a model ecosystem around this GEM, including ecYeast8, a model incorporating enzyme constraints; panYeast8 and coreYeast8, representing the pan and core metabolic networks of 1011 \textit{S. cerevisiae} strains; and proYeast8\textsuperscript{30}, a database containing 3D structures of metabolic proteins. This model ecosystem has the ability to meet wide application demands from the large scientific yeast community in systems and synthetic biology of yeast.

Yeast8 is a consensus GEM maintained in an open and version-controlled way. Through ecYeast8 and proYeast8\textsuperscript{30}, multiple parameters related to protein kinetics and 3D structures could be integrated based on gene–protein–reaction relations. Furthermore, with panYeast8 and coreYeast8, 1011 strain-specific GEMs were reconstructed and compared. Thus, with Yeast8 and its model ecosystem, we demonstrate that the metabolism of yeast can be characterised and explored in a systematic way.

### Results

**Recording community developments of yeast GEMs with GitHub.** We devised a general pipeline to record updates to the model using Git (https://git-scm.com/), a version control system, and GitHub (https://github.com/), a hosting service for Git repositories (Fig. 1a and Supplementary Fig. 1). Hereby, we record everything related to updates of the GEM, including datasets, scripts, corrections and each released version of the GEM (Supplementary Fig. 1). This Git version-controlled model enables open and parallel collaboration for a wide community of scientists. With Git and GitHub, each version of the yeast GEM can be released periodically, which helps to promote the simultaneous development of a model ecosystem around yeast GEM (Fig. 1b).

**Increasing the scope of the yeast metabolic network.** We systematically improved the yeast GEM while moving from Yeast7 to Yeast8 through several rounds of updates (Fig. 1c and Supplementary Fig. 2). To improve the genome coverage, we added additional genes from i5e926\textsuperscript{31}. Besides, all functional gene annotations of \textit{S. cerevisiae} from SGD\textsuperscript{32}, BioCyc\textsuperscript{33}, Reactome\textsuperscript{34}, KEGG\textsuperscript{35} and UniProt\textsuperscript{36} were collected and compared (Supplementary Fig. 3) to update gene–protein–reaction relations (GPRs), as well as adding more GPRs. With Biolog experiments, i.e. evaluation of growth on a range of different carbon and nitrogen sources (Supplementary Data 1), and metabolomics mapping (see methods), extra reactions were added to enable the model to ensure growth on the related substrates, as well as connecting those metabolites with high confidence with the GEM. The biomass equation was modified by adding nine trace metal ions and eight cofactors. Additionally, 37 transport reactions were added in order to eliminate 45 dead-end metabolites. To improve lipid constraints, we reformulated reactions of lipid metabolism using the SLiMER formalism, which Splits Lipids Into Measurable Entities\textsuperscript{37}. As SLiMER imposes additional constraints on both the lipid classes and the acyl chain distribution from metabolomics data, it improved the model performances in lipid metabolism.

In each round of model updates, standard quality-control tests, such as reaction mass balance check and ATP yield analysis, were performed. The results in Supplementary Fig. 2 and Supplementary Fig. 4 indicate that the gene (reaction) coverage in the model and its performance were improved during the iterative update process, which was also shown by comparing Yeast8 to Yeast7 (Fig. 1d–f). To facilitate the multi-omics integrative analysis and visualisation, we established a map of yeast metabolic pathways in SBGN (System Biology Graphical Notation) format (Supplementary Fig. 5) using CellDesigner\textsuperscript{38}.

**Expanding Yeast8 to enable enzyme constraints.** To enhance model prediction capabilities of Yeast8, ecYeast8 was generated by accounting for enzyme constraints using the GECKO framework\textsuperscript{26} (Fig. 1b, Supplementary Table 1). In enzyme-constrained models, metabolic rates are constrained by the intracellular concentration of the corresponding enzyme multiplied by its turnover number (\( k_{\text{cat}} \) value) to ensure that fluxes are kept at physiologically possible levels. Compared with Yeast8, ecYeast8 has a large reduction in flux variability for most of the metabolic reactions in both glucose-limited and glucose-excess conditions (Fig. 2a).

To further evaluate the predictive strength of ecYeast8, we compared the predicted maximum growth rate to that obtained from experiments in which we cultivated yeast (strain S288c) on 322 different combinations of carbon and nitrogen sources using microtiter plates (Fig. 2c). Here, the measured maximum specific growth rate for each carbon source under all nitrogen sources was...
scaled based on the growth on ammonia as a reference, with results displaying that the mean error for the predicted and scaled measured cell growth rate in ecYeast8 was 41.9%, which is significantly lower than using Yeast8 and traditional flux balance analysis alone (Fig. 2d–f). With ecYeast8, it is also possible to estimate flux control coefficients (FCC) of enzymes for growth on different carbon sources (Fig. 2b), i.e. enzymes exerting the majority of flux control on growth. For instance, growing on glucose, the major flux controlling enzyme is one of the isoforms of glyceraldehyde-3-phosphate dehydrogenase (Tdh1) whereas on fermentative carbon sources, such as ethanol and acetate, the flux controlling enzyme is one of the isoforms of glyceraldehyde-3-phosphate dehydrogenase (Tdh1) whereas on fermentative carbon sources, such as ethanol and acetate.

**Generation of panYeast8 and coreYeast8.** To investigate the correlation between phenotype and genotype, we used genomics data of 1011 S. cerevisiae strains from a recent genome-sequencing project and designed a pipeline to reconstruct GEMs for each strain (Supplementary Fig. 6a)40. Similar to the CarveMe method41, a pan model for all yeast strains was firstly reconstructed based on Yeast8 and the comprehensive pangenome annotation (Supplementary Note 1). Using panYeast8 and the gene presence matrix for the 1,011 strains42, strain-specific coreGEMs (ssGEMs) could be generated automatically (Supplementary Fig. 7)43. While coreYeast8 re

---

**Fig. 1 Framework of the yeast GEM project.** a Recording model updates in a community way using GitHub. b In Yeast8, genes, metabolites and reactions are annotated with their corresponding IDs from different databases, which simplifies translation between namespaces. Yeast8 forms the basis of the model ecosystem from which proYeast8, ecYeast8, etc., are derived. c Major steps of development from Yeast7 to Yeast8. d Subsystem statistical analysis for the reactions added to Yeast8. e Metabolomics mapping between Yeast7, Yeast8 and the YMDB database. f Comparison of Yeast7 and Yeast8 in percent accuracy of gene essentiality and substrate usage analysis, as well as in memote test total scores (divided by 100).
be seen as a benchmark for reconstructing GEMs for other yeast strains. Similar to the core E. coli GEM, we observed that coreYeast8 could not grow in silico in a minimal medium since it lacks the ability to synthesise all required precursors for biomass production (Supplementary Note 1, Supplementary Table 2). This can be partly related to the definition of 'core biomass compositions' in our model, which still hints that the ability of particular yeast strains to grow on a minimal medium seems to be an acquired phenotype.

Evaluation of 1011 strain-specific GEMs. Using the 1011 GEMs, we estimated in silico substrate usage (including carbon, nitrogen and phosphorus) as well as the yield of 26 metabolites and biomass for growth on minimal media (Fig. 3b–f). From the substrate usage analysis, we found that some of the domesticated strains from the ecological origin 'Industrial' had a smaller range of substrates usage, suggesting that the domestication process may have resulted in the loss of functions that were not routinely used. As for the yield of biomass, it could be observed that the yield varied greatly between strains from different ecological origins. By comparison, the biomass yield was relatively lower for strains with the ecological origin 'Human', i.e. strains that have been isolated from humans (Fig. 3b), likely due to an adaptation towards growth on complex media, as many building blocks are provided by the host and the yeasts have adapted to these conditions.

To further compare the metabolic potential of different strains, we calculated the maximum yields of 20 amino acids and six key precursors in a minimal medium from 1011 ssGEMs (Fig. 3f) and the results illustrated that the gene background can have a remarkable effect on the ranges of maximum yields of desired products. Additionally, taken strains from 'Industrial' ecological origin as an example, a considerable variation in the strains’
Fig. 3 Conservation and diversity analysis of yeast metabolism through reconstruction of 1,011 strain-specific GEMs (ssGEM). a Amount of reactions of ssGEMs for strains from different ecological origins. The top column graph describes the strain number from each ecological origin. Different strains from one ecological origin could have the equal number of reactions. In the box plots of this work, if not mentioned, the bold line in the box represents the median. The lower and upper bounds of box show the first and third quartiles, respectively, and whiskers indicate ±1.5× the interquartile range (IQR). The color dots overlaid represent the corresponding data points. b Prediction of biomass yield using ssGEMs on minimal media with glucose as the carbon source. c Number of substrates which can be used in silico for strains from different ecological origins. In the simulation using ssGEMs, 58 carbon sources, 46 nitrogen sources, 41 phosphate sources and 12 sulphate sources were used respectively in the minimal media. d Decision tree classification of strains according to the in silico maximum growth rate on different carbon sources (S5: Lactic acid, S11: Serine, S17: Sorbitol, S49: Trehalose). The in silico uptake rate for each carbon source was set at 10 mmol (g Biomass)−1 h−1. e Comparison of the yield of 20 amino acids from glucose for 30 strains from the group of “industrial strain”. f Comparison of in silico range of maximum yields of 20 amino acids and six key chemicals for all 1,011 ssGEMs.

capability to synthesise these amino acids was found (Fig. 3e). Combining genotype information, model simulation and phenotype data revealed that variations in the simulated maximum yields of amino acids are primarily due to differences in the energy pathways used for ATP regeneration and the absence of some essential genes in amino acid synthesis (Supplementary Note 1). This shows that simulations with ssGEMs can be instrumental to evaluate the potential of a strain in producing specific chemicals, as well as enable analysis of the relation between genotype and phenotype.

Next, we classified the strains based on the genes and reactions contained in the 1,011 ssGEMs, as well as the in silico substrate usage. Using PCA analysis we found that the strains could be subdivided into several distinct groups according to the existence of genes and reactions (Supplementary Fig. 7e, f). However, strains from different ecological origins can be clustered together, reflecting the high conservation in yeast metabolism. A decision tree algorithm was further employed to classify strains based on their substrate usage, while this initially indicated that strains from ‘Human’ can be separated from ‘Wine’ only based on the maximum growth rate on two substrates (lactic acid and sorbitol, Fig. 3d). When using hierarchical cluster analysis based on the reaction existence in ssGEMs for yeast strains from specific ecological origins, taking ‘Human’ as an example, these strains could be further divided into smaller groups based on the variable reactions shared between these strains (Supplementary Fig. 7g). The above analysis makes it clear that the model itself and the related simulation can classify different yeast strains, which in turn can be complementary to classical strain classifications based on single-nucleotide polymorphism (SNP) data.
To explore the potential functions of protein residue mutations, all the 16,258,509 homozygous SNPs from the 1011 S. cerevisiae genome-sequencing project were collected (Supporting Fig. 9a and Fig. 4a). Low quality PDB files were filtered out in order to only maintain high quality PDB files for our following analysis (Supplementary Note 2, Supplementary Fig. 9b–e).

To get further insight into which parts of metabolism may be sensitive or prone to mutations we chose the top 30 genes with the smallest and largest number of relative nsSNPs and conducted a GO-term analysis using DAVID 6.7 (Supplementary Table 3 and 4). Genes with the least amount of nsSNPs were enriched in thiamin metabolic process ($P = 0.0001$) and glycolysis ($P = 0.0003$), indicating that these pathways are conserved in the evolution of S. cerevisiae. By comparison, genes with the most nsSNPs were enriched in organic acid biosynthetic process ($P = 0.02$), carboxylic acid biosynthetic process ($P = 0.0002$), etc.

Following the findings mentioned above, we then mapped all nsSNPs onto the 3D protein structures of proYeast8DB using CLUMPS method, which could find proteins with significant SNP clusters within the protein structures. As an example, using the above nsSNP data as input, we ran the CLUMPS pipeline for three different groups of S. cerevisiae strains: Wild, Wine, and Bioethanol strains (Fig. 4b). With a cut-off $P$ value of 0.05, this transpires that genes catalyzing abundant enzymes or enzymes with many metabolic functions are likely to be more conserved in evolution. In addition, enzymes with high flux control coefficients related to the growth on different carbon sources (Fig. 2d) had fewer nsSNPs (Supplementary Fig. 11), which shows genes with a high impact on cell growth may be more conserved.

Characteristics of metabolism displayed with proYeast8DB. Phenotype is not only determined by absence or presence of genes, but also by SNPs. As SNPs occur randomly, we need a way to test which SNPs are significant, by looking at their distribution over the protein structure. To enable such analysis, we collected PDB files for all the metabolic proteins in Yeast8, generating the proYeast8DB database (Supplementary Fig. 9a and Fig. 4a). Low quality PDB files were filtered out in order to only maintain high quality PDB files for our following analysis (Supplementary Note 2, Supplementary Fig. 9b–e).

To explore the potential functions of protein residue mutations, all the 16,258,509 homozygous SNPs from the 1011 S. cerevisiae genome-sequencing project were collected (Supporting Fig. 9a and Fig. 4a). Low quality PDB files were filtered out in order to only maintain high quality PDB files for our following analysis (Supplementary Note 2, Supplementary Fig. 9b–e).

To get further insight into which parts of metabolism may be sensitive or prone to mutations we chose the top 30 genes with the smallest and largest number of relative nsSNPs and conducted a GO-term analysis using DAVID 6.7 (Supplementary Table 3 and 4). Genes with the least amount of nsSNPs were enriched in thiamin metabolic process ($P = 0.0001$) and glycolysis ($P = 0.0003$), indicating that these pathways are conserved in the evolution of S. cerevisiae. By comparison, genes with the most nsSNPs were enriched in organic acid biosynthetic process ($P = 0.02$), carboxylic acid biosynthetic process ($P = 0.002$), etc.

Following the findings mentioned above, we then mapped all nsSNPs onto the 3D protein structures of proYeast8DB using CLUMPS method, which could find proteins with significant SNP clusters within the protein structures. As an example, using the above nsSNP data as input, we ran the CLUMPS pipeline for three different groups of S. cerevisiae strains: Wild, Wine, and Bioethanol strains (Fig. 4b). With a cut-off $P$ value of 0.05, this transpires that genes catalyzing abundant enzymes or enzymes with many metabolic functions are likely to be more conserved in evolution. In addition, enzymes with high flux control coefficients related to the growth on different carbon sources (Fig. 2d) had fewer nsSNPs (Supplementary Fig. 11), which shows genes with a high impact on cell growth may be more conserved.

To get further insight into which parts of metabolism may be sensitive or prone to mutations we chose the top 30 genes with the smallest and largest number of relative nsSNPs and conducted a GO-term analysis using DAVID 6.7 (Supplementary Table 3 and 4). Genes with the least amount of nsSNPs were enriched in thiamin metabolic process ($P = 0.0001$) and glycolysis ($P = 0.0003$), indicating that these pathways are conserved in the evolution of S. cerevisiae. By comparison, genes with the most nsSNPs were enriched in organic acid biosynthetic process ($P = 0.02$), carboxylic acid biosynthetic process ($P = 0.002$), etc.

Following the findings mentioned above, we then mapped all nsSNPs onto the 3D protein structures of proYeast8DB using CLUMPS method, which could find proteins with significant SNP clusters within the protein structures. As an example, using the above nsSNP data as input, we ran the CLUMPS pipeline for three different groups of S. cerevisiae strains: Wild, Wine, and Bioethanol strains (Fig. 4b). With a cut-off $P$ value of 0.05,
proteins with significant mutation enrichment from each group of strains were identified (Fig. 4b). From the Wine group, we found GO terms like ‘electron transport chain’ (P value = 0.018) and ‘glutathione metabolic process’ (P value = 0.006, Supplementary Table 5), with glutathione metabolism being known to be distinct for wine yeasts, where it plays a role in anti-oxidation of various aromatic compounds.46,47

Interestingly, we discovered that the Bioethanol group contained four proteins (YML100W, YAR035W, YJL068C, YMR246W) belonging to the GO term ‘cellular response to heat’ (P value = 0.041, Supplementary Table 6), among which, YML100W is a large subunit of the trehalose 6-phosphate synthase/phosphatase complex, which is known to contribute to survival from acute heat stress. Two mutations were identified in the protein encoded by YML100W, both are located in a small cluster in the protein’s 3D structure, despite being separated by seven residues in the primary protein sequence (Fig. 4c). The annotation from the mutfunc database48 (http://mutfunc.com/) suggested that mutation at site of 422 in YML100W occurred in a conserved zone (Fig. 4c). We further wanted to know whether the mutations in these four proteins affected the phenotype of S. cerevisiae strains. Firstly, we classified all the strains based on residue mutations from these four proteins (Fig. 4e). It could be found that the relative growth rate at 42 °C was significantly higher for strains with more than six mutations in these four proteins than for strains that had less than two mutations. Similarly, the relative growth rates based on mutations only from YML100W were also analyzed, which indicated that the contribution of the two mutations from this single protein is smaller than over six mutations from all four proteins (Fig. 4f), further demonstrating that the ability to grow at elevated temperatures requires epistatic interactions of mutations in several genes (or proteins).

Systematic analysis of ecYeast8 and proYeast8DB. To further display the value of the yeast model ecosystem, we combined ecYeast8 and proYeast8DB (Fig. 5a) to identify genes and their mutations that are related to a specific phenotype. Based on growth rate data provided in Peter et al.40 (Fig. 5b), we selected 50 strains with the highest, medium, and lowest relative growth rate on complex medium with glycerol as the carbon source (growth with glucose was used as reference) (Fig. 5c). We then ran the hotspot analysis pipeline based on proYeast8DB to find hotspot zones of proteins for each group of strains. In parallel, we performed in silico flux control analysis to calculate FCCs for each protein by using ecYeast8 with growth as the objective function on the same medium. Hotspot analysis demonstrated that each group of strains has a set of unique hotspot zones in different proteins (Fig. 5d).

With the FCCs as filtering criteria, four proteins were identified with relatively high FCC (>0.01) from 50 strains with highest relative growth rate, among which YJL052W, one of the isoforms of glyceraldehyde-3-phosphate dehydrogenase (GAPDH), is located in the key pathway related to glycerol utilisation (Fig. 5f). When we listed all mutations of YJL052W, there exist six residue mutations at site of 31, 73, 24, 70, 125 and 248. Interestingly, we found four mutations at site of 24, 31, 70 and 73 were close to each other, far away from the other two mutations at site of 125 and 248 (Fig. 5e). The mutation at position of 31 and 73 were identified successfully using our hotspot analysis pipeline and the remaining two mutations at position of 24 and 70 could be filtered out according to the definition of important clusters in a hotspot zone, which should be made up of the significant pairs of two residues, separated by at least 20 amino acids in the original protein sequence.49 When the CLUMPS analysis was employed for the different combinations of mutations in the above site, the P value for mutations at site of 24, 31, 70 and 73 is 0.0247 and P value for mutations at site of 31 and 73 is 0.0195 (Supplementary Table 7). However, if all the six mutated positions are considered, the P value is 0.7817. Therefore, mutations at position of 24, 31, 70 and 73 as a whole could form into a larger hotspot zone. From the protein functional annotation46, we found that this hotspot zone was very close to the binding site of NAD+ at site of 33 and 78.

We queried whether there were correlations between mutations occurring in hotspot zones and the growth phenotype. From the remaining 746 strains which were not used for the above hotspot analysis, we found eight strains that exhibit more than one mutation in the hotspot zone and another 15 strains having two mutations in YJL052W, but in the non-hotspot zone. Comparing the relative growth rate of these two groups of strains with the remaining, we saw that the former 8 strains grows faster on complex medium with glycerol as the carbon source than both the latter 15 strains, as well as all other strains not containing the mutations in hotspot zone (Fig. 5g), indicating that the mutations that happened in the hotspot zone is possibly beneficial for cell growth. Therefore, such an comprehensive analysis with our model ecosystem can enable identification of the potential mutation targets related to a specific phenotype.

Discussion

As part of this study, we have developed Yeast8 aided by version control and open collaboration, which has provided a platform for a continued community-driven expansion of the model. This platform can greatly accelerate iterative updates of the model, and we believe that this approach should become the future standard for developing GEMs for other organisms. Yeast8 is the currently most comprehensive reconstruction of yeast metabolism, but it also represents a model that can be used for simulations. The platform provided through the GitHub repository enables addition of new knowledge when it is acquired as well as using this for further improving the model for simulations. Yeast8 is in line with the latest trend of performing model quality-control analysis in a standardised manner with memote50. Integrating consistent model evaluation with community model development will be instrumental to accelerate high quality development of GEMs.

Through developing Yeast8, we have significantly improved the metabolic scope of the consensus GEM of S. cerevisiae. As more evidences from experiments and bioinformatics analyses are revealed and utilised to update GEMs, these models will move closer to the in vivo network. Based on Yeast8, we developed strain specific GEMs, enzyme-constrained GEMs (ecYeast8), etc., which together form a model ecosystem around the yeast GEM and improve cellular phenotype predictions. As an example, ecYeast8 verifies that the yeast phenotypes are to a large extent determined by protein resources allocation, which is consistent with recent research39. We expect predictions of ecYeast8 to further improve as more organism-specific kinetic data becomes available, hopefully generated in a high-throughput and systematic way51. By comparing panYeast8, coreYeast8 and 1,011 strain specific models, it can be concluded that metabolic capabilities are largely conserved for all S. cerevisiae strains, which is consistent with a recent study52. However, through strain specific GEM simulations, we have found subtle metabolic differences among the strains in the utilisation of substrates and the maximum yield of 26 chemicals. Exploring these differences constrained with more physiological data can guide future metabolic engineering and help to evaluate the potential of any given strain for any desired product, as well as provide clues about the mechanisms of evolutionary adaption. Currently, only
in silico simulations were conducted using 1011 yeast ssGEMs; therefore, future experimental evidence for other non-reference strains will be important to evaluate the reliability of our model predictions.

With the increasing use of adaptive laboratory evolution in metabolic engineering it is important to have methods for rapid assessment of mutations in strains with improved phenotypes. Our multi-scale model analysis will be easily applicable in this area. As a further extension of Yeast8, we developed proYeast8DB by collecting and evaluating the yeast metabolic protein 3D structures from public databases. This enabled identification of mutational hotspots associated with specific phenotypes. Furthermore, through combining the predicted targets from ecYeast8 and proYeast8DB, we demonstrated how to identify mutations in enzymes with high flux control over a given pathway, which may also be associated with desirable phenotypes. It should be noted that although proYeast8DB is useful for connecting GEMs to protein structure information like PDB identifiers and protein parameters, it is still challenging to directly predict phenotypes using the model with protein structure variations as input. High quality 3D protein structures at genome scale are still scarce, thus the breakthrough in 3D proteins structure simulations and the related residues functional predictions are strongly expected. Nevertheless, like Recon3D and the E. coli GEM-PRO, the proYeast8DB holds value as a means to explore the relation between the cell genotype and phenotype with clear evidence.

Yeast8 will continue to be developed together with its ecosystem of models. As a whole, they are expected be a solid basis for developing a whole cell model of an eukaryal cell, which may serve as a stepping stone to a wider use of model simulations in...
life sciences, resulting in reducing the costs of developing biotechnology processes and drug discovery.

Methods
Tracking model changes with version control. Git and GitHub were used to develop yeast-GEM in a traceable way. Git is used to track any changes of yeast-GEM. The changes in GitHub or Git local clients. The.xml (SBML) format makes it easy to import the model across different toolboxes and programming languages.

Developers, and even other people from the community, can create new branches in the development branch, and in turn the changes in the development branch are used, including a

import the model across different toolboxes and programming languages.

As a standard step, a commit is needed when updating yeast-GEM. To make commits easy to understand, semantic commit messages are used (Supplementary Fig. 1c). To enable parallel model development, different branches of yeast-GEM are used, including a ‘master’ branch and a ‘devel’ (development) branch. Developers, and even other people from the community, can create new branches from the development branch to introduce their changes, and then request to merge them back through pull-requests. These changes are only merged to the development branch, and in turn the changes in the development branch are merged periodically to the master branch, which contains the stable releases of the model.

General procedures used to standardise annotation of metabolites and reactions. For the newly added reactions, their MetaNetX IDs were obtained according to a direct search in the MetaNetX database using the related metabolite name or EC number information. MetaNetX IDs were also obtained by reaction ID mapping from the KEGG53, Rhea52 and BioCyc53 databases. The reaction reversibility was corrected based on the BioCyc and BiGG databases58. MetaNetX IDs were also used to obtain the EC number for the corresponding reactions. As the MetaNetX database does not have the reaction name information, the name of each new reaction was obtained based on the reaction ID mapping in databases of KEGG, ModelSeed and BioCyc.

The compartment annotation of new reactions was refined based on information from the UniProt and SGD databases. The subsystem annotation was firstly obtained from KEGG53, and if no subsystems were found there, information from BioCyc or Reactome54 was used instead. If the reaction had no gene relations, we assumed that it occurred in the cytoplasm.

For all the metabolites contained in newly added reactions, the related MetaNetX IDs were obtained based on the corresponding MetaNetX IDs. If no MetaNetX IDs were obtained, they were obtained by ID mapping based on KEGG IDs or ChEBI IDs. Once the metabolite MetaNetX IDs were obtained, the charge, formula, KEGG IDs and ChEBI IDs were obtained for the correspondent metabolite based on metabolites annotation in MetaNetX.

Model update from Yeast7 to Yeast8. Firstly, all the annotations regarding metabolite ChEBI IDs and KEGG IDs (Supplementary Table 8) were corrected in the latest version of the consensus GEM of yeast (version 7.6) based on the metabolite annotation available in KEGG and ChEBI59. Additionally, several genes from iSac92631 that were not included in yeast 7.6 were added, as with all genes related to metabolic processes and transport in SGD, BioCyc, Reactome, KEGG and UniProt. The main databases used for model curation could be found in Supplementary Table 9.

In the Biolog experiments, the strain S288c was grown on 190 carbon sources, 95 nitrogen sources, 59 phosphorus sources, and 35 sulphur sources. The result showed that S288c could grow on 28 carbon sources, 44 nitrogen sources, 48 phosphorus sources and 19 sulphur sources. Based on these results new essential genes were added to make the model capable of predicting growth on the related substrates. Meanwhile, all the metabonomics data contained in the YMDB database (measured metabolites) and the latest metabolomics research (Supplementary Table 10) were collected and compared with that in yeast GEM. A standard annotation was given for all these metabolites and a pipeline was designed to add the metabolites to the GEM without bringing any new dead-end metabolites. Detailed procedures in model curation are available in the Supplementary Methods.

Model validation with varied experimental data sources. To compare the metabolites coverage, the YMDB database60 was parsed. There are 204 metabolites for yeast, among which 871 were measured in S. cerevisiae. For each metabolite, ChEBI ID and KEGG ID were assigned, and based on them the corresponding MetaNetX ID was matched. For metabolites from Yeast7 and Yeast8, the MetaNetX ID of each metabolite was also obtained based on ID mapping.

The in silico growth on 190 carbon sources, 59 nitrogen sources, 59 phosphorus sources, and 35 sulphur sources were calculated and compared with phenotype data. The results can be divided into a confusion matrix, which contains: (1) G/G: in vivo growth/in silico growth (true positive); (2) NG/NG: in vivo no growth/in silico no growth (true negative); (3) G/NG: in vivo growth/in silico no growth (false negative); (4) NG/G: in vivo no growth/in silico growth (false positive).

The model quality is then evaluated based on accuracy (Eq. 1) and the Matthews’ Correlation Coefficient (MCC)51 (Eq. 2). Accuracy ranges from 0 (worst accuracy) to 1 (best accuracy). MCC ranges from −1 (total disagreement between prediction and observation) to +1 (perfect prediction).

\[
\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}
\]

\[
\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}
\]

To conduct gene essentiality analysis, we used the essential gene list from the Yeast Deletion Project, available at http://www.sequence.stanford.edu/groups/yeast_deletion_project/downloads.html, which was generated from experiments using a complete medium. Accuracy and MCC were computed as described above.

The simulated aerobic and anaerobic growth under glucose-limited and nitrogen-limited conditions were compared with reference data62. The following procedure was employed to simulate chemostat growth in glucose-limited conditions. Firstly set the lower bound of glucose and O2 uptake reactions using experimental values. Glucose and oxygen uptake fluxes are negative and therefore the lower bounds are fixed to represent the maximum uptake rates. Secondly maximise the growth rate.

As for nitrogen-limited conditions, since protein content in biomass drops dramatically under nitrogen-limited conditions, the biomass composition was rescaled according to reference conditions63, then set the lower bound as measured for NH3 and O2 uptake reactions using experimental values and finally maximise the growth rate.

Visualisation of Yeast8. The maps of yeast-GEM were drawn for each subsystem using cellDesigner 4.4.58 (Supplementary Fig. 5). In-house R scripts were used to produce the map of each subsystem automatically based on Yeast8. Afterwards, the graph layout was adjusted manually in cellDesigner 4.4 to improve its quality and the whole yeast map in SBGN format could be found in https://github.com/SysBioChalmers/Yeast-maps/tree/master/SBMLfiles.

Generation of ecYeast8. The ecYeast8 model was generated based on the latest release of the GECKO toolbox, available at https://github.com/SysBioChalmers/GECKO. For each reaction, the algorithm queries all the necessary k_{cat} values from the BRENDA database64, according to gene annotation and a hierarchical set of criteria, giving priority to substrate and organism specificity. The k_{cat} values are then added to reactions according to:

\[
- \frac{1}{k_{cat}} v_i + c_i = 0
\]

\[
0 \leq c_i \leq [E_i]
\]

\[
v_i = k_{cat}^i [E_i]
\]

where v represents the flux through reaction i, c_i represents the amount of enzyme allocated for reaction i, E_i represents the total concentration of enzyme i, and k_{cat} represents the highest turnover number available for enzyme i and reaction j. The detailed procedure to generate ecYeast8 can be found in the supplementary material of the GECKO paper36.

Simulations with ecYeast8. To predict the maximum growth rate under different carbon and nitrogen sources using ecYeast8, the following procedure was used. Firstly remove any constraints for the related uptake rates of carbon and nitrogen sources, set minimal initial uptake of the related carbon and nitrogen sources. Lastly, simulate a growth rate maximisation, whereby the optimal value is fixed for posterior minimisation of the total protein usage. This provides a parsimonious flux distribution.

For comparative FVA between Yeast8 and ecYeast8, the maximum growth rate and the optimal glucose uptake rates obtained with ecYeast8 are used as fixed value and upper bound, respectively, in the original GEM in order to perform a fair comparison of flux variability for the same growth phenotype.

Flux control coefficients (FCCs) are defined as a ratio between a relative change in the flux of interest and a relative change in the correspondent k_{cat} of 0.1%, which can be described by:

\[
\text{FCCC} = \frac{v_{in} - v_{out}}{v_{out}} \cdot \frac{1.001k_{cat}^i - k_{cat}^i}{k_{cat}^i}
\]

where v_{in} and v_{out} are the original flux and new fluxes respectively when the k_{cat} is increased by 0.1%.
Re-annotation of the pan-genome from the 1011 yeast genome-sequencing project.

To construct the pan model of yeast (panYeast8), the latest genomics research on S. cerevisiae has been conducted. In Peter's study, 1011 yeast strains genomics had been sequenced and analysed. A pan-genome was obtained from all these strains, made up of by 6081 non-redundant ORFs from S. cerevisiae 2888C reference genome, and 1715 non-reference ORFs (nROFs) from the other strains. For the 7796 ORFs, a panID was given for each of them. By comparison, 4940 ORFs were found in these strains while 2850 across all these strains. The annotation of non-redundant 6081 ORFs can be taken directly from the latest S. cerevisiae 2888C genome annotation, while related gene–protein–reactions (GPR) can be obtained from Yeast8 directly.

As mentioned in Peter's article there are 774 nROFs with the orthology genes from S. cerevisiae 2888C. The blast analysis, along with the KEGG annotation of KEGG web service35, and EggNOG web service36, were employed to check and improve the original ortholog relation. To evaluate the ortholog gene relations qualitatively, the bi-directional blast hit (BBH) analysis was further conducted using Diamond66. Here the best hit in BBH analysis with identity larger than 80% were initially chosen and prepared for a panYeast8 formulation.

To further search reliable new reactions connected with nROFs, the annotation results from KEGG and the EggNOG web service were used. According to the format request for the two web services, the protein fasta files of pan-genome were uploaded onto KEGG (https://www.genome.jp/tools/kaas/) and EggNOG (https://eggnogdbs.embl.de/). For the KEGG annotation, a BBH (bi-directional best hit) assignment method with the default parameters was used. For the EggNOG annotation, the HMMER with the default parameters was used. In this work, mainly four import parameters, that are sequence identity (SI), sequence similarity (SS), resolution, and QMEAN, were used to classify the PDB structures. Among these parameters, QMEAN is the most important parameter. It ranges from 0 to 1, and the higher the value, the better the quality.

The relative numbers of sSNPs and nsSNPs for each gene were calculated. In this work, the Pfam32.0 database was mainly used to annotate the domain information of proteins from S. cerevisiae 2888C. If a structure covered all residues of any given domain, it was assigned to the domain. For each domain, the coordinates and the annotation information of the domain function description, the domain type, e-value, the related PDB ID, and protein ID, were all summarised. According to the GPRs of Yeast8, the relation between gene ID and reaction ID could be obtained. Following this, the domain information could be connected with each pair of gene and reaction based on the ID mapping.

SNP collection and relative coordinates mapping. Starting from the vcf file provided by the recent 1011 yeast strains genomics sequencing projects the homoyzogous SNP from the massive data file (Supplementary Fig. 10a) were firstly filtered. The SNP with depth below 5 and Genotype quality <40, genotype quality <30, and Genotype depth <5 were filtered out based on a series of standard parameters according to the Broad Institute Genome analysis Toolkit (GATK)35.

After filtration, the reliable SNP can be obtained for each strain. The data furthermore contains each SNP’s strain name, chromosome, coordinates, ref, and alt nucleotide base. In the annotation phase, the SNP type and related gene names were further annotated based on the coordinates and the annotation information of S. cerevisiae S. cerevisiae 2888C reference genome (version R64-1-1) from NCBI. If the SNP was not located on CDS zone of gene, it was classed as a non-synonymous single nucleotide polymorphism (nsSNP).

To build the PDB files for this gap, the SWISS-MODEL web service was further used to build the PDB_homos for 217 proteins. As a result, each of metabolic protein contains a common set of domain in its original protein. After removing gaps and mutations, the residues sequence and protein length were downloaded from the SGD database.

Once the PDB files were collected, the parameters of PDBs were extracted and calculated for quality analysis. For the PDB_homos, the default parameters from the ftp of the SWISS-MODEL database were obtained, and included the protein UniProt ID, the protein length, the related PDB ID (connected with chainID), the structure sources, the coordinates of protein residues covered with PDB structures, the coverage, the resolution, and QMEAN. For PDB_homos, besides the above default parameters from the SWISS-MODEL database, a greater number of parameters were obtained by using the PDB_homo atom files provided by the SWISS-MODEL with an in-house python script, which included the methods used to obtain the PDB files, the model template, the protein oliga state, the GMQE, QM4, sequence identity (SID), and sequence similarity (SIM). In summary, each PDB_homos contains 18 parameters for further PDB quality analysis.

For the PDB_ex parameters, like coverage and template ID can also be found from the SWISS-MODEL database. The other important parameters like resolution, ligands, and oliga state were obtained by parsing PDB_ex files from RCSB PDB database using (https://github.com/williamgilpin/pypdb). The chainID for each PDB_ex was downloaded from the SIFTS database35. Quality analysis of protein 3D structure. As one protein could be connected with several PDB files in different quality levels, it is essential to filter out the PDB of low quality. In this work, mainly four import parameters, that are sequence identity (SI), sequence similarity (SS), resolution, and QMEAN, were used to classify the PDB structures. Among these parameters, QMEAN is the most important parameter. It ranges from 0 to 1, and the higher the value, the better the quality.
calculated, which is equal to the total sSNPs or mSNPs divided by the related protein length.

Before mapping, the coordinates of mutated residues from each mSNP need to be calculated. Firstly, the relative coordinates of mutated residues on the original protein sequence can be obtained based on the coordinates of mSNP on the chromosome. Following this, according to the coordinates mapping between the original protein sequences and the relative residues coordinates in the proteins structure, the relative coordinates of the mutated residues in the protein structures can be estimated and used in the following calculation.

CLUMPS method to calculate p-values of mutation enriched PDB files.

Referring to Kamburov’s method45, a WAP score to calculate the pairwise distances between mutated residues for a protein 3D structure.

\[
W_{\text{AP}} = \sum_{i} n_{i} n_{e} \frac{d_{i}}{N_{p}}
\]

Where \(d_{i}\) in this article is defined as the Euclidean distance (in Å) between α carbons of any two mutated residues. \(i\) is defined as a ‘soft’ distance threshold, which equals to 6 Å. \(n_{i}\) and \(n_{e}\) are the normalised numbers of samples contains the mutations using the followed sigmoidal Hill function:

\[
n_{q} = \frac{N_{m}}{\theta^{m} + N_{m}^{m}}
\]

Where \(N_{q}\) is the number of samples with a missense mutation impacting residue \(q\) of the protein and \(\theta = 2\) and \(m = 3\) are parameters of the Hill function controlling the critical point (centre) and steepness of the sigmoid function, respectively. Formula (2) was used to normalise the sample number contained in residue mutations \(q\) and \(r\), both of which can avoid the impact of higher frequent mutated residues in the samples. A detailed description of each formula can be found in Kamburov’s article45.

The CLUMPS method can be divided into four steps. Firstly, prepare the needed SNP information and structure information of one protein. Secondly, with the normalised mutation number occurring in specific positions, calculate the WAP scores of the samples. Next, assuming that the uniform distribution of mutations across the protein residues covers the given structure, calculate each WAP score in 104 randomisations to obtain the null distribution. During the sampling process, the mutation number of residues occurring in random locations was kept the same as the original values. Lastly, calculate the right tailed \(P\) value in the null distribution for the given mutated protein structures based on the original WAP score and all the sampled WAP scores. The right tailed \(P\) value is defined as the number of samples with WAP scores larger than the original WAP scored, divided by the total number of samples.

For proteins with \(P\) value smaller than 0.05 from strains group of “Bioethanol” and “Wine”, GO-enrichment analysis using DAVID6.7 on-line web service27 was carried out.

Hotspot analysis of nsSNP mutation.

The hotspot analysis pipeline for yeast mainly refers to Niu et al’s work49. All the SNP and structure information (similar to CLUMPS’ analysis method) were prepared for a group of strains with specific phenotypes. Before carrying out the cluster analysis, the mutated paired residues of significance were filtered according to reference49. These important paired residues should meet the followed three criteria: the distance between two residues should be smaller than 10 Å for all the intramolecular clusters analysis; the two residues should be separated by at least 20 residues in the original protein sequence; and a permutation method ought to be used to calculate the \(P\) value for each paired residues (Eq. 9), with a threshold set at 0.05.

\[
P_{\text{value}} = \frac{n_{1}}{n_{2}}
\]

Where \(n_{1}\) is the number of paired residues with the distance smaller than that in the paired residues of target and \(n_{2}\) is the total number of paired residues.

Once the paired residues of significance have been obtained, the clusters made up of paired residues were obtained through graph theory, which was realised using the function ‘decompose.graph’ from the R package igraph (https://igraph.org/). For each cluster, its closeness can be calculated using the function of ‘closeness.residual’ from the R package entsime23. The detailed principle could be also find in the original reference23. As the last step, when a cluster was estimated, the \(P\) value was calculated based on the CLUMPS analysis pipeline in this work.

Prediction of mutations function.

To look into the effect of the mutations, mutfunc48 (http://mutfunc.com/) and SIFT (http://sift.jcvi.org/) were employed to predict the potential effects of mSNP on protein function.

Growth test using Biolog with different substrate sources.

The Phenotype MicroArray (PM) system was used to test growth on every carbon, nitrogen, phosphorus and sulphur sources. A total of 190 carbon sources, 95 nitrogen sources, 95 phosphorus, and sulphur sources were tested. The PM procedures for S. cerevisiae $S_{288C}$ were based on the protocol of Yeast version of the PM system.

Growth profiling in different media.

A total of 14 carbon sources and 23 nitrogen sources were combined by orthogonal experiments. Every carbon source and nitrogen source used in the medium were the same C-mole and N-mole as glucose (20 g L$^{-1}$ glucose) and ammonium sulphate (7.5 g L$^{-1}$ (NH$_4$)$_2$SO$_4$), respectively. For all other substrate sources, the same minimal medium was used (14.4 g L$^{-1}$ KH$_2$PO$_4$, 0.5 g L$^{-1}$ MgSO$_4$7H$_2$O, trace metal and vitamin solutions)35. Strains were cultivated in 96-well plates, and growth performance was determined with Growth Profiler 960 (Euryscreen B.V., Heemstede, The Netherlands). The maximum specific growth rate (\(\mu_{\text{max}}\)) was calculated with the R package—growthrates (https://github.com/tpetzoldt/growthrates).

Statistical analysis.

For two group comparison in this work, a two tailed Wilcoxon rank sum test was used.

Reporting summary.

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability.

The yeast-GEM project with all related data sources can be found at https://github.com/SysBioChalmers/yeast-GEM, and a full documentation on how the repository works and how to contribute is available at https://github.com/SysBioChalmers/yeast-GEM/blob/master/github/CONTRIBUTING.md. The panYeast with all related data sources can be found in https://github.com/SysBioChalmers/panYeast-GEM. The genomic data used in this work is from40.

Code availability.

Matlab scripts for development of yeast GEM and panYeast can be found in the above two corresponding repositories. R scripts to carry out the CLUMPS and hotspot analysis can be found at: https://github.com/SysBioChalmers/proYeast-GEM.

Received: 19 March 2019 Accepted: 17 July 2019

Published online: 08 August 2019

References.

1. Garza, D. R., van Verk, M. C., Huynen, M. A. & Dutilh, B. E. Towards predicting the environmental metabolome from metagenomics with a mechanistic model. Nat. Microbiol. 3, 456–460 (2018).
2. Robinson, I. L. & Nielsen, J. Anticancer drug discovery through genome-scale metabolic modeling. Current Opinion in. Syst. Biol. 4, 1–8 (2017).
3. Chen, X. et al. DCEO biotechnology: tools to design, construct, evaluate, and optimize the metabolic pathway for biosynthesis of chemicals. Chem. Rev. 118, 70–72 (2017).
4. Schellenberger, J. et al. Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0. Nat. Protoc. 6, 1290–1307 (2011).
5. Wang, G., Huang, M. & Nielsen, J. Exploring the potential of Saccharomyces cerevisiae for biopharmaceutical production protein. Curr. Opin. Biotechnol. 48, 77–84 (2017).
6. Brunk, E. et al. Recon3D enables a three-dimensional view of gene variation in human metabolism. Nat. Biotechnol. 36, 272–281 (2018).
7. Hevner, B. D. & Price, N. D. Comparative analysis of yeast metabolic network models highlights progress, opportunities for metabolic reconstruction. PLoS Comput. Biol. 11, e1004530 (2015).
8. Monk, J. M. et al. IME1515, a knowledgebase that computes Escherichia coli traits. Nat. Biotechnol. 35, 904–908 (2017).
9. Ate, M. et al. Traceability, reproducibility and wiki-exploration for “a-la-carte” reconstructions of genome-scale metabolic models. PLoS Comput Biol. 14, e1006146 (2018).
10. Spaulding, A. et al. Pathway Tools version 19.0 update: software for pathway/gene network models highlights progress, opportunities for metabolic reconstruction. Metab. Eng. 21, 103–113 (2014).
11. Sun, S. et al. An extended set of yeast-based functional assays accurately identifies human disease mutations. Genome Res. 26, 670–680 (2016).
12. Dubs, P., Kumar, D. & Sharma, N. In Yeast Diversity in Human Welfare. (eds. T. Sutaiyanayana & G. Kunze) 191–214 (Springer Singapore, Singapore, 2017).
13. DiCarlo, J. E. et al Genome engineering in Saccharomyces cerevisiae using CRISPR-Cas systems. Nucleic Acids Res. 41, 4336–4343 (2013).
16. Gasch, A. P. et al. Single-cell RNA sequencing reveals intrinsic and extrinsic regulatory heterogeneity in yeast responding to stress. Proc. Natl. Acad. Sci. USA 112, 5486–5495 (2015).

17. Forster, J., Famili, I., Fu, P., Palsson, B. O. & Nielsen, J. Genome-scale reconstruction of the Saccharomyces cerevisiae metabolic network. Genome Res. 13, 244–253 (2003).

18. Lahtvee, P. J. et al. Absolute quantification of protein and mRNA abundances demonstrate variations in gene-specific translation efficiency in yeast. Cell Syst. 4, 495–504 (2017).

19. Tian, M. & Reed, J. I. Integrating proteomic or transcriptomic data into metabolic models using linear bound flux balance analysis. Bioinformatics 34, 3882–3888 (2018).

20. Lopes, H. & Rocha, I. Genome-scale modeling of yeast: chronology, applications and critical perspectives. FEBS Yeast Res. 17, 1–13 (2017).

21. Pereira, R., Nielsen, J. & Rocha, I. Improving the flux distributions simulated with genome-scale metabolic models of Saccharomyces cerevisiae. Metab. Eng. Commun. 3, 153–163 (2015).

22. Aung, H. W., Henry, S. A. & Walker, L. P. Revising the representation of fatty acid, glycerolipid, and glycerophospholipid metabolism in the consensus model of yeast metabolism. Ind. Biotechnol. (New Rochelle N. Y.) 9, 215–228 (2013).

23. Yang, L., Yurkovich, J. T., King, Z. A. & Palsson, B. O. Modeling the multi-scale mechanisms of macromolecular resource allocation. Curr. Opin. Microbiol. 45, 8–15 (2018).

24. Osterlund, T., Nookaew, I., Bordel, S. & Nielsen, J. A genome-scale model of metabolism and gene expression extend and refine growth phenotype prediction. Mol. Syst. Biol. 9, 693 (2013).

25. Sanchez, B. J. et al. Improving the phenotype predictions of a yeast genome-scale metabolic model by incorporating enzymatic constraints. Mol. Syst. Biol. 13, 935 (2017).

26. Chang, R. L. et al. Structural systems biology evaluation of metabolic thermotolerance in Escherichia coli. Science 340, 1220–1223 (2013).

27. Conesa, F. et al. MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic Acids Res. 44, 471–480 (2016).

28. Fabregat, A. et al. The reactome pathway knowledgebase. Nucleic Acids Res. 42, 649–655 (2014).

29. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 28, 27–30 (2000).

30. The UniProt Consortium. UniProt: the universal protein knowledgebase. Nucleic Acids Res. 45, 158–169 (2017).

31. Sanchez, B. J., Li, F., Kerkhoven, E. J. & Nielsen, J. SLIMer: probing flexibility of lipid metabolism in yeast with an improved constraint-based modeling framework. BMC Syst. Biol. 13, 4 (2019).

32. Funahashi, A. et al. CellDesigner 3.5: a versatile modeling tool for biochemical networks. Proc. IEEE 96, 1254–1265 (2008).

33. Nilsson, A. & Nielsen, J. Metabolic trade-offs in yeast are caused by F1F0-ATP synthase. Phytochem. Res. 6, 202 (2016).

34. Peters, J. et al. Metabolic engineering of yeast: a comprehensive resource. Bioresource 18, 20338–20343 (2013).

35. Jordan, I. K., Rogozin, I. B., Wolf, Y. I. & Koonin, E. V. Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. Genome Res. 12, 962–968 (2002).

36. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat. Protoc. 4, 44–57 (2008).

37. Kamburov, A. et al. Comprehensive assessment of cancer missense mutation clustering in protein structures. Proc. Natl. Acad. Sci. USA 112, 5486–5495 (2015).
Acknowledgements
We would like to thank Pınar Kocabaş, Hao Wang, Jonathan Robinson, and Rui Pereira for their valuable input. This project has received funding from the Novo Nordisk Foundation (grant no. NNF10CC1016517), the Knut and Alice Wallenberg Foundation, and the European Union’s Horizon 2020 research and innovation program with projects DD-DeCaF and CHASSY (grant agreements No 686070 and 720824). Open access funding provided by Chalmers University of Technology.

Author contributions
H.L., F.L., B.J.S., E.J.K. and J.N. conceived the study. H.L., F.L., B.J.S., Z.Z. and E.J.K. contributed to the yeast GEM update. B.J.S. and I.D. developed ecYeast8 and performed the corresponding simulations. H.L., F.L. and E.J.K. designed and generated panYeast8 and the 1,011 strain specific GEMs. H.L. designed and contributed to the completion of proYeast8DB. G.L. assisted in the protein structure data collection for proYeast8DB. Z.Z. and F.L. designed and performed the experiments. Z.Z. and H.L. designed and contributed to the yeast map. S.M., P.M.A., D.L., C.L., M.E.B., N.S. and E.J.K. provided support and ideas for yeast GEM development using version control. H.L., F.L. and J.N. wrote the original manuscript. All authors read, edited, and approved the final paper.

Additional information
Supplementary Information accompanies this paper at https://doi.org/10.1038/s41467-019-11581-3.

Competing interests: The authors declare no competing interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

Peer review information: Nature Communications thanks the anonymous reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Publisher’s note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.