

# Deciphering complex traits with deep combinatorial genetic analysis

Albi Celaj<sup>1,2,3</sup>, Marinella Gebbia<sup>1</sup>, Louai Musa<sup>1</sup>, Atina G. Cote<sup>2</sup>, Jamie Snider<sup>1</sup>, Victoria Wong<sup>1</sup>, Minjeong Ko<sup>1,2,4</sup>, Tiffany Fong<sup>1</sup>, Paul Bansal<sup>1,2</sup>, Joseph C. Mellor<sup>1</sup>, Gireesh Seesankar<sup>1</sup>, Maria Nguyen<sup>1</sup>, Shijie Zhou<sup>1,3</sup>, Igor Stagljar<sup>1,3,5</sup>, Yo Suzuki<sup>6</sup>, Nozomu Yachie<sup>1,6,7,8,9,10,\*</sup>, and Frederick P. Roth<sup>1,2,3,6,9,11,12 \*</sup>

[Author list and order is not final]

<sup>1</sup>Donnelly Centre, University of Toronto, Toronto, Ontario, Canada.

<sup>2</sup>Lunenfeld-Tanenbaum Research Institute, Mount Sinai Hospital, Toronto, Ontario, Canada.

<sup>3</sup>Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada.

<sup>4</sup>Department of Computer Science, University of Toronto, Toronto, Ontario, Canada.

<sup>5</sup>Department of Biochemistry, University of Toronto, Toronto, Ontario, Canada.

<sup>6</sup>Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, MA, USA.

<sup>7</sup>Synthetic Biology Division, Research Center for Advanced Science and Technology, the University of Tokyo, Tokyo, Japan.

<sup>8</sup>Department of Biological Sciences, School of Science, the University of Tokyo, Tokyo, Japan.

<sup>9</sup>Institute for Advanced Biosciences, Keio University, Tsuruoka, Japan.

<sup>10</sup>PRESTO, Japan Science and Technology Agency, Tokyo, Japan.

<sup>11</sup>Canadian Institute for Advanced Research, Toronto, Ontario, Canada.

<sup>12</sup>Lead contact

\*Correspondence: [fritz.roth@utoronto.ca](mailto:fritz.roth@utoronto.ca) (F.P.R.), [yachie@synbiol.rcast.u-tokyo.ac.jp](mailto:yachie@synbiol.rcast.u-tokyo.ac.jp) (N.Y.)

## Abstract

Complex genotype-to-trait relationships can be understood by perturbing genes in many different combinations and observing the impact. Here, we describe a method to efficiently engineer, genotype, and phenotype complex multi-gene variant combinations within a target gene set, enabling a deep combinatorial genetic analysis (DCGA). Applying DCGA to 16 yeast ABC transporters, we generated and genotyped >5,000 strains bearing deletions of random subsets of transporters, and profiled each strain for resistance to each of 16 bioactive compounds ('drugs'). DCGA revealed a complex genotype-to-resistance landscape, showing high-order drug-dependent genetic interactions for 13 of the 16 transporters studied. We developed a neural network to derive intuitive system models from these complex genetic relationships. Guided by modeling, we further characterized a quadruple knockout strain (*snq2Δ yor1Δ ybt1Δ ycf1Δ*) which unexpectedly showed high *PDR5*-dependent resistance to fluconazole. Results showed that high-order genotype-to-trait relationships discovered by DCGA can help dissect complex biological systems.

## Introduction

Genes often encode products which form interdependent and functionally-overlapping molecular systems, such that combinations of genetic perturbations can yield surprising and informative phenotypes (Hartman et al., 2001). This phenomenon defines genetic interaction. Observing the phenotypic effects of genes varied in combination, i.e., combinatorial genetic analysis (CGA), has systematically uncovered functional similarities and dependencies amongst genes in yeast (Costanzo et al., 2016) and human cells (Horlbeck et al., 2018; Shen and Ideker, 2018), improving our understanding of gene function (Costanzo et al., 2010, 2016) and order-of-action in biological pathways (St Onge et al., 2007).

Complex genetic interactions—surprising phenotypes arising from simultaneous perturbation of more than two genes—further inform gene function. Indeed, three-gene interactions (for which a triple mutant phenotype cannot be simply explained by the underlying single and double mutant phenotypes) are expected to outnumber two-gene interactions by 100-fold (Kuzmin et al., 2018). Even when two-gene interactions are present, important functional inferences cannot be made until triple-knockout phenotypes are observed and analyzed (Braberg et al., 2014; Haber et al., 2013; Kuzmin et al., 2018). Single knockout alleles have been shown to have environment-dependent complex genetic interactions with natural variants that typically involve three or more additional loci (Mullis et al., 2018). More complex interactions are known, e.g., involving five (Taylor et al., 2014), seven (Beh et al., 2001), and over twenty genes (Wieczorke et al., 1999). However, the inherent experimental and analytic challenges have limited systematic profiling and functional inference involving highly-complex genetic interactions.

Functionally deciphering the potentially complex effects of multiple genetic perturbations requires extension of two-gene CGAs for more exhaustive ‘deep’ combinatorial genetic analysis (DCGA). Here we describe a strategy for DCGA and demonstrate it on the entire set of 16 yeast ABC transporters that have been implicated in multi-drug resistance. Although deletion of drug efflux pumps might naively be expected to increase drug sensitivity, examples of ABC transporter deletions that can influence one another in complex ways have been reported, such that disruption of a transporter can impart drug *resistance* (Kolaczkowska et al., 2008; Snider et al., 2013; Suzuki et al., 2011). Substrate sharing and complex interdependence between these ABC transporters can limit their analysis with two-gene CGA. We used DCGA to uncover and visualize a complex genetic landscape in each of 16 bioactive compounds (‘drugs’), revealing condition-dependent genetic interactions involving as many as five genes. To functionally decipher this system of transporters, we then developed a neural network which could learn an intuitive system model directly from the observed genotype-to-phenotype relationships. Modeling guided further mechanistic exploration of a quadruple knockout combination (*snq2Δ yor1Δ ybt1Δ ycf1Δ*) which conferred unexpectedly high resistance to fluconazole. Taken together, we illustrate an experimental and analytic approach that uncovers many unexpected high-order genetic relationships to shed light on complex molecular systems.

## Results

### Engineered population profiling: a scheme for generating and characterizing combinatorially complex mutants

Here we briefly describe the overall DCGA strategy and its component parts (Figure 1), then show results of the strategy as applied to a set of 16 yeast ABC transporters.

A simple yet powerful way to generate a complex population is to cross two outbred individuals differing at multiple unlinked loci. Progeny of the cross ('segregants') will inherit a random subset of parental variants (Bloom et al., 2013), and each segregant can then be genotyped and profiled for traits such as gene expression (Brem and Kruglyak, 2005) or drug resistance (Perlstein et al., 2007). Although this approach has been applied extensively to pairs of outbred parents, it has limitations. First, many genes involved in a process remain undetected due to the absence of functional natural variation in the parental strains (Lee et al., 2014). Second, parents can differ at  $\sim 10^5$  positions. This both makes it difficult to pinpoint the causal variants at each associated locus, and introduces a multiple testing burden such that a prohibitive number of individuals would be required for a comprehensive DCGA to have an acceptable statistical power.

To extend cross-based approaches beyond natural variation in outbred parents, we designed a population engineering strategy. In this strategy, targeted variation is engineered into individual parental strains which are crossed to yield an 'engineered population' with random segregation of only the engineered variants.

As recently reviewed (Kebschull and Zador, 2018), it is possible to track many individual strains in a complex heterogenous population using DNA barcodes. We therefore introduced a complex pool of random barcodes into one of the two haploid parental strains, such that each cell of that parental type bears a specific random barcode, as described previously (Díaz-Mejía et al., 2018; Yan et al., 2008) (Figure S1, see Methods for details).

After generating a barcoded pool in the wild-type parental strain background, we crossed this pool *en masse* to a previously-generated "ABC-16 strain" bearing knockouts for all sixteen of the ABC transporters that have been implicated in multidrug resistance (Suzuki et al., 2011). The ABC-16 strain had previously been engineered to contain all markers necessary to perform mating, sporulation, and selection for haploid cells, while the barcoded wild-type parent provided the marker necessary to select for cells carrying a barcoded *HO* locus (Methods). After mating, sporulation, and selection for barcoded haploid progeny of the cross, we used automated colony-picking to isolate an arrayed collection of  $\sim 5,000$  MAT $\alpha$  and  $\sim 5,000$  MAT $\alpha$  segregants in 384-well plates. This step generated an engineered population in which each individual haploid strain bears a random subset of knockout alleles for the target set of 16 ABC transporters.

For each strain in the collection of barcoded multi-knockout progeny, we obtained the genotype at all sixteen knockout loci and determined the identifier barcode. For genotyping, we exploited the fact that each knockout locus in the ABC-16 strain bears a locus-identifying barcode, as described previously for the YKO yeast deletion collection (Giaever et al., 2002; Suzuki et al., 2011). We adapted the previously-described row-column-plate PCR (RCP-PCR) strategy (Yachie et al., 2016), in which barcodes in each segregant are amplified with introduction of additional index

tags that identify the plate, row, and column of origin for each amplification product (Methods; Figure 1). Thus, a single next generation sequencing experiment can reveal both the strain-specific tracking barcode at the *HO* locus and the identity of every gene deleted in the segregant at each plate location (Methods; Figure 1).

To validate and calibrate the genotypes determined by high-throughput-sequencing, multiple replicates of 40 individually-genotyped ‘gold standard’ strains, as well as two additional control strains with known genotypes, were added to the collection at defined positions (Methods; Data S2). Using data from calibration strains, we estimated an overall per-locus genotyping accuracy of 93.2% (Figure S2A, Methods). An independent method relying on the distribution of knockouts in the pool estimated a similar overall per-locus accuracy of 93.8% (Figure S2B, Methods). Based on correlation analysis of the genotyping data, all genes were either unlinked or weakly linked except for *BPT1* and *YBT1* (Figure S2C;  $r = 0.49$ ), which are separated by 70.1kb on chromosome XII. Three pairs of unlinked genes—*YOR1-YCF1*, *YOR1-BPT1*, and *SNQ2-PDR5*—exhibited weak but significant negative correlation in the appearance of KO genotypes ( $-0.04 \geq r \geq -0.08$ ) (Figure S2C). This effect may have arisen via negative genetic interactions conferring lower growth for the corresponding double-knockout genotypes during the sporulation, haploid selection, or automated colony picking steps.

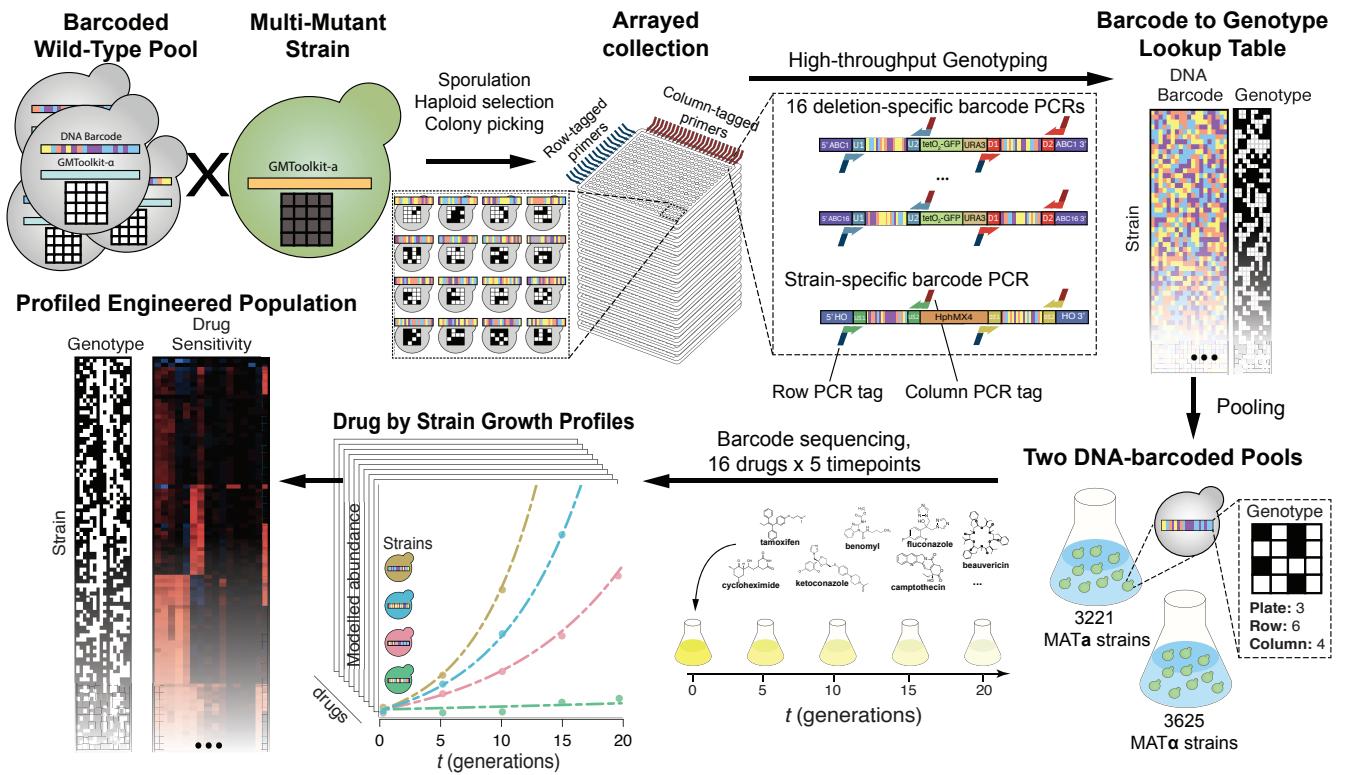
Considering only those strains with both high-quality genotyping data and at least one unique tracking barcode, this yielded 6,826 uniquely barcoded and genotyped strains, encompassing 6,087 unique genotypes. These strains were grouped by mating type to yield one pool of 3,231 MAT $\alpha$  strains and another pool of 3,595 MAT $\alpha$  strains.

Knowledge of the tracking barcode for each segregant enabled us to profile each strain’s resistance or sensitivity to particular drugs (Smith et al., 2009). Strain pools were grown competitively in each of 16 different anticancer and antifungal drugs (Data S3), and in a solvent (DMSO) control condition. Using high-throughput strain barcode sequencing, strain frequency was measured at five time points (corresponding to 0, 5, 10, 15, and 20 generations of overall pool growth, Figure 1), allowing us to compute a growth rate for each strain (Data S5; Methods).

We limited analyses to strains that were well-represented in the pre-selection pool ( $\geq 30$  barcode counts at t=0 in the solvent control - 5,790 [85%] of 6,826 strains), as these offered the best opportunity to detect changes in subsequent time points. To identify associations between each knockout and baseline growth rate (in the DMSO control condition), we applied a generalized linear model and found *yor1* $\Delta$ , *snq2* $\Delta$ , *ybt1* $\Delta$ , and *bpt1* $\Delta$  to have a statistically significant impact in both the MAT $\alpha$  and MAT $\alpha$  pools (Data S6, Figure S3). The impacts of *snq2* $\Delta$ , *ybt1* $\Delta$ , and *bpt1* $\Delta$  were small (<2% decrease in baseline growth rate), while *yor1* $\Delta$  had a stronger, but still modest effect (7-15% decrease). Before calculating drug resistance (growth rate in drug relative to that in solvent control) we further excluded all 437 strains exhibiting a strong baseline growth defect (i.e., showing <70% of the median baseline growth rate). In total, drug resistance was calculated for each of 2,367 MAT $\alpha$  and 2,986 MAT $\alpha$  strains, for each of 16 drugs (Data S5).

We applied a generalized linear model to identify and quantitatively model associations between individual knockouts and drug resistance (see Methods). Knockouts that significantly changed the estimated resistance to a drug by +/- 10%, were considered strong. All other significant associations were defined to be weak. Of the 62 drug-knockout associations we found, 19 were

**Figure 1**



strong (Data S6). Because 58 (87%) of these 62 single-gene associations and 100% of the 19 strong associations involved only five ABC transporters—*snq2Δ*, *pdr5Δ*, *yor1Δ*, *ycf1Δ*, and *ybt1Δ*—we initially restricted our attention to these ‘frequently-associated’ transporters. Among these were 18 drug-knockout associations involving the vacuolar ABC transporters *YCF1* and *YBT1*, all of which were novel (Figure S4, Data S6). For these five ‘frequently-associated’ transporters, we detected 89% of 18 previous associations between drugs and individual knockouts, while revealing 40 new associations (33 weak and 7 strong; Figure S4; Data S7).

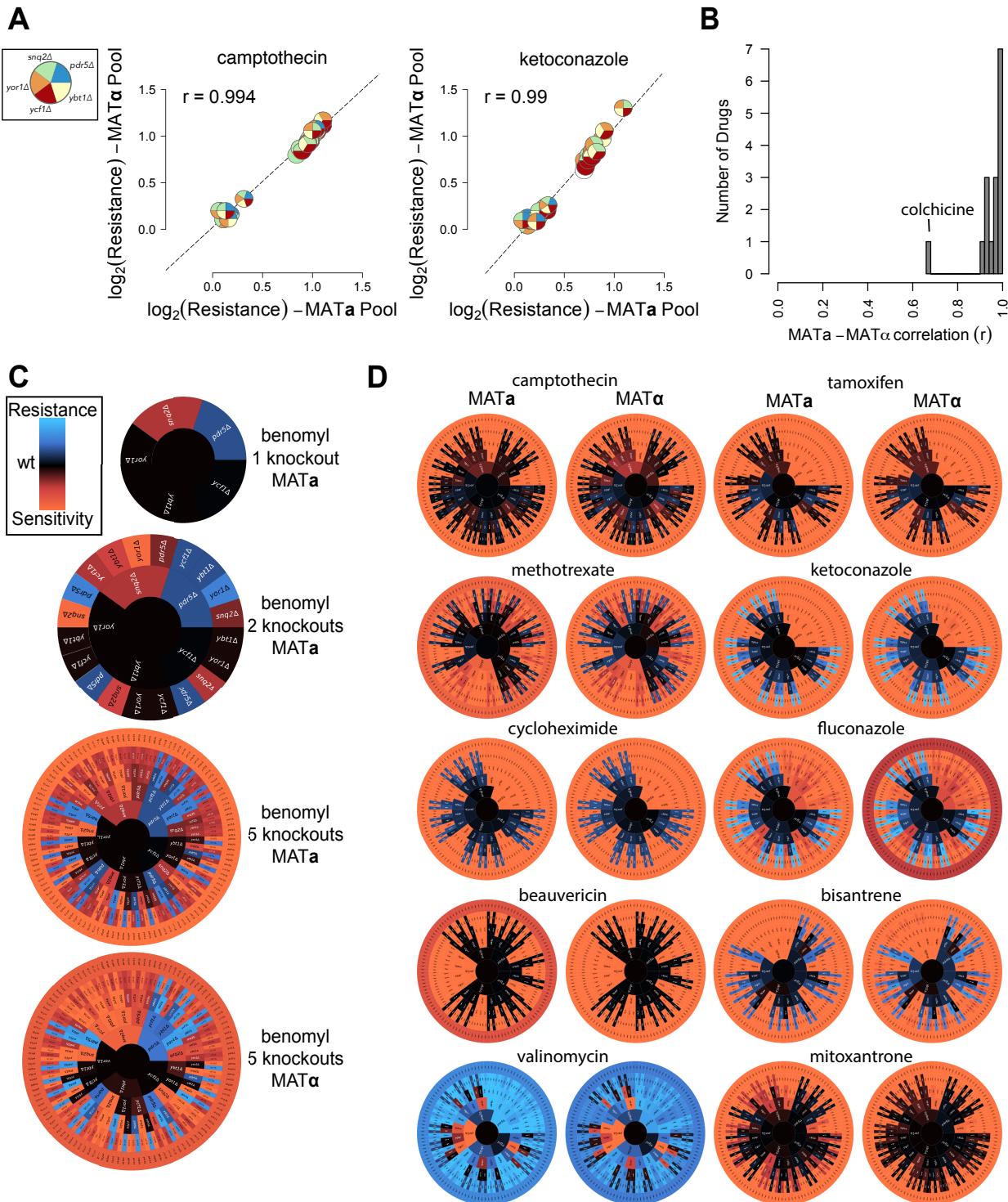
Considering only the five frequently-associated transporters, there are 32 possible combinatorial genotypes. We derived a phenotypic profile for each by calculating, for each drug, the average resistance over all strains matching this genotype at all five genes. These profiles were initially calculated separately for MAT $\alpha$  and MAT $\alpha$  strains (Figure S5). Detailed correlation analysis is shown for camptothecin and tamoxifen (Figure 2A). With the exception of colchicine ( $r = 0.77$ ), all drugs showed high reproducibility ( $r \geq 0.94$ ) between independent biological replicate pools (Figure 2B). We developed a radial visualization of this complex phenotypic landscape, in which the consequences of knocking out increasingly-many ABC transporters can be explored by tracing paths leading outward from the central wild-type genotype (Figure 2C). Graphs were visually similar between independent biological replicate MAT $\alpha$  and MAT $\alpha$  populations for many drugs, while showing large differences only for colchicine (Figure 2D and S6). Given high reproducibility, we merged MAT $\alpha$  and MAT $\alpha$  data for subsequent analyses, except where noted (Methods).

### Engineering population profiling reveals a complex drug-dependent genetic landscape

Visualizing the knockout profiles for each genotype in a fitness landscape representation (Figure 3A), we first verified that the knockout profiles could capture previously-reported relationships between ABC transporters and benomyl resistance. Our knockout profiles clearly captured the sensitivity of *snq2Δ* deletions to benomyl (Figure 3A left panel; 20% decreased resistance,  $p = 5.8e-80$ ; Wilcoxon rank sum test), which was expected given that Snq2 is known to be the primary efflux pump for benomyl (Kolaczkowski et al., 1998). We also observed several previously-reported phenomena, including increased benomyl resistance in *pdr5Δ* knockouts (13% increased resistance;  $p = 1.5e-96$ ) and a further increased benomyl resistance of the *pdr5Δ yor1Δ* double-mutant (21% increased resistance;  $p = 1.3e-72$ ). In keeping with (Kolaczkowska et al., 2008; Snider et al., 2013), these increases were dependent on the presence of *SNQ2*, with *pdr5Δsnq2Δ* yielding only a 5% increase in resistance relative to *snq2Δ* (and a 14% decrease relative to the wild-type). A comparable 6% relative increase was observed with *pdr5Δyor1Δsnq2Δ* relative to *snq2Δ* (these changes were significant,  $p = 1.4e-45$  and  $1.2e-38$ , respectively, relative to the knockout effects observed in a wild-type background, Figure 3A left panel). Although we did not observe *yor1Δ* to confer benomyl resistance ( $p = 0.09$ ), this was previously reported as a weak phenomenon (Snider et al., 2013). In summary, engineered population profiling largely recapitulated previously-reported effects of ABC transporter knockouts on benomyl resistance, including the effects of two- and three-gene combinatorial deletions.

Many of the multi-knockout effects we observed suggested the expected phenomenon of multiple partially-redundant efflux pumps acting in parallel. Specifically, we saw gene sets where each individual knockout shows sensitivity to a drug, and each higher-order knockout combination exhibits drug sensitivity that is higher than any of the individual component knockouts. Examples of this include the set {*snq2Δ*, *pdr5Δ*} under camptothecin (Figure S7), and the set {*snq2Δ*, *pdr5Δ*,

# Figure 2



$ybt1\Delta$ ,  $yor1\Delta$ } under mitoxantrone (Figure 3A middle panel, S7). These sensitivity patterns are consistent with a simple scenario in which each transporter can efflux a given drug.

In other cases, the fitness landscapes showed more surprising multi-knockout patterns conveying both drug resistance and sensitivity. For example, knocking out  $pdr5\Delta$ ,  $snq2\Delta$ ,  $ybt1\Delta$ , and  $ycf1\Delta$  individually or in any combination led to more valinomycin resistance than the wild-type strain (Figure 3A right panel). Indeed, the successive deletion of ABC transporters led to greater resistance for surprisingly many drugs (Figure 2D and S7).

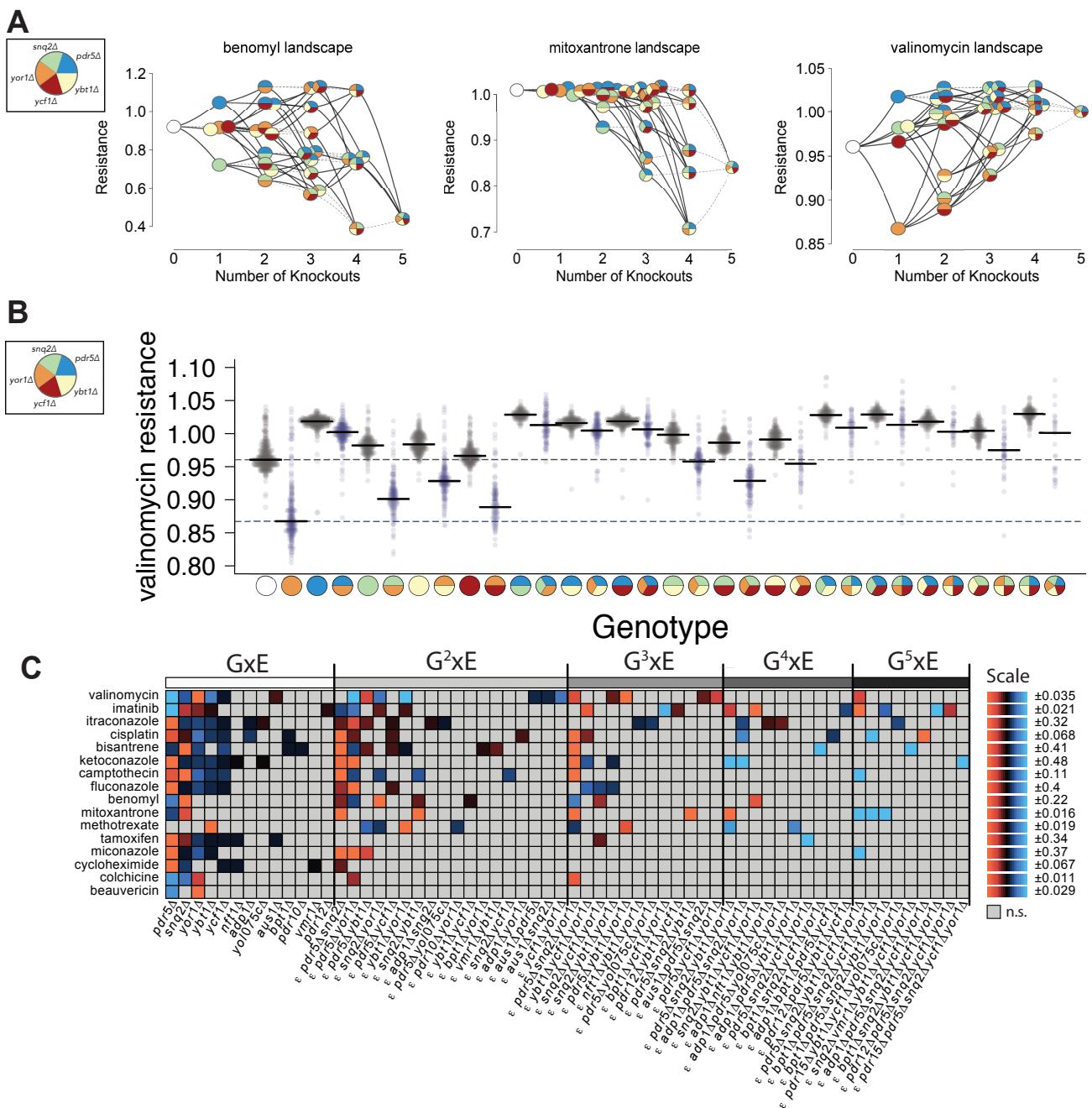
When considering only the five frequently-associated genes, the set of strains matching a specific genotype may in fact have heterogeneous genotypes owing to the variable presence of additional knockouts at the other 11 targeted transporter loci. We therefore visualized the distribution of valinomycin resistance for each of the 5-gene genotypes (grouping the results to show the effects of deleting  $YOR1$  in each genetic background [Figure 3B]). There was clearly high phenotypic variability within strains matching many of the five-gene genotypes. We therefore systematically expanded our search for multi-gene effects to include all 16 genes, using an extension (see Methods) of the linear model described above in the context of single-gene effects. All single and multi-gene interactions that passed the significance test ( $p < 0.05$  after adjusting for multiple testing) are shown in Figure 3C.

Our analysis yielded genetic interactions involving two or more genes for fifteen out of sixteen (94%) of the drugs examined (Figure 3C), with the exception of beauvericin for which we only recovered the previously-reported sensitivity of  $yor1\Delta$  knockouts (Shekhar-Guturja et al., 2016). Higher-order genetic interactions (involving three or more genes) were observed for fourteen of sixteen (88%) of drugs tested (Figure 3C). Here the exception (beyond beauvericin) was cycloheximide. For cycloheximide, we observed the previously-known strong single-gene  $pdr5\Delta$  effect (Ernst et al., 2008; Katzmann et al., 1994; Kolaczkowski et al., 1998; Snider et al., 2013), many weak single-knockout effects, and only one weak two-gene interaction between  $pdr5\Delta$  and  $snq2\Delta$  (Figure 3C). Thus, DCGA revealed higher-order genetic interaction involving three or more genes for nearly all drug resistance phenotypes studied.

In total, genetic interactions were found for 14 of the 16 genes that we targeted in our engineered population. Of these 14 genes, 13 were involved in at least one complex interaction involving three or more genes. Remarkably, 11 of the 16 targeted genes were involved in at least one 5-gene interaction. This analysis uncovered strong complex interactions that had been excluded from our initial manual exploration. For example, there were several complex positive interactions in which deleting one or more of  $PDR15$ ,  $BPT1$ ,  $ADP1$  or  $VMR1$  in a drug-sensitive multi-knockout strain background conferred drug resistance (Figure 3C).

Formalizing the identification of complex genetic interactions captured many of the effects that had been readily-apparent by manual examination of the fitness landscapes, while yielding additional effects. For example,  $yor1\Delta$  was found to have no main effect under benomyl, to have a positive genetic interaction with  $pdr5\Delta$  and, surprisingly, to have a negative genetic interaction with  $snq2\Delta$  (Figure 3C, Data S6). In camptothecin,  $pdr5\Delta$  and  $snq2\Delta$  each had a minor individual negative effect on resistance, and a strong negative interaction was observed between them (Figure 3C, Data S6).

# Figure 3



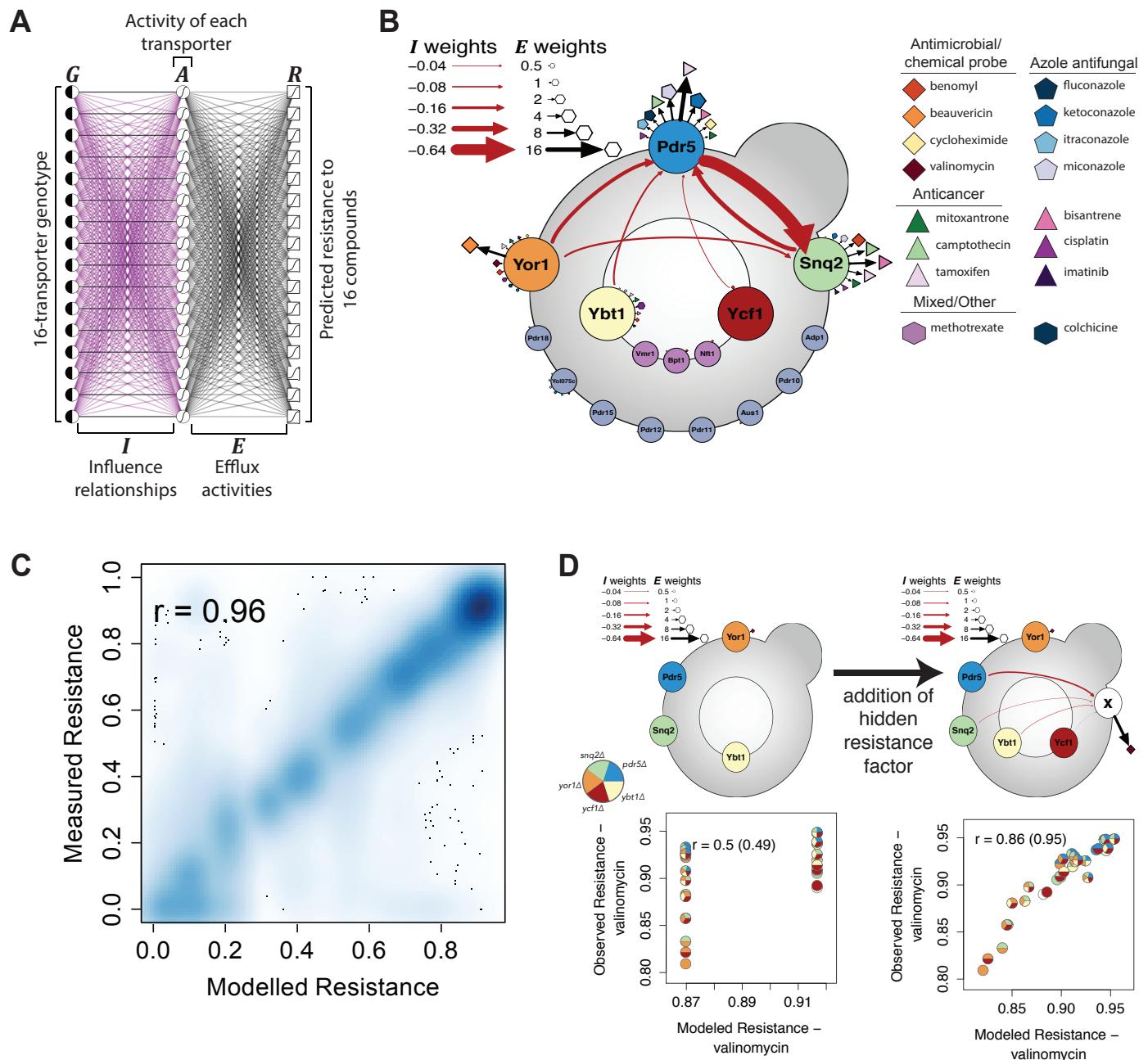
Formal complex genetic interaction analysis allowed finer parsing of the relationship between genes involved in a higher-order interaction. For example, the striking mitoxantrone sensitivity of the *snq2Δ pdr5Δ ybt1Δ yor1Δ* quadruple mutant was modelled as the combination of small marginal effects of *snq2Δ* and *pdr5Δ* alone, a two-gene negative interaction between *snq2Δ* and *pdr5Δ*, two three-gene negative interactions (between *snq2Δ pdr5Δ* and each of *ybt1Δ* and *yor1Δ*), and a four-gene {*snq2Δ, pdr5Δ, ybt1Δ, yor1Δ*} negative interaction (reflecting the fact that the quadruple mutant is more sensitive than would be expected given the observed resistance of any of the three-deletion subset genotypes; Figure 3C, Data S6). Together, these complex negative genetic interaction patterns suggest that the four genes enable mitoxantrone efflux in parallel. A similar ‘parallel action’ genetic interaction pattern was observed for {*pdr5Δ, snq2Δ, yor1Δ*} in cisplatin (Figure 3C, Data S6).

### Objectively modeling the ABC transporter system

The generalized linear models that were trained for each drug resistance phenotype do achieve the important goal of capturing a complex genotype-phenotype relationship, while also describing single-gene effects and genetic interactions. However, these models do not efficiently convey useful intuition about the system. Above, we manually reasoned that the observation of negative genetic interactions amongst a set of transporter genes suggests that each transporter is independently capable of drug efflux. Similarly, the manual application of classical epistasis analysis might lead us to conclude that the presence of one transporter can activate or repress another (either directly or indirectly). However, manually derived intuition from a complex system is laborious, error-prone, and potentially subjective. To more systematically derive intuitive models of the system from complex genotype-phenotype relationships, we developed a neural network model.

We structured a neural network model (Methods, Figure 4A) to have three layers: 1) an input layer encoding the binary genotype ( $\mathbf{G}$ ) for each of the 16 targeted transporters; 2) a middle ‘hidden’ layer with values ( $\mathbf{A}$ ; ranging from 0 to 1) that estimate the activity of each of the 16 transporters; and 3) an output layer that quantitatively describes resistance ( $\mathbf{R}$ ; ranging from 0 to 1) to each of 16 drugs. To represent pairwise regulatory influence relationships between transporters, the links between genotype and activity layers have (initially unknown) weights ( $\mathbf{I}$ ), with positive weights where gene presence increases activity and negative weights where gene presence decreases activity. The links between activity and resistance layers have (also initially unknown) non-negative weights ( $\mathbf{E}$ ) that capture the extent to which each transporter can catalyze the efflux (or otherwise reduce the intracellular activity) of each drug. Using our complete set of drug resistance phenotypes for each genotype as training data, we learned the network weights using back-propagation with stochastic gradient descent (Methods). The cost function that was used to optimize network weights contained a penalty which acts to limit the number of non-zero weights, and has the effect of favoring more parsimonious models (Methods, Figure S8A-B). After the learning procedure, parsimonious models were further favored by setting non-zero weights to zero if they did not consistently depart from zero between repeated runs with different initial parameter settings, or if doing so did not cause a significant difference in model predictions (Methods). Training this model on an input dataset of 97,392 training examples (6,087 unique genotypes  $\times$  16 drugs), we learned an interpretable neural network with only 73 non-zero fitted parameters (6  $I$  weights, 51  $E$  weights, and 16  $E$  bias terms).

# Figure 4



Despite its relatively parsimonious nature, the resulting neural network model largely recapitulated the input data ( $r = 0.96$ , Figure 4C). However, because over-fitted models may exaggerate performance when tested using data that was also used in training, we also generated the model on data from one mating type and tested it on the other. To further ensure independence of these biological replicate datasets, we removed all strains with shared genotypes between the two pools before training. We found similar performance when the model was tested with data that had not been used in training ( $r = 0.95$  and  $r = 0.96$  when using either mating type **a** or **a'** as training, respectively [Figure S8C]). Training using each of these datasets also yielded strong agreement in the parameter values ( $r = 0.98$ , Figure S8D), suggesting that model parameters were robustly determined.

The objectively-trained model provided intuition that was largely in agreement with manual interpretations. For example, the manual genetic interpretation that Pdr5, Snq2, Yor1, and Ybt1 are each independently able to efflux mitoxantrone was also supported by positive  $E$  links connecting each of these transporters to mitoxantrone (Figure 4B). The model showed Snq2 to have the highest mitoxantrone efflux activity ( $E = 2.3$ ) followed by Pdr5, Yor1, and Ybt1 ( $E = 1.9, 0.6, 0.6$ , respectively; Figure 4B, Data S8). These differences were reflected in the fitness landscape: For example, while mitoxantrone resistance of a  $pdr5\Delta ybt1\Delta yor1\Delta$  deletion strain was not significantly different than the wild-type ( $p = 0.25$ ), deletion of genes encoding the two transporters with the highest inferred efflux (Snq2 and Pdr5) yielded a 9% decrease in resistance ( $p = 1.2e-70$ ). The  $I$  weights also pointed to differential inhibitory effects between transporters: For example, Snq2 activity is predicted to be more strongly inhibited by *PDR5* than by *YOR1* ( $I = -0.96$  vs  $-0.39$ , Figure 4B, Data S8). Although this might have been gleaned from the observation that *pdr5* $\Delta$  yields greater benomyl resistance than does *yor1* $\Delta$  (Figure 3A), the neural network model provides a clearer statement of these inferred biological relationships.

### Iterative refinement of genotype-to-phenotype models

While the neural network model was accurate overall, some drugs yielded predictions that departed systematically from observation (Figure S9), suggesting the need for model refinements. For example, valinomycin resistance was quite poorly predicted by the neural network model ( $r = 0.5$ , Figure 4D, left panel). While DCGA showed that deletion of four of the five frequently-associated genes would be more resistant than wild-type (Figure 3A, right panel), the neural network did not predict increased valinomycin resistance for any gene in any background. A previous report of valinomycin resistance upon deletion of all 16 transporters (Suzuki et al., 2011) suggested the possibility that one or more of the targeted 16 transporters inhibits a valinomycin resistance mechanism that is not encoded by any of the 16 targeted transporter genes. To formalize this possibility, we added one additional ‘mystery valinomycin resistance factor’ (always present) and its corresponding activity node to the neural network. Training this extended neural network using only valinomycin data substantially improved recapitulation of the observed phenotypes ( $r = 0.86$ , Figure 3A, right panel) and yielded a model in which *SNQ2*, *PDR5*, *YBT1*, and *YCF1* each negatively influence an unknown valinomycin resistance factor. Moreover, this model improvement depended on the inclusion of this mystery factor (Figure S10A).

### ***SNQ2*, *YBT1*, *YCF1*, and *YOR1* synergistically inhibit Pdr5 fluconazole efflux activity**

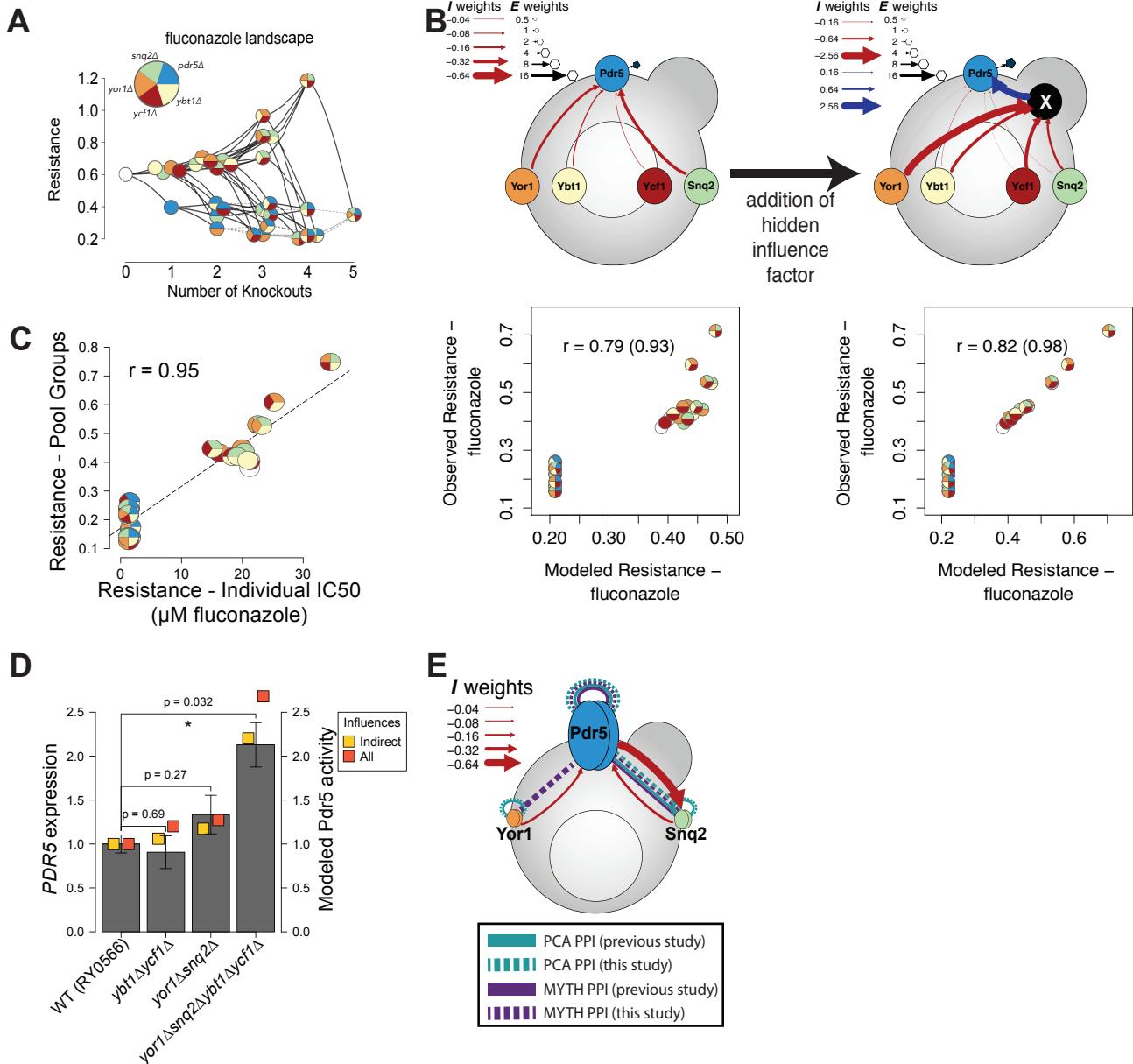
One striking phenotype revealed by DCGA was a quadruple deletion—*snq2Δ ybt1Δ ycf1Δ yor1Δ*—with high resistance to fluconazole (Figure 5A) and ketoconazole (Figure S7). Interestingly, the quintuple mutant *pdr5Δ snq2Δ ybt1Δ ycf1Δ yor1Δ* (adding a *pdr5Δ* deletion to the quadruple mutant background) showed fluconazole sensitivity that was comparable to *pdr5Δ* alone. Beyond one- and two- gene effects, generalized linear regression modeled this phenomenon as the combination of three positive three-gene interactions (all of the three-knockout combinations of {*yor1Δ, snq2Δ, ybt1Δ, ycf1Δ*} except *snq2Δ ybt1Δ ycf1Δ* - Figure 3C). The apparent dependence of the resistance resulting from these multi-knockout combinations on the presence of *PDR5* was modeled as three two-way negative interactions: {*pdr5Δ, snq2Δ*}, {*pdr5Δ, ycf1Δ*}, and {*pdr5Δ, yor1Δ*}. These results are consistent with a previous report that deletions of *SNQ2* or *YOR1* (either alone or together) increase resistance to fluconazole (Kolaczkowska et al., 2008), and extend these findings in at least three ways: 1) fluconazole resistance is increased further by *ybt1Δ* and *ycf1Δ* knockouts in addition to *snq2Δ* and *yor1Δ*; 2) the resistance provided by *ybt1Δ* and *ycf1Δ* is synergistic with that provided by *snq2Δ* and *yor1Δ*; and 3) resistance of the *snq2Δ ybt1Δ ycf1Δ yor1Δ* knockout strain depends on the presence of a wild-type *PDR5*.

Each of these observations is congruent with the genotype-to-phenotype model learned by the neural network. Indeed, the neural network modeled negative influence on Pdr5 from *SNQ2*, *YBT1*, *YCF1*, and *YOR1* (Figure 4B), and thus could be said to capture the phenomenon that *snq2Δybt1Δycf1Δyor1Δ* should be more resistant to fluconazole than strains carrying any subset of these knockouts. However, this neural network under-estimated the resistance of this four-knockout strain, and each of the three triple knockout combinations which were found to have complex positive interactions (Figure 5B left panel). As the complex genetic interactions suggested synergistic rather than additive influence on *PDR5*, we formally modeled this possibility. To model non-additive combinations of influences in a neural network, hidden layers can be added between two nodes. We first considered the simplest extension of the ‘additive influence’ model, adding a single hidden layer with a single neuron (always present) to mediate influence on Pdr5 from *SNQ2*, *YBT1*, *YCF1*, and *YOR1*. Using only fluconazole data, training a network with this single ‘influence mediator’ neuron yielded a better recapitulation of the observed fluconazole resistance for the three- and four-knockout strains (Figure 5B right panel) than training with the original neural network structure (Figure S10 B). These results suggested that at least some of the negative influence of these four genes on Pdr5 activity results indirectly via a hidden factor (Figure 5B right panel).

Before exploring this phenomenon further, we first wished to replicate the initial observations within a single genetic background. Therefore, we generated a single strain for each of the 32 possible combinations of *pdr5Δ*, *snq2Δ*, *yor1Δ*, *ybt1Δ*, and *ycf1Δ* knockouts in a common genetic background (Methods). Fluconazole resistance as estimated by DCGA correlated well with measures of resistance obtained for individual strains —  $r = 0.95$  for the fluconazole concentration expected to yield 50% inhibition (IC50; Figure 5C) and  $r = 0.89$  for total growth in fluconazole relative to no-drug conditions (Figure S11; Methods). Consistent with DCGA results, *snq2Δyor1Δybt1Δycf1Δ* had the highest resistance.

We considered two potential mechanisms by which *SNQ2*, *YOR1*, *YBT1*, and *YCF1* might negatively influence Pdr5 activity: 1) indirect reduction via reduced *PDR5* transcript levels; and 2) direct inhibition of Pdr5 via inhibitory protein interactions.

# Figure 5



In line with a hidden factor mediating negative influence, inhibition of Pdr5 activity by *SNQ2* and *YOR1* was previously reported, explained by reduced activity of the transcription factor Pdr1 via an unknown mechanism, which in turn yields reduced *PDR5* transcript levels (Kolaczkowska et al., 2008). Using qRT-PCR, we found *snq2Δyor1Δ* to have a  $\sim 1.3 \times$  increased *PDR5* mRNA level relative to wild-type. Although this was numerically consistent with the previously-reported  $\sim 1.5 \times$  increase for *snq2Δyor1Δ* (Kolaczkowska et al., 2008), the change did not reach statistical significance in our hands ( $p = 0.27$ , Figure 4E) and the previous report did not contain a statistical test. No evidence of mRNA induction in the *ybt1Δycf1Δ* strain was observed (0.9 fold expression,  $p = 0.69$ , Figure 4E). More persuasively, we found that *PDR5* mRNA levels were significantly higher in *snq2Δyor1Δybt1Δycf1Δ* than the wild type ( $2.1 \times$  increase;  $p = 0.032$ ; Figure 4D). Interestingly, relative expression of *PDR5* tracked well with the relative activity expected by the neural network model, especially when considering only ‘indirect’ influences from the hidden mediating factor (Figure 5D). Although these results can provide only weak evidence for the previous finding that some combination of *SNQ2* and *YOR1* yield reduced *PDR5* mRNA levels, our results more confidently support the idea that the fluconazole resistance *snq2Δyor1Δybt1Δycf1Δ* can be explained (at least in part) by increased *PDR5* transcript levels.

Insight into whether Snq2 might inhibit Pdr5 directly via protein interaction comes from a previous study investigating an Snq2-dependent decrease in benomyl resistance imparted by *PDR5* and *YOR1* found evidence for a modest ( $1.5 \times$ ) *SNQ2* mRNA induction only in *pdr5Δyor1Δ*, but not for either *pdr5Δ* or *yor1Δ*, despite the fact that each of these single mutants showed increased benomyl resistance (Snider et al., 2013). This result suggested the possibility that transporter genes can negatively influence one another by non-transcriptional means, and *PDR5*-mediated repression of Snq2 activity was hypothesized to result at least in part from a direct protein-protein interaction between Pdr5 and Snq2 (Snider et al., 2013). This ‘direct repression’ model, in which heterodimerization of Pdr5 and Snq2 transporters draws subunits away from the homodimeric Snq2 complex and thereby reduces total efflux activity of Snq2, drew support from the observed homodimeric interactions of Pdr5, Snq2, and a heterodimeric interaction between Pdr5 and Snq2. Although the possibility that Snq2 can reciprocally inhibit Pdr5 has not been explored, the possibility of this mechanism is strongly suggested by the previous benomyl resistance study.

We next further investigated whether the observed Pdr5-dependent decrease in fluconazole resistance provided by *YOR1* might also be mediated by direct physical interactions. This model predicted a heterodimeric interaction between Pdr5 and Yor1, and is made more plausible by the fact that Pdr5 and Yor1 are paralogs, and can each form a homodimer (Snider et al., 2013; Tarassov et al., 2008). Because all known protein interaction-testing methods miss the majority of real interactions (Braun et al., 2009), we re-tested for the model-predicted Pdr5-Yor1 interaction using two distinct protein interaction assays: MYTH (Snider et al., 2013) and PCA (Tarassov et al., 2008). All previously-known MYTH and PCA interactions amongst Pdr5, Snq2, and Yor1 were recovered in their corresponding assays (Figure 4F, S11, S12). Although PCA (Figure S11) did not detect the predicted Pdr5-Yor1 interaction, it was revealed by MYTH (Figure 4F, S12). Given a much-higher baseline abundance of Pdr5 than Snq2 (Newman et al., 2006), a ‘heterodimeric repression’ model is consistent with the prediction that negative influence of Snq2 by Pdr5 will be greater than negative influence of Pdr5 by Snq2 ( $I = -0.81$  vs  $-0.25$ , Figure 4B). This is because a greater proportion of Snq2 would be affected by each heterodimeric interaction than would Pdr5.

Taken together, these experiments provide support for two different mechanisms whereby gene deletions can relieve Pdr5 inhibition, one in which deletion of four genes relieves *PDR5* expression, and another in which deletion of *snq2Δ*, *yor1Δ* or both can relieve direct physical inhibition of Pdr5 by Snq2 or Yor1.

## Discussion

We illustrated a method to systematically engineer, profile, and interpret the effects of high-order combinations of genetic perturbations. Applying this method to 16 ABC transporters uncovered phenomena that were not evident when knocking out one or two genes at a time. Computational analysis of the revealed complex genetic relationships was used to derive an intuitive system model of these transporters. It is straightforward to continue profiling this engineered population to understand ABC-transporter-mediated drug resistance of other compounds. Broadly, these results illustrate the potential for carrying out DCGA in other sets of functionally-related genes.

The demonstrated ‘cross-based’ DCGA approach can be used with multi-variant strains in many model organisms. DCGA in yeast could be performed, for example, with an existing 16-deletion mutant for GPCR pathway-related genes (Shaw et al., 2018). Using CRISPR, variation in yeast can be engineered at up to five loci with a single transformation (Jakočiūnas et al., 2015), allowing construction of other multi-variant strains as required. CRISPR-based methods for simultaneous variant engineering at 3 – 6 loci have been described in other model organisms such as mouse (Wang et al., 2013), zebrafish (Jao et al., 2013), *C. elegans* (Xu et al., 2016), and *Arabidopsis* (Zhang et al., 2016). Multiple simultaneous genetic perturbations in mice have been used, for example, to model myeloid malignancies (Cancer Genome Atlas Research Network et al., 2013; Heckl et al., 2014), and induce pluripotent stem cells from adult fibroblasts (Takahashi and Yamanaka, 2006). Mutations can be distributed between two parents, allowing DCGA in cases where they cannot all be introduced into a single individual. Crosses can be made more complex, involving multiple matings between multiple parental strains carrying different subsets of targeted variation, and subsequent inter-crosses between F1 populations as needed.

Direct engineering of trackable large-scale multi-allele diversity into a population presents more challenges than crossing multi-variant parental strains, but technical advances in this area continue to be made (Wong et al., 2016; Zeitoun et al., 2017). For example, pooled approaches to profile two-gene combinatorial mutants in both yeast (Díaz-Mejía et al., 2018) and human cells (Horlbeck et al., 2018; Najm et al., 2017; Shen et al., 2017; Wong et al., 2016) may be expanded to allow higher combinatorial complexity at large scale. In human cells, methods developed for single-cell sequencing also show promise for ‘directly engineered’ DCGA. For example, RNA levels (Adamson et al., 2016; Dixit et al., 2016), and even chromatin state (Rubin et al., 2018) can be profiled in single cells following combinatorial gene disruptions. In addition to permitting richer phenotyping (e.g. DCGA of a transcriptional regulatory system), single-cell barcoding methods may permit more in-depth sampling of the vast combinatorial space – here, we sampled ~8% of 65,536 knockout combinations at 16 genes, which grows to over  $10^7$  combinations if heterozygotes were to be further profiled.

DCGA of yeast ABC transporters suggested future extensions of the targeted genes and considered phenotypes. For example, we hypothesized that some negative influences on Pdr5 may be mediated by a Pdr1-dependent transcriptional response (Kolaczkowska et al., 2008). The pooled

design allows for *en masse* deletion of *PDR1*, thereby expanding the DCGA targets to explicitly test the extent of its mediating role. A DCGA with high-content phenotyping would provide a richer profile of the cellular response to ABC transporter perturbation (Khakhina et al., 2015). For example, while no evidence of a transcriptional response was found when deleting only *YBT1* and *YCF1*, both genes have been shown to catalyze the movement of substrates from the cytosol to the vacuole (Sousa et al., 2015), and might act to competitively sequester Pdr5 efflux products. The use of fluorescence-conjugated drugs (Benhamou et al., 2017), coupled with the ability to image and genotype multi-knockout strains (Emanuel et al., 2017) could be used to test for this at scale. Interestingly, the complex influence between ABC transporters described here is also evident in mammals. For example, ABCC3 increases in expression when ABCC2 is disrupted in Dubin-Johnson Syndrome (Donner and Keppler, 2001; König et al., 1999), and ABCG5/ABCG8 both increase in expression when ABCG2 (a protein that confers breast cancer xenobiotic resistance in humans) is knocked out in mice (Huls et al., 2008). An analogous DCGA of human ABC transporters may permit better understanding of their roles not only in the drug response and chemotherapeutic resistance, but in numerous diseases (Huls et al., 2008).

We showed that computational tools can be used to construct and evaluate genotype-to-phenotype models from complex genetic relationships. This is in contrast to insight extracted from automated epistasis analysis, which often aims to mine genetic interactions for evidence of biologically-ordered pairs rather than construct explanatory models of all observed interactions (Angeles-Albores et al., 2018; Boettcher et al., 2018; Fischer et al., 2015; Phenix et al., 2011; St Onge et al., 2007). In addition to ordered pairs, many genetic interactions can also arise from the effects of genes on unobserved ‘hidden’ variables that mediate the measured phenotype (Otwinowski et al., 2018). For example, we could model that complex negative interactions may occur when deleting multiple ABC transporters that each independently affect an unobserved ‘efflux’ variable. Neural networks offer an expressive way to define genotype-to-phenotype models that can learn directed non-linear relationships amongst potentially vast numbers of biologically-motivated hidden variables (Ma et al., 2018). While even biologically-motivated neural network models can easily grow complex and challenging to decipher, here we showed that they can be carefully constructed using directly-interpretable parameters (e.g. ‘activity’, ‘influence’, and ‘efflux’), and then iteratively expanded as needed. The model presented here can be further expanded, for example to incorporate the several genes with complex positive genetic interactions but no network weights. Further deep combinatorial profiling, as well as richer phenotyping that measures states of variables mediating the measured phenotype rather than relying on indirect inference, can be used to learn more detailed genotype-to-phenotype relationships. In general, future availability of DCGA data can be used to develop rich genotype-to-phenotype models that may help functionally dissect and understand systems in many living organisms.

## Materials and Methods

### Yeast strains

RY0622/GM512 (Green Monster MAT $\alpha$ ):

MAT $\alpha$  *adp1Δ snq2Δ ycf1Δ pdr15Δ yor1Δ vmr1Δ pdr11Δ nft1Δ bpt1Δ ybt1Δ pdr18Δ yol075cΔ aus1Δ pdr5Δ pdr10Δ pdr12Δ can1Δ::GMToolkit-a (CMVpr-rtTA KanMX4 STE2pr-Sp-his5) his3Δ1 leu2Δ0 ura3Δ0 met15Δ0*

RY0146 (Toolkit-a strain):

*MAT $\alpha$  lyp1Δ his3Δ1 leu2Δ0 ura3Δ0 met15Δ0 can1Δ::GMToolkit-a (CMVpr-rtTA KanMX4 STE2pr-Sp-his5)*

RY0148 (Barcoder Strain MAT $\alpha$ ):

*MAT $\alpha$  lyp1Δ his3Δ1 leu2Δ0 ura3Δ0 met15Δ0 can1Δ::GMToolkit-a (CMVpr-rtTA NatMX4 STE3pr-LEU2) hoΔ::loxP UP-tag HphMX4 DN-tag lox2272*

Media

SC (SC-His, SC-Leu, SC-Ura)

YPD (+HygroB, +Clonnat, +G418)

### Creating the barcoder plasmid

We added a barcoder locus flanked by *loxP* and *lox2272* into a pSH47 plasmid backbone expressing GAL1pr-CRE. This barcoder locus consisted of a random 25bp DNA sequence ('UP tag') in between two common primer regions (US1 and US2), followed by a HphMX4 cassette, and another random 25bp DNA sequence ('DN tag') in between two common primer regions (DS1 and DS2).

First, a barcoded HphMX4 construct was created. HphMX4 was amplified from a pIS420 plasmid using the STEP1F and STEP1R primers containing HphMX4 homology and US2/DS1 overhangs (Data S1). The PCR program used for this step was 98°C for 30sec; 25 cycles of 98°C for 10sec, 59°C for 10sec, 72°C for 60sec; 72°C for 5min; 4°C forever. These PCR products were purified using a Qiagen Qiaspin kit and confirmed using 2% gel electrophoresis. To the resulting purified products, the STEP2F and STEP2R primers were used to add the random barcodes and US1/DS2 regions with the following PCR program: 98°C for 30sec; 25 cycles of 98°C for 10sec, 68°C for 10sec, 72°C for 60sec; 72°C for 5min; 4°C forever. These resulting products were again purified using a Qiagen Qiaspin kit and ~1.5-1.6kb products were confirmed using 2% gel electrophoresis. To add *loxP*/*lox2272* sites, PCR was performed with the STEP2 products using the SacI-loxP-HphMX4-Barcode-F / SacI-lox2272-HphMX4-Barcode-R primers. The PCR program used for this step was: 98°C for 30sec; 26 cycles of 98°C for 15sec, 64°C for 20sec, 72°C for 65sec; 72°C for 5min; 4°C forever. The resulting PCR products were purified using a Qiagen Qiaspin Kit, and ~1950bp products were confirmed using 2% gel electrophoresis. Two PCR reactions were performed on the resulting products to confirm correct synthesis. The first PCR reaction was performed with the SacI Reamp F/US2 primer pairs, and the second was performed using DS1/SacI Reamp R primer pairs. The PCR program used for both of these reactions was: 98°C for 30sec; 25 cycles of 98°C for 10sec, 59°C for 15sec, 72°C for 30sec; 72°C for 5min; 4°C forever. Expected sizes (~132bp, 137bp) were confirmed using 4% gel electrophoresis. All above PCR reactions were performed using High Fidelity Phusion Master Mix (NEB).

To prepare for cloning of the barcoder locus, pSH47 was digested with SacI using 100μl of 250ng/μl pSH47, 100μl NEB Buffer 4, 10μl BSA, 10μl SacI-HF in 1ml sterile water. 100μl of this mixture was incubated at 37°C for two hours, and inactivated by incubation at 65°C for 20min. Digest products were purified using a Qiagen Qiaspin kit, and confirmed using 0.8% gel electrophoresis.

### Generating a barcoder strain

A linear URA3 cassette flanked by *loxP* and *lox2272* sites and homology to the *HO* gene was amplified from purified pIS418 with the 5'HO-loxP-URA and URA-lox2272-3'HO primers using the following PCR program: 98°C for 30sec; 25 cycles of 98°C for 10sec, 60°C for 10sec, 72°C for 70sec; 72°C for 5min; 4°C forever. This PCR reaction was performed using High Fidelity Phusion Master Mix (NEB) and was purified using Qiagen Qiaspin. This cassette was integrated into the *HO* locus of the RY0148 strain through transformation to serve as the ‘landing pad’ for barcode integration using an EZ transformation kit. Transformants selected for growth in SC – Ura plates, and were later verified to exhibit no growth in 5-FOA. A transformant was selected to confirm *HO* locus integration using three PCR reactions with the following primer pairs: 5'HO-URAreamp + midURA-5'; 5'HO-URAreamp + midURA-3'; 5'HO-URAreamp + 3'HO-URAreamp. All PCR reactions were performed using High Fidelity Phusion Master Mix (NEB) with the following program: 98°C for 30sec; 25 cycles of 98°C for 10sec, 50°C for 10sec, 72°C for 70sec; 72°C for 5min; 4°C forever. Expected PCR product size was confirmed using 2% gel electrophoresis.

The *HO::loxP-URA3-lox2272* integrant strain was then transformed with a mixture of digested pSH47 and purified PCR products to enable in-yeast-assembly (Gibson et al., 2009). Transformation was carried out using a previously established protocol (Gietz and Schiestl, 2007), with a ~1:6 mixture of digested pSH47:HphMX4 barcode cassette (~12µg digested pSH47 and 15µg cassette). Transformants were grown at 30°C in YPG +HygroB plates for 3 days, allowing both selection of successful transformants and Gal1p-Cre induction. These cells were then scraped and grown overnight in 5-FOA plates to select against non-recombinant strains and strains containing the barcoder plasmids.

Twenty colonies were confirmed to have barcode integration using PCR and Sanger sequencing. Lysates were made by mixing a sample of each colony with 2µl Sterile DNA Free Water, 2µl 0.2M pH 7.4 Sodium Phosphate Buffer, 0.5 µl 5U/µl Zymoresearch zymolyase and incubated at 37°C for 25min and 95°C for 10 min, and stopped by adding 125µl of sterile DNA-free Water. To each lysed colony, two sets of primer pairs to verify the strain barcode-specific UP and DN tag - US2 and a sequence complementary to 5' of the *HO* gene (5'HO); DS1 and a sequence complementary to the 3' of the *HO* gene (3'HO), using the following program: 98°C for 30sec; 25 cycles of 98°C for 10sec, 59°C for 15sec, 72°C for 30sec; 72°C for 5min; 4°C forever. PCR reactions were performed using High Fidelity Phusion Master Mix (NEB) and analyzed using 4% gel electrophoresis to verify the presence of 263bp and 251bp bands. EXOSAP purification was performed on the PCR products by adding 10µl EXOSAP mix (0.025µl ExoI (0.5U), 0.1µl Antarctic Phosphatase (0.5U), 3.5µl 10X Antarctic Phosphatase Buffer, 6.375µl dH<sub>2</sub>O) to 25µl of PCR products and incubating at 37°C for 30min; 80°C for 20min, then diluting with 35µl of DNA-free H<sub>2</sub>O to stop the reaction. Diluted EXOSAP products were Sanger sequenced with the 5'HO seq and 3'HO seq primers to confirm the correct barcode construct.

### **Creating a ‘gold standard’ genotyped set**

To create a ‘Gold Standard’ genotyped set, 40 progeny strains (19 MAT $\alpha$  and 21 MAT $\alpha$ ) were subject to individual strain genotyping. For these 40 strains, and for an RY0148 isolate, the strain-specific UP and DN tags were also PCR amplified using two sets of primers and subject to Sanger sequencing as above.

To genotype each strain at the 16 ABC transporter loci, two PCR reactions were performed for each locus - one to determine the presence of a GFP integration cassette, and another to determine the presence of the wild type gene, as previously described (Suzuki et al., 2011). For the cassette confirmation reactions, locus-specific PCR primers from the 5' flanking sequences of each gene were paired with a common primer complementary to the *GFP* cassette (Data S2). Gene presence confirmation primers were designed individually for each gene (Data S2). PCR reactions were performed with a Platinum HiFi mix using the following program: 94°C for 2min; 34 cycles of 94°C for 30sec, 55°C for 30sec, 68°C for 60sec; 68°C for 10min; 4°C forever. PCR products were analyzed using gel electrophoresis.

### **Generating barcoded random knockout progeny**

Mating, sporulation, and haploid selection was performed between the RY0622 ‘Green Monster’ strain (**MAT $\alpha$** ) and the RY0148 barcoder strain (**MAT $\alpha$** ) as previously described (Suzuki et al., 2011), selecting for **MAT $\alpha$**  and **MAT $\alpha$**  progeny separately. The two pools were then grown in YPD +HygroB to select for barcoded haploids. The SC–Leu pool was further grown in SC–Ura to select against barcoder strain parents that may have escaped diploid selection. Using a QPix colony picker, 5,461 **MAT $\alpha$**  and 5,461 **MAT $\alpha$**  colonies were picked onto 384 well plates. In addition, 299 known positions in both the **MAT $\alpha$**  and **MAT $\alpha$**  arrayed collections consisted of known strains – either one of 40 ‘Gold Standard’ genotyped strains, RY0148, or RY0622 – to act as genotyping controls (Data S2).

### **Pooled strain genotyping**

A previously-developed Row-Column-Plate (RCP)-PCR protocol (Yachie et al., 2016) was adapted in order to perform *en-masse* genotyping of the random knockout progeny using high throughput sequencing. This protocol first uniquely tags PCR products originating from the same well on a given plate, by the use of a 5' tag encoding the well row (R) in forward primers, a 5' tag encoding the well column (C) in the reverse primers (Yachie et al., 2016). Additionally, these primers contain a linker sequence (PS1 or PS2) which primes a second reaction encoding the plate of origin (Data S2).

For each well in the collection, lysates were made on a new set of plates. 4  $\mu$ L of overnight yeast culture was mixed with 8  $\mu$ L 0.2 M sodium phosphate buffer (pH 7.4), 4  $\mu$ L DNA free dH<sub>2</sub>O, 0.05  $\mu$ L 5 U/ $\mu$ L Zymoresearch zymolyase and incubated at 37 °C for 35 minutes. 64  $\mu$ L DNA free dH<sub>2</sub>O was added to each well to prepare PCR template.

Four ‘Row-Column’ PCR reactions were performed on the lysates with the following primer pairs: PS1+R+U1 and PS2+C+U2 to amplify DNA barcodes encoding the UP tags for each gene deletion; PS1+R+D1 and PS2+C+D2 to amplify the deletion-specific DN tags; PS1+R+US1 and PS2+C+US2 to amplify the strain-specific UP tag; PS1+R+DS1 and PS2+C+DS2 to amplify the strain-specific DN tag (Data S2). PCR reactions were performed with 2  $\mu$ L of lysed colonies using a Hydrocycler with the following program: 95 °C for 5 min; 23 cycles of 95 °C for 60 sec, 57 °C for 35 sec, 72 °C for 45 sec; 72 °C for 2 min; 4 °C forever. Row-Column PCR products from each plate were pooled and size was verified on a 4% agarose gel. PCR products from each plate were pooled and 260  $\mu$ L was purified using a Qiagen Qiaquik Spin kit. DNA yield was quantified using a Nanoquant. From the resulting products from each plate, Illumina adapters containing plate tags were added using an additional PCR reaction as previously described (Yachie et al., 2016). A pair

of PXX\_PE1.0 and PYY\_PE2.0 primers (Data S2) were added to 3-6 µl pooled products (calibrated to ~150 ng) from each plate to encode the plate of origin, and were amplified using the following PCR program: 98 °C for 30 sec; 15 cycles of 98 °C for 10 sec, 59 °C for 15 sec, 72 °C for 40 sec; 72 °C for 2 min; 4 °C forever. All PCR reactions above were performed using High Fidelity Phusion Master Mix (NEB).

Expected product size from the plate tags was confirmed on 4% agarose gel. PCR products were purified using a Qiagen Qiaquick Spin kit. qPCR was performed on all plate tag PCR products using a light cycler and KAPA Illumina sequencing quantification kit. qPCR results were used to pool approximately equal amounts of all samples, and 100 µl of this multiplexed sample were run on a 4% gel. Products of the desired size (260-290 bp) were isolated from each lane, and purified using a Qiagen gel purify kit and another qPCR was run on the purified sample.

### **Analysis of pooled strain genotyping data**

Pooled strain genotyping PCR products were sequenced using an Illumina HiSeq, and the reads were demultiplexed into individual samples corresponding to a plate and well of origin using a Perl script.

For each sample, a genotype calling pipeline determined the strain-specific tag sequences and genotype from the reads. The parameters of this pipeline were trained based on known reference strains. Cross-validated accuracy for each gene is reported in Figure S2A.

UP or DN tag identity and a corresponding genotype was successfully determined for 7,195 samples. For 7,030 samples, the UP or DN tag was unique, and for 165 samples, both the UP and DN tag sequences were redundant with another sample where the called genotype was isogenic or highly similar ( $\leq 2$  differences), indicating the presence of a single strain in multiple wells. If a single strain existed in multiple wells, one of the matching identifiers was randomly assigned to represent each unique combination of UP and DN tag sequences.

### **Examining putative wild-type pool strains**

For 73 MAT $\alpha$  and 131 MAT $\alpha$  strains, pooled sequencing analysis had called the genotype as wild-type. Many of these strains were isolated and tested for the presence of one or more gene knockout cassettes by growth in SC-Ura. Out of 96 MAT $\alpha$  strains, 74 exhibited no detectable growth in SC-Ura, and likely arose from remaining barcoder parents which had escaped a previous SC-Ura selection step. The genotypes for these 74 strains were kept as is, while the other 23 strains, as well as 46 untested strains were discarded from the analysis. Out of 45 MAT $\alpha$  strains, all exhibited growth in SC-Ura. Individual genotyping was performed for these MAT $\alpha$  strains, and was successful for 40 of 45 strains, confirming the lack of true wild types. These strains had their genotype corrected (Data S2). The 5 unsuccessfully genotyped strains, as well as 28 additional strains were discarded from analysis. When calculating linkage and distribution of gene knockouts (Figure S2), the wild-type MAT $\alpha$  strains were excluded from analysis as they were likely parental strains rather than progeny arising from mating.

### **Estimating genotyping accuracy by knockout distribution**

To lend independent support to the genotyping accuracy determined by gold standard strains, an alternate method based on the distribution of knockouts in the population was used. Since *en*

*masse* genotyping associates barcode sequences with ABC transporter knockouts, the absence of a given barcode implies either a wild-type genotype at that locus or a failure in amplification, sequencing, or calling. Conversely, cases where a wild-type is called as a mutant are expected to be comparably rare. Excess wild-type calls lead to a reduction in the average number of knockouts in the pool, and can be used to estimate genotyping accuracy. The average number of knockouts in the pool was 7.0, lower than the 8 expected with perfect genotyping. If there are no wild-type to mutant miscalls, this number is most likely with an ‘asymmetric’ genotyping accuracy of 93.8%, compared to the 93.2% estimated by comparison to gold standards (Figure S2C).

### **Individual liquid growth profiling**

To measure individual strain growth, the a 0.0625 OD<sub>600 nm</sub> starting culture was measured in the appropriate medium every 15 mins using a GENios microplate reader (Tecan).

### **Drug testing for growth inhibition**

The effects of 16 different drugs on strain growth were tested to find a concentration which inhibits wild type growth by approximately 20% (Data S3). All drugs used were dissolved in 2% DMSO, which was used as a solvent control. Here, growth was determined by the Average\_G metric (Proctor et al., 2011), which represents the average generation time.

### **Population growth profiling by high-throughput sequencing**

Progeny with at least one mapped strain-specific barcode (Data S2) were combined into two separate liquid YPD + glycerol pools separated by mating type, and kept at -80°C. Samples from the original YPD + glycerol pool were thawed and added to the appropriate drug or solvent containing medium at a final concentration of 0.0625 OD<sub>600 nm</sub> in 10ml. For the solvent control, a 0 generation sample was immediately harvested for sequencing. After growth to approximately 2 OD<sub>600 nm</sub>, a sample was taken from each drug for sequencing and cells were resuspended in fresh medium to a final concentration of 0.0625 OD<sub>600 nm</sub>. This process was repeated until 4 generations of samples were collected. Collected samples corresponded approximately to 5, 10, 15, and 20 generations of growth. Harvested samples were subject to genomic DNA extraction using a YeaStar™ Genomic DNA Kit, quantified using a Qubit® 2.0 fluorometer, and diluted to a final concentration of 20 ng/µl. Approximately 350 ng of isolated DNA was extracted from each sample and added to 20 µl of 2x Platinum PCR SuperMix High Fidelity, 1 µL of 10 µM F primer, and 1 µl of 10 µM R primer. F and R primer pairs were PXX+US1/PYY+US2 and PXX+DS1/PYY+DS2 for the strain-specific UP and DN tag, respectively. PXX and PYY correspond to sequences containing plate-specific Illumina sequencing adapters, as well as tags which were used to demultiplex the samples (See Data S2). PCR products amplified using the following PCR program: 98 °C for 30 sec; 24 cycles of 98 °C for 10 sec, 60 °C for 10 sec, 72 °C for 1 min; 72 °C for 5 min; 4 °C forever. PCR products were subject to gel electrophoresis, and ~210 bp bands were isolated, purified and subject to quantification using a KAPA qPCR assay kit. Samples were pooled to yield approximately equal amounts of DNA, and subject to sequencing using an Illumina NextSeq500 Mid Output kit.

### **Sequence data processing**

Paired-end Illumina sequencing data were first de-multiplexed using a custom Python script which searches for an exact match to the tag regions of the PXX and PYY primers within each pair of

reads. For each strain in each de-multiplexed sample (corresponding to a combination of mating type, timepoint, and drug), strain identification is attempted. To perform this identification, a search is performed for all barcodes matching the sample mating type. If an exact match is not found, up to two ungapped mismatches are permitted to assign a putative strain identity, which is then accepted if there are at least 2 additional mismatches separating this identity with the next closest match (e.g. if 2 mismatches are present with the closest match, then the next closest match must have 4 or more mismatches). This process was performed for both the forward and reverse reads (corresponding to the UP and DN tags) for each strain, and potential cases where the putative strain identity differed between tags were discarded. All samples with less than 200,000 reads in either the UP or DN tag were discarded from the analysis.

### Defining a resistance metric

Following processing of the sequence data, a count  $c$  was assigned for each strain  $s_x$  in a pool under drug  $d$  sequenced at time  $t$  ( $c_{s_x,d,t}$ ). The counts in each sample were then converted to a frequency  $f_{s_x,d,t}$  by division with the total count for all strains in that sample:

$$f_{s_x,d,t} = \frac{c_{s_x,d,t}}{\sum_{i=1}^n c_{s_i,d,t}}$$

If both an UP and DN tag for a given strain were successfully linked to a genotype,  $f_{s_x,d,t}$  is estimated from the UP and DN counts and averaged, otherwise the available tag is used. The frequency of each strain was then converted into a ‘area under the growth curve (AUC) by first multiplying the frequency at each time point by the expected overall pool growth at that time ( $2^t$ , since  $t$  is defined by the number of generations) to estimate the individual abundance over time of each strain, then taking the integral over all measured timepoints 0 to  $T$  (the total number of pool generations measured). Frequencies between measured timepoints were linearly interpolated.

$$AUC_{s_x,d} = \int_0^T f_{s_x,d,t} 2^t dt$$

We modelled each strain as growing constantly from an initial abundance ( $A_0$ ) by an exponential growth rate  $g$  in each drug over time  $t$ , such that:

$$A_{s_x,d,t} = A_{s_x,d,0} 2^{g_{s_x,d} t}$$

In this constant exponential growth model, integrating  $A_t$  over all time points results in the following relationship with growth rate:

$$\int_0^T A_{s_x,d,0} 2^{gt} dt = A_{s_x,d,0} \frac{g_{s_x,d} T - 1}{g_{s_x,d} T \log(2)}$$

We substitute  $AUC_{s_x,d}$  for  $\int_0^T A_{s_x,d,0} 2^{gt} dt$  and  $f_{s_x,d,0}$  for  $A_{s_x,d,0}$ . We then numerically solve for the  $g_{s_x,d}$  which satisfies this relationship using the `optimize()` function in R, setting a minimum of 0 and maximum of 10 for the interval. To obtain the resistance for each strain in each drug ( $r_{s_x,d}$ ),  $g_{s_x,d}$  is divided by growth in the DMSO control ( $g_{s_x,DMSO}$ ):

$$r_{s_x,d} = \frac{g_{s_x,d}}{g_{s_x,DMSO}}$$

We note that experimental uncertainty in the collected generation times  $t$  can introduce biases in the estimation of  $r_{s_x,d}$ , such that resistance estimates from the MAT $\alpha$  and MAT $\alpha$  pool can be highly correlated, but may differ in range in some drugs (Figure S5). To avoid un-necessary batch effects in  $r_{s_x,d}$  estimated from the MAT $\alpha$  and MAT $\alpha$  pools, we use the line of best fit derived in Figure S5 to rescale  $r_{s_x,d}$  estimates from the MAT $\alpha$  pool to match the MAT $\alpha$  pool before merging.

### Finding complex genetic interactions using a general linear model

The multiplicative model of genetic interactions (Mani et al., 2008) was applied to the  $r$  metric. In this model, the expected resistance of a double knockout strain  $x\Delta y\Delta$  in a given drug ( $\hat{r}_{x\Delta y\Delta,d}$ ) is the product of the resistances of the corresponding single knockout strains:

$$1) \quad \hat{r}_{x\Delta y\Delta,d} = r_{x\Delta,d} r_{y\Delta,d}$$

To express this model equivalently in an additive form, we can state this relationship as an exponentiated sum of the log-resistances of the single knockouts -  $\log(r_{s_i,d}) = l_{s_i,d}$ , so that:

$$2) \quad \hat{r}_{x\Delta y\Delta,d} = \exp(l_{x\Delta,d} + l_{y\Delta,d})$$

We defined a two-gene interaction term  $\varepsilon_{x\Delta y\Delta,d}$  as the log-ratio of the observed fitness to the fitness expected by single-gene effects, rather than the traditional linear difference from a multiplicative estimate.

$$3) \quad \varepsilon_{x\Delta y\Delta,d} \equiv \log\left(\frac{r_{x\Delta y\Delta,d}}{\hat{r}_{x\Delta y\Delta,d}}\right)$$

This interaction term can be added to 2) to express the observed rather predicted double mutant fitness:

$$4) \quad r_{x\Delta y\Delta,d} = \exp(l_{x\Delta,d} + l_{y\Delta,d} + \varepsilon_{x\Delta y\Delta,d})$$

When modelling the expected triple mutant fitness, all relevant two-gene interaction terms are added as such:

$$5) \quad \hat{r}_{x\Delta y\Delta z\Delta,d} = \exp(l_{x\Delta,d} + l_{y\Delta,d} + l_{z\Delta,d} + \varepsilon_{x\Delta y\Delta,d} + \varepsilon_{x\Delta z\Delta,d} + \varepsilon_{y\Delta z\Delta,d})$$

Similarly, a three gene interaction term is the deviation from the one- and two- gene expectation:

$$6) \quad \varepsilon_{x\Delta y\Delta z\Delta,d} \equiv \log\left(\frac{r_{x\Delta y\Delta z\Delta,d}}{\hat{r}_{x\Delta y\Delta z\Delta,d}}\right)$$

This definition can be extended analogously for interactions of arbitrary complexity, with  $\varepsilon$  terms denoting interactions between the corresponding knockouts. Specifically, in each drug we fit a general linear model which aims to predict the fitness of each given its knockout genotype  $\Delta G$ , which consists of a subset of ABC transporter knockouts  $\{ABC1_\Delta \dots ABC16_\Delta\}$ :

$$7) \hat{r}_{\Delta G | \Delta G \subseteq \{ABC1_{\Delta} \dots ABC16_{\Delta}\}} = \exp(\sum_{i \in \Delta G} l_i + \sum_{j \subseteq \Delta G | |j| \geq 2} \varepsilon_j + c)$$

To train this model,  $\Delta G$  is encoded as a set of 16 binary variables, where 0 represents a wild-type and 1 represents a knockout at a given gene. Therefore, to predict phenotype from  $\Delta G$ , the relevant  $l_i$  coefficients are added only if the corresponding gene  $i$  is knocked out, and the  $\varepsilon_j$  coefficients are added only if all the genes in subset  $j$  are knocked out. For each drug, we fit this model using the `glm()` function in R, with  $\varepsilon$  terms to a chosen level of complexity.

To perform the marginal association in Figure S4, we fit a model with only  $l_i$  terms, and performed stepwise feature elimination (eliminating the gene with the highest p-value at each step) until all included terms had a significance level of  $p \leq 0.05/16$ . Linear model term significance was tested using the Type III Sums of Squares ANOVA implementation given in the `car` package in R. The same method was used to perform the marginal association in Figure S3, substituting  $g$  for  $r$ .

We expanded this approach to train models containing  $\varepsilon$  terms of up to  $n$ -way complexity using a “stepwise search” approach. First, we use the marginal association procedure above to initialize the model at  $n = 1$ . Then,  $n$  is incremented by 1, and all possible  $n$ -way interactions between the genes contained in the existing (i.e.  $n - 1$ ) model are added as additional  $\varepsilon_j$  features. Each term in this proposed “ $n$ -way” model is tested for significance using Type III Sums of Squares ANOVA, those with  $p \geq 0.05$  are discarded, and the model is updated. This “stepwise addition” procedure is repeated until either  $n$  reaches 5, or the number of genes in the  $n - 1$  model is less than  $n$  (i.e. there are no more possible interaction terms). After the stepwise addition procedure is finished, the remaining terms are more rigorously tested for statistical significance by performing stepwise feature elimination (as in the marginal association procedure) until all included terms have a significance level of  $p \leq 0.05/k$ , where  $k$  is the number of all possible 1-5 gene combinations amongst the marginally associated genes.

### Defining a non-linear system model

We will define an ‘efflux and compensatory activation’ schematic of ABC transporter function which we will later fit as a neural network. First, we normalize resistance data in each drug by dividing with the maximum observed resistance in that drug:

$$r_{norm_d} \equiv \frac{r_d}{\max(r_d)}$$

We then model a sigmoidal relationship between drug concentration and normalized resistance:

$$\hat{r}_{norm_d} = \frac{1}{1 + e^{k[d] - a}}$$

Here  $[d]$  is the concentration of the given drug, and  $k, a$  are unknown constants which define the dose-response curve (such that  $\frac{a}{k}$  yields the expected IC50). In addition, we model each transporter as encoding a resistance factor which acts to additively lower the effective concentration of a drug (for example, by efflux out of the cell):

$$\hat{r}_{norm_{G,d}} = \frac{1}{1 + e^{k[d] - a - \sum_{i \in G} C_{i,d}}}$$

Here,  $G$  is the set of ABC transporters present in a genotype:  $\{ABC1^+ \dots ABC16^+\}$ , and  $C_{i,d}$  is the clearance coefficient of a given ABC transporter for a given drug (i.e.  $C_{i,d} = k[\Delta d_i]$ ). Importantly, a dose response curve in this form can be expressed as the activation of a sigmoid neuron, where  $k[d] - a$  is collapsed into a single bias term  $B$ , and  $C_{i,d}$  are the weights learned as inputs to this neuron from the ABC transporters. As each transporter must act to lower effective drug concentration in this model, we constrain  $C_{i,d}$  to be positive.

We then model compensatory activation between ABC transporters. To do this, we first decompose the clearance coefficient of each ABC transporter  $C_{i,d}$ . That is, each ABC transporter is also given a degree of activity (a value between 0 and 1) which depends on the genotype -  $A_{i,G}$ . This activation variable is modelled as being dependent on genotype  $G$ , but not the drug  $d$ . In this extension,  $C_{i,d}$  is the product of  $A_{i,G}$  and  $E_{i,d}$ , a ‘maximal’ efflux/clearance capacity of a given transporter for a given drug ( $C_{i,d} = A_{i,G}E_{i,d}$ ):

$$\hat{r}_{norm_{G,d}} = \frac{1}{1 + e^{-\sum_{i \in G} A_{i,G}E_{i,d} - B}}$$

We then allow  $A_{i,G}$  to capture compensatory activation. That is,  $A_{i,G}$  can be influenced by other ABC transporters:

$$A_{i,G} = f(\sum_{j \in G, j \neq i} I_j)$$

Where  $I_j$  are the ‘influences’ from other ABC transporters. While the form of  $A_{i,G}$  may depend on the inhibition mechanism, here we also modelled it as having a sigmoidal form for simplicity:

$$A_{i,G} = \frac{1}{1 + e^{-\sum_{j \in G, j \neq i} I_j - B}}$$

### Learning a non-linear system model as a neural network

To create the above model and learn the  $I$ ,  $E$ , and  $A$  parameters from our data, we used the keras library in R to construct a neural network of the appropriate form.

We first provide the genotype of each strain as the input layer to the neural network by encoding  $G$  in binary form. That is, we create an input layer of length 16, where each input value will be either 1 for ABC transporter presence, or 0 for a knockout for each of  $\{ABC1 \dots ABC16\}$ .

We then provide a second layer of length 16 to keras, where the weights from the input layer to the second layer encode the influence weights from and to each transporter-transporter pair  $i - j$ , ( $I_{i-j}$ ), and the second layer acts to compute the activity state  $A$  for each transporter. Specifically, we create a second sigmoid layer of length 16, and connect each transporter in the first layer to each transporter  $j$  in the second layer, except where  $i = j$ , as a transporter cannot inhibit itself in this model. The activity state  $A$  for each transporter is then computed by the neurons in the second

layer in this network from their inbound influence connection  $I_j$ , and a learned bias term  $b_j$ . Notably, the neural network multiplies each outgoing inhibitory connection  $I_i$  by its corresponding genotype value in  $G$ , such that all outgoing inhibitory weights from transporter  $i$  are set to 0 if it is knocked out. To analogously set the activation state of each transporter in the second layer  $A_j$  to 0 if it is knocked out, each neuron in the second layer is then multiplied element-wise by its corresponding value in  $G$  using the `layer_multiply()` function.

To encode the efflux weights for each transporter-drug pair  $E_{i-d}$ , we then added another sigmoidal layer of length 16, which was fully connected to the genotype-multiplied second layer. The `kernel_constraint` argument was used with this layer to ensure that only positive  $E$  parameters are learned. Each neuron in this third layer predicts the normalized resistance to each compound  $\hat{r}_{norm_{G,d}}$  by multiplying the activation state of each transporter  $A_j$  with the learned efflux weights  $E_{i-d}$  to compute the clearance coefficients  $C_{i,d}$  for each compound-transporter pair, and furthermore learns a bias term which defines the shape of the dose-response curve.

In addition to the above schematic, L1 regularization with coefficient  $\lambda$  was added to both the  $I$  weights and the bias term which defines  $A$  for each transporter, using the `kernel_regularizer` and `bias_regularizer` parameters in the second layer. Regularization on  $I$  achieves sparsity in their weights, as it is otherwise possible, for example, to add  $I$  to a transporter which has no  $E$  weights, thus learning  $I$  parameters which are not supported by any phenotypes. Because the clearance coefficient of each gene for each drug is defined by a product  $C = AE$ , regularization of the bias term acts to keep  $A$  close to 0.5, effectively setting a prior on  $A$ . This prior on  $A$  avoids parameterizing  $C \approx 0$  by setting a large bias such that  $A \approx 0$ , which then allows  $E$  weights to be added to transporters without affecting phenotype predictions. Thus, regularization of  $A$  indirectly enforces sparsity in the  $E$  parameters, as each  $E$  directly impacts resistance predictions when  $A$  is not close to 0. While more complex regularization schemes can potentially impose three separate regularization weights for the  $I$  terms, the bias on  $A$ , and the  $E$  terms, here we found that using a single weight for regularizing both  $I$  and the bias for  $A$  without any further regularization to the  $E$  terms was sufficient for learning a sparse predictive model.

The neural network model was compiled with the mean-squared error ('mse') loss function, using the adam optimizer with a learning rate of 0.05. Training was performed for 10,000 epochs, using a batch size of 1,000 and 10% split between training and validation (`validation_split = 0.1`). Model initialization and training was repeated 10 times, and the weights to the final model were set to the mean weights learned from these 10 iterations. In addition, standard deviation was calculated between these 10 iterations, and an absolute Z score was computed for each parameter:

$$|Z_{param}| = \frac{|\mu_{param}|}{\sigma_{param}}$$

Given the non-deterministic nature of the algorithm, we wanted to confidently ensure that non-zero parameters are not a result of stochastic noise, and therefore non-zero weights with  $|Z_{param}| < 4$  were set to 0.

We searched for an appropriate regularization rate by performing the above training and averaging procedure using a range of rates from  $10^{-6}$  to  $10^{-1}$ . We first searched 13 intervals between  $10^{-6}$

to  $10^0$  (Figure S8A-B). After observing high mean-squared error (MSE) and a lack of reproducible parameters at regularization rates below  $\sim 10^{-4.5}$  (Figure S8B) and a smaller ‘jump’ in MSE around  $\sim 10^{-3}$  (Figure S8 A), we searched another 11 intervals between  $10^{-4}$  to  $10^{-3}$  (Figure S8 A-B). We then chose a regularization rate of  $5 \times 10^{-4}$ , as any rate higher than this resulted in a jump in mean-squared error in both the MAT $\alpha$  and MAT $\alpha$  pools (Figure S8B), while lowering this rate did not have a clear mean-squared error impact but increased the number of non-zero parameters (Figure S8A).

After using the training and averaging procedure to learn model weights, we tested each non-zero weight for predictive support. First, we compute the vector of squared residuals in the initial learned model over  $i$  strains and  $j$  drugs, given the set of  $k$  initial non-zero weights  $W_{\{1\dots k\}}$ :

$$(\varepsilon_{initial})^2 = \left( r_{norm_{G_{\{1\dots i\}}, d_{\{1\dots j\}}}} - \hat{r}_{norm_{G_{\{1\dots i\}}, d_{\{1\dots j\}}}} \mid W_{\{1\dots k\}} \right)^2$$

Then, for each  $l \in \{1\dots k\}$ , we set  $W_l := 0$ , and compute the squared residuals in the proposed reduced model:

$$(\varepsilon_{reduced})^2 = \left( r_{norm_{G_{\{1\dots i\}}, d_{\{1\dots j\}}}} - \hat{r}_{norm_{G_{\{1\dots i\}}, d_{\{1\dots j\}}}} \mid W_{\{1\dots (l-1), (l+1)\dots k\}} \right)^2$$

Considering only data where settings  $W_l$  to 0 made a predictive difference ( $\varepsilon_{initial} \neq \varepsilon_{reduced}$  at a numerical tolerance of  $10^{-4}$ ), we then compute the paired Mann-Whitney U statistic between  $(\varepsilon_{initial})^2$  and  $(\varepsilon_{reduced})^2$  to derive a p-value for  $l$ , and keep all features with  $p < 0.05/k$  in the final model.

### Extensions to the Neural Network

For all training in single drugs, learning rate was decreased to 0.01, as we found this to lower variance in parameter values between different initializations.

To extend the neural network for valinomycin (Figure 4D), we added an extra ‘X’ variable to the neural network input to the and set its value to 1 for each strain.

To extend the neural network for fluconazole (Figure 5B), we performed several additional steps. First, for each transporter in the  $A$  layer, we added a single extra ‘always-present’ neuron ( $A'$ ) which provides an additional indirect connection between the  $G$  and  $A$  layers. Thus, while each transporter retains original influence connections between  $G$  and  $A$ , additional  $G$  to  $A'$  influence connections were added. Each  $A'$  node is then connected only to its corresponding transporter in  $A$  (i.e. each indirect node can only influence a single transporter). As in the original network, we omitted both direct and indirect connections between a transporter and itself. These additional influence connections, as well as the bias on  $A'$ , were also subject to L1 regularization with rate  $\lambda$ .

As this three-layer network presented additional training challenges compared to the original two-layer model (e.g. more than twice as many potential parameters), we made additional *a priori* modeling decisions for simplification. First, based on the two-layer results, we restricted  $G$  to

encode only the presence of *PDR5*, *SNQ2*, *YBT1*, *YCF1*, and *YOR1*, and restricted the *A* layer to encode only the presence of *PDR5*. This reduced the number of potential parameters from 545 to 13. Second, while parameters were reproducible between most initializations, we noticed that a small number of runs converged to a different set of parameters with a higher mean-squared error even at high regularization rates (data not shown). Therefore, to increase robustness we used the median between 10 runs rather than the average to parameterize the final neural network, and did not employ the  $|Z_{param}|$  filter described above. We performed a separate  $\lambda$  search for this network, searching 13 intervals between  $10^{-6}$  to  $10^0$  (Figure S10C). We found that a rate of up to  $10^{-5}$  did not negatively impact MSE (Figure S10C) and used this  $\lambda$  to train the network in Figure 5B.

### Targeted mating and selection to obtain 32 knockouts

The TWAS21230902 strain (*pdr10Δ pdr18Δ pdr5Δ snq2Δ ybt1Δ ycf1Δ yor1Δ*; Data S2) was subject to individual strain genotyping, confirming the genotype generated using the RCP-PCR based method. This strain (MAT $\alpha$ ) was mated with RY0146 (MAT $\alpha$ ), and was subject to sporulation and MAT $\alpha$  haploid selection (Suzuki et al., 2011). Individuals from this cross were arrayed onto a 384 well plate, and individually genotyped at *PDR10* and *PDR18*. Strains with no deletions at these genes were further genotyped at *PDR5*, *SNQ2*, *YBT1*, *YCF1*, and *YOR1*. PCR reactions for individual genotyping of these progeny used the Qiagen Mix with the following program: 95°C for 5min; 34 cycles of 95°C for 30sec, 57°C for 30sec, 72°C for 30sec; 68°C for 10min; 4°C forever. After analysis of genotyping results, one strain of each genotype combination was chosen to create the 32-strain collection. These chosen 32 strains were again individually genotyped at these 5 loci for validation.

### Analysis of Liquid Growth Data

Individual strains with 32 knockout combinations at *PDR5*, *SNQ2*, *YBT1*, *YCF1*, and *YOR1* were each grown in fluconazole at concentrations of 1.3, 1.9, 3.9, 7.8, 15.6, 23.4, 31.2, 35 and 40 $\mu$ M. Each genotype was grown an average of 2.7 times in each concentration (Data SXX). For each growth experiment, a culture was started at 2% DMSO at the same time to act as a solvent control. Each culture was started at an initial cell concentration of 0.0625 OD600. OD600 was measured every 10 minutes using a Tecan plate reader for a minimum of 20 hours. To calculate resistance, we divided the OD measured in the drug by the OD measured in the solvent at the time which the culture first saturated in the solvent. To automatically determine a saturation timepoint, we took the second derivative of the growth curve (using a window size of 4 tecan measurements to calculate the first derivative) and determined the time which it is maximized. Automatically determined saturation times were checked visually. Multiple replicates were averaged to yield the values in Figure S11. To determine the fitted IC50 values in Figure 4D, averaged resistance values were linearly interpolated between measured concentrations.

### MYTH testing of protein-protein interactions

*PDR5*, *YOR1*, and *SNQ2* were cloned into the L2 AMBV MYTH bait vector to add a Cub-LexA-VP16 MYTH tag as previously described<sup>25</sup>. A previously-cloned artificial MYTH-tagged bait plasmid was retrieved, and acted as a negative interaction control. NubG-PDR5 (PDR5 prey) and NubI-PDR5 (PDR5 positive interaction control) strains were retrieved from a previously constructed genomic prey library<sup>25</sup>. Previously-constructed Ost1p-NubG (negative interaction

control) and Ost1p-NubI (positive interaction control) strains were also retrieved. All prey-bait combinations were obtained using individual transformations and selected for growth in SD –Trp (SD –W)(Snider et al., 2010). Colonies of transformed strains were grown in solid medium for 5 days in SD –W, SD –Trp–Ade–His (SD –WAH), SD –WAH +25 $\mu$ M fluconazole + 2% DMSO, SD –WAH +50 $\mu$ M fluconazole + 2% DMSO, and SD –WAH + 2% DMSO.

### **PCA testing of protein-protein interactions**

*PDR5*, *YOR1*, and *SNQ2* MAT $\alpha$  (mDHFR-F[1,2]-NatMX fusions) and MAT $\alpha$  (mDHFR-F[3]-HphMX fusions) PCA strains were obtained from a previous genome-wide screen(Tarassov et al., 2008). Additional strains acting required to recreate positive and negative controls were also obtained from this screen (Figure S11). Strains were individually mated and diploids were selected on solid YPD supplemented with Hygromycin B and Nourseothricin (YPD +Hyg +Nat). Diploid strains were spotted on solid YPD +Hyg +Nat supplemented with either 2% DMSO, 2% DMSO + 200  $\mu$ g/mL methotrexate, or 2% DMSO + 200  $\mu$ g/mL methotrexate + 46.8 $\mu$ M fluconazole. Strains were grown for 72 hours at 30°C.

### **Quantitative RT-PCR**

RNA was extracted from cultures growing exponentially in 23.43 $\mu$ M fluconazole using the QIAGEN RNeasy® kit. 1 $\mu$ g of isolate was treated with DNase and analyzed using an Agilent Bioanalyzer to quantify nucleic acid concentration and verify purity. cDNA synthesis was performed using a combination of oligo-DT and random hexamer primers. qPCR on these samples was then performed using a SensiFAST™ Real-Time PCR Kit and Ct values were quantified using a CFX machine. cDNA synthesis and qPCR was performed for *PDR5* and *UBC6* (acting loading control).

## **Availability of Data and Materials**

### **Competing Interests**

The authors declare that they have no competing interests.

### **Acknowledgements**

This work was supported by XX.

### **Author Contributions**

N.Y, F.P.R & A.C conceived the experiments. N.Y, M.G, L.M, S.Z & T.F performed experiments. A.C and N.Y analyzed the data. A.C, F.P.R, & N.Y. wrote the paper.

### **Additional Data Files**

**Additional Data S1.** List of primers used in this study. Includes the primers used to construct the barcoder strain, perform genotyping, RCP-PCR overhangs, and pool multiplexing primers.

**Additional Data S2.** Genotyping data in the engineered population. Includes a list of control strains used in high-throughput genotyping, initial genotyping results, re-genotyping of putative wild-type strains, and the final set of genotyping data used.

**Additional Data S3.** Drugs used in this study and their concentration in the pooled growth data.

**Additional Data S4.** List of primer pairs used to multiplex pooled growth sequencing data.

**Additional Data S5.** Growth and resistance metrics obtained for all strains in both the MAT $\alpha$  and MAT $\alpha$  pools.

**Additional Data S6.** Summary of linear modelling results obtained in this study.

**Additional Data S7.** Previously-known drug knockout associations within the 16 ABC transporters and 16 drugs studied.

**Additional Data S8.** Functional interpretations of genetic interactions present in the data.

## References

- Adamson, B., Norman, T.M., Jost, M., Cho, M.Y., Nuñez, J.K., Chen, Y., Villalta, J.E., Gilbert, L.A., Horlbeck, M.A., Hein, M.Y., et al. (2016). A Multiplexed Single-Cell CRISPR Screening Platform Enables Systematic Dissection of the Unfolded Protein Response. *Cell* *167*, 1867–1882.e21.
- Angeles-Albores, D., Puckett Robinson, C., Williams, B.A., Wold, B.J., and Sternberg, P.W. (2018). Reconstructing a metazoan genetic pathway with transcriptome-wide epistasis measurements. *Proc. Natl. Acad. Sci. U. S. A.* *115*, E2930–E2939.
- Beh, C.T., Cool, L., Phillips, J., and Rine, J. (2001). Overlapping functions of the yeast oxysterol-binding protein homologues. *Genetics* *157*, 1117–1140.
- Benhamou, R.I., Bibi, M., Steinbuch, K.B., Engel, H., Levin, M., Roichman, Y., Berman, J., and Fridman, M. (2017). Real-Time Imaging of the Azole Class of Antifungal Drugs in Live Candida Cells. *ACS Chem. Biol.* *12*, 1769–1777.
- Bloom, J.S., Ehrenreich, I.M., Loo, W.T., Lite, T.-L.V., and Kruglyak, L. (2013). Finding the sources of missing heritability in a yeast cross. *Nature* *494*, 234–237.
- Boettcher, M., Tian, R., Blau, J.A., Markegard, E., Wagner, R.T., Wu, D., Mo, X., Biton, A., Zaitlen, N., Fu, H., et al. (2018). Dual gene activation and knockout screen reveals directional dependencies in genetic networks. *Nat. Biotechnol.* *36*, 170–178.
- Braberg, H., Alexander, R., Shales, M., Xu, J., Franks-Skiba, K.E., Wu, Q., Haber, J.E., and Krogan, N.J. (2014). Quantitative analysis of triple-mutant genetic interactions. *Nat. Protoc.* *9*, 1867–1881.
- Braun, P., Tasan, M., Dreze, M., Barrios-Rodiles, M., Lemmens, I., Yu, H., Sahalie, J.M., Murray, R.R., Roncari, L., de Smet, A.-S., et al. (2009). An experimentally derived confidence score for binary protein-protein interactions. *Nat. Methods* *6*, 91–97.
- Brem, R.B., and Kruglyak, L. (2005). The landscape of genetic complexity across 5,700 gene expression traits in yeast. *Proc. Natl. Acad. Sci. U. S. A.* *102*, 1572–1577.
- Cancer Genome Atlas Research Network, Ley, T.J., Miller, C., Ding, L., Raphael, B.J., Mungall, A.J., Robertson, A.G., Hoadley, K., Triche, T.J., Laird, P.W., et al. (2013). Genomic and Epigenomic Landscapes of Adult De Novo Acute Myeloid Leukemia. *N. Engl. J. Med.* *368*, 2059–2074.
- Costanzo, M., Baryshnikova, A., Bellay, J., Kim, Y., Spear, E.D., Sevier, C.S., Ding, H., Koh, J.L.Y., Toufighi, K., Mostafavi, S., et al. (2010). The genetic landscape of a cell. *Science* *327*, 425–431.

- Costanzo, M., VanderSluis, B., Koch, E.N., Baryshnikova, A., Pons, C., Tan, G., Wang, W., Usaj, M.M.M., Hanchard, J., Lee, S.D., et al. (2016). A global genetic interaction network maps a wiring diagram of cellular function. *Science* 353.
- Díaz-Mejía, J.J., Celaj, A., Mellor, J.C., Coté, A., Balint, A., Ho, B., Bansal, P., Shaeri, F., Gebbia, M., Weile, J., et al. (2018). Mapping DNA damage-dependent genetic interactions in yeast via party mating and barcode fusion genetics. *Mol. Syst. Biol.* 14, e7985.
- Dixit, A., Parnas, O., Li, B., Chen, J., Fulco, C.P., Jerby-Arnon, L., Marjanovic, N.D., Dionne, D., Burks, T., Raychowdhury, R., et al. (2016). Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. *Cell* 167, 1853–1866.e17.
- Donner, M., and Keppler, D. (2001). Up-regulation of basolateral multidrug resistance protein 3 (Mrp3) in cholestatic rat liver. *Hepatology* 34, 351–359.
- Emanuel, G., Moffitt, J.R., and Zhuang, X. (2017). High-throughput, image-based screening of pooled genetic-variant libraries. *Nat. Methods* 14, 1159–1162.
- Ernst, R., Kueppers, P., Klein, C.M., Schwarzmüller, T., Kuchler, K., and Schmitt, L. (2008). A mutation of the H-loop selectively affects rhodamine transport by the yeast multidrug ABC transporter Pdr5. *Proc. Natl. Acad. Sci.* 105, 5069–5074.
- Fischer, B., Sandmann, T., Horn, T., Billmann, M., Chaudhary, V., Huber, W., and Boutros, M. (2015). A map of directional genetic interactions in a metazoan cell. *Elife* 4, e05464.
- Giaever, G., Chu, A.M., Ni, L., Connelly, C., Riles, L., Véronneau, S., Dow, S., Lucau-Danila, A., Anderson, K., André, B., et al. (2002). Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature* 418, 387–391.
- Gibson, D.G., Young, L., Chuang, R.-Y., Venter, J.C., Hutchison, C.A., and Smith, H.O. (2009). Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods* 6, 343–345.
- Gietz, R.D., and Schiestl, R.H. (2007). High-efficiency yeast transformation using the LiAc/SS carrier DNA/PEG method. *Nat. Protoc.* 2, 31–34.
- Haber, J.E., Braberg, H., Wu, Q., Alexander, R., Haase, J., Ryan, C., Lipkin-Moore, Z., Franks-Skiba, K.E., Johnson, T., Shales, M., et al. (2013). Systematic triple-mutant analysis uncovers functional connectivity between pathways involved in chromosome regulation. *Cell Rep.* 3, 2168–2178.
- Hartman, J.L., Garvik, B., and Hartwell, L. (2001). Principles for the Buffering of Genetic Variation. *Science* 291.
- Heckl, D., Kowalczyk, M.S., Yudovich, D., Belizaire, R., Puram, R. V, McConkey, M.E., Thielke, A., Aster, J.C., Regev, A., and Ebert, B.L. (2014). Generation of mouse models of myeloid malignancy with combinatorial genetic lesions using CRISPR-Cas9 genome editing. *Nat. Biotechnol.* 32, 941–946.
- Horlbeck, M.A., Xu, A., Wang, M., Bennett, N.K., Park, C.Y., Bogdanoff, D., Adamson, B., Chow, E.D., Kampmann, M., Peterson, T.R., et al. (2018). Mapping the Genetic Landscape of Human Cells. *Cell* 174, 953–967.e22.
- Huls, M., Brown, C.D.A., Windass, A.S., Sayer, R., van den Heuvel, J.J.M.W., Heemskerk, S., Russel, F.G.M., and Masereeuw, R. (2008). The breast cancer resistance protein transporter ABCG2 is expressed in the human kidney proximal tubule apical membrane. *Kidney Int.* 73, 220–225.
- Jakočiūnas, T., Bonde, I., Herrgård, M., Harrison, S.J., Kristensen, M., Pedersen, L.E., Jensen, M.K., and Keasling, J.D. (2015). Multiplex metabolic pathway engineering using CRISPR/Cas9 in *Saccharomyces cerevisiae*. *Metab. Eng.* 28, 213–222.

- Jao, L.-E., Wente, S.R., and Chen, W. (2013). Efficient multiplex biallelic zebrafish genome editing using a CRISPR nuclease system. *Proc. Natl. Acad. Sci.* *110*, 13904–13909.
- Katzmann, D.J., Burnett, P.E., Golin, J., Mahé, Y., and Moye-Rowley, W.S. (1994). Transcriptional control of the yeast PDR5 gene by the PDR3 gene product. *Mol. Cell. Biol.* *14*, 4653–4661.
- Kebschull, J.M., and Zador, A.M. (2018). Cellular barcoding: lineage tracing, screening and beyond. *Nat. Methods* *15*, 871–879.
- Khakhina, S., Johnson, S.S., Manoharlal, R., Russo, S.B., Blugeon, C., Lemoine, S., Sunshine, A.B., Dunham, M.J., Cowart, L.A., Devaux, F., et al. (2015). Control of Plasma Membrane Permeability by ABC Transporters. *Eukaryot. Cell* *14*, 442–453.
- Kolaczkowska, A., Kolaczkowski, M., Goffeau, A., and Moye-Rowley, W.S. (2008). Compensatory activation of the multidrug transporters Pdr5p, Snq2p, and Yor1p by Pdr1p in *Saccharomyces cerevisiae*. *FEBS Lett.* *582*, 977–983.
- Kolaczkowski, M., Kolaczkowska, A., Luczynski, J., Witek, S., and Goffeau, A. (1998). In Vivo Characterization of the Drug Resistance Profile of the Major ABC Transporters and Other Components of the Yeast Pleiotropic Drug Resistance Network. *Microb. Drug Resist.* *4*, 143–158.
- König, J., Rost, D., Cui, Y., and Keppler, D. (1999). Characterization of the human multidrug resistance protein isoform MRP3 localized to the basolateral hepatocyte membrane. *Hepatology* *29*, 1156–1163.
- Kuzmin, E., VanderSluis, B., Wang, W., Tan, G., Deshpande, R., Chen, Y., Usaj, M., Balint, A., Mattiazzi Usaj, M., van Leeuwen, J., et al. (2018). Systematic analysis of complex genetic interactions. *Science* (80-. ). *360*, eaao1729.
- Lee, A.Y., St Onge, R.P., Proctor, M.J., Wallace, I.M., Nile, A.H., Spagnuolo, P.A., Jitkova, Y., Gronda, M., Wu, Y., Kim, M.K., et al. (2014). Mapping the cellular response to small molecules using chemogenomic fitness signatures. *Science* *344*, 208–211.
- Ma, J., Yu, M.K., Fong, S., Ono, K., Sage, E., Demchak, B., Sharan, R., and Ideker, T. (2018). Using deep learning to model the hierarchical structure and function of a cell. *Nat. Methods* *15*, 290–298.
- Mani, R., St Onge, R.P., Hartman, J.L., Giaever, G., and Roth, F.P. (2008). Defining genetic interaction. *Proc. Natl. Acad. Sci. U. S. A.* *105*, 3461–3466.
- Mullis, M.N., Matsui, T., Schell, R., Foree, R., and Ehrenreich, I.M. (2018). The complex underpinnings of genetic background effects. *Nat. Commun.* *9*, 3548.
- Najm, F.J., Strand, C., Donovan, K.F., Hegde, M., Sanson, K.R., Vaimberg, E.W., Sullender, M.E., Hartenian, E., Kalani, Z., Fusi, N., et al. (2017). Orthologous CRISPR–Cas9 enzymes for combinatorial genetic screens. *Nat. Biotechnol.* *36*, 179–189.
- Newman, J.R.S., Ghaemmaghami, S., Ihmels, J., Breslow, D.K., Noble, M., DeRisi, J.L., and Weissman, J.S. (2006). Single-cell proteomic analysis of *S. cerevisiae* reveals the architecture of biological noise. *Nature* *441*, 840–846.
- Otwinowski, J., McCandlish, D.M., and Plotkin, J.B. (2018). Inferring the shape of global epistasis. *Proc. Natl. Acad. Sci. U. S. A.* *115*, E7550–E7558.
- Perlstein, E.O., Ruderfer, D.M., Roberts, D.C., Schreiber, S.L., and Kruglyak, L. (2007). Genetic basis of individual differences in the response to small-molecule drugs in yeast. *Nat. Genet.* *39*, 496–502.
- Phenix, H., Morin, K., Batenchuk, C., Parker, J., Abedi, V., Yang, L., Tepliakova, L., Perkins, T.J., and Kærn, M. (2011). Quantitative Epistasis Analysis and Pathway Inference from Genetic

- Interaction Data. PLoS Comput. Biol. 7, e1002048.
- Proctor, M., Urbanus, M.L., Fung, E.L., Jaramillo, D.F., Davis, R.W., Nislow, C., and Giaever, G. (2011). The Automated Cell: Compound and Environment Screening System (ACCESS) for Chemogenomic Screening. (Humana Press), pp. 239–269.
- Rubin, A.J., Parker, K.R., Satpathy, A.T., Qi, Y., Wu, B., Ong, A.J., Mumbach, M.R., Ji, A.L., Kim, D.S., Cho, S.W., et al. (2018). Coupled Single-Cell CRISPR Screening and Epigenomic Profiling Reveals Causal Gene Regulatory Networks. *Cell*.
- Shaw, W.M., Yamauchi, H., Mead, J., Gowers, G.F., Oling, D., Larsson, N., Wigglesworth, M., Ladds, G., and Ellis, T. (2018). Engineering a model cell for rational tuning of GPCR signaling. *BioRxiv* 390559.
- Shekhar-Guturja, T., Tebung, W.A., Mount, H., Liu, N., Köhler, J.R., Whiteway, M., and Cowen, L.E. (2016). Beauvericin Potentiates Azole Activity via Inhibition of Multidrug Efflux, Blocks *C. albicans* Morphogenesis, and is Effluxed via Yor1 and Circuitry Controlled by Zcf29. *Antimicrob. Agents Chemother.* 60, AAC.01959-16.
- Shen, J.P., and Ideker, T. (2018). Synthetic Lethal Networks for Precision Oncology: Promises and Pitfalls. *J. Mol. Biol.* 430, 2900–2912.
- Shen, J.P., Zhao, D., Sasik, R., Luebeck, J., Birmingham, A., Bojorquez-Gomez, A., Licon, K., Klepper, K., Pekin, D., Beckett, A.N., et al. (2017). Combinatorial CRISPR–Cas9 screens for de novo mapping of genetic interactions. *Nat. Methods* 14, 573–576.
- Smith, A.M., Heisler, L.E., Mellor, J., Kaper, F., Thompson, M.J., Chee, M., Roth, F.P., Giaever, G., and Nislow, C. (2009). Quantitative phenotyping via deep barcode sequencing. *Genome Res.* 19, 1836–1842.
- Snider, J., Kittanakom, S., Damjanovic, D., Curak, J., Wong, V., and Stagljar, I. (2010). Detecting interactions with membrane proteins using a membrane two-hybrid assay in yeast. *Nat. Protoc.* 5, 1281–1293.
- Snider, J., Hanif, A., Lee, M.E., Jin, K., Yu, A.R., Graham, C., Chuk, M., Damjanovic, D., Wierzbicka, M., Tang, P., et al. (2013). Mapping the functional yeast ABC transporter interactome. *Nat. Chem. Biol.* 9, 565–572.
- Sousa, C.A., Hanselaer, S., and Soares, E. V. (2015). ABCC Subfamily Vacuolar Transporters are Involved in Pb (Lead) Detoxification in *Saccharomyces cerevisiae*. *Appl. Biochem. Biotechnol.* 175, 65–74.
- St Onge, R.P., Mani, R., Oh, J., Proctor, M., Fung, E., Davis, R.W., Nislow, C., Roth, F.P., and Giaever, G. (2007). Systematic pathway analysis using high-resolution fitness profiling of combinatorial gene deletions. *Nat. Genet.* 39, 199–206.
- Suzuki, Y., St Onge, R.P., Mani, R., King, O.D., Heilbut, A., Labunskyy, V.M., Chen, W., Pham, L., Zhang, L. V, Tong, A.H.Y., et al. (2011). Knocking out multigene redundancies via cycles of sexual assortment and fluorescence selection. *Nat. Methods* 8, 159–164.
- Takahashi, K., and Yamanaka, S. (2006). Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors. *Cell* 126, 663–676.
- Tarassov, K., Messier, V., Landry, C.R., Radinovic, S., Serna Molina, M.M., Shames, I., Malitskaya, Y., Vogel, J., Bussey, H., and Michnick, S.W. (2008). An in vivo map of the yeast protein interactome. *Science* 320, 1465–1470.
- Taylor, M.B., Ehrenreich, I.M., Rothstein, R., Hu, T., and Mast, J. (2014). Genetic Interactions Involving Five or More Genes Contribute to a Complex Trait in Yeast. *PLoS Genet.* 10, e1004324.
- Wang, H., Yang, H., Shivalila, C.S., Dawlaty, M.M., Cheng, A.W., Zhang, F., and Jaenisch, R.

- (2013). One-Step Generation of Mice Carrying Mutations in Multiple Genes by CRISPR/Cas-Mediated Genome Engineering. *Cell* *153*, 910–918.
- Wieczorke, R., Krampe, S., Weierstall, T., Freidel, K., Hollenberg, C.P., and Boles, E. (1999). Concurrent knock-out of at least 20 transporter genes is required to block uptake of hexoses in *Saccharomyces cerevisiae*. *FEBS Lett.* *464*, 123–128.
- Wong, A.S.L., Choi, G.C.G., Cui, C.H., Pregering, G., Milani, P., Adam, M., Perli, S.D., Kazer, S.W., Gaillard, A., Hermann, M., et al. (2016). Multiplexed barcoded CRISPR-Cas9 screening enabled by CombiGEM. *Proc. Natl. Acad. Sci. U. S. A.* *113*, 2544–2549.
- Xu, S., Wang, Z., Kim, K.W., Jin, Y., and Chisholm, A.D. (2016). Targeted Mutagenesis of Duplicated Genes in *Caenorhabditis elegans* Using CRISPR-Cas9. *J. Genet. Genomics* *43*, 103–106.
- Yachie, N., Petsalaki, E., Mellor, J.C., Weile, J., Jacob, Y., Verby, M., Ozturk, S.B., Li, S., Cote, A.G., Mosca, R., et al. (2016). Pooled-matrix protein interaction screens using Barcode Fusion Genetics. *Mol. Syst. Biol.* *12*, 863.
- Yan, Z., Costanzo, M., Heisler, L.E., Paw, J., Kaper, F., Andrews, B.J., Boone, C., Giaever, G., and Nislow, C. (2008). Yeast Barcoders: a chemogenomic application of a universal donor-strain collection carrying bar-code identifiers. *Nat. Methods* *5*, 719–725.
- Zeitoun, R.I., Pines, G., Grau, W.C., and Gill, R.T. (2017). Quantitative Tracking of Combinatorially Engineered Populations with Multiplexed Binary Assemblies. *ACS Synth. Biol.* *6*, 619–627.
- Zhang, Z., Mao, Y., Ha, S., Liu, W., Botella, J.R., and Zhu, J.-K. (2016). A multiplex CRISPR/Cas9 platform for fast and efficient editing of multiple genes in *Arabidopsis*. *Plant Cell Rep.* *35*, 1519–1533.

## Figures

**Figure 1.** Overview of the engineered population profiling process.

We created a barcoded wild-type pool (Figure S1) to enable construction of an engineered population using any multi-mutant strain. In this study, this pool was mated with a 16 ABC transporter knockout strain (ABC-16). The genotype at 16 ABC transporters is indicated by the squares drawn in each cell (black = knockout, white = wild type). Diploids from this cross were subject to sporulation and barcoded haploids were then selected. Each haploid inherits either a wild-type or knockout allele at these 16 loci. Single colonies were picked and arrayed onto a series of 384-well plates. *En masse* genotyping was performed on this collection using an RCP-PCR (Yachie et al., 2016) strategy, which uses a combination of row and column tags to allow identification of PCR products arising from the same well in each plate (Methods). An additional PCR reaction adds a plate tag (not shown). High throughput sequencing of pooled RCP-PCR products allows large scale genotyping and identification of a strain-specific DNA barcode for many strains. Strains with a successfully determined barcode and genotype are transferred into two liquid pools based on mating type (MAT $\alpha$  or MAT $\alpha$ ), and grown under each of 16 drugs, as well as a solvent control. High throughput sequencing of strain-specific DNA barcodes at t = 0, 5, 10, 15, and 20 generations of growth reconstructs the resistance of each strain to each drug, profiling the engineered population.

**Figure 2.** An exploration and assessment of multi-knockout fitness landscapes within a 6-gene group.

**A** Comparison of MAT $\alpha$  and MAT $\alpha$  group resistance profiles in camptothecin and ketoconazole. Individuals were grouped by their genotype at *pdr5* $\Delta$ , *snq2* $\Delta$ , *ybt1* $\Delta$ , *ycf1* $\Delta$ , and *yor1* $\Delta$ . The 5-locus genotype of each group is indicated by the legend. Individuals in each group vary at the remaining 11 loci. Each point represents the mean resistance of the indicated group in the MAT $\alpha$  pools (x-axis) and MAT $\alpha$  pools (y-axis). Profiles for the remaining drugs are shown in Figure S5.

**B** Distribution of MAT $\alpha$ –MAT $\alpha$  correlations of the grouped resistance profiles amongst all drugs tested.

**C** A radial landscape of benomyl resistance. The graph is centered by the 5-gene wild-type group, with outward extensions adding cumulative knockouts. Each section is coloured by the average resistance of the corresponding 5-gene group relative to the 5-gene wild type. Extensions to 1, 2, and 5 total knockouts are shown. Sections are coloured by the mean resistance of each group relative to the 5-gene wild-type. The colour scale is centered by the mean 5-gene wild-type resistance and extends by half of the observed difference between the 5% and 95% percentile resistance in that drug in both directions (blue for increased resistance, orange for decreased resistance).

**D** As in B, showing radial fitness landscapes for 10 additional drugs. The remaining 5 drugs are shown in Figure S6.

**Figure 3.** Exploration and formalization of surprising multi-gene knockout phenotypes.

**A** A linear landscape of resistance to benomyl, mitoxantrone, and valinomycin amongst 5-gene groups. The 5-gene genotype of each group is indicated by the legend. Groups are arranged on the x-axis by the number of knockouts (with jitter added to improve clarity), and the y-axis by average drug resistance. Groups separated by a single additional knockout are connected by lines. Solid lines indicate significant differences in resistance (Bonferroni-adjusted  $p < 0.05$ , Mann-Whitney U test), otherwise dashed lines are used. Linear landscapes for all pools are drawn in Figure S7.

**B** Distribution of valinomycin resistance amongst all *ybt1* $\Delta$ , *yor1* $\Delta$ , *snq2* $\Delta$ , *ycf1* $\Delta$ , and *pdr5* $\Delta$  knockout groups. Group genotype is indicated for each line using the same legend as in A). All *pdr5* $\Delta$  groups (dark blue) are paired with their corresponding *PDR5* $+$  equivalent (grey).

**C** A linear model was used to formally determine significant gene knockout and genetic interaction effects mediating resistance to the tested drugs (see Methods). Linear model terms which were significant (Bonferroni adjusted  $p < 0.05$ ) in a given drug are coloured according the legend on the right. Maximum and minimum scale values are determined by the median absolute deviation of the log(resistance) in that drug. Non-significant terms are coloured in grey.  $\epsilon$  terms represent n-way interactions (see Methods). Coefficients are sorted by term complexity. Term complexity is also indicated by the grey colour scale on the top of the heatmap.

**Figure 4.** Modeling and interpreting a complex genetic landscape.

**A** A neural network model was created to infer transporter-drug and transporter-transporter relationships from the engineered population profiles. The 16-transporter genotype ( $\mathbf{G}$ ), is given as input to the model as a binary variable (1 = presence, 0 = absence for each transporter), and the activity of each transporter ( $\mathbf{A}$ ) is computed by the set of learned transporter-transporter influence weights ( $\mathbf{I}$ ), and is multiplied element-wise by  $\mathbf{G}$ . Resistance to each of the 16 tested compounds ( $\mathbf{R}$ ) is then computed by transporter-drug efflux weights ( $\mathbf{E}$ ). Appropriate weights for  $\mathbf{I}$  and  $\mathbf{E}$  are learned using stochastic gradient descent and backpropagation using the engineered population

profiling data such that mean-squared error is minimized between  $\mathbf{R}$  and measured resistance. In addition, a positive constraint is placed on  $\mathbf{E}$  and regularization is added to the model (Methods).

**B** Weights learned by the neural network model after training and pruning are shown. All non-zero  $\mathbf{I}$  weights learned by the model were negative.

**C** Comparing the normalized resistance of each strain measured by engineered population profiling to resistances modelled by the neural network.

**D** Comparing the neural model in valinomycin to the observed resistances for each five-gene knockout group. The neural network weights (top) are shown for the original model (top-left) and one trained with an extra always-present node in the activity layer to model potential influence of a hidden resistance factor (top right). At the bottom, strains were grouped by knockout genotypes at  $pdr5\Delta$ ,  $snq2\Delta$ ,  $ybt1\Delta$ ,  $ycf1\Delta$ , and  $yor1\Delta$ . Each point represents the mean resistance of a group of strains containing the 5-locus genotype indicated by the legend, either as modeled by the corresponding neural network (x-axis) or as measured in the data (y-axis). Correlation in the top left is shown for all data, then only for the 5-locus groups in parentheses.

**Figure 5.** Further modeling and exploring of ABC-16 mediated fluconazole resistance.

**A** As in Figure 3A, a linear landscape of fluconazole resistance is shown .

**B** Comparing the neural model in fluconazole to the observed resistances for each five-gene knockout group. The neural network weights (top) are shown for the original model (top-left) and one trained with an extra always-present ‘hidden’ node between the  $G$  and  $A$  layer to model potential non-linear influence of Pdr5 (see Methods for details, top right). At the bottom, strains were grouped by knockout genotypes at  $pdr5\Delta$ ,  $snq2\Delta$ ,  $ybt1\Delta$ ,  $ycf1\Delta$ , and  $yor1\Delta$ . Each point represents the mean resistance of a group of strains containing the 5-locus genotype indicated by the legend, either as modeled by the corresponding neural network (x-axis) or as measured in the data (y-axis). Correlation in the top left is shown for all data, then only for the 5-locus groups in parentheses.

**C** Comparing the IC50 of fluconazole derived from single-strain growth experiments to the normalized resistance expected by in the grouped pool data (mean resistance is shown for each group). Strain genotype is indicated by the legend.

**D** Measuring the mRNA expression of *PDR5* in wild-type (RY0566),  $ybt1\Delta ycf1\Delta$ ,  $snq2\Delta yor1\Delta$ , and  $snq2\Delta yor1\Delta ybt1\Delta ycf1\Delta$  strains. *PDR5* mRNA expression was measured using qRT-PCR and normalized relative to *UBC6*. Values represent the ratio of *PDR5* expression compared to the average in the wild-type. Error bars indicate standard deviation. Three replicates were used in each experiment. p-values were calculated using a t-test. Overlaid are the corresponding Pdr5 activity values from the neural network in the top-right panel of Figure 5B, considering only influences going through the hidden node (yellow), or all influences (orange).

**E** Comparing the modeled *PDR5* repression by *YOR1* and *SNQ2* with with protein-protein interactions found using MYTH and PCA. Interactions were measured in both this study (Figure S11, S12) and previous studies(Snider et al., 2013; Tarassov et al., 2008). Learned  $\mathbf{I}$  weights from 4B are overlaid.

**Figure S1.** Creation of a parent barcoder pool.

**A** Engineering of a barcoder pool cassette. An HphMX4 cassette was amplified from pIS420, with overhangs adding the US2 and DS1 sites. A second PCR reaction was performed to add 25

random base pairs for use as UP and DN tags, as well as two constant US1 and DS2 regions. A third PCR reaction then adds *loxP*/*lox2272* sites, and homology to the pSH47 SacI site.

**B** Transforming a pool of barcoder parents. RY0148 was modified to add a *loxP-URA3-lox2272* site and was co-transformed with the barcoder pool cassette and SacI-digested pSH47 to enable reconstitution of a pSH47-based barcoder plasmid construct through in-yeast assembly. Transformants were selected by growth in YPG +Hyg for 3 days to allow for both selection of successful in-yeast assembly products, as well as induction of Cre to enable recombination and replacement of URA3 with the barcoder pool cassette. Loss of URA3 through Cre-enabled recombination is selected by subsequent growth in 5-FOA.

**Figure S2.** Analysis of pool genotyping quality.

**A** Expected genotyping accuracy at the 16 ABC transporters surveyed. Accuracy was estimated by evaluating the performance of the RCP-PCR genotyping protocol on a set of ‘gold standard’ strains (Methods, Data S2).

**B** Distribution of knockouts in the combined MAT $\alpha$  and MAT $\alpha$  pools. The observed number of strains with a given number of knockouts are indicated in grey. The expected number of strains with a given number of knockouts at 93.8% genotyping accuracy under a random assortment model are indicated in black.

**C** Tests of gene linkage within the MAT $\alpha$  pools (upper triangle) and MAT $\alpha$  pools (lower triangle). The Pearson correlation coefficient of the corresponding genotype pairs are indicated on the right. Pairs without significant correlation (Bonferroni-corrected  $p$  value  $\geq 0.05$ ) are shaded in grey.

**Figure S3.** Reproducible marginal gene knockout growth effects in the pool.

A linear model was used to formally determine significant gene knockout effects mediating growth in the tested drugs. Linear model terms which were significant (Bonferroni adjusted  $p < 0.05$ ) in both MAT $\alpha$  and MAT $\alpha$  pools for their given drug are coloured according the legend on the left. Other terms are coloured in grey.

**Figure S4.** Reproducible marginal gene knockout resistance effects in the pool.

A linear model was used to formally determine significant gene knockout effects mediating resistance to the tested drugs. Linear model terms which were significant (Bonferroni adjusted  $p < 0.05$ ) in both MAT $\alpha$  and MAT $\alpha$  pools for their given drug are coloured according the legend on the left. Other terms are coloured in grey.

**Figure S5.** Reproducibility of grouped genotype resistance.

Strains were grouped on knockout genotypes at *pdr5* $\Delta$ , *snq2* $\Delta$ , *ybt1* $\Delta$ , *ycf1* $\Delta$ , and *yor1* $\Delta$ . Each point represents a group of strains containing the 5-locus genotype indicated by the legend. Strains in each group vary at the remaining 11 loci. Each point represents the mean resistance of each group in the MAT $\alpha$  (x-axis) and MAT $\alpha$  (y-axis) pools.

**Figure S6.** A radial fitness landscape in six additional drugs.

A radial fitness landscape in six drugs showing all multi-knockout paths. Each graph is centered by the 5-gene wild-type group, with outward extensions adding cumulative knockouts. Each section is coloured by the average resistance of its corresponding knockout group relative to the 5-gene wild type. Extensions to 1, 2, and 5 total knockouts are shown. Sections are coloured by

the mean resistance of each group relative to the 5-gene wild-type. The colour scale is centered by the mean 5-gene wild-type resistance and extends by half of the observed difference between the 5% and 95% percentile resistance in that drug in both directions (blue for increased resistance, orange for decreased resistance).

**Figure S7.** A linear landscape of resistance to 16 drugs.

**A** A linear landscape of resistance to all tested drugs in the amongst 5-gene groups. The 5-gene genotype of each group is indicated by the legend. Groups are arranged on the x-axis by the number of knockouts (with jitter added to improve clarity), and on the y-axis by average drug resistance. Groups separated by single knockouts are connected by lines. Solid lines indicate significant differences in resistance (Bonferroni-adjusted  $p < 0.05$ , Mann-Whitney U test), otherwise dashed lines are used.

**Figure S8.** Neural network evaluation

**A** Number of reproducible network parameters ( $Z \geq 4$  estimated from 10 iterations, Methods) as a function of the regularization rate  $\lambda$ . 13 intervals are plotted from  $10^{-6}$  to  $10^0$  (left), and 11 intervals are plotted from  $10^{-4}$  to  $10^{-3}$  (right). Values between intervals are linearly interpolated.

**B** As in S8A, showing the overall mean squared error of the neural network.

**C** Comparing the normalized resistance of each strain measured by engineered population profiling to resistances modelled by the neural network. Results are shown when the network is trained on either the MAT $\alpha$  or MAT $\alpha$  population, and then tested on either the MAT $\alpha$  or MAT $\alpha$  population.

**D** Comparing the learned network weights when the network is trained on either the MAT $\alpha$  or MAT $\alpha$  population separately.

**Figure S9.** Neural network performance for single drugs

Strains were grouped on knockout genotypes at  $pdr5\Delta$ ,  $snq2\Delta$ ,  $ybt1\Delta$ ,  $ycf1\Delta$ , and  $yor1\Delta$ . Each point represents the mean resistance of a group of strains containing the 5-locus genotype indicated by the legend, either as modeled by the neural network (x-axis) or as measured in the data (y-axis). Correlation in the top left is shown for all data, then only for the 5-locus groups in parentheses.

**Figure S10.** Neural networks trained in single environments

**A** Comparing the neural model in valinomycin to the observed resistances for each five-gene knockout group. The neural network weights (top) are shown for a model trained only on valinomycin data. At the bottom, strains were grouped by knockout genotypes at  $pdr5\Delta$ ,  $snq2\Delta$ ,  $ybt1\Delta$ ,  $ycf1\Delta$ , and  $yor1\Delta$ . Each point represents the mean resistance of a group of strains containing the 5-locus genotype indicated by the legend, either as modeled by the corresponding neural network (x-axis) or as measured in the data (y-axis). Correlation in the top left is shown for all data, then only for the 5-locus groups in parentheses.

**B** As in A, showing a neural network trained only on fluconazole data.

**Figure S11.** Comparing drug resistance measured from single-strain experiments to the grouped pool data.

Resistance of individual strains containing each of 32 knockout combinations at *pdr5Δ*, *snq2Δ*, *ybt1Δ*, *ycf1Δ*, and *yor1Δ* was measured and compared to the resistance to the pool data. Pool strains were grouped based on genotype at these 5 loci, median log<sub>2</sub>-resistance was determined for each group in MAT $\alpha$  and MAT $\alpha$  pools, and these values were averaged to obtain a single pool value. Strain genotype is indicated by the legend. Growth of individual strains was measured at 1.9, 3.9, 7.8, 15.6, 23.4, 31.2, 35, and 40 $\mu$ m of fluconazole.

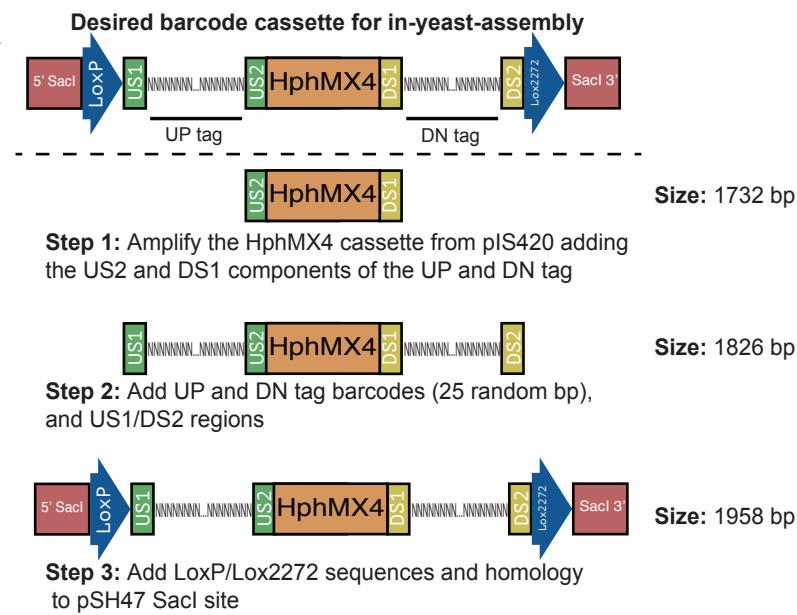
**Figure S12.** Measuring all protein-protein interactions between Pdr5, Snq2, and Yor1 using mDHFR PCA.

*PDR5*, *YOR1*, and *SNQ2* MAT $\alpha$  (mDHFR-F[1,2]-NatMX fusions) and MAT $\alpha$  (mDHFR-F[3]-HphMX fusions) PCA strains were obtained from a previous genome-wide screen(Tarassov et al., 2008). Strains were individually mated to obtain the indicated diploids. Diploid strains were spotted on YPD containing either DMSO, DMSO + methotrexate (MTX), or DMSO + MTX + 46.8 $\mu$ M fluconazole. MTX selects for successful reconstruction of mDHFR from the F[1,2] and F[3] fragments via a protein-protein interaction. Link-F[1,2]/ Link-F[3] is a diploid strain which tests against interaction of the universal linker regions when fused to the mDHFR fragments. Zip-F[1,2]/ Zip-F[3] is a diploid strain which tests for interaction between two leucine Zipper sequences fused to the mDHFR fragments. Strains were grown for 3 days at 30°C.

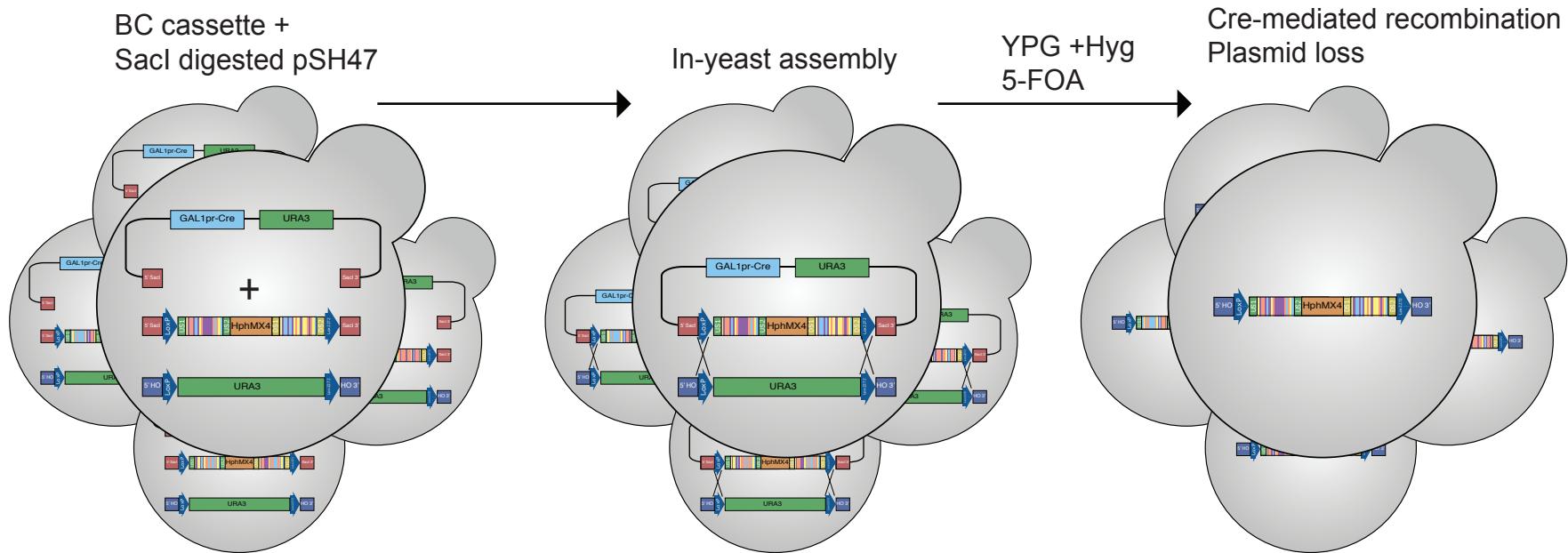
**Figure S13.** Measuring protein-protein interactions of Pdr5 with Snq2 and Yor1 using MYTH. NubG-PDR5, NubI-PDR5, Ost1-NubG, and Ost1-NubI strains were retrieved from a previously constructed genomic prey library<sup>25</sup> and were each transformed with plasmids containing clones of *PDR5*, *YOR1*, *SNQ2*, or an artificial bait fused to Cub (YOR1-L2, PDR5-L2, SNQ2-L2, Artificial L2 bait). NubI fusions are expected to spontaneously reconstitute ubiquitin with Cub, while NubG fusions are expected to require a protein-protein interaction for reconstitution. Ost1 is a component of the oligosaccharyltransferase complex localized to the endoplasmic reticulum membrane and is not expected to interact with any baits tested. Colonies of transformed strains were spotted on SD –Trp (SD –W), SD –Trp–Ade–His (SD –WAH), SD –WAH +25 $\mu$ M fluconazole + 2% DMSO, SD –WAH +50 $\mu$ M fluconazole + 2% DMSO, and SD –WAH + 2% DMSO. SD –WAH conditions select for reconstitution of ubiquitin.

# Figure S1

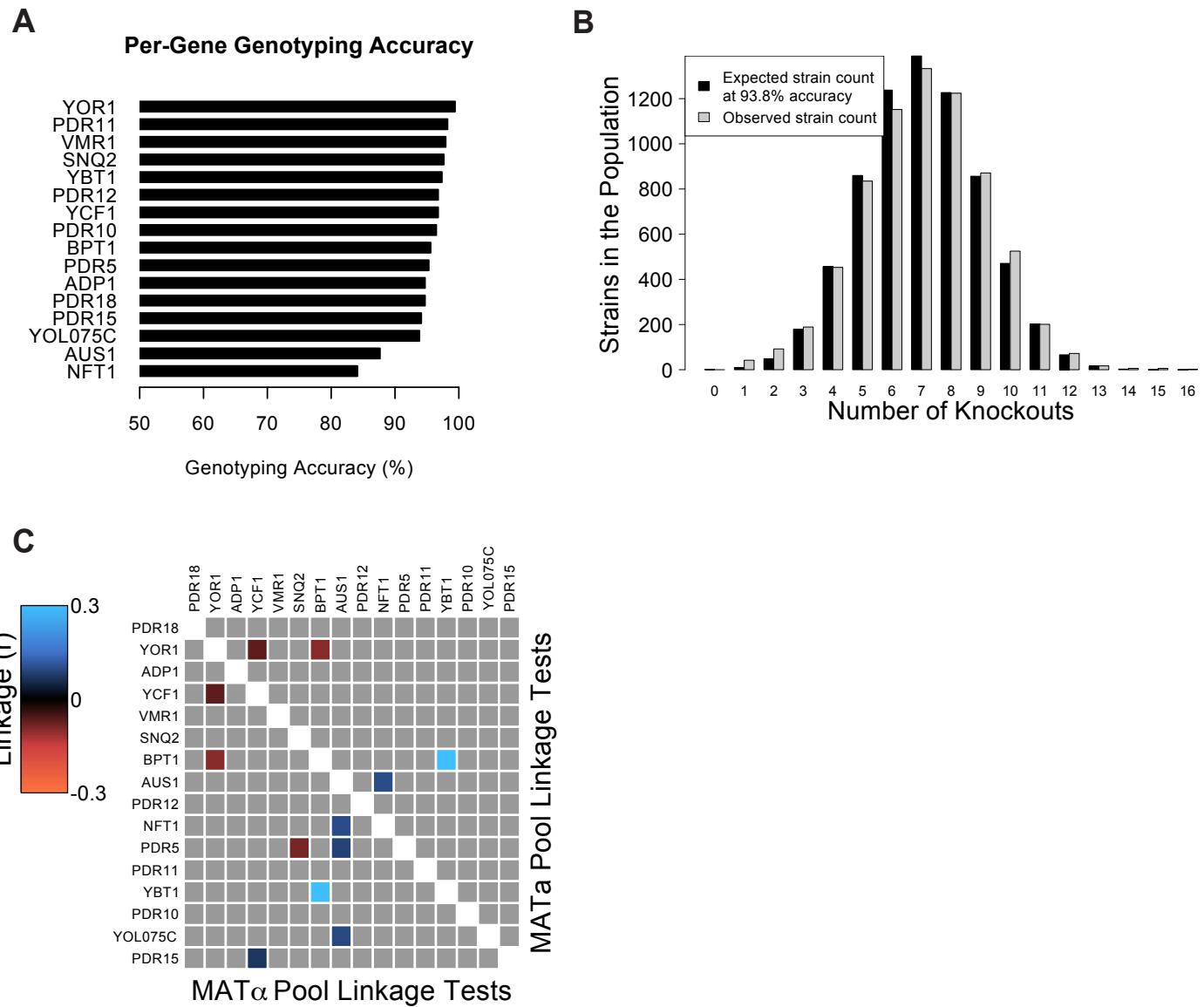
A



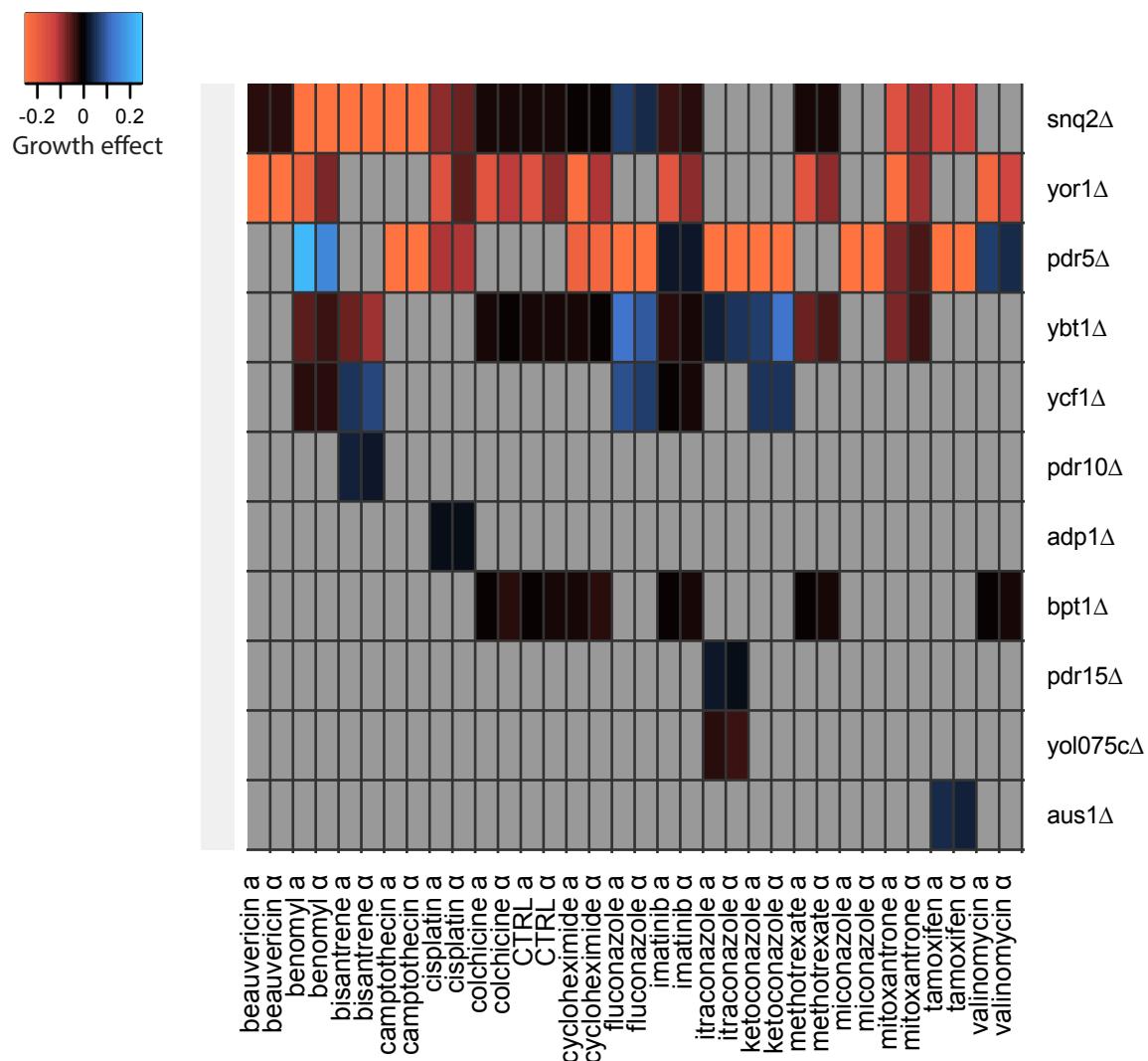
B



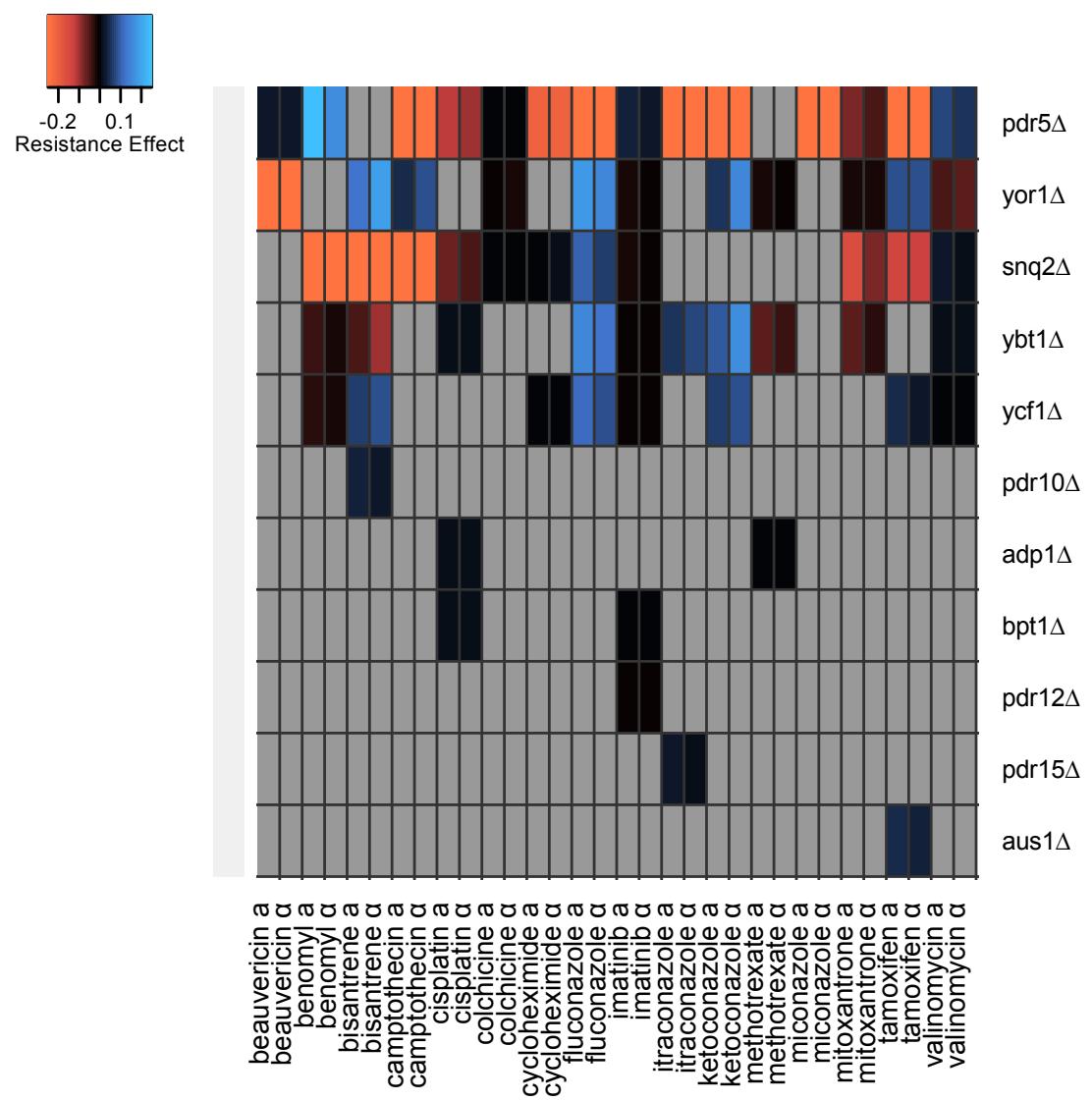
# Figure S2



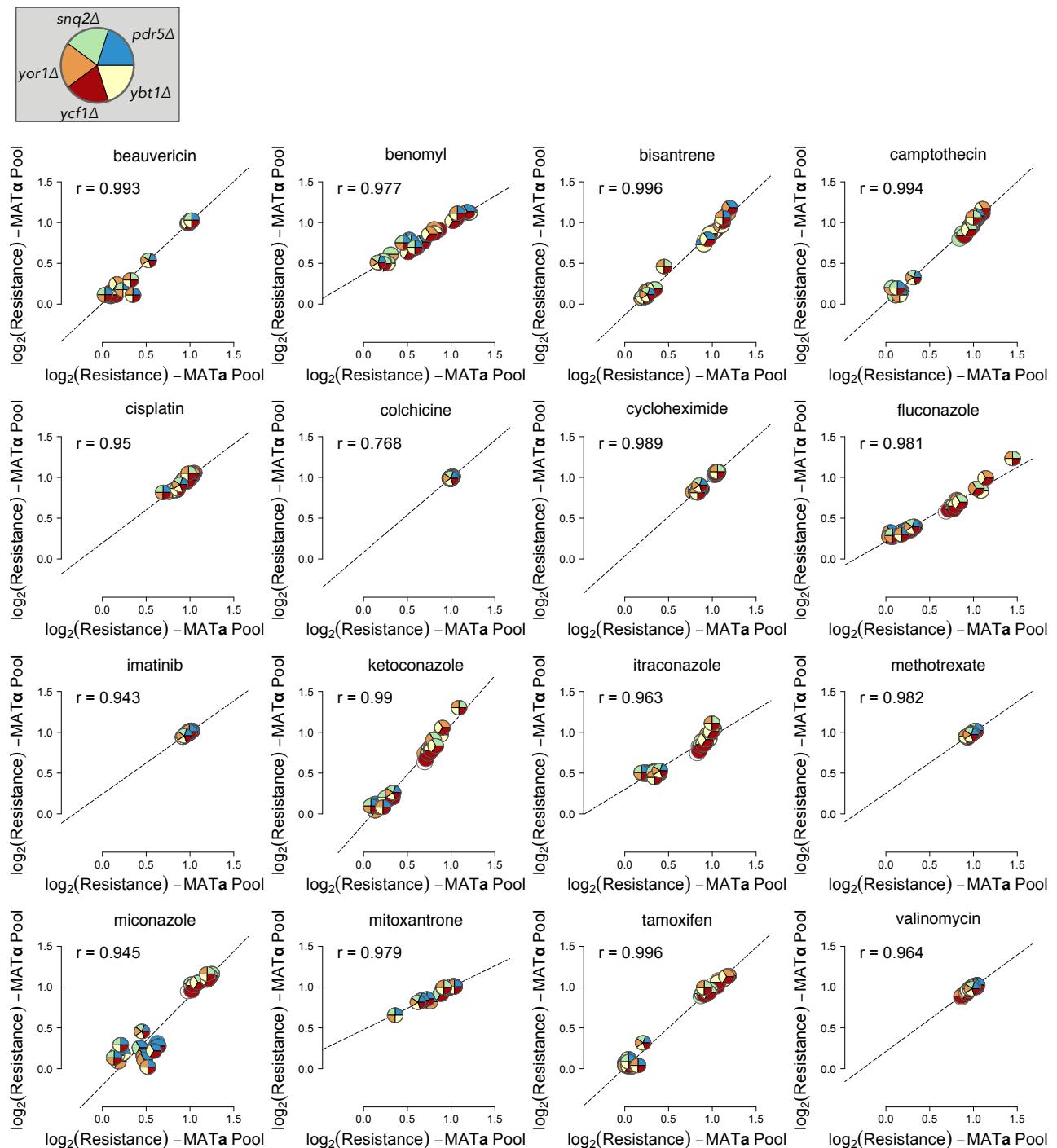
# Figure S3



# Figure S4

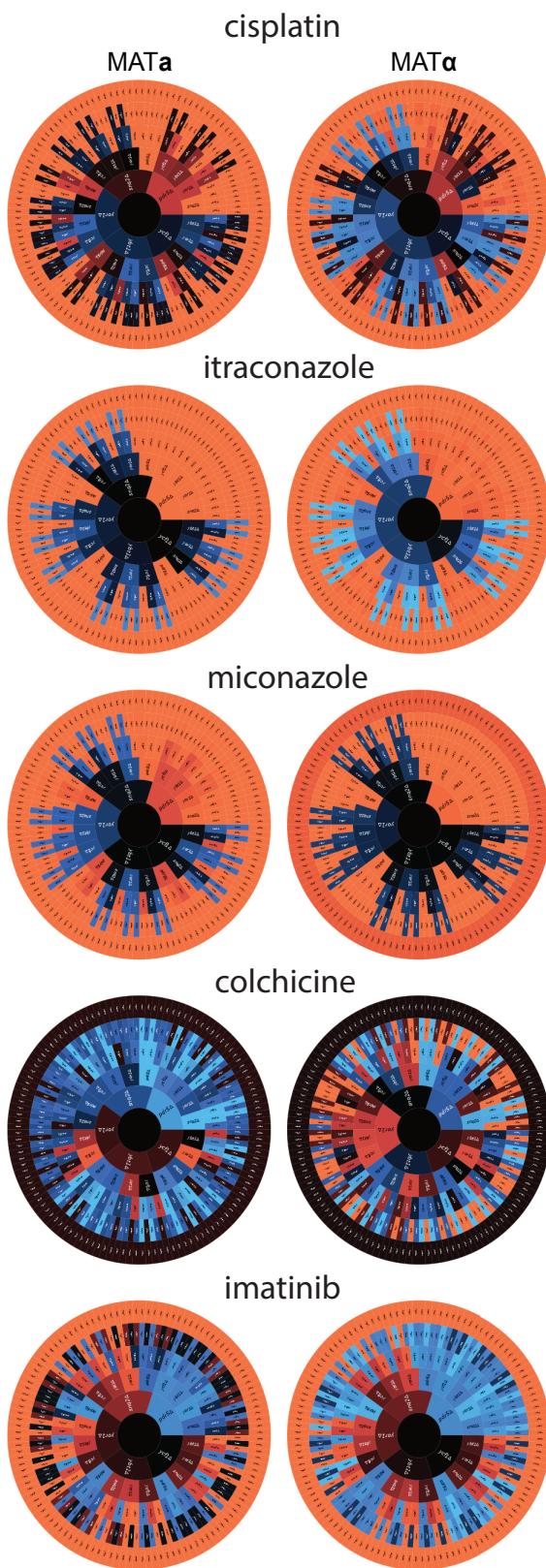


# Figure S5



## Figure S6

Resistance  
wt  
Sensitivity



**Figure S7**

**A**

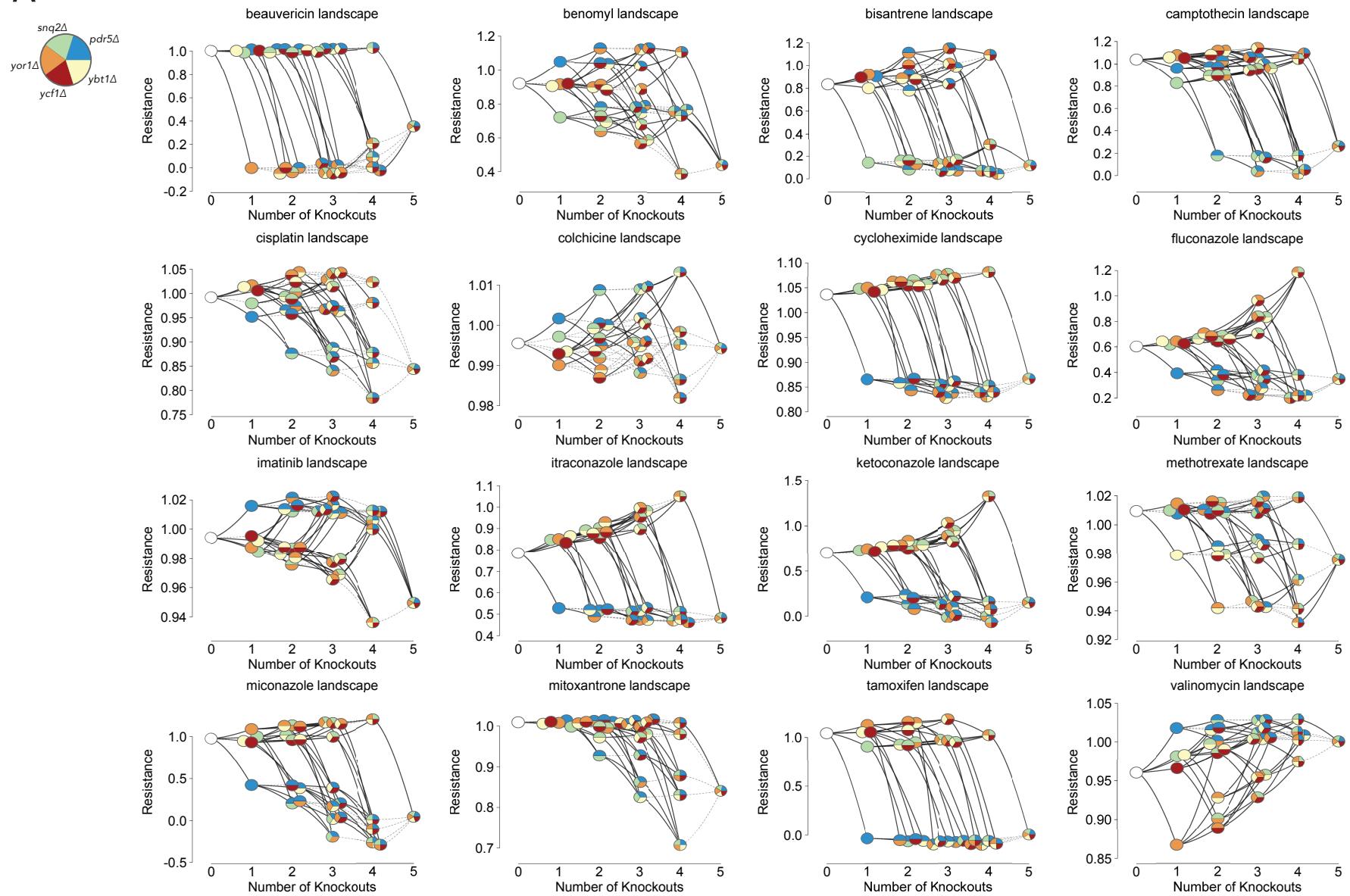
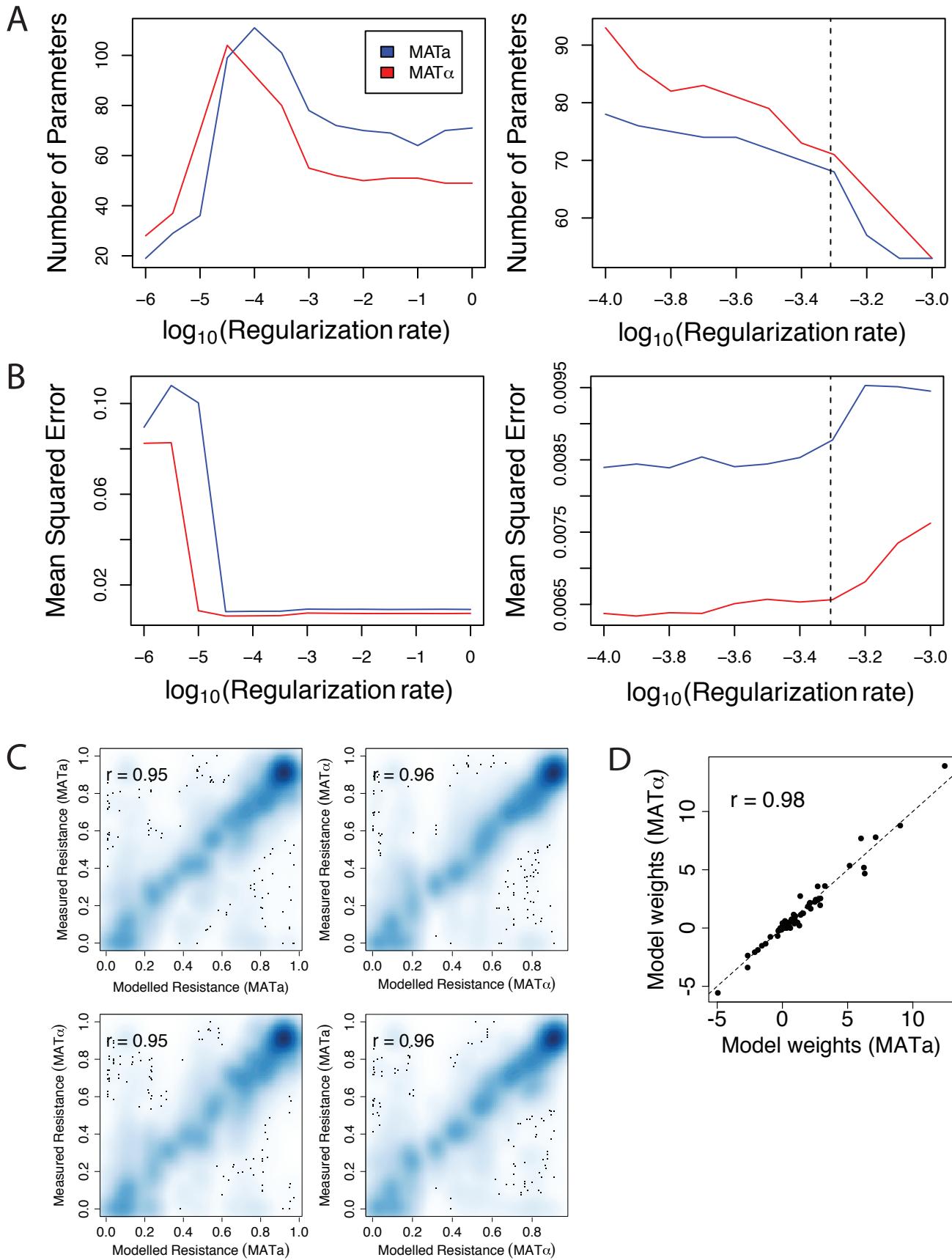


Figure S8



# Figure S9

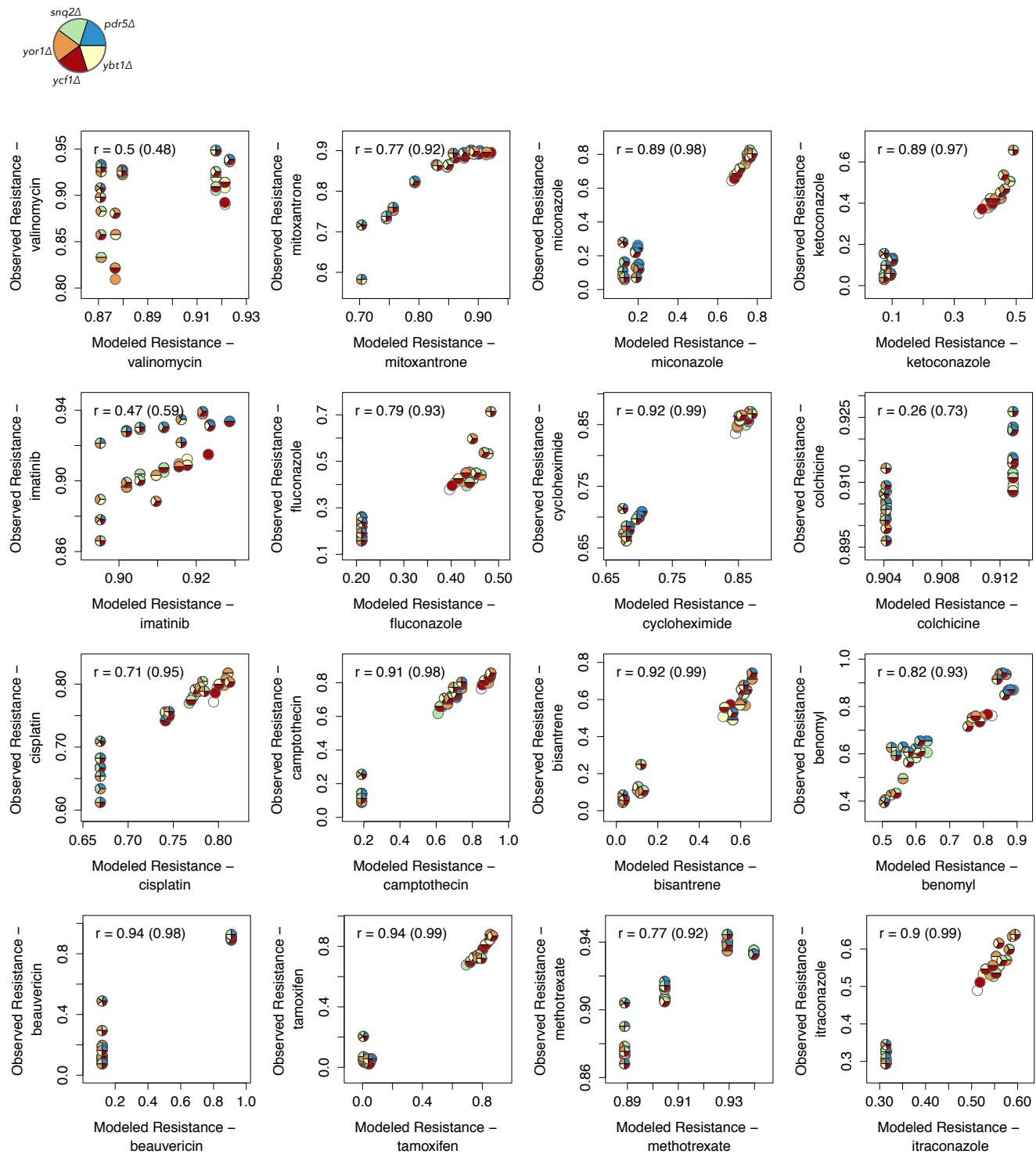
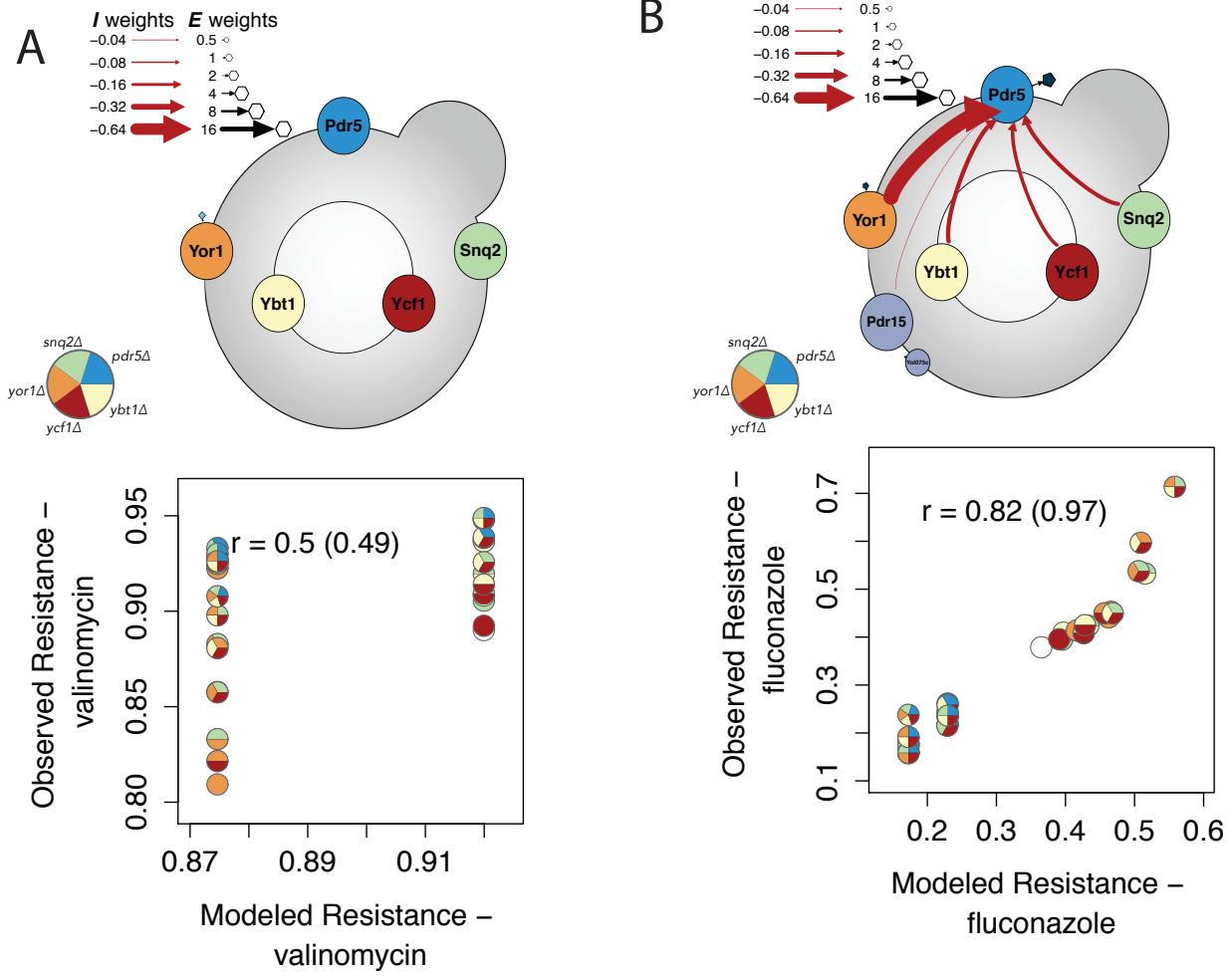
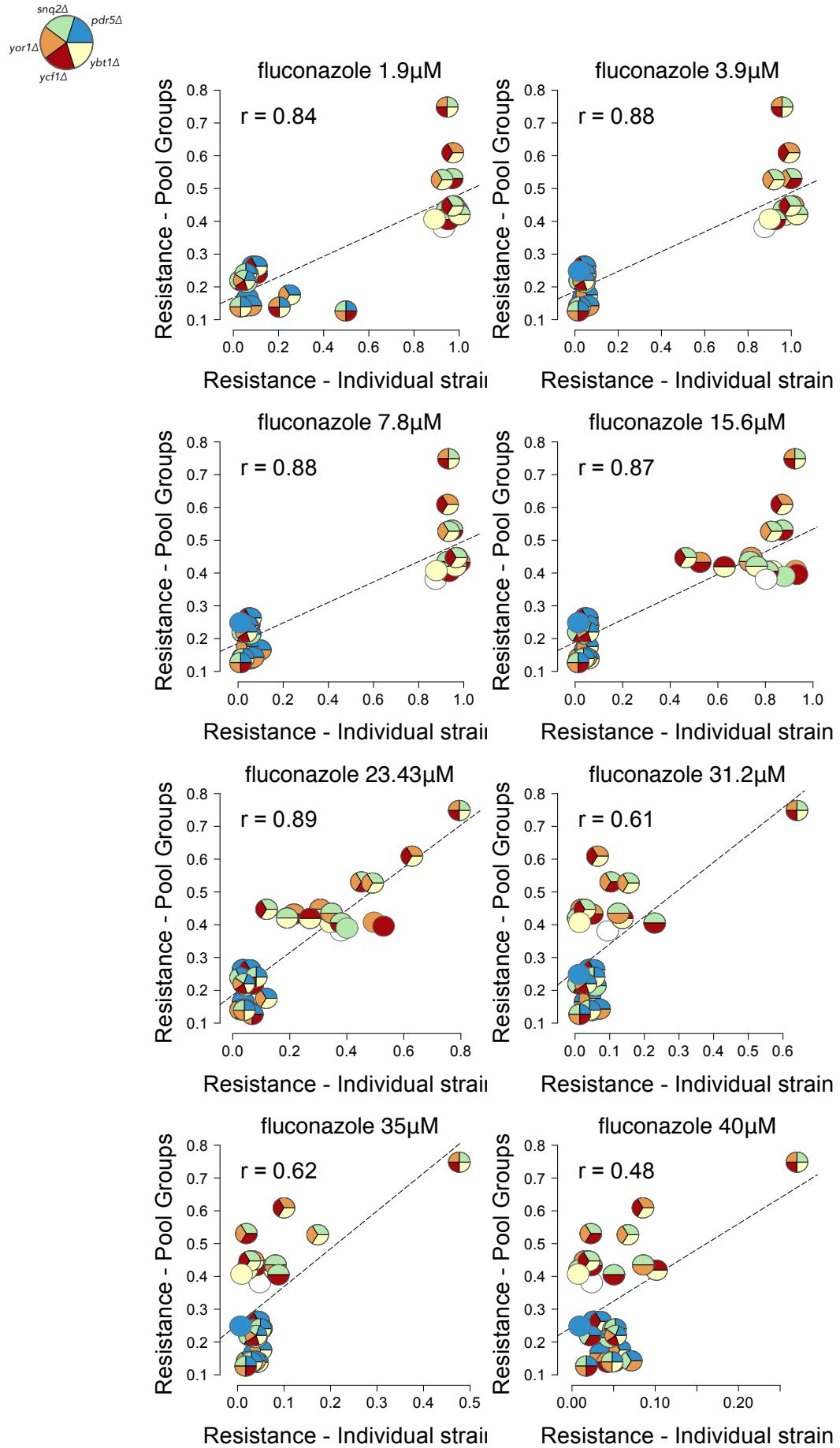


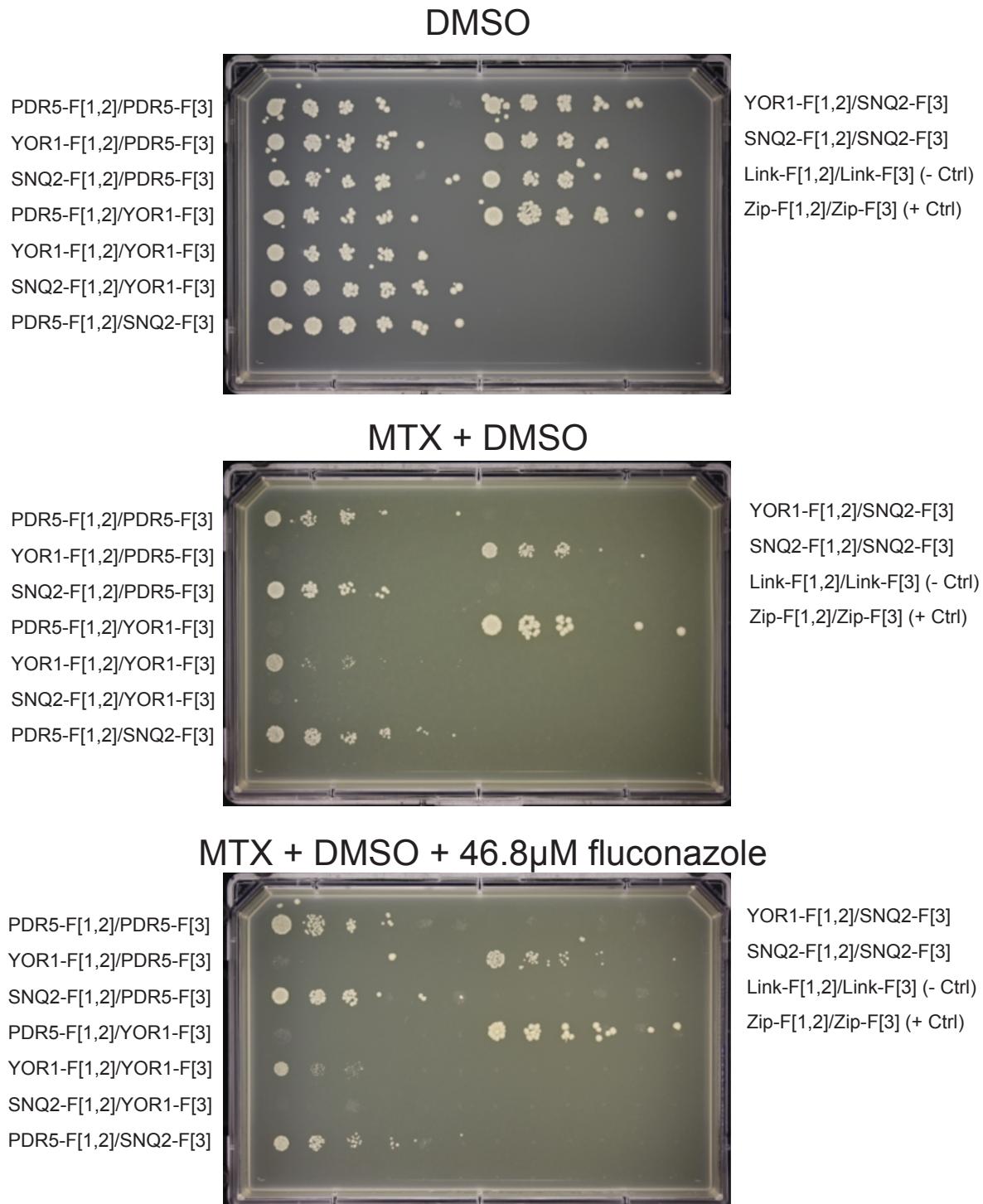
Figure S10



**Figure S11**



# Figure S12



**Figure S13**

