Science Translational Medicine
AAAS

Systematic Analysis of Challenge-Driven Improvements in Molecular Prognostic Models for Breast Cancer

Adam A. Margolin *et al.*

Sci Transl Med **5**, 181re1 (2013);
DOI: 10.1126/scitranslmed.3006112

Science Translational Medicine (print ISSN 1946-6234; online ISSN 1946-6242) is published weekly, except the last week in December, by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005.

CROWDSOURCING

Systematic Analysis of Challenge-Driven Improvements in Molecular Prognostic Models for Breast Cancer

Adam A. Margolin,[1*†] Erhan Bilal,[2†] Erich Huang,[1,3,4†] Thea C. Norman,[1] Lars Ottestad,[5] Brigham H. Mecham,[1,6] Ben Sauerwine,[7] Michael R. Kellen,[1] Lara M. Mangravite,[1] Matthew D. Furia,[1,8] Hans Kristian Moen Vollan,[5,9,10,11] Oscar M. Rueda,[11] Justin Guinney,[1] Nicole A. Deflaux,[1] Bruce Hoff,[1] Xavier Schildwachter,[1] Hege G. Russnes,[9,10,12] Daehoon Park,[13] Veronica O. Vang,[9,10] Tyler Pirtle,[7] Lamia Youseff,[7] Craig Citro,[7] Christina Curtis,[14] Vessela N. Kristensen,[9,10,15] Joseph Hellerstein,[7] Stephen H. Friend,[1*] Gustavo Stolovitzky,[2] Samuel Aparicio,[16,17,18†] Carlos Caldas,[11,19,20†] Anne-Lise Børresen-Dale[9,10†]

Although molecular prognostics in breast cancer are among the most successful examples of translating genomic analysis to clinical applications, optimal approaches to breast cancer clinical risk prediction remain controversial. The Sage Bionetworks–DREAM Breast Cancer Prognosis Challenge (BCC) is a crowdsourced research study for breast cancer prognostic modeling using genome-scale data. The BCC provided a community of data analysts with a common platform for data access and blinded evaluation of model accuracy in predicting breast cancer survival on the basis of gene expression data, copy number data, and clinical covariates. This approach offered the opportunity to assess whether a crowdsourced community Challenge would generate models of breast cancer prognosis commensurate with or exceeding current best-in-class approaches. The BCC comprised multiple rounds of blinded evaluations on held-out portions of data on 1981 patients, resulting in more than 1400 models submitted as open source code. Participants then retrained their models on the full data set of 1981 samples and submitted up to five models for validation in a newly generated data set of 184 breast cancer patients. Analysis of the BCC results suggests that the best-performing modeling strategy outperformed previously reported methods in blinded evaluations; model performance was consistent across several independent evaluations; and aggregating community-developed models achieved performance on par with the best-performing individual models.

INTRODUCTION

Breast cancer is the leading female malignancy in the world (*1*) and is one of the first malignancies for which molecular biomarkers have exhibited promise for clinical decision making (*2*–*5*). Biomarkers can be used to divide the disease into predictive (the likelihood that a patient responds to a particular therapy) or prognostic (a patient's risk for a defined clinical endpoint independent of treatment) subcategories. Such molecular subcategorization highlights the possibilities for precision medicine (*6*), in which biomarkers are leveraged to identify disease taxonomies that distinguish biologically relevant groupings beyond standard clinical measures and can potentially inform treatment strategies.

A decade after early achievements in the development of prognostic molecular classifiers (called signatures) of breast cancer based solely on gene expression analysis (*4*, *5*, *7*), a large number of signatures proposed as markers of clinical risk prediction either fail to surpass the performance of conventional clinical covariates or await meaningful prospective validation [for example, the MINDACT (*8*, *9*) and TAILORx/RxPONDER (*10*) trials]. Although gene expression–based breast cancer prognostic tests have been successfully implemented in routine clinical use, the application of molecular data to guide clinical decision making remains controversial (*11*).

Slow progress to evolve useful molecular classifiers may relate to poor study design, inconsistent findings, or improper validation studies (*12*). Data and code that underlie a potential new disease classifier are often unavailable for diligence. In addition, the rigor and objectivity of assessing molecular models are confounded by the tendency of data generation, data analysis, and model validation to be combined within the same study. This leads to the "self-assessment trap," in which the desire to demonstrate improved performance of a researcher's own methodology may cause inadvertent bias in elements of study design, such as data set selection, parameter tuning, or evaluation criteria (*13*).

[1]Sage Bionetworks, 1100 Fairview Avenue North, MS: M1-C108, Seattle, WA 98109, USA. [2]Functional Genomics and Systems Biology, IBM Computational Biology Center, P. O. Box 218, Yorktown Heights, NY 10598, USA. [3]Institute for Genome Sciences & Policy, Duke University, Durham, NC 27708, USA. [4]Department of Surgery, Duke University School of Medicine, Durham, NC 27710, USA. [5]Department of Oncology, Division of Cancer, Surgery and Transplantation, Oslo University Hospital, 0450 Oslo, Norway. [6]Trialomics, LLC, Seattle, WA 98103, USA. [7]Google Inc., 651 North 34th Street, Seattle, WA 98103, USA. [8]Genomics Institute of the Novartis Research Foundation, San Diego, CA 92121, USA. [9]Department of Genetics, Institute for Cancer Research, Oslo University Hospital, The Norwegian Radium Hospital, Montebello, 0310 Oslo, Norway. [10]K.G. Jebsen Centre for Breast Cancer Research, Institute for Clinical Medicine, Faculty of Medicine, University of Oslo, 0313 Oslo, Norway. [11]Cancer Research UK, Cambridge Research Institute, Li Ka Shing Centre, Robinson Way, Cambridge CB2 0RE, UK. [12]Department of Pathology, Oslo University Hospital, 0450 Oslo, Norway. [13]Department of Pathology, Drammen Hospital, Vestre Viken HF, 3004 Drammen, Norway. [14]Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, CA 90033, USA. [15]Department of Clinical Molecular Oncology, Division of Medicine, Akershus University Hospital, 1478 Ahus, Norway. [16]Molecular Oncology, British Columbia Cancer Research Centre, Vancouver, British Columbia V5Z 1L3, Canada. [17]Department of Pathology and Laboratory Medicine, University of British Columbia, Vancouver, British Columbia V6T 1Z4, Canada. [18]Genome Sciences Centre, BC Cancer Agency, 675 West 10th Avenue, Vancouver, British Columbia V5Z 1L3, Canada. [19]Cambridge Breast Unit, Addenbrooke's Hospital, Cambridge University Hospital NHS Foundation Trust and NIHR Cambridge Biomedical Research Centre, Cambridge CB2 2QQ, UK. [20]Cambridge Experimental Cancer Medicine Centre, Cambridge CB2 0RE, UK.
*Corresponding author. E-mail: margolin@sagebase.org (A.A.M.); friend@sagebase.org (S.H.F.)
†These authors contributed equally to this work.

A potential response to such problems is leveraging the Internet, social media, and cloud computing technologies to make it possible for physically distributed researchers to share and analyze the same data in real time and collaboratively iterate toward improved models based on predefined objective criteria applied in blinded evaluations. The Netflix competition (*14*) and X-Prize (*15*) have successfully demonstrated that crowdsourcing novel solutions for data-rich problems and technology innovation is feasible when substantial monetary incentives are offered to engage the competitive instincts of a community of analysts and technologists. Alternatively, a game-like environment has been used to seek solutions to biological problems, such as protein (*16*) and RNA folding (http://eterna.cmu.edu/web), thereby appealing to a vast community of gamers. Others have demonstrated that both fee-for-service [Kaggle (*17*) and Innocentive (*18*)] and open-participation computational biology challenges (*19*) [CASP (*20*) and CAFA (*21*)] can be used as potential new research models for data-intensive science. Despite the wide breadth of areas covered by these competitions, a key finding is that the best models from a competition usually outperform analogous models generated using more traditional isolated research approaches (*14*, *17*).

Building on the success of crowdsourcing efforts such as DREAM to solve important biomedical research problems, we developed and ran the Sage Bionetworks–DREAM Breast Cancer Prognosis Challenge (BCC) to determine whether predefined performance criteria, real-time feedback, transparent sharing of source code, and a blinded final validation data set could promote robust assessment and improvement of breast cancer prognostic modeling approaches. The BCC was designed to make use of the METABRIC data set, which contains nearly 2000 breast cancer samples with gene expression and copy number data and clinical information. The availability of such a large data set affords the statistical power required to assess the robustness of performance of many models evaluated in independent tests, but is subject to the trade-off of using (overall) survival time in a historical cohort as the clinical endpoint, rather than potential endpoints that could be driven by clinically actionable criteria, such as response to targeted therapies. Therefore, the aim of this Challenge was not direct clinical deployment of a full-fledged suite of complex biomarkers. Indeed, we expect this study to serve as a pilot that lays the groundwork for future breast cancer challenges designed at the outset to answer clinically actionable questions. With this in mind, the BCC resulted in the development of predictive models with better performance than standard clinicopathological measures for prediction of overall survival (OS). The performance of these models was highly consistent across multiple blinded evaluations, including a novel validation cohort generated specifically for this Challenge.

RESULTS

The BCC included 354 registered participants from more than 35 countries. Participants were tasked with developing computational models that predict OS of breast cancer patients based on clinical information (for example, age, tumor size, and histological grade; see Table 1), mRNA expression data, and DNA copy number data. The BCC used genomic and clinical data from a cohort of 1981 women diagnosed with breast cancer (the METABRIC data set) (*22*) and provided participants with authorized Web access to data from 1000 samples as a training data set, and held back the remaining samples as a test data set (see Materials and Methods for more details on the Challenge design). Participants used the data to train computational models on their own standardized virtual machine (donated to the Challenge by Google) and submitted their trained models to the Synapse computational platform (*23*) as an R binary object (*24*) and rerunnable source code, where they were immediately evaluated. The predictive value of each model was scored by calculating the concordance index (CI) of predicted death risk compared to OS in a held-out data set, and the CIs were posted on a real-time leaderboard (http://leaderboards.bcc.sagebase.org). The CI is a standard performance measure in survival analysis that quantifies the quality of ranking risk predictors with respect to survival (*25*). In essence, given two randomly drawn patients, the CI represents the

Table 1. Clinical characteristics of METABRIC and OsloVal data sets. NA, not available.

Categories	METABRIC	OsloVal
Cohort size	1981	184
Age, years (%)		
≤50	21.4	33.1
50–60	22.5	18.5
≥60	56.1	48.4
Tumor size, cm (%)		
≤2	43.3	38.0
2–5	48.2	42.4
≥5	7.5	7.1
NA	1.0	12.5
Node status (%)		
Node negative	52.3	49.5
1–3 nodes	31.4	21.2
4–9 nodes	11.4	9.8
≥10 nodes	4.6	8.2
NA	0.3	11.3
ER status (%)		
ER^+	76.3	60.9
ER^-	23.7	39.1
PR status (%)		
PR^+	52.7	21.2
PR^-	47.3	78.8
HER2 copy status (%)		
HER2 amplification	22.1	13.6
HER2 neutral	72.6	86.4
HER2 loss	5.0	0.0
NA	0.3	0.0
Tumor grade (%)		
1	8.6	6.5
2	39.1	37.0
3	48.1	30.4
NA	4.2	26.1

probability that a model will correctly predict which of the two patients will experience an event before the other (for example, a CI of 0.75 for a model means that if two patients are randomly drawn, the model will order their survival correctly three of four times).

Throughout the 3-month orientation and training phases of the Challenge (phases 1 and 2, respectively), participants collectively submitted more than 1400 predictive models, 1400 of which successfully executed from the submitted binary object and were assigned CI scores in the test data set (Fig. 1). One unique characteristic of this Challenge with respect to previous biomedical research Challenges was that each participant's source code was available for others to view and adapt in new models. At the end of the training phase, participants were given the opportunity to train five models each on all 1981 METABRIC samples. The overall Challenge was determined in the final phase (phase 3) by assessing up to five models per participant or team (see table S3 for a listing of those members of the Breast Cancer Challenge Consortium who submitted at least one model to phase 3 and requested to have their name and affiliation listed) against a newly generated breast cancer data set consisting of genomic and clinical data from 184 women diagnosed with breast cancer (the "OsloVal" data set) (see Materials and Methods for details on how to access the METABRIC and OsloVal data sets). The participant or team with the best-performing model in the new data set was invited to publish an article about the winning model in *Science Translational Medicine*, provided it exceeded the scores of preestablished benchmark models, including the first-generation 70-gene risk predictor (*4*, *7*) and the best model developed by a group of expert researchers in a precompetition (*26*). This pioneering mechanism of Challenge-assisted peer review assessed performance metrics in a blinded validation test, and these results were the foremost criterion for publication in the journal.

The METABRIC and the OsloVal data sets were comparable in terms of their clinical characteristics (Table 1). Roughly three-quarters of the patients were estrogen receptor–positive (ER^+), and about half were lymph node–negative (LN^-). The patients from the METABRIC and OsloVal cohorts received combinations of hormonal, radio-, and chemotherapy, and none were treated with more modern drugs such as trastuzumab, which specifically targets the human epidermal growth factor receptor 2/neu (HER2/neu) receptor pathway (*27*).

In phase 2 of the Challenge, participants trained on 1000 samples and obtained real-time feedback on model scores (CIs) in a held-out 500-sample test set (phase 1 results are displayed on the original leaderboard and are not discussed here). A reference model using a random survival forest (*28*) trained on the clinical covariates and genomic data was provided as a baseline. Models submitted by Challenge participants quickly exceeded the performance of this baseline model and steadily improved over time (Fig. 2A), although we note that the improvement was modest compared to the baseline model (Fig. 2A, inset).

At the end of phase 2, all models in the leaderboard were evaluated against 481 hidden validation samples from the METABRIC data set. Given that more than 1400 models were submitted, there was concern that improvement in model scores resulted from overfitting to the test

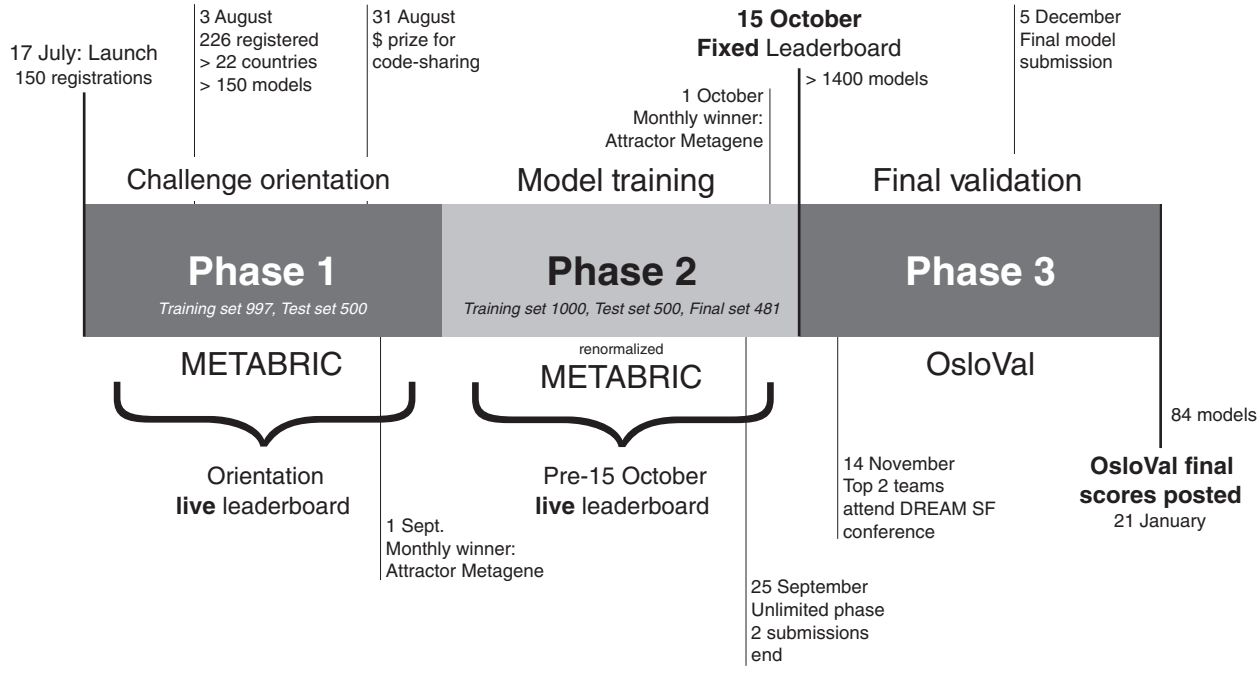

Fig. 1. Timeline and phases of BCC. At initiation (phase 1), a subset of the METABRIC data set was provided along with orientation on how to use the Synapse platform for accessing data and submitting models and source code. Phase 2 provided a new randomization of samples, to eliminate biases in the distribution of clinical variables across training and test data, and renormalization of METABRIC mRNA expression and DNA copy number data, to reduce batch effects and harmonize data with the OsloVal data used in phase 3. During phase 2, there was a live "pre–15 October 2012" leaderboard that provided real-time scores for each submission against the held-out test set of 500 samples. At the conclusion of phase 2 on 15 October 2012, all models in the leaderboard were tested against the remaining held-out 481 samples. In the final validation round (phase 3), participants were invited to retrain up to five models on the entire METABRIC data set. Each model was then assigned a final CI score and consequently a rank based on the model's performance against the independent OsloVal test set.

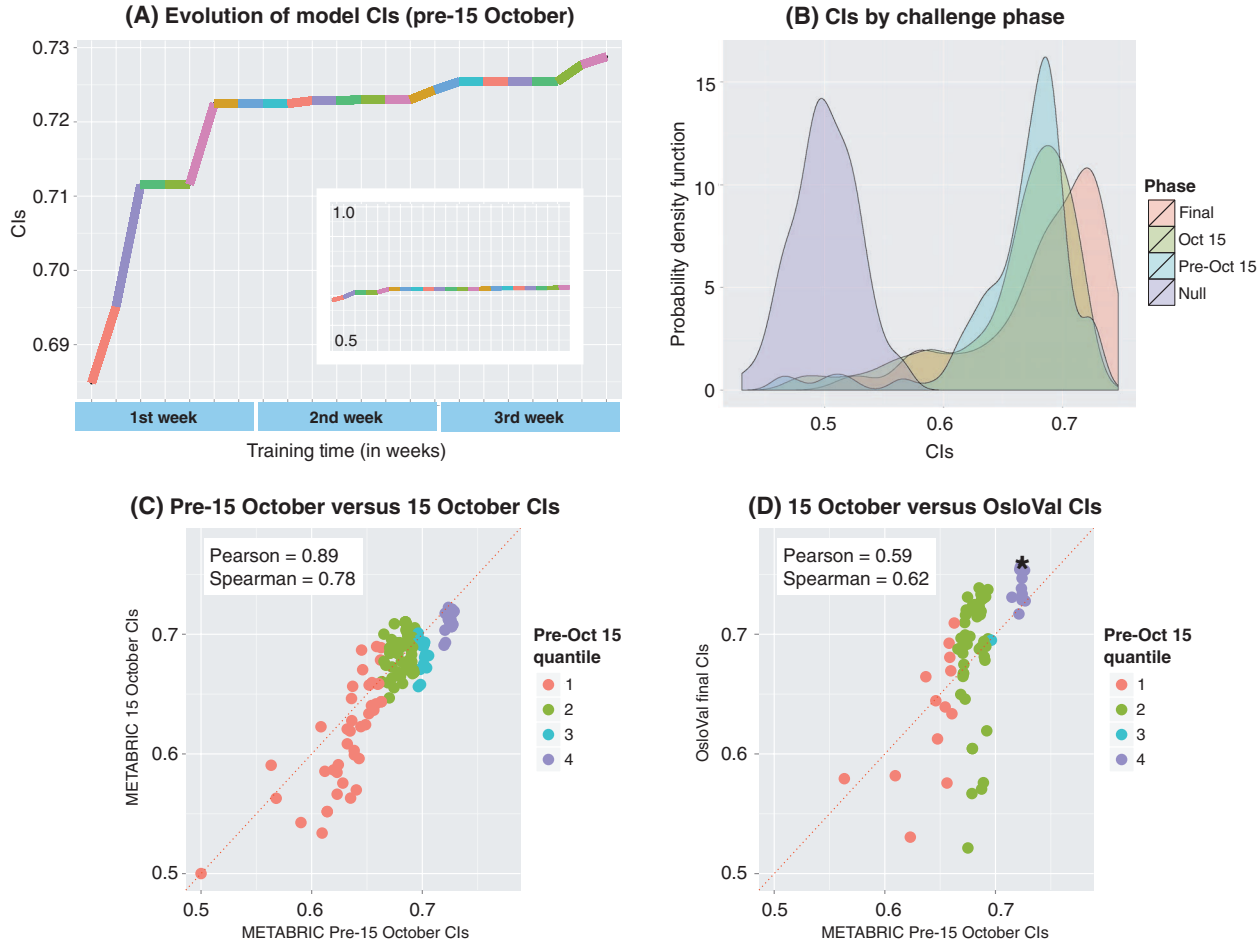

Fig. 2. BCC through time. (**A**) During phase 2, the highest-scoring model scores were recorded for each date until the leaderboard was closed. Each colored segment represents a top-scoring team at any given point for the period extending from late September 2012 until the final deadline of phase 2 (15 October 2012). The plot records only the times when there was an increase in the best score, whereas the teams that achieved this score are labeled with different colors. The sequence of colors highlights an important aspect of the real-time feedback, where teams were encouraged to improve their models after being bested on the leaderboard by another team. Inset, the same plot with a *y*-axis scale ranging from of 0.5 to 1.0 maximum CI. (**B**) Probability density function plots of model scores posted on the live pre–15 October leaderboard evaluated against (i) the first test set of 500 samples (blue), (ii) the second test set of 481 samples (yellow), and (iii) the OsloVal data set (red). The null hypothesis probability density, which corresponds to random predictions evaluated against the OsloVal data set, is shown in purple. (**C**) Scatter plot of pre–15 October 2012 model performance versus 15 October 2012 performance. Colors represent quantiles, meaning that the ordered data are divided into four equal groups numbered consecutively from the bottom-scoring models (1) to the top-scoring models (4) for pre–15 October model performance. (**D**) Scatter plot of pre–15 October 2012 model performance versus final OsloVal performance. Colors represent quantiles of pre–15 October model performance. Asterisk represents the highest-scoring submitted model.

set used to provide real-time feedback. However, the performance of most models in the new test set was consistent with the performance of the same models in the previous test set, with comparable score ranges (Fig. 2, B and C; Pearson correlation: 0.90). A similar outcome was observed for the five models from every team that were evaluated against the OsloVal data set, after being trained on the 1981 samples in the METABRIC data set. Again, there was little evidence of overfitting when compared to the previous METABRIC test scores (Fig. 2D; Pearson correlation: 0.59), especially for models ranked in the top quantile.

After participants trained and selected their final models, and after eliminating the ones that could not be evaluated because of run time errors, a total of 83 models were assessed and scored against the newly generated OsloVal data set. The winning team had three similar models that performed consistently better than all other scored models. The robustness of the final ranking was evaluated by sampling, without replacement, 80% of the validation set 100 times. Figure 3 shows box plots of the rankings obtained by each model across all trials ordered by the initial scores on all 184 OsloVal samples. The top three models belong to the same team and are ranked significantly better than the rest. The top model achieved a CI of 0.7561 [$P = 5.1 \times 10^{-28}$ compared to the fourth-ranked model, by Wilcoxon rank-sum test (*29*)].

The BCC top-scoring model also significantly outperformed the top model developed during a pilot precompetition phase (*26*) in which BCC organizers tested 60 different models based on state-of-the-art

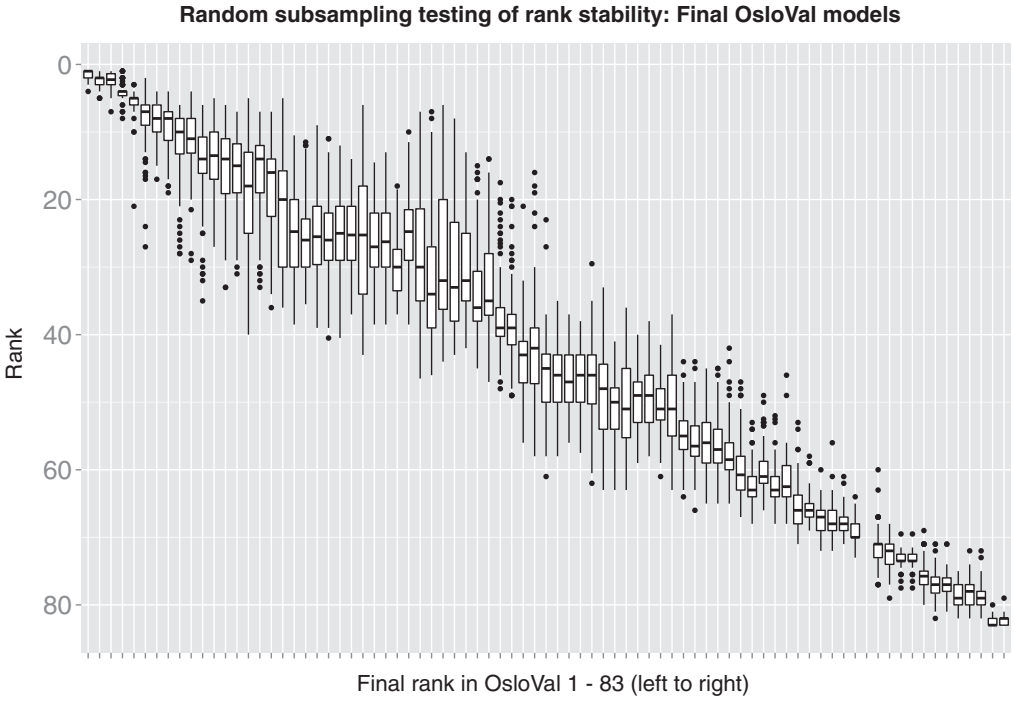

Fig. 3. Rank stability of final models. The OsloVal test data were randomly subsampled 100 times using 80% of the samples. Model rank was recalculated at each iteration. Models are ordered by their final posted leaderboard score (*P* values for the top three models, which were submitted by the same team, versus the fourth place model were as follows = 5.1×10^{-28}, 1.8×10^{-22}, and 1.7×10^{-20} by Wilcoxon rank-sum tests). With these box plots, the middle horizontal line represents the median, the upper whisker extends to the highest value within a 1.5× distance between the first and third quantiles, and the lower whisker extends to the lowest value within a 1.5× distance. Data beyond the ends of the whiskers are outliers plotted as points.

machine learning approaches and clinical feature selection strategies. The best model from the precompetition used a random survival forest trained on the clinical feature data in addition to a genomic instability index derived from the copy number data.

The top-scoring model from the precompetition would have ranked as the sixth best model in the Challenge and achieved a CI of 0.7408 (Wilcoxon paired test, $P = 4.33 \times 10^{-18}$ compared to the winning model from the Challenge) when trained on the full METABRIC data set and evaluated in the OsloVal data. For comparison, a research version of a 70-gene risk signature (*4*) was also evaluated in the OsloVal cohort and achieved a CI of 0.60. In addition, two more test models were developed that included only the clinical covariates available for the two data sets (listed in tables S1 and S2). The first model was based on boosted regression (*30*) and achieved a CI of 0.7001 on the validation data set, whereas the second model used random forest regression (*28*) and achieved a score of 0.6964. The winning Challenge model achieved a score of 0.7562, significantly higher than the two clinical-only models (Wilcoxon paired test, $P = 6.1 \times 10^{-32}$ for both).

Meta-analyses of predictions submitted to past DREAM Challenges have systematically demonstrated (*31*–*34*) that the ensemble predictions resulting from the aggregation of the predictions of some or all the models usually perform similarly or even better than the best model. This phenomenon has been called the "wisdom of the crowds" and highlights one of the advantages of enabling research communities to work collaboratively to analyze the same data sets. The wisdom of the crowds was also at play in BCC (Fig. 4). For both cohorts, participants' predictions were aggregated by calculating a community prediction formed by taking the average predicted rank of each patient across top *n* models for $n = 1 \ldots 83$ (Fig. 4, A and B). Even when adding very poor predictions to the aggregate, the resulting score was robust and comparable with the top models. Robust predictors were also achieved by constructing community scores based on random subsamples of models (Fig. 4, C and D).

We evaluated model performance to determine whether there were specific clinical cases that are inherently more difficult to predict in terms of prognosis. We conducted analysis of variance (ANOVA) analyses by separating the patients from the OsloVal cohort based on the following clinical variables: age (≤50 years versus >50 years), tumor size (≤2 cm versus >2 cm), tumor grade (grades 1, 2, and 3), LN status (LN^-, 1 to 3 positive LNs, 4 to 9 positive LNs, >9 positive LNs), ER status (ER^+ versus ER^-), progesterone receptor status (PR^+ versus PR^-), HER2/neu receptor amplification status ($HER2^+$, $HER2^-$), and OS (<5 years, 5 to 10 years, or >10 years). Of all variables considered (Fig. 5), tumor grade, LN status, OS, age, and tumor size were significantly associated with model performance (*F* test *P* values: 1.7×10^{-36}, 1.1×10^{-38}, 8.0×10^{-26}, 9.0×10^{-7}, and 5.0×10^{-5}, respectively).

Breast cancer patients with high-grade tumors and large numbers of positive LNs (>9) were associated with low CI scores (Fig. 5, B and C) and explain more than 40% of CI variance (Fig. 5A), suggesting that the OS of breast cancer patients with aggressive tumors is harder to predict. By contrast, there was no significant association between model performance and ER, PR, or HER2 status (Fig. 5E).

DISCUSSION

The BCC was an exercise in crowdsourcing that constitutes an open distributed approach to develop predictive models with the future potential to advance precision medicine applications. By creating a large, standardized breast cancer genotypic and phenotypic resource readily accessible via Web services and a common cloud computing environment, we were able to explore whether the open sharing of predictors of disease within a Challenge environment encourages the sharing of models and ideas before publication and whether decoupling of data creation, data analysis, and evaluation would minimize analytical and self-assessment biases.

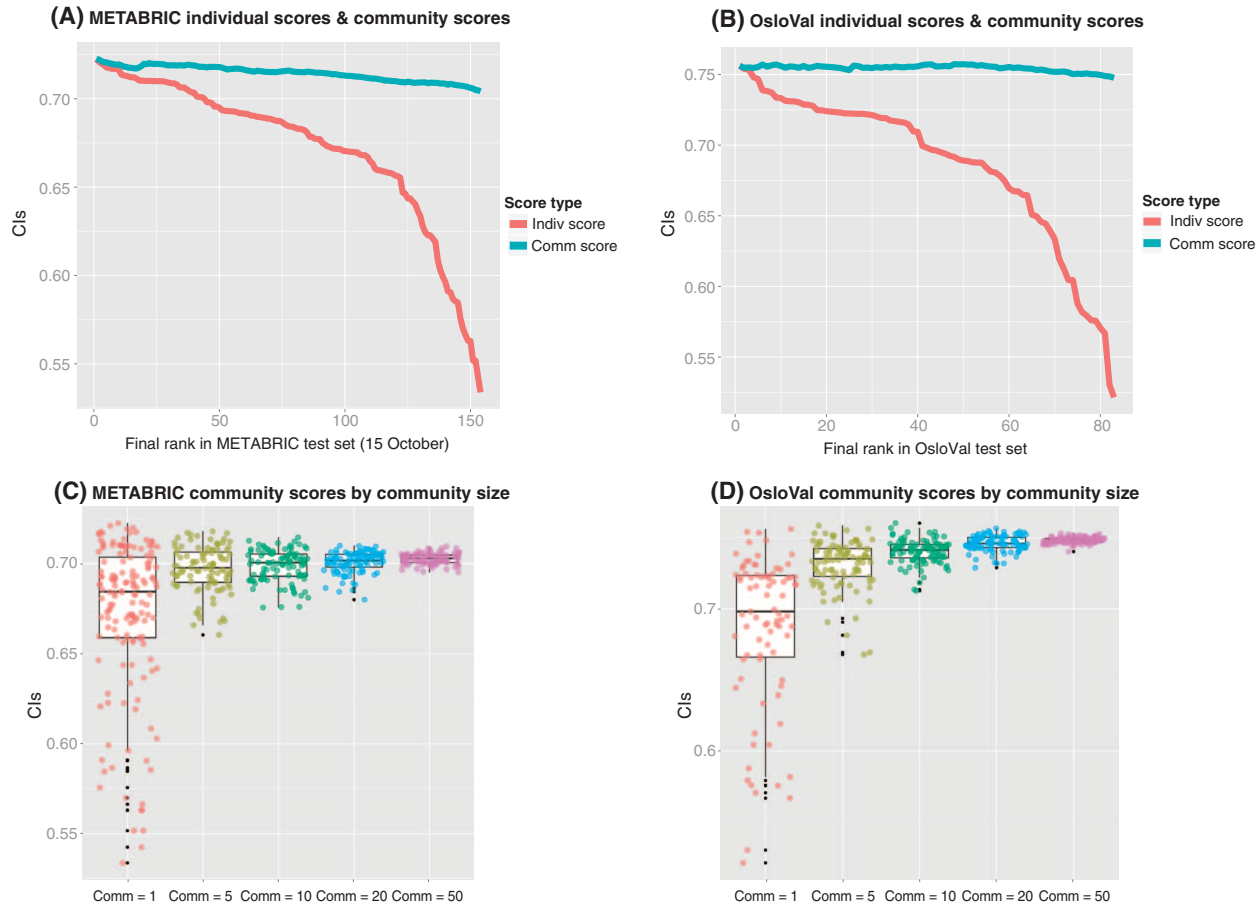

Fig. 4. Individual and community scores for METABRIC and OsloVal. (**A**) Individual model scores are ordered by their rank on the pre–15 October 2012 METABRIC leaderboard (red line). For each model rank (displayed on the *x* axis), the blue line plots the aggregate model score based on combining all models less than or equal to the given rank. (**B**) Individual and aggregate model scores based on evaluation in the OsloVal data set. (**C**) Individual model scores (that is, community = 1) from the pre–15 October 2012 METABRIC leaderboard (http://leaderboards.bcc.sagebase.org/pre_oct15/index.html) are plotted alongside the community aggregate scores obtained when 5, 10, 20, and 50 randomly chosen models were considered. (**D**) Individual model scores (that is, community = 1) from the final OsloVal leaderboard (http://leaderboards.bcc.sagebase.org/final/index.html) are plotted alongside the community aggregate scores obtained when 5, 10, 20, and 50 randomly chosen predictions were considered. The colors correspond to community size: red = 1, yellow = 5, green = 10, blue = 20, purple = 50.

Scientific conclusions from the Challenge

Improved model performance. The BCC results show that the best-performing model achieved significant CI improvements over currently available best-in-class methodologies, including the best model developed by a group of experts in a precompetition and a 70-gene risk signature. We note that the first-generation 70-gene risk signature used for comparison was designed as a binary risk stratifier for a specific patient subpopulation and should be viewed as a baseline and sanity check rather than a direct comparison. More significantly, the best-performing model (as well as slight variants of the same model submitted by the same BCC team) consistently outperformed all other approaches in three independent rounds of assessment and across multiple data sets. The top-scoring models used a methodology that minimized overfitting to the METABRIC training set by defining a "Metagene" feature space based on robust gene expression patterns observed in multiple external cancer data sets (*35*).

Robustness and generalizability of model performance. Our post hoc analysis suggests that the potential for training models overfit to the test set was not a significant confounding factor, even when participants were allowed to submit an unlimited number of models and obtain real-time scores for each submission. Comparison of the same models across multiple rounds of evaluation suggests a surprising degree of consistency between the unlimited submission phase and independent evaluations in the held-out METABRIC data and newly generated OsloVal data. This is especially remarkable because the OsloVal samples were collected from a different geographical location, by a different team, at a different time, and for a different purpose—in contrast to studies such as MAQC-II (*36*), in which the test sets and training sets were collected and processed by the same team and organization.

The distribution of CI scores improved with each round of evaluation, and the highest-scoring models in the OsloVal evaluation achieved higher CI scores than any of the 1400 models evaluated in either of the METABRIC phases. This apparent counterintuitive result can likely be explained by the fact that the average follow-up time for the OsloVal cohort is 4541 days, much longer than the average follow-up time of 2951 days for METABRIC. The rate of censored events is also

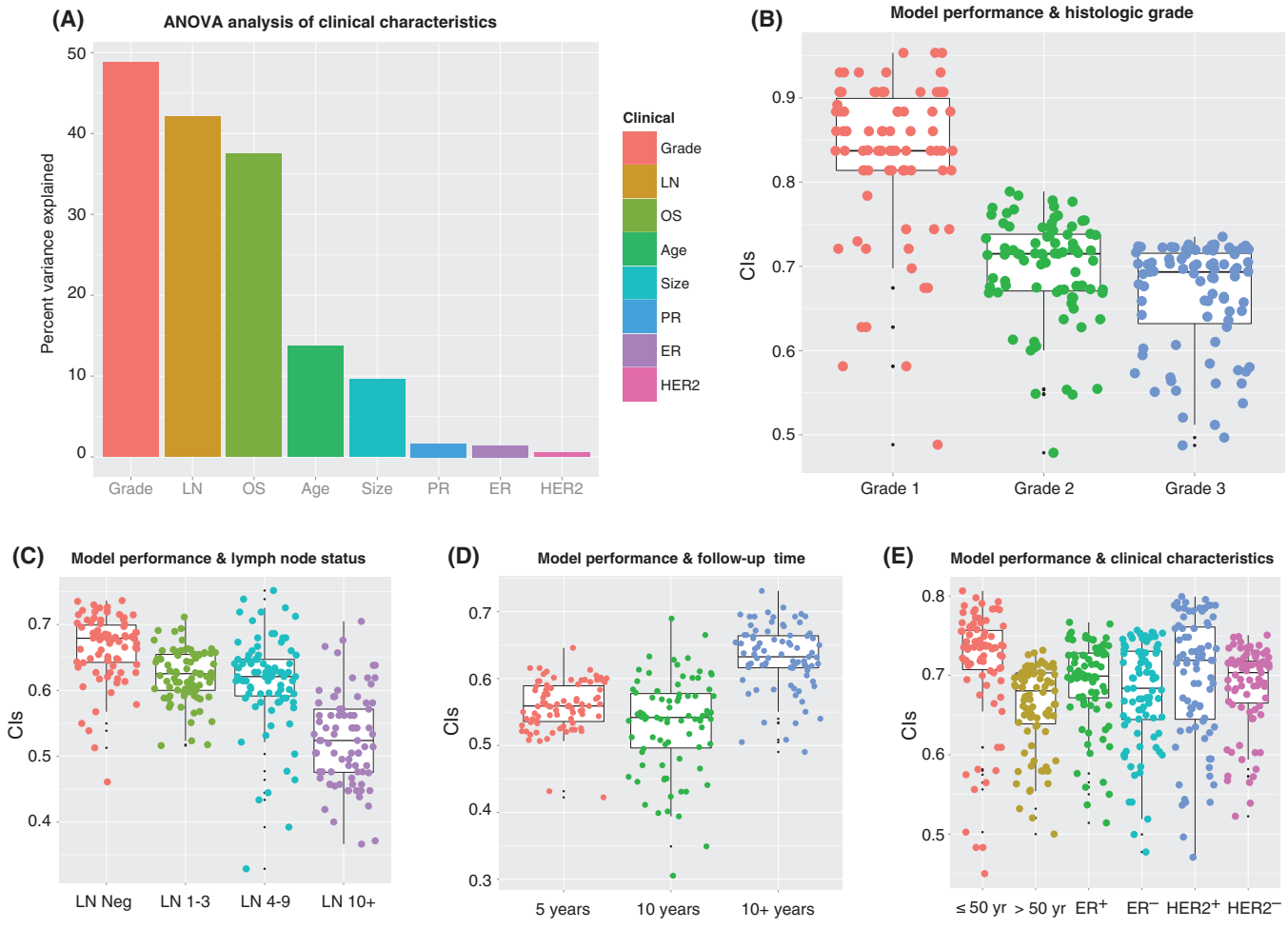

Fig. 5. Model performance and clinical characteristics. (**A**) Percentage of CI variance explained by each clinical variable. (**B**) CIs were calculated for OsloVal models according to subsets of patients by histological grade. (**C**) CIs were calculated for OsloVal models according to subsets of patients by LN status. (**D**) CIs were calculated for OsloVal models according to subsets of patients by follow-up time (OS). (**E**) CIs were calculated for OsloVal models according to subsets of patients by age, ER status, and HER2 status. Patients were divided into subsets according to each of the above clinical characteristics. Individual model predictions were generated for patients belonging to each subset, and the CI was calculated by comparison with the actual survival for each patient.

lower in OsloVal (27.7%) than in METABRIC (55.2%); therefore, the proportion of long-term survivors scored in OsloVal is larger than that in the METABRIC validation set. Because the long-time survivors are better predicted (Fig. 5D) than the short-term survivors, the CIs associated with the METABRIC data are biased to shorter-time survivors and lower CIs.

The consistency of model scores across our three evaluations does not conclusively rule out some degree of overfitting, and, as with any scientific study, the generality of our findings will be continuously refined through sustained scrutiny in subsequent studies throughout the research community. However, the consistency of model performance across independent evaluations provides strong initial evidence that the findings of our study are likely to generalize to unseen data.

Robust performance of community models. Consistent with results of previous DREAM Challenges (*31*–*34*), the current study suggests that community models, constructed by aggregating predictions across many models submitted by participants, achieve performance on par with the highest-scoring individual models, and this high performance is remarkably robust to the inclusion of many low-scoring models into the ensemble. This result suggests that crowdsourcing a biomedical prediction question as a Challenge and using the community prediction as a solution is a fast and sound strategy to attain a model with strong performance and robustness. Intuitively, such an approach leverages the best ideas from many minds working on the same problem. More specifically, different approaches that accurately model biologically relevant signals are likely to converge on similar predictions related to the true underlying biology, whereas errant predictions resulting from erroneous approaches are less likely to be consistent. Thus, ensemble models may amplify the true signals that remain consistent across multiple approaches while decreasing the effects of less-correlated errant signals.

Contributions of BCC to the community

Challenge design. In designing BCC to incentivize crowdsourced solutions that might improve breast cancer prognostic signatures, our aims included continuous participant engagement throughout the

Challenge, computational transparency, and rewarding of model generalizability. To accomplish these aims, respectively, we provided a framework that allowed users to make submissions and obtain real-time feedback on their performance, required submission of source code with each model, and provided multiple rounds of evaluation in independent data sets.

A common framework for comparing and sharing models. The model evaluation framework and CI scores calculated for models submitted during the Challenge provide a baseline set of model scores against which emerging tests, such as the PAM50 (*37*) risk of recurrence score and future prognostic models, may be compared.

The requirement for submission of publicly available source code provided an additional level of transparency compared to the typical Challenge design, which requires users to submit a prediction vector output from their training algorithm. We envision that sharing source code from multiple related predictive modeling Challenges will give rise to a community-developed resource of predictive models that is extended and applied across multiple projects and that might facilitate cumulative scientific progress, in which subsequent innovations build off of previous ones. It was encouraging to discover that participants learned from both well-performing and poorer-performing models. For example, the top-performing team consistently used the discussion forum to share with the other teams the prognostic ability of their Metagene features. Some of the other teams used aspects of the best-performing team's code to improve their submissions, which in turn gave feedback to the best-performing team on the use of their methods by the other challenge participants.

Innovations from the community. Through the transparent code submission system and communication tools, such as a community discussion forum, the Challenge resulted in numerous examples of sharing and borrowing of scientific and technical insights between participants. At one point, we tested whether a cash incentive could be used to promote collaborative model improvement and offered a $500 incentive to any participant who could place atop the leaderboard by borrowing code submitted by another participant (in addition to $500 to the participant whose code was borrowed). In less than 24 hours, a participant achieved the highest-scoring model by combining the computational approaches of the previously highest-scoring model with his clinical insight of modeling LN status as a continuous, rather than binary, variable. Unanticipated innovations also emerged organically from the community, including an online game within a game (http://genegames.org/cure), in which a player and a computer avatar successively select genes as input features to a predictive model, until one model is deemed statistically significantly superior to the other in predicting survival in the held-out data sets. This game attracted 120 players within 1 week, who played more than 2000 hands.

Limitations and extensions for future Challenges

The design of BCC included a number of simplifying assumptions intended to define a tractable prediction problem and evaluation criterion. However, it would be beneficial to account for biological and analytical complexities that were obscured by simplifications made in our experimental design.

First, we evaluated all models on the basis of a single metric, CI, which represents the most widely used statistic for evaluating survival models. However, a more complete assessment of advantages and disadvantages of each model would include additional criteria, such as model run time, trade-offs between sensitivity and specificity, or metrics more closely tied to the clinical relevance of a prognosticator.

Second, our choice to evaluate survival predictions across all samples in the cohort may obscure identification of models with advantages in particular breast cancer subtypes.

Third, by providing participants with normalized data, we tested only for modeling innovations given predefined input data, but did not assess different methods for data quality control and standardization that could contribute substantially to model improvements. A useful future Challenge design may allow participants to submit alternative methods for preprocessing raw data.

Fourth, we chose to evaluate models on the basis of OS. Other clinical endpoints, such as progression-free survival, are not currently available in the data sets used in BCC but may represent better evaluation endpoints if made available in the future. We chose OS over the other clinical endpoint available in the data set, disease-specific (DS) survival, to be consistent with decisions (*38*) by regulatory agencies such as the U.S. Food and Drug Administration and the European Medicines Agency. However, because of informal feedback we received from participants—that use of DS survival yielded more accurate models—we reevaluated all models and supported a separate exploratory leaderboard based on this metric. The model performance on this leaderboard suggested that using DS rather than OS as the clinical endpoint yielded improved CI scores, increased correlation with molecular features, and decreased correlation with confounding variables such as age. However, the best-performing models were consistent across both metrics.

Fifth, although prognostic models are one translational question, future Challenges that focus more directly on inferring predictive models of response to therapy may more directly affect clinical decision making. A laudable goal for future Challenges would be to directly engage the patient community and provide means to submit their own samples, help define questions, work alongside Challenge participants, and provide more direct feedback on how Challenge results can yield insights able to translate to improved patient care. In addition, providing large high-quality data sets (often prepublication) to a "crowd" of analysts is not common practice in academia and industry. However, such contributions by data generators would provide the necessary substrate for running such Challenges, with a potentially high impact on biomedical discovery.

Sixth, although statistically significant, the BCC results show that the improvement of the best-performing model is moderate with respect to the score achieved by aggregating standard clinical information. Thus, whereas molecular prognostic models derived from BCC warrant further investigation into their clinical utility, our results also suggest a new benchmark for future predictive methods derived from incorporating clinical covariates into state-of-the-art ensemble methods such as boosting or random forests. Future Challenges also may investigate the use of additional types of genomic information.

Finally, the sharing of ideas enabled by requiring submissions as rerunnable source code may ironically inhibit the diversity of innovations, effectively encouraging a monoculture as the community converges on a local optimum, modifying and extending approaches with high performance in the early stages of feedback (*39*). Improvements for future Challenges may include short embargo periods before sharing source code, with release possibly associated with declaring

winners in stages of sub-Challenges with slightly modified data or prediction criteria. In addition, future Challenges that promote code sharing should establish well-defined criteria for assigning proper attribution (for example, in publication of winning models resulting from the Challenge) to all participants who made material intellectual contributions that were incorporated into the final winning model. More generally, it is important to develop a reward system that favors collaborative research practices that balance the currently prevalent winner-takes-all reward system.

Our results reinforce the trend from efforts such as CASP, DREAM, and others that Challenges incentivize collaboration and rapid learning, create iterative solutions in which one Challenge may feed into follow-up Challenges, motivate the generation of new data to answer clinically relevant questions, and provide the means for objective criteria to prioritize approaches likely to translate basic research into improved benefit to patients. We envision that expanding such mechanisms to facilitate broad participation from the scientific community in solving impactful biomedical problems and assessing molecular predictors of disease phenotypes could be integral to building a transparent, extensible resource at the heart of precision medicine.

MATERIALS AND METHODS

Challenge timeline

The BCC comprised three phases spanning a period of 3 months: (i) an orientation phase, (ii) a training phase, and (iii) a validation phase.

In phase 1, participants were provided mRNA and copy number aberration (CNA) data from 997 samples from METABRIC alongside the clinical data to train predictive models of patient OS. Samples contained in this training set corresponded to those used as a training set in the original publication of the METABRIC data set (*22*). During this phase of the Challenge (17 July to 22 September 2012), models were evaluated in real time against a held-out data set of 500 METABRIC samples, and a leaderboard was developed to display CI scores for each model, based on predicted versus observed survival times in these 500 held-out samples.

In phase 2 (25 September to 15 October 2012), lessons learned from the previous phase were implemented: The clinical variable "LN status" was changed from binary (positive or negative) values to integer values, representing the number of affected LNs; the gene expression and CNA data were renormalized (as described below) to better correct for batch effects; and the full cohort was randomly split into a new training and a new testing set to correct for sample biases from the original split (for example, in the original split, nearly all missing clinical covariates were in the test set). During this phase, models were trained on a training set of 1000 samples and evaluated in real time against a held-out test data set of 500 samples. The resulting scores were posted on a new leaderboard and, at the end of phase 2, were evaluated against the remaining held-out 481 samples. The winner of this phase was determined on the basis of evaluation in this second test set. All official scoring was performed using the OS endpoint, although based on requests from Challenge participants, we also configured a leaderboard that allowed participants to assess their model scores against DS survival as a secondary "unofficial" evaluation.

Before submission of their final models in phase 3, participants were given the opportunity to retrain their models on the entire METABRIC data set of 1981 (for convenience, with clinical variables reduced to only those present in the OsloVal data set). Furthermore, participants were asked to select a maximum of five models per team for assessment against the OsloVal data set. All participants who submitted a model in phase 3 are listed in table S3. If a team did not choose their preferred five models, their five top-scoring models were selected by default. These models were scored on the basis of the CI predicted versus observed ranks of survival times in the OsloVal data set. The overall winner of the Challenge was determined by the top CI score in OsloVal. The significance of the top-scoring models compared to the rest was assessed on the basis of scores for multiple random subsamples without replacement of 80% of patients from the OsloVal cohort. This process was repeated 100 times, and the resulting rankings were compared with a Wilcoxon rank-sum test (*29*).

Data governance

The data generators' institutional ethics committee approved use of the clinical data within the BCC. Expression and copy number data from METABRIC were made available to Challenge participants during the duration of phases 1 to 3. Data for clinical covariates from the METABRIC cohort had been made public previously (*40*). Each participant who accessed the data agreed to (i) use these data only for the purposes of participating in the Challenge and (ii) not redistribute the data. Data access permissions were revoked at the completion of each BCC phase, and participants were required to reaffirm their agreement to these terms to enter each phase.

The METABRIC and OsloVal data sets have been deposited in the Synapse database (https://synapse.prod.sagebase.org/#!Synapse:syn1710250) and will be available to readers for a 6-month "validation phase" of the BCC. Those interested in accessing these data for use in independent research are directed to the following links. Expression and CNA data from OsloVal are available through an open access mechanism. All of other data are available through a controlled access mechanism.

METABRIC: https://synapse.prod.sagebase.org/#!Synapse:syn1688369
OsloVal: https://synapse.prod.sagebase.org/#!Synapse:syn1688370

The final Challenge scoring was based on the CIs of models, which was the same metric used in the leaderboard at earlier phases of the Challenge. The CI was predefined from the beginning of the Challenge as the performance metric to be used at final scoring. For all final submissions, every effort was made by the Sage Bionetwork team to run the submitted code in the Synapse platform. However, because of different problems in the submitted code, some of the submissions did not run successfully. Each submitted model that could be successfully run yielded a final survival prediction for the OsloVal cohort. For each of these predictions, the CI was mathematically computed in an unambiguous way by a computer program, from which the performance ranking was generated in order of descending CI. All final models can be accessed at http://leaderboards.bcc.sagebase.org/final/index.html.

Data normalization

The Affymetrix Genome-Wide Human SNP 6.0 and Illumina HT12 Bead Chip data were normalized according to the supervised normalization of microarrays (snm) framework and Bioconductor package (*41*, *42*). Following this framework, models were devised for each data set that expressed the raw data as functions of biological and adjustment variables. The models were built and implemented through an iterative process designed to learn the identity of important variables defined by latent factors identified via singular value decomposition. Once these

variables were identified, we used the snm R package to remove the effects of the adjustment variables while controlling for the effects of the biological variables of interest.

For example, to normalize the METABRIC mRNA data, we used a model that included ER status as a biological variable and both scan date and intensity-dependent array effects as adjustment variables. The resulting normalized data consisted of the residuals from this model fit plus the estimated effects of ER status. For the relevant data sets, we list the biological and adjustment variables as follows: METABRIC mRNA: biological variable = ER status, adjustment variables = scan date and intensity-dependent array effects; METABRIC SNP: biological variable = none, adjustment variables = scan date and intensity-dependent array effects; OsloVal mRNA: biological variable = ER status, adjustment variables = Sentrix ID (*43*), intensity-dependent array effects; OsloVal SNP: biological variable = none, adjustment variables = scan date, intensity-dependent effects.

Summarization of probes to genes for the SNP6.0 copy number data was done as follows. First, probes were mapped to genes with information obtained from the pd.genomewidesnp.6 Bioconductor package (*44*). For genes measured by two probes, we defined the gene-level values as an unweighted average of the data from the two probes. For genes measured by a single probe, we defined the gene-level values as the data for the corresponding probe. For those measured by more than two probes, we devised an approach that weighted probes based on their similarity to the first eigengene as defined by taking a singular value decomposition of the probe-level data for each gene. The percent variance explained by the first eigengene was then calculated for each probe. The summarized values for each gene were then defined as the weighted mean, with the weights corresponding to the percent variance explained.

Compute resources

A total of 2000 computational cores in the Google Cloud were provided to participants for the ~5-month duration of the Challenge, corresponding to a maximum of 7.5 million core hours if used at capacity. Specifically, each participant was provisioned an 8-core, 16-GB RAM machine, preconfigured and tested with an R computing environment and required libraries used in the Challenge. Compute resources were provisioned to each participant for dedicated use throughout the Challenge, and once capacity was reached, resources assigned to inactive users were recycled to new registrants such that all active users were provisioned compute resources. Users were also allowed to work on their own computing resources.

OsloVal data generation

The OsloVal cohort consisted of fresh-frozen primary tumors from 184 breast cancer patients collected from 1981 to 1999 (148 from 1981 to 1989 and 36 from 1994 to 1999) at the Norwegian Radium Hospital. Tumor material collection, clinical characterization, and DNA and mRNA extraction methods are described in the Supplementary Materials and Methods.

SUPPLEMENTARY MATERIALS

www.sciencetranslationalmedicine.org/cgi/content/full/5/181/181re1/DC1

Materials and Methods

Table S1. Univariate and multivariate Cox regression statistics for the clinical covariates in the METABRIC data set.

Table S2. Univariate and multivariate Cox regression statistics for the clinical covariates in the OsloVal data set.

Table S3. The Breast Cancer Challenge Consortium: Challenge participants who submitted a model to phase 3 of the BCC.

REFERENCES AND NOTES

1. F. Bray, J. S. Ren, E. Masuyer, J. Ferlay, Global estimates of cancer prevalence for 27 sites in the adult population in 2008. *Int. J. Cancer* **132**, 1133–1145 (2013).
2. C. M. Perou, T. Sørlie, M. B. Eisen, M. van de Rijn, S. S. Jeffrey, C. A. Rees, J. R. Pollack, D. T. Ross, H. Johnsen, L. A. Akslen, O. Fluge, A. Pergamenschikov, C. Williams, S. X. Zhu, P. E. Lønning, A. L. Børresen-Dale, P. O. Brown, D. Botstein, Molecular portraits of human breast tumours. *Nature* **406**, 747–752 (2000).
3. T. Sørlie, C. M. Perou, R. Tibshirani, T. Aas, S. Geisler, H. Johnsen, T. Hastie, M. B. Eisen, M. van de Rijn, S. S. Jeffrey, T. Thorsen, H. Quist, J. C. Matese, P. O. Brown, D. Botstein, P. E. Lønning, A. L. Børresen-Dale, Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl. Acad. Sci. U.S.A.* **98**, 10869–10874 (2001).
4. M. J. van de Vijver, Y. D. He, L. J. van't Veer, H. Dai, A. A. M. Hart, D. W. Voskuil, G. J. Schreiber, J. L. Peterse, C. Roberts, M. J. Marton, M. Parrish, D. Atsma, A. Witteveen, A. Glas, L. Delahaye, T. van der Velde, H. Bartelink, S. Rodenhuis, E. T. Rutgers, S. H. Friend, R. Bernards, A gene-expression signature as a predictor of survival in breast cancer. *N. Engl. J. Med.* **347**, 1999–2009 (2002).
5. S. Paik, S. Shak, G. Tang, C. Kim, J. Baker, M. Cronin, F. L. Baehner, M. G. Walker, D. Watson, T. Park, W. Hiller, E. R. Fisher, D. L. Wickerham, J. Bryant, N. Wolmark, A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *N. Engl. J. Med.* **351**, 2817–2826 (2004).
6. *Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease* (National Academies Press, Washington, DC, 2011); http://dels.nas.edu/Report/Toward-Precision-Medicine-Building-Knowledge/13284.
7. L. J. van 't Veer, H. Dai, M. J. van de Vijver, Y. D. He, A. A. M. Hart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Marton, A. T. Witteveen, G. J. Schreiber, R. M. Kerkhoven, C. Roberts, P. S. Linsley, R. Bernards, S. H. Friend, Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415**, 530–536 (2002).
8. E. Rutgers, M. J. Piccart-Gebhart, J. Bogaerts, S. Delaloge, L. Van't Veer, I. T. Rubio, G. Viale, A. M. Thompson, R. Passalacqua, U. Nitz, A. Vindevoghel, J. Y. Pierga, P. M. Ravdin, G. Werutsky, F. Cardoso, The EORTC 10041/BIG 03-04 MINDACT trial is feasible: Results of the pilot phase. *Eur. J. Cancer* **47**, 2742–2749 (2011).
9. F. Cardoso, M. Piccart-Gebhart, L. Van't Veer, E. Rutgers; TRANSBIG Consortium, The MINDACT trial: The first prospective clinical validation of a genomic tool. *Mol. Oncol.* **1**, 246–251 (2007).
10. J. A. Sparano, TAILORx: Trial assigning individualized options for treatment (Rx). *Clin. Breast Cancer* **7**, 347–350 (2006).
11. L. Marchionni, R. F. Wilson, A. C. Wolff, S. Marinopoulos, G. Parmigiani, E. B. Bass, S. N. Goodman, Systematic review: Gene expression profiling assays in early-stage breast cancer. *Ann. Intern. Med.* **148**, 358–369 (2008).
12. R. Simon, Roadmap for developing and validating therapeutically relevant genomic classifiers. *J. Clin. Oncol.* **23**, 7332–7341 (2005).
13. R. Norel, J. J. Rice, G. Stolovitzky, The self-assessment trap: Can we all be better than average? *Mol. Syst. Biol.* **7**, 537 (2011).
14. R. M. Bell, Y. Koren, Lessons from the Netflix prize challenge. *ACM SIGKDD Explor. Newsl.* **9**, 75–79 (2007).
15. L. Kedes, E. T. Liu, The Archon Genomics X PRIZE for whole human genome sequencing. *Nat. Genet.* **42**, 917–918 (2010).
16. S. Cooper, F. Khatib, A. Treuille, J. Barbero, J. Lee, M. Beenen, A. Leaver-Fay, D. Baker, Z. Popović, F. Players, Predicting protein structures with a multiplayer online game. *Nature* **466**, 756–760 (2010).
17. J. Carpenter, May the best analyst win. *Science* **331**, 698–699 (2011).
18. R. J. Allio, CEO interview: The InnoCentive model of open innovation. *Strategy Leadership* **32**, 4–9 (2004).
19. D. Marbach, J. C. Costello, R. Küffner, N. M. Vega, R. J. Prill, D. M. Camacho, K. R. Allison; DREAM5 Consortium, M. Kellis, J. J. Collins, G. Stolovitzky, Wisdom of crowds for robust gene network inference. *Nat. Methods* **9**, 796–804 (2012).
20. J. Moult, J. T. Pedersen, R. Judson, K. Fidelis, A large-scale experiment to assess protein structure prediction methods. *Proteins* **23**, ii–v (1995).
21. P. Radivojac, W. T. Clark, T. R. Oron, A. M. Schnoes, T. Wittkop, A. Sokolov, K. Graim, C. Funk, K. Verspoor, A. Ben-Hur, G. Pandey, J. M. Yunes, A. S. Talwalkar, S. Repo, M. L. Souza, D. Piovesan, R. Casadio, Z. Wang, J. Cheng, H. Fang, J. Gough, P. Koskinen, P. Törönen, J. Nokso-Koivisto, L. Holm, D. Cozzetto, D. W. A. Buchan, K. Bryson, D. T. Jones, B. Limaye, H. Inamdar, A. Datta, S. K. Manjari, R. Joshi, M. Chitale, D. Kihara, A. M. Lisewski, S. Erdin, E. Venner, O. Lichtarge, R. Rentzsch, H. Yang, A. E. Romero, P. Bhat, A. Paccanaro, T. Hamp, R. Kaßner, S. Seemayer, E. Vicedo, C. Schaefer, D. Achten, F. Auer, A. Boehm, T. Braun, M. Hecht, M. Heron, P. Hönigschmid, T. A. Hopf, S. Kaufmann, M. Kiening, D. Krompass, C. Landerer,

Y. Mahlich, M. Roos, J. Björne, T. Salakoski, A. Wong, H. Shatkay, F. Gatzmann, I. Sommer, M. N. Wass, M. J. E. Sternberg, N. Skunca, F. Supek, M. Bošnjak, P. Panov, S. Džeroski, T. Smuc, Y. A. I. Kourmpetis, A. D. J. van Dijk, C. J. F. Ter Braak, Y. Zhou, Q. Gong, X. Dong, W. Tian, M. Falda, P. Fontana, E. Lavezzo, B. Di Camillo, S. Toppo, L. Lan, N. Djuric, Y. Guo, S. Vucetic, A. Bairoch, M. Linial, P. C. Babbitt, S. E. Brenner, C. Orengo, B. Rost, S. D. Mooney, I. Friedberg, A large-scale evaluation of computational protein function prediction. *Nat. Methods* **10**, 221–227 (2013).

22. C. Curtis, S. P. Shah, S. F. Chin, G. Turashvili, O. M. Rueda, M. J. Dunning, D. Speed, A. G. Lynch, S. Samarajiwa, Y. Yuan, S. Gräf, G. Ha, G. Haffari, A. Bashashati, R. Russell, S. McKinney; METABRIC Group, A. Langerød, A. Green, E. Provenzano, G. Wishart, S. Pinder, P. Watson, F. Markowetz, L. Murphy, I. Ellis, A. Purushotham, A. L. Børresen-Dale, J. D. Brenton, S. Tavaré, C. Caldas, S. Aparicio, The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486**, 346–352 (2012).
23. J. M. J. Derry, L. M. Mangravite, C. Suver, M. D. Furia, D. Henderson, X. Schildwachter, B. Bot, J. Izant, S. K. Sieberts, M. R. Kellen, S. H. Friend, Developing predictive molecular maps of human disease through community-based modeling. *Nat. Genet.* **44**, 127–130 (2012).
24. R language definition; http://lib.stat.cmu.edu/R/CRAN/doc/manuals/R-lang.html.
25. F. E. Harrell, *Regression Modeling Strategies* (Springer, New York, 2001), p. 600; http://www.amazon.com/Regression-Modeling-Strategies-Frank-Harrell/dp/0387952322.
26. E. Bilal, J. Dutkowski, J. Guinney, I. S. Jang, B. A. Logsdon, G. Pandey, B. A. Sauerwine, Y. Shimoni, H. K. M. Vollan, B. H. Mecham, O. M. Rueda, J. Tost, C. Curtis, M. J. Alvarez, V. N. Kristensen, S. Aparicio, A. L. Børresen-Dale, C. Caldas, A. Califano, S. H. Friend, T. Ideker, E. E. Schadt, G. A. Stolovitzky, A. A. Margolin, Improving breast cancer survival analysis through competition-based multidimensional modeling. *PLoS Comput. Biol.* **9**, e1003047 (2013).
27. K. P. Garnock-Jones, G. M. Keating, L. J. Scott, Trastuzumab: A review of its use as adjuvant treatment in human epidermal growth factor receptor 2 (HER2)-positive early breast cancer. *Drugs* **70**, 215–239 (2010).
28. H. Ishwaran, U. B. Kogalur, E. H. Blackstone, M. S. Lauer, Random survival forests. *Ann. Appl. Stat.* **2**, 841–860 (2008).
29. D. Bauer, Constructing confidence sets using rank statistics. *J. Am. Stat. Assoc.* **67**, 687–690 (1972).
30. J. H. Friedman, Stochastic gradient boosting. *Comput. Stat. Data Anal.* **38**, 367–378 (2002).
31. R. J. Prill, J. Saez-Rodriguez, L. G. Alexopoulos, P. K. Sorger, G. Stolovitzky, Crowdsourcing network inference: The DREAM predictive signaling network challenge. *Sci. Signal.* **4**, mr7 (2011).
32. D. Marbach, R. J. Prill, T. Schaffter, C. Mattiussi, D. Floreano, G. Stolovitzky, Revealing strengths and weaknesses of methods for gene network inference. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 6286–6291 (2010).
33. R. J. Prill, D. Marbach, J. Saez-Rodriguez, P. K. Sorger, L. G. Alexopoulos, X. Xue, N. D. Clarke, G. Altan-Bonnet, G. Stolovitzky, Towards a rigorous assessment of systems biology models: The DREAM3 challenges. *PLoS One* **5**, e9202 (2010).
34. G. Stolovitzky, R. J. Prill, A. Califano, Lessons from the DREAM2 challenges. *Ann. N. Y. Acad. Sci.* **1158**, 159–195 (2009).
35. W. Y. Cheng, T. H. O. Yang, D. Anastassiou, Development of a prognostic model for breast cancer survival in an open challenge environment. *Sci. Transl. Med.* **5**, 181ra50 (2013).
36. L. Shi, G. Campbell, W. D. Jones, F. Campagne, Z. Wen, S. J. Walker, Z. Su, T. M. Chu, F. M. Goodsaid, L. Pusztai, J. D. J. Shaughnessy, A. Oberthuer, R. S. Thomas, R. S. Paules, M. Fielden, B. Barlogie, W. Chen, P. Du, M. Fischer, C. Furlanello, B. D. Gallas, X. Ge, D. B. Megherbi, W. F. Symmans, M. D. Wang, J. Zhang, H. Bitter, B. Brors, P. R. Bushel, M. Bylesjo, M. Chen, J. Cheng, J. Cheng, J. Chou, T. S. Davison, M. Delorenzi, Y. Deng, V. Devanarayan, D. J. Dix, J. Dopazo, K. C. Dorff, F. Elloumi, J. Fan, S. Fan, X. Fan, H. Fang, N. Gonzaludo, K. R. Hess, H. Hong, J. Huan, R. A. Irizarry, R. Judson, D. Juraeva, S. Lababidi, C. G. Lambert, L. Li, Y. Li, Z. Li, S. M. Lin, G. Liu, E. K. Lobenhofer, J. Luo, W. Luo, M. N. McCall, Y. Nikolsky, G. A. Pennello, R. G. Perkins, R. Philip, V. Popovici, N. D. Price, F. Qian, A. Scherer, T. Shi, W. Shi, J. Sung, D. Thierry-Mieg, J. Thierry-Mieg, V. Thodima, J. Trygg, L. Vishnuvajjala, S. J. Wang, J. Wu, Y. Wu, Q. Xie, W. A. Yousef, L. Zhang, X. Zhang, S. Zhong, Y. Zhou, S. Zhu, D. Arasappan, W. Bao, A. B. Lucas, F. Berthold, R. J. Brennan, A. Buness, J. G. Catalano, C. Chang, R. Chen, Y. Cheng, J. Cui, W. Czika, F. Demichelis, X. Deng, D. Dosymbekov, R. Eils, Y. Feng, J. Fostel, S. Fulmer-Smentek, J. C. Fuscoe, L. Gatto, W. Ge, D. R. Goldstein, L. Guo, D. N. Halbert, J. Han, S. C. Harris, C. Hatzis, D. Herman, J. Huang, R. V. Jensen, R. Jiang, C. D. Johnson, G. Jurman, Y. Kahlert, S. A. Khuder, M. Kohl, J. Li, M. Li, Q. Z. Li, S. Li, Z. Li, J. Liu, Y. Liu, Z. Liu, L. Meng, M. Madera, F. Martinez-Murillo, I. Medina, J. Meehan, K. Miclaus, R. A. Moffitt, D. Montaner, P. Mukherjee, G. J. Mulligan, P. Neville, T. Nikolskaya, B. Ning, G. P. Page, J. Parker, R. M. Parry, X. Peng, R. L. Peterson, J. H. Phan, B. Quanz, Y. Ren, S. Riccadonna, A. H. Roter, F. W. Samuelson, M. M. Schumacher, J. D. Shambaugh, Q. Shi, R. Shippy, S. Si, A. Smalter, C. Sotiriou, M. Soukup, F. Staedtler, G. Steiner, T. H. Stokes, Q. Sun, P. Y. Tan, R. Tang, Z. Tezak, B. Thorn, M. Tsyganova, Y. Turpaz, S. C. Vega, R. Visintainer, J. von Frese, C. Wang, E. Wang, J. Wang, W. Wang, F. Westermann, J. C. Willey, M. Woods, S. Wu, N. Xiao, J. Xu, L. Xu, L. Yang, X. Zeng, J. Zhang, L. Zhang, M. Zhang, C. Zhao, R. K. Puri, U. Scherf, W. Tong, R. D. Wolfinger, The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. *Nat. Biotechnol.* **28**, 827–838 (2010).
37. J. S. Parker, M. Mullins, M. C. U. Cheang, S. Leung, D. Voduc, T. Vickery, S. Davies, C. Fauron, X. He, Z. Hu, J. F. Quackenbush, I. J. Stijleman, J. Palazzo, J. S. Marron, A. B. Nobel, E. Mardis, T. O. Nielsen, M. J. Ellis, C. M. Perou, P. S. Bernard, Supervised risk predictor of breast cancer based on intrinsic subtypes. *J. Clin. Oncol.* **27**, 1160–1167 (2009).
38. Guidance for industry clinical trial endpoints for the approval of cancer drugs and biologics; http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/ucm071590.pdf.
39. D. Williams, J. W. Davidson, J. D. Hiser, J. C. Knight, A. Nguyen-Tuong, Security through diversity: Leveraging virtual machine technology. *IEEE Security Privacy Mag.* **7**, 26–33 (2009).
40. P. J. Stephens, P. S. Tarpey, H. Davies, P. Van Loo, C. Greenman, D. C. Wedge, S. Nik-Zainal, S. Martin, I. Varela, G. R. Bignell, L. R. Yates, E. Papaemmanuil, D. Beare, A. Butler, A. Cheverton, J. Gamble, J. Hinton, M. Jia, A. Jayakumar, D. Jones, C. Latimer, K. W. Lau, S. McLaren, D. J. McBride, A. Menzies, L. Mudie, K. Raine, R. Rad, M. S. Chapman, J. Teague, D. Easton, A. Langerød, M. T. M. Lee, C. Y. Shen, B. T. K. Tee, B. W. Huimin, A. Broeks, A. C. Vargas, G. Turashvili, J. Martens, A. Fatima, P. Miron, S. F. Chin, G. Thomas, S. Boyault, O. Mariani, S. R. Lakhani, M. van de Vijver, L. van 't Veer, J. Foekens, C. Desmedt, C. Sotiriou, A. Tutt, C. Caldas, J. S. Reis-Filho, S. A. J. R. Aparicio, A. V. Salomon, A. L. Børresen-Dale, A. L. Richardson, P. J. Campbell, P. A. Futreal, M. R. Stratton, The landscape of cancer genes and mutational processes in breast cancer. *Nature* **486**, 400–404 (2012).
41. R. C. Gentleman, V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A. J. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J. Y. H. Yang, J. Zhang, Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biol.* **5**, R80 (2004).
42. B. H. Mecham, P. S. Nelson, J. D. Storey, Supervised normalization of microarrays. *Bioinformatics* **26**, 1308–1315 (2010).
43. Gene expression profiling with Sentrix focused arrays; http://www.illumina.com/Documents/products/techbulletins/techbulletin_rna.pdf.
44. B. Carvalho, pd.genomewidesnp.6: Platform design info for Affymetrix GenomeWideSNP_6; http://www.bioconductor.org/packages/2.12/data/annotation/html/pd.genomewidesnp.6.html.
45. The International Agency for Research on Cancer, *World Health Organization: Tumours of the Breast and Female Genital Organs (WHO/IARC Classification of Tumours)*, F. A. Tavassoéli, P. Devilee, Eds. (IARC Press, Lyon, 2003), p. 432; http://www.amazon.com/World-Health-Organization-Tumours-Classification/dp/9283224124.
46. C. W. Elston, I. O. Ellis, Pathological prognostic factors in breast cancer. I. The value of histological grade in breast cancer: Experience from a large study with long-term follow-up. *Histopathology* **41**, 151–152 (2002).

Acknowledgments: We thank I. R. Bergheim, E. U. Due, A. H. T. Olsen, C. Pedersen, M. La. Skrede, and P. Vu (Department of Genetics, Institute for Cancer Research) and L. I. Håseth (Department of Pathology, Oslo University Hospital, The Norwegian Radium Hospital) for their great efforts in the laboratory in collecting, cataloging, preparing, and analyzing the breast cancer tissue samples in a very short time; I. S. Jang for help on data analysis; H. Dai for assistance with the 70-gene signature evaluation; and L. Van't Veer for helpful discussions in formulating the Challenge. The Breast Cancer Challenge Consortium members (listed in table S3) are each acknowledged for their submission of at least one computational model to phase 3 of the BCC. **Funding:** This work was funded by NIH/National Cancer Institute grant 5U54CA149237 and Washington Life Science Discovery Fund grant 3104672. Generation of the OsloVal data set was funded by the Avon Foundation (Foundation for the NIH grant FRIE12PD), The Norwegian Cancer Society, and The Radium Hospital Foundation. Computational resources provided to Challenge participants were donated by Google. **Author contributions:** A.A.M., E.B., E.H., T.C.N., H.K.M.V., O.M.R., J.G., V.N.K., S.H.F., G.S., S.A., C. Caldas, and A.-L.B.-D. conceived and designed the study. A.A.M., E.B., E.H., T.C.N., S.H.F., and G.S. drafted the manuscript. A.A.M., E.B., and E.H. performed data analysis. L.O., H.K.M.V., H.G.R., D.P., V.O.V., and A.-L.B.-D. organized clinical data and samples for the OsloVal data set. B.H.M. performed data normalization. B.S., M.R.K., M.D.F., N.A.D., B.H., and X.S. developed software infrastructure. L.M.M. organized data governance. T.P., L.Y., C. Citro, and J.H. developed Google compute resources. C. Curtis, S.A., and C. Caldas developed the METABRIC data set.

Submitted 3 March 2013
Accepted 29 March 2013
Published 17 April 2013
10.1126/scitranslmed.3006112

Citation: A. A. Margolin, E. Bilal, E. Huang, T. C. Norman, L. Ottestad, B. H. Mecham, B. Sauerwine, M. R. Kellen, L. M. Mangravite, M. D. Furia, H. K. M. Vollan, O. M. Rueda, J. Guinney, N. A. Deflaux, B. Hoff, X. Schildwachter, H. G. Russnes, D. Park, V. O. Vang, T. Pirtle, L. Youseff, C. Citro, C. Curtis, V. N. Kristensen, J. Hellerstein, S. H. Friend, G. Stolovitzky, S. Aparicio, C. Caldas, A.-L. Børresen-Dale, Systematic analysis of challenge-driven improvements in molecular prognostic models for breast cancer. *Sci. Transl. Med.* **5**, 181re1 (2013).