COMPUTATIONAL MODELING

Development of a Prognostic Model for Breast Cancer Survival in an Open Challenge Environment

Wei-Yi Cheng, Tai-Hsien Ou Yang, Dimitris Anastassiou*

The accuracy with which cancer phenotypes can be predicted by selecting and combining molecular features is compromised by the large number of potential features available. In an effort to design a robust prognostic model to predict breast cancer survival, we hypothesized that signatures consisting of genes that are coexpressed in multiple cancer types should correspond to molecular events that are prognostic in all cancers, including breast cancer. We previously identified several such signatures—called attractor metagenes—in an analysis of multiple tumor types. We then tested our attractor metagene hypothesis as participants in the Sage Bionetworks–DREAM Breast Cancer Prognosis Challenge. Using a rich training data set that included gene expression and clinical features for breast cancer patients, we developed a prognostic model that was independently validated in a newly generated patient data set. We describe our model, which was based on three attractor metagenes associated with mitotic chromosomal instability, mesenchymal transition, or lymphocyte-based immune recruitment.

INTRODUCTION

Medical tests that incorporate molecular profiling of tumors for clinical decision-making (predictive tests) or prognosis (prognostic tests) are typically based on models that combine values associated with particular molecular features, such as the expression levels of specific genes. These genes are selected after analyzing rich gene expression data sets (acquired from testing patient tumors) annotated with clinical phenotypes such as drug responses or survival times. The data sets used to define a model are referred to as "training data sets." A computational technique is typically used to identify a number of genes that, when properly combined, are associated with a phenotype of interest in a statistically significant manner. The predictive power of the resulting model is later confirmed in independent "validation data sets."

There are, however, vast numbers—tens or hundreds of thousands—of potentially relevant molecular features to choose from when developing a model, making it difficult to precisely identify those at the core of the biological mechanisms responsible for the phenotype of interest. Spurious or suboptimal predictions may occur, and the end result may be a model that only partly reflects physiological reality. Such a model may still be clinically useful, but there is room for improvement.

One way to address this problem is by using molecular features preselected on the basis of previous knowledge. In such an approach, a training data set is used mainly for pinpointing the combination of preselected features that is most associated with the phenotype of interest. We used this approach during our participation in the Sage Bionetworks–DREAM Breast Cancer Prognosis Challenge, an open challenge to build computational models that accurately predict breast cancer survival (hereinafter referred to as the Challenge) (*1*). Specifically, we hypothesized that selected gene coexpression signatures present in multiple cancer types should be useful for prediction of survival in breast cancer. We had derived these signatures, which we call "attractor metagenes," previously through a multicancer analysis of gene expression data (*2*).

Attractor metagenes are signatures of coexpressed genes identified in rich gene expression data sets by an iterative approach (*2*) starting from a "seed" gene and converging to a metagene—a hypothetical gene whose expression levels are a weighted average of actual genes—that points to the "heart" (core) of the coexpression mechanism. Each attractor metagene results from the convergence of multiple seed genes. By independently analyzing data sets from several different cancer types, we found that there exist several such attractor metagenes to which this iterative algorithm converges in nearly identical form, regardless of the cancer type that gave rise to the gene expression data set. The differences between similar metagenes identified in different cancer types may sometimes be smaller than the differences between those found by analyzing data sets of the same cancer type (*2*). This observation suggests that several attractor metagenes represent "pan-cancer" (cancer type–independent) biomolecular events in cancer.

Through participation in the Challenge, we found that these attractor metagenes were also strong prognostic features for breast cancer survival. This phenotypic association was present despite the fact that these signatures (i) were discovered by a purely unsupervised method (that is, without reference to any phenotypic association) and (ii) were determined without using the Challenge training data set. Instead, we used our previously identified attractor metagenes (*2*) as prognostic features in the Challenge. Here, we describe our Challenge-winning model (*1*), which combined three universal metagenes and several additional clinical and molecular features to predict patient ranking in terms of their survival.

RESULTS

The three universal attractor metagenes used to develop our final model contain genes associated with mitotic chromosomal instability (CIN), mesenchymal transition (MES), and lymphocyte-specific immune recruitment (LYM). Because cancer is thought to be characterized by a few unifying "hallmarks" (*3*), we think of these gene signatures as "bioinformatic hallmarks of cancer" that are associated with the ability of cancer cells to divide uncontrollably, to invade surrounding tissues, and, with the effort of the organism to fight cancer with a particular immune response. In addition, our model makes use of another molecular feature that we identified during our participation in the Challenge: a metagene whose expression is associated with good prognosis and that contains the expression values of two genes—*FGD3* and *SUSD3*—that are genomically adjacent to each other.

Center for Computational Biology and Bioinformatics and Department of Electrical Engineering, Columbia University, New York, NY 10027, USA.
*Corresponding author. E-mail: da8@columbia.edu

Participation in the Challenge

The initial phases of the Challenge (*1*) were based on partitioning of the rich METABRIC breast cancer data set (*4*) (which includes molecular, clinical, and survival information from 1981 patients) into two subsets: a training set and a validation set. Participants' computational models were developed on the training set and evaluated on the validation set, using a real-time leaderboard to record the performance [as determined with concordance index (CI) values, defined below] of all submitted models. During the final phase of the Challenge, participants were given access to the full set of the METABRIC data, which had been renormalized for uniformity by Sage Bionetworks using eigen probe set analysis (*5*). At that time, the computational models could be trained on that full set and submitted for evaluation against a newly generated validation data set of patients, referred to as the Oslo Validation (OsloVal) data set (*1*). Therefore, the numerical values for the results that we present here use the full METABRIC data set to maximize accuracy, whereas our computational models were developed using the originally available training data sets.

The setup accounted for a vibrant environment in which each participant was encouraged to observe and use others' computer programs and to post comments and suggestions in the Sage Bionetworks Synapse forum. Indeed, as we were discovering the prognostic ability of each attractor metagene (CIN, MES, and LYM), we immediately used the forum to make all participants aware of our findings, in case others wanted to use them in the development of their own models (for example, see http://support.sagebase.org/sagebase/topics/mitotic_chromosomal_instability_attractor_metagene, http://support.sagebase.org/sagebase/topics/mesenchymal_transition_attractor_metagene-znl1g, and http://support.sagebase.org/sagebase/topics/lymphocyte_specific_attractor_metagene).

It was also quite helpful for us to observe the results of other participants' use of our code, as we were able to avoid trying to incorporate methods that we saw were not working well for others.

Selection of a numerical score for evaluating prognostic models

A CI (*6*) was the numerical measure used to score all Challenge submissions on the leaderboards. In this context, the CI is a score that applies to a cohort of patients (rather than an individual patient) and evaluates the similarity between the actual ranking of patients in terms of their survival and the ranking predicted by the computational model. CI measures the relative frequency of accurate pairwise predictions of survival over all pairs of patients for which such a meaningful determination can be achieved and, therefore, is a number between 0 and 1. The average CI for random predictions is 0.5. If a model achieves a CI of 0.75, then the model will correctly order the survival of two randomly chosen patients three of four times. Our final model had a CI of 0.756 in the OsloVal data set.

The METABRIC data set included both disease-specific (DS) survival data, in which all reported deaths were determined to be due to breast cancer (otherwise, a patient was considered equivalent to a hypothetical still living patient with reported survival equal to the time to actual death from other causes), and overall survival (OS) data, in which all deaths are reported even though they could potentially be due to other causes. Our research performed in the context of the Challenge used mainly DS survival–based data, and unless otherwise noted, the CI scores referring to the METABRIC data set presented in this paper were evaluated using DS survival data. This is because we found that the CIs for models developed using DS survival–based data from the METABRIC data set were significantly higher than those obtained when the OS survival–based data were used. Furthermore, we found that DS survival–based modeling did not need to include age as a prognostic feature as much as OS survival–based modeling did, which suggests that OS survival–based modeling cannot predict survival using molecular features as accurately as DS survival–based modeling, and instead needed to make use of age, which is an obvious feature for predicting survival even in healthy people.

The first phases of the Challenge consisted of participants training their prognostic computational models using a subset of samples from the full METABRIC data set as a training set, whereas the remaining subset was used to test the models by evaluating the CI scores in a real-time leaderboard. The survival data and the corresponding scoring of the OsloVal data set were OS survival–based. Accordingly, the Kaplan-Meier survival curves presented in this paper involving OsloVal are OS survival–based.

CI scores for individual genes. As a first task, we quantified the prognostic ability of the expression level of each individual gene by computing the CI between the expression levels of the gene in all patients and the survival of those patients (Table 1). Specifically, the CIs reported in Table 1 are the CIs that we would calculate if the prognostic model consisted exclusively of the expression level of only one specific gene. For example, consider the *CDCA5* gene (listed at the top of the left-hand column of Table 1). If we ranked all patients in terms of their *CDCA5* expression levels, from highest to lowest, and then ranked all patients in terms of their survival times, from shortest to longest, these two rankings would yield a CI of 0.651. This means that if we randomly select two patients from the METABRIC data set, the one whose expression of *CDCA5* is higher will have the shorter survival time 65.1% of the time. Because *CDCA5* expression is associated with poor prognosis (that is, the higher the expression, the shorter the survival), we refer to *CDCA5* as a poor survival–inducing gene (or simply, an "inducing gene," which is one that displays a CI that is significantly greater than 0.5).

At the opposite end of the spectrum was the *FGD3* gene, which had a CI of 0.352 (Table 1, right-hand column). This CI indicates that if one randomly chooses two patients from the METABRIC data set, then the one with lower *FGD3* expression levels will have the shorter survival time 64.8% (100% minus 35.2%) of the time. Because high levels of *FGD3* expression were associated with a good prognosis (that is, the higher the expression, the longer the survival), we refer to *FGD3* as a survival-protective gene (or simply, a "protective" gene, which is one that displays a CI that is significantly less than 0.5). Table 1 shows two expanded lists of ranked genes: one with the most inducing genes (those with the highest CIs) and one with the most protective genes (those with the lowest CIs).

In the following, all references to gene expression levels, including average values and numbers on scatter-plot axes, are assumed to be $\log_2$-normalized as provided to us. For each attractor metagene, when we refer to its top-ranked genes, we mean those that had the highest mutual information (*7*) with the attractor metagene, as provided in our previous work (*2*).

Mitotic CIN attractor metagene

In the Challenge, we represented the mitotic CIN attractor metagene with the average of the expression levels of the 10 top-ranked genes from our previously evaluated (*2*) attractor metagene: *CENPA*, *DLGAP5*,

Table 1. CIN expression and survival. We ranked individual genes in terms of their CIs with respect to gene expression and survival data in the METABRIC data set. The CI measures the similarity of patient rankings based on the expression level of the gene compared to the actual rankings based on DS survival data. Shown on the left are the most "inducing" genes with the highest CIs. Shown on the right are the most protective genes with the lowest CIs. The underlined genes are among the top 100 genes of the CIN attractor metagene defined in (*2*). The probe IDs are identifiers for probes designed by Illumina. If a gene was profiled by multiple probes, we chose the probe with the highest difference from the average CI for random predictions, 0.5. Genes identified by asterisks are among the 10 top-ranked genes of the CIN attractor metagene and were used in the model.

Probe ID	Gene symbol	CI	Probe ID	Gene symbol	CI
ILMN_1683450	CDCA5	0.651	ILMN_1772686	FGD3	0.352
ILMN_1714730	UBE2C	0.644	ILMN_1785570	SUSD3	0.358
ILMN_1801939	CCNB2*	0.643	ILMN_2310814	MAPT	0.372
ILMN_1700337	TROAP	0.643	ILMN_2353862	LRRC48	0.374
ILMN_2357438	AURKA	0.642	ILMN_2397954	PARP3	0.374
ILMN_1781943	FAM83D	0.640	ILMN_1674661	CIRBP	0.375
ILMN_2212909	MELK*	0.640	ILMN_1801119	BCL2	0.376
ILMN_1695658	KIF20A*	0.639	ILMN_1708983	CASC1	0.377
ILMN_1673721	EXO1	0.639	ILMN_1772588	CCDC170	0.377
ILMN_1786125	CCNA2*	0.638	ILMN_1849013	HS.570988	0.378
ILMN_1801257	CENPA*	0.638	ILMN_1809639	TMEM26	0.378
ILMN_1796949	TPX2	0.637	ILMN_1657361	CBX7	0.380
ILMN_1771039	GTSE1	0.637	ILMN_1713162	GSTM2	0.380
ILMN_1716279	CENPE	0.637	ILMN_1806456	C14orf45	0.380
ILMN_1808071	KIF14	0.636	ILMN_1790315	C7orf63	0.381
ILMN_2077550	RACGAP1	0.636	ILMN_1667716	TMEM101	0.382
ILMN_1736176	PLK1	0.636	ILMN_1907649	HS.144312	0.382
ILMN_1703906	HJURP	0.636	ILMN_1811014	PGR	0.382
ILMN_1663390	CDC20	0.636	ILMN_1807211	NICN1	0.382
ILMN_1751776	CKAP2L	0.635	ILMN_1805104	ABAT	0.382
ILMN_2344971	FOXM1	0.635	ILMN_1655117	WDR19	0.383
ILMN_1751444	NCAPG*	0.635	ILMN_1696254	CYB5D2	0.383
ILMN_1747016	CEP55	0.634	ILMN_1777342	PREX1	0.383
ILMN_2042771	PTTG1	0.634	ILMN_2183692	PHYHD1	0.384
ILMN_1740291	POLQ	0.633	ILMN_2128795	LRIG1	0.384
ILMN_2202948	BUB1*	0.633	ILMN_1784783	NME5	0.384
ILMN_1685916	KIF2C*	0.633	ILMN_1862217	HS.532698	0.384
ILMN_2413898	MCM10	0.632	ILMN_1815705	LZTFL1	0.384
ILMN_1713952	C1orf106	0.632	ILMN_1670925	CYB5D1	0.385
ILMN_1684217	AURKB	0.632	ILMN_1684034	STAT5B	0.386
ILMN_1815184	ASPM	0.632	ILMN_1664922	FLNB	0.387
ILMN_1737728	CDCA3	0.632	ILMN_1794213	ABHD14A	0.387
ILMN_1702197	SAPCD2	0.630	ILMN_1776967	DNAAF1	0.387
ILMN_1728934	PRC1	0.630	ILMN_1736184	GSTM3	0.387
ILMN_1739645	ANLN	0.629	ILMN_1760574	RAI2	0.387
ILMN_2049021	PTTG3	0.629	ILMN_2341254	STARD13	0.387
ILMN_1670238	CDC45	0.628	ILMN_1651364	PCBD2	0.387
ILMN_1799667	KIF4A*	0.628	ILMN_1769382	KBTBD3	0.387
ILMN_1788166	TTK	0.628	ILMN_1697317	DYNLRB2	0.387
ILMN_1771734	GMPSP1	0.627	ILMN_1790350	TPRG1	0.388
ILMN_1811472	KIF23	0.627	ILMN_1664348	PNPLA4	0.389
ILMN_1666305	CDKN3	0.627	ILMN_2125763	ZMYND10	0.389
ILMN_1731070	ORC6	0.627	ILMN_2323385	TRIM4	0.389
ILMN_2413650	STIL	0.626	ILMN_1657451	SRPK2	0.389
ILMN_1770678	CBX2	0.626	ILMN_1779416	SCUBE2	0.390
ILMN_1749829	DLGAP5*	0.625	ILMN_1719622	RABEP1	0.391
ILMN_1789510	STIP1	0.624	ILMN_1687351	ANKRA2	0.391
ILMN_1814281	SPC25	0.624	ILMN_1691884	STC2	0.391
ILMN_1709294	CDCA8	0.624	ILMN_2140700	CRIPAK	0.393
ILMN_1671906	MND1	0.624	ILMN_1858599	HS.20255	0.393

MELK, *BUB1*, *KIF2C*, *KIF20A*, *KIF4A*, *CCNA2*, *CCNB*, and *NCAPG*. We refer to the metagene defined by this average as the "CIN feature." It contains many genes that encode proteins that are part of the kinetochore—a structure at which spindle fibers attach during cell division to segregate sister chromatids—particularly those involved in the microtubule-kinetochore interface, suggesting a biological mechanism by which mitotic chromosomal instability in dividing cancer cells gives rise to daughter cells with genomic modifications, some of which pass the test of natural selection. We showed previously that the mitotic CIN attractor metagene is strongly associated with tumor grade (a classification system that measures how abnormal a cancer cell appears when assessed microscopically) in multiple cancers (*2*).

We essentially rediscovered the mitotic CIN attractor metagene by identifying the genes for which expression was most associated with poor prognosis in the METABRIC data set. Indeed, all 10 genes (listed above) of the CIN feature that we used in the Challenge were among the 50 genes listed in the left column of Table 1; furthermore, 40 of the 50 genes listed in the left column of Table 1 were among the top 100 genes of the CIN attractor metagene identified previously (*2*) (the P value for such overlap is less than 1.04×10^{-97} based on Fisher's exact test).

Our results regarding this and other attractor metagenes were validated in a statistically significant manner in the OsloVal data set despite its relatively small size (184 samples). For example, Fig. 1 shows the Kaplan-Meier cumulative survival curves (*8*) of the CIN feature for the METABRIC ($P < 2 \times 10^{-16}$ using log-rank test) and OsloVal ($P = 0.0041$ using log-rank test) data sets, comparing tumors with high and low values of the CIN feature. These data confirmed that poor prognosis was associated with expression of the mitotic CIN attractor metagene.

MES attractor metagene

In the Challenge, we represented the MES attractor metagene with the average of the expression levels of the 10 top-ranked genes from our previously evaluated (*2*) attractor metagene: *COL5A2*, *VCAN*, *SPARC*, *THBS2*, *FBN1*, *COL1A2*, *COL5A1*, *FAP*, *AEBP1*, and *CTSK*. We refer to the metagene defined by this average as the MES feature. We had discovered a nearly identical signature previously (*9*) from its association with tumor stage (a measure of the extent to which the cancer has spread to adjacent lymph nodes or distant sites in the body). Specifically, the signature is expressed in high amounts only in tumor samples from patients whose cancer has exceeded a defined stage threshold, which is cancer type–specific. For example, in breast cancer, the MES signature appears early, when in situ carcinoma becomes invasive (stage I); in colon cancer, it is expressed when stage II is reached; and in ovarian cancer, it is expressed when stage III is reached. Identification of stage-specific differentially expressed genes in these three cancers reveals strong enrichment of the signature. We found that this differential expression results from the fact that the signature is present in some, but not all, samples in which the stage threshold is exceeded, but never in samples in which the stage threshold has not been reached. That is, the presence of the signature implies tumor invasiveness, but its absence is uninformative.

Related versions of the MES signature were found to be prognostic in various cancers, such as oral squamous cell carcinoma (*2*) and ovarian cancer (*10*). In breast cancer, however, we found that the prognostic ability of the MES feature individually was not significant. We reasoned that this lack of prognostic power is explained by the fact that the presence of the MES signature in breast cancer implies that the tumor is invasive, but this was the case anyway for nearly all patients in the METABRIC data set.

Therefore, we hypothesized that the MES signature would be prognostic only for very early stage breast cancer patients, which we defined by the absence of positive lymph nodes combined with a tumor size less than 30 mm. This restriction improved prognostic ability, but it still did not reach the level of statistical significance. However, we found that, in combination with the other features that we used, this restricted version of the MES signature was helpful toward the performance of our final model. This was confirmed, as we describe below, by the fact that the prognostic power of our final model was reduced when eliminating the MES feature.

LYM attractor metagene

In the Challenge, we represented the LYM attractor metagene with the average of the expression levels of the 10 top-ranked genes from our previously evaluated (*2*) attractor metagene: *PTPRC* (*CD45*), *CD53*, *LCP2* (*SLP-76*), *LAPTM5*, *DOCK2*, *IL10RA*, *CYBB*, *CD48*, *ITGB2* (*LFA-1*), and *EVI2B*. We refer to the metagene defined by this average as the LYM feature. The composition of this gene signature indicates that a signaling pathway that includes the protein tyrosine phosphatase receptor type C (also called CD45; encoded by *PTPRC*) and leukocyte surface antigen CD53 has a role in patient survival. Some of the top-ranked genes in the LYM attractor metagene, including *ADAP* (*FYB*), are known to participate in a particular type of immune response (*11*) in which the LFA-1 integrin mediates costimulation of T lymphocytes that are regulated by the SLP-76–ADAP adaptor molecule.

By itself, the LYM feature was slightly protective (CI < 0.5) in the METABRIC data set but was not significantly associated with prognosis. Therefore, we used a "trial and error" approach by testing the prognostic power of the feature on various subsets of patients grouped on the basis of histology, estrogen receptor (ER) status, etc. The LYM feature was strongly protective in ER-negative breast cancer in the METABRIC data set, and this observation was validated in the OsloVal data set; Fig. 2A shows Kaplan-Meier survival curves for ER-negative patients from the METABRIC data set ($P = 0.0024$ using log-rank

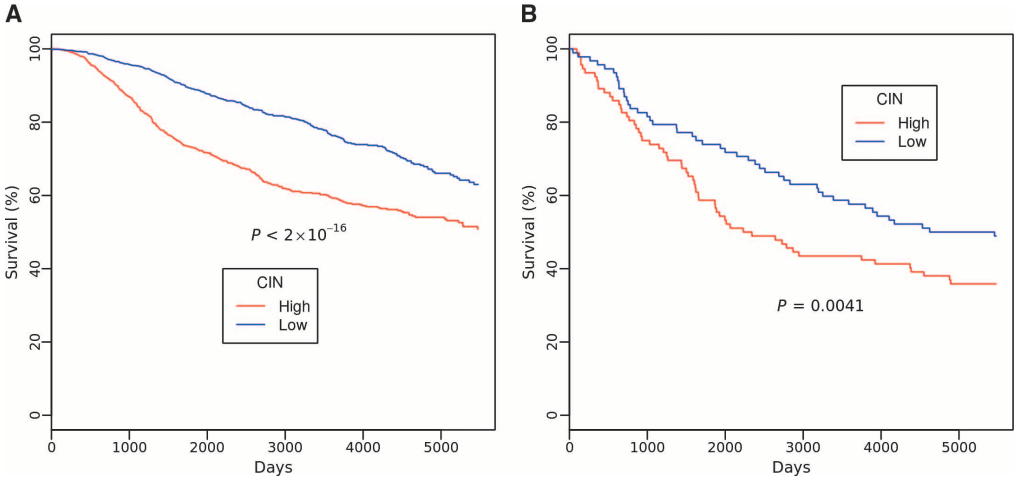

Fig. 1. Mitotic CIN attractor metagene. (**A** and **B**) Kaplan-Meier cumulative survival curves of breast cancer patients over a 15-year period on the basis of the mitotic CIN attractor metagene expression—represented by the CIN feature—in the (A) METABRIC and (B) OsloVal data sets. The patients were divided into equal-sized "high" and "low" CIN-expressing subgroups according to their ranking with respect to expression values of the CIN feature. High expression of the mitotic CIN attractor metagene was associated with poorer survival in both data sets. *P* values derived using the log-rank test in the two data sets were less than 2×10^{-16} and 0.041, respectively.

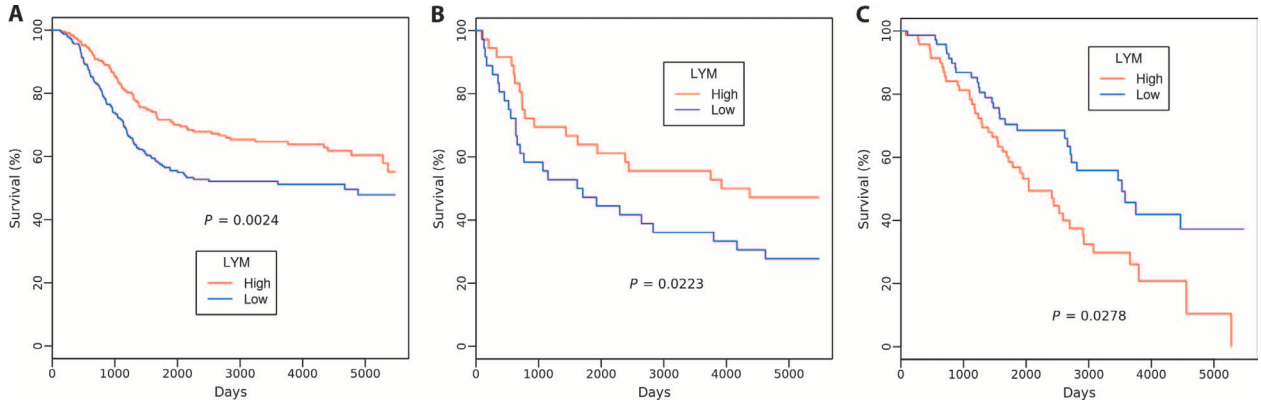

Fig. 2. LYM attractor metagene. (**A** and **B**) Kaplan-Meier cumulative survival curves of ER-negative breast cancer patients over a 15-year period on the basis of LYM attractor metagene expression—represented by the LYM feature—in the (A) METABRIC and (B) OsloVal data sets. The ER-negative breast cancer patients were divided into equal-sized high and low LYM-expressing subgroups according to their ranking with respect to expression values of the LYM feature. High expression of the LYM attractor metagene was associated with improved survival in both data sets. *P* values derived using the log-rank test in the two data sets were 0.0024 and 0.0223, respectively. (**C**) Kaplan-Meier cumulative survival curves of ER-positive breast cancer patients with more than four positive lymph nodes over a 15-year period on the basis of LYM attractor metagene expression—represented by the LYM feature—in the METABRIC data set. ER-positive breast cancer patients with more than four positive lymph nodes were divided into equal-sized high and low LYM-expressing subgroups according to their ranking with respect to expression values of the LYM feature. In contrast to (A), high expression of the LYM attractor metagene was associated with poorer survival in this patient subset. The *P* value derived using the log-rank test was 0.0278. There were only 19 corresponding samples in the OsloVal data set, insufficient for validation of this reversal relative to (B).

test); Fig. 2B shows Kaplan-Meier survival curves for ER-negative patients from the OsloVal data set ($P = 0.0223$ using log-rank test). In both cases, the curves compare tumors with high and low values of the LYM feature.

By contrast, the effect on prognosis was reversed for patients who had ER-positive cancers and multiple cancer cell–positive lymph nodes; Fig. 2C shows the Kaplan-Meier survival curves for METABRIC patients with ER-positive status and more than four positive lymph nodes, comparing tumors with high and low values of the LYM feature ($P = 0.0278$ using log-rank test). There were only 19 corresponding samples in the OsloVal data set, insufficient for validation of this reversal.

FGD3-SUSD3 metagene

As shown in Table 1, the *FGD3* and *SUSD3* genes were found to be the most protective ones in the METABRIC data set, with CIs equal to 0.352 and 0.358, respectively. Therefore, we considered them to be promising candidates to be included as features in our prognostic model. The two genes are genomically adjacent to each other at chromosome 9q22.31. In our final prognostic model, we used the *FGD3-SUSD3* metagene, which was defined by the average of the two expression values.

A scatter plot (Fig. 3A) of the METABRIC expression levels of *FGD3* versus *SUSD3* showed that the two genes did not appear to be co-regulated when one or the other gene was highly expressed, but the genes did appear to be simultaneously silent (that is, low expression of one gene implies low expression of the other). The CIs for the *FGD3-SUSD3* metagene and the *estrogen receptor 1* (*ESR1*) gene in the METABRIC data set were 0.346 and 0.403, respectively, indicating that the lack of *FGD3-SUSD3* expression was more strongly associated with poor prognosis compared with lack of expression of *ESR1*. Furthermore, a scatter plot (Fig. 3B) of the METABRIC expression levels of the *FGD3-SUSD3* metagene versus *ESR1* revealed that the two features were associated in the sense that ER-negative breast cancers tended to express low levels of the *FGD3-SUSD3* metagene, but the reverse was not necessarily true.

The poor prognosis associated with low expression of the *FGD3-SUSD3* metagene was validated in the OsloVal data set. Figure 3C shows the Kaplan-Meier curves for the *FGD3-SUSD3* metagene in the METABRIC data set ($P < 2 \times 10^{-16}$ using log-rank test). Figure 3D shows the Kaplan-Meier survival curves for the *FGD3-SUSD3* metagene in the OsloVal data set ($P = 0.0028$ using log-rank test). In both cases, the curves compare tumors with high and low expression of the *FGD3-SUSD3* metagene.

Breast Cancer Prognosis Challenge model

The development of our breast cancer prognosis model for the Challenge is described in detail in Materials and Methods. It used, as potential features, several metagenes that we had identified previously (*2*), the *FGD3-SUSD3* metagene (identified during the Challenge), and the clinical phenotypes that were available to the Challenge participants. During the course of the Challenge, we tried several combinations of prognostic algorithms (based on various statistical and machine-learning techniques), each of which defined a computational model that automatically selected some of the potential features and achieved

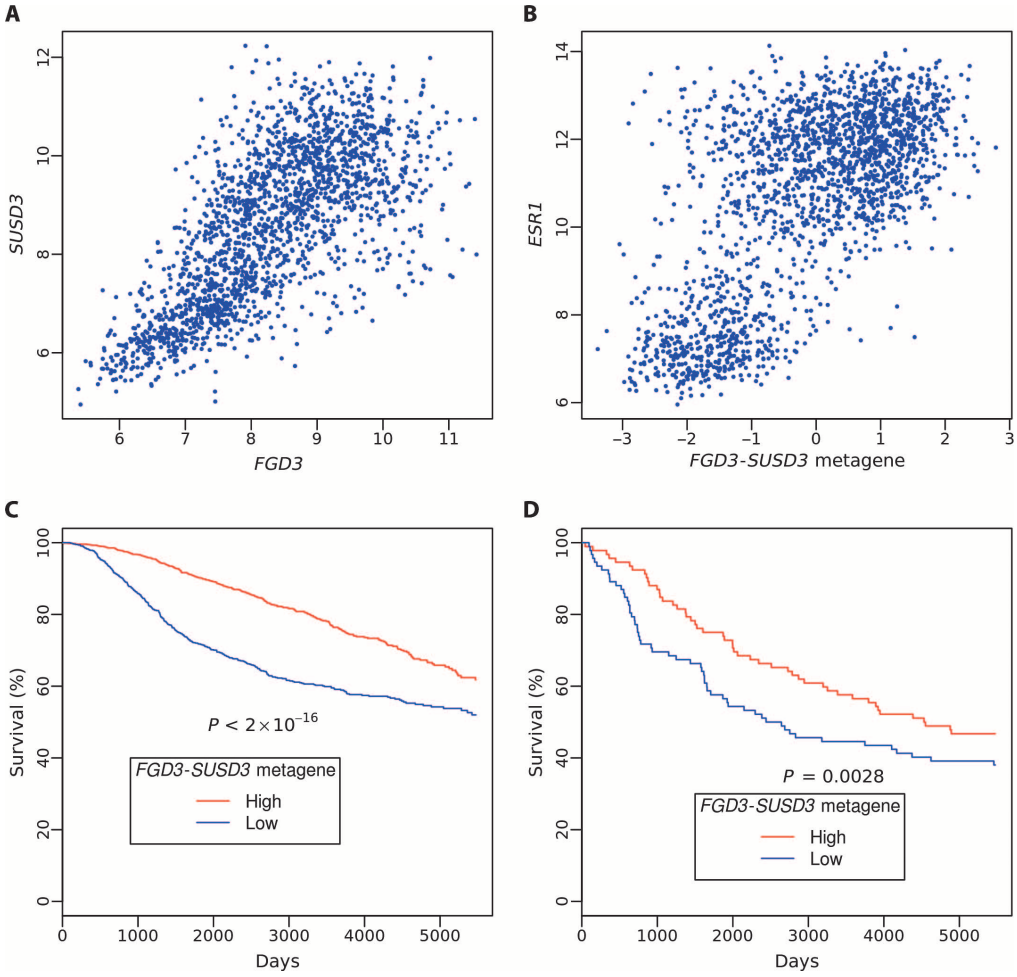

Fig. 3. *FGD3-SUSD3* metagene. (**A**) A scatter plot of the expression of *SUSD3* versus *FGD3* in the METABRIC data set shows a high variance in the expression of both genes at high expression levels. On the other hand, low expression of one strongly suggests low expression of the other in breast tumors. (**B**) ER-negative breast tumors tended not to express the *FGD3-SUSD3* metagene, whereas ER-positive breast tumors may or may not express the *FGD3-SUSD3* metagene. (**C** and **D**) Kaplan-Meier cumulative survival curves of breast cancer patients over a 15-year period on the basis of *FGD3-SUSD3* metagene expression in the (C) METABRIC and (D) OsloVal data sets. Patients were divided into equal-sized high and low subgroups according to their ranking with respect to *FGD3-SUSD3* metagene expression values. Low levels of *FGD3-SUSD3* metagene expression were associated with poor survival in both data sets. *P* values derived using the log-rank test in the two data sets were less than 2×10^{-16} and 0.0028, respectively.

prediction of survival. We refer to these as "submodels," which were eventually combined into one "ensemble" model.

The choices of parameters and prognostic methods used in the development of each submodel were made by trial and error search. Specifically, we made several submissions to the leaderboard using initial guesses about which combinations of features would be most prognostic and observed the resulting CIs; we also carried out our own cross-validation experiments (that is, we randomly partitioned the available samples into a training set and a validation set, trained the model accordingly, and recorded the CI in the validation set).

Figure 4 shows the Kaplan-Meier cumulative survival curves for our final ensemble prognostic model using the OsloVal data set (the *P* value derived from the log-rank test was lower than the minimum computable one, which was 2×10^{-16} using log-rank test), comparing patients with "poor" and "good" predicted survival according to the ranking assigned by the model, which was trained on the METABRIC data set.

The corresponding CI of the final ensemble model in the OsloVal data set was 0.7562. To test whether three of our features—CIN, MES, and LYM—contributed toward increasing the CI for our model using the OsloVal data set, we evaluated the CIs after removing each feature separately and retraining the model on the METABRIC data set without it. The resulting CI after removing the CIN feature and keeping the MES and LYM features was 0.7526, the CI after removing the MES feature and keeping the CIN and LYM features was 0.7514, and the CI after removing the LYM feature and keeping the CIN and MES features was 0.7488. In all cases, the CI was lower than that of the ensemble model. These results are consistent with our hypothesis that each of these three attractor metagenes provides information useful for breast cancer prognosis.

Comparison with random gene expression signatures

Venet *et al.* recently observed that randomly chosen gene expression signatures may often be significantly associated with breast cancer outcome (*12*). To explain this phenomenon, the authors introduced a specially defined proliferation signature—called meta-PCNA—which consists of 127 genes whose expression levels were most positively correlated with that of the proliferation marker *PCNA*, as determined from a gene expression data set of normal tissues. They observed that the meta-PCNA signature, although derived from an analysis of normal tissues, was prognostic for breast cancer outcome, and that the expression levels of many other genes were also associated with the meta-PCNA signature to varying degrees. Thus, they explained the observed association of random signatures with breast cancer outcome by the fact that several member genes of such random signatures are likely to be associated with those prognostic genes.

The meta-PCNA signature is highly similar to our own mitotic CIN attractor metagene. Indeed, 39 of the 127 genes in the meta-PCNA signature are among the 100 top-ranked genes of the CIN attractor metagene (*2*) (the *P* value for such overlap is 1.07×10^{-54} based on Fisher's exact test). Furthermore, 7 of the 10 genes (*CENPA*, *MELK*, *KIF2C*, *KIF20A*, *KIF4A*, *CCNA2*, and *CCNB2*) of our CIN feature used in the Challenge are among the 127 genes of the meta-PCNA signature.

Therefore, both the meta-PCNA signature, which was derived from normal tissue analysis, and the mitotic CIN attractor metagene, which was derived from a multicancer analysis, can be used to explain the observed phenomenon that random gene expression signatures are associated with breast cancer outcome. To compare the mitotic CIN attractor metagene with the meta-PCNA signature, we evaluated the corresponding CIs for the two breast cancer data sets (NKI and Loi) used in the meta-PCNA study (*12*), for the METABRIC data set using both DS- and OS-based survival data, and for the OsloVal data set. In all five cases, the CIs of the CIN feature were slightly higher than those of the meta-PCNA signature (Table 2). We hypothesize that the large "mitotic" component of the mitotic CIN attractor metagene is not exclusively cancer-associated, but it is also found in normal cells. By contrast, we hypothesize that the "chromosomal instability" component of the mitotic CIN attractor metagene is cancer-related and may account for the observed slightly higher association with survival compared with the meta-PCNA signature. Furthermore, the performance of our ensemble model with the OsloVal data set was higher than that of the CIN metagene alone.

DISCUSSION

Even though we used features discovered previously from an unsupervised and multicancer analysis without using the METABRIC

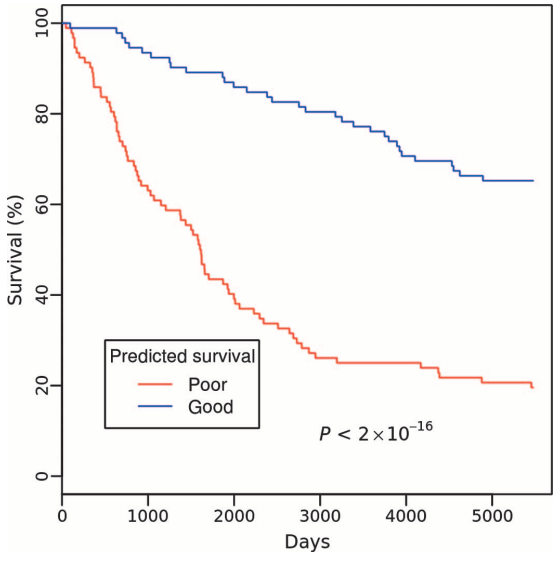

Fig. 4. Final ensemble model. Shown are Kaplan-Meier cumulative survival curves of breast cancer patients over a 15-year period on the basis of the predictions made by the final ensemble model in the OsloVal data set. The patients were divided into equal-sized poor and good predicted survival subgroups according to the ranking assigned by the final model, which was trained on the METABRIC data set. The *P* value derived using the log-rank test was less than 2×10^{-16}.

Table 2. CIs of the CIN feature and meta-PCNA index in four breast cancer data sets.

	CIN feature	Meta-PCNA index
NKI	0.725	0.717
Loi	0.675	0.662
METABRIC: DS-based	0.648	0.635
METABRIC: OS-based	0.605	0.595
OsloVal	0.579	0.554

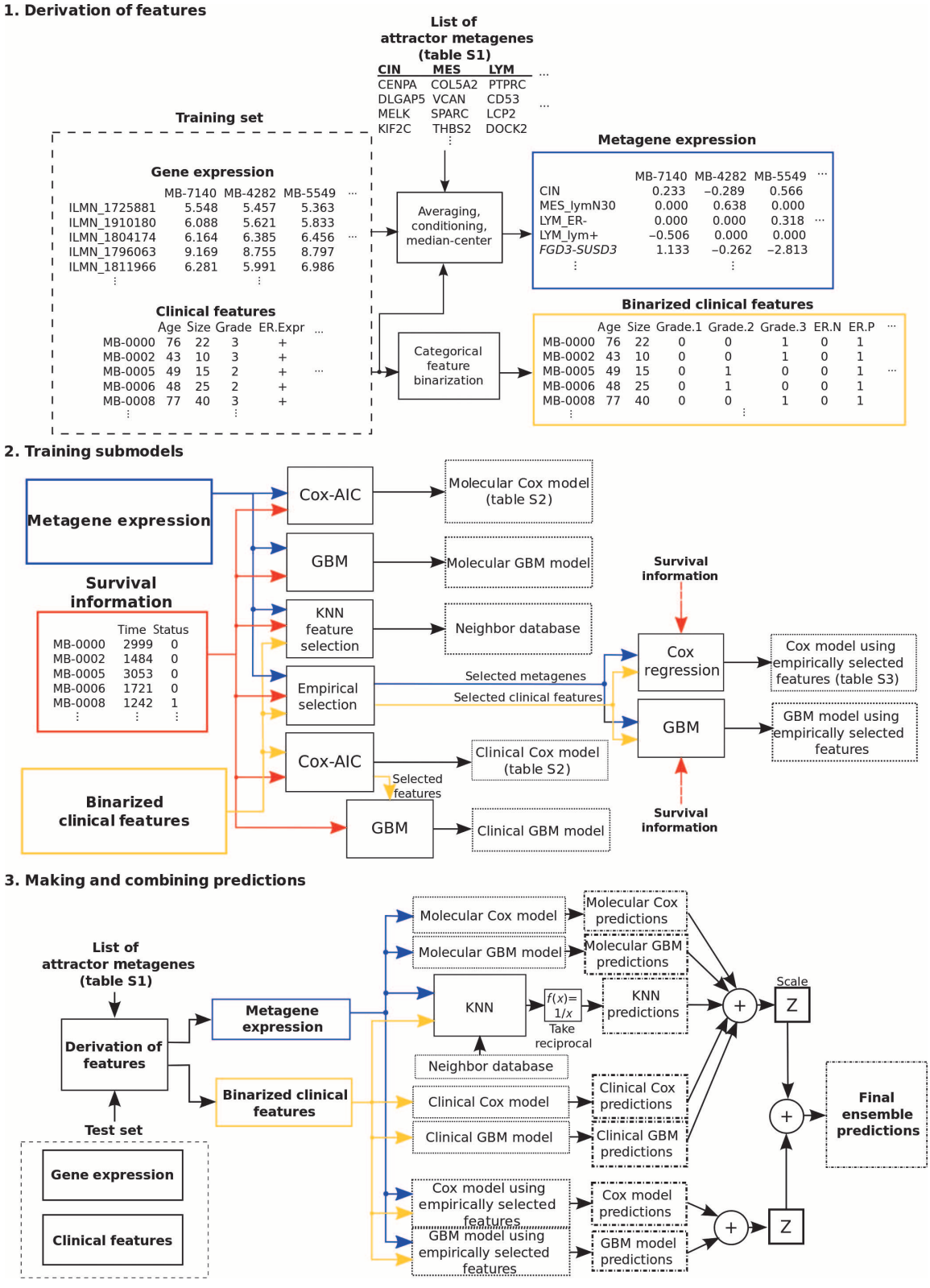

Fig. 5. Schematic of model development. Shown are block diagrams that describe the development stages for our final ensemble prognostic model. Building a prognostic model involves derivation of relevant features, training submodels and making predictions, and combining predictions from each submodel. Our model derived the attractor metagenes using gene expression data, combined them with the clinical information through Cox regression, GBM, and KNN techniques, and eventually blended each submodel's prediction.

data set for training, our model proved highly predictive of survival in breast cancer within the context of the Challenge. Therefore, we hypothesize that these features represent important molecular events in cancer development and might be associated with cancer-related phenotypes other than survival, such as response to drugs. A clinical trial would be required to test whether our computational predictions have utility in the realm of patient care.

Several cancer-related gene signatures that share similarity with the mitotic CIN and MES attractor metagenes have been reported (*13*–*16*). The key advantage of the attractor metagenes is that they are sharply defined by independent analyses, after being discovered separately and in nearly identical form in multiple cancer types, and can thus point to the few top-ranked genes for each attractor metagene. In the short term, these select genes can be tested for their ability to improve the performance of current cancer biomarker products. Existing clinical biomarker products include some genes that are components of attractor metagene signatures but do not rank at the top of their corresponding ranked list of genes. For example, the *CENPA*, *PRC1*, and *ECT2* genes are among those used in Agendia's MammaPrint (*17*) breast cancer assay, and *CCNB1*, *BIRC5*, *AURKA*, *MKI67*, and *MYBL2* are used in Genomic Health's Oncotype DX (*18*) assay for breast cancer. All eight of these genes are included in the ranked list of the top 100 genes of the CIN attractor metagene (*2*). It would be reasonable to test whether replacing such genes with a choice that more closely represents the mitotic CIN attractor metagene would improve the accuracy of these products.

In the longer term, study of these top-ranked genes may provide opportunities for uncovering new molecular mechanisms of cancer biology. Cross-disciplinary collaborations among molecular biologists and molecular geneticists, cancer researchers and clinicians, systems biologists, immunologists, and drug discovery scientists that are aimed at scrutinizing the attractor metagenes may reveal new methods for therapeutic intervention.

We previously found that the MES attractor metagene is strongly expressed in human cancer cells, but never in mouse stromal cells, after implanting pure human (neuroblastoma) cancer cell lines in immunodeficient mice (*19*). For this reason, we hypothesize that related gene expression signatures [often associated with drug response (*16*) or survival (*10*) in various cancer types] that appear to be of stromal origin because they contain fibroblastic markers may, in fact, be associated with the invasive cancer cells themselves.

Breast cancer has been classified into four main subtypes (*20*), which were also provided as clinical annotation features in the Challenge. However, we found that subtype identification did not impart any additional prognostic power to our model, despite our best efforts to incorporate subtype features in our model in various ways.

Notably absent from our selected features are copy number variations (CNVs), although such data were provided in uniformly renormalized form for both the METABRIC and OsloVal data sets. We tried to include CNVs and found that they did not improve performance in the presence (but not in the absence) of the CIN attractor metagene. We were aware that a CNV-based "genomic instability index" (GII) was used as part of a milestone performance before the start of the Challenge. However, we found that the inclusion of the CIN expression–based feature nullified the prognostic ability of GII as well as of all the individual CNVs that we tried. Even for the amplicons, we found that the corresponding expression-based attractor metagenes consistently had higher prognostic ability compared to any kind of CNV-based features that we tried. Therefore, we speculate that (i) the components of the mitotic CIN metagene play fundamental biological roles that function upstream of biological aberrations caused by genomic alterations in cancer, and (ii) the biological effects of CNVs are more directly manifested by the expression of a few highly ranked genes in the corresponding amplicon attractor than by the presence of CNVs in the corresponding genomic region.

MATERIALS AND METHODS

Data used by Challenge participants can be found at:
METABRIC: https://synapse.prod.sagebase.org/#Synapse:syn1688369
OsloVal: https://synapse.prod.sagebase.org/#Synapse:syn1688370

A general overview of building a prognostic model

Building our prognostic model involved derivation and selection of relevant features, training the submodels using the derived features based on survival information, and combining predictions from the submodels to produce a robust ensemble prediction. Figure 5 shows block diagrams describing our model. Each subhead in the figure corresponds to the section with the same subhead that follows. The source code of the model is available on Sage Synapse under ID syn1417992 and in table S4.

Derivation of features

We reduced the number of potential molecular features by preselecting the 12 features shown in table S1. We chose these features by trial and error after several experiments, including and removing features and evaluating the performance on the Challenge leaderboard and in cross-validation. The set of features used in the final model included (i) the three attractor metagenes and the *FGD3-SUSD3* metagene described in Results; (ii) the chr8q24.3 amplicon attractor metagene (because we had found it to be the most prominent amplicon in all cancer types we had considered and in the METABRIC training data set) (*2*) and the chr15q26.1 amplicon attractor metagene (because we had found it to be the most prognostic amplicon in the METABRIC training data set); (iii) three breast cancer–specific attractor metagenes (the ER metagene, the adipocyte metagene, and the *HER2* metagene); and (iv) two additional metagenes—*ZMYND10* metagene and the *PGR-RAI2* metagene—because we observed that their inclusion often improved performance. Both the *ZMYND10* and *PGR-RAI2* metagenes were protective (their individual CIs in all breast cancer data sets were less than 0.5). The rationale for considering these metagenes was that we wished to include additional protective features, and these ones were highly protective and at the same time not positively correlated with the most protective feature, the *FGD3-SUSD3* metagene.

Each metagene feature used in our model was defined by the average expression value of each of the 10 top-ranked genes in each attractor metagene. If, however, some of these 10 genes had mutual information with the metagene—as defined in (*2*)—that was less than 0.5, it was removed from consideration when deriving the metagene feature. If a gene was profiled by multiple probes—a collection of micrometer beads that bind a specific nucleic acid sequence—we selected the probe with the highest degree of coexpression with the metagene. The selection was done by applying the iterative attractor-finding algorithm (*2*) on all the probes for the top 10 genes and selecting the top-ranked probe for each gene. The expression values of each metagene feature were median-centered by subtracting their median value.

All the categorical—nonnumerical, such as histological type—variables in the clinical data were binarized by representing each category by a binary variable. In that case, missing values were assigned zero in each binary variable. For example, the categorical variable ER_IHC_status (a variable that describes the immunochemistry status of ER) was binarized into two binary variables: ER-positive (ER.P) and ER-negative (ER.N). ER-positive patients were assigned [1, 0] for these two variables, ER-negative patients were assigned [0, 1], and patients with missing ER status were uniquely assigned [0, 0]. Missing values in numerical variables were imputed by the average of the nonmissing values across all samples.

Conditioning of metagene features. We used three conditioned metagene features in our model: the MES feature conditioned on tumor sizes of less than 30 mm and no positive lymph nodes, the LYM feature conditioned on ER-negative patients, and the LYM feature conditioned on patients with more than three positive lymph nodes. We conditioned the features by median-centering the metagene's expression values of the subgroup of samples, satisfying the condition using the subgroup's median, and setting the values of the remaining samples to zero.

Training submodels and making predictions

A prognostic model selects particular features out of the set of derived features and combines them using an algorithm for optimally fitting the given survival information. Our ensemble model consisted of several such submodels. The choice of these models, described below, was made on a trial and error basis depending on the occasional leaderboard scores of other Challenge participants and our own cross-validation scores.

Cox regression based on Akaike Information Criterion. The Cox proportional hazards model relates the effect of a unit increase in a covariate to the hazard ratio (*21*). To select from derived features as covariates in the regression model, we performed stepwise selection based on Akaike Information Criterion (AIC) (*22*). In each step, we selected the feature with the lowest AIC measure. The Cox-AIC model makes predictions by computing fitted values of the given features to the regression model.

We used AIC for feature selection on molecular features and clinical features separately to fit Cox proportional hazards models. The molecular and clinical features selected by the Cox-AIC model applied to the METABRIC data set are given in table S2. The predictions made by the two separate models were combined by summation.

Generalized boosted regression models. The generalized boosted regression model (GBM) adopts the exponential loss function used in the AdaBoost algorithm (*23*) and uses Friedman's gradient descent algorithm accompanied by subsampling to improve predictive performance and reduce computational time (*24*).

We trained GBMs on molecular features and clinical features separately, as we did for the Cox-AIC models. We used only the clinical features that were selected by the Cox-AIC model (listed in table S2) as input to the GBM. We performed fivefold cross-validation to determine the best number of trees in the model. The tree depth was set to the number of significant explanatory variables in the Cox-AIC model ($P < 0.05$ based on t test). The predicted values made by the two separated models were combined by summation.

K-nearest neighbor model. We used a modified version of the K-nearest neighbor (KNN) model (*25*) for survival prediction in our model. We selected the features whose values defined patients' ranking with CI greater than 0.6 or less than 0.4 in the training set.

When making predictions, we computed the Euclidean distance in the selected feature space between the patient with unknown survival and each deceased patient in the training set. The top 10% of the deceased patients with smallest distances, defined as the "nearest neighbors," were used to make predictions. The predictions were made by taking the weighted average of the survival times of the nearest neighbors, where the weight of a neighbor was the reciprocal of the distances between the neighbor and the patient with unknown survival.

Combination of Cox regression and GBM applied on empirically selected features. We observed that the performance of the overall model was improved by incorporating a submodel constrained to include the four fundamental molecular features described in Results (CIN, MES constrained to a tumor size less than 30 mm with no positive lymph node, LYM constrained to ER-negative patients, and the *FGD3-SUSD3* metagene) together with very few clinical features, which, by trial and error search, we determined to be the number of positive lymph nodes and the age at diagnosis. The selected features were used to fit a Cox regression model and a GBM, whose predictions were combined by summation. The Cox proportional hazards model trained on these features is given in table S3.

Combination of predictions

Our final model contained the submodels described above. We added the resulting predictions from Cox-AIC and GBM, as well as the reciprocal of the predicted survival time given by the KNN model, and we divided the result by the corresponding SD. We also did the same normalization on the predictions derived from submodel 4 above, and the final ensemble prediction was the summation of these two.

Combination of OS- and DS-based predictions

Our best performance on the leaderboard was achieved when we trained our models twice, once using OS-based survival data and again using DS-based survival data, and then combining the two predictions. Therefore, we adopted the ensemble model depicted in Fig. 5. We combined these two sets of predictions by taking the weighted average of the two. The weights were determined by maximizing the CI with OS in the training set with a heuristic optimization technique.

SUPPLEMENTARY MATERIALS

www.sciencetranslationalmedicine.org/cgi/content/full/5/181/181ra50/DC1

Table S1. Molecular features used in the model.

Table S2. Cox proportional hazards model trained on molecular and clinical features on the basis of AIC.

Table S3. Cox proportional hazards models trained on empirically selected features.

Table S4. Source code of the final model.

REFERENCES AND NOTES

1. 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. Borresen-Dale, Systematic analysis of challenge-driven improvements in molecular prognostic models for breast cancer. *Sci. Transl. Med.* **5**, 181re1 (2013).
2. W. Y. Cheng, T. H. Yang, D. Anastassiou, Biomolecular events in cancer revealed by attractor metagenes. *PLoS Comput. Biol.* **9**, e1002920 (2013).
3. D. Hanahan, R. A. Weinberg, Hallmarks of cancer: The next generation. *Cell* **144**, 646–674 (2011).
4. 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. Borresen-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).
5. B. H. Mecham, P. S. Nelson, J. D. Storey, Supervised normalization of microarrays. *Bioinformatics* **26**, 1308–1315 (2010).
6. M. J. Pencina, R. B. D'Agostino, Overall C as a measure of discrimination in survival analysis: Model specific population value and confidence interval estimation. *Stat. Med.* **23**, 2109–2123 (2004).
7. T. M. Cover, J. O. Y. A. Thomas, *Elements of Information Theory* (Wiley-Interscience, Hoboken, ed. 2, 2006), pp. xxiii, 748.
8. E. L. Kaplan, P. Meier, Nonparametric estimation from incomplete observations. *J. Am. Stat. Assoc.* **53**, 457–481 (1958).
9. H. Kim, J. Watkinson, V. Varadan, D. Anastassiou, Multi-cancer computational analysis reveals invasion-associated variant of desmoplastic reaction involving INHBA, THBS2 and COL11A1. *BMC Med. Genomics* **3**, 51 (2010).
10. R. W. Tothill, A. V. Tinker, J. George, R. Brown, S. B. Fox, S. Lade, D. S. Johnson, M. K. Trivett, D. Etemadmoghadam, B. Locandro, N. Traficante, S. Fereday, J. A. Hung, Y. E. Chiew, I. Haviv, D. Gertig, A. DeFazio, D. D. Bowtell, Novel molecular subtypes of serous and endometrioid ovarian cancer linked to clinical outcome. *Clin. Cancer Res.* **14**, 5198–5208 (2008).
11. H. Wang, B. Wei, G. Bismuth, C. E. Rudd, SLP-76-ADAP adaptor module regulates LFA-1 mediated costimulation and T cell motility. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 12436–12441 (2009).
12. D. Venet, J. E. Dumont, V. Detours, Most random gene expression signatures are significantly associated with breast cancer outcome. *PLoS Comput. Biol.* **7**, e1002240 (2011).
13. C. Sotiriou, P. Wirapati, S. Loi, A. Harris, S. Fox, J. Smeds, H. Nordgren, P. Farmer, V. Praz, B. Haibe-Kains, C. Desmedt, D. Larsimont, F. Cardoso, H. Peterse, D. Nuyten, M. Buyse, M. J. Van de Vijver, J. Bergh, M. Piccart, M. Delorenzi, Gene expression profiling in breast cancer: Understanding the molecular basis of histologic grade to improve prognosis. *J. Natl. Cancer Inst.* **98**, 262–272 (2006).
14. S. L. Carter, A. C. Eklund, I. S. Kohane, L. N. Harris, Z. Szallasi, A signature of chromosomal instability inferred from gene expression profiles predicts clinical outcome in multiple human cancers. *Nat. Genet.* **38**, 1043–1048 (2006).
15. E. Fredlund, J. Staaf, J. K. Rantala, O. Kallioniemi, A. Borg, M. Ringnér, The gene expression landscape of breast cancer is shaped by tumor protein p53 status and epithelial-mesenchymal transition. *Breast Cancer Res.* **14**, R113 (2012).
16. P. Farmer, H. Bonnefoi, P. Anderle, D. Cameron, P. Wirapati, V. Becette, S. André, M. Piccart, M. Campone, E. Brain, G. Macgrogan, T. Petit, J. Jassem, F. Bibeau, E. Blot, J. Bogaerts, M. Aguet, J. Bergh, R. Iggo, M. Delorenzi, A stroma-related gene signature predicts resistance to neoadjuvant chemotherapy in breast cancer. *Nat. Med.* **15**, 68–74 (2009).
17. L. J. van 't Veer, H. Dai, M. J. van de Vijver, Y. D. He, A. A. 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).
18. 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).
19. D. Anastassiou, V. Rumjantseva, W. Cheng, J. Huang, P. D. Canoll, D. J. Yamashiro, J. J. Kandel, Human cancer cells express Slug-based epithelial-mesenchymal transition gene expression signature obtained in vivo. *BMC Cancer* **11**, 529 (2011).
20. The Cancer Genome Atlas Network, Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61–70 (2012).
21. P. K. Andersen, R. D. Gill, Cox's regression model for counting processes: A large sample study. *Ann. Stat.* **10**, 1100–1120 (1982).
22. Y. Sakamoto, M. Ishiguro, G. Kitagawa, *Akaike Information Criterion Statistics* (D. Reidel Publishing Company, Dordrecht, 1986).
23. Y. Freund, R. E. Schapire, A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.* **55**, 119–139 (1997).
24. J. H. Friedman, Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **29**, 1189–1232 (2001).
25. W. N. Venables, B. D. Ripley, *Modern Applied Statistics with S* (Springer, New York, ed. 4, 2002).

Acknowledgments: We are grateful to the organizers and donors who made the Sage Bionetworks–DREAM Breast Cancer Prognosis Challenge possible. We thank J. Kandel and D. Yamashiro of Columbia University's Medical Center for helpful discussions and insightful suggestions. **Funding:** This work was funded by Columbia University's unrestricted-purpose allocation of inventor's (D.A.) research of royalties resulting from intellectual property totally unrelated to the work described in this paper. **Author contributions:** Study concept and design: W.-Y.C. and D.A.; acquisition of data and statistical analysis: W.-Y.C. and T.-H.O.Y.; analysis and interpretation of data: W.-Y.C., T.-H.O.Y., and D.A.; drafting of the manuscript: W.-Y.C. and D.A. **Competing interests:** D.A. is coauthor on a patent application (PCT/US11/32356) for biomarkers based on a multicancer invasion–associated mechanism; D.A. and W.-Y.C. are coauthors on a patent application (PCT/US12/45958) for biomarkers, methods, and compositions for inhibiting a multicancer mesenchymal transition mechanism, both filed by Columbia University.

Submitted 20 February 2013
Accepted 2 April 2013
Published 17 April 2013
10.1126/scitranslmed.3005974

Citation: 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).

Science Translational Medicine

Development of a Prognostic Model for Breast Cancer Survival in an Open Challenge Environment

Wei-Yi Cheng, Tai-Hsien Ou Yang and Dimitris Anastassiou

Sci Transl Med **5**, 181ra50181ra50.
DOI: 10.1126/scitranslmed.3005974

DREAMing of Biomedicine's Future

Although they no longer live in the lab, scientific editors still enjoy doing experiments. The simultaneous publication of two unusual papers offered *Science Translational Medicine*'s editors the chance to conduct an investigation into peer-review processes for competition-based crowdsourcing studies designed to address problems in biomedicine. In a Report by Margolin *et al*. (which was peer-reviewed in the traditional way), organizers of the Sage Bionetworks/DREAM Breast Cancer Prognosis Challenge (BCC) describe the contest's conception, execution, and insights derived from its outcome. In the companion Research Article, Cheng *et al*. outline the development of the prognostic computational model that won the Challenge. In this experiment in scientific publishing, the rigor of the Challenge design and scoring process formed the basis for a new style of publication peer review.

DREAM—Dialogue for Reverse Engineering Assessments and Methods—conducts a variety of computational Challenges with the goal of catalyzing the "interaction between theory and experiment, specifically in the area of cellular network inference and quantitative model building in systems biology." Previous Challenges involved, for example, modeling of protein-protein interactions for binding domains and peptides and the specificity of transcription factor binding. In the BCC—which was a step in the translational direction—participants competed to create an algorithm that could predict, more accurately than current benchmarks, the prognosis of breast cancer patients from clinical information (age, tumor size, histological grade), genome-scale tumor mRNA expression data, and DNA copy number data. Participants were given Web access to such data for 1981 women diagnosed with breast cancer and used it to train computational models that were then submitted to a common, open-access computational platform as re-runnable source code. The predictive value of each model was assessed in real-time by calculating a concordance index (CI) of predicted death risks compared to overall survival in a held-out data set, and CIs were posted on a public leaderboard.

The winner of the Challenge was ultimately determined when a select group of top models were validated in a new breast cancer data set. The winning model, described by Cheng *et al*., was based on sets of genes (signatures)—called attractor metagenes—that the same research group had previously shown to be associated, in various ways, with multiple cancer types. Starting with these gene sets and some other clinical and molecular features, the team modeled various feature combinations, selecting ones that improved performance of their prognostic model until they ultimately fashioned the winning algorithm.

...