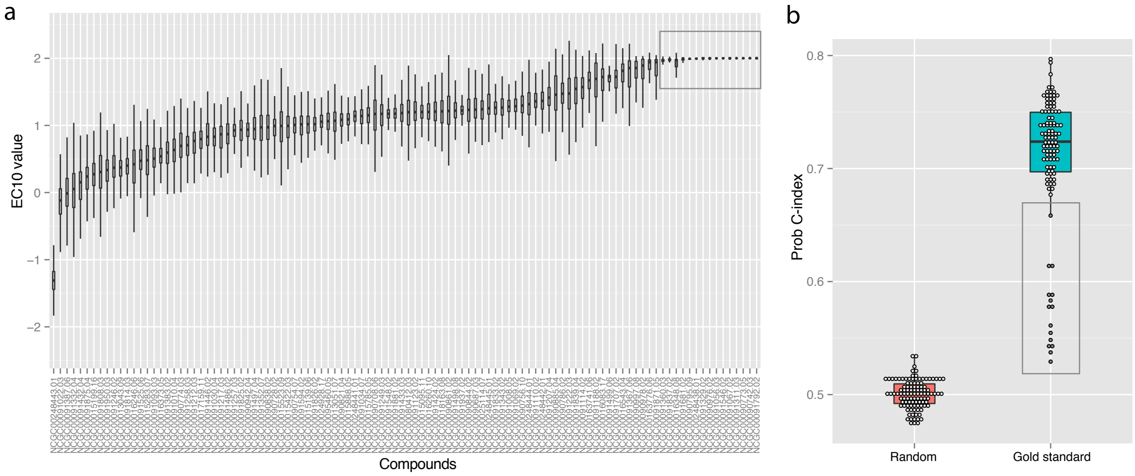
NIEHS-NCATS-UNC DREAM Toxicogenetics Challenge Manuscript Supplementary information

Please see the online supplement on synapse.org:
http://dx.doi.org/10.7303/syn1840307

1 - Supplementary Figures

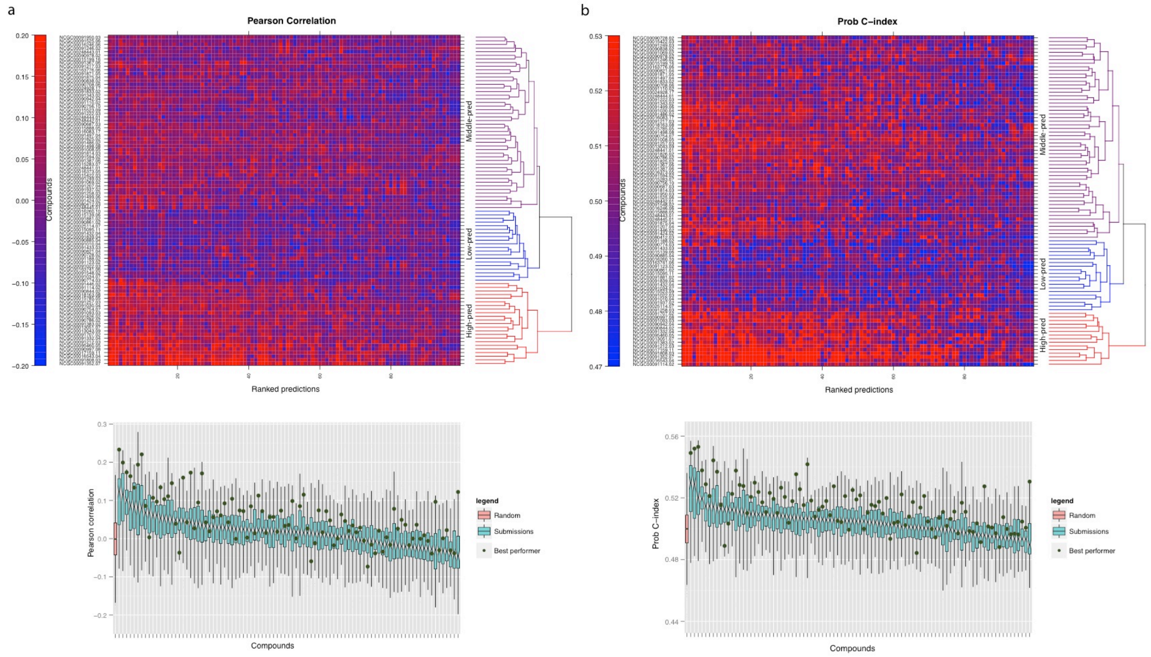
Supplementary Fig. 1. Compounds used for scoring



Out of the 106 predicted compounds, 15 were not used for scoring because they are non-cytotoxic and they are not predictable. (a) Boxplot of the EC10 values for each compound ordered from the more cytotoxic to the less cytotoxic, the 15 less cytotoxic compounds are marked with a grey box. (b) Probabilistic C-index computed for each compound for random predictions and for the gold standard. The 15 less toxic compounds (grey box) show very poor predictability.

See source code for the figure, files source_code/SuppFigure1.R

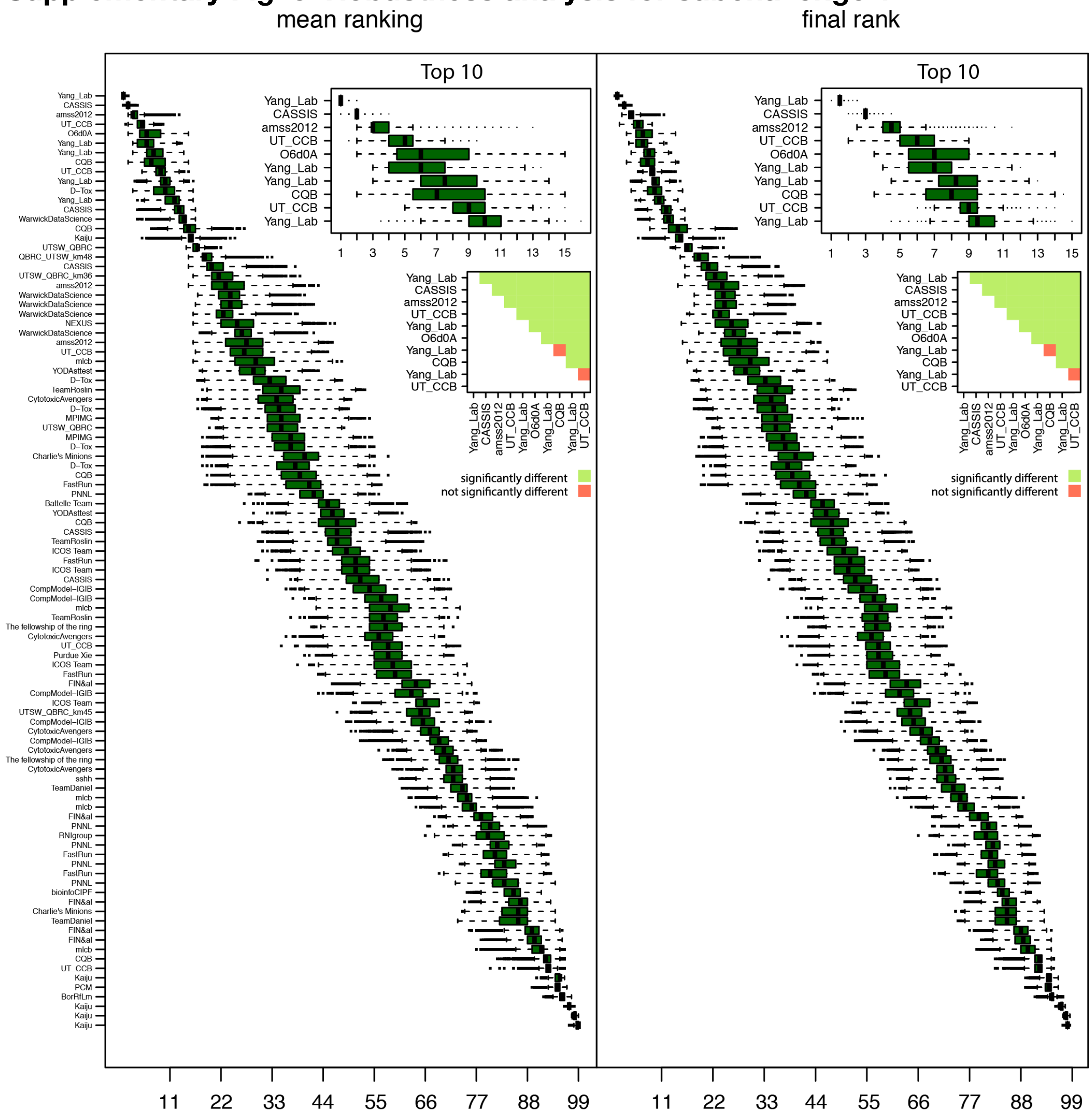
Supplementary Fig. 2. Performances of submissions for subchallenge 1



Predictions to subchallenge 1 were compared to the gold standard based on (a) Pearson Correlation and (b) probabilistic C-index. The heatmap in the top panel illustrates performances of all predictions for all compounds used for evaluation: predictions are ranked and compounds are clustered. Performance values are saturated at -0.2 and 0.2 for Pearson correlation and at 0.53 and 0.47 for probabilistic C-index. In the bottom panel boxplot of performances of predictions for each compound, are shown along with the null distribution.

See source code for the figure, files source_code/SuppFigure2.R

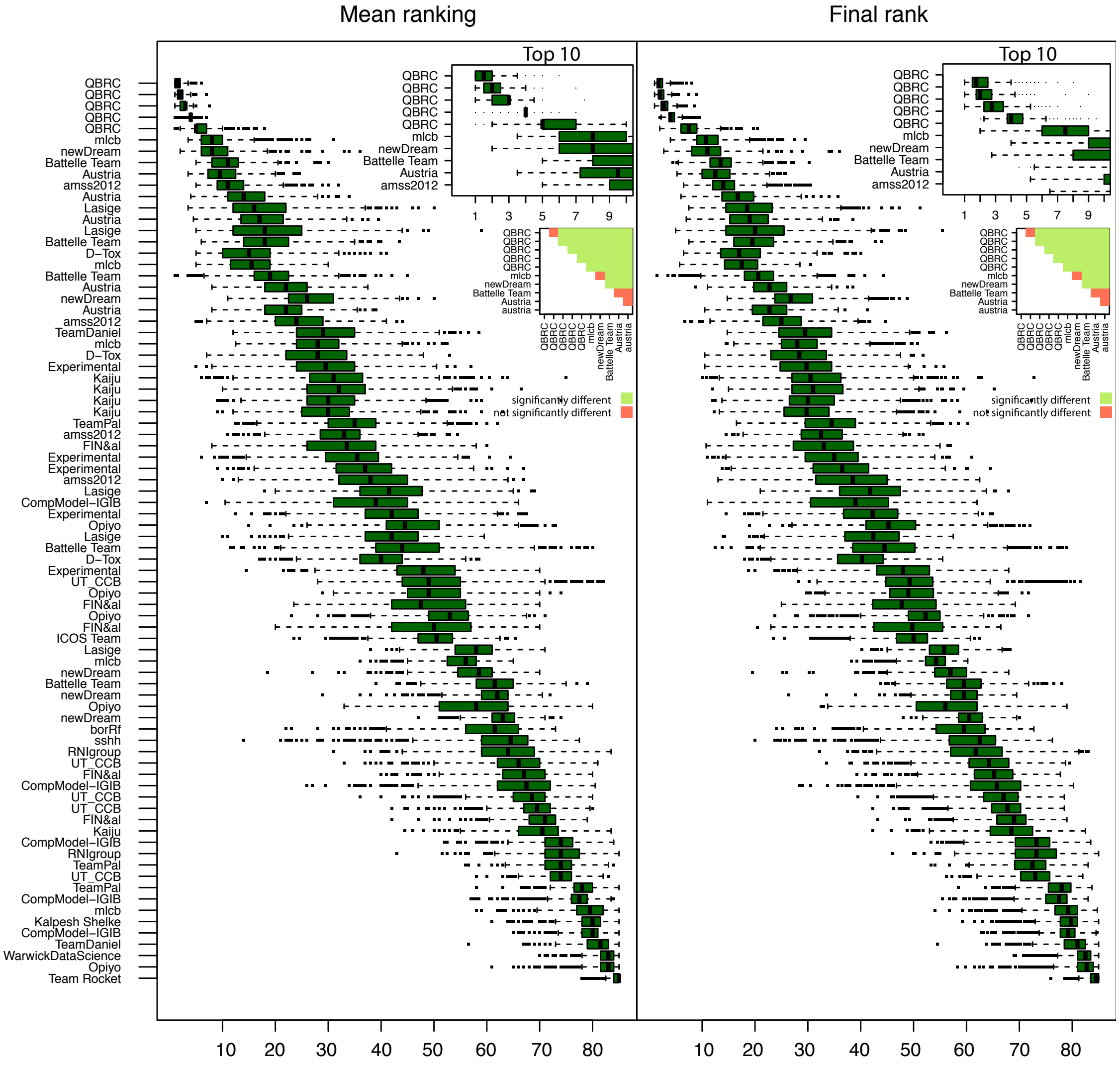
Supplementary Fig. 3. Robustness analysis for subchallenge 1



A robustness (sampling) analysis was performed for subchallenge 1 in order to assess the robustness of teams ranking with respect to compounds used for scoring. For 10000 iterations, both the mean ranking (left panel) and the final rank (right panel) were recomputed using each time only 80 randomly selected compounds (out of the 91) for scoring. A zoom of the 10 best performers is also shown, and distributions are compared using a one-sided Wilcoxon signed-rank test corrected for multiple hypothesis using the Benjamini-Hochberg correction. False discovery rates (FDR) are considered significant when $FDR<10^{-10}$.

See source code for the figure, files source_code/SuppFigure3.R

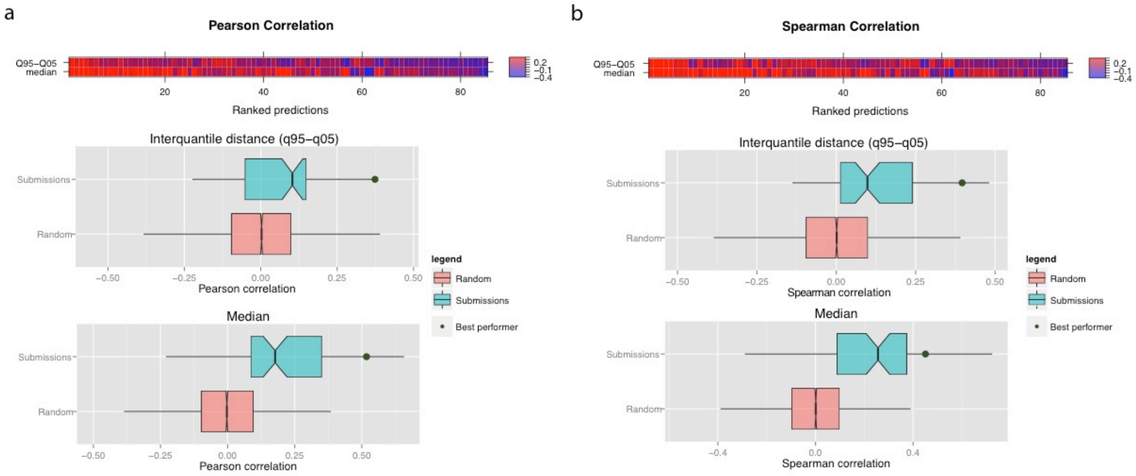
Supplementary Fig. 4. Robustness analysis for subchallenge 2



A robustness (sampling) analysis was performed for subchallenge 2 in order to assess the robustness of teams ranking with respect to compounds used for scoring. For 10000 iterations, both the mean ranking (left panel) and the final rank (right panel) were recomputed using each time only 45 randomly selected compounds (out of the 50) for scoring. A zoom of the 10 best performers is also shown, and distributions are compared using a one-sided Wilcoxon signed-rank test corrected for multiple hypothesis using the Benjamini-Hochberg correction. False discovery rates (FDR) are considered significant when $FDR<10^{-10}$.

See source code for the figure, files source_code/SuppFigure4.R

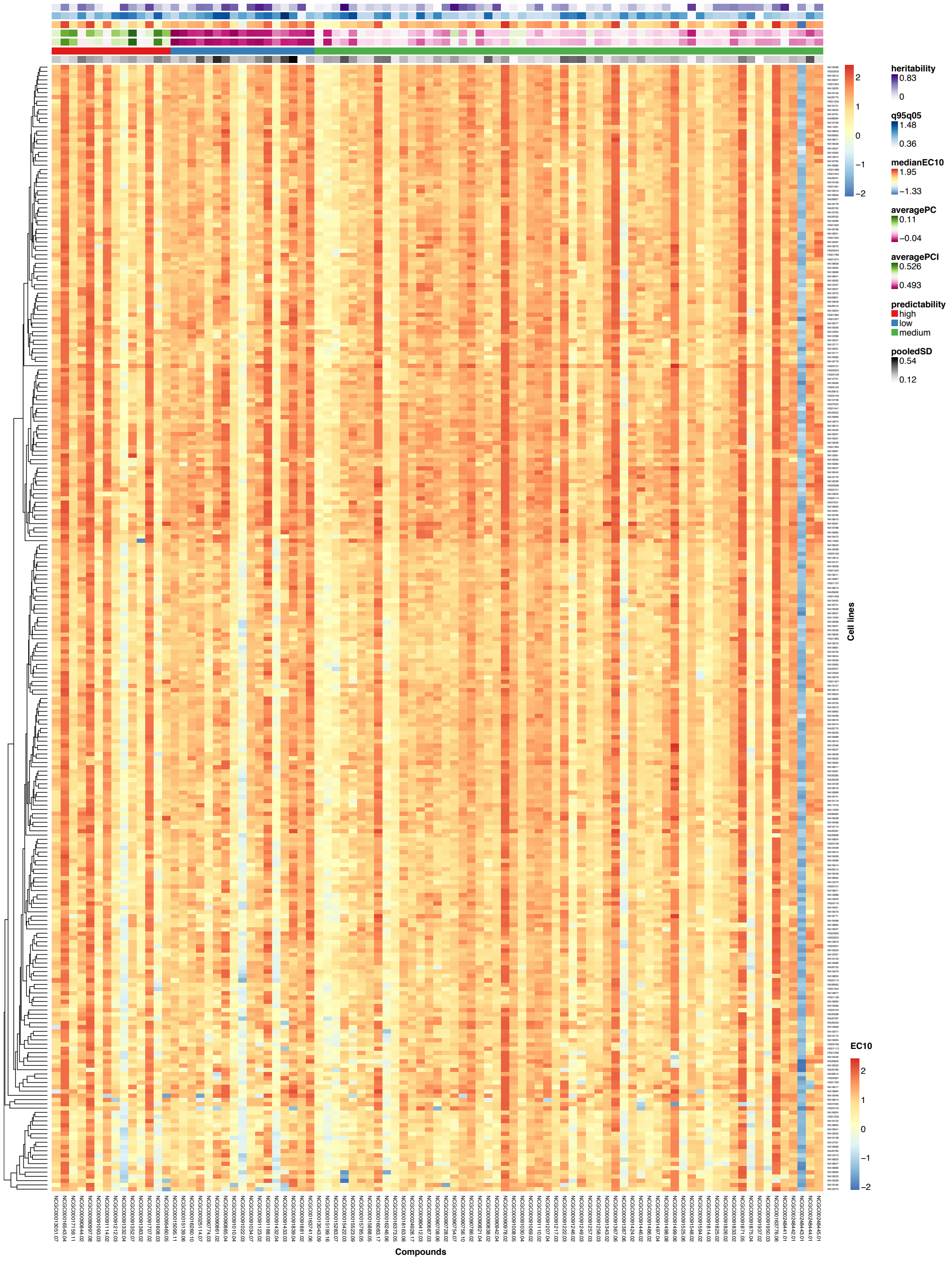
Supplementary Fig. 5. Performances of submissions for subchallenge 2



Predictions to subchallenge 2 were compared to the gold standard based on (a) Pearson Correlation and (b) Spearman correlation. The heatmap in the top panel illustrates performances of all ranked predictions for predicted median and interquantile range (q95-q05). In the bottom distribution of performances is shown for each predicted value along with the random distribution.

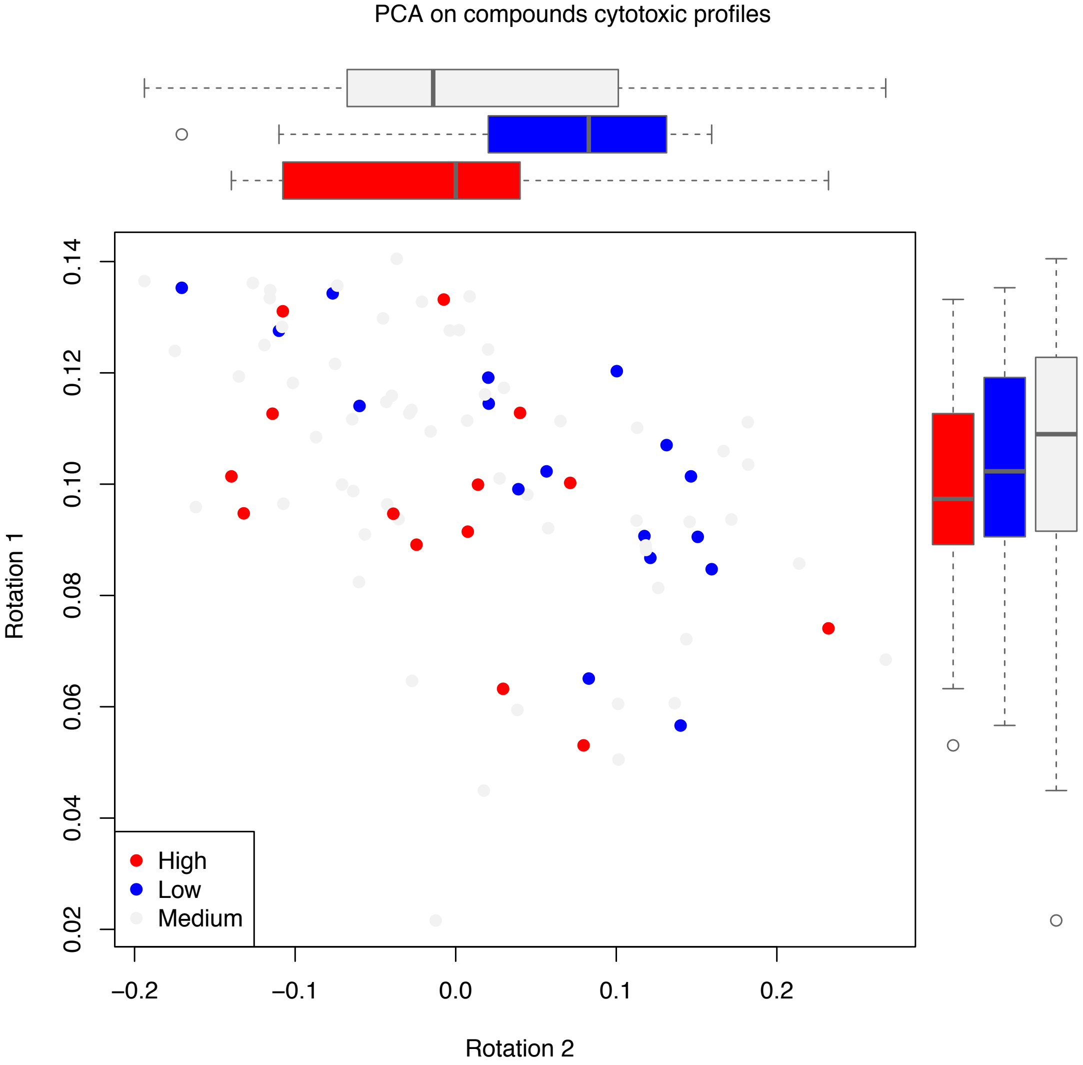
See source code for the figure, files source_code/SuppFigure5.R

Supplementary Fig. 6. Heatmap of the EC10 values clustered based on compounds predictability



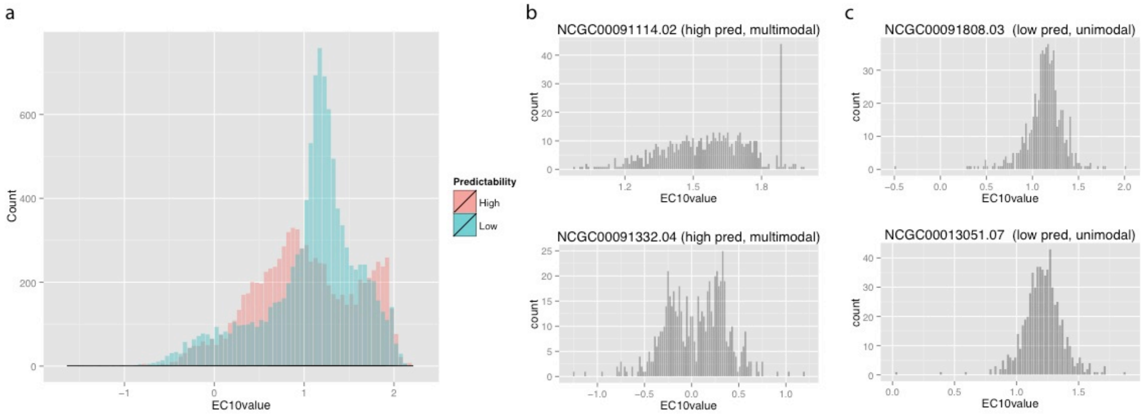
Cytotoxicity values are shown for all individuals in the test in response to the 91 compounds used for scoring. Compounds are clustered based on predictability, however clusters show no relationship between predictability and heritability, variability of toxicity across cell lines (in terms of interquartile distance, q95-q05), median toxicity across cell lines (medianEC10). As expected, some of the poorly predicted compounds show higher variability among replicates (in terms of pooled standard deviation).

Supplementary Fig. 7. PCA on the compounds cytotoxic profiles



Principal component analysis (PCA) was performed on the cytotoxic profiles of all compounds across cell lines. Compounds are coloured based on the level of predictability shown across submissions: well predicted compounds (high predictability, in red) are clearly separated from poorly predicted compounds (low predictability, in blue). The corresponding boxplot for the first and the second eigenvector are shown at the right and at the top of the scatter plot respectively.

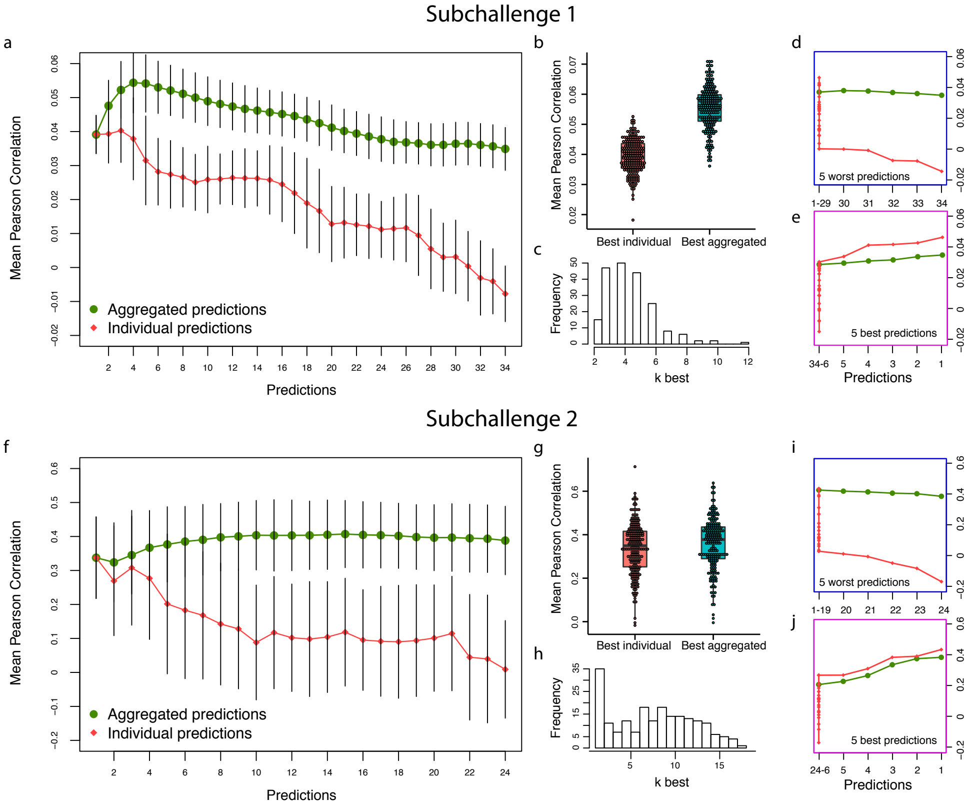
Supplementary Fig. 8. Multimodal and unimodal distribution of cytotoxicity values



Distribution of EC10 values across individuals shown for (a) all compounds showing high and low predictability; (b) two examples of highly predictable compounds with multimodal distribution according to Hartigans's dip test for unmodality (p-value<0.05) and (c) two examples of poorly predictable compounds with unimodal distribution according to Hartigans's dip test for unmodality (p-value>0.05).

See source code for the figure, files source_code/SuppFigure8.R

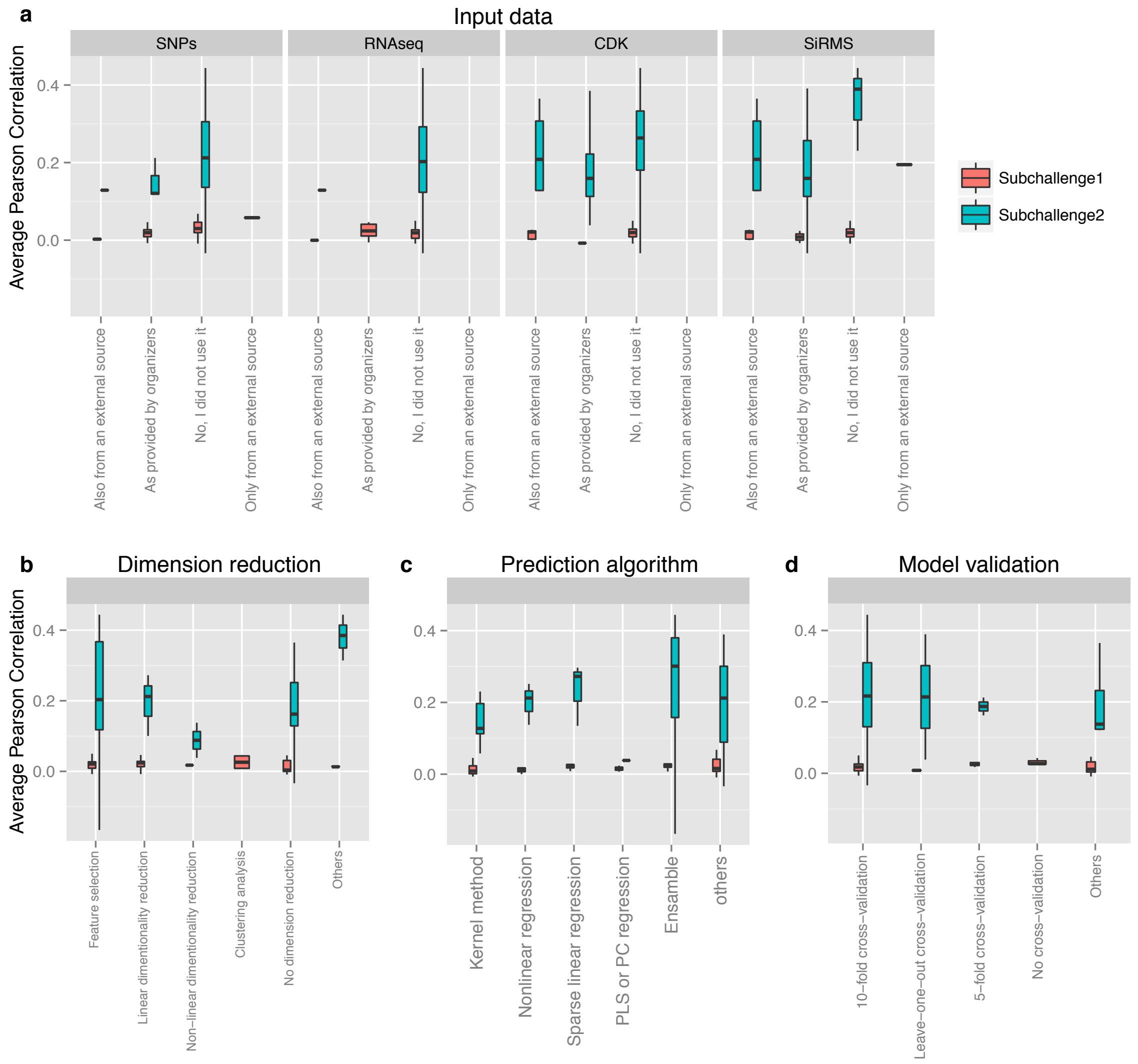
Supplementary Fig. 9. The wisdom of the crowds



Performances of individual predictions for all teams (in red) were compared with the performances of aggregated predictions (in green) for both (a-e) subchallenge 1 and (f-j) subchallenge 2. (a,f) Individual predictions are ordered from the best to the worse and aggregated: the first green point represents only the best prediction, the second is the aggregation of the top two predictions until the last point that is obtained by aggregating all predictions. (b,g) Average performances obtained by splitting the test data set 200 times in two subgroups, S1 and S2: one of the groups (S2) was used to evaluate the optimal number of teams in the aggregate, and the other (S1) to score their performance against that of individual teams. (c,h) Histogram of the optimal number of predictions to be aggregated to obtain the best aggregated prediction. (d,i) Zoom from panels a,f respectively, showing only aggregation of the five worse predictions. (e,j) Plot of the aggregation of the five best predictions obtained by first averaging the worse N-5 (where N is the total number of submissions for each subchallenge), first point, and then including the 4th, 3rd, 2nd and 1st predictions. Performances are shown in terms of average Pearson correlation computed between predicted and measured values separately for each compound. Predictions were aggregated by averaging them. In order to aggregate only independent predictions, only one submission for each team was considered as the average of all predictions submitted by the team.

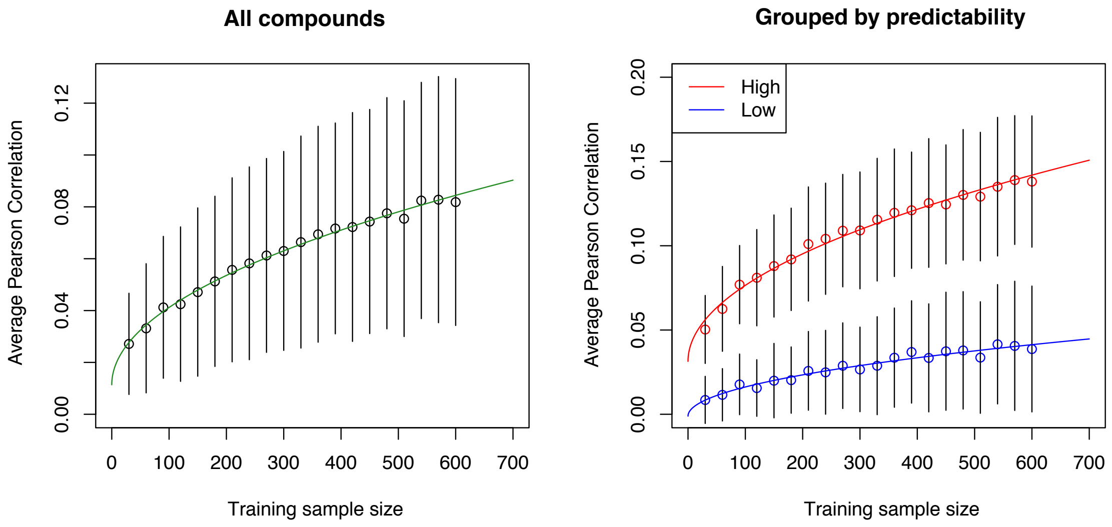
See source code for the figure, files sourcecode/SuppFigure9sch1.R and sourcecode/SuppFigure9sch2.R

Supplementary Fig. 10. Performances based on methods and data used to solve the challenges



Performances of predictions grouped based on used input data, data reduction techniques, prediction algorithms, and model validation techniques are shown using Pearson correlation computed separately for each compound and then averaged across compounds. Each team submitted up to 5 submissions that are typically not independent. In order to compare performances only for independent predictions, predictions are considered independent if they use different data or approaches (i.e., at least one different answer to the survey), non-independent predictions for the same team are averaged and considered as one prediction. We obtain 49 independent submissions for subchallenge 1 (out of the 75 for which the survey was filled) and 28 for subchallenge 2 (out of the 51 for which the survey was filled).

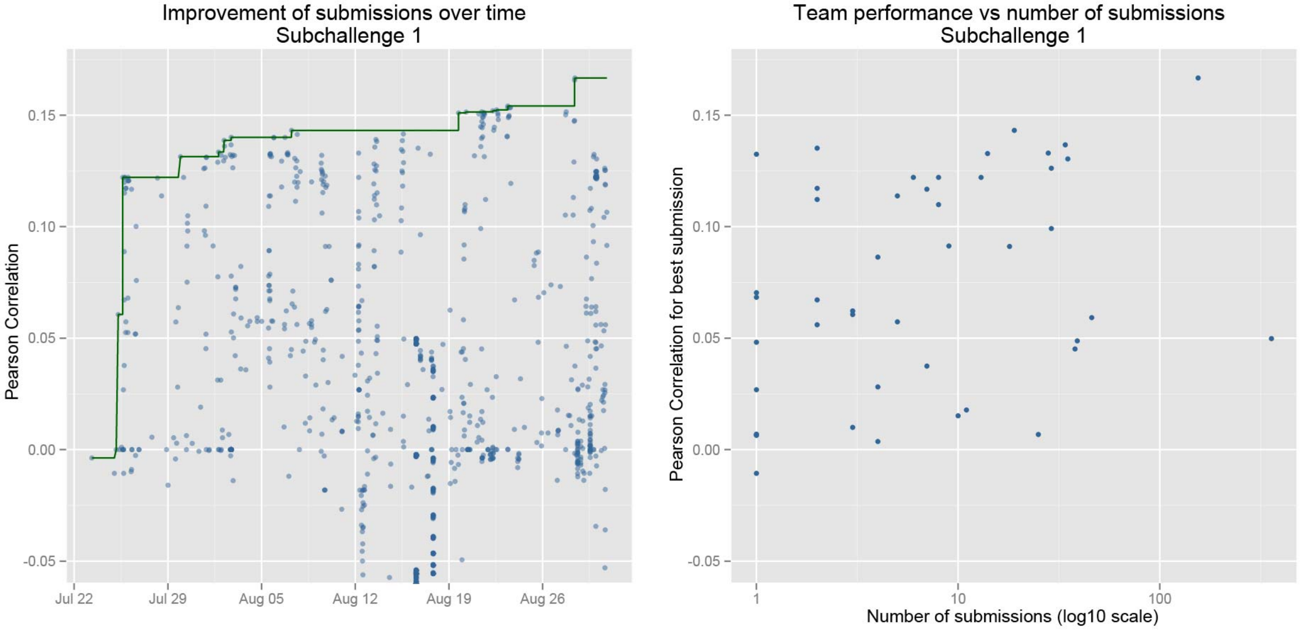
Supplementary Fig. 11. Down-sampling analysis for subchallenge 1



Down-sampling analysis was performed for subchallenge 1 using the best performing method to analyse how the sample size in the training dataset affects the prediction accuracy. The training set was randomly sampled 100 times for each sample size, with sample size ranging from 30-600 (increment of 30). Each randomly selected training set was used to train the model in order to predict the corresponding testing set (264 cell lines, different for each iteration). Left panel: The increase in performances with sample size is shown as average Pearson Correlation across all compounds, with error bars representing standard deviation across iterations. Right panel: Separate plots for compounds showing high and low predictability, with error bars representing standard deviation across compounds in the same group.

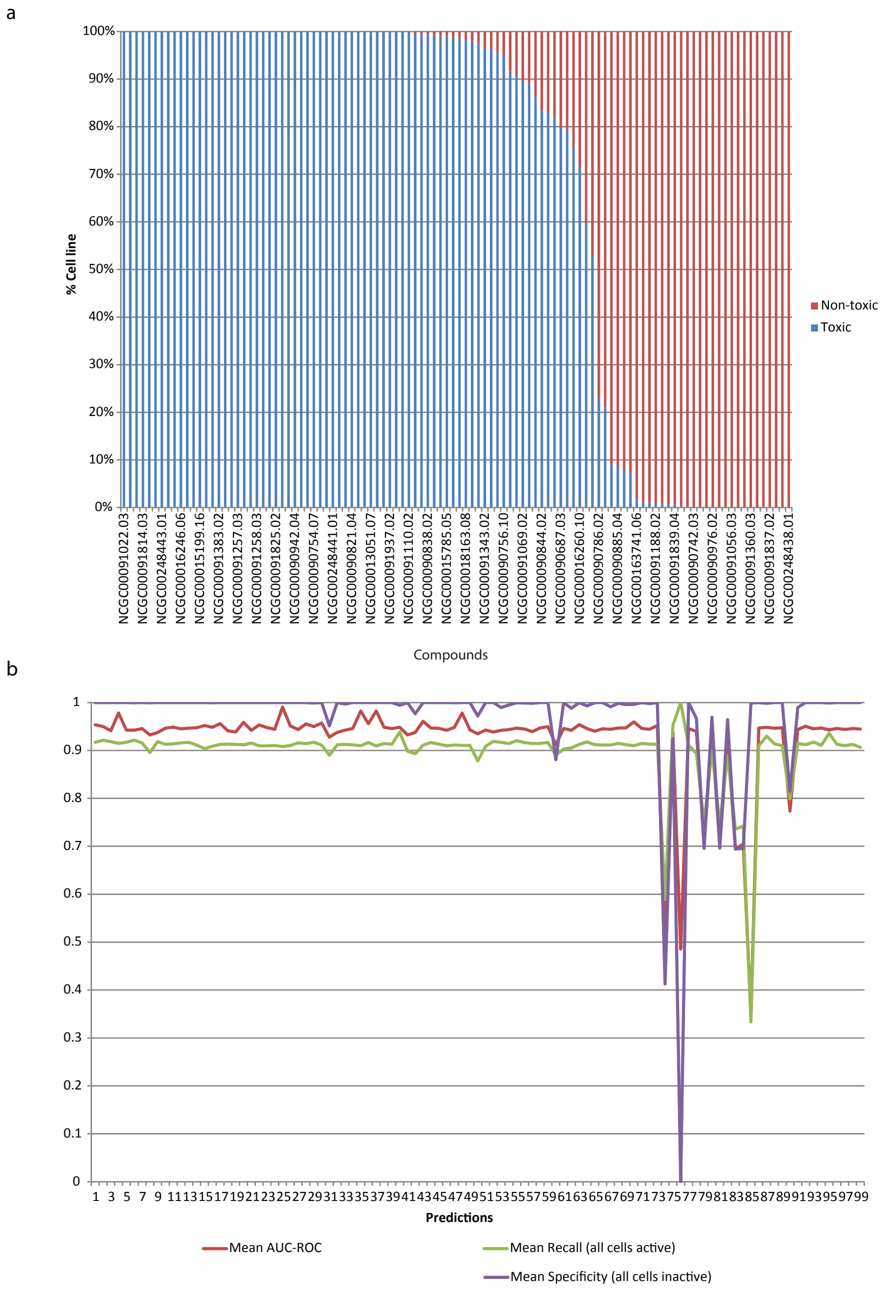
Nature Biotechnology: doi:10.1038/nbt.3299

Supplementary Fig. 12. Performances for Subchallenge 1 real-time leaderboard



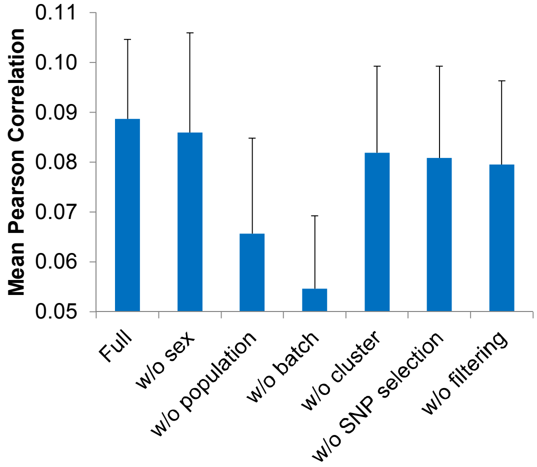
A real-time leaderboard was run for subchallenge 1 over a three-month period on a test set (133 cell lines) that was then released as part of the final training set (see Figure 1). During this phase, teams could submit as many predictions as desired. In the left panel, each submission is represented by a blue dot and the improvement in the best overall prediction is shown in green. On the right panel, each team's best prediction is plotted against the number of submissions made by that team on a log10 scale. Pearson's correlation between the number of submissions and their score is weakly positive at 0.0849.

Supplementary Fig. 13. Reexamination of subchallenge 1 as a classification problem



(a) Activity distribution of test set compounds across all cell lines. (b) Predicted EC10 values from all teams were applied to predict the activity outcome (toxic or non-toxic) of compounds in each cell line. Performances of predictions were measured by the area under the receiver operating characteristic curve (AUC-ROC) (red line: AUC-ROC values were calculated for each cell line then averaged across all cell lines), specificity (purple line; calculated for compounds that had non-toxic calls in all cell lines then averaged), and recall (green line; calculated for compounds that had toxic calls in all cell lines then averaged).

Supplementary Fig. 14. Analysis of the prediction procedure of the best performing team of sub challenge 1



The performance of prediction model judged by mean Pearson Correlation across all compounds. “Full” denotes the prediction model built using sex, population, experimental batch and the “genetic cluster” variable. The next four columns are prediction models lacking each of the four predictors. The “w/o SNP selection” column shows the prediction model whose “genetic cluster” variable is generated based on a randomly sampled set of 0.15 million SNPs from the 0.61 million SNP set after the first round of narrowing down. The “w/o filtering” columns shows the prediction model whose “genetic cluster” variable is generated based on all the 0.61 million SNPs. All predictions for each group are repeated 20 times and each time a random split of the whole dataset into a 620-cell line training subset and a 264-cell line testing subset is applied.