A community computational challenge to predict the activity of pairs of compounds

Mukesh Bansal[1,2,62], Jichen Yang[3,62], Charles Karan[4,62], Michael P Menden[5], James C Costello[6,61], Hao Tang[3], Guanghua Xiao[3], Yajuan Li[7], Jeffrey Allen[3], Rui Zhong[3], Beibei Chen[3], Minsoo Kim[3,8], Tao Wang[3], Laura M Heiser[9], Ronald Realubit[4], Michela Mattioli[10], Mariano J Alvarez[1,2], Yao Shen[1,2], NCI-DREAM Community[11], Daniel Gallahan[12], Dinah Singer[12], Julio Saez-Rodriguez[5], Yang Xie[3,8], Gustavo Stolovitzky[13] & Andrea Califano[1,2,14–17]

© 2014 Nature America, Inc. All rights reserved.

Recent therapeutic successes have renewed interest in drug combinations, but experimental screening approaches are costly and often identify only small numbers of synergistic combinations. The DREAM consortium launched an open challenge to foster the development of *in silico* methods to computationally rank 91 compound pairs, from the most synergistic to the most antagonistic, based on gene-expression profiles of human B cells treated with individual compounds at multiple time points and concentrations. Using scoring metrics based on experimental dose-response curves, we assessed 32 methods (31 community-generated approaches and SynGen), four of which performed significantly better than random guessing. We highlight similarities between the methods. Although the accuracy of predictions was not optimal, we find that computational prediction of compound-pair activity is possible, and that community challenges can be useful to advance the field of *in silico* compound-synergy prediction.

Recent success in the study of synergistic combinations, such as the use of CHK1 inhibitors in combination with several DNA damaging agents[1] or of the PARP inhibitor olaparib in combination with the PI3K inhibitor BKM120 (ref. 2), have generated significant interest in the systematic screening of compound pairs to identify synergistic pairs for combination therapy. Compound synergy can be measured by multiple endpoints, including reducing or delaying the development of resistance to treatment[3] (for instance by abrogating the emergence of resistant clones[4–6]), improving overall survival[7,8] or lowering toxicity by decreasing individual compound dose[9].

Similarly, at the molecular level, synergistic interactions can be implemented by several distinct mechanisms. For instance, a compound may sensitize cells to another compound by regulating its absorption and distribution, modulating the cell's growth properties[10], inhibiting compound degradation[11], inhibiting pathways that induce resistance[6] or reducing the other compound's toxicity[12]. When used in combination, two compounds may elicit one of three distinct responses: (i) additive, when the combined effect is equivalent to the sum of the independent effects; (ii) synergistic, when the combined effect is greater than additive; and (iii) antagonistic, when the combined effect is smaller than additive. The goal of combination therapy is thus to attain a synergistic or at least an additive yet complementary effect.

Most approaches to identify synergistic compound pairs are still exploratory[13,14]. In cancer research, synergy assays are usually performed by treating cell lines *in vitro* with all possible compound combinations from a diverse library or with candidate combinations selected on the basis of mechanistic principles. Unfortunately, such experimental screens impose severe limits on the practical size of compound diversity libraries. Computational methods to predict compound synergy can potentially complement high-throughput synergy screens, but the few that have been published lack rigorous experimental validation or are appropriate only for compounds that modulate well-studied molecular pathways[15] or that are equivalent to previously established combinations[16]. Current algorithms are not generalizable to arbitrary compound combinations unless molecular profile data following compound-pair treatment are available[17], which is clearly impractical. Thus, there is a need for new methods to predict compound synergy from molecular profiles of single compound activity, as well as for assays designed to objectively and systematically evaluate the accuracy and specificity of such predictions.

To address this issue, the DREAM Challenges initiative (an effort run by a community of researchers that poses fundamental questions in systems biology and translational science in the form of

[1]Department of Systems Biology, Columbia University, New York, New York, USA. [2]Center for Computational Biology and Bioinformatics, Columbia University, New York, New York, USA. [3]Quantitative Biomedical Research Center, Department of Clinical Sciences, University of Texas Southwestern Medical Center, Dallas, Texas, USA. [4]Columbia Genome Center, High Throughput Screening Facility, Columbia University, New York, New York, USA. [5]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK. [6]Howard Hughes Medical Institute, Department of Biomedical Engineering and Center of Synthetic Biology, Boston University, Boston, Massachusetts, USA. [7]Department of Immunology, University of Texas Southwestern Medical Center, Dallas, Texas, USA. [8]Simmons Comprehensive Cancer Center, University of Texas, Southwestern Medical Center, Texas, USA. [9]Department of Biomedical Engineering, Oregon Health and Science University, Portland, Oregon, USA. [10]Center for Genomic Science of IIT@SEMM, Fondazione Istituto Italiano di Tecnologia (IIT), Milan, Italy. [11]Full lists of members and affiliations appear below. [12]Division of Cancer Biology, National Cancer Institute, Bethesda, Maryland, USA. [13]IBM Computational Biology Center, IBM, T.J. Watson Research Center, Yorktown Heights, New York, USA. [14]Department of Biomedical Informatics, Columbia University, New York, New York, USA. [15]Department of Biochemistry and Molecular Biophysics, Columbia University, New York, New York, USA. [16]Institute for Cancer Genetics, Columbia University, New York, New York, USA. [17]Herbert Irving Comprehensive Cancer Center, Columbia University, New York, New York, USA.

Received 19 February; accepted 25 September; published online 17 November 2014; doi:10.1038/nbt.3052

crowdsourced challenges), in collaboration with the National Cancer Institute, organized a community-based challenge to systematically and objectively test methods to computationally predict compound-pair activity in human B cells. Challenge participants were asked to rank 91 compound pairs (all pairs of 14 compounds) from the most synergistic to the most antagonistic in the OCI-LY3 human diffuse large B-cell lymphoma (DLBCL) cell line (**Fig. 1**), based on the gene expression profiles of cells perturbed with the individual compounds. Predictions were then evaluated against an experimentally assessed gold standard, generated by systematic evaluation of compound-pair synergy *in vitro*. This data set was originally intended to experimentally validate the SynGen algorithm, which we introduce for the first time in this paper. However, we chose to first give the community the opportunity to develop *in silico* methods for synergy predictions. Therefore, we also evaluated SynGen, which, by introducing original ideas of synergy prediction, complements the 31 methods that participated in the DREAM challenge.

We present a comparative blind-assessment of all 31 methods submitted to the DREAM Challenge as well as a nonblind assessment of SynGen. Comparative analyses suggest that some *de novo*, *in silico* compound synergy prediction methods can achieve a performance that is statistically significantly better than random guessing. Moreover, integrating the methods can further increase performance. Although these results are encouraging, there is still much room for performance improvement.

RESULTS

Summary of data set and challenge

Participants were provided with (i) dose-response curves for viability of OCI-LY3 cells following perturbation with 14 distinct compounds (**Supplementary File 1**), including DMSO as a control media, (ii) gene expression profiles (GEP) in triplicates of the same cells untreated (baseline) and at 6 h, 12 h and 24 h following perturbation with each of the 14 compounds, and (iii) the previously reported[18] baseline genetic profile of the OCI-LY3 cell line (**Fig. 1**). Two compound concentrations were used, including the compound's IC_{20} (concentration of drug needed to kill 20% of cells) at 24 h and the compound's IC_{20} at 48 h, as assessed from nine-point titration curves. Any additional baseline data from the literature or experimental assays were considered admissible in the challenge, but direct measurement of compound synergy, even in limited format, was expressly prohibited. Challenge participation required ranking each of the 91 compound pairs from most synergistic to most antagonistic.

The 31 predictions submitted to this challenge showed considerable diversity in the methods and data used (**Table 1** and **Supplementary Table 1**). This reflects the lack of standard approaches for predicting compound-pair activity from transcriptomic data and the lack of training data (that is, pairs of compounds known to be synergistic or nonsynergistic), intentionally preventing use of established machine learning methods. Despite broad methodological diversity, of the 31 teams, 10 based their predictions on the hypothesis that compounds with higher transcriptional profile similarity (different similarity definitions were used) were more likely to be synergistic (similarity hypothesis). In contrast, eight teams assumed the opposite (dissimilarity hypothesis). The remaining teams either used a combination of similarity and dissimilarity hypotheses (combination hypothesis, $n = 4$) or used more complex hypotheses ($n = 9$). Only two teams made explicit use of OCI-LY3 genetic profiles, suggesting either that genomic data are deemed not useful to this analysis or that their use in predicting compound synergy is not yet developed. Finally, 12 teams relied only on provided information, whereas the others used additional literature information, such as generic pathway knowledge, compound structure or targets, and substrates of these compounds.

Performance evaluation

To objectively evaluate challenge submissions, we generated a gold-standard data set based on the experimental assessment of OCI-LY3 cell viability for the 91 compound pairs used in the challenge, at 60 h. The joint compound-pair activity was estimated using excess over Bliss (EOB) (**Supplementary Fig. 1**), which determines whether the combined effect of two compounds is significantly greater or smaller than the naive (independent) combination of their individual effects. These activity estimates were used to rank all pairs from most synergistic to most antagonistic (**Supplementary Table 2**).

Predictions were scored using a modified version of the concordance index[19] called the probabilistic concordance-index (PC-index, **Supplementary Note 1**). This metric quantifies the concordance between the ranking of compound pairs in the gold standard (**Fig. 2a**) and the predicted ranking in each submission, accounting for experimental measurement errors in the estimation of the EOB, that is, it estimates the average fraction of compound pairs, over all experimental replicates, ranked correctly when rankings of pairs of

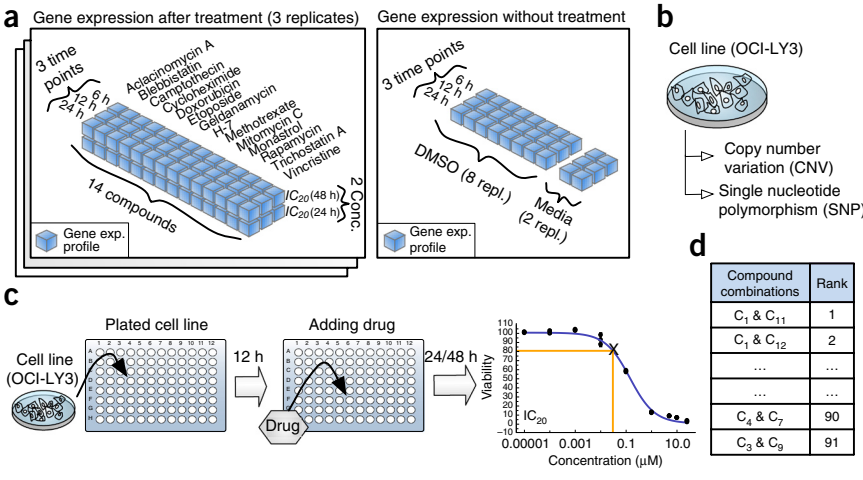

Figure 1 Overview of data sets used in the NCI-DREAM compound-pair activity challenge. (**a**) Gene expression profiles of baseline samples, DMSO-treated and 14 single compound–treated samples are generated at three different time points (6, 12 and 24 h) and two different compound concentrations (IC_{20} at 24 and 48 h, where IC_{20} is defined as the compound concentration that kills 20% of cells). Compound-treated samples were generated in triplicate, baseline samples in duplicate and DMSO-treated samples in octuplicate. (**b**) The baseline genetic profile of the OCI-LY3 cell line obtained previously[18] was provided to the participants. (**c**) Participants were also provided with the dose-response curve following single treatment. The curves were derived from a single-agent treatment of OCI-LY3 for the indicated time. X represents IC_{20} concentration of a compound. (**d**) Participants were required to rank each of the 91 pairwise compound combinations of 14 compounds from the most synergistic to the most antagonistic. Any additional data derived by participants through analysis of the literature were considered admissible in the challenge. Assays to experimentally test compound synergy, even in a limited format, were expressly prohibited.

© 2014 Nature America, Inc. All rights reserved.

compound pairs are compared with their respective experimental rank. Other methods such as concordance index or correlation assume no ambiguity in the observed ranks. However, experimental noise causes uncertainty in ranking compound pairs, thus making these methods unsuitable for scoring the predictions. Source code for the PC-index can be found in **Supplementary Software 1**.

Of 31 predictions (SynGen was evaluated separately as it was developed by one of the Challenge organizers and therefore it

Table 1 Summary of methods and data used by each participant

Rank	PC-index	Summary	Data type used
Similarity in compound activity leads to synergy			
2[a]	0.60518	Identified a set of core genes defined by statistically significant DEGs in at least one compound treatment and used these genes to estimate interaction score by calculating number of overlapping genes, taking direction of regulation into account.	GD
4[a]	0.57529	Computed a Pearson correlation between gene expression profiles of two compounds using genes DEGs in at least one compound treatment.	G
7[a]	0.56219	Used support vector machine, trained using chemical properties, chemogenomic profiling and gene-expression data of a set of synergistic fungicidal compounds[13].	GA
8[a]	0.5507	Designed a scoring function that combines target and transporter information of each compound, DEGs, their *t*-score and the number of common DEGs between two compounds[38,39].	GDA
9[a]	0.5327	Used the rank-aggregation method to combine results obtained from compound-pair similarity using correlation, common compound affected pathways, set of common compound-gene interactions (from ChEMBL), compound-genes interaction for one compound that are significantly affected by other compounds and compound pairs in the same clinical trial[40,41].	GPDA
11[a]	0.52848	Determined cell viability by predicting activation of biological pathways in response to a single compound treatment and combined this with dose-response curves[42].	GPD
12[a]	0.52779	Used score combining overlap of gene expression signatures of individual compound treatments and cell line-specific signature derived from external datasets, taking direction of regulation into account.	GA
14[a]	0.51854	Constructed probable pathways connecting compound targets and DEGs and used the Jaccard score based on gene co-occurrences in these pathways.	GPA
15[a]	0.51624	Used weighted Euclidean distance, weighted by activity of each compound.	GD
31[a]	0.41993	Computed correlation between gene expression profiles of two compounds using genes DEGs in at least one compound treatment.	G
Dissimilarity in compound activity leads to synergy			
5[a]	0.56637	Computed the Manhattan distance between pathways significantly enriched by each compound.	GP
6[a]	0.56495	Designed a geometric-based score using the number of significant DEGs, the number of common DEGs between two compounds, the correlation between their gene expression profile and the dose-response curve.	GD
18[a]	0.50653	Applied the Pareto ranking strategy using compound activity as well as chemical and target similarity[43].	GA
19[a]	0.50501	Built a model that measures the effect on each of the 15 core signaling pathways by considering the number of significant DEGs, the number of common DEGs between two compounds and the direction of regulation.	GP
20[a]	0.49602	Built a cooperative score by combining the number of significant DEGs, the number of common DEGs between two compounds and the correlation between weighted gene expression profiles.	G
24[a]	0.46791	Identified a set of core genes defined by statistically significant DEGs in at least one compound treatment and used these genes to estimate interaction score by calculating the number of overlapping genes, taking direction of regulation into account.	G
25[a]	0.45415	Estimated deviation between correlation using gene expression profile and correlation using GO terms enriched by two compounds and used that as a measure of synergy.	GP
26[a]	0.44467	Built a model combining the IC_{20} concentration of two compounds and the correlation between their gene expression profiles.	G
Combination of similarity and dissimilarity in compound activity leads to synergy			
1[a]	0.61303	Drug Induced Genomic Residual Effect (DIGRE) model (see main text).	GPDA
3[a]	0.59981	Drug Induced Genomic Residual Effect (DIGRE) model (see main text; different cut-off for feature selection).	GPDA
21[a]	0.48988	Estimated the similarity between compound pairs using DEG's and pathway information and combined this similarity with dose-response curves.	GPDA
29[a]	0.42992	Linear interpolation between two dose points on dose-response curve using similarity between each compound pair, calculated by overlap of DEGs.	GPDA
Complex synergistic relationship			
10[a]	0.52974	Used expression of genes, identified from the public dataset whose expression are correlated with overall survival, to predict cell viability.	GDA
13[a]	0.51952	Used OCI-LY3 virtual baseline created from The Cellworks proprietary Tumor Cell Technology, trained using the known mode of actions of the compounds[44].	GPDA
16[a]	0.50927	Built a model linking gene expression and cell viability. Used predicted gene expression profile after a compound combination in this model to infer the cell viability of compound pairs.	GD
17	0.50703	Identified potential effective targets by comparing expression profiles of effective and ineffective compounds and computed the sum of log-odd ratio for each compound pair under the naïve Bayes assumption.	G
22[a]	0.48568	Built a bagged regression trees model using features obtained from known synergistic and antagonistic compound pairs from published literature[45].	NA
23[a]	0.47183	Used expression of genes, identified from the public dataset whose expression are correlated with overall survival, to predict cell viability.	GDA
27[a]	0.44346	Used the Bayesian estimation of temporal regulation and the nearest template prediction algorithm with cosine distance to associate significantly DEGs between pairs of drugs[46].	GD
28[a]	0.43479	Predicted expression profile after the treatment with 2 compounds using ANOVA based liner regression and built a model linking gene expression and cell viability.	GD
30	0.42297	Used model trained using target and chemical structure of known compound combinations in cancer therapy along with protein-protein interactions.	GPA

All methods are categorized into four groups based on distinct hypotheses used by various teams. G, gene expression profile; A, additional information not provided in the challenge; D, dose-response curve; P, pathway.
[a]Detailed method description is available in the **Supplementary Note 2**. The summary reported in this table was obtained directly from the participants.

© 2014 Nature America, Inc. All rights reserved.

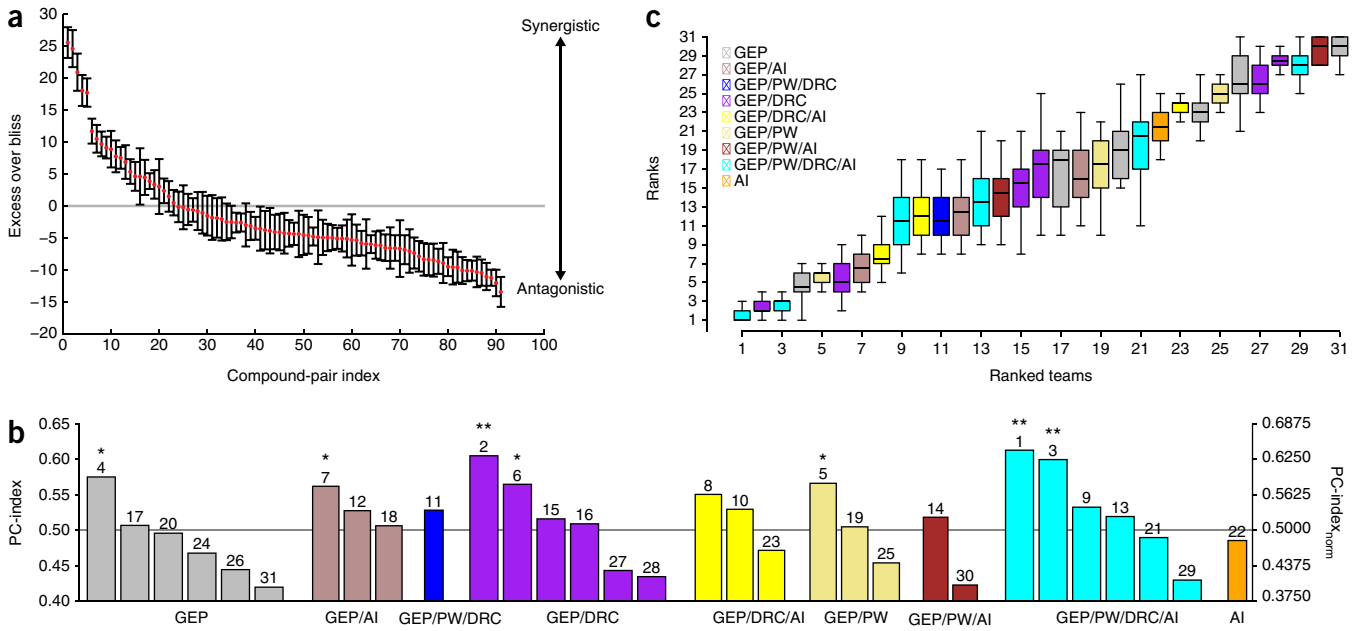

Figure 2 Gold standard data for evaluation and performance of predictions. (**a**) The results of excess over Bliss for all compound pairs ranked from most synergistic to most antagonistic. Error bars represent the s.e.m. of excess over Bliss, estimated from five experimental replicates. The solid gray line at excess over Bliss equals 0 and represents a line over and below which compound pairs are generally considered synergistic and antagonistic, respectively. (**b**) PC-index for all participants grouped by the kind of data or information used by their method. There is no apparent correlation between the final score with the kind of data or information used. AI, additional information other than pathway information used; DRC, dose-response curve used; GEP, gene expression profile used; PW, pathway information used. The rank of each team is reported on the top of the bar. The gray line represents random performance. The *y* axis on right shows the PC-index$_{norm}$ where PC-index is normalized to have a score between 0 and 1. *FDR ≤ 0.20; **FDR ≤ 0.05. (**c**) Box plot showing the median, quartile and range of ranks for each team in leave-one-out test. All teams are sorted by their PC-index. Teams are color coded with the kind of data or information used by their methods.

did not participate in the Challenge on the same terms as the other 31 methods), three methods (DIGRE, IUPUI_CCBB and DPST) produced predictions that were statistically significant at a conservative false-discovery rate threshold (FDR = 0.05) (**Fig. 2b**). Independent of whether these methods may help in planning large synergy screens, this suggests that compound synergy prediction is possible. Furthermore, our challenge-based blind assessment of these methods' performance provides a realistic and effective baseline for further methodological development.

We found little obvious association between method performance and data utilization (**Fig. 2b** and **Supplementary Fig. 2**). Only use of gene expression profiles at 24 h following treatment showed a minimal effect on performance (**Supplementary Fig. 2**). However, this trend was not statistically significant and additional data will be required to evaluate it. Additionally, distinct hypotheses used by the teams may have had an influence on performance (**Supplementary Fig. 3** and **Table 1**). For instance, teams using similarity or combination hypotheses achieved overall a higher PC-index compared to other teams using other hypotheses. However, these differences are not statistically significant and are reported here for completeness.

To test performance consistency, we scored each prediction using a second metric (resampled Spearman correlation). Both metrics yielded virtually identical performance evaluation (correlation $r = 0.99$), with only small differences in rank for a few methods that did not perform better than random (**Supplementary Fig. 4**). The robustness of the prediction ranking was tested by removing one compound at a time from the set and considering the remaining 13 compounds (leave-one-out). This analysis revealed that the predictions from two best-performing methods consistently ranked in the top 5 across each of the 14 different rankings obtained by removing each compound, suggesting that their predictions are only weakly biased by any specific drug selection (**Fig. 2c** and **Supplementary Fig. 5**). The remaining methods showed much greater variation in their performance.

Best performing methods

The best performing method, DIGRE (drug-induced genomic residual effect) hypothesizes that when cells are sequentially treated with two compounds, the transcriptional changes induced by the first contribute to the effect of the second (**Fig. 3a**). This is consistent with the observation that sequential drug administration affects outcome[20,21]. Thus, although compounds were administered simultaneously in the experimental assays, the algorithm models synergy sequentially. DIGRE implements three major steps. The first step involves comparing transcriptional changes following individual compound treatment to derive a compound-pair similarity score. This is obtained, first, by overlapping differentially expressed genes after treatment with the two compounds with eight cell growth–related KEGG pathways (focused view), and second, by considering genes upstream of the differentially expressed genes in 32 cancer-relevant KEGG pathways (global view). In the second step, the effects of compound-induced transcriptional changes on cell survival are approximated using a compound similarity score *r*, defined as: $(1 - f_{B+A'}) = (1 - rf_{2B})[1 - (1 - r)f_B]$, assuming that samples were treated first with compound A (where ′ suggested primary treatment) followed by compound B (**Fig. 3b**). Here, $f_{B+A'}$ represents the cell viability reduction after B treatment, following the transcriptional changes induced by A, r is the compound-pair similarity score, f_B is the viability reduction after B treatment, and f_{2B} is the viability reduction for a double dose of B, estimated from the dose-response curve. The final step introduces a combined score defined

© 2014 Nature America, Inc. All rights reserved.

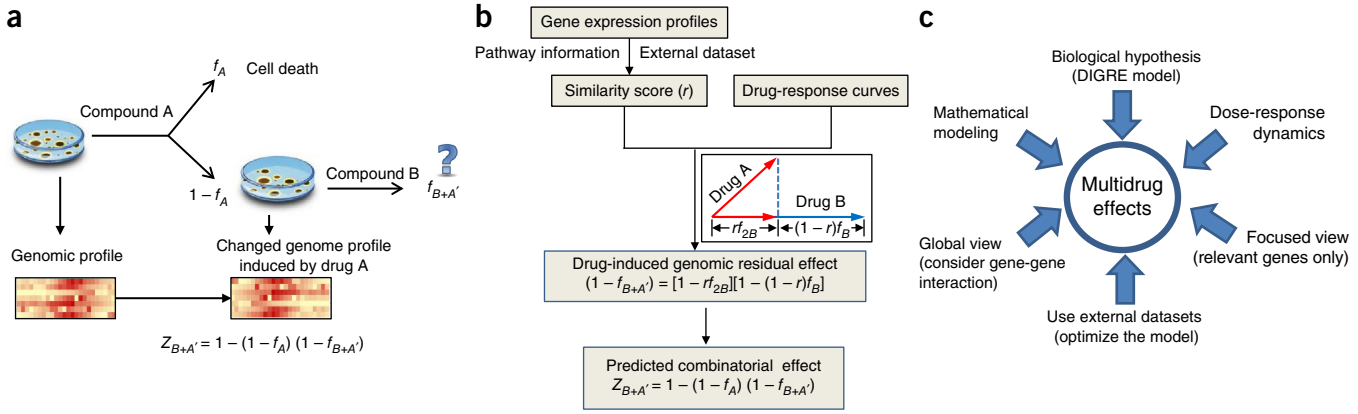

Figure 3 The DIGRE model. (**a**) Biological hypotheses of the DIGRE model submitted by the best-performing team. The combined compound effect for compounds A and B is hypothesized to result from the compound-induced genomic residual effect. If cells were treated by compounds A and B sequentially, the genomic changes induced by compound A will further contribute to the effect induced by compound B. Here, f_X denotes the percentage of cells killed by compound X and $f_{B+A'}$ represents the cell viability reduction after B treatment, following the transcriptional changes induced by A. Based on this hypothesis, the estimation of the combinatorial compound effect ($Z_{B+A'}$) reduces to the estimation of the compound-induced genomic residual effect ($f_{B+A'}$) (**Supplementary Note 2**). (**b**) Workflow of DIGRE. (Step 1) The genomic or transcriptome changes induced by two compounds are compared. The similarity score is refined by using pathway information and an external training data set. (Step 2) A mathematical model incorporates the similarity score and the dose-response curves to estimate the compound-induced genomic residual effect. (Step 3) A combined score is estimated for each of the two possible sequential orders of treatment and finally the synergistic score is estimated as the average combined score obtained by two possible sequential orders of treatment. (**c**) Key ingredients of the DIGRE model.

as $Z_{B+A'} = 1 - (1 - f_A)(1 - f_{B+A'})$, where f_A is the viability reduction after A treatment. Finally, the fraction of dead cells Z, also defined as synergistic score, is estimated as the average between the two possible sequential orders of treatment (**Supplementary Note 2**).

Our analyses suggest that the following factors contribute to DIGRE performance (**Fig. 3c**): (i) the hypothesis that compound synergy is at least partially due to compound-induced transcriptomic residual effects, which are the transcriptional changes induced by the first compound that contribute to the cell inhibition effect of both compounds; (ii) using explicit mathematical models to quantify the relationships between transcriptomic changes and compound synergy (i.e., analysis of compound-induced transcriptomic residual effects and compound similarity score); (iii) using information from the full dose-response curve instead of just the IC_{20} data; (iv) incorporating pathway information (focused view) and gene-gene interactions (global view) to measure similarity between transcriptomic changes induced by different compounds; and (v) using external data sets to optimize pathway selection and model parameters. When each of these factors was systematically removed from the analysis, the algorithm performance decreased. In particular, the residual effect hypothesis is critical as its removal completely abrogates the algorithm's predictive power (**Supplementary Note 2** and **Supplementary Fig. 6**).

The second-best-performing method (IUPUI_CCBB) hypothesized that the activity of a compound can be estimated from its effect on the genes that are significantly differentially expressed following treatment with highly toxic compounds versus control media. Compound synergy or antagonism is then determined by computing whether the effect of two compounds on this set of genes is concordant or discordant, by means of a compound-pair interaction score.

For methods that are not based on machine learning and cannot thus rely on positive and negative examples, performance is determined by how well they model the underlying biology of the process. As such, the best-performing algorithms exploited the dose-response curve, the concept that one compound may have a faster pharmacodynamics than the other, and also the fact that synergy was estimated by excess over Bliss additivity. Further information about all participating methods and the source codes for the DIGRE and IUPUI_CCBB methods can be found in **Supplementary Note 2** and **Supplementary Software 2** and **3**.

Community-based methods

Participants used quite distinct computational strategies resulting in at least partially statistically independent predictions that may be complementary. This suggests that their integration may outperform individual methods. Similar integrative methods have been successful in a variety of biological challenges, such as predicting the disorder of proteins[22], identifying monoamine oxidase inhibitors[23], inferring gene regulation[24], and in cancer prognostics[25] and diagnostics[26].

To test the predictive power of integrative approaches, we divided the gold standard data set into three subsets (S1, S2 and S3). S1 was used to sort the methods from best to worst performers, S2 for determining how many of the top-performing methods should be integrated to achieve optimal performance, and S3 for determining the final performance of the best individual and best integrative methods independent of training bias, thus avoiding overfitting. This provides a practical approach to implement multimethod integration as a crowdsourcing activity. Based on 1,000 distinct S1, S2 and S3 splits, we found that integrative methods consistently and significantly outperformed the best single methods obtained from S1 ($P \leq 10^{-36}$, by Wilcoxon rank sum test, **Fig. 4a**) in more than 75% of splits (**Supplementary Fig. 7**). When using only single best method's performance (ordered according to set S1), average integrative performance (when evaluated independently in both set S2 and S3) peaked at ~7% improvement, when the first seven methods were integrated, and decreased monotonically when more than seven methods were used (**Fig. 4b**). When also selecting the optimal number of methods (based on evaluations in S2), integration of the first 4–6 methods, on average, produced the best result (**Supplementary Fig. 8**). Critically, challenge submissions were evaluated using the full 91 compound-pair set, whereas predictive power of the crowdsourcing approach was evaluated using only a third of these (i.e., 30 compound pairs). This consideration should be taken into account when comparing the PC-index in **Figures 2b** and **4**.

© 2014 Nature America, Inc. All rights reserved.

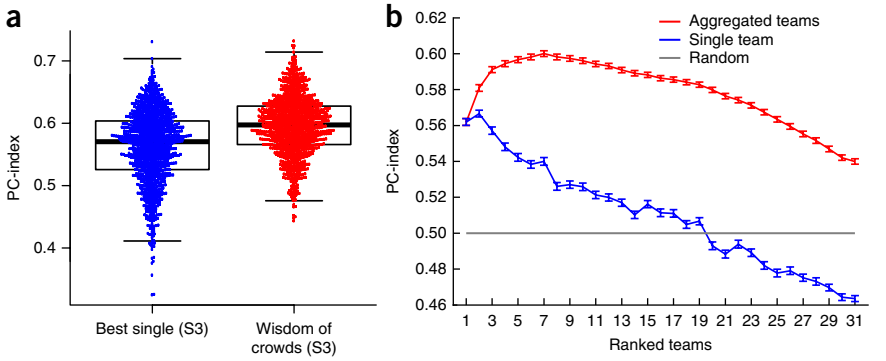

Figure 4 Community predictions. (**a**) Bee Swarm plot showing the performance of ensemble models and single best model, inferred from over 1,000 different three-set splits (S1/S2/S3) of the 91 drug pairs in the challenge. The first set S1 was used to determine an order of performance. The second set S2 was used to choose the optimal number of top methods to aggregate to attain best performance of the aggregate. Finally, the last set S3, which was not used to choose the order of aggregation or the optimal number of predictions to aggregate, was used to determine the performance of the best method (according to set S1) and of the "wisdom of crowds" aggregate. The latter is consistently better than the former. (**b**) Average and standard error over the 1,000 splits shown in **a** of the PC-index as computed in set S3 of individual teams (blue) and aggregates of the top-performing teams (red). The order of the teams in the *x* axis was determined according to set S1, but the performance was evaluated in set S3. The gray solid line represents random performance. Error bar represents the s.e.m. of the PC-index.

Such an increase in performance when integrating disparate methods has been called the "wisdom of crowds"[24].

Methods' advantages and limitations

Because multimetric evaluation provides a broader assessment of a method's bias and value[27], we used two additional metrics, sensitivity versus specificity (ROC) analysis and precision/sensitivity analysis, for performance assessment. The first is a threshold-free metric designed to assess a method's tradeoff between sensitivity and specificity in predicting synergistic or antagonistic combinations, whereas the second tests how precise in predicting the intended pairs the methods are at a specific cutoff. If we choose the cutoff to be the number of pairs of interest (that is, the number of synergistic and antagonistic pairs when studying synergy and antagonism, respectively), then precision coincides with sensitivity.

We first defined a criterion and identified 16 synergistic and 36 antagonistic compound pairs (**Fig. 5a**). We then evaluated the sensitivity versus specificity tradeoff using the area under the receiver operating characteristic (ROC) curve (AUC)[28] independently for synergistic and antagonistic compound pairs (**Supplementary Fig. 9**), resulting in distinct rankings. Based on AUC ranking, DIGRE was the best algorithm for antagonistic pair prediction and the fourth for synergistic ones. Conversely, the second team was the best performer in predicting synergistic pairs but did not perform well on antagonistic ones (**Fig. 5b**). Using the ROCs, we could also compute the statistical significance of the difference in performance of any two methods using the Hanley-McNeil method. This was done separately for synergistic and antagonistic compound pairs. We considered method A to outperform method B if its ROC-based performance was statistically significantly better, either in predicting antagonism or synergy ($P \leq 0.05$). This analysis revealed no statistically significant differences in the direct performance comparison of the top methods (**Supplementary Fig. 10**) but confirmed that the top three methods were statistically significantly better than the others.

We performed precision/sensitivity analysis by selecting the top 16 and the bottom 36 predictions from each method (**Online Methods**). Using this metric, the fifth overall best method and DIGRE were the best at predicting synergistic and antagonistic combinations, respectively (**Fig. 5c**). When a similar analysis was performed to test for the misclassification rate (i.e., synergistic pairs predicted as antagonistic and vice versa), we found that DIGRE's misclassification rate was very low, despite their weak performance in predicting synergistic pairs (**Supplementary Fig. 11**); that is, although the algorithm was not effective at identifying synergistic pairs, it virtually never misclassified a synergistic pair as antagonistic and vice versa (they were misclassified as additive). Across these metrics, the methods' hypotheses had a trending effect on predicting synergy (methods using a similarity hypothesis trended to have better sensitivity, **Supplementary Fig. 12a**) based on precision/sensitivity analysis but hardly any effect on predicting antagonism (**Supplementary Fig. 12b**), suggesting that hypotheses needed to correctly predict synergy and antagonism may be different. We need more extensive studies to confirm if such a trend generalizes.

As we did for the PC-index, we also evaluated the performance of aggregating predictions by various methods using these metrics to test "wisdom of crowds." Due to the limited number of synergistic (16) and antagonistic (36) compound pairs, we could not divide the gold standard data set into three subsets (S1, S2 and S3) to train and evaluate the performance of aggregating predictions by various methods on these metrics. Therefore, we used the training outcome based on the PC-index used in the previous section and averaged the performance of the optimized number of top-performing methods according to S2 across 1,000 partitions to determine the average precision and AUC metrics to test "wisdom of crowds." Similar to the PC-index, average integrated performance of the top seven methods (when evaluated in S3) showed 14% and 7% improvement in AUC for predicting synergistic and antagonistic compound pairs, respectively, compared to only single-best method's performance (ordered according to set S1). Results showed that "wisdom of crowds" results in high and consistent performance across all metrics, further supporting the notion of integrative strategies in scientific research. Indeed, no individual method outperformed the others across all metrics, suggesting that multiple hypotheses may need to be combined to globally address context-dependent compound synergy and antagonism. In particular, although several methods (**Fig. 5c**) were clearly statistically significant in predicting compound synergy, overall sensitivity was relatively modest (the highest being 37.5% ($P \leq 0.02$), compared to 17.6% by random selection, **Fig. 5c**). Performance using "wisdom of crowds" did especially well, achieving greater than 46% sensitivity for synergy and 51% for antagonism, suggesting that methods for *in silico* assessment of compound synergy are starting to achieve predictive value.

The SynGen algorithm

The experimental data set was originally intended to validate SynGen, a method explicitly designed to predict synergy and not antagonism. Following on results from several publications[6,29–33], SynGen assumes that the activity of the Master Regulators (MRs) of a specific cellular

© 2014 Nature America, Inc. All rights reserved.

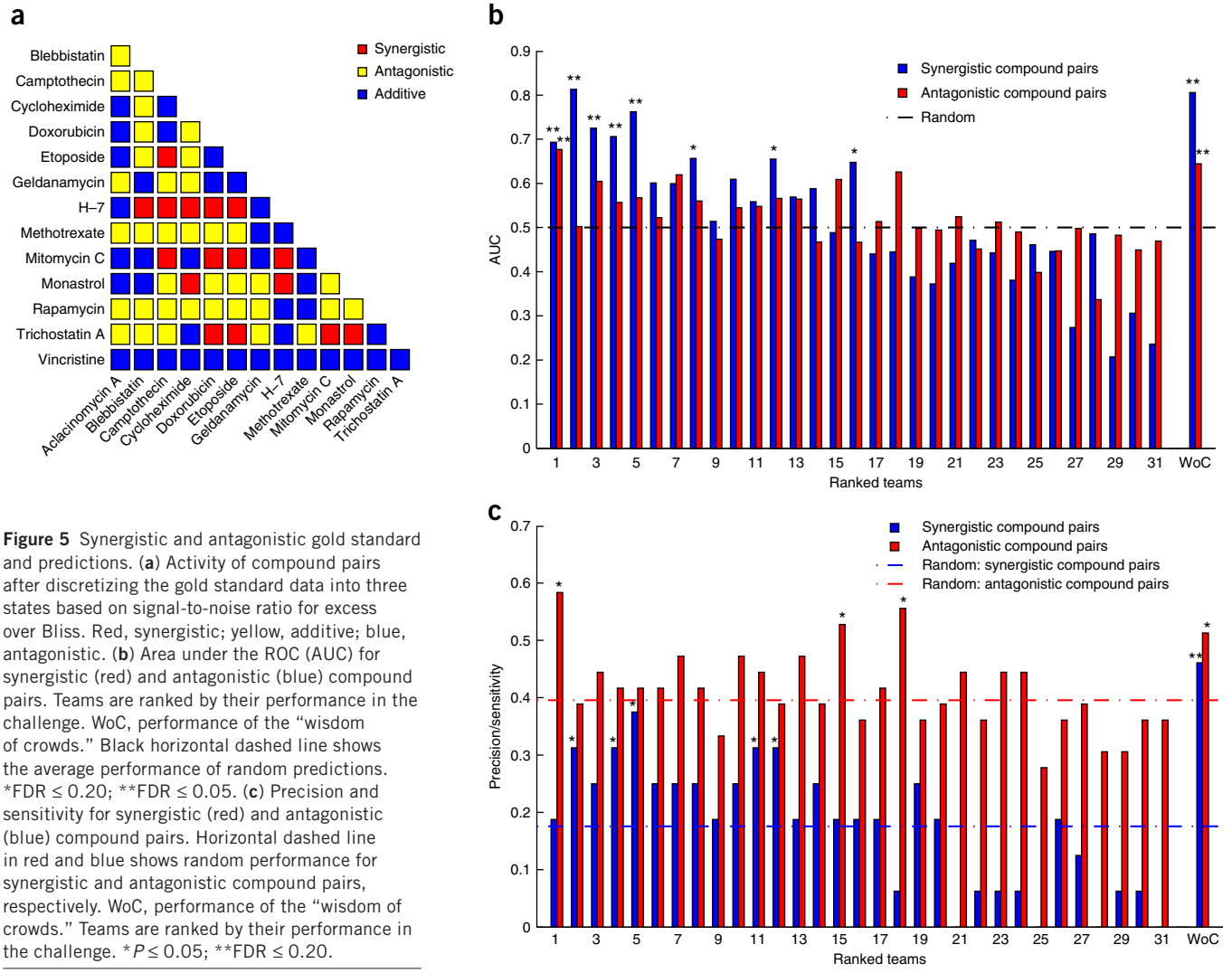

Figure 5 Synergistic and antagonistic gold standard and predictions. (**a**) Activity of compound pairs after discretizing the gold standard data into three states based on signal-to-noise ratio for excess over Bliss. Red, synergistic; yellow, additive; blue, antagonistic. (**b**) Area under the ROC (AUC) for synergistic (red) and antagonistic (blue) compound pairs. Teams are ranked by their performance in the challenge. WoC, performance of the "wisdom of crowds." Black horizontal dashed line shows the average performance of random predictions. *FDR ≤ 0.20; **FDR ≤ 0.05. (**c**) Precision and sensitivity for synergistic (red) and antagonistic (blue) compound pairs. Horizontal dashed line in red and blue shows random performance for synergistic and antagonistic compound pairs, respectively. WoC, performance of the "wisdom of crowds." Teams are ranked by their performance in the challenge. *$P \leq 0.05$; **FDR ≤ 0.20.

© 2014 Nature America, Inc. All rights reserved.

phenotype, as inferred by the Master Regulator Inference algorithm MARINa[29,30], is essential for cell viability (akin to oncogene addiction[34]). MRs are defined as regulators that are causally necessary and sufficient for the maintenance of a phenotype-specific gene expression signature. Thus, perturbations that either (i) abrogate the activity pattern of cell state MRs or (ii) activate MRs of cell death phenotypes, as also inferred by MARINa, may induce loss of cell viability. Based on this hypothesis, SynGen first infers relevant MR patterns for OCI-LY3 cell death and cell state and then identifies compounds that are most complementary in inducing the former and abrogating the latter (**Supplementary Note 2**). Two signatures used for MR inference were (i) a 'cell death' signature based on GEP following perturbation by the 14 compounds at 24 h, which induce appreciable toxicity levels (IC_{20}); and (ii) a 'cell addiction' signature, associated with the activated B-cell subtype of DLBCL cells (which include OCI-LY3) versus germinal center B-cell subtype, as we have shown that MRs of tumor subtype elicit addiction[30]. The latter signature was computed using publically available GEPs[35] for germinal center B-cell subtype cell lines (OCI-LY1, OCI-LY7, OCI-LY8, OCI-LY18 and SUDHL5) and for the activated B-cell subtype line OCI-LY3. SynGen then predicted synergistic compound combinations by selecting the compound pairs that are most complementary in implementing or abrogating these MR patterns, respectively. SynGen predicted synergistic compound pairs with high sensitivity (56%, $P \leq 0.001$). However, its ability to predict the full compound-pair ranking was not statistically significant, as the algorithm was not designed to predict compound antagonism. Source code for the SynGen algorithm can be found in **Supplementary Software 4**.

Compound- and cell-dependent bias

To analyze whether specific compound categories are more likely to elicit synergy or antagonism, and whether successful predictions were biased toward specific compounds, we ranked all compounds using the area under recall curve, AURC, for their specific combinations (**Supplementary Fig. 13**). High AURCs indicate compound proclivity toward synergy, whereas low AURCs indicate antagonism. Analysis of gold standard data suggests that pleiotropic compounds, exhibiting significant polypharmacology, such as H-7 and mitomycin C, were enriched in synergistic pairs. Conversely, compounds with more targeted mechanisms, such as rapamycin and blebbistatin, were least synergistic.

Finally, to determine whether synergy or antagonism is a universal property of the compound pairs or is context specific, we performed additional experiments to assess synergistic activity for 142 compound pairs in MCF7 breast cancer cells and LNCaP prostate cancer cells and compared them (**Supplementary Table 3**).

The analysis revealed no significant correlation between compound pairs ranked from the most synergistic to the most antagonistic ($\rho = -0.06$, $P = 0.45$, **Supplementary Fig. 14**). This shows that synergy and antagonism are highly context specific and are thus not universal properties of the compounds' chemical, structural or substrate information. As a result, predictive methods that account for the genetics and regulatory architecture of the context will become increasingly relevant to generalize results across multiple contexts.

DISCUSSION

This challenge provides a systematic and comparative evaluation of compound synergy and antagonism prediction methods based on blind experimental data. There are at least four reasons supporting the value and significance of this effort. First, although there are no previous experimentally validated efforts to predict synergy or antagonism of arbitrary compound pairs from single-compound perturbation data, our analysis shows that several laboratories have developed methodologies whose predictive ability is significantly better than random. Second, synergy and antagonism emerge as strongly context-dependent compound-pair properties. Thus, the value of synergy prediction methods is even more relevant, as experimental high-throughput synergy screen results cannot be generalized from one cellular context to others. Third, despite a complete lack of publications and established methodologies in this area, 31 teams from more than 13 countries participated in the challenge, thus effectively creating major interest in this field that over the long run is likely to further enhance our abilities to predict compound synergy and antagonism. Fourth, we established rigorous evaluation metrics for the assessment of synergy and antagonism prediction methods, thus allowing identification of three individual methods whose predictions significantly outperformed random guessing.

Although it is premature to claim that these advances will have an immediate and dramatic impact on the design of high-throughput screening assays for compound synergy assessment, the top-performing methods identified by this challenge already provide substantial potential reductions of the search space, suggesting that further improvements may increase the practical value of these techniques. For instance, the best-performing synergy-prediction method would have allowed screening only half of the compound combinations without missing any synergistic pair (**Supplementary Fig. 15**). Furthermore, many large-scale data sets representing individual compound perturbations are being generated and put in the public domain, such as those generated by the Library of Integrated Network-based Cellular Signatures (LINCS), which produced over 300,000 gene expression profiles following single-compound perturbations across multiple cell lines. It is reasonable to expect that availability of these data sets will lead to additional advances in the predictive power of these methods.

Introduction of additional, more specific metrics suggests that different methods did not score consistently across all of them, and that none of the methods is effective in predicting both synergy and antagonism. This suggests that the specific hypotheses used to predict synergy may not necessarily apply to antagonism prediction, and vice versa. This further suggests a valuable path for approaches that integrate different hypotheses for synergy, additivity and antagonism.

Even though the SynGen method, for which the data were originally generated, was highly effective in predicting compound synergy with higher sensitivity than other methods, its validation followed the more common procedure of prediction followed by evaluation against experimental data. However, despite the fact that SynGen is not based on machine learning methods that may be trained from experimental data, one cannot absolutely rule out potential overfitting. As such, direct comparison of SynGen's performance to the community-submitted algorithms is not appropriate and was deliberately avoided in this manuscript.

Our analysis also suggested that compounds exhibiting significant polypharmacology were enriched in synergistic pairs, whereas compounds with targeted mechanisms were more likely antagonistic. This may be due to the increase in the probability of modulating specific synergistic genetic dependencies in the cell, when using polypharmacology compounds[36,37]. Thus, these experimental assays provide an initial basis to guide future development of rational methodologies for the study of synergistic compound combinations in ABC-DLBCL lymphomas, providing further insight about relevant pathways that may be exploited in synergy experiments.

Despite these advances, there is ample room for both algorithm and evaluation metric improvements. For instance, none of the methods achieved near-optimal predictive power. Indeed, even though this challenge shows that current methodologies can perform significantly better than chance, there is still a large gap between ground truth (PC-index = 0.90) and the best prediction algorithms (PC index = 0.61). Methodological improvements are thus still required and could be achieved by several approaches, including (i) testing additional or more complex hypotheses about the mechanistic basis for compound synergy; (ii) generating larger perturbational profile data sets, for instance, using more concentrations and time points, to assess both early and late response to compound perturbation; (iii) exploring methodologies that better exploit the time-dependent nature of perturbational profiles; (iv) measuring complementary, context-specific molecular profiles, such as proteomic and epigenomic landscapes, to perform cross-data modality integrative analyses; (v) further integrating different methods within a unified framework; and (vi) addressing synergy, additivity and antagonism using distinct conceptual frameworks and hypotheses.

Compound synergy and antagonism were assessed only at the IC_{20} concentration of individual compound, using the excess over Bliss additivity. In future challenges, however, synergy may need to be tested over a wider range of concentrations and using additional methodologies (e.g., isobolograms). Results from gold standard data and predictions from top teams suggests that while designing new synergy experiments, it is important to make a larger selection of mechanistically diverse small molecules (targeted and pleiotropic) to compensate for the small number of potentially synergistic pathways.

Our findings suggest that DREAM challenges can provide a valuable mechanism to accelerate the development of predictive models for combination therapy, by providing an objective platform for the identification of model strengths and limitations through unbiased evaluations of model performance.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. Data provided to all participants in this challenge can be downloaded from http://www.the-dream-project.org/challenges/nci-dream-drug-sensitivity-prediction-challenge. Raw CEL files for gene expression profiles are in GEO: GSE51068.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

© 2014 Nature America, Inc. All rights reserved.

ACKNOWLEDGMENTS

The data for this challenge were kindly provided prepublication by A. Califano's laboratory. The general idea of the challenge was conceived in a workshop co-organized by National Cancer Institute (NCI) and Dialogue for Reverse Engineering Assessments and Methods (DREAM) on April 23, 2012. We acknowledge the contributions of all the participants in that summit, whose list of attendees can be accessed by the following link http://tinyurl.com/n3zp443. We would like to thank L. Pasqualucci and A. Holmes for processing the genetic profile of the OCI-LY3 cell line; W.K. Lim for normalizing GEPs; S. Dhindaw for proofreading the manuscript; F.M. Giorgi for helping with the figures of the manuscript; H. Li for tissue culture; P. Guarnieri, A. Ambesi, S. Anand, P. Subramaniam and S. Irshad for their valuable comments and feedback during the preparation of this manuscript. GeneTitan hybridizations were run at Rutgers University facility (Bionomics Research and Technology Center - BRTC). This work is supported in part by the Multiscale Analysis of Genomic and Cellular Networks (MAGNet) grant (5U54CA121852-08) and Library of Integrated Network-based Cellular Signatures Program (LINCS) grants (1U01CA164184-02 and 3U01HL111566-02) to A.C.; National Institutes of Health (NIH) grant (5R01CA152301) and Cancer Prevention and Research Institute of Texas (CPRIT) grant (RP101251) to Y.X. and NIH, NCI grant (U54 CA112970) to J.W.G.

AUTHOR CONTRIBUTIONS

M.B., M.P.M., J.C.C., L.M.H., J.S.-R., D.G., D.S., A.C. and G.S. designed and organized the challenge. The top-performing method was designed by J.Y., H.T., G.X., Y.L., J.A., R.Z., B.C., M.K. and T.W., and Y.X. C.K., R.R., M.M. and A.C. generated data for the challenge. M.B., J.C.C. and G.S. evaluated the predictions of the challenge. M.B., J.C.C., G.S. and A.C. interpreted the results of the challenge and performed the follow-up analysis. SynGen method was developed by M.J.A., Y.S., A.C., M.B., G.S., A.C. and Y.X. wrote the paper. The NCI-DREAM community provided predictions and **Supplementary Note 2** description. The corresponding author for the second-best-performing team is L. Li (lali@iu.edu).

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

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

1. Rawlinson, R. & Massey, A.J. Gamma H2AX and Chk1 phosphorylation as predictive pharmacodynamic biomarkers of Chk1 inhibitor-chemotherapy combination treatments. *BMC Cancer* **14**, 483 (2014).
2. Ibrahim, Y.H. *et al.* PI3K inhibition impairs BRCA1/2 expression and sensitizes BRCA-proficient triple-negative breast cancer to PARP inhibition. *Cancer Discov.* **2**, 1036–1047 (2012).
3. Yonesaka, K. *et al.* Activation of ERBB2 signaling causes resistance to the EGFR-directed therapeutic antibody cetuximab. *Sci. Transl. Med.* **3**, 99ra86 (2011).
4. Al-Lazikani, B., Banerji, U. & Workman, P. Combinatorial drug therapy for cancer in the post-genomic era. *Nat. Biotechnol.* **30**, 679–692 (2012).
5. Keith, C.T., Borisy, A.A. & Stockwell, B.R. Multicomponent therapeutics for networked systems. *Nat. Rev. Drug Discov.* **4**, 71–78 (2005).
6. Piovan, E. *et al.* Direct reversal of glucocorticoid resistance by AKT inhibition in acute lymphoblastic leukemia. *Cancer Cell* **24**, 766–776 (2013).
7. Vermorken, J.B. *et al.* Platinum-based chemotherapy plus cetuximab in head and neck cancer. *N. Engl. J. Med.* **359**, 1116–1127 (2008).
8. Bokemeyer, C. *et al.* Fluorouracil, leucovorin, and oxaliplatin with and without cetuximab in the first-line treatment of metastatic colorectal cancer. *J. Clin. Oncol.* **27**, 663–671 (2009).
9. Nelson, H.S. Advair: combination treatment with fluticasone propionate/salmeterol in the treatment of asthma. *J. Allergy Clin. Immunol.* **107**, 397–416 (2001).
10. Fernandes, D.J. & Bertino, J.R. 5-Fluorouracil-methotrexate synergy—enhancement of 5-fluorodeoxyuridylate binding to thymidylate synthase by dihydropteroylpolyglutamates. *Proc. Natl. Acad. Sci. USA* **77**, 5663–5667 (1980).
11. Stein, G.E. & Gurwith, M.J. Amoxicillin-potassium clavulanate, a beta-lactamase-resistant antibiotic combination. *Clin. Pharm.* **3**, 591–599 (1984).
12. Zhao, S. *et al.* Systems pharmacology of adverse event mitigation by drug combinations. *Sci. Transl. Med.* **5**, 206ra140 (2013).
13. Cokol, M. *et al.* Systematic exploration of synergistic drug pairs. *Mol. Syst. Biol.* **7**, 544 (2011).
14. Puri, N. & Salgia, R. Synergism of EGFR and c-Met pathways, cross-talk and inhibition, in non-small cell lung cancer. *J. Carcinog.* **7**, 9 (2008).
15. Fitzgerald, J.B., Schoeberl, B., Nielsen, U.B. & Sorger, P.K. Systems biology and combination therapy in the quest for clinical efficacy. *Nat. Chem. Biol.* **2**, 458–466 (2006).
16. Zhao, X.M. *et al.* Prediction of drug combinations by integrating molecular and pharmacological data. *PLoS Comput. Biol.* **7**, e1002323 (2011).
17. Jin, G.X., Zhao, H., Zhou, X.B. & Wong, S.T.C. An enhanced Petri-net model to predict synergistic effects of pairwise drug combinations from gene microarray data. *Bioinformatics* **27**, i310–i316 (2011).
18. Green, M.R. *et al.* Integrative analysis reveals selective 9p24.1 amplification, increased PD-1 ligand expression, and further induction via JAK2 in nodular sclerosing Hodgkin lymphoma and primary mediastinal large B-cell lymphoma. *Blood* **116**, 3268–3277 (2010).
19. Harrell, F.E. Jr., Lee, K.L. & Mark, D.B. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat. Med.* **15**, 361–387 (1996).
20. Shah, M.A. & Schwartz, G.K. Cell cycle-mediated drug resistance an emerging concept in cancer therapy. *Clin. Cancer Res.* **7**, 2168–2181 (2001).
21. Recht, A. *et al.* The sequencing of chemotherapy and radiation therapy after conservative surgery for early-stage breast cancer. *N. Engl. J. Med.* **334**, 1356–1361 (1996).
22. Schlessinger, A., Punta, M., Yachdav, G., Kajan, L. & Rost, B. Improved disorder prediction by combination of orthogonal approaches. *PLoS ONE* **4**, e4433 (2009).
23. Helguera, A.M. *et al.* Combining QSAR classification models for predictive modeling of human monoamine oxidase inhibitors. *Eur. J. Med. Chem.* **59**, 75–90 (2013).
24. Marbach, D. *et al.* Wisdom of crowds for robust gene network inference. *Nat. Methods* **9**, 796–804 (2012).
25. Margolin, A.A. *et al.* Systematic analysis of challenge-driven improvements in molecular prognostic models for breast cancer. *Sci. Transl. Med.* **5**, 181re181 (2013).
26. Alexe, G. *et al.* A robust meta-classification strategy for cancer diagnosis from gene expression data. *Proc. IEEE Comput. Syst. Bioinform. Conf.* **2005**, 322–325 (2005).
27. Norel, R., Rice, J.J. & Stolovitzky, G. The self-assessment trap: can we all be better than average? *Mol. Syst. Biol.* **7**, 537 (2011).
28. Prill, R.J. *et al.* Towards a rigorous assessment of systems biology models: the DREAM3 challenges. *PLoS ONE* **5**, e9202 (2010).
29. Lefebvre, C. *et al.* A human B-cell interactome identifies MYB and FOXM1 as master regulators of proliferation in germinal centers. *Mol. Syst. Biol.* **6**, 377 (2010).
30. Carro, M.S. *et al.* The transcriptional network for mesenchymal transformation of brain tumours. *Nature* **463**, 318–325 (2010).
31. Aytes, A. *et al.* Cross-species regulatory network analysis identifies a synergistic interaction between FOXM1 and CENPF that drives prostate cancer malignancy. *Cancer Cell* **25**, 638–651 (2014).
32. Chen, J.C. *et al.* Regulatory network based analysis of genetic alterations reveals deletion of KLHL9 E3 ligase complex adapter protein as a driver of mesenchymal signature in glioblastoma. *Cell* (in the press).
33. Chudnovsky, Y. *et al.* ZFHX4 interacts with the NuRD core member CHD4 and regulates the glioblastoma tumor-initiating cell state. *Cell Reports* **6**, 313–324 (2014).
34. Weinstein, I.B. Cancer. Addiction to oncogenes–the Achilles heal of cancer. *Science* **297**, 63–64 (2002).
35. Zhang, J. *et al.* Genetic heterogeneity of diffuse large B-cell lymphoma. *Proc. Natl. Acad. Sci. USA* **110**, 1398–1403 (2013).
36. Araujo, R.P., Petricoin, E.F. & Liotta, L.A. A mathematical model of combination therapy using the EGFR signaling network. *Biosystems* **80**, 57–69 (2005).
37. Lehár, J. *et al.* Synergistic drug combinations tend to improve therapeutically relevant selectivity. *Nat. Biotechnol.* **27**, 659–666 (2009).
38. Gottlieb, A., Stein, G.Y., Ruppin, E. & Sharan, R. PREDICT: a method for inferring novel drug indications with application to personalized medicine. *Mol. Syst. Biol.* **7**, 496 (2011).
39. Jia, J. *et al.* Mechanisms of drug combinations: interaction and network perspectives. *Nat. Rev. Drug Discov.* **8**, 111–128 (2009).
40. Pedregosa, F. *et al.* Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
41. Guyon, I., Weston, J., Barnhill, S. & Vapnik, V. Gene selection for cancer classification using support vector machines. *Mach. Learn.* **46**, 389–422 (2002).
42. Di Camillo, B. *et al.* Function-based discovery of significant transcriptional temporal patterns in insulin stimulated muscle cells. *PLoS ONE* **7**, e32391 (2012).
43. van Westen, G.J.P. & Overington, J.P. A ligand's-eye view of protein similarity. *Nat. Methods* **10**, 116–117 (2013).
44. Rajendran, P. *et al.* Suppression of signal transducer and activator of transcription 3 activation by butein inhibits growth of human hepatocellular carcinoma in vivo. *Clin. Cancer Res.* **17**, 1425–1439 (2011).
45. Vilar, S. *et al.* Drug-drug interaction through molecular structure similarity analysis. *J. Am. Med. Inform. Assoc.* **19**, 1066–1074 (2012).
46. Hoshida, Y. Nearest template prediction: a single-sample-based flexible class prediction with confidence assessment. *PLoS ONE* **5**, e15543 (2010).

© 2014 Nature America, Inc. All rights reserved.

NCI-DREAM Community

Jean-Paul Abbuehl[18], Jeffrey Allen[3], Russ B Altman[19], Shawn Balcome[20], Mukesh Bansal[1,2,62], Ana Bell[21,22], Andreas Bender[23], Bonnie Berger[24], Jonathan Bernard[18], Andrew A Bieberich[25], Giorgos Borboudakis[26,27], Andrea Califano[1,2,14–17], Christina Chan[28–30], Beibei Chen[3], Ting-Huei Chen[31], Jaejoon Choi[32], Luis Pedro Coelho[33], James C Costello[6,61], Chad J Creighton[34], Will Dampier[35], V Jo Davisson[25], Raamesh Deshpande[20], Lixia Diao[36], Barbara Di Camillo[37], Murat Dundar[38], Adam Ertel[39], Cellworks Group[40], Daniel Gallahan[12], Chirayu P Goswami[41], Assaf Gottlieb[19], Michael N Gould[42], Jonathan Goya[22], Michael Grau[43], Joe W Gray[9], Laura M Heiser[9], Hussein A Hejase[28], Michael F Hoffmann[42], Krisztian Homicsko[18], Max Homilius[21,22], Woochang Hwang[32], Adriaan P Ijzerman[44], Olli Kallioniemi[45], Bilge Karacali[46], Charles Karan[4,62], Samuel Kaski[47,48], Junho Kim[32], Minsoo Kim[3,8], Arjun Krishnan[22], Junehawk Lee[32,49], Young-Suk Lee[21,22], Eelke B Lenselink[44], Peter Lenz[43], Lang Li[41], Jun Li[36,50], Yajuan Li[7], Han Liang[36,51], Michela Mattioli[10], Michael P Menden[5], John-Patrick Mpindi[45], Chad L Myers[20], Michael A Newton[52], John P Overington[53], Juuso Parkkinen[47], Robert J Prill[54], Jian Peng[24], Richard Pestell[39], Peng Qiu[36,61], Bartek Rajwa[55], Ronald Realubit[4], Anguraj Sadanandam[18], Julio Saez-Rodriguez[5], Francesco Sambo[37], Dinah Singer[12], Gustavo Stolovitzky[13], Arvind Sridhar[56], Wei Sun[31,57], Hao Tang[3], Gianna M Toffolo[37], Aydin Tozeren[35], Olga G Troyanskaya[21,22], Ioannis Tsamardinos[26,27], Herman W T van Vlijmen[58], Tao Wang[3], Wen Wang[20], Joerg K Wegner[58], Krister Wennerberg[45], Gerard J P van Westen[53], Tian Xia[20], Guanghua Xiao[3], Yang Xie[3,8], Jichen Yang[3,62], Yang Yang[36,59], Victoria Yao[21,22], Yuan Yuan[36,51], Haoyang Zeng[24], Shihua Zhang[60], Junfei Zhao[60], Jian Zhou[22] & Rui Zhong[3]

[18]Swiss Institute for Experimental Cancer Research (ISREC), Swiss Federal Institute of Technology Lausanne (EPFL), Lausanne, Switzerland. [19]Departments of Genetics, Stanford University, Stanford, California, USA. [20]University of Minnesota, Minneapolis, Minnesota, USA. [21]Department of Computer Science, Princeton University, Princeton, New Jersey, USA. [22]Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, New Jersey, USA . [23]Unilever Centre, Cambridge University, Cambridge, UK. [24]Computer Science and Artificial Intelligence Laboratory, MIT, Cambridge, Massachusetts, USA. [25]Department of Medicinal Chemistry and Molecular Pharmacology, Purdue University, W. Lafayette, Indiana, USA. [26]Computer Science Department, University of Crete, Crete, Greece. [27]Institute of Computer Science, FORTH, Crete, Greece. [28]Department of Computer Science and Engineering, Michigan State University, East Lansing, Michigan, USA. [29]Department of Chemical Engineering and Materials Science, Michigan State University, East Lansing, Michigan, USA. [30]Department of Biochemistry and Molecular Biology, Michigan State University, East Lansing, Michigan, USA. [31]Department of Biostatistics, University of North Carolina, Chapel Hill, North Carolina, USA. [32]Korea Advanced Institute of Science and Technology, Daejeon, Korea. [33]Instituto de Medicina Molecular, Faculdade de Medicina da Universidade de Lisboa, Lisbon, Portugal. [34]Department of Medicine, Dan L. Duncan Center Division of Biostatistics, Baylor College of Medicine, Houston, Texas, USA. [35]Center for Integrated Bioinformatics, Drexel University, Philadelphia, Pennsylvania, USA. [36]Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, Houston, Texas, USA. [37]Department of Information Engineering, University of Padova, Padova, Italy. [38]Department of Computer and Information Science, IUPUI, Indianapolis, Indiana, USA. [39]Jefferson Kimmel Cancer Center, Philadelphia, Pennsylvania, USA. [40]Cellworks Group Inc., San Jose, California, USA. [41]Center for Computational Biology and Bioinformatics, IU School of Medicine, Indianapolis, Indiana, USA. [42]Department of Oncology and Carbone Cancer Center, University of Wisconsin, Madison, Wisconsin, USA. [43]Department of Physics, University of Marburg, Marburg, Germany. [44]Leiden Academic Center for Drug Research, University of Leiden, Leiden, the Netherlands. [45]Institute for Molecular Medicine Finland FIMM, University of Helsinki, Helsinki, Finland. [46]Izmir Institute of Technology, Izmir, Turkey. [47]Helsinki Institute for Information Technology HIIT, Department of Information and Computer Science, Aalto University, Espoo, Finland. [48]Institute for Information Technology HIIT, Department of Computer Science, University of Helsinki, Helsinki, Finland. [49]Korea Institute of Science and Technology Information, Daejeon, Korea. [50]CAS-MPG Partner Institute for Computational Biology, Key Laboratory of Computational Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai, P. R. China. [51]Graduate Program in Structural and Computational Biology and Molecular Biophysics, Baylor College of Medicine, Houston, Texas, USA. [52]Departments of Statistics and of Biostatistics and Medical Informatics, University of Wisconsin, Madison, Wisconsin, USA. [53]European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge, UK. [54]IBM Almaden Research Center, San Jose, California, USA. [55]Bindley Bioscience Center, Purdue University, W. Lafayette, Indiana, USA. [56]Embedded Systems Laboratory (ESL), Institute of Electrical Engineering, Swiss Federal Institute of Technology Lausanne (EPFL), Laussane, Switzerland. [57]Department of Genetics, UNC Chapel Hill, Chapel Hill, North Carolina, USA. [58]Janssen Pharmaceutica, Beerse, Belgium. [59]Division of Biostatistics, School of Public Health, The University of Texas Health Science Center at Houston, Houston, Texas, USA. [60]National Center for Mathematics and Interdisciplinary Sciences, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, China. [61]Present addresses: Department of Pharmacology, School of Medicine, University of Colorado Anschutz Medical Campus, Aurora, Colorado, USA (J.C.C.) and Department of Biomedical Engineering, Georgia Institute of Technology and Emory University (P.Q.). [62]These authors contributed equally to this work. Correspondence should be addressed to G.S. (gustavo@us.ibm.com) or A.C. (califano@c2b2.columbia.edu) or Y.X. (Yang.Xie@UTSouthwestern.edu).

© 2014 Nature America, Inc. All rights reserved.

ONLINE METHODS

Cell culture and compound treatment. The OCI-LY3 diffuse large B-cell lymphoma (DLBCL) cell line was obtained from University Health Network (Toronto, Canada) and was cultured under standard conditions (37 °C in humidified atmosphere, with 5% CO_2) in IMDM supplemented with 10% FCS. Each compound was titrated in the OCI-LY3 cell line in a 20-point titration curve. Cell viability following compound treatment was determined using the CellTiter-Glo (Promega Corporation). An IC_{20} value for each compound was calculated by using Dose Response Fit and Calculate ECx components from the Pipeline Pilot Plate Data Analytics collection. For compounds in which more than 20% viability reduction could not be reached, a default concentration of 100 μM was used. For generation of GEPs, the OCI-LY3 cells were seeded in tissue culture–treated 96-well plates at a density of 50,000 cells per well (100 μl) and treated at the IC_{20} concentrations of each of the compounds at 24 h and 48 h. In the assay, three time points (6, 12 and 24 h) were analyzed for gene expression profiling. All profiles were generated in triplicate biological replicates except DMSO-treated samples which were hybridized in octuplicate as they were used as internal controls for each time point. To confirm viability data at each step, identical plates were produced and cell viability assessed using the CellTiter-Glo reagent (Promega Corporation).

Gene expression profiling. Total RNA was isolated with the Janus automated liquid handling system (PerkinElmer Inc.) using the RNAqueous-96 Automated Kit (Ambion), quantified by NanoDrop 6000 spectrophotometer and quality checked by Agilent Bioanalyzer. 300 ng of each of the samples with RIN value >7 was converted to biotinylated cRNA with the Illumina TotalPrep-96 RNA Amplification Kit (Ambion) using a standard T7-based amplification protocol and hybridized on the Human Genome U219 96-Array Plate (Affymetrix). Hybridization, washing, staining and scanning of the array plates were performed on the GeneTitan Instrument (Affymetrix) according to the manufacturer's protocols.

Experimental determination of synergy. For each compound, IC_{20} was determined assessed from 20-point titration curves (as described above) at 60 h following compound treatment by measuring cell viability and generating a dose-response curve. Each compound combination was then tested at the respective IC_{20} (or 100 μM) concentration of the individual compounds in five replicates. All compounds and combinations are diluted in DMSO, with a final DMSO concentration of 0.4%. Cells were placed at a density of 2,000 cells per well in 384-well plates, and compounds were added at 12-h intervals after seeding by compound transfers of serially diluted compounds. Assay plates were then incubated for 60 h followed by addition of 25 μl of CellTiter-Glo (Promega Corp.) at room temperature. Plates were read on the Envision (PerkinElmer Inc.) using enhanced luminescence protocol.

Data processing. All gene expression samples were quality controlled and normalized with the RMA normalization method using Bioconductor package in R. The baseline genetic profile of the OCI-LY3 cell line was obtained from reference[18] and was processed using the CBS algorithm, as published[47]. The final segmentation file was filtered for any germline aberrations between 1.74 and 2.3 and segments with less than eight markers. Segments with aberrations less than 1.74 or greater than 2.3 were assigned as deleted and amplified, respectively.

Excess over Bliss as a measurement for synergy. The Bliss additivism (or Bliss independence) model[48] predicts that if compound D_x and D_y, with experimentally measured fractional inhibitions f_x and f_y, have an additive effect, then the expected fractional inhibition, f_{xy}, induced by their combination should be:

$$f_{xy} = 1-(1-f_x)\times(1-f_y) = f_x + f_y - f_x \times f_y$$

Excess over Bliss is determined by computing the difference in fractional inhibition induced by compound combination, f_z, and the expected fractional inhibition, f_{xy}

$$eob = f_z - f_{xy}$$

A compound pair for which $eob \approx 0$ has an additive behavior, whereas a compound pair with positive (or negative) *eob* values has synergistic (or antagonistic) behavior. We used propagation of errors using s.e.m. of fractional inhibitions to compute the s.e.m. of *eob*.

Resampled spearman correlation. To assess that the ranking of participants is not biased by our scoring methods, we used another independent approach to score all participants. This method assumed that the experimental measurements of the mean excess over Bliss for a given compound pair is noisy, following a normal distribution, $N(\mu,\sigma)$, with mean, μ, equal to the mean EOB and s.d., σ, equal to the s.e.m. of excess over Bliss. For every compound pair, *i*, we randomly sample a possible measurement of the mean EOB_i from the distribution associated with that compound pair $N(\mu_i,\sigma_i)$, resulting in a new sampled observed score for all compound pairs $\{O_1^{rand}, O_2^{rand}, \ldots O_{91}^{rand}\}$. We compute a Spearman correlation between these new sampled EOB values and the predicted EOB ranks to generate $scorr^{rand}$. We repeat this step 10,000 times creating 10,000 different $scorr^{rand}$ and finally calculate an average over all 10,000 $scorr^{rand}$ to assign a final score, *scorr*, to each participant.

***P*-value estimation.** We assessed the statistical significance of scores generated by both probabilistic c-index and resampled Spearman correlation methods by assigning a *P*-value to each score. To compute a *P*-value we generated 10,000 random predictions and scored them independently using PC-index and *scorr* resulting in the generation of an empirical null distribution ($\text{PC-index}_{\text{null}}$ and $\text{scorr}_{\text{null}}$). We used this empirical null distribution to estimate *P*-values for each participant, which are calculated as the fraction of scores in the null distribution higher than the participant's score

$$P-\text{value}_{PC-\text{index}} = \frac{\#(\text{PC}-\text{index}_{\text{null}}) \geq \text{PC}-\text{index}}{10{,}000}$$

$$P-\text{value}_{\text{scorr}} = \frac{\#(\text{score}_{\text{null}}) \geq \text{scorr}}{10{,}000}$$

Leave-one-out test. To ensure that participant ranktings were robust, we calculated a score for each participant by systematically removing one compound and considering 13 compounds for scoring and assigning them new ranks. This resulted in 14 different tests for each participant, each after removing one of the 14 compounds. In the end each team was assigned 14 ranks based on its performance using the remaining 13 compounds.

AUC and precision/sensitivity analyses. We estimated the significance of predicting synergistic and antagonistic compound pairs using the area under the receiver operating characteristic curve (AUC), which was called the sensitivity versus specificity analysis in the main text. To compute the AUC for synergistic predictions, first we sort the predictions of each participant from the most to the least synergistic (predicted list). Second, from the gold standard, we define the compound pairs that are synergistic and antagonistic. To identify such compound pairs we computed the signal to noise ratio (*snr*) of each compound pair, defined as the ratio of the mean excess over Bliss (EOB) over the s.e.m. of EOB. We defined any compound pair as synergistic if its mean EOB was positive and its *snr* is greater than 2, which yielded 16 synergistic compound pairs. Similarly, a compound pair is defined to be antagonistic if its EOB is negative and its *snr* is greater than 2, yielding 36 antagonistic compound pairs. The rest of the pairs are considered to be additive. From the predicted list, we select the top *i* predictions and calculate the true positive rate (TPR_i) and false-positive rate (FPR_i). To estimate the TPR_i and FPR_i, we calculate the number of true positives (TP_i), defined as the number of correct synergistic pairs in the top *i* predictions, the number of false positives (FP_i), defined as number of false synergistic predictions in the top *i* predictions, the number of true negatives (TN_i), defined as the number of correct nonsynergistic compound pairs predicted below the top *i* predictions and the number of false negatives (FN_i), defined as the number of synergistic compound pairs predicted below the top *i* predictions.

Finally, TPR_i and FPR_i are calculated as

$$TPR_i = \frac{TP_i}{TP_i + FN_i};\ FPR_i = \frac{FP_i}{FP_i + TN_i}$$

© 2014 Nature America, Inc. All rights reserved.

We varied i from 1 to 91 and plotted the TPR_i (or sensitivity) versus FPR_i (or 1–specificity) to generate the receiver operating characteristic (ROC) curve. Finally, we calculated area under the ROC curve using a trapezoidal method to integrate the ROC curve. The AUC for antagonistic compound pairs is estimated by ranking predictions from the most to the least antagonistic and by selecting the true antagonistic compound pairs from the gold standard.

The precision/sensitivity analysis was performed as follows. After sorting the predictions of each participant from the most to the least synergistic, we compute the precision of synergistic predictions as the fraction of synergistic compound pairs in the top 16 predictions, that is

$$\text{Precision(synergy)} = \frac{TP_{16}}{16}$$

Similarly, precision for antagonistic compound pairs was calculated by sorting the predictions of each participant from the most to least antagonistic and computing the fraction of antagonistic compound pairs in the top 36 predictions. Sensitivity is defined as the number of TP (true positives) divided by the total number of positives, P (e.g., synergistic or antagonistic drug pairs). Because we selected the top P drug pairs to compute precision, our calculation of precision coincides with the evaluation of sensitivity.

Cross-validated ensemble models. To build ensemble models using predictions from different methods, we averaged the rank for each compound pair predicted by all models being aggregated, and re-ranked the compound pairs according to the average rank. To evaluate the merits of aggregation, we used a model-selection assessment approach. We randomly divided all compound pairs into three subsets of equal sizes and used the first group, S1, for sorting the models from the best to worst performance, the second group, S2, to estimate the number of models to combine to attain best performance, and, finally, the third group, S3, for an unbiased test of the individual or aggregate models. We repeated this process 1,000 times to evaluate the statistical significance of the differences between aggregate versus individual performance. More precisely, in the *ith* split (where i is varied from 1 to 1,000), we compute the PC-index for each participant using the compound pairs in $S1_i$, and based on their performance, create a team list $T1_i$ ordered from best to worst performing teams. Next we aggregate the k best methods in $T1_i$ and use subset $S2_i$ to compute the PC-index (using the compound pairs in S2i), $PC2_{ik}$, and vary k from 1 to 31. We identify k^* such that $PC2_{ik} \geq PC2_{ik}$ for all k, giving us the number of participants whose aggregate gives the maximum PC-index (in $S2_i$). Finally, using $S3_i$, we compute the PC-index, $PC3_{ik^*}$, to determine the performance of top k^* participants identified in the previous step but using the compound pairs in subset S3i. In this way, subsets S1 were used to determine which models to add cumulatively, subsets S2 were exclusively required for determining the optimal number of methods to achieve maximum performance and subsets S3 were solely used to estimate an unbiased performance of the aggregate that was eventually reported in **Figure 4**. For assessing a comparable single method performance, we chose the best performer determined on S1 and evaluated its performance on S3.

Area under the recall curve (AURC). For each compound, the area under the recall curve, AURC, is generated by first calculating the fraction of the 13 combinations that the chosen compound can participate in, contained in the top i compound pairs, ranked from the most synergistic to the most antagonistic pairs. We varied i from 1 to 91 and plotted that fraction versus i to generate the recall curve and finally calculated the area under the recall curve using the trapezoidal method. A high area under the recall curve is a predictor of the proclivity of a compound toward synergy, whereas a low area under the recall curve is a predictor toward antagonism.

Hanley-McNeil method. We estimated the statistical significance of the difference in performance by any two methods (i,j) by calculating the significance of the difference in the area under their ROC curve using the Hanley and McNeil method[49]. We calculated this significance separately for synergistic and antagonistic predictions. To estimate this significance for any method i, first we calculated the area under the ROC curve, A_i, using the trapezoidal method. Next we estimated the standard error, SE_i, likely to be associated in the estimation of A_i

$$SE_i = \sqrt{\frac{A_i(1-A_i) + [((n)]_p - 1)(Q_1 - A_i^2) + (n_n - 1)(Q_2 - A_i^2)}{n_p n_n}}$$

where

$$Q_1 = \frac{A_i}{2 - A_i};\ Q_2 = \frac{2A_i^2}{1 + A_i}$$

n_p = number of synergistic or antagonistic compound pairs and $n_n = 91 - n_p$.

Finally we used the normal cumulative distribution function

$$\frac{1}{2}\left[1 + \text{erf}\left(\frac{A_i - A_j}{\sqrt{SE_i^2 + SE_j^2}}\right)\right]$$

to estimate the P-value where 'erf' is the error function. Note that this procedure assumes independence between the predicted AUC, an assumption that can be violated if there are hidden biases in the ordering of the compound pairs.

Compound-pair activity in the MCF7 and LNCAP cell lines. We tested the pairwise combinations of 71 compounds with a proteasome inhibitor MG 132 and a HDAC inhibitor Trichostatin A (a total of 142 combinations) in the MCF7 breast cancer cell line and the LNCAP prostate cell line by using a cell viability assay. For each pair of compounds, we performed 16 experiments with four different dosages for each compound. To compute the synergy for each pair, we calculated the excess over Bliss score for each of the 16 experiments and took the average of 16 scores as the synergy score for the compound pair.

47. Olshen, A.B., Venkatraman, E.S., Lucito, R. & Wigler, M. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* **5**, 557–572 (2004).
48. Borisy, A.A. *et al.* Systematic discovery of multicomponent therapeutics. *Proc. Natl. Acad. Sci. USA* **100**, 7977–7982 (2003).
49. Hanley, J.A. & McNeil, B.J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **143**, 29–36 (1982).

© 2014 Nature America, Inc. All rights reserved.

doi:10.1038/nbt.3052