Large-scale interpretation of chemical-genetic interaction profiles using a genetic interaction network

Scott W. Simpkins[1], Justin Nelson[1], Raamesh Deshpande[2], Sheena C. Li[3], Jeff S. Piotrowski[3,4], Erin H. Wilson[2], Abraham A. Gebre[5], Reika Okamoto[3], Yoshikazu Ohya[5], Hiroyuki Osada[3], Minoru Yoshida[3], Charles Boone[3,6†], Chad L. Myers[1,2†]

1. University of Minnesota-Twin Cities, Bioinformatics and Computational Biology Graduate Program, Minneapolis, Minnesota, USA
2. University of Minnesota-Twin Cities, Department of Computer Science and Engineering, Minneapolis, Minnesota, USA
3. RIKEN Center for Sustainable Resource Science, Wako, Saitama, Japan
4. Yumanity Therapeutics, Cambridge, MA, USA
5. University of Tokyo, Department of Integrated Biosciences, Graduate School of Frontier Sciences, Kashiwa, Chiba, Japan
6. University of Toronto, Donnelly Centre, Toronto, Ontario, Canada

†Correspondence to: chadm@umn.edu, charlie.boone@utoronto.ca

ABSTRACT

Genetic interactions provide a key for interpreting the functional information contained in chemical-genetic interaction profiles. However, they have remained underutilized in this capacity across recent chemical-genetic interaction screening efforts and their ability to interpret chemical-genetic interaction profiles on a large scale has not been tested. We developed a method, which we refer to as CG-TARGET (Chemical Genetic Translation via A Reference Genetic nETwork), that integrates the data from large-scale chemical-genetic interaction screens with genetic interaction data to predict the biological processes perturbed by compounds. CG-TARGET compared favorably to a standard enrichment approach across a variety of benchmarks, achieving similar performance on measures of accuracy and substantial improvement in the ability to control the false discovery rate of its predictions. We found that one-third to one-half of gene mutants in the data contribute to the highest-confidence biological process predictions and that these contributions overwhelmingly come from negative chemical-genetic interactions. This method was used to prioritize over 1500 out of over 13,000 compounds for further study in a recently-completed chemical-genetic interaction screen in *Saccharomyces cerevisiae*, enabling the rapid functional annotation of unknown compounds to biological processes through targeted biological validations. We present here a detailed characterization of

36 the method and further biological validations to demonstrate the utility of genetic interactions in
37 the interpretation of chemical-genetic interaction profiles and the effectiveness of our
38 implementation of this concept.

39 **INTRODUCTION**

40 The ability to discover chemical compounds with desirable and/or interesting biological
41 activity is essential to understanding the way compounds and biological systems interact. One
42 way to characterize the biological activity of a compound in an unbiased manner is to profile its
43 activity across a genome-wide array of genetic mutants, also known as chemical-genetic
44 interaction screening [1]. In the resulting chemical-genetic interaction profiles, the identities of
45 the gene mutations that confer sensitivity or resistance to a compound provide functional
46 information regarding the actions performed by that compound inside the cell.

47 Genetic interaction profiles provide analogous information regarding gene function, and as
48 such can be used to interpret the functional information contained in chemical-genetic interaction
49 profiles [2]. Specifically, shared interactions between chemical-genetic and genetic interaction
50 profiles may implicate a particular gene or group of genes (e.g. a biological process or protein
51 complex) as the target of a compound's actions in the cell (Figure 1). This scheme for
52 interpreting chemical-genetic interaction profiles does not depend on the existence of chemical-
53 genetic interaction profiles for well-characterized compounds, and thus enables the discovery of
54 compounds with novel modes of action.

55 Recent advances in whole-genome chemical-genetic interaction screening technology have
56 opened the possibility of using chemical genomics as a high-throughput screening approach [3–
57 5]. This would, for example, enable functional profiling of compound bioactivity at earlier points
58 in the drug discovery process, providing an additional means of prioritizing promising
59 compounds (and discarding compounds with less obvious, yet undesirable activities) before
60 investing into them large amounts of resources. However, genetic interactions remained
61 essentially unused in these screens for the systematic interpretation of chemical-genetic
62 interaction profiles. As such, a systematic interpretation of chemical-genetic interaction profiles
63 using genetic interaction profiles has not been demonstrated on a large scale (1,000s to 10,000s
64 of compounds). A study of this type would provide insights into the compatibility between
65 chemical-genetic and genetic interaction profiles and the ability of a genetic interaction-based

method to prioritize compounds with high-confidence predictions while controlling for issues typically associated with high-throughput chemical screening.

In this manuscript, we present the use of genetic interaction profiles to systematically interpret chemical-genetic interaction profiles on a large-scale. Specifically, we developed a method, called CG-TARGET (Chemical Genetic Translation via A Reference Genetic nETwork), that incorporates genetic interaction data and different sources of experimental variation to predict the biological processes perturbed by compounds. We applied this method to a high-throughput chemical-genetic interaction screen of more than 13,000 compounds in *S. cerevisiae* [6], using profiles from the corresponding yeast genetic interaction network [7,8] to interpret the chemical-genetic interaction profiles. CG-TARGET recapitulated known information for well-characterized compounds and showed a marked improvement in the ability to control the false discovery rate – and as a result, prioritize interesting compounds – compared to a baseline approach. We also confirmed, through a global analysis, the compatibility between chemical-genetic and genetic interaction profiles for the purpose of predicting perturbed biological processes. CG-TARGET is available, free for academic use and licensed for commercial use, at github.com/csbio/CG-TARGET.

RESULTS

Predicting perturbed biological processes from chemical-genetic interaction profiles

When developing a method that uses genetic interaction profiles to interpret chemical-genetic interaction profiles obtained at scale, it was important to consider scenarios in which experimental artifacts or common signatures in the chemical-genetic interaction profiles could strongly influence the similarities between chemical-genetic and genetic interaction profiles, leading to biased and inaccurate process predictions. For example, common similarity measures are blind to the variance of individual gene mutants across all chemical-genetic interaction profiles. As a result, gene mutants with highly variable interaction scores in chemical-genetic interaction experiments possess the potential to drive the prediction of processes in a nonspecific manner. While this can be addressed with negative experimental controls, it is not inconceivable that certain gene mutants would respond nonspecifically only in the presence of compound, requiring a correction derived from the dataset itself. Additionally, spurious correlations introduced by normalized similarity measures on weak chemical-genetic interaction profiles

96 (Pearson correlation coefficient, cosine correlation) can be further amplified by the redundancy
97 in the genetic interaction network, leading to false discoveries.
98 We developed CG-TARGET (Chemical Genetic Translation via A Reference Genetic
99 nETwork) to address these concerns surrounding the prediction of perturbed biological processes
100 at scale (Figure 1). Predicting a compound's perturbed biological processes using CG-TARGET
101 requires three input datasets (chemical-genetic interaction profiles, genetic interaction profiles,
102 and a mapping from the query genes in the genetic interaction profiles to biological processes)
103 and involves four distinct steps. First, a set of resampled chemical-genetic interaction profiles is
104 generated, each of which consists of one randomly sampled interaction score for each gene
105 mutant across all compound treatment profiles in the chemical-genetic interaction dataset.
106 Second, scores reflecting both the strength of each compound's chemical-genetic interaction
107 profile and its similarity to the profile of each gene mutant are obtained by computing a dot
108 product between all chemical-genetic interaction profiles (comprising compound treatment,
109 experimental control, and resampled profiles) and all L_2-normalized query genetic interaction
110 profiles. These "gene-level" prediction scores, which possess per-compound ranks equivalent to
111 those obtained using cosine similarity but prioritize compounds with stronger profiles, are then
112 aggregated into process predictions; the z-score and empirical p-value for each compound-
113 process prediction are obtained by mapping the gene-level prediction scores to the genes in the
114 process of interest and comparing these scores to those from shuffled gene-level prediction
115 scores and to distributions of the scores derived from experimental control and resampled
116 profiles. Finally, the false discovery rates for these predictions are estimated by calculating the
117 frequency at which experimental control and resampled profiles predict processes across a range
118 of significance thresholds, compared to the compound treatment profiles.

119 Application to and evaluation on large-scale chemical-genetic interaction screening 120 data

121 We applied CG-TARGET to the problem of predicting biological target processes from two
122 recent large-scale chemical-genetic interaction screens in *S. cerevisiae*. The first screen was
123 performed on 9850 compounds from the RIKEN Natural Product Depository [9] (the "RIKEN"
124 screen) and the second was performed on 4116 compounds from the NCI Open Chemical
125 Repository's plated compound libraries, the NIH Clinical Collection, and GlaxoSmithKline's
126 Published Kinase Inhibitor Set (the "NCI/NIH/GSK" screen) [10]. The number of chemical-

127 genetic interaction profiles obtained from each screen was 8418 and 3565, respectively. In both
128 screens, the chemical-genetic interaction profiles were determined across a diagnostic set of
129 approximately 300 haploid gene deletion mutants. Genetic interaction profiles were obtained
130 from a compendium of genetic interaction profiles in *S. cerevisiae* [7], with the query genes
131 mapped to propagated Gene Ontology biological process terms [11,12] to define the process
132 targets.

133 To provide a baseline approach for benchmarking the performance of CG-TARGET on these
134 large screens, we implemented a standard enrichment approach that tests for the enrichment of
135 processes in each compound's top-k gene-level prediction scores. Using experimental control
136 and resampled profiles, we assessed the ability of the enrichment-based prediction method to
137 control the false discovery rate at various values of k. The best-performing value of k was then
138 used to further benchmark the accuracy of the process predictions made by CG-TARGET.

139 CG-TARGET was successful in controlling the false discovery rate across both chemical-
140 genetic interaction screens, identifying 848 out of 8418 compounds from the RIKEN screen
141 (10%) and 705 compounds from the NCI/NIH/GSK screen (20%) with at least one prediction
142 that achieved false discovery rates of 25 and 27%, respectively (Table 1, Figure 2A-D). In
143 contrast, the best-performing top-k enrichment approach (k=100) identified only 57 compounds
144 with an equivalent false discovery rate when applied to the profiles from the RIKEN screen
145 (Figure 2E-F). In all cases, the false discovery rates derived from resampled profiles were more
146 conservative than those derived from experimental controls, suggesting that some sources of
147 variance in each gene mutant's interaction scores arise only upon treatment with compound and
148 therefore cannot be corrected using only negative experimental controls. The compounds from
149 the RIKEN and NCI/NIH/GSK screens with at least one prediction at or below the respective
150 false discovery rate cutoff set for each screen will be referred to as the RIKEN and
151 NCI/NIH/GSK "high-confidence sets," respectively.

152 We benchmarked the accuracy of CG-TARGET's process predictions against a set of 35
153 gold-standard compound-process annotations curated from the literature and observed favorable
154 performance to that on predictions generated using the top-100 enrichment approach. More
155 specifically, we computed the rank of each compound's gold-standard process within its set of
156 process predictions, and compared this rank to those obtained from randomly shuffled
157 predictions. CG-TARGET performed slightly worse overall when comparing the ranks of the

158 gold-standard processes (12 versus 14 in the top ten, Figure 3A) and the number of compounds
159 with significant ranks (22 versus 23, Table 2). However, only 2 out of 23 significantly-ranked
160 gold-standard predictions made using the top-100 enrichment approach achieved a false
161 discovery rate of 25% or less, while 16 out of 22 predictions from CG-TARGET achieved this
162 same false discovery rate. As such, CG-TARGET discovered 8-fold more compounds with
163 significantly-ranked gold-standard process annotations within the RIKEN high-confidence set.
164 This result provides further evidence supporting the utility of genetic interaction profiles in the
165 interpretation of chemical-genetic interaction profiles, while simultaneously demonstrating that
166 the predictive power of genetic interaction profiles improved when combined with additional
167 experimental data.

168 In a more global benchmarking effort, we also observed that CG-TARGET improved the
169 prioritization of process predictions when applied to simulated chemical-genetic interaction
170 profiles. The set of simulated profiles was designed contain three compounds that target each
171 query gene in the genetic interaction dataset; each simulated profile thus inherited the process
172 annotations of its parent genetic interaction profile, providing a gold standard with which to
173 evaluate their predictions. Evaluation performed on the top process prediction for each simulated
174 compound revealed that CG-TARGET captured 15% more gold-standard annotations than did
175 top-100 enrichment. While both methods only captured a gold-standard annotation in the top
176 process prediction for approximately 30% of the simulated compounds, this still represents a 56-
177 fold enrichment over the background expectation of 0.00533. In addition, CG-TARGET more
178 successfully prioritized its true positive annotations, as shown by the consistent improvement in
179 precision over top-100 enrichment, especially at low recall values (Figure 3B). The diversity of
180 the true positive process predictions, when mapped to a set of 17 broad functional
181 neighborhoods, was also improved using CG-TARGET (Shannon index = 2.58 for CG-TARGET
182 versus 2.33 for top-100 enrichment predictions), likely due to a substantial reduction in the
183 number of compounds mapped to the "vesicle traffic" neighborhood (150 for CG-TARGET vs.
184 367 for top-100 enrichment).

185 In addition to benchmarking, we investigated the potential to expand the use cases of CG-
186 TARGET to the prediction of perturbed protein complexes. For protein complex prediction on
187 the RIKEN screen data, 714 compounds were identified with at least one prediction achieving a
188 false discovery rate of 25% or less. 603 of these 714 compounds were also identified in the high

189 confidence RIKEN process predictions, suggesting the potential to map prioritized process
190 predictions to more specific, defined predictions of compound targets in the cell. For example,
191 the top protein complex prediction for NPD1409 was "Kornberg's mediator (SRB) complex,"
192 which plays an important role in the initiation of transcription as well as chromatin looping
193 [13,14]. This prediction agrees with the compound's top process predictions from the RIKEN
194 screen to perturb "chromosome organization," "DNA metabolic process," and/or "RNA
195 polymerase II transcriptional preinitiation complex assembly," and points to a more direct target
196 for testing in experimental validation efforts.

197 **Insights into the compatibility of chemical-genetic and genetic interaction profiles**

198 Given previous demonstrations [2,7] and the evaluations presented here, it should be clear
199 that the use of genetic interaction profiles to interpret chemical-genetic interaction profiles is
200 both appropriate and useful. However, a further investigation of the inner workings of this
201 approach was warranted to more comprehensively understand the extent to which these two
202 types of profiles can be combined and how this affects the prediction of processes. Here we
203 present visualizations that reveal insights into 1) the interpretation of a chemical-genetic
204 interaction profile to predict a biological process and 2) the different ways in which a process can
205 be predicted using chemical-genetic and genetic interaction profiles. Finally, we quantify, across
206 the RIKEN high confidence set of compounds, the relationship between chemical-genetic
207 interactions and their importance to the prediction of perturbed biological processes.

208 To better understand the interpretation of chemical-genetic interaction profiles, we
209 quantified, for every compound, the contribution of each gene mutant to the prediction of
210 individual biological processes. For a single compound and predicted process, these "importance
211 scores" were obtained by 1) computing the Hadamard product (elementwise multiplication)
212 between the compound's chemical-genetic interaction profile and each L_2-normalized query
213 genetic interaction profile mapped to the predicted process and 2) for each gene mutant,
214 computing the mean of this product across the genetic interaction profiles. These scores can be
215 positive, indicating agreement in the sign of chemical-genetic and genetic interactions for a
216 particular gene mutant, or they can be negative, indicating that the interactions do not agree for
217 that gene mutant. As such, the importance scores summarize the concordance between chemical-
218 genetic and genetic interaction profiles, conditioned on an individual compound and a perturbed
219 process of interest.

220 The prediction of NPD4142, a compound from the RIKEN Natural Product Depository, to
221 the "mRNA transport" process can be used to illustrate how the overlap between chemical-
222 genetic and genetic interactions leads to process predictions (Figure 4A). A qualitative
223 examination revealed that, indeed, NPD4142 possesses a pattern of chemical-genetic interactions
224 similar to the genetic interactions for the query genes annotated to mRNA transport. However, a
225 quantitative assessment achieved more nuance in this comparison. While the *POM152* deletion
226 mutant possessed the strongest negative interaction with NPD4142, the importance scores
227 revealed that it was not the most important gene mutant for making this prediction; instead, the
228 deletion mutant for *NUP133*, which possessed a weaker chemical-genetic interaction score but
229 more genetic interactions with the mRNA transport-annotated query genes, emerged as the most
230 important for predicting mRNA transport.
231 We also compared the concordance of chemical-genetic and genetic interaction profiles
232 across multiple compounds predicted to the same process, revealing that individual processes
233 were predicted by both homogenous and heterogeneous sets of chemical-genetic interaction
234 profiles. For example, all predictions made to "proteasome assembly" depended almost entirely
235 on a strong negative chemical-genetic interaction with *RPN4*, which was captured most clearly
236 by the relevant importance scores (Figure 4B). This uniformity in the prediction of a process is
237 contrasted by the diversity of profiles captured within "fungal-type cell wall organization"
238 predictions (Figure 4C). Here, filtering on the importance scores showed that chemical-genetic
239 interactions with four genes – *GAS1*, *SMI1*, *ABP1*, and *DFG5* – were primarily responsible for
240 predictions to this term, but with low agreement regarding their relative importance for each
241 compound's prediction. In the lattermost case, the concordance of chemical-genetic and genetic
242 interactions was not particularly obvious, yet was sufficient to enable the prediction of a
243 perturbed process.
244 More globally, we found broad contribution across a large fraction of observed chemical-
245 genetic interactions – primarily negative interactions – to the prediction of perturbed processes
246 (Figure 4D). By comparing the chemical-genetic interactions for each compound to their
247 corresponding importance scores for that compound's top process prediction, we observed that
248 nearly one-third (5398 / 16464) of chemical-genetic interactions contributed to top process
249 predictions, the fraction of which increased to nearly one-half (5087 / 10281) when considering
250 only negative interactions. While positive chemical-genetic interactions were much less

251 frequently observed overall (and only 7% positively contributed to a top process prediction), they
252 were 3.5 times more likely to contribute negatively to a process prediction than were negative
253 interactions (1.47% vs. 0.42%). Overall, 199 gene mutants (72%) contributed to at least one top
254 process prediction, while half (143) contributed to at least five predictions (importance score >
255 0.1). Based on a more stringent threshold on importance scores (> 1.0), 65 gene mutants (23% of
256 mutants) were observed to be very strong contributors to certain process predictions, with 17 of
257 these contributing strongly to the top process prediction of at least five compounds. While some
258 gene mutants certainly contributed disproportionately to a subset of predictions, the broad
259 majority of predictions required contributions from a much larger fraction of gene mutants.

260 Experimental validation of compound-process predictions

261 *Phenotypic analysis of cell cycle progression*

262 Several compounds from the RIKEN dataset were predicted to perturb the process related ot
263 the cell cycle. We chose to test 19 of these compounds, 14 of which had high-confidence to the
264 "spindle assembly checkpoint" process active in the M phase of the cell cycle, to determine if
265 our predictions captured the biological activity of these compounds. Indeed, we observed that 7
266 of the 19 compounds induced a cell cycle phenotype, with 6 of the 14 compounds annotated to
267 spindle assembly checkpoint inducing abnormally large buds on cells, changes in the budding
268 index of cells, and increases in cellular DNA content, indicative of arrest in G2/M phase (Figure
269 5A-C). These phenotypes were not observed when performing the same experiments on a set of
270 10 active compounds from the high-confidence set whose predictions were to processes
271 unrelated to the cell cycle. This difference in the rate of validation between predicted active
272 compounds and negative controls (6 / 14 vs. 0 / 10) was statistically significant ($p < 0.03$,
273 proportion test). Two of the selected compounds were predicted to perturb "cell cycle phase,"
274 one of which induced phenotypes consistent with G1 arrest (Fig 5A-C). This provided one
275 demonstration of CG-TARGET's ability to prioritize compounds that perturb a particular
276 function in the cell.

277 *Inhibition of tubulin polymerization*

278 Compounds that disrupt microtubules are useful for studying cell organization and division,
279 and remain promising candidates as antitumor agents [15–17]. We therefore chose to
280 experimentally validate our predictions in a way that might identify compounds that possess such

281 activities. All compounds with the strongest predictions (FDR = 0%) to "tubulin complex
282 assembly" were selected for biochemical validation in an *in vitro* tubulin polymerization assay
283 (Figure 5D). Similar to the previous validation, a negative control set of compounds was selected
284 to contain active compounds (process predictions with FDR ≤ 25%) whose predictions were not
285 related to microtubules or related processes. We observed that the novel compound NPD2784
286 strongly inhibited tubulin polymerization nearly as well as the drug nocodazole and more
287 strongly than the microtubule probe benomyl. In addition, the entire set of compounds predicted
288 to perturb tubulin complex assembly showed significantly increased inhibition of tubulin
289 polymerization when compared to the negative control compounds ($p < 0.005$, Wilcoxon
290 ranksum test). These confirmatory results showed the translation of our predictions to a specific
291 biochemical validation, even in the context of a different species.

292 **DISCUSSION**

293 The scaling of chemical-genetic interaction screens from tens or hundreds of compounds to
294 tens of thousands of compounds has provided the opportunity, and the necessity, to more
295 comprehensively characterize appropriate methods for interpreting the interaction profiles and
296 prioritizing high-confidence compounds. We developed a method, CG-TARGET, to address this
297 need and used it to predict perturbed biological processes for more than 13,000 interaction
298 profiles from a recent high-throughput chemical-genetic interaction screen [6]. CG-TARGET
299 demonstrated the ability to recapitulate known compound function while controlling the false
300 discovery rate, prioritizing 1522 compounds for further study. Further investigation of the
301 profiles from these high-confidence compounds revealed broad compatibility between chemical-
302 genetic and genetic interaction profiles. In addition to these findings, the predictions made using
303 CG-TARGET were experimentally validated on a large scale for 67 compounds in an orthogonal
304 cell cycle assay and revealed insights into the distribution of functions perturbed by compounds
305 in large compound libraries [6].

306 In high-throughput chemical screens, it is important to prioritize the compounds most likely
307 to demonstrate desired biological activity in further studies. While CG-TARGET and a baseline
308 approach performed similarly on the task of ranking gold-standard compound-process
309 annotations, CG-TARGET was 8 times better at prioritizing these compounds as high-confidence
310 predictions. Surprisingly, CG-TARGET outperformed the same baseline method when
311 predicting and prioritizing perturbed processes for simulated chemical-genetic interaction

312 profiles derived from genetic interaction profiles across the genome, providing evidence for its
313 value in discovering compounds with modes of action not previously characterized in the
314 literature. This is particularly important, as our gold standard set of compound-process
315 annotations, consisting of 35 compounds across 17 biological processes (4 of which are DNA-
316 related), does not enable prediction across the range of biological processes present in the cell.
317 Genetic interactions thus provide the most comprehensive reference for interpreting chemical-
318 genetic interaction profiles in an unbiased, genome-wide manner.

319 While we demonstrated the ability to predict perturbed processes for compounds and
320 prioritize the highest-confidence predictions, many further steps are required to identify lead
321 compounds and ultimately develop molecular probes or even pharmaceutical agents. Perturbing a
322 biological process does not necessarily require perturbing a specific protein target, and as such,
323 further refinements to our methods are needed to prioritize the compounds most likely to perturb
324 a small number of defined targets in the cell. Different modes of chemical-genetic interaction
325 screening can provide support in this endeavor, as the profiling of heterozygous diploid strains
326 provides evidence for the direct, essential cellular target(s) of a compound [1,4]. Regardless of
327 these limitations in predicting defined targets, information about the processes perturbed by an
328 entire library of would be useful in selecting the compounds most amenable to activity
329 optimization and off-target effect minimization in the development of a pharmaceutical agent or
330 molecular probe.

331 Moving forward, this work supports the idea of performing both genetic and chemical-
332 genetic interaction screens in other species for which obtaining functional information on
333 compounds would be useful. For example, genome-wide deletion collections have been
334 developed for *Escherichia coli* [18] and *Schizosaccharymyces pombe* [19] and used to perform
335 chemical-genetic interaction screens [20,21] as well as genetic interaction mapping [22–25].
336 Such efforts are even underway in human cell lines, enabled by genome-wide CRISPR knockout
337 screens [26–29]. Furthermore, future efforts to interpret chemical-genetic interaction profiles in a
338 new species need not wait for the completion of an all-by-all genetic interaction network, as this
339 work highlights the ability of a diagnostic set of gene mutants to capture functional information
340 and predict perturbed biological processes. From the discovery of urgently-needed antibacterial
341 or antifungal agents, to the treatment of orphan diseases or a better understanding of drug and
342 chemical toxicity, the combination of chemical-genetic and genetic interactions in a high-

343 throughput format, with appropriate analysis tools, offers a means to achieve these goals via the
344 discovery of new compounds with previously uncharacterized mechanisms of action.

345 MATERIALS AND METHODS

346 Datasets

347 *Chemical-genetic interaction data*

348 Chemical-genetic interaction profiles were obtained from a recent study [6], in which nearly
349 14,000 compounds were screened for chemical-genetic interactions across approximately 300
350 haploid yeast gene deletion strains. Each profile contains z-scores that reflect the deviation of
351 each strain's observed fitness from its expected fitness in the presence of compound. The
352 chemical-genetic interaction profiles were obtained in two batches and divided as such into: 1)
353 the "RIKEN" dataset, which contains chemical-genetic interaction profiles across 289 deletion
354 strains for 8418 compounds from the RIKEN Natural Product Depository [9] and 5724 negative
355 experimental controls (solvent control, DMSO) ; and 2) the "NCI/NIH/GSK" dataset, which
356 contains chemical-genetic interactions across 282 deletion strains for 3565 compounds from the
357 NCI Open Chemical Repository, the NIH Clinical Collection, and the GSK kinase inhibitor
358 collection [10], as well as 2128 negative experimental control profiles. The solvent control
359 profiles consisted of biological and technical replicate profiles.

360 *Genetic interaction data*

361 The genetic interaction dataset consisted of quantitative fitness observations for the double
362 mutants obtained upon crossing between 1505 high-signal query gene mutants into an array of
363 3827 array gene mutants, obtained from a recently assembled *S. cerevisiae* genetic interaction
364 map [7,8]. These observations are represented as epsilon scores, which quantify the difference
365 between each double mutant's observed and its expected fitness values and are analogous to the
366 chemical-genetic interaction z-scores. The procedure for selecting the 1505 high-signal query
367 genes out of the larger pool of 4382 is described in [6]. Briefly, each query profile was required
368 to possess at least 40 significant genetic interactions, a sum of cosine similarity scores with all
369 other query profiles greater than 2, and a sum of dot products with all other query profiles greater
370 than 2. When comparing with chemical-genetic interactions, the genetic interaction dataset was
371 filtered to contain only array strains present in the chemical-genetic interaction datasets.

372 *GO Biological Processes and protein complexes*

373 A subset of terms from the "biological process" ontology within the Gene Ontology
374 annotations [12] were used as the processes. Query genes from the *S. cerevisiae* genetic
375 interaction dataset were mapped to genes to biological process terms using annotations from the
376 Saccharomyces cerevisiae Genome Database [11]. Both gene ontology and *S. cerevisiae*
377 annotations were downloaded from their respective databases via Bioconductor in R [30]. Terms
378 were propagated using "is_a" relationships, such that all child terms were annotated to their
379 parent terms as well. The final set of processes consisted of the terms with 4 – 200 gene
380 annotations from the set of 1505 high-signal query genes in the genetic interaction dataset.

381 Protein complex annotations were obtained from [8]. Complexes with 3 or more genes
382 annotated to them were used as the input biological processes for CG-TARGET-based protein
383 complex predictions.

384 *Gold standard compound-process annotations*

385 Biological processes were assigned to 35 primarily antifungal compounds with chemical-
386 genetic interaction profiles in the RIKEN dataset, based on known information about their
387 mechanisms of action. Process terms were selected to be specific to the compound's mechanisms
388 of action where applicable.

389 *Mapping biological processes to functional neighborhoods*

390 We expanded an initial standard of 488 Gene Ontology biological process terms annotated to
391 17 functional neighborhoods [31] using a k-nearest-neighbors approach. For each previously
392 unannotated process in our set of processes, we assigned similarity scores for the 3 most similar
393 (Jaccard overlap on gene annotations) processes in the process-neighborhood standard to their
394 respective functional neighborhoods and annotated the new process to the functional
395 neighborhood with the highest sum of similarity scores. In the case of a tie, the process was
396 annotated to both functional neighborhoods.

397 **Predicting the biological processes perturbed by compounds**

398 Our method to predict biological processes perturbed by compounds is described in the
399 recent study from which the chemical-genetic interaction profiles were obtained [6]. We describe
400 here the modifications to this approach to implement the top-k enrichment method for
401 benchmarking.

402 Given a set of gene-level similarity scores for each compound and a set of gene-process
403 annotations, the enrichment of each process in the set of the top-k most similar genes for each
404 compound was computed. This is reflected in an enrichment factor (the fraction of the *k* selected
405 genes annotated to a particular process divided by the fraction of total genes annotated to that
406 process) and a p-value obtained using a hypergeometric test. These enrichment factors and p-
407 values were substituted in place of the z-scores and p-values obtained using CG-TARGET for
408 subsequent analyses.

409 **Computational evaluation of process predictions**

410 *Performance on gold-standard compounds*

411 The predicted perturbed processes for each of the gold standard compounds were sorted, first
412 by their p-value (ascending) and then by their z-score (for CG-TARGET, descending) or
413 enrichment factor (top-100 enrichment, descending), and the rank of each of their gold-standard
414 process annotations was recorded. To assess the significance of each rank, each pair of p-value
415 and z-score was assigned to a new process, the lists re-ordered, and the ranks of each
416 compound's target process re-computed. The empirical p-value for each gold-standard
417 compound-process pair was computed as the number of times the rank from the shuffled
418 processes achieved the same or better rank as the observed rank.

419 *Performance on genetic interaction profiles*

420 We generated a set of simulated chemical-genetic interaction profiles derived from the
421 genetic interaction profiles [6]. Each simulated chemical-genetic interaction profile was a query
422 genetic interaction profile augmented with noise sampled from a Gaussian distribution with a
423 mean of 0 and a variance for each array gene twice that of the same array gene in the genetic
424 interaction dataset. Three simulated profiles were generated based on each query gene, resulting
425 in 4515 total profiles. Because each simulated chemical-genetic interaction profile was derived
426 from a query genetic interaction profile, it inherited the gold standard process annotations from
427 its parent genetic interaction profile in subsequent benchmarking efforts.

428 We then used CG-TARGET and the top-100 enrichment method to predict perturbed
429 processes for this set of 4515 simulated chemicals x 289 deletion mutants. For each simulated
430 chemical, its top process prediction was compared to the set of inherited gold-standard process
431 annotations, counting as a true positive if the top prediction matched an existing annotation and a

432 false positive if it did not. Precision-recall curves were then generated by sorting the list of each
433 simulated chemical's top process predictions (p-value ascending, z-score or enrichment factor
434 descending) and computing the precision (true positives / (true positives + false positives)) and
435 recall (true positives) at each point in this list.
436 The set of true positive process predictions from both methods was mapped to functional
437 neighborhoods via the expanded process-neighborhood mapping ("Mapping biological processes
438 to functional neighborhoods"). The proportion of processes mapped to each neighborhood was
439 used to compute diversity via Shannon Index.

440 *Analysis of process prediction drivers in chemical-genetic interaction data*

441 Given a compound and a predicted process, a profile of "importance scores" describes the
442 contribution of each gene mutant that compound's process prediction. To obtain this score, a
443 Hadamard product (elementwise multiplication) is first computed between the compound's
444 genetic interaction profile and each L_2-normalized genetic interaction profile for which the dot
445 product between them is 2 or greater and the genetic interaction profile is annotated to the
446 process of interest. The final importance profile consists of the mean of each gene's elementwise
447 products across all selected genetic interaction profiles.

448 **Experimental validation of process predictions**

449 *Phenotypic assessment of cell cycle*

450 To examine the effect of compounds on arresting cells in G2/M phase, we looked for
451 differences in budding index and cell DNA content between compounds predicted to perturb the
452 cell cycle versus negative control compounds. Nineteen compounds with high-confidence
453 predictions to cell cycle-related biological processes, 14 of them to "spindle assembly
454 checkpoint" were selected for validation, while ten compounds with predictions of false
455 discovery rate ≤ 25% to processes not mapped to the functional neighborhoods of "Cell Cycle
456 Signaling and Progression" and "Mitosis and Chromosome Segregation" (see "Mapping
457 biological processes to functional neighborhoods") were selected as bioactive negative controls.
458 Two compounds predicted to perturb "cell cycle phase" were also tested in these experiments.
459 All compounds were tested at a concentration of 10 μg/mL, which was also the concentration
460 used to obtain their chemical-genetic interaction profiles [6].

461 To quantify budding index, logarithmically-growing *pdr1*Δ*pdr3*Δ*snq2*Δ cells were
462 transferred to fresh galactose-containing medium (YPGal) containing compounds and incubated
463 at 25^0C for 4 h. The budding status of at least 200 cells was visually determined under the
464 microscope. The percentage of the budded cells in no compound or compound-treated cells was
465 counted.
466 For flow cytometry analysis, the *pdr1*Δ*pdr3*Δ*snq2*Δ cells were grown in YPGal media in the
467 presence or absence of a compound until log phase and then fixed in 70% ethanol for 1 h at
468 25^0C. Cells were collected by centrifugation, treated with RNase A solution (0.25 mg/mL in
469 50 mM Tris pH7.5) for 1.5 h, and 20 µl of 20 mg/ml proteinase K further incubated at 50^0C for
470 1h. Samples were then stained with propidium iodide, briefly sonicated, and measured using
471 FACSCalibur ver 2.0 (Becton Dickinson, CA, USA).
472 The proportions of predicted active compounds and negative controls with positive
473 phenotypic results were compared using the prop.test function in R to assess significance.

474 *Tubulin polymerization assay and analysis*

475 We performed in vitro tubulin polymerization assays using the Cytoskeleton fluorescent-
476 based porcine tubulin polymerization assay (BK011P) following manufacturer specifications.
477 We tested the compounds at a concentration of 10 µg/ml, which was identical to the
478 concentration at which they were screened to generate their chemical-genetic interaction profiles.
479 Nine out of the 10 compounds predicted to perturb "tubulin complex assembly" with an
480 estimated false discovery rate of 0% were selected for testing in the tubulin polymerization
481 assay. Twelve compounds with predictions of false discovery rate ≤ 25% to processes not
482 mapped to the "Mitosis and Chromosome Segregation" functional neighborhood were selected
483 as bioactive negative controls.
484 We used the V_{max} of tubulin polymerization between the tubulin-predicted compounds and
485 the negative controls to determine if the tubulin-predicted compounds inhibited polymerization
486 to a greater degree than the controls. V_{max} for each compound's fluorescence time-course was
487 calculated as the maximum mean change in fluorescence across 8 consecutive time points. V_{max}
488 values were then compared using a Wilcoxon rank-sum test to determine the significance of
489 polymerization inhibition increases caused by the compounds predicted to perturb tubulin
490 complex assembly.
491

492 **ACKNOWLEDGEMENTS**

493 This work was partially supported by the National Institutes of Health (R01HG005084,
494 R01GM104975) and the National Science Foundation (DBI 0953881). SWS is supported by an
495 NSF Graduate Research Fellowship (00039202), an NIH Biotechnology training grant
496 (T32GM008347), and a BICB one-year fellowship. CLM and CB are fellows in the Canadian
497 Institute for Advanced Research (CIFAR) Genetic Networks Program. Computing resources and
498 data storage services were partially provided by the Minnesota Supercomputing Institute and the
499 UMN Office of Information Technology, respectively.

500 **REFERENCES**

501 1. Giaever G, Shoemaker DD, Jones TW, Liang H, Winzeler EA, Astromoff A, et al. Genomic
502 profiling of drug sensitivities via induced haploinsufficiency. Nat Genet. 1999
503 Mar;21(3):278–83.

504 2. Parsons AB, Brost RL, Ding H, Li Z, Zhang C, Sheikh B, et al. Integration of chemical-
505 genetic and genetic interaction data links bioactive compounds to cellular target pathways.
506 Nat Biotechnol. 2004 Jan;22(1):62–9.

507 3. Hoepfner D, Helliwell SB, Sadlish H, Schuierer S, Filipuzzi I, Brachat S, et al. High-
508 resolution chemical dissection of a model eukaryote reveals targets, pathways and gene
509 functions. Microbiol Res. 2014 Mar;169(2-3):107–20.

510 4. Lee AY, St Onge RP, Proctor MJ, Wallace IM, Nile AH, Spagnuolo PA, et al. Mapping the
511 cellular response to small molecules using chemogenomic fitness signatures. Science. 2014
512 Apr 11;344(6180):208–11.

513 5. Smith AM, Heisler LE, Mellor J, Kaper F, Thompson MJ, Chee M, et al. Quantitative
514 phenotyping via deep barcode sequencing. Genome Res. 2009 Oct;19(10):1836–42.

515 6. Piotrowski JS, Li SC, Deshpande R, Simpkins SW, Nelson J, Yashiroda Y, et al. Functional
516 Annotation of Chemical Libraries across Diverse Biological Processes. bioRxiv [Internet].
517 2017 Feb 28; Available from: http://biorxiv.org/content/early/2017/02/28/112557.abstract

518 7. Costanzo M, Baryshnikova A, Bellay J, Kim Y, Spear ED, Sevier CS, et al. The genetic
519 landscape of a cell. Science. 2010 Jan 22;327(5964):425–31.

520 8. Costanzo M, VanderSluis B, Koch EN, Baryshnikova A, Pons C, Tan G, et al. A global
521 genetic interaction network maps a wiring diagram of cellular function. Science. 2016 Sep
522 23;353(6306).

523 9. Kato N, Takahashi S, Nogawa T, Saito T, Osada H. Construction of a microbial natural
524 product library for chemical biology studies. Curr Opin Chem Biol. 2012 Apr;16(1-2):101–8.

525 10. Drewry DH, Willson TM, Zuercher WJ. Seeding collaborations to advance kinase science
526 with the GSK Published Kinase Inhibitor Set (PKIS). Curr Top Med Chem. 2014;14(3):340–
527 2.

528 11. Cherry JM, Hong EL, Amundsen C, Balakrishnan R, Binkley G, Chan ET, et al.
529 Saccharomyces Genome Database: the genomics resource of budding yeast. Nucleic Acids
530 Res. 2012 Jan;40(Database issue):D700–5.

531 12. Gene Ontology Consortium. Gene Ontology Consortium: going forward. Nucleic Acids Res.
532 2015 Jan;43(Database issue):D1049–56.

533 13. Flanagan PM, Kelleher RJ, Sayre MH, Tschochner H, Kornberg RD. A mediator required for
534 activation of RNA polymerase II transcription in vitro. Nature. 1991 Apr 4;350(6317):436–8.

535 14. Allen BL, Taatjes DJ. The Mediator complex: a central integrator of transcription. Nat Rev
536 Mol Cell Biol. 2015 Mar;16(3):155–66.

537 15. Denning DP, Hirose T. Anti-tubulins DEPendably induce apoptosis. Nat Cell Biol. 2014
538 Aug;16(8):741–3.

539 16. Jackson JR, Patrick DR, Dar MM, Huang PS. Targeted anti-mitotic therapies: can we
540 improve on tubulin agents? Nat Rev Cancer. 2007 Feb;7(2):107–17.

541 17. La Regina G, Bai R, Coluccia A, Famiglini V, Pelliccia S, Passacantilli S, et al. New pyrrole
542 derivatives with potent tubulin polymerization inhibiting activity as anticancer agents
543 including hedgehog-dependent cancer. J Med Chem. 2014 Aug 14;57(15):6531–52.

544 18. Baba T, Ara T, Hasegawa M, Takai Y, Okumura Y, Baba M, et al. Construction of
545 Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. Mol Syst
546 Biol. 2006;2:2006.0008.

547 19. Kim D-U, Hayles J, Kim D, Wood V, Park H-O, Won M, et al. Analysis of a genome-wide
548 set of gene deletions in the fission yeast Schizosaccharomyces pombe. Nat Biotechnol. 2010
549 Jun;28(6):617–23.

550 20. Kapitzky L, Beltrao P, Berens TJ, Gassner N, Zhou C, Wüster A, et al. Cross-species
551 chemogenomic profiling reveals evolutionarily conserved drug mode of action. Mol Syst
552 Biol [Internet]. 2010 Dec 21 [cited 2017 Feb 28];6. Available from:
553 http://msb.embopress.org/cgi/doi/10.1038/msb.2010.107

554 21. French S, Mangat C, Bharat A, Côté J-P, Mori H, Brown ED. A robust platform for chemical
555 genomics in bacterial systems. Mol Biol Cell. 2016 Mar 15;27(6):1015–25.

556 22. Babu M, Arnold R, Bundalovic-Torma C, Gagarinova A, Wong KS, Kumar A, et al.
557 Quantitative genome-wide genetic interaction screens reveal global epistatic relationships of
558 protein complexes in Escherichia coli. PLoS Genet. 2014 Feb;10(2):e1004120.

559 23. Roguev A, Bandyopadhyay S, Zofall M, Zhang K, Fischer T, Collins SR, et al. Conservation
560 and rewiring of functional modules revealed by an epistasis map in fission yeast. Science.
561 2008 Oct 17;322(5900):405–10.

562 24. Frost A, Elgort MG, Brandman O, Ives C, Collins SR, Miller-Vedam L, et al. Functional
563 repurposing revealed by comparing S. pombe and S. cerevisiae genetic interactions. Cell.
564 2012 Jun 8;149(6):1339–52.

565 25. Ryan CJ, Roguev A, Patrick K, Xu J, Jahari H, Tong Z, et al. Hierarchical modularity and
566 the evolution of genetic interactomes across species. Mol Cell. 2012 Jun 8;46(5):691–704.

567 26. Estoppey D, Hewett JW, Guy CT, Harrington E, Thomas JR, Schirle M, et al. Identification
568 of a novel NAMPT inhibitor by CRISPR/Cas9 chemogenomic profiling in mammalian cells.
569 Sci Rep. 2017 Feb 16;7:42728.

570 27. Blomen VA, Májek P, Jae LT, Bigenzahn JW, Nieuwenhuis J, Staring J, et al. Gene
571 essentiality and synthetic lethality in haploid human cells. Science. 2015 Nov
572 27;350(6264):1092–6.

573 28. Wang T, Birsoy K, Hughes NW, Krupczak KM, Post Y, Wei JJ, et al. Identification and
574 characterization of essential genes in the human genome. Science. 2015 Nov
575 27;350(6264):1096–101.

576 29. Hart T, Chandrashekhar M, Aregger M, Steinhart Z, Brown KR, MacLeod G, et al. High-
577 Resolution CRISPR Screens Reveal Fitness Genes and Genotype-Specific Cancer Liabilities.
578 Cell. 2015 Dec 3;163(6):1515–26.

579 30. Huber W, Carey VJ, Gentleman R, Anders S, Carlson M, Carvalho BS, et al. Orchestrating
580 high-throughput genomic analysis with Bioconductor. Nat Methods. 2015 Jan 29;12(2):115–
581 21.

582 31. Baryshnikova A. Systematic Functional Annotation and Visualization of Biological
583 Networks. Cell Syst. 2016 Jun 22;2(6):412–21.

584

585 **FIGURE LEGENDS**

586 **Figure 1**. **Overview of genetic interaction-based interpretation of chemical-genetic**
587 **interactions on a large scale.** Chemical-genetic interaction profiles, obtained by measuring the
588 sensitivity or resistance of a library of gene mutants to a particular compound, are compared
589 against genetic interaction profiles consisting of double mutant interaction scores. The resulting
590 similarities are aggregated at the level of biological processes to predict the process(es) perturbed
591 by the compound. Better agreement between chemical-genetic and genetic interaction profiles
592 leads to stronger process predictions.
593
594 **Figure 2. Rate of compound discovery and control of the false discovery rate for the**
595 **prediction of biological processes from chemical-genetic interaction profiles.** Biological
596 processes were predicted for compounds, negative controls (DMSO), and resampled compound
597 profiles from the RIKEN and NCI/NIH/GSK datasets. (A,C,E) The number of compounds,

experimental controls, and resampled compound profiles discovered across different significance thresholds. (B,D,F) The respective translation of panels A, C, E into estimates of the false discovery rate of biological process predictions. (A,B) Estimation of false discovery rate for predictions using CG-TARGET on the RIKEN dataset. (C,D) Same as (A,B), but for the NCI/NIH/GSK dataset. (E,F) Same as (A,B), but for predictions made using the top-100 enrichment-based process prediction method on the RIKEN data.

Figure 3. Performance of biological process prediction for well-characterized and simulated compounds. (A) Biological processes were predicted for compounds in the RIKEN dataset using both CG-TARGET and the top-100 enrichment method. For each of 35 well-characterized compounds in the RIKEN dataset with gold-standard biological process annotations, we determined the rank of its gold-standard process within its list of predictions. The number of compounds for which a rank (or better) was achieved is plotted for each process prediction method. (B) Biological processes were also predicted using both prediction methods for a set of 4500 simulated compounds, each of which was designed to target one query gene in the genetic interaction network. As such, each simulated compound inherited gold-standard process annotations from its target gene, and the ability to recapitulate gold-standard annotations within the set of rank 1 predictions was assessed using precision and recall measures.

Figure 4. Detailed analysis of the contribution of individual gene mutants to biological process predictions. (A) Schematic showing prediction of the "mRNA transport" process for NPD4142. An abbreviated chemical-genetic interaction profile for NPD4142 is shown, aligned with the corresponding array genes from the genetic interaction data and scores that reflect the importance of each of these gene mutants in making the mRNA transport prediction. Query genes were selected from the set of genes mapped to mRNA transport and were sorted from left to right in order of increasing similarity to NPD4142 (gene-level prediction score ≥ 2). Array genes were included if they possessed a chemical-genetic interaction score < -2.5 or variance > 0.02 across the selected query genes. Compounds are sorted from left to right in order of decreasing process prediction strength. (B) Schematic showing the prediction of "proteasome assembly" for several compounds. Query genes (columns) were selected from the set of genes mapped to proteasome assembly and sorted from left to right in order of increasing average similarity with the compounds predicted to "proteasome assembly" (gene-level prediction score ≥ 2). Array genes were included if they possessed a mean chemical-genetic interaction score < -2.5 or variance > 0.02 across the selected query genes in the genetic interaction data. Compounds are sorted from left to right in order of decreasing process prediction strength. Importance scores, as in (A), reflect the importance of each strain for each compound's prediction to proteasome assembly. (C) same as (B), but for the "fungal-type cell wall organization" process. (D) Global characterization of the relationship between chemical-genetic interactions and their importance to process predictions for the RIKEN dataset. Each point compares the interaction score between one gene mutant and one compound to the importance of that gene mutant in predicting the top biological process for that compound. Black points represent gene-compound pairs that possess chemical-genetic interactions (z-score $\geq \pm 2.5$) and contribute nontrivially (importance score $\geq \pm 0.1$) to the compound's top process prediction. Data for predictions made with high confidence (FDR $\leq 25\%$) were included.

643 **Figure 5. In vivo and in vitro experimental validations of biological process predictions.**
644 (A,B,C) Phenotypic validation of cell cycle-related predictions, performed on drug-
645 hypersensitive yeast treated with solvent control (DMSO) or compounds predicted to perturb the
646 cell cycle. (A) Differential interference contrast microscopy (DIC) and fluorescence upon DAPI
647 staining showing bud size and DNA localization, respectively, after compound treatment. (B)
648 FACS analysis of cell populations in different cell cycle phases at 0h, 2h, and 4h after compound
649 treatment. The green curve overlay represents the estimated cell population in G1, S and G2/M
650 phases. (C) Budding index percentages induced by treatment with compound or solvent control.
651 (D) In vitro inhibition of tubulin polymerization by compounds predicted to perturb "tubulin
652 complex assembly" (red) compared to negative control compounds (blue). Vmax values
653 reflecting the maximum rate of tubulin polymerization for each compound and negative control
654 are plotted. Assay positive and negative control compounds are colored grey. A Wilcoxon rank-
655 sum test was used to assess the significance of the difference in Vmax between the predicted
656 active compounds and the negative controls.

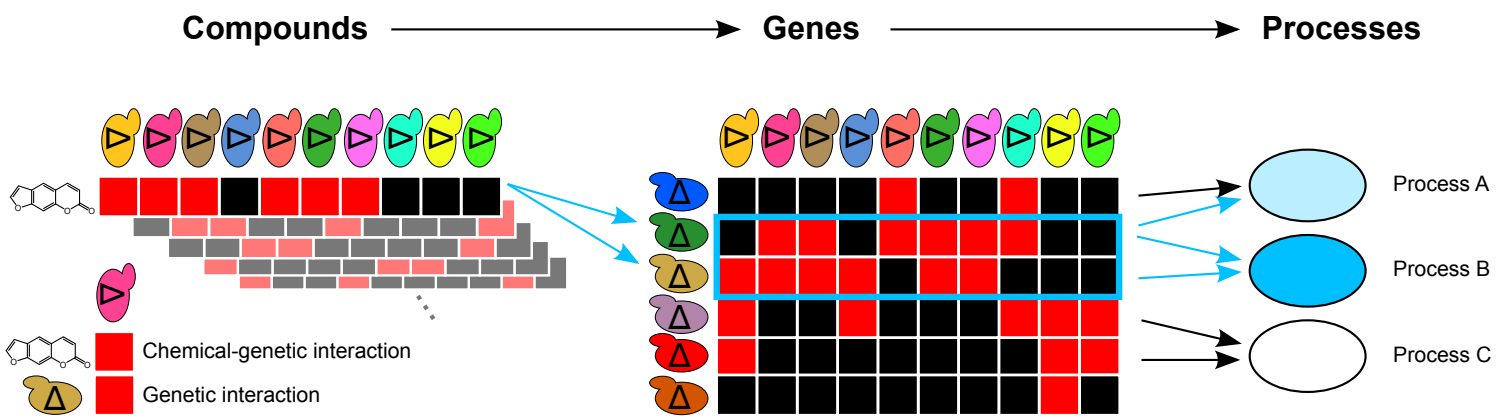

Compounds
Genes
Processes
Chemical-genetic interaction
Genetic interaction
Process A
Process B
Process C

Figure 1

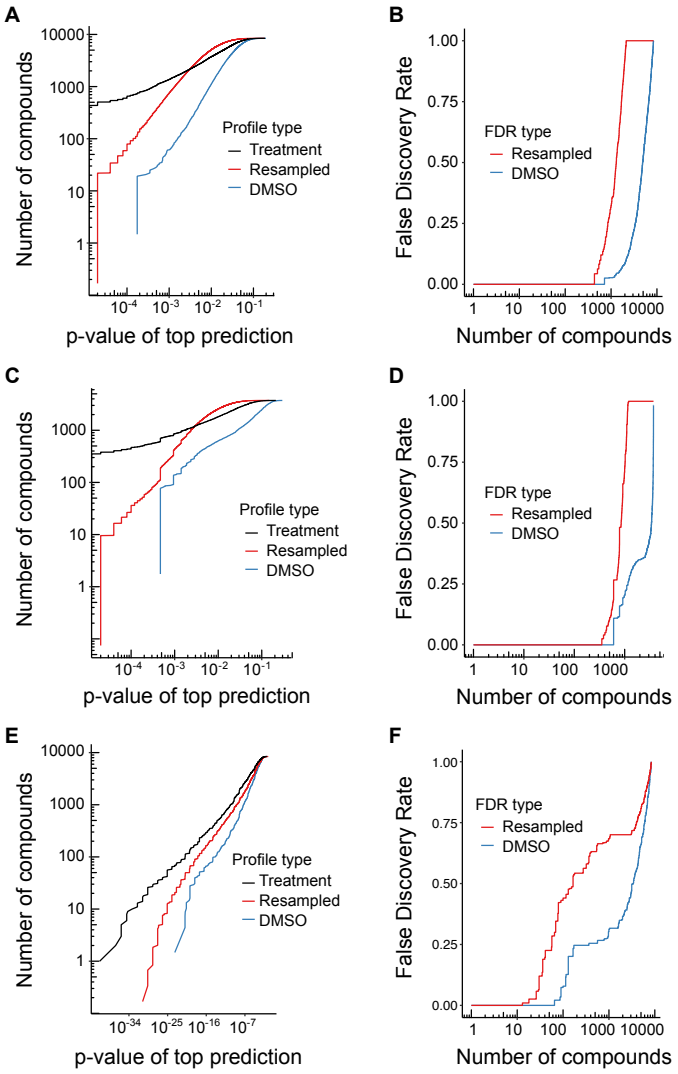

A
Number of compounds
10000
1000
100
10
1
Profile type
Treatment
Resampled
DMSO
10^{-4} 10^{-3} 10^{-2} 10^{-1}
p-value of top prediction
B
False Discovery Rate
1.00
0.75
0.50
0.25
0.00
FDR type
Resampled
DMSO
1 10 100 1000 10000
Number of compounds
C
Number of compounds
1000
100
10
1
Profile type
Treatment
Resampled
DMSO
10^{-4} 10^{-3} 10^{-2} 10^{-1}
p-value of top prediction
D
False Discovery Rate
1.00
0.75
0.50
0.25
0.00
FDR type
Resampled
DMSO
1 10 100 1000
Number of compounds
E
Number of compounds
10000
1000
100
10
1
Profile type
Treatment
Resampled
DMSO
10^{-34} 10^{-25} 10^{-16} 10^{-7}
p-value of top prediction
F
False Discovery Rate
1.00
0.75
0.50
0.25
0.00
FDR type
Resampled
DMSO
1 10 100 1000 10000
Number of compounds

Figure 2

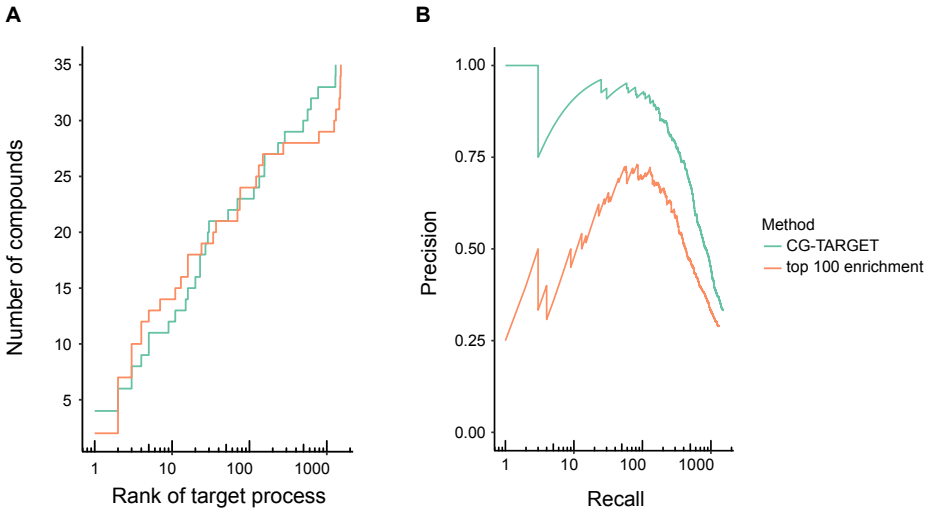
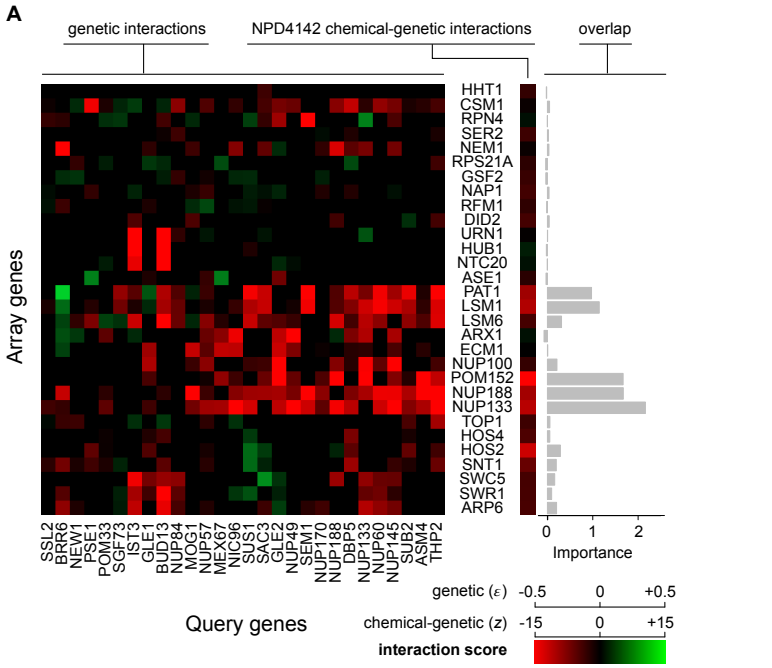
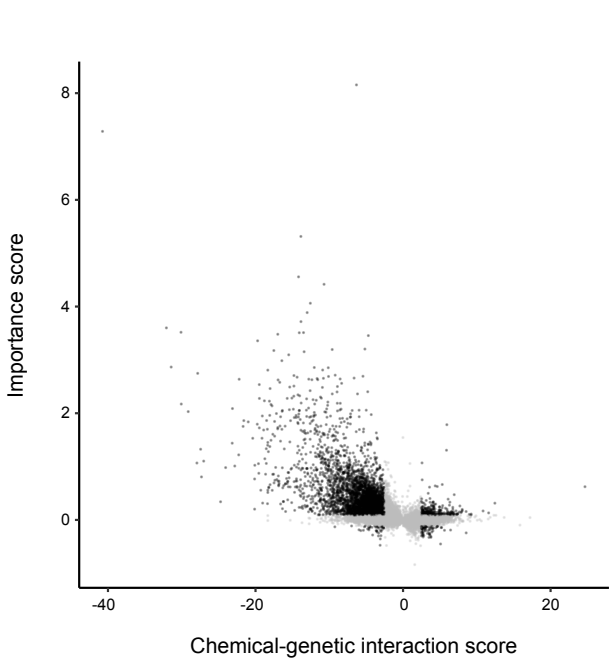
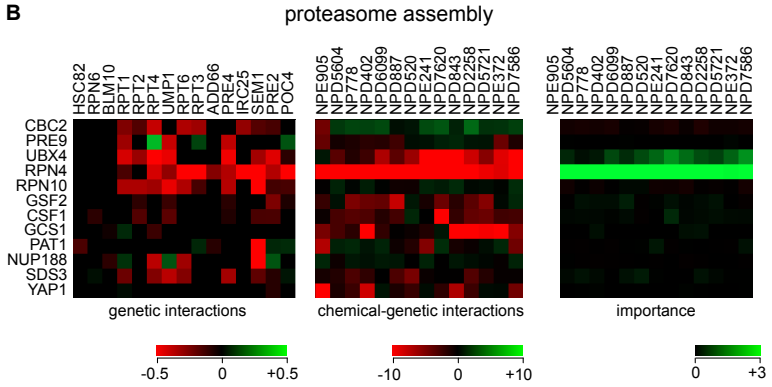

A
Number of compounds
35
30
25
20
15
10
5
1
10
100
1000
Rank of target process
B
Precision
1.00
0.75
0.50
0.25
0.00
1
10
100
1000
Recall
Method
CG-TARGET
top 100 enrichment

Figure 3


A
genetic interactions
NPD4142 chemical-genetic interactions
overlap
Array genes
Query genes
HHT1
CSM1
RPN4
SER2
NEM1
RPS21A
GSF2
NAP1
RFM1
DID2
URN1
HUB1
NTC20
ASE1
PAT1
LSM1
LSM6
ARX1
ECM1
NUP100
POM152
NUP188
NUP133
TOP1
HOS4
HOS2
SNT1
SWC5
SWR1
ARP6
SSL2
BRR6
NEW1
PSE1
POM33
SGF73
IST3
GLE1
BUD13
NUP84
MOG1
NUP57
MEX67
NIC96
SUS1
SAC3
GLE2
NUP49
SEM1
NUP170
NUP188
DBP5
NUP133
NUP60
NUP145
SUB2
ASM4
THP2
0
1
2
Importance
genetic (ε) -0.5 0 +0.5
chemical-genetic (z) -15 0 +15
interaction score


D
Importance score
8
6
4
2
0
-40
-20
0
20
Chemical-genetic interaction score


B
proteasome assembly
HSC82
RPN6
BLM10
RPT1
RPT2
RPT4
UMP1
RPT6
RPT3
ADD66
PRE4
IRC25
SEM1
PRE2
POC4
NPE905
NPD5604
NP778
NPD402
NPD6099
NPD887
NPD520
NPE241
NPD7620
NPD843
NPD2258
NPD5721
NPE372
NPD7586
CBC2
PRE9
UBX4
RPN4
RPN10
GSF2
CSF1
GCS1
PAT1
NUP188
SDS3
YAP1
genetic interactions
chemical-genetic interactions
importance
-0.5
0
+0.5
-10
0
+10
0
+3

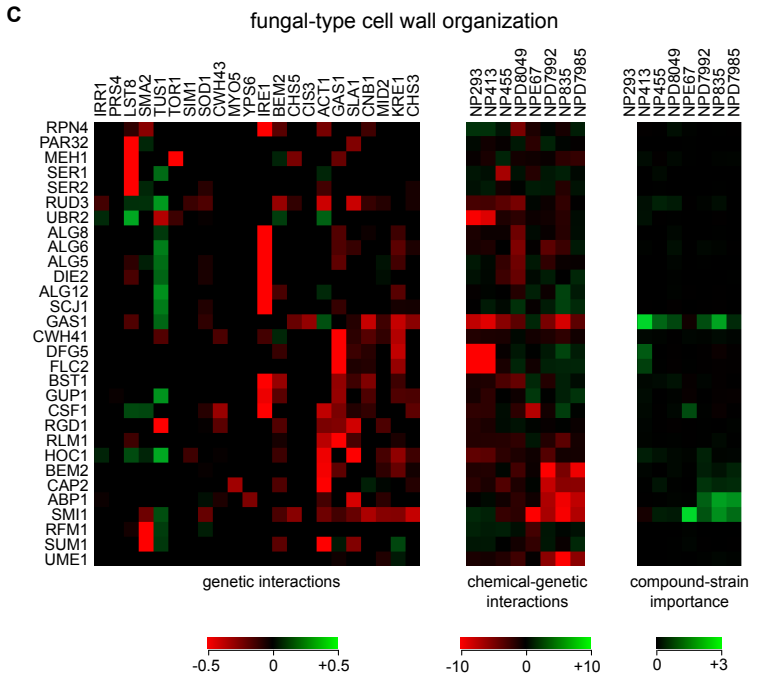
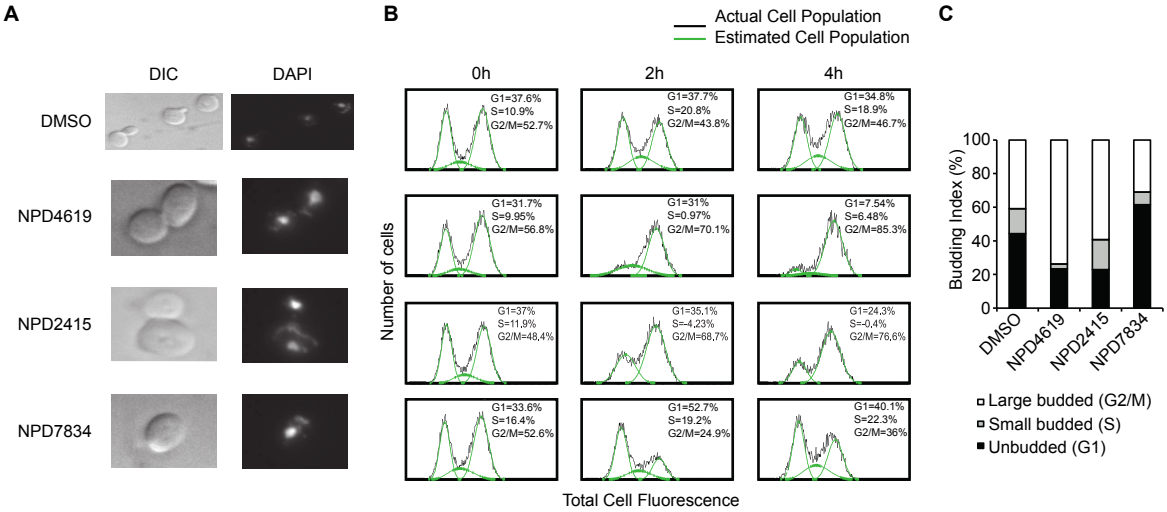

C
fungal-type cell wall organization
IRR1
PRS4
LST8
SMA2
TUS1
TOR1
SIM1
SOD1
CWH43
MYO5
YPS6
IRE1
BEM2
CHS5
CIS3
ACT1
GAS1
SLA1
CNB1
MID2
KRE1
CHS3
NP293
NP413
NP455
NPD8049
NPE67
NPD7992
NP835
NPD7985
RPN4
PAR32
MEH1
SER1
SER2
RUD3
UBR2
ALG8
ALG6
ALG5
DIE2
ALG12
SCJ1
GAS1
CWH41
DFG5
FLC2
BST1
GUP1
CSF1
RGD1
RLM1
HOC1
BEM2
CAP2
ABP1
SMI1
RFM1
SUM1
UME1
genetic interactions
chemical-genetic interactions
compound-strain importance
-0.5
0
+0.5
-10
0
+10
0
+3

Figure 4


A
DIC
DAPI
DMSO
NPD4619
NPD2415
NPD7834
B
Actual Cell Population
Estimated Cell Population
0h
2h
4h
G1=37.6%
S=10.9%
G2/M=52.7%
G1=37.7%
S=20.8%
G2/M=43.8%
G1=34.8%
S=18.9%
G2/M=46.7%
G1=31.7%
S=9.95%
G2/M=56.8%
G1=31%
S=0.97%
G2/M=70.1%
G1=7.54%
S=6.48%
G2/M=85.3%
G1=37%
S=11,9%
G2/M=48,4%
G1=35,1%
S=-4,23%
G2/M=68,7%
G1=24,3%
S=-0,4%
G2/M=76,6%
G1=33.6%
S=16.4%
G2/M=52.6%
G1=52.7%
S=19.2%
G2/M=24.9%
G1=40.1%
S=22.3%
G2/M=36%
Number of cells
Total Cell Fluorescence
C
Budding Index (%)
100
80
60
40
20
0
DMSO
NPD4619
NPD2415
NPD7834
Large budded (G2/M)
Small budded (S)
Unbudded (G1)

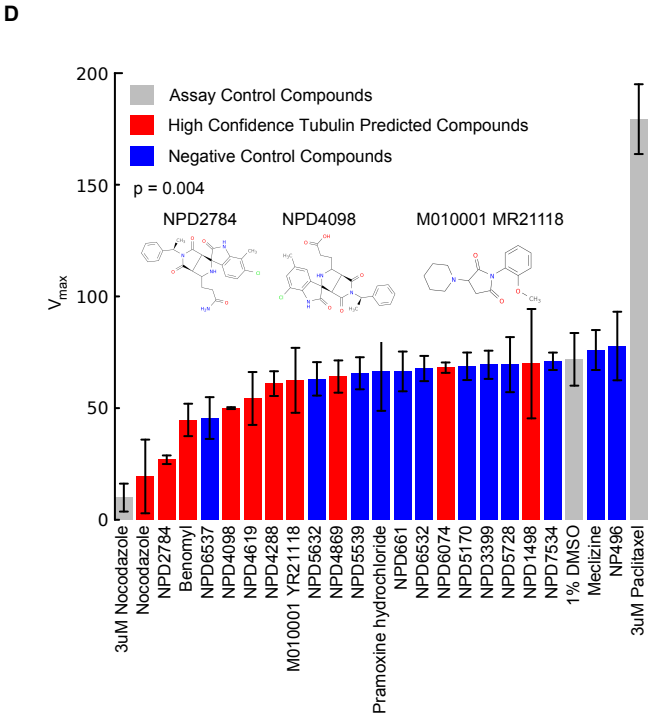

D
200
150
100
50
0
V_{max}
Assay Control Compounds
High Confidence Tubulin Predicted Compounds
Negative Control Compounds
p = 0.004
NPD2784
NPD4098
M010001 MR21118
3uM Nocodazole
Nocodazole
NPD2784
Benomyl
NPD6537
NPD4098
NPD4619
NPD4288
M010001 YR21118
NPD5632
NPD4869
NPD5539
Pramoxine hydrochloride
NPD661
NPD6532
NPD6074
NPD5170
NPD3399
NPD5728
NPD1498
NPD7534
1% DMSO
Meclizine
NP496
3uM Paclitaxel

Figure 5

Table 1. Comparison of number of compounds discovered at selected false discovery rate thresholds for CG-TARGET vs. the best-performing enrichment method (top 100 gene target candidates). The CG-TARGET method for predicting chemical-process targets was applied to two large-scale chemical-genetic interaction screens, one of compounds from the RIKEN Natural Product Depository (RIKEN) and the other consisting of 6 chemical compound collections from the National Cancer Institute, National Institutes of Health, and GlaxoSmithKline (NCI/NIH/GSK).

Dataset	RIKEN				NCI/NIH/GSK	
Prediction method	top-100 enrichment		CG-TARGET			
FDR cutoff	p-value	number of compounds	p-value	number of compounds	p-value	number of compounds
0.00	7.4E-32	13	0.00E+00	586	0.00E+00	352
0.05	9.3E-30	26	2.00E-05	610	4.00E-05	405
0.10	5.2E-29	30	8.00E-05	649	1.60E-04	494
0.25*	3.5E-25	57	2.80E-04	848	4.70E-04	705

*This cutoff is 0.27 for the NCI/NIH/GSK dataset

Table 2. Evaluation against literature-derived, gold standard compound-process annotations. The target process rank was determined by its position in the list of all process predictions for each gold standard compound, with the significance computed empirically by shuffling the processes and re-computing the rank (bold p-values indicate significance, $p < 0.05$). Grey boxes indicate cases in which the false discovery rate of the gold standard compound-process prediction was less than 25%. The "effective rank" represents the rank of the process prediction after considering its similarity to processes predicted with higher confidence.

		top-100 enrichment			CG-TARGET		
Chemical	Target GO process	Target process rank	Rank significance	Effective rank	Target process rank	Rank significance	Effective rank
Aclacinomycin A	DNA conformation change (GO:0071103)	11	**0.0076**	1	1	**6.00E-04**	1
Caffeine	TOR signaling cascade (GO:0031929)	16	**0.0113**	4	1	**6.00E-04**	1
Mycophenolic acid	DNA metabolic process (GO:0006259)	1	**6.00E-04**	1	1	**6.00E-04**	1
Tunicamycin	glycosylation (GO:0070085)	4	**0.004**	2	1	**6.00E-04**	1
Benomyl	microtubule-based process (GO:0007017)	2	**0.0015**	1	2	**0.0021**	2
Nocodazole	microtubule-based process (GO:0007017)	16	**0.0099**	4	2	**0.0021**	2
Bortezomib	protein catabolic process (GO:0030163)	4	**0.0033**	1	3	**0.0025**	1
MMS	DNA repair (GO:0006281)	2	**0.0015**	1	3	**0.0025**	1
Hedamycin	DNA repair (GO:0006281)	3	**0.0022**	1	4	**0.0034**	1
Tyrocidine B	cell wall organization or biogenesis (GO:0071554)	24	**0.0165**	5	5	**0.0035**	1
Haloperidol	steroid metabolic process (GO:0008202)	13	**0.0097**	2	5	**0.0043**	2
Oligomycin A	response to pH (GO:0009268)	3	**0.0022**	2	9	**0.0045**	2
Latrunculin B	cytoskeleton organization (GO:0007010)	37	**0.0243**	2	11	**0.0083**	1
Mitomycin C	DNA replication (GO:0006260)	3	**0.0022**	1	15	**0.0093**	4
Camptothecin	DNA conformation change (GO:0071103)	274	0.181	1	16	**0.0142**	4
Furazolidone	DNA replication (GO:0006260)	7	**0.0054**	1	20	**0.014**	4
FK228	chromatin organization (GO:0006325)	2	**0.0015**	1	23	**0.0168**	2
Trichostatin A	chromatin organization (GO:0006325)	5	**0.003**	2	23	**0.0168**	3
5-Fluorocytosine	RNA biosynthetic process (GO:0032774)	2	**0.0015**	2	27	**0.0188**	2
Hydroxyurea	DNA replication (GO:0006260)	76	**0.0478**	2	29	**0.02**	6
Acriflavine	DNA metabolic process (GO:0006259)	1	**6.00E-04**	1	30	**0.0214**	1
Podophyllotoxin	microtubule-based process (GO:0007017)	149	0.0965	8	53	**0.0401**	6
Daunorubicin	DNA replication (GO:0006260)	1245	0.8145	139	70	0.0551	21

Fluconazole	steroid metabolic process (GO:0008202)	70	**0.0481**	6	114	0.0877	12
Cisplatin	DNA replication (GO:0006260)	34	**0.0208**	1	134	0.1025	23
Rapamycin	TOR signaling cascade (GO:0031929)	790	0.5843	7	156	0.1164	8
Nigericin	Golgi vesicle transport (GO:0048193)	2	**0.0015**	1	157	0.1226	13
Itraconazole	steroid metabolic process (GO:0008202)	122	0.08	18	234	0.1748	29
Gramicidin S	cell wall organization or biogenesis (GO:0071554)	75	0.0506	11	286	0.2142	39
Micafungin	cell wall organization or biogenesis (GO:0071554)	132	0.0891	21	495	0.3809	47
Brefeldin A	ER to Golgi vesicle-mediated transport (GO:0006888)	1452	0.9541	196	565	0.4198	32
Calcofluor White	cell wall organization or biogenesis (GO:0071554)	1498	0.9815	177	624	0.4577	90
Blasticidin S	translation (GO:0006412)	1519	0.9962	258	772	0.5742	57
Griseofulvin	microtubule-based process (GO:0007017)	1313	0.8642	244	1291	0.9723	227
Polyoxin D	cell wall organization or biogenesis (GO:0071554)	1490	0.976	127	1302	0.9773	225
	Number with significant rank		23			22	
	Number with significant rank and prediction FDR <= 25%		2			16	