In format provided by Saez-Rodriguez *et al.* (doi:10.1038/nrg.2016.69)

Supplementary information S1 (box): Scoring Metrics

An appropriate metric is essential to score a challenge. Broadly speaking, there are two main challenge questions, classification (typically binary classification) and regression.

Classification is the task of assigning elements of a dataset into two (binary) or more groups (e.g. patient into responder or non-responder to a treatment). For the binary case, there are four possible outcomes that can be arranged in a 2x 2 matrix, the so-called contingency matrix:

	Gold Standard Positive Set (P=TP+FN)	Gold Standard Negative Set (N=FP+TN)
Predicted Positive	True Positive (TP)	False Positive (FP)
Predicted Negative	False Negative (FN)	True Negative (TN)

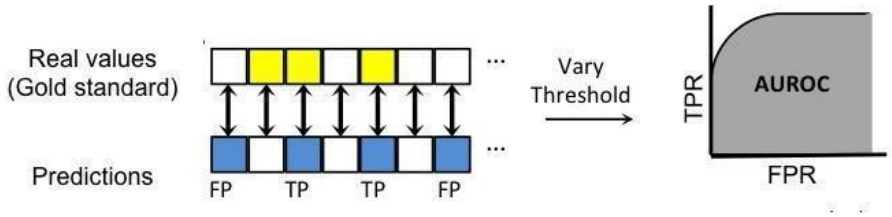
Multiple metrics can be derived from a contingency table. Some of the most common are

- True positive rate (TPR), also known as sensitivity or recall: TPR = TP/ P
- True negative rate (TNR) or specificity: TNR = specificity = TN/N
- False negative rate (FNR): FNR = FN /P
- False positive rate (FPR): FPR=FP/N
- Precision = TP/ (TP +FP)
- False discovery rate (FDR): FDR = FP /(TP + FP)

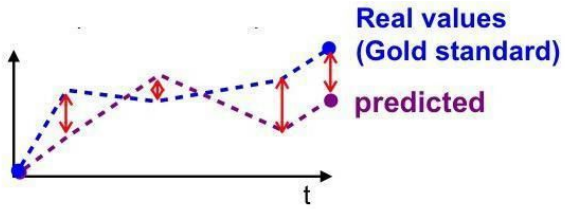
Generally there is trade-off between precision (being right in the calls made) and recall (identifying the calls that can be made) in a classification problem. Hence, these are sometimes combined in metrics that aim to balance them, such as the F-score (the harmonic mean of precision and recall), and Mathew's correlation coefficient.

Often a classification algorithm can provide more or less calls with more or less confidence, and such a confidence is often asked in the context of the Challenges. By computing precision and recall for different levels of confidence, plotting them, and then joining those points, one can compute the Precision-Recall (PR) curve. Similarly, by computing and plotting TPR and FPR, one obtains the closely related ROC (Receiving Operating Characteristics) curve. Both represent the capacity of a given algorithm for different levels of confidence, and are often summarized by computing the area under their curve (AUPR and AUROC, respectively). While both give very similar information (Davis and Goadrich 2006), AUPR is more accurate for cases where the number of positives and negatives is very unbalanced.

In format provided by Saez-Rodriguez *et al.* (doi:10.1038/nrg.2016.69)



In a *regression problem* the task is to predict the numerical values for a number of variables (dependent variables), based on certain features (independent variables). A common metric is the root-mean squared error (RMSE) that averages the quadratic errors of the individual measurements. Another common metric to compare predicted vs. measured values is the Pearson correlation.



There is a simple relationship between the RMSE and the Pearson correlation coefficient ϱ:

$$RMSE^2 = (\mu_{pred} - \mu_{exp})^2 + (\sigma_{pred} - \sigma_{exp})^2 + 2\sigma_{pred}\sigma_{exp}(1 - \varrho)$$

where σ_{pred} and, σ_{exp}, μ_{pred} and μ_{exp} are the standard deviations and μ_{pred} and μ_{exp} the means in the predictions and experimental (gold standard) data, respectively. This relationship nicely shows that RMSE is aggregating the comparison of predictions and measurements in several facets simultaneously, namely the average (μ), the range (σ) and how they covary (p). This may be undesirable if one of the terms dominates over the others, which makes it difficult to separate subtle performance differences between teams. Sometimes it is desirable to compare the order (rank) of the predictions and gold standard rather than the actual values, when the actual ordering is the important thing to predict (e.g. prioritize drugs from more to less efficacious as treatment(Costello et al. 2014)). The analogous metric to Pearson's correlation considering ranks is Spearman's rank correlation coefficient. Another useful rank-based metric is the Concordance Index.

When the Gold Standard is noisy, the regression metrics should take into account the experimental variability, weighting the predictions so as to give more importance to data points whose ground truth we are more certain about. For example, the RMSE was divided by the experimental noise in some challenges(Prill et al. 2011; Hill et al. 2016), or the Concordance Index modified into the so-called probabilistic c-index(Costello et al. 2014).

Different metrics highlight different aspects of an algorithm performance. Therefore a thorough evaluation of the strengths and weaknesses of an algorithm requires looking at it under the light of different metrics. To cover the multiple aspects of prediction evaluation a combination of several scoring metrics is desired. In the end, a final score based on the combination of different metrics can provide an integrated evaluation of the quality of the predictions.

All these scoring metrics have then to be compared with a null model (for example random predictions), to assess the statistical significance of predictions. It is important to ensure that the final ranking in a Challenge is robust to subtle changes in the test set. This can be achieved by generating an ensemble of new submissions by bootstrapping the test set and assessing if the difference in ranking between teams (e.g., first and second, or second or third) is statistically significant.

A collection of all metrics used in the DREAM Challenges is available in the package DREAMTools.(Cokelaer et al. 2015)

References

Cokelaer, Thomas, Mukesh Bansal, Christopher Bare, Erhan Bilal, Brian M. Bot, Elias Chaibub Neto, Federica Eduati, et al. 2015. “DREAMTools: A Python Package for Scoring Collaborative Challenges.” *F1000Research* 4 (October). http://f1000research.com/articles/4-1030/v1.

Costello, James C., Laura M. Heiser, Elisabeth Georgii, Mehmet Gönen, Michael P. Menden, Nicholas J. Wang, Mukesh Bansal, et al. 2014. “A Community Effort to Assess and Improve Drug Sensitivity Prediction Algorithms.” *Nature Biotechnology*, June. Nature Publishing Group. http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=24880487&retmode=ref&cmd=prlinks.

Davis, Jesse, and Mark Goadrich. 2006. “The Relationship Between Precision-Recall and ROC Curves.” In *Proceedings of the 23rd International Conference on Machine Learning*, 233–40. ICML ’06. New York, NY, USA: ACM.

Hill, Steven M., Laura M. Heiser, Thomas Cokelaer, Michael Unger, Nicole K. Nesser, Daniel E. Carlin, Yang Zhang, et al. 2016. “Inferring Causal Molecular Networks: Empirical Assessment through a Community-Based Effort.” *Nature Methods*, February. doi:10.1038/nmeth.3773.

Prill, Robert J., Julio Saez-Rodriguez, Leonidas G. Alexopoulos, Peter K. Sorger, and Gustavo Stolovitzky. 2011. “Crowdsourcing Network Inference: The DREAM Predictive Signaling Network Challenge.” *Science Signaling* 4 (189). AAAS: mr7.

In format provided by Saez-Rodriguez *et al.* (doi:10.1038/nrg.2016.69)

Supplementary information S2 (table). Examples of collaborative competitions. A set of nineteen Challenges organized in the past six years (see also the additional case studies in the main text). This table is an expanded version of Table 1, in which additional information is provided, primarily regarding the solvability of the Challenges based on the data provided to participants and on scoring metrics used. Challenges are coloured according to the research area.

Challenge name, Reference, Year of challenge, Participation	**Challenge question**	**Gold standard and scoring**	**Solvability: does the underlying data provide information for successful predictions?**	**Winning methodology or algorithm**	**Scientific advance (What we learned scientifically or biologically)**	**Legacy (e.g. databases, biomarkers in use, spin-out companies)**
Gene regulation and signalling network Challenges						
DREAM5 Gene regulatory network inference[1] (2010) 29 teams	Infer a transcription factor-to-target gene regulatory network	GS: RegulonDB for *E. coli* ; GeneNetWeaver known interactions for *in silico*; ChIP binding and evolutionary conservation for *S. cerevisiae* Scoring: area under the ROC and area under the PR curves	Performance for the *in silico* and *E. coli* networks were high, but *S. cerevisiae* inferences were poor. A network was constructed using all the teams and experimental validation was used to verify overall precision of 50%.	The top method to predict *E. coli* interactions was based on a two-way ANOVA. The top method to predict *in silico* interactions used group lasso regression and bootstrapping.	Network motifs were predicted differently based on the underlying model. The 'wisdom of crowds' model was the most robust across all individual models.	Challenge Publication[1]. The GenePattern-DREAM server can be used to run individual methods and build an ensemble prediction (http://dream.broadinstitute.org/)
DREAM TF–DNA Motif Recognition Challenge[2](2010) 14 teams	Model the DNA binding sites of a transcription factor (TF) based on protein binding microarray (PBM) data.	GS: The measured degree of binding of each of the TF in the test set in an independent PBM. Scoring: Correlation between predicted and measured signals, Precision/Recall analysis.	Quality of the predictions was dependent on both the algorithm used and the specific TF. In general, the best algorithms produced highly accurate predictions (AUROC > 0.95)	The best method was based on a k-mer model. Several position-weight-matrix (PWM)-based methods also performed well for most TFs.	PWMs work well for most TFs. *In vitro*-based TF binding measurements can be used to effectively distinguish *in vivo*-bound sequences from random sequences. Most TFs recognize highly 'degenerate' sequences.	Challenge Publication[2]. Server to enable continuous benchmarking of methods (http://www.ebi.ac.uk/saezrodriguez-srv/d5c2/cgi-bin/TF_web.pl). All data are also available (http://cisbp.ccbr.utoronto.ca/)

In format provided by Saez-Rodriguez *et al.* (doi:10.1038/nrg.2016.69)

DREAM Gene Expression Prediction Challenge[3] (2011) 21 teams	Predict the expression levels of genes downstream of ribosomal promoters based on the DNA sequence of the promoter	GS: Fluorescence of GFP downstream of promoters. Scoring: RMSD and correlations between measured and predicted expression	Correlations were above 0.8. Post-challenge model considering prior knowledge (TF and RNA polymerase binding site information) fared better than original submissions.	SVM with a previous search for the best adapted feature was complemented by a previous physical model of TF and RNA polymerase interaction with DNA.	General models for promoter expression prediction did not fare well for predicting a specific family of promoters (ribosomal genes)	Challenge Publication[3]. Data produced is available for benchmarking models in https://www.synapse.org/GeneExpressionChallenge
DREAM Network Topology and Parameter Estimation Challenges[4] (2011-2012) 31 teams (19 in 2011, 12 in 2012)	SubC1: Infer kinetic parameters in *in silico* gene regulatory networks. SubC2: Predict protein time courses under perturbed conditions. SubC3: Find missing network edges based on a limited set of data.	SubC1: GS: Actual kinetic parameters from *in silico* model. Scoring: RMSD in log scale SubC2. GS: Simulated time courses. Scoring: Normalized RMSD between predicted and simulated protein values. SubC3: GS: known missing edges. Scoring. Number of edges and nodes correctly predicted.	The solutions for parameter estimation and dynamic predictions were very good. The solutions for the network topology problem were not very good, probably due to the difficulty of the problem, rather than the lack of adequate data.	Maximum likelihood fit of the model parameters given observed data obtained from *in silico* experiments and construction of a game tree of possible sequences of most informative data to use and experiments to perform.	The main conclusion is that given a model, a low amount of well-chosen data is enough to have a good estimate of parameters and dynamics of the GRN. The difficulty in solving the network topology problem confirms the essential problem of finding the correct GRN topology.	Challenge Publication[4] Networks and Data produced are available for benchmarking paramater estimation approaches in https://www.synapse.org/NetworkTopologyChallenge

In format provided by Saez-Rodriguez *et al.* (doi:10.1038/nrg.2016.69)

HPN-DREAM Breast Cancer Network Inference Challenge[5] (2013) 178 final submissions	SubC1: Infer signalling networks in breast cancer cell lines using protein time-course data obtained after intervention on specific proteins SubC2: Predict phosphoprotein time-course data given a specific intervention. SubC3: Develop tools to visualize the Challenge data.	SubC1. GS: Measured perturbed protein downstream of the intervention in the withheld data set. Scoring: AUROC of standardized data. SubC2. GS: Measured time course of the phosphorylation levels resulting from intervention. Scoring: RMSD between predicted and true time courses. SubC3. No GS. Scoring: All participants voted for their favourite visualization.	SubC1: several teams attained statistically significant AUROC scores. Performance varied across cellular contexts. In some cases, only marginal improvements over prior information alone, while in others there was a clear gain in performance. SubC2: teams did not do well at predicting protein abundance time courses following specific protein inhibition. SubC3: Including a visualization subchallenge to a data challenge can motivate the development of better data representations.	SubC1. Granger causality, extended to include future time points, combined with prior networks based on known biological pathways. Another top method (FunChisq) used a chi-squared test to examine functional dependencies among variables without using any prior information. SubC2. One top method used a regression model with truncated singular value decomposition. A second method used Generalized Linear Models informed by networks inferred in SubC1. SubC3: Biowheel visualization (dream8.dibsbiotech.com).	Results suggest that causal network inference is feasible in complex mammalian settings. Scoring by empirically assessing inferred causal networks using withheld interventional data can be applied in other settings. Incorporation of prior information was broadly beneficial. Data-driven learning offered the most utility in those contexts where prior information alone performed less well. Submissions included novel approaches.	Challenge Publication[5]. All challenge data, included open source code, participant prior networks and crowdsourced aggregate networks have been made available as a community resource (www.synapse.org/HPN_DREAM_Network_Challenge). The best performing method is implemented in the Cytoscape tool Cyni. The visualization tool Biowheel is available at dream8.dibsbiotech.com

In format provided by Saez-Rodriguez *et al.* (doi:10.1038/nrg.2016.69)

Translational and clinical challenges						
FlowCap/ DREAM Molecular Classification of AML Challenge[6] (2011) 39 teams	Classify AML versus normal blood samples from flow cytometry data	GS: Actual diagnosis of healthy versus AML in the test dataset of blood samples. Scoring: AUPR.	Challenge was fairly easy and multiple participants got a perfect score.	Not relevant in this context, as many algorithms got a perfect score.	If the signal is clearly contained in the data, the choice of machine learning algorithms is inessential to identify correlates of clinical outcomes in flow cytometry data	Challenge Publication[6] Dataset is public at FlowRepository.org and has been used in multiple independent articles.
DREAM-Phil Bowen ALS Prediction Prize4Life Challenge[7] (2012) >1000 registrants 37 unique teams 10 teams made final submissions.	Predict the progression of patients with ALS from clinical trial data	GS: Slope of change in ALS Functional Rating Scale (a measure of disease status) per unit time Scoring: RMSD and correlations between measured and predicted slopes.	The relatively small size of the data set (while the biggest available at the time) probably took away from performance. The best performing team only improved beyond a baseline algorithm by a small but significant amount	Two teams were identified as winners. One of them used a Bayesian Additive Random Trees, whereas the other used random forest.	Best performers predicted ALS progression better than a group of consulted physicians. An analysis of most informative features identified potential novel biomarkers such as creatinine and creatine kinase	Challenge publication[7]. Origent Data Sciences, (http://www.origent.com) was spanned out from Sentrana to predict disease behaviour of individual patients. This Challenge was the basis of the subsequent DREAM Challenge on ALS.
Sage Bionetworks-DREAM Breast Cancer Prognosis Challenge[8] (2012) 354 registrants	Predict the survival of patients with breast cancer on the basis of gene expression data, genomic copy number data, and clinical covariates.	GS was the actual survival of patients in the test set. Scoring: Concordance index (CI) between the predicted risk score and overall survival.	Models that used clinical covariates alone achieved a CI of ~0.70. The addition of genomic features provided an incremental improvement in CI of up to 0.76 The best performing model beat the first generation 70-gene risk predictor MammaPrint.	The winning algorithm topped the leaderboards throughout the different phases of the Challenge. The main idea that differentiated it from other methods was the use of 'Attractor Metagene'[9] features. Briefly, these are features built by combining the expression of multiple genes using a mutual-information-based iterative algorithm.	Copy number and gene expression data provided only an incremental performance improvement over clinical covariates alone, especially for aggressive high-grade tumours. This suggests that additional genomics data may be necessary to capture tumour progression.	Challenge Publication[8] The winning method Attractor Metagenes is now part of the standard bioinformatic toolboxes in R and Matlab.

In format provided by Saez-Rodriguez *et al.* (doi:10.1038/nrg.2016.69)

Alzheimer's Disease Big Data DREAM Challenge [10] (2014) 520 registrants 100 Unique Teams 1,296 Total Submissions	SubC1: Predict changes in cognitive scores 24 months after initial assessment based on genetic data. SubC2: Predict amyloid perturbation in a set of cognitively normal individuals based on genetic data SubC3: Classify individuals into diagnostic groups using magnetic resonance imaging	SubC1: GS was the actual cognitive score for patients in the test set. Scoring: Correlation between predicted and the actual change in cognitive scores SubC2: GS was the actual status of amyloid perturbation. Scoring: AUROC and Balanced Accuracy. SubC3: GS: The actual diagnosis of the patients in the test set. Scoring: Correlation and Lin's concordance correlation coefficient for agreement on a continuous measure between observed and predicted cognitive scores	SubC1 and SubC2: modest performance suggest that algorithms were not able to leverage genetic signal to predict cognition changes, or that such information was not to be found in genetics data SubC3. Modest performance that validated an established relationship between structural imaging data and cognition, but performance was low for application in a clinical setting.	SubC1: Six teams performed significantly better than the rest but were statistically indistinguishable from each other. SubC2. Participants were unable to develop algorithms with predictive performances significantly better than random SubC3. Three teams performed significantly better than the others but were indistinguishable from the each other.	Predictions of cognitive decay from genetic or structural imaging data were modest across a diverse set of modelling methods. Future efforts will benefit from a focus on methods that work to incorporate greater phenotypic complexity across diverse data sources. Today's premier publicly available data repositories for Alzheimer's disease have use restrictions that made it very difficult to collate and widely share the data for this Challenge.	Challenge Publication [10] The data used in the Challenge is available for download at https://www.synapse.org/AD_Challenge
Rheumatoid Arthritis Responder DREAM Challenge (2014) 373 registrants; 73 teams contributed final submissions	Predict the response to anti-TNF therapy in patients with rheumatoid arthritis based on genotype information.	GS. Known response of patients in the test set. Scoring: Correlation, AUPR and AUROC.	Best correlation: 0.39; Best AUPR: 0.51 (null expectation 0.36); best AUROC: 0.62. Although the best performing teams had better than random predictions, they were not of sufficient quality for clinical utility. Signals resulted mostly from clinical covariates.	Gaussian Process Regression	Community phase showed that genetic predictors did not significantly contribute to anti-TNF response prediction.	Methods and outcomes are archived through Challenge website (https://www.synapse.org/RA_Challenge). All data are available for secondary use through Synapse (https://www.synapse.org/RAchallengedata)

In format provided by Saez-Rodriguez *et al.* (doi:10.1038/nrg.2016.69)

Genotype-to-phenotype prediction Challenges						
NCI-DREAM Drug Sensitivity Prediction Challenge[11] (2012) 40 teams submitted results 127 individuals	Rank a panel of breast cancer cell lines from the most sensitive to the most resistant to a set of drugs, based on gene expression, mutation, copy number, DNA methylation, and protein quantification of the untreated cell lines.	GS: Concentration of drug that inhibits the growth to 50% of the maximum (GI50), measured over 28 drugs and 18 breast cancer cell lines. Scoring: Probabilistic CI (PCI), a weighted version of the concordance index that takes into account the noisy nature of the GS.	Many teams performed significantly better than random, suggesting that there is signal in the basal omics datasets to predict drug sensitivity, although there was much room for improvement. Some drugs and drug classes were more easily predicted than others.	The top performer, which significantly outperformed the next best team, developed a novel method that leveraged a range of machine learning approaches including Bayesian inference, multitask learning, multiview learning and kernelized regression. This nonlinear, probabilistic model aims to learn and predict drug sensitivities simultaneously from all drugs.	Integrative approaches to leverage all the available omics data pay off. Microarray data were the most informative individual dataset. Drug classes showed variation in predictability. Crowdsourcing promotes innovation, as the top performing method was a novel one.	Challenge Publication[11]. The NCI awarded contracts to the two best performing teams to strengthen the models and create a resource that can be used for the purpose of estimating drug sensitivities given multiple omics data sets. Challenge data available in http://www.synapse.org/NCI_DREAM

In format provided by Saez-Rodriguez *et al.* (doi:10.1038/nrg.2016.69)

NCI-DREAM Drug Synergy Prediction Challenge[12] (2012) 31 teams	Rank 91 compound pairs (all pairs of 14 compounds) from the most synergistic to the most antagonistic in a human lymphoma cell line, using gene expression profiles of cells perturbed with the individual compounds.	GS. Excess over Bliss (EoB), a measure of the deviation from additivity for all compound pairs. Scoring A weighted version of the concordance index, that takes into account the noisy nature of the GS.	Three teams performed better than chance (PCI ~0.61;maximum possible score: 0.9), indicating that there was signal in the data. Top methods provide substantial potential reductions of the search space for synergistic drug pairs.	The best performing method hypothesized that when cells are sequentially treated with two compounds, the transcriptional changes induced by the first contribute to the effect of the second. A synergistic score was calculated by averaging two possible sequential orders of treatment between pairs of compounds.	Compounds exhibiting polypharmacology are more often synergistic. Compounds with targeted mechanisms are more likely antagonistic. Hypotheses used to predict synergy may not necessarily apply to predicting antagonism, and vice-versa. Synergy and antagonism are highly cell-context specific.	Challenge Publication[12]. The NCI awarded contracts to the two best performers to strengthen the models and create a resource that can be used for the purpose of estimating drug synergy given gene expression data from the monotherapies. Challenge data available in http://www.synapse.org/NCI_DREAM
CAMDA Ideation Challenge: dataset from the Japanese Toxicogenomics Project (TGP)[13] (2013) ~20 teams	Question 1: Can we replace the animal study with in vitro assays? Question 2: Can we predict the liver injury in humans using toxicogenomics data from animals?	GS. Example data was provided. Scoring: 4-fold cross-validation and Matthew's correlation coefficient.	The conclusion was that the problem of predicting response with the set of compounds used was very difficult. The inclusion of non-toxic drugs in the provided dataset would have probably helped in improving results.	First recursive feature elimination (RFE) followed by classification with artificial neural network consisting of 50 input units, 10 hidden units and 1 output unit with sigmoid activation.	The prediction of liver injury in humans using toxicogenomic data from animals is possible, but more data, especially non-toxic drugs, would be necessary to obtain better predictions	Challenge Publication[13]

In format provided by Saez-Rodriguez *et al.* (doi:10.1038/nrg.2016.69)

NIEHS-NCATS-UNC DREAM Toxicogenetics challenge[14] (2013) 213 registrants 57 teams (34 teams in SubC1 and 23 teams in SubC2)	SubC1. Predict cytotoxicity of individual cell lines to a given set of compounds based on genotype information and RNA-seq data for a subset of cells. SubC2. Predict population-level cytotoxicity for different compounds based on chemical attributes.	SubC1. GS. Measured cytotoxicity data for cell lines in the test set in response to chemical compounds. Scoring: Correlation and probabilistic CI. SubC2. Average and standard deviation in the population for the compounds in the test set. Scoring: Correlation between measurements and predictions.	SubC1: predictions were overall poor even if robustly significant for best performing teams. Availability of RNA seq data for some of the cell lines (instead of only genotype data) showed improved performances. SubC2. Good performances of top teams; Correlation= 0.65 and 0.37 for prediction of median toxicity and interquartile distance across the population.	SubC1: Random forest algorithm was used to build a model for each compound using as variables genetic SNPs, sex, population and experimental batch. SubC2: Random forest models were built separately for each group of compounds using as features a selection of chemical attributes. Predictions for new compounds were based on similarity to the compounds clusters.	Genotype data are not sufficient to have meaningful predictions of cytotoxicity in individual cells. Transcriptional data are more informative. Increased sample size would probably improve predictions. Chemical attributes are good predictors of mean cytotoxicity in the population and of the variability in the response.	Challenge Publication[14] All data and methods used to solve the challenge (code and wiki with approach descriptions) are available in Synapse at (https://www.synapse.org/ToxicogeneticsChallenge).
CAGI PGP, predict individuals phenotype From their personal genomes[15] (2013) 16 teams.	From 291 subjects, 77 genomes matched a phenotype from a list of 243 phenotypic profiles, 214 were “decoys”. Participants had to match each genome to a phenotype.	For each subject, the 243 pheno- typic profiles were ranked from most probable to least probable. Evaluate based on correct top-ranked profiles and mean rank of the correct profiles for all participants Phenotypes were based on surveys.	Models were assessed by their ability to correctly rank individuals in the PGP cohort. AUC values had a p-value <10-4	Bayesian probabilistic model predicting risk of a dichotomous phenotype using population-level prevalence as a prior, integrating the contribution of rare and common variant genotypes in an individual.	Model using the combination of GWAS hits, Low penetrance genes, High penetrance genes and High penetrance variants yields the best performance.	Challenge publication by best performer[15]

In format provided by Saez-Rodriguez *et al.* (doi:10.1038/nrg.2016.69)

Next-generation sequencing data analysis						
Assemblathon 1: A competitive assessment of *de novo* short-read assembly methods[16] (2010) 17 teams.	Assemble *de novo* a simulated diploid genome from short-read sequences	GS: Simulated data. Scoring: contig accuracy, scaffold accuracy, reconstruction of genes and functional features, phasing of separate haplotypes. No attempt was made to aggregate the metrics.	The solutions were qualitatively quite good. In part this was due to the fact that the simulated genome was only 112 Mb (~4% the size of the human genome).	Several best-performing methodologies were identified. Many of the methods used variants of de Bruijn graphs. What distinguished the best methods were the heuristics used for error correction, bubble removal, contig resolution, scaffolding, etc.	The best sequence assemblers could reconstruct at high coverage and with good accuracy large sequences of a *de novo* genome.	Challenge publication[16] Lessons from Assemblathon 1 were used to create the Assemblathon 2[17] and the Alignathon[18]. The Assemblathon 1 data and code are published online and free to use in www.assemblathon.org/assemblathon1
RGASP, RNA-seq Read Mapping[19] (2011) 11 computational methods 26 protocol variants	Align RNA-seq reads to reference genomes, identifying loci of origin and reporting alignments with correctly placed introns, mismatches, and small insertions and deletions (indels).	GS: RNA-seq from simulated transcriptome data. Scoring: Several metrics specific to short-read alignment problem: alignment yield, basewise accuracy, mismatch and gap placement, exon junction discovery and suitability of alignments for transcript reconstruction.	High degree of solvability. Different top methods had different strengths and weaknesses. MapSplice was conservative in mismatch frequency, indel and exon junction calls. GSNAP, GSTRUCT and STAR had many false exon junctions.	GSNAP, GSTRUCT, MapSplice and STAR compared favourably to other methods tested.	Benefits of two-pass read mapping were revealed. Remaining challenges for RNA-seq alignment were identified: reduce false intron discovery rate, benefits of unbiased use of gene annotation, accurate placement of mismatches and indels.	Challenge Publication[19] Metrics for evaluating RNA-seq aligners. Open-source codebase, test data and program output available in the public domain at http://www.gencodegenes.org/rgasp_archive.html

In format provided by Saez-Rodriguez *et al.* (doi:10.1038/nrg.2016.69)

RGASP, RNA-seq transcript assembly[20] (2011) 14 computational methods 25 protocol variants	Identification and quantification of transcript isoforms based on RNA-seq data, assessed against well-curated reference genome annotation	GS: RNA-seq and NanoString data. Scoring: Several metrics rather specific to transcript assembly problem. Exon level: Precision and recall. Transcript level: percentage of reported splice transcripts. Gene level: Matching of at least one correct isoform in the given locus	Results were modest. Short-read sequencing limitations resulted in serious computational challenges in transcript reconstruction and quantification. For most transcripts, many of the constituent exons were not detected. No single protocol had a satisfactory performance at all metrics.	AUGUSTUS, GSTRUCT and Transomics demonstrated high precision. mGene exhibited diminished performance on human RNA-seq data, suggesting that method performance can depend on the organism under study.	Transcript assembly remains an outstanding challenge for whole-transcriptome shotgun sequencing. The study revealed that accuracy can be substantially improved by combining RNA-seq data with analysis of the genome sequence.	Challenge Publication[20] Metrics for evaluating transcript reconstruction methods. Open-source codebase, test data and program output available in the public domain at http://www.gencodegenes.org/rgasp_archive.html
ICGC-TCGA DREAM Somatic Mutation Calling (SMC) Challenge[21] (2012) 400 registrants 40 teams	Identify cancer-associated somatic mutations (single nucleotide variants (SNVs) and structural variants) from whole-genome next-generation sequencing data. Simulated data and patient data were provided.	Simulated Leaderboard Rounds: GS: *in silico* genomes. Scoring: sensitivity, specificity and balanced accuracy. Real Tumour Final Round: predictions were based on validation experiments based on the submitted predictions.	The leaderboard played a critical role. Teams were able to rapidly improve, particularly in precision, once they had an initial performance estimate. This suggests that real-time feedback can yield improved methods with low risk of overfitting.	Consensus model from the first three simulated data rounds resulted in a 'meta' algorithm that is far superior to any single algorithm used in genomic data analysis to date, highlighting the importance of considering a wisdom of crowds approach.	This challenge was useful to compare and promote innovation in methods for cancer somatic mutation calling. The new tool 'Bam Surgeon' used in this Challenge to simulate tumour genomes was tested and improved with input from participants.	Challenge Publication[21] 10 Patient-derived tumour–normal paired genomes from prostate and pancreatic cancers. Living benchmarks leaderboards open indefinitely to allow rapid comparison of methods. Simulator of a tumour genome, Bam Surgeon, is open source.

Abbreviations:AML, acute myeloid leukaemia; ALS, amyotrophic lateral sclerosis; AUPR, Area Under the Precision-Recall curve; AUROC, Area Under the Receiving Operating Characteristics curve; AMDA, Critical Assessment of Massive Data Analysis; CI, Concordance Index; DREAM, Dialogue for Reverse Engineering Assessment and Methods; FlowCAP, Flow Cytometry Critical Assessment of Population Identification Methods; GRN, gene regulatory network; GS, Gold Standard; GSNAP, Genomic Short-read Nucleotide Alignment Program; HPN, Heritage Provider Network; ICGC, International Cancer Genome Consortium; NCI, US National Cancer Institute; RGASP, RNA-seq Genome Annotation Assessment Project; RMSD, Root Mean Square Deviation; STAR, Spliced Transcripts Alignment to a Reference; SubC, SubChallenges; SVM, support vector machine; TCGA, The Cancer Genome Atlas; TF, transcription factor; TNF, tumour necrosis factor.

In format provided by Saez-Rodriguez *et al.* (doi:10.1038/nrg.2016.69)

References for supplementary information S2 (table)

1. Marbach, D. *et al.* Wisdom of crowds for robust gene network inference. *Nat. Methods* 9, 796–804 (2012).
2. Weirauch, M. T. *et al.* Evaluation of methods for modeling transcription factor sequence specificity. *Nat. Biotechnol.* **31,** 126–134 (2013).
3. Meyer, P. *et al.* Inferring gene expression from ribosomal promoter sequences, a crowdsourcing approach. *Genome Res.* **23,** 1928–1937 (2013).
4. Meyer, P. *et al.* Network topology and parameter estimation: from experimental design methods to gene regulatory network kinetics using a community based approach. *BMC Syst. Biol.* **8,** 13 (2014).
5. Hill, S. M. *et al.* Inferring causal molecular networks: empirical assessment through a community-based effort. *Nat. Methods* (2016). doi:10.1038/nmeth.3773
6. Aghaeepour, N. *et al.* Critical assessment of automated flow cytometry data analysis techniques. *Nat. Methods* **10,** 228–238 (2013).
7. Küffner, R. *et al.* Crowdsourced analysis of clinical trial data to predict amyotrophic lateral sclerosis progression. *Nat. Biotechnol.* 33, 51–57 (2015).
8. Margolin, A. A. *et al.* Systematic Analysis of Challenge-Driven Improvements in Molecular Prognostic Models for Breast Cancer. *Sci. Transl. Med.* **5,** 181re1–181re1 (2013).
9. Cheng, W.-Y., Ou Yang, T.-H. & Anastassiou, D. Development of a prognostic model for breast cancer survival in an open challenge environment. *Sci. Transl. Med.* **5,** 181ra50 (2013).
10. Allen, G. I. *et al.* Crowdsourced estimation of cognitive decline and resilience in Alzheimer's disease. *Alzheimers. Dement.* (2016). doi:10.1016/j.jalz.2016.02.006
11. Costello, J. C. *et al.* A community effort to assess and improve drug sensitivity prediction algorithms. *Nat. Biotechnol.* (2014).
12. Bansal, M. *et al.* A community computational challenge to predict the activity of pairs of compounds. *Nat. Biotechnol.* **32,** 1213–1222 (2014).
13. Uehara, T. *et al.* The Japanese toxicogenomics project: application of toxicogenomics. *Mol. Nutr. Food Res.* **54,** 218–227 (2010).
14. Eduati, F. *et al.* Prediction of human population responses to toxic compounds by a collaborative competition. *Nat. Biotechnol.* **33,** 933–940 (2015).
15. Chen, Y.-C. *et al.* A probabilistic model to predict clinical phenotypic traits from genome sequencing. *PLoS Comput. Biol.* **10,** e1003825 (2014).
16. Earl, D. *et al.* Assemblathon 1: a competitive assessment of de novo short read assembly methods. *Genome Res.* **21,** 2224–2241 (2011).
17. Bradnam, K. R. *et al.* Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *Gigascience* **2,** 10 (2013).
18. Earl, D. *et al.* Alignathon: a competitive assessment of whole-genome alignment methods. *Genome Res.* **24,** 2077–2089 (2014).
19. Engström, P. G. *et al.* Systematic evaluation of spliced alignment programs for RNA-seq data. *Nat. Methods* 10, 1185–1191 (2013).
20. Steijger, T. *et al.* Assessment of transcript reconstruction methods for RNA-seq. *Nat. Methods* 10, 1177–1184 (2013).
21. Ewing, A. D. *et al.* Combining tumor genome simulation with crowdsourcing to benchmark somatic single-nucleotide-variant detection. *Nat. Methods* **12,** 623–630 (2015).