

# Deciphering Complex Traits with High-Order 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>, Nishka Kishore<sup>1,2</sup>, Igor Stagljari<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.)

## Summary

Pairwise (‘two-gene’) genetic interaction profiling has proven useful for functional inference, but misses all functional relationships revealed by higher-order interactions involving multiple genes. Here, we describe a high-order ‘*X*-gene’ genetic analysis (XGA) that uses many combinations of genetic perturbations at multiple target loci to functionally decipher complex traits. We demonstrate XGA on yeast ABC transporters by engineering 5,352 strains bearing random combinations of deletions at 16 transporters, and profiling each strain’s resistance to 16 bioactive compounds (‘drugs’). These 85,632 genotype-to-resistance measurements revealed high-order drug-dependent genetic interactions for 13 of the 16 transporters studied. We trained a neural network to derive intuitive system models from these complex genetic relationships. XGA-based modeling guided further characterization of a five-gene fluconazole resistance phenomenon. Together, results demonstrate that XGA can help dissect complex functional relationships within multi-gene systems.

## Introduction

Genes often encode interdependent and functionally-overlapping molecular systems, such that combinations of genetic variants can yield surprising phenotypes (Hartman et al., 2001). This phenomenon defines genetic interaction. Observing an interaction when knocking out two genes in pairs, e.g., using synthetic genetic array analysis (SGA), has systematically uncovered functional relationships 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., 2016) and order-of-action in biological pathways (St Onge et al., 2007).

More complex genetic interactions, e.g. surprising phenotypes arising from three-gene knockouts, can reveal additional important functions (Haber et al., 2013; Kuzmin et al., 2018). Indeed, three-gene interactions are expected to outnumber two-gene interactions by 100-fold (Kuzmin et al., 2018). Even higher-order gene variant combinations have yielded interesting effects – e.g., those involving four (Takahashi and Yamanaka, 2006), five (Taylor et al., 2014), seven (Beh et al., 2001), and over twenty genes (Wieczorke et al., 1999). Systematic mapping of higher-order interactions between multiple variants at one locus has further demonstrated their importance in yielding functional understanding (Baeza-Centurion et al., 2019; Domingo et al., 2018; Sarkisyan et al., 2016). However, higher-order interactions between multiple genes remain poorly characterized, limiting genetic inference within systems that show complex multi-gene functional dependencies.

To permit in-depth functional inference within multi-gene systems, we present a high-order ‘X-gene’ genetic analysis (XGA) strategy that generalizes beyond the use of one- and two-gene effects (i.e. ‘1GA’ and ‘2GA’). We demonstrate an XGA on yeast ABC transporters, involved in cellular efflux of small molecules, for which several informative higher-order genetic effects have been reported (Khakhina et al., 2015; Kolaczowska et al., 2008; Suzuki et al., 2011). More specifically, we apply XGA systematically to the entire set of 16 ABC transporters that have been implicated in multi-drug resistance. By revealing a multi-knockout genetic landscape for 16 bioactive compounds (‘drugs’), XGA uncovered many drug-dependent high-order genetic interactions involving as many as five genes. A neural network trained on XGA data provided an intuitive genotype-to-phenotype model and functional insights into this system of ABC transporters. Taken together, our results highlight the ability of XGA to systematically uncover high-order genetic relationships, and use them shed light on complex molecular systems.

## Results

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

Here we briefly describe our overall XGA strategy (Figure 1), then show results of applying XGA to sixteen 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 (‘segregants’) will inherit a random subset of parental variants and each segregant can then be genotyped and profiled for traits such as gene expression or drug resistance (Bloom et al., 2013). Although applied extensively to pairs of outbred parents, this approach has limitations. First, many genes involved in a process will be missed for lack of functional natural variation in the parental strains (Lee et al., 2014). Second, the fact that parents can differ at  $>10^5$  positions, coupled with genetic linkage of proximal variants, can complicate identification of the causal variants at each associated locus.

To exploit the power of cross-based approaches while avoiding limits of using natural variation, we designed a population engineering strategy in which targeted polygenic variation is engineered into individual parental strains. Parental strains are then crossed to yield an ‘engineered population’. Thus, a genetic cross yields random segregation only for engineered variants.

As recently reviewed (Kebschull and Zador, 2018), many individual strains can be tracked in a complex heterogenous population using DNA barcodes. We therefore introduced a complex pool of random barcodes into a haploid parental strain (which was wildtype for all ABC transporter genes of interest in this study), as described previously (Díaz-Mejía et al., 2018) (Figure S1, see Methods for details). 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 SGA 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 ~5,000 MAT $\alpha$  and ~5,000 MAT $a$  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 this arrayed population, we determined the genotype at all sixteen knockout loci and identified the barcode. To genotype, we exploited the fact that each knockout locus in the ABC-16 strain was derived from the YKO yeast deletion collection (Giaever et al., 2002; Suzuki et al., 2011) and is therefore flanked by a deletion-identifying barcode. We adapted the previously-described row-column-plate PCR (RCP-PCR) strategy (Yachie et al., 2016), in which barcodes in each segregant are amplified together with additional PCR-introduced index tags that identify the plate, row, and column of origin for each amplification product (Methods; Figure 1). Thus, a single sequencing experiment revealed 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).

Two independent methods estimated overall per-locus genotyping accuracy to be from 93.2%-93.8% (Figure S2A and 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. Considering only those strains with both high-quality genotyping data and at least one unique tracking barcode, our engineered strain population included 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.

### [Fritz Stopped Here]

To profile each strain's resistance or sensitivity to 16 different antifungal or anticancer agents ('drugs'; Data S3), we grew the strain pools competitively in each drug, and in a solvent (DMSO) condition. We used high-throughput strain barcode sequencing at five points (corresponding to 0, 5, 10, 15, and 20 generations of overall pool growth, Figure 1), to estimate growth rate and resistance for each strain in each drug (Data S5; Methods). We performed additional filtering steps, limiting analysis 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), and further excluding 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 2,985 MAT $\alpha$  strains, for each of the 16 drugs (Data S5).

### Grouped combinatorial profiles illustrate a complex drug-dependent genetic landscape

For an initial analysis, we identified and quantitatively modeled associations between individual ABC transporter knockouts and drug resistance phenotypes using a generalized linear model (see Methods). We found 62 resistance-knockout associations that were reproducible in both MAT $\alpha$  and MAT $\alpha$  pools (Figure S3A). Most (58/62) of these associations involved five 'frequently-associated' ABC transporters—*snq2 $\Delta$* , *pdr5 $\Delta$* , *yor1 $\Delta$* , *ycf1 $\Delta$* , and *ybt1 $\Delta$*  (Figure S3A). For these five frequently-associated transporters, we detected 16/18 previous associations between drug resistance and individual knockouts, while revealing 40 new associations (Figure S3A; Data S7). For example, we detected 18 drug-knockout associations involving the vacuolar ABC transporters *YCF1* and *YBT1*, all of which were novel (Figure S3A, Data S6). In contrast, only weak associations were found between each knockout growth rate in the DMSO control condition ( $<2\%$  decrease), with the exception of *yor1 $\Delta$* , which had a modest association (7-15% decrease; Figure S3B, Data S6).

Considering only the five frequently-associated transporters, we profiled resistance for the set of 32 ( $2^5$ ) combinatorial genotypes. We derived a resistance profile for each combinatorial genotype by calculating, for each drug, the average resistance over all strains matching this genotype at all five genes. These grouped resistance profiles showed striking correspondence when calculated separately for MAT $\alpha$  and MAT $\alpha$  pools (Figure S4). For example, camptothecin and ketoconazole had a very high profile correlation of  $r \geq 0.99$  (Figure 2A). Indeed, with the exception of colchicine ( $r = 0.77$ ), all drugs showed high reproducibility ( $r \geq 0.94$ ) between independent biological replicate pools (Figure 2B). The five-gene resistance profiles could be used to provide a visual overview of multi-knockout resistance in each drug. We created a 'radial combinatorial signature' for each drug, in which the consequences of knocking out increasingly-many ABC transporters

are shown as paths leading outward from the central wild-type genotype (Figure 2C-D). As expected, radial combinatorial signatures were visually similar between independent MATa and MATα populations for many drugs (Figure 2D and S5). Given high reproducibility, we merged MATa and MATα data for subsequent analyses, except where noted (Methods).

To analyze the five-gene combinatorial resistance profiles in more detail, we visualized them as fitness landscapes (Figure 3 and S6). For some drugs, we observed a clear sensitivity effect from knocking out only one transporter – e.g. *pdr5Δ* for cycloheximide and tamoxifen (Figure 3). In other drugs, we saw increased sensitivity resulting from knocking multiple transporters – e.g. the set {*snq2Δ*, *pdr5Δ*} under camptothecin, and the set {*snq2Δ*, *pdr5Δ*, *ybt1Δ*, *yor1Δ*} under mitoxantrone (Figure 3). These sensitivity patterns are consistent with a scenario in which one or more transporters can efflux a given drug.

In other drugs, the fitness landscapes showed more surprising multi-knockout patterns conveying both drug resistance and sensitivity. In benomyl, for example, we not only observed sensitivity from knocking out the known primary efflux pump *snq2Δ* (20% decreased resistance,  $p = 5.8\text{e-}80$ ; Wilcoxon rank sum test), but 13% increased resistance in *pdr5Δ* knockouts ( $p = 1.5\text{e-}96$ ), and further resistance in the *pdr5Δ yor1Δ* double-mutant (21% increased resistance;  $p = 1.3\text{e-}72$ ). All of these effects had been previously reported (Kolaczowska et al., 2008; Snider et al., 2013), and were explained by increased *SNQ2*-mediated resistance upon deleting *pdr5Δ* and *yor1Δ*. Consistent with this explanation, these resistance effects were more modest in a *snq2Δ* background (Figure 3). A similar landscape was found in bisantrene, which also showed a strong *snq2Δ* effect (Figure S6). Indeed, the successive deletion of ABC transporters led to greater resistance for surprisingly many drugs (Figure 2D and S5). For example, knocking out *pdr5Δ*, *snq2Δ*, *ybt1Δ*, and *ycf1Δ* individually or in any combination led to more valinomycin resistance than the wild-type strain (Figure 3).

### **XGA reveals many complex drug-dependent genetic interactions**

To identify and model multi-gene knockout effects at all 16 transporters, we extended the above generalized linear model used to capture both single knockout effects and multi-gene interactions (see Methods). All associations and interactions that passed the significance test ( $p < 0.05$  after adjusting for multiple testing) are shown in Figure 3A.

This analysis yielded genetic interactions involving two or more genes for fifteen out of sixteen (94%) of the drugs examined (Figure 3A). (The exception was beauvericin for which we only recovered the previously-reported sensitivity of *yor1Δ* knockouts (Shekhar-Guturja et al., 2016).) Higher-order genetic interactions (involving three or more genes) were observed for fourteen of sixteen (88%) drugs tested (Figure 3A). Thus, XGA revealed higher-order genetic interactions involving three or more genes for nearly all drug resistance phenotypes studied. In total, genetic interactions were found for 14 of the 16 genes targeted by XGA. Of these 14 genes, 13 were involved in at least one interaction involving three or more genes. Remarkably, 11 of the 16 targeted genes were involved in at least one 5-gene interaction.

Formally identifying complex genetic interactions captured many of the effects that had been readily-apparent by manual examination of the five-gene fitness landscapes, while yielding additional effects. For example, *pdr5Δ* was found to have a resistance effect under benomyl, to

have a positive genetic interaction with *yor1Δ* and, to have a negative genetic interaction with *snq2Δ* (Figure 3A, Data S6).

This analysis uncovered strong complex interactions involving genes outside the five-frequently associated transporters. In both cisplatin and mitoxantrone, for example, a five-way positive interaction pointed to the phenomenon that addition of *bpt1Δ* was found to confer resistance in a sensitive *pdr5Δ snq2Δ ycf1Δ yor1Δ* background (Figure 3B). A five-way positive interaction in bisantrene pointed to a similar, but more modest effect with *vmr1Δ* in a *snq2Δ ybt1Δ ycf1Δ yor1Δ* background (Figure 3B).

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 (Figure 3) was modeled as the combination of small negative 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 3B, Data S6). Together, these complex negative genetic interaction patterns suggest that the four genes enable mitoxantrone efflux in parallel. Similar ‘parallel efflux’ genetic interaction pattern were observed, e.g. for  $\{pdr5Δ, snq2Δ\}$  in camptothecin, and  $\{pdr5Δ, snq2Δ, yor1Δ\}$  in cisplatin (Figure 3A-B, Data S6).

### Using XGA to learn intuitive genotype-to-phenotype models of the ABC transporter system

Above, we used generalized linear models to capture complex genotype-phenotype relationships as a set of single-gene effects and  $X$ -way genetic interactions. However, these models did not efficiently convey useful intuition about the system. For example, we reasoned (without benefit of computational modeling) a set of transporter genes showing patterns of within-set negative genetic interactions suggests that each transporter is independently capable of drug efflux. Other genetic interaction patterns led us to conclude that the presence of one transporter can positively or negatively influence the activity of another (e.g. influence on Snq2 activity from *PDR5* and *YORI* in benomyl). However, manually derived genetic intuition from a complex system is laborious, and it is difficult to objectively evaluate how well these explanations correspond to observed data. To demonstrate that such intuitions can be derived more systematically from complex genotype-phenotype relationships, we developed a neural network model.

We structured the 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 regulatory influence relationships between transporters, the links between genotype and activity layers have (initially unknown) ‘influence’ 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 (initially unknown) non-negative ‘efflux’ weights ( $\mathbf{E}$ ) that capture the extent to which each transporter can catalyze the efflux (or otherwise reduce the activity) of each drug. The model also allowed for offset terms for both  $\mathbf{A}$  and  $\mathbf{R}$  (Methods).

We aimed to learn these network weights with back-propagation with stochastic gradient descent using our complete set of drug resistance phenotypes as training data (see Methods). To favor more parsimonious models, the cost function that was used to optimize network weights contained a penalty which acts to limit the number of non-zero weights (Methods, Figure S7A-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 our input dataset of 85,632 genotype-phenotype measurements, we learned an interpretable neural network with only 73 non-zero fitted parameters (6 *I* weights, 51 *E* weights, no *A* offset terms and 16 *R* offset terms). Despite its parsimonious nature, the trained neural network model largely recapitulated the observed genotype-phenotype relationships ( $r = 0.96$ , Figure 4C). To test that this performance generalizes to unseen data, we also trained the model using data from one mating type then tested it using independent data from the other mating type. This yielded similar performance ( $r = 0.95$  and  $r = 0.96$  when using either mating type **a** or  **$\alpha$**  as training, respectively [Figure S7C]), and the resulting independently-trained models also showed strong agreement in parameter values ( $r = 0.98$ , Figure S7D), suggesting that model parameters were robustly determined.

The first striking result from this model was that all influence (*I*) values were either zero or negative. More specifically, only 6 out of 240 influence values were negative, while all others were zero. Thus, while some ABC transporters exhibited negative influence on other ABC transporters, our study of 16 transporters in 16 drugs found no evidence that any ABC transporter can positively influence any other ABC transporter.

The objectively-trained neural network 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 ( $p = 1.2\text{e-}70$ ). 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), which is reflected, for example, by the observation that *pdr5 $\Delta$*  yields greater benomyl resistance than does *yor1 $\Delta$*  (Figure 3).

While the neural network model was accurate overall, we could also evaluate where predictions departed systematically from observation (Figure S8), guiding model refinements. For example, while XGA showed that many multi-transporter deletions result in *increased* valinomycin resistance (Figure 3), the neural network did not predict increased valinomycin resistance for any genotype, yielding poor predictions overall ( $r = 0.5$ , Figure 4D, left panel). Together with previous reports of valinomycin resistance upon deletion of all 16 transporters (Suzuki et al., 2011), we hypothesized that one or more of the targeted 16 transporters inhibits a valinomycin

resistance factor outside of the 16 ABC transporters. To formally test whether this explanation better captures the observed data, we added one additional hypothesized valinomycin resistance factor and its corresponding activity node to the neural network (see Methods). Training this extended neural network using only valinomycin data substantially improved recapitulation of the observed phenotypes ( $r = 0.86$ , Figure 4D, right panel) and yielded a model in which *SNQ2*, *PDR5*, *YBT1*, and *YCF1* each negatively influence an unknown valinomycin resistance factor. This model improvement depended on the inclusion of this hypothesized factor (Figure S9A).

#### **Four genes for which deletion yields synergistic *PDR5*-dependent fluconazole resistance**

One striking phenotype revealed by XGA was a quadruple deletion—*snq2Δ ybt1Δ ycf1Δ yor1Δ*—with high resistance to both fluconazole (Figure 5A, top panel) and ketoconazole (Figure S6). 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. Generalized linear regression had modeled the quadruple-knockout resistance phenomenon as the combination of three positive three-gene interactions (all of the three-knockout combinations of  $\{yor1\Delta, snq2\Delta, ybt1\Delta, ycf1\Delta\}$  except *snq2Δ ybt1Δ ycf1Δ*), while its dependence on *PDR5* was encoded by three two-way negative interactions:  $\{pdr5\Delta, snq2\Delta\}$ ,  $\{pdr5\Delta, ycf1\Delta\}$ , and  $\{pdr5\Delta, yor1\Delta\}$ .

Overall, these observations were consistent 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), thereby capturing that *snqΔybt1Δycf1Δyor1Δ* 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 with 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. We extended the ‘additive influence’ model by adding a neuron that acts to mediate influence on Pdr5 from *SNQ2*, *YBT1*, *YCF1*, and *YOR1* (see Methods). Using only fluconazole resistance data, training a network with an added ‘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 S9 B). These results also suggested that at least some of the negative influence of these four genes on Pdr5 activity is mediated via a hidden factor.

To confirm these initial observations in a more uniform genetic background, we generated a single strain for each of the 32 possible combinations of *pdr5Δ*, *snq2Δ*, *yor1Δ*, *ybt1Δ*, and *ycf1Δ* knockouts, each in a common genetic background (Methods). Fluconazole resistance as estimated by pooled data 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 S10; Methods). Consistent with pooled results, *snq2Δyor1Δybt1Δycf1Δ* had the highest resistance.



We explored two 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.

Consistent with the idea that a hidden factor mediates (at least some of) the negative influence from the four genes to Pdr5, previously-reported inhibition of Pdr5 activity by *SNQ2* and *YOR1* was explained by reduced *PDR5* transcript levels mediated by the transcription factor Pdr1 (Kolaczowska 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* (Kolaczowska et al., 2008), the change did not reach statistical significance in our hands ( $p = 0.27$ , Figure 5D) and the previous report did not contain a statistical test. No evidence of increased *PDR5* mRNA level in the *ybt1Δycf1Δ* strain was observed (0.9 fold expression,  $p = 0.69$ , Figure 5D). More persuasively, we found that *PDR5* mRNA levels were significantly higher in *snq2Δyor1Δybt1Δycf1Δ* than in wild type ( $2.1\times$  increase;  $p = 0.032$ ; Figure 5D). 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* yields reduced *PDR5* mRNA levels, they are consistent with the idea that combinations of *snq2Δ*, *yor1Δ*, *ybt1Δ*, *ycf1Δ* non-linearly influence fluconazole resistance by modulating *PDR5* transcript levels.

However, it has been suggested that transporter genes can also negatively influence one another by non-transcriptional means. For example, a previously study found that while *pdr5Δ* and *yor1Δ* each showed increased benomyl resistance, evidence for mRNA induction was found only in *pdr5Δyor1Δ* (Snider et al., 2013). Given the known protein-protein interaction between Pdr5 and Snq2 (Snider et al., 2013), one might hypothesize that *PDR5*-mediated repression of Snq2 activity results at least in part from a direct interaction. This ‘direct repression’ model, in which heterodimerization of Pdr5 and Snq2 transporters draws subunits away from the homodimeric Snq2 complex, thereby reducing the total efflux activity of Snq2, draws support from the observed homodimeric interactions of Pdr5, Snq2, and a heterodimeric interaction between Pdr5 and Snq2.

We next explored whether a direct repression model could explain the observed Pdr5-dependent decrease in fluconazole resistance provided by *YOR1*. This model predicted a heterodimeric interaction between Pdr5 and Yor1. This interaction 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 used two distinct protein interaction assays: MYTH (Snider et al., 2013) and PCA (Tarassov et al., 2008) to test the Pdr5-Yor1 interaction. All previously-known MYTH and PCA interactions amongst Pdr5, Snq2, and Yor1 were recovered (Figure 5E, S11, S12). Although PCA (Figure S11) did not detect the predicted Pdr5-Yor1 interaction, MYTH could detect this interaction (Figure 5E, S12), thus confirming a key prediction of the direct repression model for the Pdr5-dependent decrease in fluconazole resistance provided by *YOR1*.

Given a much-higher baseline abundance of Pdr5 than Snq2 (Wang et al., 2015), a heterodimeric repression model is also 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 (Celaj et al., 2017). 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.

## Discussion

Here we presented an implementation of XGA, an extension of the SGA approach, that enables systematic high-order genetic analysis by engineering polygenic variation into a population and profiling the resulting phenotypes.

XGA of 16 ABC transporters uncovered complex genetic phenomena that were not evident from one- and two- gene knockout effects, and the resulting data could be used to train a neural network to derive an intuitive system model of ABC transporters, and provide functional insight. The engineered population provides a readily-available resource for continued ABC-transporter XGA. Broadly, we illustrate the potential for carrying out XGA in other sets of functionally-related genes.

The demonstrated cross-based XGA approach can be adapted with multi-variant strains in many model organisms. An XGA of GPCR signaling, for example, could be performed using an existing yeast mutant with 16 GPCR pathway-related genes deleted (Shaw et al., 2018). With CRISPR, variation in yeast can be engineered at up to five loci with a single transformation (Jakočiūnas et al., 2015), facilitating construction of multi-variant strains. Indeed, simultaneous variant engineering at 3 – 6 loci has been described in multiple model organisms: e.g. mouse (Wang et al., 2013), zebrafish (Jao et al., 2013), *C. elegans* (Xu et al., 2016), and *Arabidopsis* (Zhang et al., 2016).

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 for higher combinatorial complexity at large scale. In human cells, methods developed for single-cell sequencing also show promise for directly-engineered XGA. Indeed, it is already possible to profile single-cell RNA levels (Dixit et al., 2016) and chromatin state (Rubin et al., 2018) following combinatorial gene disruptions, although this has not been applied for large numbers of *X*-gene variants. In addition to permitting richer phenotyping, 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 would grow to  $>10^7$  combinations if heterozygotes were considered.

Many future expansions can be envisioned even for the XGA study of yeast ABC transporters. For example, we hypothesized that some negative influences on Pdr5 may be mediated by a Pdr1-dependent transcriptional response (Kolaczowska et al., 2008). Future work could include *en masse* deletion of *PDR1* in each pool, thereby expanding the XGA targets to explicitly test the

extent of its mediating role. Higher-content phenotyping could also 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, there is also evidence for complex influence between ABC transporters 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 XGA of human ABC transporters could yield 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 modeling of the complex genetic landscape can be used to construct and evaluate genotype-to-phenotype models. 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; 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 (Otwinski et al., 2018). For example, our neural network could model complex negative interactions in terms of the influence of ABC transporters on an indirectly-observed drug efflux activity. In the case of valinomycin, we could greatly improve models by hypothesizing the influence of multiple ABC transporters via an unknown valinomycin resistance factor. Neural networks 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 interpret, here we showed that they can be simply constructed using directly-interpretable parameters (e.g. ‘activity’, ‘influence’, and ‘efflux’), and then iteratively expanded as needed.

In summary, we showed that measuring and modeling rich genotype-to-phenotype relationships via XGA can help to functionally dissect and understand a complex system.

## Star Methods

### Key Resource Table

Reagent or Resource	Source	Identifier
<b>Experimental Models: Organisms/Strains</b>		
RY0622	Suzuki et al., 2011	N/A
RY0146	Suzuki et al., 2011	N/A
RY0566	Suzuki et al., 2011	N/A
RY0148	Suzuki et al., 2011	N/A

Barcoded RY0148 pool	This paper	N/A
<b>Chemicals, Peptides, and Recombinant Proteins</b>		
fluconazole	Sigma-Aldrich	F8929
ketoconazole	Sigma-Aldrich	K1003
miconazole	Sigma-Aldrich	1443409
itraconazole	Sigma-Aldrich	I6657
beauvericin	Sigma-Aldrich	B7510
tamoxifen	Sigma-Aldrich	T5648
benomyl	Sigma-Aldrich	45339
cycloheximide	Sigma-Aldrich	C1988
methotrexate	Sigma-Aldrich	M9929
camptothecin	Sigma-Aldrich	C9911
cisplatin	Sigma-Aldrich	P4394
bisantrene	Sigma-Aldrich	B4563
mitoxantrone	Sigma-Aldrich	6545
colchicine	Sigma-Aldrich	9754
imatinib	Sigma-Aldrich	270784
valinomycin	Sigma-Aldrich	V3639
<b>Oligonucleides</b>		
All DNA primers used, see Data S1	This paper	N/A
<b>Recombinant DNA</b>		
Plasmid: pSH47	Euroscarf	P30119
Plasmid: pIS420	Euroscarf	P30575
<b>Software and Algorithms</b>		
Analysis pipeline (written in R)	This paper	<a href="https://github.com/a3cel2/xga">https://github.com/a3cel2/xga</a>

## Contact For Reagent and Resource Sharing

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Frederick P. Roth (fritz.roth@utoronto.ca).

## Experimental Model and Subject Details

### *Saccharomyces cerevisiae* strains

RY0622 (ABC-16/‘Green Monster’ MATa):

*MATa 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):

*MATa lyp1Δ his3Δ1 leu2Δ0 ura3Δ0 met15Δ0 can1Δ::GMToolkit-a (CMVpr-rtTA KanMX4 STE2pr-Sp-his5)*

RY0566 (“Toolkit-a” strain with tet-inducible GFP-URA3):

*MATa lyp1Δ his3Δ1 leu2Δ0 ura3Δ0 met15Δ0 can1Δ::GMToolkit-a (CMVpr-rtTA KanMX4 STE2pr-Sp-his5) hoΔ::tetO2-GFP-URA3*

RY0148 (“Toolkit-**α**” strain):

*MATa lyp1Δ his3Δ1 leu2Δ0 ura3Δ0 met15Δ0 can1Δ::GMToolkit-α (CMVpr-rtTA NatMX4 STE3pr-LEU2)*

Barcoded RY0148 pool:

*MATa lyp1Δ his3Δ1 leu2Δ0 ura3Δ0 met15Δ0 can1Δ::GMToolkit-α (CMVpr-rtTA NatMX4 STE3pr-LEU2) hoΔ::loxP UP-tag HphMX4 DN-tag lox2272*

## Methods Details

### 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 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 a Qiagen Qiaspin kit. 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 gel electrophoresis. EXOSAP purification was performed on the PCR products, and they 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 (MATa) and the barcoded RY0148 pool (MATα) as previously described (Suzuki et al., 2011), selecting for MATa and MATα 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 MATa and 5,461 MATα colonies were picked onto 384 well plates. In addition, 299 known positions in both the MATa and MATα 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 µl of overnight yeast culture was mixed with 8 µL 0.2 M sodium phosphate buffer (pH 7.4), 4 µl DNA free dH<sub>2</sub>O, 0.05 µl 5 U/µl Zymoresearch zymolyase and incubated at 37 °C for 35 minutes. 64 µl 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 µl 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

combined, and 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  $\mu$ l 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 MinElute Gel Extraction kit (Cat. No.28604), and 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  $\mu$ l of this multiplexed sample were run on a 4% gel. Products of the desired size (260-290 bp) were isolated from each lane, purified using a Qiagen MinElute Gel Extraction 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, genotyping data from a single well was assigned to the barcode at random.

### **Refining Genotype for Putative Wild-Type Strains**

For 131 MAT $\alpha$  and 73 MATa 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 (indicating the absence of any knockout cassettes), 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 MATa strains, all exhibited growth in SC-Ura (indicating a knockout cassette at one or more loci). Individual genotyping was performed for these MATa 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 untested strains were discarded from analysis. When calculating linkage and distribution of gene knockouts (Figure S2C), 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 wild-type to mutant miscalls are negligible, 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 S2B).

### **Individual liquid growth profiling**

To measure individual strain growth, 100 $\mu$ l of starting culture at 0.0625 OD<sub>600nm</sub> was grown in a 96 well-plate in a temperature-controlled shaking spectrophotometer (Tecan GENios microplate reader). Growing cultures were shaken at 800 rpm at 30°C and OD<sub>600nm</sub> of each well was measured every 15 min.

### **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 + 15% 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. In addition, a ‘0 generation’ sample was immediately harvested from the YPD + glycerol pool and processed for DNA extraction and sequencing. After growth to approximately 2 OD<sub>600 nm</sub> (~5 generations), cells were collected and processed for sequencing, and a small aliquot was diluted in fresh media (at a final concentration of 0.0625 OD<sub>600 nm</sub> in 10ml) in presence of drug or solvent to be grown for an additional 5 generations. This process was repeated for a total of 20 generations of growth, with collected samples corresponding to approximately 5, 10, 15, and 20 generations.

Harvested samples were subject to genomic DNA extraction using a YeaStar™ Genomic DNA Kit, quantified using the Invitrogen Quant-IT dsDNA BR Assay kit (Cat No. Q32853), and diluted to a final concentration of 25 ng/ $\mu$ l. 350ng of DNA from each sample was indexed with the following PCR mixture: 20  $\mu$ l of 2x Platinum PCR SuperMix High Fidelity, 1  $\mu$ l of 10  $\mu$ M F primer, and 1  $\mu$ l of 10  $\mu$ M 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 were amplified using the following 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. After indexing, equal volumes of UP-tag and DN-tag PCR products from each pool were run on a 3% agarose gel. The expected 210bp bands were isolated and purified using a Qiagen MinElute Gel Extraction kit. DNA size and purity was confirmed by using the Agilent Bioanalyzer High Sensitivity DNA kit. DNA yield was quantified in triplicate using a KAPA SYBR FAST Universal qPCR kit (Cat. No. KK4824). Approximately equal amounts of each sample were combined and sequenced using an Illumina NextSeq 500 High Output v2 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.

We modeled 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:

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

Here,  $f_{s_x,d,0}$  is proportional to a desired ‘absolute’ starting abundance metric for each strain (e.g. number of cells),  $A_{s_x,d,0}$ , such that:

$$A_{s_x,d,t} = (k_{s_x} f_{s_x,d,0}) 2^{g_{s_x,d}t}$$

Similarly, we also obtain a proportional estimate for  $A_{s_x,d,t}$  by using the frequency at each time point, multiplied by expected relative pool growth at that time compared to time 0 ( $2^t$ , since  $t$  is defined by the number of generations since  $t = 0$ )

$$A_{s_x,d,t} = (k_{s_x} f_{s_x,d,t}) 2^t$$

Therefore, integrating  $f_{s_x,d,t} 2^t$  over time can be used to compute a metric that is proportional to the absolute ‘area under the growth curve’ ( $AUC$ ) from timepoints 0 to  $T$  (the total number of pool generations measured). Here, frequencies between measured timepoints were linearly interpolated:

$$\int_0^T A_{s_x,d,0} 2^{gt} = AUC_{s_x,d} = k_{s_x} \int_0^T f_{s_x,d,t} 2^t dt$$

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

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

Given the above relationships of the frequency-based metrics with absolute abundance-based metrics, we can substitute  $AUC_{s_x,d}$  with  $k_{s_x} \int_0^T f_{s_x,d,t} 2^t dt$  and  $A_{s_x,d,0}$  with  $k_{s_x} f_{s_x,d,0}$ . Therefore, the unknown constant relating frequency-based abundance estimates to absolute abundance estimates for each strain ( $k_{s_x}$ ) is not needed to compute growth rate ( $g_{s_x,d}$ ):

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

Because  $\int_0^T f_{s_x,d,t} 2^t dt$  and  $f_{s_x,d,0}$  are both known, we then numerically solve for the  $g_{s_x,d}$  which satisfies this relationship using the `optimize()` function in R, setting a minimum of -10 and maximum of 10 for the interval (there is no simple closed-form solution for this equation).  $g_{s_x,d}$  is an estimate of the growth rate relative to the pool as a whole (i.e. a strain with  $g_{s_x,d} = 1$  perfectly ‘keeps up’ with the pool). In practice,  $g_{s_x,d}$  estimates the average relative exponential growth rate from 0 to  $T$  (e.g. a prolonged lag phase effectively lowers the average exponential growth rate). For linear regression and neural network training, the minimum  $g$  is set to 1e-10 to avoid numerical errors in the respective algorithms. 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 range uncertainty in  $r_{s_x,d}$ , such that resistance estimates from the MATa and MATa pool may be highly correlated, but may differ in range in some drugs (Figure S5). To correct for any potential pool-of-origin effects in  $r_{s_x,d}$  arising from merging the MATa and MATa populations, we use the line of best fit derived in Figure S5 to rescale  $r_{s_x,d}$  estimates from the MATa pool to match the MATa pool before merging. Any small remaining effects are handled by plate-of-origin terms in linear modeling (below).

## 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:

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

To express this model 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:

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

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 (a difference-based metric is not readily generalizable to high-order effects).

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

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

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

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

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

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

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

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  $\{ABC1_\Delta \dots ABC16_\Delta\}$ :

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

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. Equation 7) contains a ‘plate of origin’ term for each strain to correct for any potential batch effects.

To perform the marginal association in Figure S4, we fit a model with only  $l_i$  terms (and plate-of-origin effects), 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 expansion” method. 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 (all plate-of-origin effects are re-added at each step). 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 expansion procedure is repeated until either  $n$  reaches 5, or the number of unique genes in the  $n - 1$  model is less than  $n$  (i.e. there are no more possible interaction terms to search for). After the stepwise expansion 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. Plate-of-origin effects are re-added to the model before stepwise elimination, and are subject to the same significance criteria.

### Defining a neural network system model

We will define an ‘efflux and influence’ 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 an 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 non-negative.

We then model influence 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 -  $A_{i,G}$ . We initially set a common set of influence relationships across drugs, so that the activity variable is modeled 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 influence. That is,  $A_{i,G}$  can be influenced by other ABC transporters:

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

Where  $I_j$  are the ‘influences’ from other ABC transporters. While the form of  $A_{i,G}$  is itself unknown, here we also modeled 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 non-negative  $E$  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 baseline  $A \approx 0$ , 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 $\beta$  pools (Figure S8B), while lowering this rate did not have a clear impact on mean-squared error 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  $\lambda$  (more complex models can again set a separate regularization rate here).

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, as we aimed to simply extend 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 highly reproducible between most initializations, we noticed that a small number of random initializations resulted in convergence to a different set of parameters with a higher mean-squared error, even at high regularization rates (data not shown). To better ensure robust parameterization we used the median between 10 runs rather than the average to assign weights to the final neural network, and did not employ the



$|Z_{param}|$  filter described above. Here, training with  $\lambda = 5 \times 10^{-4}$  resulted in a similar model as the two-layer network (data not shown). Therefore, we performed a separate ‘three-layer  $\lambda$ ’ search for this network, searching 13 intervals between  $10^{-6}$  to  $10^0$  (Figure S10C). For three-layer training, we found that  $\lambda > 10^{-5}$  negatively impacts MSE (Figure S10C), and therefore used a less-restrictive  $\lambda = 10^{-5}$  to train the three-layer network in Figure 5B. Learned weights were subject to the same statistical significance test as for the two-layer network.

### Targeted mating and selection to obtain 32 knockouts

The TWAS21230902 strain (genotyped as *pdr10Δ pdr18Δ pdr5Δ snq2Δ ybt1Δ ycf1Δ yor1Δ* by RCP-PCR; Data S2) was subject to individual strain genotyping (Suzuki et al., 2011), which confirmed the expected wild-type and knockout PCR products at each locus. This strain (MAT $\alpha$ ) was mated with RY0566 (MATa), and was subject to sporulation and MATa haploid selection (Suzuki et al., 2011). Individuals from this cross were arrayed onto 96 well plates, 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 Multiplex PCR Plus Kit 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 1.9, 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 (Snider et al., 2010). 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 (Snider et al., 2010). 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 recreated 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<sup>®</sup> 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 using the Thermo Scientific<sup>™</sup> Maxima<sup>™</sup> H Minus First Strand cDNA Synthesis Kit. qPCR on these samples was then performed using a Bioline SensiFAST<sup>™</sup> SYBR No-ROX qPCR kit and Ct values were quantified using a CFX machine. cDNA synthesis and qPCR was performed for *PDR5* and *UBC6* (which acted as loading control). Relative expression of *PDR5* in all strains was calculated as  $2^{C_qUBC6 - C_qPDR5}$ . For each strain,  $C_q$  values for the cDNA samples were quantified multiple times to assess technical variability ( $C_qPDR5$  was measured in triplicate, and  $C_qUBC6$  was measured in triplicate), and these multiple measurements were averaged before calculating relative expression. qRT-PCR was performed for three individual cultures of each strain in each genetic background. RY0566 was used as the wildtype.

### Data and Software Availability

R scripts used to perform computational analyses are available at: <https://github.com/a3cel2/xga>

### Competing Interests

The authors declare that they have no competing interests.

### Acknowledgements

We are grateful for helpful comments from [XX] at the outset of this project, to B. Andrews and C. Boone for providing reagents. We gratefully acknowledge support by the Canadian Excellence Research Chairs (CERC) Program and by National Human Genome Research Institute of the National Institutes of Health (NIH/NHGRI) grants CA130266 and HG004756. FPR was also supported by a Canadian Institutes of Health Research Foundation Grant, an NIH/NHGRI Center of Excellence in Genomic Science award (HG004233), by NIH/NHGRI Grant HG001715, and by the One Brave Idea Foundation

### Author Contributions

F.P.R., Y.S., N.Y., & A.C. conceived of this study. N.Y., M.G., L.M., T.F., P.B., G.S., S.Z., M.N., and V.W. performed experiments, with advice and assistance from A.G.C., J.S., I.S. and J.C.M.. A.C., M.K., and N.Y. analyzed the data, with advice from F.P.R.. 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 locus, perform genotyping, carry out RCP-PCR, and for adding index tags to sequencing libraries.

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

**Additional Data S3.** Chemical compounds used in this study and their concentration in the pooled growth experiments.

**Additional Data S4.** Growth and drug resistance measurements for all strains as inferred from barcode-sequencing readout of pooled growth experiments.

**Additional Data S5.** Summary of single-gene effects and genetic interactions as obtained by generalized linear modeling.

**Additional Data S6.** Previously-known associations for the five “frequently-associated” ABC transporters with resistance or sensitivity to the 16 drugs studied.

**Additional Data S7.** Neural network model parameters.

## **References**

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.

Baeza-Centurion, P., Miñana, B., Schmiedel, J.M., Valcárcel, J., and Lehner, B. (2019). Combinatorial Genetics Reveals a Scaling Law for the Effects of Mutations on Splicing. *Cell* *176*, 549–563.e23.

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.
- 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.
- Celaj, A., Schlecht, U., Smith, J.D., Xu, W., Suresh, S., Miranda, M., Aparicio, A.M., Proctor, M., Davis, R.W., Roth, F.P., et al. (2017). Quantitative analysis of protein interaction network dynamics in yeast. *Mol. Syst. Biol.* *13*, 934.
- 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* (80-. ). 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.
- Domingo, J., Diss, G., and Lehner, B. (2018). Pairwise and higher-order genetic interactions during the evolution of a tRNA. *Nature* *558*, 117–121.
- 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.
- 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* (80-. ). 291.
- 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.

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.

Kolaczowska, A., Kolaczowski, 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.

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.

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.

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.

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*.

Sarkisyan, K.S., Bolotin, D.A., Meer, M. V., Usmanova, D.R., Mishin, A.S., Sharonov, G. V., Ivankov, D.N., Bozhanova, N.G., Baranov, M.S., Soylemez, O., et al. (2016). Local fitness landscape of the green fluorescent protein. *Nature* 533, 397–401.

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.

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.
- Wang, M., Herrmann, C.J., Simonovic, M., Szklarczyk, D., and von Mering, C. (2015). Version 4.0 of PaxDb: Protein abundance data, integrated across model organisms, tissues, and cell-lines. *Proteomics* *15*, 3163–3168.
- 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., Pregernig, 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.
- 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

A population is engineered by mating the barcoded wild-type pool with a multi-mutant strain (here, the ABC-16 strain). Barcoded haploid progeny inherit a random combination of knockout (black) and wild-type (white) alleles at multiple loci (16 in this study), and are arrayed from single colonies into 384-well plates. The genotype of each strain in this collection was associated with a DNA barcode *en masse* using a tag-based PCR indexing strategy. Strains with a successfully determined barcode and genotype are pooled together by mating type, and grown under each of 16 drugs, as well as a solvent control (DMSO). High throughput sequencing of strain-specific DNA barcodes at  $t = 0, 5, 10, 15$ , and 20 generations of pool growth reconstructs the resistance of each strain to each drug. See also Figures S1 and S2.

### Figure 2. Illustrating a Reproducible 5-gene Resistance Landscape

(A) Comparison of grouped resistance profiles in camptothecin and ketoconazole in MATa and MATα pools. Individual strains were grouped by their genotype at *pdr5Δ*, *snq2Δ*, *ybt1Δ*, *ycf1Δ*, and *yor1Δ*, as indicated by the legend. See also Figures S3 and S4.

(B) Distribution of correlations of grouped resistance profiles between biological replicates (MATa and MATα pools) for all drugs tested.

(C) A radial combinatorial signature of benomyl resistance. The graph center represents strains that are wild-type at all 5 loci, while radial segments extending outwards represent strains carrying successively more knockout alleles. Each section is colored by the mean resistance of the corresponding strain group relative to the 5-gene wild-type. Extensions to 1, 2, and 5 total knockouts are illustrated. Color scale extends by  $\pm 1$  standard deviation of all drug resistance values observed in a pool. See also Figure S5.

(D) 5-knockout radial signatures are shown for 8 drugs.

### Figure 3. 5-gene Resistance Landscapes Illustrate Complex Drug Resistance Phenomena

A resistance landscape of 5-gene groups is shown for six drugs. Groups differing 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. See also Figure S6.

### Figure 4. Environment-Dependent Complex Genetic Interactions amongst 16 ABC Transporters

(A) All single-gene knockout effects and  $X$ -way genetic interactions ( $\epsilon$  terms) mediating resistance to each compound that passed significance criteria (Bonferroni adjusted  $p < 0.05$ ). Magnitude of genetic effects were determined by generalized linear model, then rescaled for each drug.

(B) Illustration of modeled genetic effects in cisplatin, mitoxantrone, and bisantrene. For three selected 5-way genetic interactions, top panels show the distribution of drug resistance phenotypes for grouped strains sharing the 5-gene knockout genotype, and for strains grouped by other relevant genotypes. (In each panel, genotypes are defined only by the five genes of interest, and may vary at other loci.) Differences in median resistance (black lines) between the indicated 4- and 5-gene groups are evaluated with a Mann-Whitney  $U$  test. The bottom row shows genetic effects extracted from Figure 4A.



### Figure 5. An Intuitive Neural Network Model of Complex Genotype-Phenotype Relationships

(A) Schematic of the neural network trained to infer transporter-drug and transporter-transporter influence relationships from complex genotype-to-phenotype relationships. The 16-transporter genotype ( $G$ ) is given as binary input (1 = presence, 0 = absence for each transporter). The activity of each transporter ( $A$ ) is computed using the set of learned transporter-transporter influence weights ( $I$ ).  $A$  is set to zero for transporters that are absent in  $G$ . Resistance to each of the 16 tested compounds ( $R$ ) is then computed by the learned transporter-drug efflux weights ( $E$ ), each multiplied by the  $A$  values of the corresponding transporters.

(B)  $I$  and  $E$  weights learned by the neural network model after training. See also Figures S7A-B and S8.

(C) Agreement between measured and neural-network-modeled drug resistance. See also Figures S7C-D and S8.

(D) Extending the model of valinomycin resistance improves agreement with measurement. For simplicity, data is shown only for the five ‘frequently-associated’ transporters. The trained neural network weights (top) are shown for the original model (top-left) and one with an extra node in the activity layer to model potential influence on a hypothesized resistance factor (top right). See also Figure S8.

### Figure 6. Deciphering a Complex Fluconazole Resistance Trait

(A) Illustrating the complex genetics of fluconazole resistance. Top panel - A landscape of normalized fluconazole resistance for five ABC transporter knockouts (as in Figure 3). Bottom panel - Significant single-knockout effects and  $x$ -gene genetic interactions mediating fluconazole resistance (as in Figure 4).

(B) Extending the model of fluconazole resistance, by adding an extra node between the  $G$  and  $A$  layer to model non-additive influence of four transporters on Pdr5 activity, improves agreement with measurement. Limitation to frequently-associated five transporters and the method for calculating correlation values are as described in Figure 5D.

(C) Close agreement between measures of fluconazole resistance derived by *en masse* analysis of pooled strains with individually-measured resistance (IC50) values in independently-constructed strains.

(D) The synergistic effect of deleting four ABC transporters on Pdr5 activity is primarily explained by an indirect influence on *PDR5* transcript levels. Gray bars represent *PDR5* expression in the specified genotype, relative to that of wild-type, with error bars indicating standard error ( $n=3$ ). Significance was assessed by *t*-test. Colored bars show model-inferred Pdr5 activity (see Figure 5B, top-right panel) for each genotype, relative to that of wild-type, considering all (red) or only the indirect (orange) influences.

(E) Interaction evidence supporting the plausibility of direct repression of Pdr5 by Snq2 and Yor1. This study confirmed all previously known-interactions shown (Snider et al., 2013; Tarassov et al., 2008), and revealed a novel Pdr5-Yor1 protein interaction (Figure S11, S12). Direct influence of Yor1 on Pdr5 activity was predicted by both original and extended neural network models for fluconazole (Figure 5B). Influences from the neural network model (Figure 4B) are shown here with red edges. Whole-organism protein levels (Wang et al., 2015) are indicated by node size.

### **Figure S1 Creating a Pool of Uniquely Barcoded Strains ('Barcode Pool'), Related to Figure 1**

**A** Engineering 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 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, Related to Figure 1**

**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 $a$  pools. The observed number of strains with a given number of knockouts are indicated in gray. 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 $a$  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 gray. Three pairs of unlinked genes— *YORI-YCF1*, *YORI-BPT1*, and *SNQ2-PDR5* – exhibited weak but significant negative correlation in the appearance of KO genotypes ( $-0.04 \geq r \geq -0.08$ ). 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.

### **Figure S3 Marginal Single-Gene Knockout Effects in the Pool, Related to Figure 2**

**A** A linear model was used to formally identify significant gene knockout effects mediating resistance (growth in drug relative to growth in drug) for each of the tested drugs. Linear model terms which were significant (Bonferroni adjusted  $p < 0.05$ ) in both MAT $a$  and MAT $\alpha$  pools for their given drug are colored according the legend on the left. Other terms are colored in gray.

**B** As in A, showing knockout effects that mediate growth

### **Figure S4 Reproducibility of Grouped Genotype Resistance, Related to Figure 2**

As in Figure 2A, showing all 16 drugs

### **Figure S5 Radial Combinatorial Signatures in Additional Drugs, Related to Figure 2**

As in Figure 2C, showing 5 additional drugs

### **Figure S6 Resistance Landscapes for all Drugs, Related to Figure 2**

As in Figure 2D, showing all 16 drugs

### Figure S7 Neural Network Construction and Evaluation, Related to Figure 4

**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 S7A, 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 modeled by the neural network. Results are shown when the network is trained on either the MAT $\alpha$  or MAT $a$  population, and then tested on either the MAT $a$  or MAT $\alpha$  population.

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

### Figure S8 Neural Network Performance for Single Drugs, Related to Figure 4

As in Figure 4D, showing neural network performance for each drug (considering 5-gene groups)

### Figure S9 Neural Network Training with Single-Drug Data, Related to Figures 4 and 5

**A** As in Figure 4D and Figure S8. Performance is shown for the original network trained only on valinomycin data. No substantial improvement over the original neural network is evident.

**B** As in A, showing the original neural network trained only on fluconazole data. No substantial predictive improvement over the original neural network is evident for *PDR5*<sup>+</sup> groups.

**C** As in Figure S7B, showing the mean-squared error of the ‘indirect influence’ fluconazole neural network trained in Figure 5C (right panel) as a function of the regularization rate  $\lambda$ . 13 intervals are plotted from  $10^{-6}$  to  $10^0$ .

### Figure S10 Comparing Single-Strain Fluconazole Resistance to 5-gene Groups, Related to Figure 5

Fluconazole resistance of individual strains containing 32 knockout combinations at *pdr5* $\Delta$ , *snq2* $\Delta$ , *ybt1* $\Delta$ , *ycf1* $\Delta$ , and *yor1* $\Delta$  are compared to the normalized resistance of the grouped pool data (Figure 5A). 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. The pool was grown at 23.4  $\mu$ M.

### Figure S11 Measuring Protein-Protein interactions between Pdr5, Snq2, and Yor1 using mDHFR PCA, Related to Figure 5

*PDR5*, *YOR1*, and *SNQ2* MAT $a$  (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 S12 Measuring Protein-Protein Interactions of Pdr5 with Snq2 and Yor1 using MYTH, Related to Figure 5**

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. 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 +25μM fluconazole + 2% DMSO, SD –WAH +50μM fluconazole + 2% DMSO, and SD –WAH + 2% DMSO. SD –WAH conditions select for reconstitution of ubiquitin.