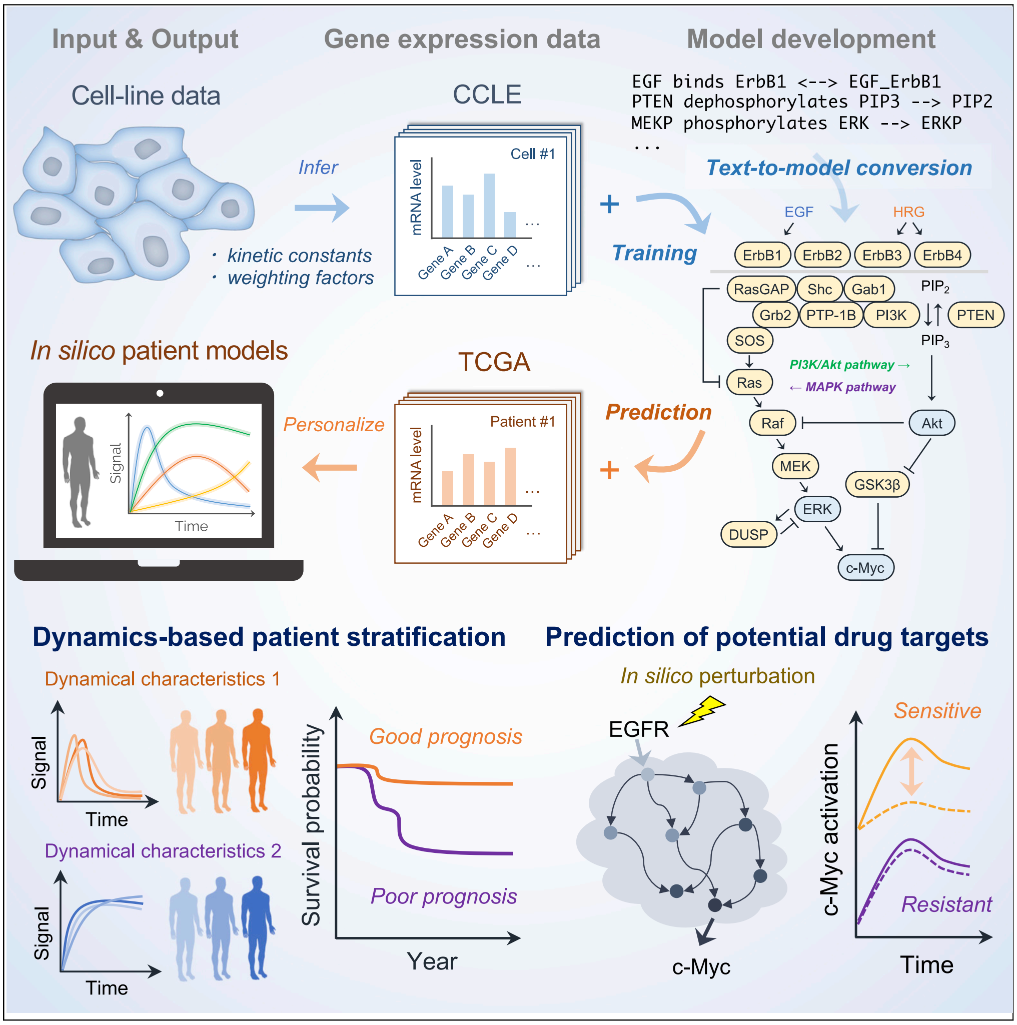
iScience

CellPress OPEN ACCESS

Article

A text-based computational framework for patient -specific modeling for classification of cancers



```
EGF binds ErbB1 <--> EGF_ErbB1
PTEN dephosphorylates PIP3 --> PIP2
MEKP phosphorylates ERK --> ERKP
...
```

Hiroaki Imoto, Sawa Yamashiro, Mariko Okada

mokada@protein.osaka-u.ac.jp

Highlights

A text file describing biochemical systems is converted into an executable model

Patient-specific models incorporate individual gene expression profiles

In silico signaling dynamics can be utilized as prognostic biomarkers

Personalized kinetic models are capable of predicting potential drug targets

Imoto et al., iScience *25*, 103944
March 18, 2022 © 2022 The Author(s).
https://doi.org/10.1016/j.isci.2022.103944

iScience

CellPress OPEN ACCESS

Article

A text-based computational framework for patient -specific modeling for classification of cancers

Hiroaki Imoto,[1,3] Sawa Yamashiro,[1,3] and Mariko Okada[1,2,4,*]

SUMMARY

Patient heterogeneity precludes cancer treatment and drug development; hence, development of methods for finding prognostic markers for individual treatment is urgently required. Here, we present Pasmopy (Patient-Specific Modeling in Python), a computational framework for stratification of patients using *in silico* signaling dynamics. Pasmopy converts texts and sentences on biochemical systems into an executable mathematical model. Using this framework, we built a model of the ErbB receptor signaling network, trained in cultured cell lines, and performed *in silico* simulation of 377 patients with breast cancer using The Cancer Genome Atlas (TCGA) transcriptome datasets. The temporal dynamics of Akt, extracellular signal-regulated kinase (ERK), and c-Myc in each patient were able to accurately predict the difference in prognosis and sensitivity to kinase inhibitors in triple-negative breast cancer (TNBC). Our model applies to any type of signaling network and facilitates the network-based use of prognostic markers and prediction of drug response.

INTRODUCTION

Cancer is a heterogeneous disease in terms of mutation signatures, gene expression profiles, and response to drug treatments (Dagogo-Jack and Shaw, 2018). Innovations in sequencing, genome-wide measurements of mutations and transcriptomics profiles (Gusev et al., 2016; Ozaki et al., 2002) have brought more attention to inter-patient heterogeneity. Accordingly, different types of data-driven algorithms, such as machine learning methods (Kourou et al., 2015; Van't Veer et al., 2002), have been developed to identify correlations between these gene signatures and clinical outcomes. Despite these efforts, the molecular mechanisms by which different genomic and transcriptomic profiles predict distinct patient-specific prognostic outcomes remain poorly understood.

Mechanistic descriptions of biological network using ordinary differential equations (ODEs) is considered one of the promising approaches to uncover the regulatory mechanisms in biological systems (Clarke and Fisher, 2020; Kholodenko, 2006). Several attempts have focused on pan-cancer signaling networks to explore the mechanisms underlying heterogeneous responses in cancer (Fröhlich et al., 2018; Hass et al., 2017), by combining mechanistic modeling with transcriptome profiles obtained from the cancer cell lines (Barretina et al., 2012). In these studies, experimental data on signaling activities, cell growth, and drug response from more than 100 cell lines were used for model prediction, and training the model with the datasets allowed it to accurately predict cell-specific drug response from the untrained data (Fröhlich et al., 2018; Hass et al., 2017). Accordingly, these studies using cell line profiles suggest the potential of "patient-specific models" (Saez-Rodriguez and Blüthgen, 2020) that can determine personalized prognosis and drug response using the patient's signaling and transcriptome profiles. However, there are several challenges to overcome. Although clinical transcriptome data are available from public databases, obtaining signaling activity from each patient is not feasible due to the difficulty of culturing cells from cancer tissues (Inoue et al., 2017; Whittle et al., 2015; Yoshida, 2020). Additionally, the drug responses predicted by patient-specific models cannot be immediately tested in living patients. In addition, as another fundamental issue, mathematical modeling usually requires specific mathematical expertise of users. To be able to apply mathematical modeling to patient data analysis, we need a simpler, readable format tool that many biologists can use for cancer classification.

To resolve these problems, we developed a computational framework called Pasmopy (Patient-Specific Modeling in Python). Pasmopy enables the conversion of text describing biochemical reactions (such as

[1] Institute for Protein Research, Osaka University, Suita, Osaka 565-0871, Japan

[2] Center for Drug Design and Research, National Institutes of Biomedical Innovation, Health and Nutrition, Ibaraki, Osaka 567-0085, Japan

[3] These authors contributed equally

[4] Lead contact

*Correspondence: mokada@protein.osaka-u.ac.jp

https://doi.org/10.1016/j.isci.2022.103944

association, phosphorylation, and degradation) in signaling networks into ordinary differential equation (ODE) models, without mathematical knowledge of users. It also offers several biologist-friendly functions, such as parameterization of patient models against the learning datasets obtained from cultured cell lines, individualization of mechanistic models by incorporating cell-line- or patient-specific gene expression data, prediction of patient prognosis based on simulation outputs, the ability to investigate the molecular mechanisms underlying patient outcomes, and the ability to identify potential drug targets for individual patients.

Using this tool, we developed a personalized model of ErbB receptor signaling network. The model includes a series of biochemical reactions involved in ErbB receptor activation and c-Myc induction (Arteaga and Engelman, 2014; Xu et al., 2010). By combining 377 individual patient transcriptome datasets obtained from The Cancer Genome Atlas (TCGA) (Weinstein et al., 2013) and personalized models, we succeeded in classifying patients with triple-negative breast cancer (TNBC) into poor and better prognosis groups, based solely on *in silico* Akt, extracellular signal-regulated kinase (ERK), and c-Myc dynamics of each patient. Our models suggested that these subclusters can be classified by a simple metric: the epidermal growth factor receptor (EGFR / ErbB1) expression ratio to other ErbB receptor families. Further analysis of the models implied that patients with poorer prognoses are more resistant to treatments targeting the EGFR. We also confirmed that the same model could stratify patients with colon cancer (Muzny et al., 2012) based on predicted *in silico* signaling dynamics, indicating that these two cancers share common regulatory mechanisms in the signaling network that determine prognosis.

RESULTS

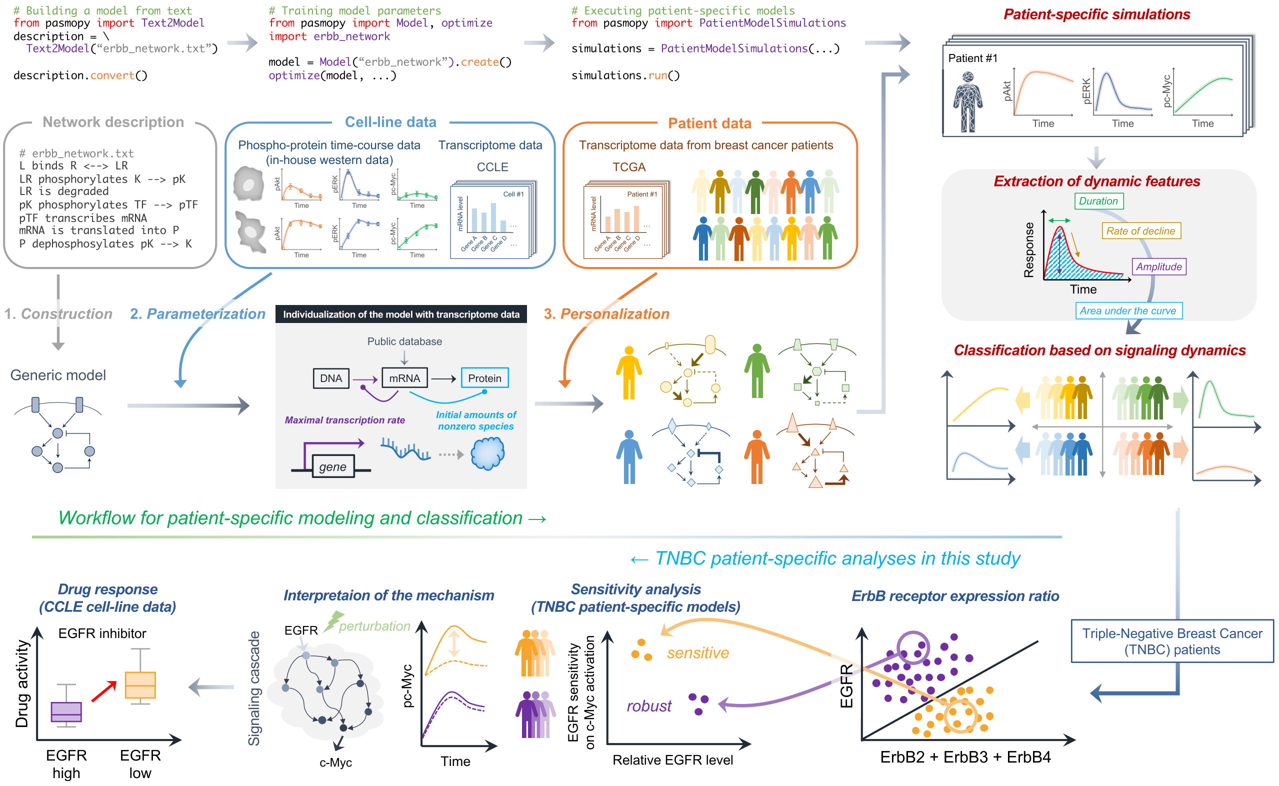
Development of Pasmopy: a scalable computational toolkit for patient-specific modeling

The dynamics of signaling pathways play key roles in determining cell fate and cancer progression (Purvis and Lahav, 2013). Therefore, the experimental analysis of patient response data is primarily required for development of drugs targeting these pathways (Zhong et al., 2021). However, analyzing signaling dynamics using the patient tissues is generally difficult even using advanced techniques such as the patient-derived xenograft (PDX) model. This is due to the limitations of the current PDX models, including the inability to reconstitute human immune cell systems, low success rates, and a high cost to establish cell lines and maintain the original cell properties (Inoue et al., 2017; Whittle et al., 2015; Yoshida, 2020). To tackle this problem, we developed Pasmopy, a scalable toolkit for *in silico* patient-specific mathematical modeling (Figure 1). Pasmopy offers the following unique features: (i) construction of mechanistic models from texts and sentences of gene regulatory network without a knowledge of mathematical modeling (Figure 2A), (ii) personalization of the model using transcriptome data of each patient, (iii) prediction of patient outcome based on *in silico* signaling dynamics, e.g., amplitude, duration, and area under the curve (AUC), and (iv) sensitivity analysis for prediction of potential drug targets. Pasmopy currently contains a list of 14 reaction rules on gene regulation and biochemical reactions, including binding, dissociation, phosphorylation, transcription, translation, synthesis, degradation, and translocation, which can be automatically converted into kinetic equations (Figure 2B). New terminology of a reaction rule can also be added by users. Pasmopy is compatible with a Python framework for Modeling and Analysis of Signaling Systems (BioMASS) (Imoto et al., 2020), which allows parameterization and network analysis of large scale biological models, and more specialized for personalized modeling.

In this study, we constructed a mathematical model of ErbB receptor signaling network (Birtwistle et al., 2007) and c-Myc induction (Lee et al., 2008) using this tool (Figure 2C). The model includes activation and dimerization of four ErbB receptors (ErbB1 / EGFR, ErbB2, ErbB3, and ErbB4), Ras-ERK cascade, and the Akt-PI3K pathway, which was adapted from the model of Birtwistle et al. (Birtwistle et al., 2007) and integrated the process of c-Myc induction and stabilization by ERK and Akt signals, which was newly constructed for this study. The resulting model has 319 rate equations, 228 species, and 648 parameters. Of the 648 parameters such as kinetic constants and weighting factors, 220 were estimated from phospho-proteins time-course data obtained from four breast cancer cell lines stimulated with epidermal growth factor (EGF) or heregulin (HRG) for up to 120 min (see below and STAR methods section).

Using transcriptomic data to personalize the mechanistic model

Modeling biological systems usually requires initial abundances of chemical species in the model and kinetic parameters of the reaction. To determine the kinetic parameters of patient-specific models, we first assumed that the reaction parameters are unique to the molecular species involved in a reaction event and



```
# Building a model from text
from pasmopy import Text2Model
description = \
  Text2Model("erbb_network.txt")

description.convert()
```

```
# Training model parameters
from pasmopy import Model, optimize
import erbb_network

model = Model("erbb_network").create()
optimize(model, ...)
```

```
# Executing patient-specific models
from pasmopy import PatientModelSimulations

simulations = PatientModelSimulations(...)

simulations.run()
```

```
# erbb_network.txt
L binds R <--> LR
LR phosphorylates K --> pK
LR is degraded
pK phosphorylates TF --> pTF
pTF transcribes mRNA
mRNA is translated into P
P dephosphosylates pK --> K
```

Network description
Cell-line data
Phospho-protein time-course data
(in-house western data)
Transcriptome data
CCLE
pAkt
pERK
pc-Myc
Time
Cell #1
mRNA level
Gene A
Gene B
Gene C
Gene D
Patient data
Transcriptome data from breast cancer patients
TCGA
Patient #1
Patient-specific simulations
Patient #1
Extraction of dynamic features
Duration
Rate of decline
Amplitude
Area under the curve
Response
Classification based on signaling dynamics
1. Construction
2. Parameterization
Individualization of the model with transcriptome data
Public database
DNA
mRNA
Protein
Maximal transcription rate
Initial amounts of nonzero species
gene
3. Personalization
Generic model
Workflow for patient-specific modeling and classification →
← TNBC patient-specific analyses in this study
Drug response
(CCLE cell-line data)
EGFR inhibitor
Drug activity
EGFR high
EGFR low
Interpretaion of the mechanism
EGFR
perturbation
Signaling cascade
c-Myc
pc-Myc
Time
Sensitivity analysis
(TNBC patient-specific models)
sensitive
robust
EGFR sensitivity on c-Myc activation
Relative EGFR level
ErbB receptor expression ratio
EGFR
ErbB2 + ErbB3 + ErbB4
Triple-Negative Breast Cancer (TNBC) patients

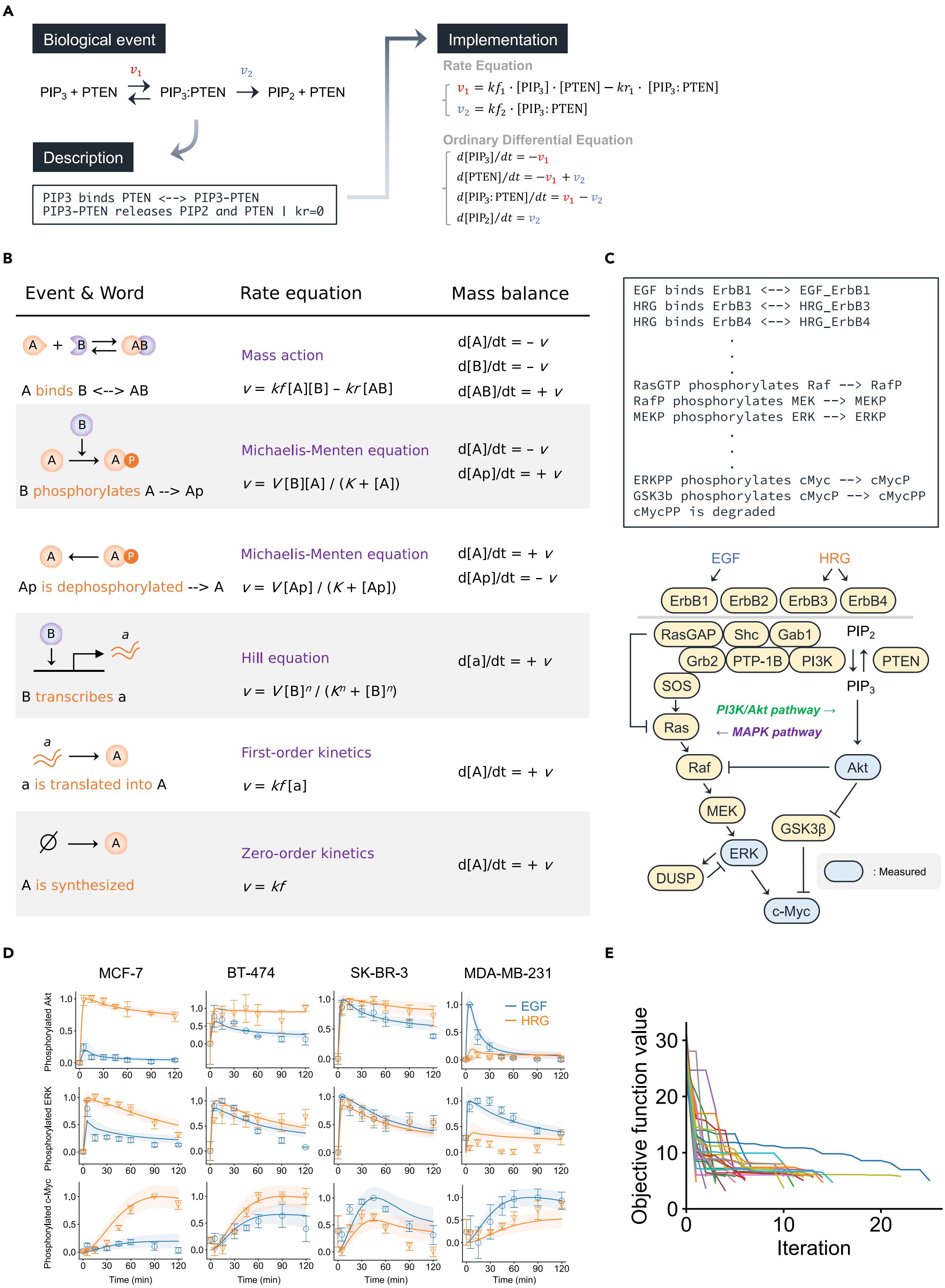
Figure 1. Overview of the workflow

A workflow for identifying cancer prognostic factors based on signaling dynamics from mechanistic modeling. A text file describing the biochemical reactions is converted into an executable model (1. Construction). The model parameters are trained on phospho-protein time-course data obtained from growth factor-stimulated cultured cell lines (2. Parameterization). The model is personalized by incorporating individual gene expression profiles (3. Personalization). The patients are classified based on in silico signaling responses from personalized simulations. In this study, the patient group with triple-negative breast cancer (TNBC) was further analyzed (bottom, right to left). Based on the examination of signaling properties, the patients with TNBC were classified into two subclusters by ErbB receptor expression ratios. Sensitivity analysis indicated that patients with higher EGFR expression ratios were less sensitive to EGFR inhibitors. This hypothesis was validated using drug-response data obtained from cancer cell lines.

remain the same even if gene mutations are present in the species. Instead, we assumed that such genomic mutations are reflected in the gene expression signatures. This assumption is empirically supported by expression quantitative trait loci (eQTLs) (Nica and Dermitzakis, 2013) analysis and transcriptome-wide association studies (TWAS) (Gusev et al., 2016) that links genomic mutations to gene expression signatures. Accordingly, unknown parameters of the model, that are common to all patients and cultured cell lines, were obtained by fitting the model to the phospho-protein time-course data obtained from the cultured cell lines.

In brief, the ErbB network model was trained against the growth factor-stimulated time-course datasets of phosphorylated Akt, ERK, and c-Myc obtained from MCF-7, BT-474, SK-BR-3, and MDA-MB-231 cancer cell lines (which represent four breast cancer subtypes: Luminal A, Luminal B, HER2+, and triple-negative, respectively) along with their corresponding the Cancer Cell Line Encyclopedia (CCLE) (Barretina et al., 2012) transcriptome data, which are used for determining nonzero initial conditions (protein levels of the species) in the model (see below and gene list in Table S1) (Figure 2D). By minimizing the objective function, i.e., the residual sum of squares between simulation and experimental measurements, 30 good fitting parameter sets were obtained that reproduced experimental observations in these four breast cancer cell lines (Figure 2E). These 30 parameters were also used as kinetic parameters for the patient model.

To personalize the model, individual TCGA transcriptome data were analyzed and used to infer the initial amount of nonzero species or maximal transcription rate for each patient model. Because we use the cultured cell line data to estimate the parameters of the model, we need to normalize the patient's



A
Biological event
PIP$_3$ + PTEN $\overset{v_1}{\rightleftarrows}$ PIP$_3$:PTEN $\xrightarrow{v_2}$ PIP$_2$ + PTEN
Description
Implementation

```
PIP3 binds PTEN <--> PIP3-PTEN
PIP3-PTEN releases PIP2 and PTEN | kr=0
```

Rate Equation

$$\begin{cases} v_1 = kf_1 \cdot [\mathrm{PIP_3}] \cdot [\mathrm{PTEN}] - kr_1 \cdot [\mathrm{PIP_3{:}PTEN}] \\ v_2 = kf_2 \cdot [\mathrm{PIP_3{:}PTEN}] \end{cases}$$

Ordinary Differential Equation

$$\begin{cases} d[\mathrm{PIP_3}]/dt = -v_1 \\ d[\mathrm{PTEN}]/dt = -v_1 + v_2 \\ d[\mathrm{PIP_3{:}PTEN}]/dt = v_1 - v_2 \\ d[\mathrm{PIP_2}]/dt = v_2 \end{cases}$$

Event & Word	Rate equation	Mass balance
A binds B <--> AB	Mass action $v = kf[A][B] - kr[AB]$	d[A]/dt = − v d[B]/dt = − v d[AB]/dt = + v
B phosphorylates A --> Ap	Michaelis-Menten equation $v = V[B][A] / (K + [A])$	d[A]/dt = − v d[Ap]/dt = + v
Ap is dephosphorylated --> A	Michaelis-Menten equation $v = V[Ap] / (K + [Ap])$	d[A]/dt = + v d[Ap]/dt = − v
B transcribes a	Hill equation $v = V[B]^n / (K^n + [B]^n)$	d[a]/dt = + v
a is translated into A	First-order kinetics $v = kf[a]$	d[A]/dt = + v
A is synthesized	Zero-order kinetics $v = kf$	d[A]/dt = + v

```
EGF binds ErbB1 <--> EGF_ErbB1
HRG binds ErbB3 <--> HRG_ErbB3
HRG binds ErbB4 <--> HRG_ErbB4
            .
            .
            .
RasGTP phosphorylates Raf --> RafP
RafP phosphorylates MEK --> MEKP
MEKP phosphorylates ERK --> ERKP
            .
            .
            .
ERKPP phosphorylates cMyc --> cMycP
GSK3b phosphorylates cMycP --> cMycPP
cMycPP is degraded
```

Figure 2. Construction and parameterization of the mechanistic model

(A) The strategy for implementing ordinary differential equations (ODEs) from the text descriptions of the biological events.

(B) Representative biological events and words that can be converted into rate equations and ODEs.

(C) Cancer signaling network and its conversion into an ODE model in this study.

(D) The model parameter was trained on time-series Akt, ERK, and c-Myc phosphorylation levels obtained from four breast cancer cell lines: MCF-7, BT-474, SK-BR-3, and MDA-MB-231 stimulated with growth factors. The points (blue squares, EGF; orange triangles, HRG) denote experimental data, solid lines denote simulations, and shaded areas denote SD. For all panels, error bars denote SE for three independent experiments. (E) Objective function traces from 30 optimization runs.

transcriptome data to the data obtained from cultured cell lines for patient modeling. First, transcriptome profiles of 413 patients with breast cancer and 51 breast cancer cell lines were obtained from TCGA (Weinstein et al., 2013) and CCLE (Barretina et al., 2012), respectively, and their batch effects were removed using ComBat-seq (Zhang et al., 2020) (Figure S1). By performing this step, cell line transcriptomes could be merged with the patient transcriptomes and the parameters estimated from phospho-protein cell line data could also be used as patient parameters. Several samples were removed due to low total read counts of the sequence data (see STAR methods and Figure S1B), finally resulting in 377 patient data for modeling (Table 1). To make the models patient-specific, the clinical transcriptomic data for 38 genes (see gene list in Table S1) were incorporated as the maximum transcription rate or the initial number of nonzero species in the model (see STAR methods). If the initial value of a model species is zero and its expression was induced by upstream signals, e.g., *c-myc* mRNA or *dusp* mRNA, maximal transcription rate was estimated from its own mRNA level. Unless otherwise stated, transcriptome data were used as the mRNA level to estimate the translated protein level. In this way, we computationally predicted the protein levels from their corresponding mRNA levels in TCGA. We confirmed that our simulated protein levels were reasonably consistent with the experimentally measured protein levels of four breast cancer subtypes in the Library of Integrated Network-based Cellular Signatures (LINCS) database (Niepel et al., 2013) (Figure S2).

Stratification of patients with TNBC based on signaling dynamics

After determining the model parameters, we performed numerical simulations to predict how each patient would respond to EGF and HRG stimulation *in silico*. We performed simulations for 377 patients, and extracted quantitative information, such as amplitude, duration, drop rate, and the cumulative response (see STAR methods for their definition), from the *in silico* dynamics of Akt, ERK, and c-Myc activation in each patient. This was followed by clustering of each patient based on dynamic features (Figure S3). Among these characteristics, we used amplitude, i.e., the maximum activation level for the classification of patients with breast cancer. Even though this dynamic feature cannot distinguish Luminal A and Luminal B subtypes, it could distinguish TNBC from other subtypes (Figure 3A). Notably, our network-based classifier divided the patients with TNBC into two clusters: cluster one and two for patients with poor and better prognoses, respectively (Figures 3B and 3C). A classical PAM50 classification method (Jiang et al., 2016; Koboldt et al., 2012; Nielsen et al., 2010), which is based on the expression signatures of 50 genes, was suitable for subtype classification but not for the prediction of TNBC prognosis (Figures S4A and S4B).

To identify crucial genes for distinguishing clusters 1 and 2, we checked 253 differentially expressed transcripts (see STAR methods for the criteria of gene selection). However, there was no clear trend between these two clusters (Figure S5). From this result, we concluded that dynamical modeling is more suitable for the stratification of patients with TNBC rather than the standard gene expression profiles, and *in silico* signaling dynamics of ErbB signaling network can be utilized as a prognostic marker for TNBC.

Table 1. The criteria used for pre-processing TCGA-BRCA/CCLE-BREAST and TCGA-COAD/CCLE-BREAST samples

TCGA/CCLE	TCGA: Upper age limit	TCGA: Stages of cancer	Total read counts
BRCA/BREAST	59	Stage I, Stage IA, Stage IB, Stage II, Stage IIA, Stage IIB	Lower: 40,000,000 Upper: 140,000,000
COAD/BREAST	79	Stage I, Stage IA, Stage IB, Stage II, Stage IIA, Stage IIB	Lower: 10,000,000 Upper: 160,000,000

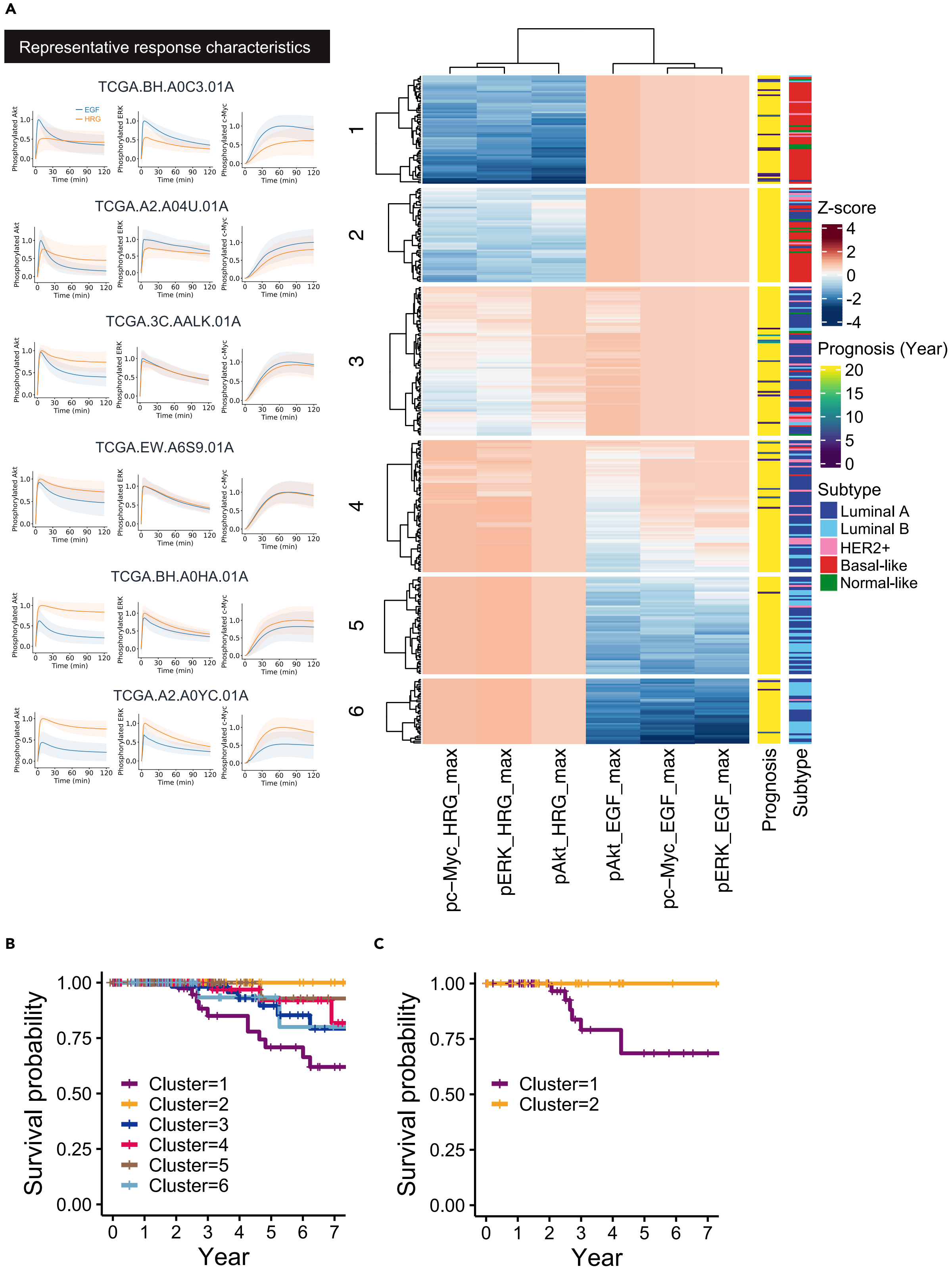

A
Representative response characteristics
TCGA.BH.A0C3.01A
TCGA.A2.A04U.01A
TCGA.3C.AALK.01A
TCGA.EW.A6S9.01A
TCGA.BH.A0HA.01A
TCGA.A2.A0YC.01A
EGF
HRG
Phosphorylated Akt
Phosphorylated ERK
Phosphorylated c-Myc
Time (min)
0 30 60 90 120
0.0 0.5 1.0
1
2
3
4
5
6
pc−Myc_HRG_max
pERK_HRG_max
pAkt_HRG_max
pAkt_EGF_max
pc−Myc_EGF_max
pERK_EGF_max
Prognosis
Subtype
Z-score
4
2
0
-2
-4
Prognosis (Year)
20
15
10
5
0
Subtype
Luminal A
Luminal B
HER2+
Basal-like
Normal-like
B
Survival probability
1.00
0.75
0.50
0.25
0.00
Cluster=1
Cluster=2
Cluster=3
Cluster=4
Cluster=5
Cluster=6
0 1 2 3 4 5 6 7
Year
C
Survival probability
1.00
0.75
0.50
0.25
0.00
Cluster=1
Cluster=2
0 1 2 3 4 5 6 7
Year

Figure 3. Stratification of patients with triple-negative breast cancer (TNBC) based on ErbB signaling dynamics

(A) The patients are classified based on personalized simulations. The prognostic score for patients who deceased within n-1 to n years are donated by n, and patients who were alive after 20 years are denoted in yellow. The representative signal response characteristics were extracted from the topmost portion of each cluster. The blue and orange solid lines denote simulations with EGF and HRG stimulation, respectively. Shaded areas denote SD.

(B and C) Kaplan-Meier survival curves of all patients for all clusters (B) and of patients with the basal-like subtype for clusters 1 and 2 (C).

Furthermore, we performed undersampling and clustering analysis of the patients to determine the least number of samples required for stratification of TNBC. Initially, the number of patients was set to 30, with an increment of 20 in each step. We found that as the number exceeded 110, the patients with TNBC were classified into two clusters with differences in their prognosis being statistically significant ($p < 0.05$) (Figure S6). This shows the efficacy of this analytical method to classify patients with TNBC using relatively small number of samples.

Identification of the mechanisms affecting patient prognosis

Our clustering results showed that patients with TNBC displaying poor prognoses were associated with lower Akt, ERK, and c-Myc activities under HRG stimulation. Together with our model structure, we hypothesized that the signaling activity of HRG receptors (ErbB3, ErbB4, and their heterodimerization partner ErbB2) in this patient group could not efficiently transmit the downstream signal due to competitive interference from higher levels of EGFR. Consistent with this hypothesis, further analysis showed that the expression ratios of EGFR to the ErbB2, 3, and 4 receptors were higher in the poor prognosis group (Figure 4A). This result was also supported by the protein abundances predicted from our models (Figure S7). To further investigate the mechanisms that distinguish prognosis, we randomly sampled patients from each group and performed a sensitivity analysis, which examined how perturbations to the initial conditions (inferred from gene expression level) of the model species affected the c-Myc activity (model output). The result indicated that higher EGFR expression is associated with lower sensitivity to the EGFR inhibitors (Figure 4B). To test this hypothesis, we used CCLE drug response data (Barretina et al., 2012) for validation analysis. First, similar to patient-specific models, cell-line-specific models were constructed using their gene expression values and classified based on their dynamic features (Figure S8). Available breast cancer cell line data (n = 2 for both clusters 1 and 2) were not enough to satisfy the statistical tests. However, breast cancer cell lines in cluster 1 (relative EGFR expression level: high) seemed less sensitive to EGFR inhibitors than cluster 2 (relative EGFR expression level: low). To further validate this, we collected all types of cancer cell lines (n = 229) from CCLE, classified them in terms of ErbB receptor expression ratio, and analyzed drug sensitivity (Figure 4C). Drug efficacy and potency were quantified by the ''activity area,'' which was the area over the dose-response curve (Barretina et al., 2012). We found that cell lines with higher EGFR expression levels showed significantly lower sensitivity to EGFR inhibitors (erlotinib and lapatinib). There was no statistical significance for other inhibitors, such as MEK inhibitors (selumetinib and PD-0325901) (Figure 4D).

Applying model-based stratification to other types of cancer

We next tried if the same ErbB network model can stratify patients with different types of cancers. We selected colon cancer (Muzny et al., 2012), in which EGFR inhibitors are clinically used (Xie et al., 2020). After individualization of the ErbB network models by adding 189 individual transcriptomic datasets provided in TCGA database (TCGA-COAD), the models successfully classified their prognoses according to the argmax (the time at which the signal intensity reached the maximum) of c-Myc dynamics (Figures 5A–5C). Patients in cluster 4 showed poorer prognosis than other clusters even though their signaling dynamics were similar to those in cluster 3. To predict the mechanistic cause of the difference between clusters 3 and 4, we performed sensitivity analysis on time-integrated c-Myc response. The results of this analysis implied that *in silico* patients in cluster 4 showed lower sensitivity in EGFR (ErbB1) than those in cluster 3 (Figure 5), indicating that patients in this cluster may be more resistant to anti-EGFR treatments. In fact, a recent study found a significant correlation between c-MYC expression and anti-EGFR antibody resistance in metastatic colorectal cancer (Strippoli et al., 2020). Thus, our mathematical analysis potentially provides a mechanistic insight to explain the anti-EGFR therapy response of each patient.

DISCUSSION

Identifying the prognostic factors and potential therapeutic drugs for individual patients is crucial for development of personalized medicine. Recent studies indicate that the temporal dynamics of signaling activities and transcription factors are critical for cell fate determination (Johnson and Toettcher, 2019;

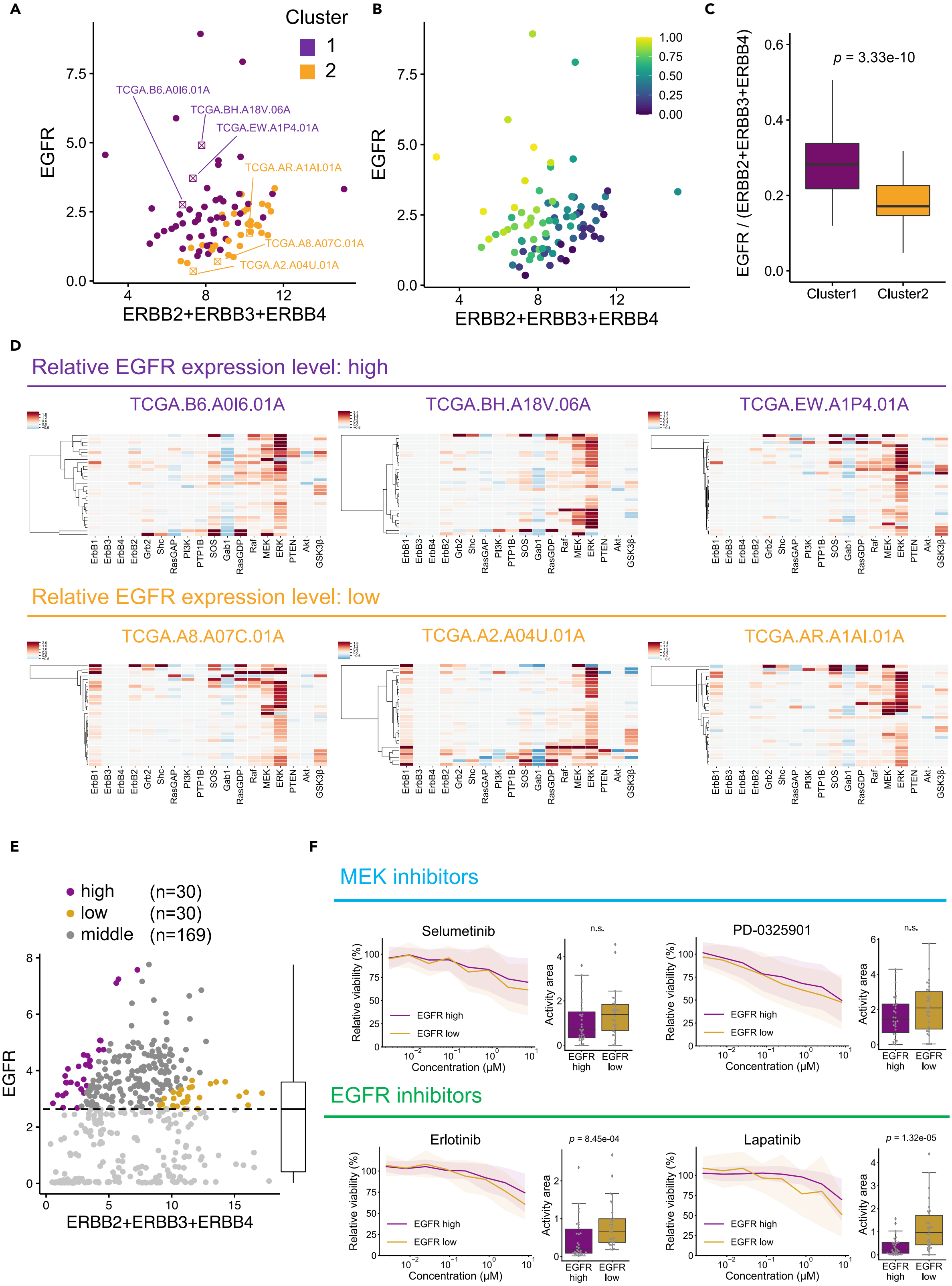

A
Cluster
1
2
TCGA.B6.A0I6.01A
TCGA.BH.A18V.06A
TCGA.EW.A1P4.01A
TCGA.AR.A1AI.01A
TCGA.A8.A07C.01A
TCGA.A2.A04U.01A
EGFR
ERBB2+ERBB3+ERBB4
B
1.00
0.75
0.50
0.25
0.00
EGFR
ERBB2+ERBB3+ERBB4
C
p = 3.33e-10
EGFR / (ERBB2+ERBB3+ERBB4)
Cluster1
Cluster2
D
Relative EGFR expression level: high
TCGA.B6.A0I6.01A
TCGA.BH.A18V.06A
TCGA.EW.A1P4.01A
ErbB1 ErbB3 ErbB4 ErbB2 Grb2 Shc RasGAP PI3K PTP1B SOS Gab1 RasGDP Raf MEK ERK PTEN Akt GSK3β
Relative EGFR expression level: low
TCGA.A8.A07C.01A
TCGA.A2.A04U.01A
TCGA.AR.A1AI.01A
E
high (n=30)
low (n=30)
middle (n=169)
EGFR
ERBB2+ERBB3+ERBB4
F
MEK inhibitors
Selumetinib
PD-0325901
EGFR high
EGFR low
Relative viability (%)
Concentration (μM)
Activity area
n.s.
EGFR inhibitors
Erlotinib
Lapatinib
p = 8.45e-04
p = 1.32e-05

Figure 4. ErbB receptor expression ratios are critical for determining drug sensitivity

(A and B) Scatterplots of the epidermal growth factor receptor (EGFR) expression ratios to the sum of other ErbB receptor families: ERBB2, ERBB3, and ERBB4. Each dot represents one patient. (A) Purple and orange dots denote individual patients in clusters 1 and 2, respectively. (B) The corresponding response scores: the sum of the maximum level of three observables (pERK, pAkt, and pc-Myc) in response to HRG stimulation. Higher scores indicate a lower maximum level when stimulated with EGF.

(C) Boxplots showing the ErbB receptor expression ratios in patients from clusters 1 and 2. The p value was calculated using the Brunner-Munzel test.

(D) Sensitivity analysis of c-Myc activation on representative *in silico* patients with TNBC with high (upper panel) and low (lower panel) EGFR expression ratios.

(E) The EGFR expression ratios to the sum of other ErbB receptor families in the cell lines provided in the CCLE. Based on the ratio, cell lines are classified into three groups, namely, "high": top 30, "low": bottom 30, and "middle": the other 169 cell lines.

(F) Analysis of response profiles against anticancer drugs (MEK inhibitors: selumetinib, PD-0325901; EGFR inhibitors: erlotinib, lapatinib). The solid lines and shaded areas in dose-response curves denote the average and SD of relative viability of 30 cell lines in each cluster, respectively. The efficacy and potency of a drug are simultaneously quantified based on the "activity area" and the p values were calculated using the Brunner-Munzel test with a significance level of 0.05.

Manning et al., 2019; Purvis et al., 2012; Sasagawa et al., 2005). Fey et al. also indicated that JNK signaling metrics can be used as prognostic factors for neuroblastoma (Fey et al., 2015).

Therefore, we hypothesized that signaling dynamics in individual patients with cancer can be used for classification and prediction of the prognosis and drug responses. To this end, we developed a scalable computational framework, Pasmopy, for patient-specific modeling and classification of cancers based on signaling dynamics. Besides developing these unique classification features, we used this framework to construct mechanistic models from texts describing biochemical reactions instead of formulating mathematical equations. In this study, we developed a model of ErbB receptor signaling network from text as a proof of concept. This method of building models will facilitate future studies investigating underlying mechanisms of various biological processes.

Notably, we found that the selected gene panels used in our current ErbB model (Table S1) were not capable of classifying TNBC prognosis (Figures S9A and S9B), and they were not sufficient to identify the molecular mechanism (e.g., EGFR/ErbBs ratio) or drug response. EGFR overexpression has been reported in up to 78% of patients with TNBC (Park et al., 2014). We randomly sampled "*in silico* patients" and performed sensitivity analyses, which surprisingly suggested that patients with higher EGFR expression ratios were less sensitive to EGFR inhibition. To test this model-based prediction, we used publicly available cell-line data and confirmed that cancer cell lines with higher EGFR expression ratios were less sensitive to anticancer drugs targeting EGFR, such as erlotinib and lapatinib. Thus, this framework not only allows us to classify patients but also provides potential mechanistic insight into the regulation of signaling pathways and drug resistance.

Another advantage of our method is that it enables the computational analysis with a small number of data inputs. In this study, the model parameter was trained against experimental data consisting of four cell lines, three observables, two growth factors, and eight time-points. The number of our training datasets was much smaller than the one used in the earlier work (Fröhlich et al., 2018), in which datasets from 120 cell lines treated with seven different drugs and up to nine concentrations of each were used to predict anticancer drug response in different cell lines. We confirmed that the model-predicted sensitive reactions in the ErbB network are highly conserved across 30 independent parameter sets. This indicates that parameter identifiability obtained from our modeling approach does not significantly affect the uncertainty of the model output.

The models were personalized for each patient by incorporating individual gene expression data. Although similar approaches have been used in previous studies for JNK signaling (Fey et al., 2015) and HGF/Met signaling pathway to stratify patients with neuroblastoma (Jafarnejad et al., 2019), they needed to rescale the protein levels based on the fold changes in their mRNA levels in the tumor and healthy tissue. However, this scaling method narrows the potential use of mRNA information. We extended this method and used transcriptome data to infer the maximal translation rate of the corresponding proteins. This method allows us to develop larger models from genome-wide transcriptome datasets.

Finally, we hypothesized that the critical molecular mechanism governing cancer prognosis might be shared, at least in part, by different types of cancers. This would explain why the same model (i.e., short term ErbB signaling dynamics within 120 min) can be used to classify patient prognosis in both breast

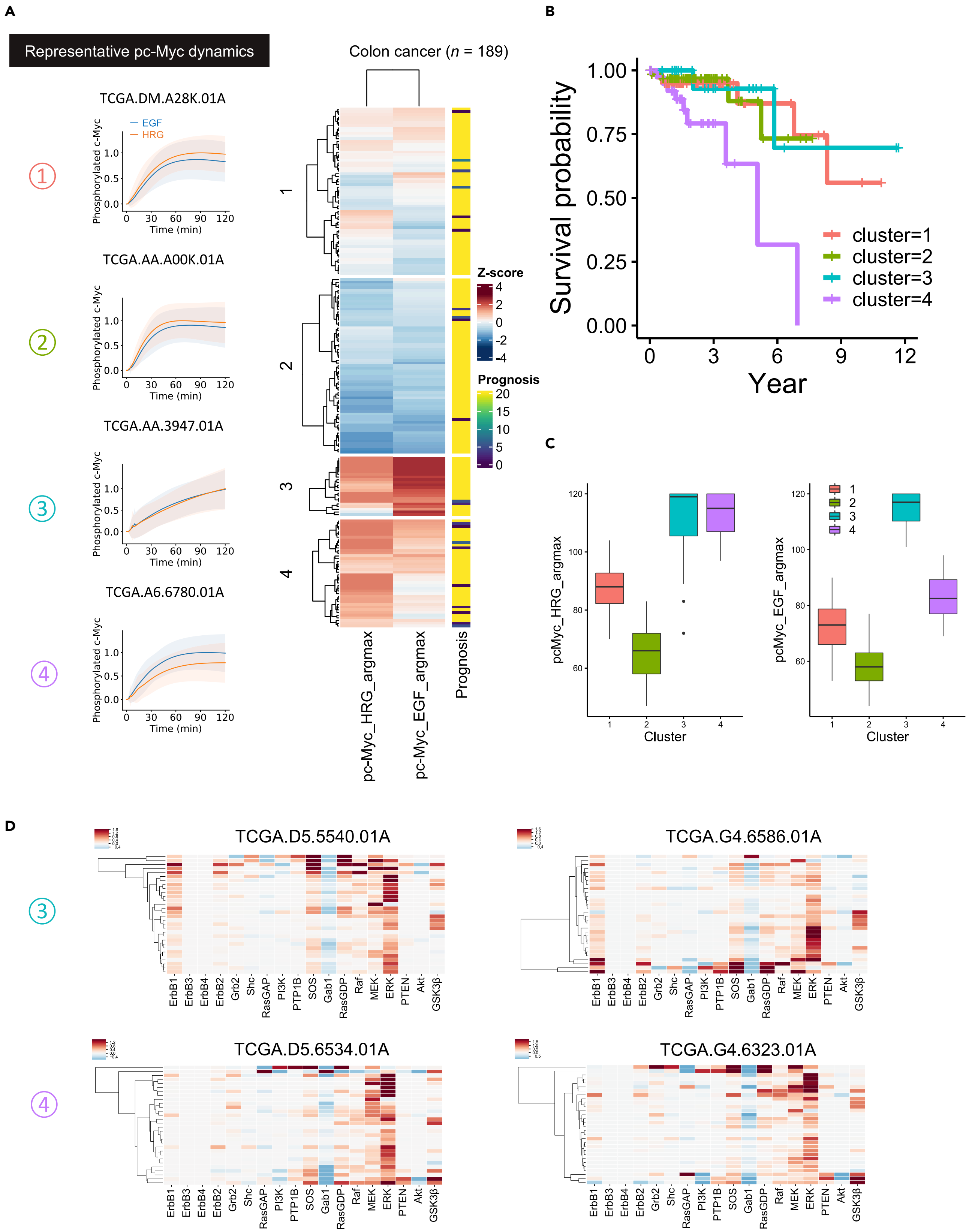

A
Representative pc-Myc dynamics
TCGA.DM.A28K.01A
— EGF
— HRG
Phosphorylated c-Myc
1.0
0.5
0.0
0 30 60 90 120
Time (min)
①
TCGA.AA.A00K.01A
②
TCGA.AA.3947.01A
③
TCGA.A6.6780.01A
④
Colon cancer (n = 189)
1
2
3
4
Z-score
4
2
0
-2
-4
Prognosis
20
15
10
5
0
pc-Myc_HRG_argmax
pc-Myc_EGF_argmax
Prognosis
B
Survival probability
1.00
0.75
0.50
0.25
0.00
0 3 6 9 12
Year
cluster=1
cluster=2
cluster=3
cluster=4
C
pcMyc_HRG_argmax
pcMyc_EGF_argmax
120
100
80
60
Cluster
1 2 3 4
D
TCGA.D5.5540.01A
TCGA.G4.6586.01A
TCGA.D5.6534.01A
TCGA.G4.6323.01A
ErbB1
ErbB3
ErbB4
ErbB2
Grb2
Shc
RasGAP
PI3K
PTP1B
SOS
Gab1
RasGDP
Raf
MEK
ERK
PTEN
Akt
GSK3β

Figure 5. Applying model-based patient stratification to colon cancer

(A) A total of 189 patients were classified based on personalized simulations. The representative pc-Myc dynamics were extracted from the topmost portion of each cluster. The blue and orange solid lines denote simulations with EGF and HRG stimulation, respectively. Shaded areas denote SD. The metric "argmax" denotes the time at which the simulated signal intensity reached the maximum.
(B) Kaplan-Meier survival curves of all patients for all clusters.
(C) Boxplots showing individual argmax values in each cluster.
(D) Sensitivity analysis on the time-integrated response of EGF-induced pc-Myc for randomly sampled *in silico* patients in clusters 3 and 4.

and colon cancers. Therefore, our study suggests that modeling approaches can be used to investigate the specificity and commonality of different types of cancer and evaluation of drug repositioning. In this study, we did not consider somatic mutations in the model. To support this, there was no clear trend on mutation types, at least for two TNBC clusters (Figure S10). However, gene mutations are tightly linked with treatment strategies in some types of cancer such as lung cancer (Collisson et al., 2014). Therefore, parameterization of mutational information to adapt the model to a wide range of cancer types can be done in future.

Limitations of the study

This study focuses on the classification of patients with cancer based on the dynamics of ErbB receptor signaling pathways. However, we cannot exclude the possibility of other signaling pathways being involved in it. Our framework will be able to address these issues by expanding the network to include other receptors and players of cell cycle regulation, apoptotic pathways, or metabolic pathways. In the current study, we used the TCGA and CCLE datasets after normalization of TPM value of the transcripts. Therefore, users need to reconsider data normalization method when other methods such as microarrays or qRT-PCR are used.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODELS AND SUBJECT DETAILS
  - Cell culture
- METHOD DETAILS
  - Model development
  - Training datasets
  - Parameter estimation
  - Individualization of the mechanistic model
  - Data processing in transcriptomic data integration
  - Clustering breast cancer patients with gene expression level
  - Gene mutation analysis
  - Extraction of response characteristics
  - Sensitivity analysis
  - Drug response data analysis
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2022.103944.

ACKNOWLEDGMENTS

We wish to thank Dr. Suxiang Zhang for the western blot experiments of breast cancer cell lines, Dr. Yuichi Shiraishi, Dr. Atsushi Ochiai, and Mr. Kyoichi Ebata for their discussions regarding modeling and TCGA analysis. M.O. was supported by the JSPS KAKENHI Grant No. 17H06299, 17H06302, and 18H04031, the JST-CREST Grant No. JPMJCR21N3, JST-Mirai Program No. JPMJMI19G7, the Moonshot R&D Grant Number JPMJMS2021 and the Uehara Memorial Foundation. H.I. was supported by the Grant-in-Aid for JSPS Fellows Grant No. 20J20192.

AUTHOR CONTRIBUTIONS

H.I. developed the conceptual idea of the computational framework for individualization of mechanistic models, implemented Pasmopy, constructed the mathematical model, performed parameter estimation, and analyzed drug response data. S.Y. analyzed CCLE and TCGA datasets, developed the pipeline for transcriptomic data integration in Pasmopy, and performed the simulation and classification of patients. M.O. coordinated and supervised the study. H.I., S.Y., and M.O. interpreted the results and wrote the manuscript.

DECLARATION OF INTERESTS

Japanese Patent Application (No. 2021–128753) related to this work was filed (H.I., S.Y., and M.O.).

Received: August 22, 2021
Revised: January 3, 2022
Accepted: February 14, 2022
Published: March 18, 2022

REFERENCES

Arteaga, C.L., and Engelman, J.A. (2014). ERBB receptors: from oncogene discovery to basic science to mechanism-based cancer therapeutics. Cancer Cell *25*, 282–303. https://doi.org/10.1016/j.ccr.2014.02.025.

Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A.A., Kim, S., Wilson, C.J., Lehár, J., Kryukov, G.V., Sonkin, D., et al. (2012). The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature *483*, 603–607. https://doi.org/10.1038/nature11003.

Birtwistle, M.R., Hatakeyama, M., Yumoto, N., Ogunnaike, B.A., Hoek, J.B., and Kholodenko, B.N. (2007). Ligand-dependent responses of the ErbB signaling network: experimental and modeling analyses. Mol. Syst. Biol. *3*, 144. https://doi.org/10.1038/msb4100188.

Cibulskis, K., Lawrence, M.S., Carter, S.L., Sivachenko, A., Jaffe, D., Sougnez, C., Gabriel, S., Meyerson, M., Lander, E.S., and Getz, G. (2013). Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat. Biotechnol. *31*, 213–219. https://doi.org/10.1038/nbt.2514.

Clarke, M.A., and Fisher, J. (2020). Executable cancer models: successes and challenges. Nat. Rev. Cancer *20*, 343–354. https://doi.org/10.1038/s41568-020-0258-x.

Colaprico, A., Silva, T.C., Olsen, C., Garofano, L., Cava, C., Garolini, D., Sabedot, T.S., Malta, T.M., Pagnotta, S.M., Castiglioni, I., et al. (2016). TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. Nucleic Acids Res. *44*, e71. https://doi.org/10.1093/nar/gkv1507.

Collisson, E.A., Campbell, J.D., Brooks, A.N., Berger, A.H., Lee, W., Chmielecki, J., Beer, D.G., Cope, L., Creighton, C.J., Danilova, L., et al. (2014). Comprehensive molecular profiling of lung adenocarcinoma: the cancer genome atlas research network. Nature *511*, 543–550. https://doi.org/10.1038/nature13385.

Dagogo-Jack, I., and Shaw, A.T. (2018). Tumour heterogeneity and resistance to cancer therapies. Nat. Rev. Clin. Oncol. *15*, 81–94. https://doi.org/10.1038/nrclinonc.2017.166.

Degasperi, A., Birtwistle, M.R., Volinsky, N., Rauch, J., Kolch, W., and Kholodenko, B.N. (2014). Evaluating strategies to normalise biological replicates of western blot data. PLoS One *9*, e87293. https://doi.org/10.1371/journal.pone.0087293.

Durinck, S., Spellman, P.T., Birney, E., and Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Nat. Protoc. *4*, 1184–1191. https://doi.org/10.1038/nprot.2009.97.

Fey, D., Halasz, M., Dreidax, D., Kennedy, S.P., Hastings, J.F., Rauch, N., Munoz, A.G., Pilkington, R., Fischer, M., Westermann, F., et al. (2015). Signaling pathway models as biomarkers: patient-specific simulations of JNK activity predict the survival of neuroblastoma patients. Sci. Signal. *8*, 1–16. https://doi.org/10.1126/scisignal.aab0990.

Fröhlich, F., Kessler, T., Weindl, D., Shadrin, A., Schmiester, L., Hache, H., Muradyan, A., Schütte, M., Lim, J.H., Heinig, M., et al. (2018). Efficient parameter estimation enables the prediction of drug response using a mechanistic pan-cancer pathway model. Cell Syst *7*, 567–579.e6. https://doi.org/10.1016/j.cels.2018.10.013.

Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B.W.J.H., Jansen, R., De Geus, E.J.C., Boomsma, D.I., Wright, F.A., et al. (2016). Integrative approaches for large-scale transcriptome-wide association studies. Nat. Genet. *48*, 245–252. https://doi.org/10.1038/ng.3506.

Hass, H., Masson, K., Wohlgemuth, S., Paragas, V., Allen, J.E., Sevecka, M., Pace, E., Timmer, J., Stelling, J., MacBeath, G., et al. (2017). Predicting ligand-dependent tumors from multi-dimensional signaling features. NPJ Syst. Biol. Appl. *3*, 27. https://doi.org/10.1038/s41540-017-0030-3.

Imoto, H., Zhang, S., and Okada, M. (2020). A computational framework for prediction and analysis of cancer signaling dynamics from RNA sequencing data—application to the ErbB receptor signaling pathway. Cancers (Basel). *12*, 2878. https://doi.org/10.3390/cancers12102878.

Inoue, T., Terada, N., Kobayashi, T., and Ogawa, O. (2017). Patient-derived xenografts as in vivo models for research in urological malignancies. Nat. Rev. *14*, 267–283. https://doi.org/10.1038/nrurol.2017.19.

Jafarnejad, M., Sové, R.J., Danilova, L., Mirando, A.C., Zhang, Y., Yarchoan, M., Tran, P.T., Pandey, N.B., Fertig, E.J., and Popel, A.S. (2019). Mechanistically detailed systems biology modeling of the HGF/Met pathway in hepatocellular carcinoma. NPJ Syst. Biol. Appl. *5*, 29. https://doi.org/10.1038/s41540-019-0107-2.

Jiang, G., Zhang, S., Yazdanparast, A., Li, M., Pawar, A.V., Liu, Y., Inavolu, S.M., and Cheng, L. (2016). Comprehensive comparison of molecular portraits between cell lines and tumors in breast cancer. BMC Genomics *17*, 525. https://doi.org/10.1186/s12864-016-2911-z.

Johnson, H.E., and Toettcher, J.E. (2019). Signaling dynamics control cell fate in the early Drosophila embryo. Dev. Cell *48*, 361–370.e3. https://doi.org/10.1016/j.devcel.2019.01.009.

Kholodenko, B.N. (2006). Cell-signalling dynamics in time and space. Nat. Rev. Mol. Cell Biol. *7*, 165–176. https://doi.org/10.1038/nrm1838.

Kiyatkin, A., and Aksamitiene, E. (2009). Multistrip western blotting to increase quantitative data output. Methods Mol. Biol. *536*, 149–161. https://doi.org/10.1007/978-1-59745-542-8_17.

Koboldt, D.C., Fulton, R.S., McLellan, M.D., Schmidt, H., Kalicki-Veizer, J., McMichael, J.F., Fulton, L.L., Dooling, D.J., Ding, L., Mardis, E.R., et al. (2012). Comprehensive molecular portraits of human breast tumours. Nature *490*, 61–70. https://doi.org/10.1038/nature11412.

Kourou, K., Exarchos, T.P., Exarchos, K.P., Karamouzis, M.V., and Fotiadis, D.I. (2015). Machine learning applications in cancer prognosis and prediction. Comput. Struct. Biotechnol. J. *13*, 8–17. https://doi.org/10.1016/j.csbj.2014.11.005.

Lee, T., Yao, G., Nevins, J., and You, L. (2008). Sensing and integration of Erk and PI3K signals by Myc. PLoS Comput. Biol. *4*, e1000013. https://doi.org/10.1371/journal.pcbi.1000013.

Manning, C.S., Biga, V., Boyd, J., Kursawe, J., Ymisson, B., Spiller, D.G., Sanderson, C.M., Galla, T., Rattray, M., and Papalopulu, N. (2019). Quantitative single-cell live imaging links HES5 dynamics with cell-state and fate in murine neurogenesis. Nat. Commun. *10*, 2835. https://doi.org/10.1038/s41467-019-10734-8.

Muzny, D.M., Bainbridge, M.N., Chang, K., Dinh, H.H., Drummond, J.A., Fowler, G., Kovar, C.L., Lewis, L.R., Morgan, M.B., Newsham, I.F., et al. (2012). Comprehensive molecular characterization of human colon and rectal cancer. Nature *487*, 330–337. https://doi.org/10.1038/nature11252.

Nica, A.C., and Dermitzakis, E.T. (2013). Expression quantitative trait loci: present and future. Philos. Trans. R. Soc. B Biol. Sci. *368*, 20120362. https://doi.org/10.1098/rstb.2012.0362.

Nielsen, T.O., Parker, J.S., Leung, S., Voduc, D., Ebbert, M., Vickery, T., Davies, S.R., Snider, J., Stijleman, I.J., Reed, J., et al. (2010). A comparison of PAM50 intrinsic subtyping with immunohistochemistry and clinical prognostic factors in tamoxifen-treated estrogen receptor-positive breast cancer. Clin. Cancer Res. *16*, 5222–5232. https://doi.org/10.1158/1078-0432.CCR-10-1282.

Niepel, M., Hafner, M., Pace, E.A., Chung, M., Chai, D.H., Zhou, L., Schoeberl, B., and Sorger, P.K. (2013). Profiles of basal and stimulated receptor signaling networks predict drug response in breast cancer lines. Sci. Signal. *6*, ra84. https://doi.org/10.1126/scisignal.2004379.

Ozaki, K., Ohnishi, Y., Iida, A., Sekine, A., Yamada, R., Tsunoda, T., Sato, H., Sato, H., Hori, M., Nakamura, Y., et al. (2002). Functional SNPs in the lymphotoxin-α gene that are associated with susceptibility to myocardial infarction. Nat. Genet. *32*, 650–654. https://doi.org/10.1038/ng1047.

Park, H.S., Jang, M.H., Kim, E.J., Kim, H.J., Lee, H.J., Kim, Y.J., Kim, J.H., Kang, E., Kim, S.W., Kim, I.A., et al. (2014). High EGFR gene copy number predicts poor outcome in triple-negative breast cancer. Mod. Pathol. *27*, 1212–1222. https://doi.org/10.1038/modpathol.2013.251.

Purvis, J.E., Karhohs, K.W., Mock, C., Batchelor, E., Loewer, A., and Lahav, G. (2012). p53 dynamics control cell fate. Science *336*, 1440–1444. https://doi.org/10.1126/science.1218351.

Purvis, J.E., and Lahav, G. (2013). Encoding and decoding cellular information through signaling dynamics. Cell *152*, 945–956. https://doi.org/10.1016/j.cell.2013.02.005.

Robinson, M.D., and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. *11*, R25. https://doi.org/10.1186/gb-2010-11-3-r25.

Saez-Rodriguez, J., and Blüthgen, N. (2020). Personalized signaling models for personalized treatments. Mol. Syst. Biol. *16*, e9042. https://doi.org/10.15252/msb.20199042.

Sasagawa, S., Ozaki, Y.I., Fujita, K., and Kuroda, S. (2005). Prediction and validation of the distinct dynamics of transient and sustained ERK activation. Nat. Cell Biol. *7*, 365–373. https://doi.org/10.1038/ncb1233.

Schoeberl, B., Pace, E.A., Fitzgerald, J.B., Harms, B.D., Xu, L., Nie, L., Linggi, B., Kalra, A., Paragas, V., Bukhalid, R., et al. (2009). Therapeutically targeting ErbB3: a key node in ligand-induced activation of the ErbB receptor-PI3K axis. Sci. Signal. *2*, ra31. https://doi.org/10.1126/scisignal.2000352.

Storn, R., and Price, K. (1997). Differential evolution - a simple and efficient heuristic for global optimization over continuous spaces. J. Glob. Optim. *11*, 341–359. https://doi.org/10.1023/A:1008202821328.

Strippoli, A., Cocomazzi, A., Basso, M., Cenci, T., Ricci, R., Pierconti, F., Cassano, A., Fiorentino, V., Barone, C., Bria, E., et al. (2020). C-myc expression is a possible keystone in the colorectal cancer resistance to egfr inhibitors. Cancers (Basel) *12*, 638. https://doi.org/10.3390/cancers12030638.

Van't Veer, L.J., Dai, H., Van de Vijver, M.J., He, Y.D., Hart, A.A.M., Mao, M., Peterse, H.L., Van Der Kooy, K., Marton, M.J., Witteveen, A.T., et al. (2002). Gene expression profiling predicts clinical outcome of breast cancer. Nature *415*, 530–536. https://doi.org/10.1038/415530a.

Van Der Walt, S., Colbert, S.C., and Varoquaux, G. (2011). The NumPy array: a structure for efficient numerical computation. Comput. Sci. Eng. *13*, 22–30. https://doi.org/10.1109/MCSE.2011.37.

Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., et al. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat. Methods *17*, 261–272. https://doi.org/10.1038/s41592-019-0686-2.

Wagner, G.P., Kin, K., and Lynch, V.J. (2012). Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. Theor. Biosci *131*, 281–285. https://doi.org/10.1007/s12064-012-0162-3.

Waskom, M. (2021). seaborn: statistical data visualization. J. Open Source Softw. *6*, 3021. https://doi.org/10.21105/joss.03021.

Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R.M., Ozenberger, B.A., Ellrott, K., Sander, C., Stuart, J.M., Chang, K., Creighton, C.J., et al. (2013). The cancer genome atlas pan-cancer analysis project. Nat. Genet. *45*, 1113–1120. https://doi.org/10.1038/ng.2764.

Whittle, J.R., Lewis, M.T., Lindeman, G.J., and Visvader, J.E. (2015). Patient-derived xenograft models of breast cancer and their predictive power. Breast Cancer Res. *17*, 17. https://doi.org/10.1186/s13058-015-0523-1.

Xie, Y.H., Chen, Y.X., and Fang, J.Y. (2020). Comprehensive review of targeted therapy for colorectal cancer. Signal Transduct. Target. Ther. *5*, 22. https://doi.org/10.1038/s41392-020-0116-z.

Xu, J., Chen, Y., and Olopade, O.I. (2010). MYC and breast cancer. Genes Cancer *1*, 629–640. https://doi.org/10.1177/1947601910378691.

Yoshida, G.J. (2020). Applications of patient-derived tumor xenograft models and tumor organoids. J. Hematol. Oncol. *13*, 4. https://doi.org/10.1186/s13045-019-0829-z.

Yu, K., Chen, B., Aran, D., Charalel, J., Yau, C., Wolf, D.M., van 't Veer, L.J., Butte, A.J., Goldstein, T., and Sirota, M. (2019). Comprehensive transcriptomic analysis of cell lines as models of primary tumors across 22 tumor types. Nat. Commun. *10*, 3574. https://doi.org/10.1038/s41467-019-11415-2.

Zhang, Y., Parmigiani, G., and Johnson, W.E. (2020). ComBat-seq: batch effect adjustment for RNA-seq count data. NAR Genom. Bioinform. *2*, lqaa078. https://doi.org/10.1093/nargab/lqaa078.

Zhong, L., Li, Y., Xiong, L., Wang, W., Wu, M., Yuan, T., Yang, W., Tian, C., Miao, Z., Wang, T., et al. (2021). Small molecules in targeted cancer therapy: advances, challenges, and future perspectives. Signal Transduct. Target. Ther. *6*, 201. https://doi.org/10.1038/s41392-021-00572-w.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
Anti-c-Myc (phospho S62)	abcam	Cat#ab51156 RRID:AB_869189
Anti-rabbit IgG, HRP-linked Antibody	Cell Signaling Technology	Cat#7074S RRID:AB_2099233
Chemicals, peptides, and recombinant proteins		
DMEM (High Glucose)	Nacalai Tesque	Cat#08458-16
Fetal Bovine Serum	Sigma-Aldrich	Cat#F7524
Penicillin-streptomycin	Nacalai Tesque	Cat#09367-34
Deposited data		
CCLE drug-response data	Barretina et al., 2012	https://sites.broadinstitute.org/ccle/
CCLE RNAseq gene expression data (read counts)	Barretina et al., 2012	https://sites.broadinstitute.org/ccle/
TCGA-BRCA gene expression data (HTseq-Counts)	NIH GDC data portal	https://portal.gdc.cancer.gov/
TCGA-COAD gene expression data (HTseq-Counts)	NIH GDC data portal	https://portal.gdc.cancer.gov/
TCGA-BRCA Somatic mutation data – MuTect2	NIH GDC data portal	https://portal.gdc.cancer.gov/
LINCS: Basal profile of receptor tyrosine kinase signaling network measured by ELISA	Niepel et al., 2013	https://lincs.hms.harvard.edu/niepel_scisignal_2013/ HMS Dataset #20137
Experimental models: Cell lines		
Human: MCF7	ATCC	Cat#HTB-22; RRID:CVCL_0031
Human: BT-474	ATCC	Cat#HTB-20 RRID:CVCL_0179
Human: SK-BR-3	ATCC	Cat#HTB-30 RRID:CVCL_0033
Human: MDA-MB-231	ATCC	Cat#CRM-HTB-26 RRID:CVCL_0062
Software and algorithms		
Python 3.7.2	Python Software Foundation	https://www.python.org
pasmopy v0.1.0	This paper	https://github.com/pasmopy/pasmopy
biomass v0.5.2	Imoto et al., 2020	https://github.com/biomass-dev/biomass
numpy v1.19.2	Van Der Walt et al., 2011	https://numpy.org
scipy v1.6.2	Virtanen et al., 2020	https://scipy.org
pandas v1.2.4	pandas – Python Data Analysis Library	https://pandas.pydata.org
seaborn v0.11.2	Waskom, 2021	https://seaborn.pydata.org
Julia 1.6.2	The Julia Programming Language	https://julialang.org
BioMASS.jl v0.5.0	Imoto et al., 2020	https://github.com/biomass-dev/BioMASS.jl
R 4.0.2	The R Foundation	https://www.r-project.org
TCGAbiolinks v2.18.0	Colaprico et al., 2016	https://bioconductor.org/packages/TCGAbiolinks/
sva v3.38.0	Zhang et al., 2020	https://bioconductor.org/packages/sva/
biomaRt v2.46.3	Durinck et al., 2009	https://bioconductor.org/packages/biomaRt/

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Mariko Okada (mokada@protein.osaka-u.ac.jp).

Materials availability

This study did not generate new unique regents.

Data and code availability

- All code for model development, parameter estimation, simulations, and analyses are available at https://github.com/pasmopy/breast_cancer. The Pasmopy core library can be found at https://github.com/pasmopy/pasmopy.
- Pasmopy requires Python 3.7 or newer versions to run and can be installed from its source by downloading the code directly from the above GitHub link, or can be installed using the pip package install manager with the following command:
- $ pip install pasmopy

EXPERIMENTAL MODELS AND SUBJECT DETAILS

Cell culture

MCF-7, BT-474, SK-BR-3, and MDA-MB-231 cells were maintained in Dulbecco's modified Eagle's medium (DMEM) supplemented with 10 % fetal bovine serum (FBS).

METHOD DETAILS

Model development

The model used in this study is based on a systems of ODEs. Most of rate equations in the model were derived by means of law of mass action. We applied the Michaelis-Menten equation and the Hill equation for reactions describing (de)phosphorylation and transcription, respectively. The upstream signaling network model included ErbB receptor activation, Ras-ERK cascade, and the Akt-PI3K pathway, which was adapted from the model of Birtwistle et al. (Birtwistle et al., 2007) and integrated the process of c-Myc regulation, which was newly constructed for this study. The resulting model has 319 rate equations, 228 species, and 648 parameters. In the original model, it was assumed that the route of membrane recruitment does not affect the function of proteins and a "membrane-localized state" was introduced to reduce the complexity of membrane recruitment. We did not use this expression but imposed parameter value constraints. In the rate equation, we assumed that the kinetic parameter describing the binding of downstream proteins to an adaptor protein, e.g., Grb2 and Gab1, is identical regardless of how the adaptor protein is recruited to the membrane. To study how the upstream ERK and Akt activity control c-Myc induction and activation, we added the process of c-Myc regulation so that ERK and Akt could activate and stabilize c-Myc, respectively (Lee et al., 2008). The text file used in building this model is available at https://github.com/pasmopy/breast_cancer/blob/master/models/erbb_network.txt.

Training datasets

We used time-series data on phosphorylated Akt, ERK, and c-Myc stimulated with EGF and HRG (eight time-points, up to 120 min) obtained from four breast cancer cell lines (MCF-7, BT-474, SK-BR-3, and MDA-MB-231). The datasets on phosphorylated Akt and ERK were obtained from a previous study (Imoto et al., 2020). The time-course data of phosphorylated c-Myc were obtained in the current study (Figure S11). Before treatment with 10 nM EGF or HRG, the cells were synchronized by serum starvation for 16 h. The cells were lysed with BioPlex Lysis buffer (Bio-Rad Laboratories, Hercules, CA, USA), cell lysates were cleared by centrifugation (13,000 rpm, 15 min, 4 °C), and the total protein concentration in the supernatants was determined using a protein assay reagent (Bio-Rad Laboratories, Hercules, CA, USA). For Western blotting, anti-phospho-c-Myc (S62, ab51156) was purchased from Abcam (Cambridge, MA, USA). We adopted the transfer and normalization methods described in previous studies (Degasperi et al., 2014; Kiyatkin and Aksamitiene, 2009) to minimize transfer errors and variability between the blots.

Parameter estimation

Of the 648 parameters such as kinetic constants and weighting factors, 220 were trained against the time series phospho-protein data and cell-line-specific transcriptomic data obtained from the CCLE (Barretina et al., 2012). Gene expression level information was used to individualize the maximal transcription rate and nonzero initial conditions in each cell line. We used BioMASS.jl (ver. 0.5.0), which provides a Julia (ver. 1.6.0) interface to the BioMASS parameter estimation. The parameters were trained using a global parameter estimation method called Differential Evolution (DE) (Storn and Price, 1997) that minimizes the residual sum of squares between experimental measurements and simulations. The optimization was stopped after the objective function value dropped below 6.0. Using BioMASS.jl, the results for all 30 independent parameter estimation runs were saved in the .dat format. For the original biomass framework in Python to recognize and read the optimized parameters, these results were converted into the standard binary file format in numpy (Van Der Walt et al., 2011).

Individualization of the mechanistic model

The gene expression profiles for each cell line or patient were incorporated through the following methods:

(i) When the initial amount of a species is zero and induced by upstream signals, e.g., *dusp* mRNA and *c-myc* mRNA:

The rate equation on transcription, *v*, is described by the Hill equation in our model:

$$v = \frac{V \cdot [TF]^n}{K^n + [TF]^n} \quad \text{(Equation 1)}$$

Where *V* is the maximal transcription rate, *K* is the concentration of transcription factor ([TF]) producing 50% maximal response, and *n* is the Hill coefficient. In this case, the transcriptomic data was incorporated as the maximal transcription rate, *V*, using the following equation:

$$V = \sum_i \alpha_i \cdot x_i \quad \text{(Equation 2)}$$

Where x and α are the transcripts per million (TPM) value (relative log expression (RLE) normalized and post-ComBat) and the corresponding weighting factor for a gene to estimate transcription rate, respectively.

(ii) When the initial condition is not zero:

The transcriptomic data is used to estimate the initial value of protein, y_0 via:

$$y_0 = \sum_i \beta_i \cdot x_i \quad \text{(Equation 3)}$$

Where β is the weighting factor to estimate the initial amounts of model protein species.

In both cases, the weighting factors were estimated during model parameterization.

Data processing in transcriptomic data integration

The criteria for sample selection in the stratification of TCGA-BRCA and TCGA-COAD, i.e., upper age limits, stages of cancer, and the range of patients' total read counts, are described in Table 1.

TCGA-BRCA and TCGA-COAD RNA-seq samples were downloaded using R TCGAbiolinks (Colaprico et al., 2016) version 2.18.0. The gene expression tables of HTSeq-based count retrieval were handled by TCGAbiolinks: GDCquery command. This repository contained 1222 RNA-seq samples that have been uniformly processed from row data. First, samples donated from patients over an upper age limit, patients entering late stages of cancer, noncancerous solid tissue, and duplicated samples were filtered from the data table.

The CCLE RNA-seq count matrix was downloaded from https://data.broadinstitute.org/ccle/CCLE_RNAseq_genes_counts_20180929.gct.gz, and samples other than those obtained from breast cancer cell-lines were excluded from the data table. After merging these two matrices by ensemble gene ID, we executed the R ComBat-seq program (Zhang et al., 2020) to adjust batch effects between the two datasets (Yu et al., 2019). Because the total read counts among samples varied widely, samples with total reads less than the lower bound (TCGA-BRCA/CCLE-BREAST: 40,000,000, TCGA-COAD/CCLE-BREAST: 10,000,000) or greater than the upper bound (TCGA-BRCA/CCLE-BREAST: 140,000,000, TCGA-COAD/CCLE-BREAST: 160,000,000) among the adjusted datasets (Figure S1B) were excluded. After normalization of the library size using the RLE method (Robinson and Oshlack, 2010), the count matrix was normalized using the TPM method. The gene lengths used in the TPM calculation (Wagner et al., 2012) were the differences between the start and end positions of each gene on the chromosomes retrieved from the R biomaRt package (Durinck et al., 2009) version 2.46.3.

To predict the ErbB signaling dynamics of the TCGA-COAD samples, we reused the parameter sets optimized in the breast cancer model (TCGA-BRCA). Since the normalized TPM values of 38 genes in four cell lines (MCF-7, BT-474, SK-BR-3, and MDA-MB-231) in the TCGA-BRCA: CCLE-BREAST matrix, *X*, are different from those in the TCGA-COAD: CCLE-BREAST matrix, *Y*, the values were transformed by multiplying a scaling factor, F_i, for ith gene to reproduce the experimental observations in the four breast cancer cell lines. F_i is calculated using the following equation:

$$F_i = \frac{\max_i X_{i,\ \{\mathrm{MCF7,\ BT474,\ SKBR3,\ MADMB231}\}}}{\max_i Y_{i,\ \{\mathrm{MCF7,\ BT474,\ SKBR3,\ MADMB231}\}}} \quad \text{(Equation 4)}$$

Clustering breast cancer patients with gene expression level

After removing genes with zero expression levels in more than half of the samples from the log2-transformed matrix, the *p*-values between the cluster 1 and cluster 2 samples were calculated using Student's t-test. The *q*-values (FDR threshold: 0.05) were estimated from the *p*-values with the R-value package (ver 2.22.0), and genes with a value less than 0.05 were designated as differentially expressed genes (DEGs). For clustering based on gene expression, the expression levels of all genes were log2-transformed and then clustered by the expression level of a given gene (genes in the model, PAM50 genes, DEGs) using the k-medoids method.

Gene mutation analysis

Mutation Annotation Format (MAF) file used to store somatic mutations per sample were summarized, analyzed, and visualized using the maftools Bioconductor package. The `GDCquery_Maf()` command of TCGAbiolinks was used to obtain the MAF file of TCGA-BRCA detected by the MuTect2 pipeline (Cibulskis et al., 2013). After extracting the patient information used for our clustering, we used the oncoplot() command to plot the mutations of the genes used in the model.

Extraction of response characteristics

Pasmopy provides the following response characteristics to classify personalized simulations: "max," "AUC," and "droprate." The maximum and time-integrated responses are calculated using the numpy.max and scipy.integrate.simpson functions, respectively. The rate of decline ("droprate"), r, is defined as follows:

$$r = -\frac{(A_{\mathrm{end}} - A_{\max})}{(T_{\mathrm{end}} - T_{\max})} \quad \text{(Equation 5)}$$

Where T_{end} and $T_{\max}$ are simulation end time (120 min) and time to reach maximum level, respectively. $T_{\max}$ was calculated with the numpy.argmax function. A_{end} and $A_{\max}$ denote the normalized simulated values at 120 min and $T_{\max}$, respectively.

We applied the maximum value of simulated dynamics for both EGF and HRG stimulation to classify samples. After calculating the standard score (*z*-score) of each feature, the patients were classified based on the Euclidean distance of each patient using the *k*-medoids method. The prognostic score of each patient was given as n, where n is the score of patients who deceased within n-1 to n years, and 20 is the score of patients who were alive.

Sensitivity analysis

The sensitivity coefficients Sy were calculated using the following equation:

$$S_y = \partial \ln M \Big/ \partial \ln y_j \qquad \text{(Equation 6)}$$

Where M is the signaling metric, i.e., the maximum level of phosphorylated c-Myc with EGF stimulation, and y_j is each nonzero species in the mechanistic model (Schoeberl et al., 2009). The sensitivity coefficients were calculated using finite difference approximations with 1 % changes in the initial conditions. To calculate sensitivity coefficients, we used the PatientModelAnalyses class in Pasmopy version 0.1.0 with the biomass_kws={''metric'': ''maximum'', ''style'': ''heatmap''} options.

In this study, the maximum value of EGF-induced c-Myc activation during the observation time from 0 min to 120 min was collected for the following analysis.

Drug response data analysis

The drug response and gene expression data from the CCLE were downloaded from https://data.broadinstitute.org/ccle_legacy_data/pharmacological_profiling/CCLE_NP24.2009_Drug_data_2015.02.24.csv and https://data.broadinstitute.org/ccle/CCLE_RNAseq_genes_counts_20180929.gct.gz, respectively. First, RNA-seq count datasets from the CCLE were normalized using the TPM method after the library size of all samples was converted with the RLE method. To study the effect of EGFR inhibitors on cancer cells, we used cell lines with high EGFR expression ratios relative to other ErbB receptor families (ErbB2, ErbB3, and ErbB4) by setting the minimum EGFR expression level to the median of all cell lines. Next, we divided the cell lines into three groups: (i) high (top 30 EGFR expression ratios), (ii) low (bottom 30), and (iii) middle (the other 169 cell lines). The efficacy and potency of a drug were simultaneously quantified by the ''activity area'' (Barretina et al., 2012), whose values were extracted from the column: ''ActArea'' in the drug response data. We used pandas v1.2.4 for loading data, scipy (Virtanen et al., 2020) v1.6.2 for the statistical test, and seaborn (Waskom, 2021) v0.11.2 for data visualization.

QUANTIFICATION AND STATISTICAL ANALYSIS

The *p*-values were calculated using the Brunner-Munzel test with a significance level of 0.05 using scipy.stats.brunnermunzel() and brunnermunzel() in Python and R, respectively. Details of the statistical methods can be found in the figure legends. The p-values in the survival curve were calculated using the log-rank test with a significance level of 0.05 using survival package version.3.2.13 in R.