Computational Approach for Deriving Cancer Progression Roadmaps from Static Sample Data

Yijun Sun[1,2,3,5,*], Jin Yao[1], Le Yang[2], Runpu Chen[2],
Norma J. Nowak[4], Steve Goodison[6,*]

[1]Department of Microbiology and Immunology
[2]Department of Computer Science and Engineering
[3]Department of Biostatistics, [4]Department of Biochemistry
The State University of New York, Buffalo, NY 14203
[5]Department of Bioinformatics and Biostatistics
Roswell Park Cancer Institute, Buffalo, NY 14201
[6]Department of Health Sciences Research
Mayo Clinic, Jacksonville, FL 32224

Abstract

As with any biological process, cancer development is inherently dynamic. While major efforts continue to catalog the genomic events associated with human cancer, it remains difficult to interpret and extrapolate the accumulating data to provide insights into the dynamic aspects of the disease. Here, we present a computational strategy that enables the construction of a cancer progression model using static tumor sample data. The developed approach overcame many technical limitations of existing methods. Application of the approach to breast cancer data revealed a linear, branching model with two distinct trajectories for malignant progression. The validity of the constructed model was demonstrated in 27 independent breast cancer datasets, and through visualization of the data in the context of disease progression we were able to identify a number of potentially key molecular events in the advance of breast cancer to malignancy.

Nucleic Acids Research
Submitted in November 2016, accepted in January 2017

*These authors contributed equally. Please address all correspondence to Dr. Yijun Sun (yijunsun@buffalo.edu) and Dr. Steve Goodison (Goodison.Steven@mayo.edu).

Introduction

Human cancer is a dynamic disease that develops over an extended time period through the accumulation of a series of genetic alterations. Once initiated from normal cells, the advance of the disease to malignancy can be viewed as a Darwinian, multistep evolutionary process at the cellular level, characterized by random genetic variations and natural selection imposed by the microenvironment [1–6]. While the majority of genetic alterations confer no specific growth advantage, tumor cells that acquire changes in genes and pathways that control key cellular processes can overwhelm less vigorous cell populations within a tumor mass, resulting in a series of clonal expansions leading to the invasion of surrounding tissues and metastasis to distant organs (**Figure S1**). Delineating the entire dynamic process, identifying pivotal molecular events that drive stepwise disease progression, and placing identified changes in a cancer development roadmap would significantly advance our understanding of tumor biology and lay a foundation for the development of improved cancer diagnostics, prognostics and targeted therapeutics.

The concept of cancer evolution was posited in the 1970s [1], and numerous studies have since been conducted that significantly expanded our understanding of the concept (see [7] for an excellent review). However, beyond conceptual models [8,9], for most cancers there is currently no established progression model derived from human tumor tissue data that describes the dynamic disease process. Traditionally, system dynamics is approached through time-course studies achieved by repeated sampling of the same cohort of subjects across an entire biological process. However, due to the need for immediate treatment upon diagnosis, it is not feasible to collect time-series data to study human cancer progression, and we have to rely on profile data obtained from excised tissue samples. Constrained by this sampling limitation, previous studies have focused on inferring disease progression through the derivation of phylogenetic trees. These are achieved by comparing DNA mutation or copy number variation (CNV) profiles from a small number of evolutionary-related tumor samples (e.g., those collected either from the same patient before and after surgery or from different regions of the same tumor) [7,10,11]. While phylogenetic analysis experimentally verified cancer evolution theory, constructed models reflect only the evolutionary histories of individual tumors at the time of sampling, and cannot be generalized to other patients since tumors even of the same phenotype can have a completely different mutational and CNV profiles [12]. A related line of research is oncogenetic modeling, which aims to estimate the statistical dependencies among genetic alterations [13, 14]. By assuming that each tumor is an independent realization of the same stochastic evolutionary process, the analysis can be applied to cross-sectional data collected from different patients. However, constructed models represent only a possible occurrence order of a small set of genetic events (usually the most abundant events, e.g., a sequential accumulation of $APC \rightarrow KRAS \rightarrow TP53$ gene mutations in colorectal carcinogenesis [15]). They cannot reveal the dynamic process of disease progression and be used to detect new cancer genes [7]. As with phylogenetic analysis, oncogenetic tree analysis can only be applied to gene mutation and CNV data [14]. While both phylogenetic analysis and oncogenetic modeling are sometimes collectively termed as progression modeling in the cancer literature [7], they were not designed to construct models that describe disease dynamics.

With the rapid development of molecular profiling techniques and the establishment of major international cancer genome consortia [12, 16, 17], an impressive catalog of molecular profile data obtained from excised tumor tissue samples is accumulating. Then, we ask the following question: can

we construct a cancer progression model by using static sample data, instead of using time-course data? Static samples each provide a snapshot of the disease process. If the number of samples is sufficiently large, the genetic footprints of individual samples populate progression trajectories, enabling us to recover the dynamic disease process from static samples using a computational approach. This idea was first proposed in [18, 19], however, early work did not consider the problem of feature selection, and could not extract branching lineages [20]. Several attempts have been made to address the two aforementioned issues [20,21], but constrained by sample number and algorithm limitations, prior work did not demonstrate the feasibility of using static samples to construct cancer progression models.

In this paper, we present a comprehensive computational pipeline for the derivation of cancer progression models and the identification of pivotal driver gene mutations. To demonstrate the utility of the developed pipeline, we applied it to the analysis of 27 independent breast cancer datasets comprised of $> 9,000$ breast tumor and normal tissue samples. Our analysis revealed a linear, branching model with two distinct trajectories for malignant progression. To demonstrate the validity of the developed model, we proposed a comprehensive validation plan and conducted a large-scale study that provided support for the proposed model. To demonstrate the utility of the constructed model, we also developed a new method to identify putative cancer driver genetic mutations within the cancer-progression framework. This study demonstrates the feasibility of using static samples to construct cancer progression models, and provides a technical foundation for the construction of high-resolution cancer progression models by integration of all available molecular and genetic data.

Materials and Methods

Figure 1 provides an overview of the presented stepwise study. It consists of three major components: (1) methodology development and cancer progression model construction, (2) model validation, and (3) detection of cancer driver gene mutations.

Datasets

Molecular profile data from 27 studies was assembled into a database comprised of $8,996$ breast tumor tissues and 285 normal breast tissues (**Table S1**). The progression modeling analysis was primarily performed on the METABRIC [17] and TCGA RNA-seq [12] datasets, which are the two largest single breast cancer datasets collected to date, containing $2,133$ and $1,176$ tumor samples, respectively. The additional 25 datasets contained a various number of tumor samples, and were used mainly for model validation. The 27 datasets include almost all breast tumor and normal tissue profile data assembled over the past 15 years. A mutation data analysis was performed on the TCGA mutation data, which cataloged $54,013$ non-silent mutations in $13,870$ genes in 958 breast tumor samples.

Comprehensive Bioinformatics Pipeline for Cancer Progression Modeling

We developed a comprehensive bioinformatics pipeline, referred to as CancerMapp, for cancer progression modeling using static tumor sample data (**Figure 2**). In line with cancer evolution theory, a cancer progression trajectory can be mathematically described as a tree-like structure with branching lineages hidden in a high-dimensional genomics space, connecting a series of clusters that represent genetically homogeneous groups (**Figure S1**). Accordingly, the developed bioinformatics pipeline consists of four major components. First, we performed feature selection to identify disease related genes. Then, by using the selected genes, we constructed a principal tree to describe the general trend of data, and performed clustering analysis to identify genetically homogenous groups. Finally, by using

the constructed principal tree as a backbone, we combined the principal tree and the detected clusters to construct a cancer progression model and extracted disease progression paths. Several algorithmic innovations were proposed to identify cancer related genes that preserve data intrinsic structures, and to extract a self-intersecting principal curve embedded in a high-dimensional space. The bioinformatics pipeline was extensively tested on both simulation and cancer datasets and compared against existing approaches.

Feature selection for identifying cancer related genes

Since only a small fraction of genes are likely to be involved in the biological processes of cancer development, the first step toward cancer progression modeling is to identify disease related genes. Early work on cancer progression modeling analysis did not consider the problem of gene selection [18, 19]. Several methods have been proposed to address the issue [20, 21], but our numerical analysis showed that existing methods did not perform well (**Sections S4**). In our previous work, we conducted a proof-of-concept study that used static sample data to study cancer dynamics [22]. Feature selection was performed within the framework of molecular classification by using patient survival data. However, survival time is a poor indicator of cancer development, and multiple confounding factors (e.g., treatment regimens, patient compliance and even lifestyles) could significantly impact patient survival. It is difficult, if not impossible, to include unknown confounding factors into a computational model. Furthermore, the goal of molecular classification is to separate patients into good or bad prognostic groups, rather than to maintain data intrinsic structure. Patients with distinct molecular characteristics but with similar clinical outcomes can be grouped together, leading to structural distortion in a constructed progression model.

To overcome these problems, we developed a new feature selection method within the molecular subtyping framework. Formulated as an unsupervised clustering problem, molecular subtyping stratifies cancer patients into subtypes with distinct clinical outcomes [23, 24]. While clustering analysis treats each cluster as an independent event, we attempted to build a model to describe the disease dynamics process. Thus, cancer progression modeling analysis can be considered a natural extension of molecular subtyping. It is reasonable to assume that the sample distribution supported by the selected genes is compliant with existing subtyping systems. A major issue associated with using cancer subtypes as a template to select relevant genes is that for breast cancer there is currently no definitive method for molecular subtyping, and several large-scale benchmark studies showed that existing methods achieved only moderate concordance [25–27]. Moreover, most molecular subtyping studies were performed on gene expression data. We want to develop a computational framework that enables us to leverage the results of existing work and extend the search over other genetic data. From the machine-learning perspective, feature selection for unsupervised learning is generally much more difficult than that for supervised learning due to the lack of labels to guide the selection of relevant features [28, 29]. We proposed a new method that transforms the problem of feature selection for unsupervised learning into that for supervised learning by using subtype labels from existing methods. Due to the use of different gene sets and clustering methods, existing subtyping methods may come up with different assessments on a patient. To address this issue, we associated each sample with a probability label vector that reflects decision uncertainty. As such, we can integrate the results of subtyping methods developed in the past decade into *one* computational framework. Although in this study we used only gene expression data for model construction, it can be easily extended to integrate

other genetic data.

Let $\{(\mathbf{x}_n, \mathbf{y}_n)\}_{n=1}^N$ be a training dataset, where $\mathbf{x}_n \in \Re^J$ is the nth sample with J features, and $\mathbf{y}_n \in \Re^m$ is a label vector recording the probabilities of $\mathbf{x}_n$ belonging to m subtypes (construction of the label vectors is detailed in **Section S2.1.4**). Our goal is to find a gene subset so that the label vectors of unseen samples can be optimally predicted. To this end, we performed feature selection within the HSIC Lasso framework [30]:

$$\min_{\mathbf{w}} \left\| \bar{\mathbf{Y}} - \sum_{j=1}^{J} w_j \bar{\mathbf{K}}_j \right\|_F + \lambda \|\mathbf{w}\|_1, \;\; \text{subject to} \;\; \mathbf{w} \geq 0 \,, \tag{1}$$

where $\| \cdot \|_F$ is the Frobenius norm, $\mathbf{w}$ is a non-negative weight vector where the magnitude of each element indicates the relevance of the corresponding feature, and λ is a regularization parameter that controls the sparseness of a solution. $\bar{\mathbf{K}}_j = \mathbf{L}\mathbf{K}_j\mathbf{L}$ and $\bar{\mathbf{Y}} = \mathbf{L}\mathbf{Y}\mathbf{L}$ are centered Gram matrices, where $K_j(n,m) = k(x_n(j), x_m(j))$, $Y(n,m) = k(\mathbf{y}_n, \mathbf{y}_m)$, $k(x,x')$ is the Gaussian kernel function, $x_n(j)$ is the jth element of $\mathbf{x}_n$, and $\mathbf{L} = \mathbf{I}_N - 1/N \mathbf{1}_N \mathbf{1}_N^T$ is a centering matrix with $\mathbf{I}_N$ being an identity matrix and $\mathbf{1}_N$ being a vector of all ones. Due to the use of the Gaussian kernel function, the nonlinear dependency between individual features and label vectors can be extracted [30].

The above formulation can be interpreted as regressing matrix $\bar{\mathbf{Y}}$ constructed by using subtype probability vectors against gene expression data through a linear combination of feature-wise matrices $\{\bar{\mathbf{K}}_j\}$. By vectorizing $\bar{\mathbf{Y}}$ and $\{\bar{\mathbf{K}}_j\}$ in the same order, it can be reformulated as a non-negative Lasso problem with N^2 samples and J features, and there are several well-known algorithms for solving a Lasso problem [31]. However, for our application, direct optimization is computationally infeasible. For example, if the METABRIC data is used, it amounts to solving a Lasso problem with $\sim 4 \times 10^6$ samples each with $\sim 2.5 \times 10^4$ features, requiring ~ 10 terabytes of memory. This is a typical big data problem. To address the computational issue, we developed a stochastic-learning based method. The basic idea is to update a solution iteratively by using a gradient calculated based on a small set of randomly picked samples, instead of using all samples [32]. Since problem (1) is a constrained convex optimization problem, in order to use gradient descent techniques, we converted the constrained problem into an unconstrained one by setting $w_j = v_j^2$, $1 \leq j \leq J$. Note that the new problem is no longer convex, and gradient descent may find a local minimizer or a saddle point. However, it was proved that it is quasi-convex for $\mathbf{v} \geq 0$, and if gradient descent starts from a nonzero initial point, the solution obtained when the gradient vanishes is a global minimizer [33]. The mathematical derivation is detailed in **Section S2.1.2**.

The proposed method has two parameters, the regularization parameter and the learning rate of stochastic learning. We employed the ten-fold cross validation method to estimate the regularization parameter. It has been proved that the asymptotic convergence rate of stochastic learning is independent of the sample size [34]. Therefore, the learning rate can readily be estimated by using a small subset of data. See **Section S2.1.4** for a detailed discussion.

Constructing a principal tree to describe dynamic disease process

Once cancer related genes are selected, the next step is to build a mathematical model to formally describe the tree structure of the cancer progression process. To this end, principal curve fitting methods were used. Formally, a principal curve is a nonlinear generalization of the first principal component line that passes through the middle of a data cloud [35] (**Figure S4**). In the past two

decades, a dozen methods have been developed for principal curve fitting [35, 36]. However, existing methods are generally limited to learn a curve that is embedded in a low-dimensional space and does not intersect itself [35, 36], which is quite restrictive for our application. We proposed a new graphic model-based method to learn a tree structure from data that addresses some limitations of prior work.

Let $\{\mathbf{x}_1, \cdots, \mathbf{x}_N\}$ be a sample dataset and $\mathbf{x}_n \in \Re^D$ be the nth sample with D features, and assume that the tree structure to be learned lies in a latent space $\mathcal{Z} \subset \Re^d$ with $d \ll D$. We used an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ to represent the structure, where $\mathcal{V} = \{v_1, \cdots, v_N\}$ is a set of vertices and $\mathcal{E}$ is a set of edges connecting the vertices. We introduced a set of latent variables $\{\mathbf{z}_1, \cdots, \mathbf{z}_N\} \subset \mathcal{Z}$ to explicitly represent the graph, and associated $\mathbf{z}_n$ with vertex v_n. Our goal is to learn a mapping function $f_{\mathcal{G}} : \Re^d \to \Re^D$ that projects data in the latent space back on to the input space so that a reconstruction error is minimized. Without explicitly specifying a form for $f_{\mathcal{G}}$, it is generally difficult to learn the structure of a graph. However, for our application, we are only interested in learning a tree structure and projection points $\{f_{\mathcal{G}}(\mathbf{z}_1), \cdots, f_{\mathcal{G}}(\mathbf{z}_N)\}$. In this case, a minimum spanning tree (MST) [37] is a natural choice to describe disease dynamics. For notional simplicity, we denoted $f_{\mathcal{G}}(\mathbf{z}_n)$ as $\boldsymbol{\theta}_n$, and solved the following optimization problem:

$$\min_{\{\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_N\}, \{p_{ij}\}, \{w_{ij}\}} \sum_{i=1}^{N} \sum_{j=1}^{N} p_{ij} \left(\|\mathbf{x}_i - \boldsymbol{\theta}_j\|^2 + \sigma \log p_{ij} \right), \text{subject to} \sum_{(v_i, v_j) \in \mathcal{E}} w_{ij} \|\boldsymbol{\theta}_i - \boldsymbol{\theta}_j\|^2 \leq \ell \,, \qquad (2)$$

where p_{ij} is the probability of assigning sample $\mathbf{x}_i$ to projection point $\boldsymbol{\theta}_j$, $\sigma \geq 0$ is a parameter for soft assignment using the negative entropy regularization [38], and $\{w_{ij}\}$ are constrained to be a feasible solution of a minimum spanning tree where the cost of an edge is computed as the squared Euclidean distance between two projection points. The above formulation can be interpreted as fitting to a given dataset a minimum spanning tree with a length constrained to be less than ℓ (**Figure S6**). It can be proved that problem (2) is a biconvex optimization problem (**Section S2.2.2**), and thus can be efficiently solved by alternate convex search [39]. Briefly, we first fixed $\{w_{ij}\}$ and $\{p_{ij}\}$ and found a solution for $\{\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_N\}$ through convex optimization. Then, we fixed $\{\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_N\}$ and found a solution for $\{w_{ij}\}$ by solving a MST problem using Kruskal's method [40] and solved $\{p_{ij}\}$ analytically. The two steps were iterated until convergence. To recover the obtained minimum spanning tree and graph structure, we only need to check the non-zero entries of $\{w_{ij}\}$. The detailed mathematical derivation of alternate convex searching and convergence analysis are given in **Sections S2.2.2** and **S2.2.3**.

The proposed method has two parameters σ and ℓ that control the model complexity of the learned tree structure. They can be estimated from data automatically by controlling the bias-variance tradeoff. For the purpose of this study, we employed the elbow method [41] to tune the parameters. The elbow method was originally developed to estimate the optimal number of clusters for a given dataset. The basic idea is to examine the percentage of variance explained as a function of the number of clusters, and choose a number of clusters so that adding another cluster does not yield much improvement in modeling the data. In our application, we fit a given dataset using a minimum spanning tree with a bounded length. It can be shown that the data fitting error decreases with respect to the total length of a tree that reflects model complexity, and at some point the rate of decrease markedly flattens off as the model begins to fit data noise. Therefore, we can use the elbow method to estimate parameters σ and ℓ that control the tree length. One issue in our application is that if the

two parameters are estimated simultaneously, it leads to a two-dimensional search and it is hard to determine an elbow in a two-dimensional surface. To address this issue, we performed the search of the two parameters separately. First, we made a guess of σ and estimated ℓ by using the elbow method. The initial guess of σ can be estimated by using the leave-one-out-maximum likelihood criterion. Then, we fixed ℓ and estimated σ. To automatically determine the elbow position, we developed a method that performs a regression analysis that fits two lines to the two arms of an elbow curve and estimates the optimal parameter as the one that generates a principal curve with a length equal to that at the intersection of the two lines (see **Figure S9**). The two-stage estimation procedure worked very well on a wide variety of simulation data and breast cancer datasets (**Figures S7**, **S9**, and **S21**). We also found that the performance of our principal tree construction method is largely insensitive to a specific choice of the parameters, which makes parameter tuning and hence the implementation of our method easy, even for researchers outside of the machine learning community.

Existing methods for detecting branching structure usually involve some manual manipulations (**Section S4**). This is highly undesirable, since for structure learning prior information (e.g., the existence of branches and the number of branches) is generally unavailable. In contrast, our method relies on automatic parameter optimization, and once a principal tree is constructed, branches can be determined trivially. Before our method was applied to breast cancer data, it was intensively tested on synthetic data (**Section S2.5**).

Clustering analysis to identify genetically homogenous groups

By using the selected cancer genes, we next performed a clustering analysis to identify groups of tumor samples with homogenous genetic profiles. For the purpose of this study, the K-means method [42] was used. The optimal number of clusters was estimated by using gap statistic [43]. It is well known that K-means may return a local optimal solution. To identify robust and stable clusters, the technique of resampling-based consensus clustering [44] was used, where K-means clustering was repeated $1,000$ times and in each time 80% samples were drawn randomly without replacement from the entire dataset. The results of the $1,000$ runs were then aggregated into a consensus matrix that gave a visual representation of the frequency of two samples being grouped into the same cluster. To further assess the clustering robustness, the silhouette width of each sample was calculated, which is defined as the difference between its average similarity with samples in the same cluster and the largest average similarity with samples in different clusters. A cluster with an average silhouette width larger than 0 is generally considered stable.

Building a cancer progression model

Finally, by using the learned tree structure as a backbone, we combined the clustering and principal curve results to build a progression model and extract disease progression paths. Specifically, we represented a progression model as an undirected graph, where the vertices were the centroids of the clusters identified in the cluster analysis and they were connected based on the progression trend inferred from the principal curve. Let $\mathcal{P} = \{\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_N\}$ be the constructed principal curve. First, we projected the tumor samples $\{\mathbf{x}_1, \cdots, \mathbf{x}_N\}$ back on to the principal curve. Since $\mathcal{P}$ contains only a finite number of data points, multiple tumor samples can be projected onto the same point, making it difficult to determine the pseudo-time order of these samples for downstream analysis. To resolve the issue, we augmented the set $\mathcal{P}$ by interpolation and extrapolation. Specifically, if a point to which multiple samples were mapped is a leaf vertex, we extended the principal curve by using polynomial

curve fitting, and if a point was an inner point of the principal curve, we locally interpolated the curve and re-projected the samples onto an affine line determined by the inner point and its nearest point on the curve. After all the samples were projected onto the principal curve, the progression paths were extracted from the curve by finding the shortest path from a designated root vertex to all the leaf vertices of the principal curve. In this study, we used the projection of the mean of the normal samples as the root vertex to represent the origin of cancer progression. By following the same procedure, the centroids of the clusters were mapped onto the principal curve, and an undirected graph was constructed. Two projected centroids were connected if there are no other centroids between them along a progression path, and the edge was weighted by the curve distance of the two centroids measured along the progression path.

Method for Identifying Gene Mutations Associated with Cancer Progression

The development of a cancer progression model can inform a range of research goals (see **Section S10** for a detailed discussion). As an example of model utility, we performed a cancer genome analysis, focusing primarily on the detection of cancer driver gene mutations. Once a cancer progression model was constructed, we projected the tumor samples back onto the identified progression paths. Here, the projection of a sample is defined as a point on a progression path that is the closest to the sample (see **Figure S4**). By using the normal samples as the baseline, the static samples were ordered along a progression path according to the extent to which the tumors progress towards malignancy, and the ordered samples can be viewed as pseudo-time series data. This provides a unique opportunity to identify driver gene mutations and put their possible roles in the context of a dynamic disease process, which is previously attainable in static sampling data analysis.

We developed a new method, referred to as MutationPattern, that combines the information of mutation abundance and disease progression to detect driver gene mutations and delineate their dynamic patterns (**Figure 5a**). It consists of three major steps. First, tumor samples were mapped on to a progression model, then the mutation rates of individual genes were estimated as a function of a progression path, and finally null models were constructed to identify genes that showed a significant change in mutation incidence along the progression path. The developed method was tested on some previously described driver genes and passenger genes (**Figures S34** and **S35**).

Estimating mutation rates as a function of a progression path. Suppose that we have N tumor samples mapped onto a progression path. Let $\mathbf{y} = [y_1, \cdots, y_N] \in \Re_+^N$ be the progression distances of the tumor samples measured with respect to the centroid of the normal samples, and $\mathbf{M} \in \{0,1\}^{N \times J}$ be a patient-by-gene mutation data matrix, where $M(n,j) = 1$ if the jth gene in the nth sample carried a non-silent mutation and 0 otherwise. Without loss of generality, assume that $y_1 \leq \cdots \leq y_N$, and the patients in $\mathbf{M}$ were organized in the same order as $\mathbf{y}$. We used the non-parametric kernel regression method [45] to estimate the mutation rate of a gene as a function of a progression path:

$$P_j(y_n) = \sum_{i=1}^{N} k(y_i, y_n|\sigma) M(i,j) \Big/ \sum_{i=1}^{N} k(y_i, y_n|\sigma), 1 \leq n \leq N, 1 \leq j \leq J \,, \tag{3}$$

where $k(y_i, y_n|\sigma)$ is the Gaussian kernel and σ is the bandwidth that can be estimated through cross validation (see **Section S9.2**). By construction, $P_j(y_n)$ takes a value between 0 and 1, and can be interpreted as the probability of a tumor sample at position y_n carrying a mutation in the jth gene.

Computing test statistics. After the mutation rate of a gene was estimated, we next determined whether it showed a significant change along a progression path. To this end, we compared the estimated mutation rate with the average mutation rate and used as a test statistic the sum of squared errors given by

$$r_j = \sum_{n=1}^{N} \left(P_j(y_n) - \sum_{i=1}^{N} M(i,j)/N \right)^2, 1 \le j \le J. \quad (4)$$

Constructing null model and determining statistical significance. We next constructed a null model to assess the statistic significance of an observed test statistic. We first considered a constant background mutation model, by assuming that if a gene is not involved in cancer progression its mutations are random events uniformly distributed along a progression path. However, it is known that cancer progression is accompanied by the accumulation of genetic alterations due to impairment of DNA repair functions [2–5]. This means that even if a gene is a passenger gene, its mutation rate can increase slightly along a path. We found that this is indeed the case (see **Figures 4d** and **5b**). Therefore, a constant background mutation model is not appropriate. One possible way to address the issue is to build a null model by randomly permuting mutation data matrix $\mathbf{M}$ along the column direction so that the total number of mutations in each position in a progression path is fixed. However, this amounts to assuming that all genes under the null model have the same mutation rate, which is clearly inappropriate. By assuming that under the null model a mutation in each nucleotide position follows a binomial distribution, the probability of a gene carrying a mutation is proportional to its length [46,47]. Let $\mathbf{a} = [a_1, \cdots, a_J]$ be the exon lengths of the genes sequenced, and $\mathbf{m} = [m_1, \cdots, m_N]$ be the total numbers of mutations in the N tumor samples. Then, the probability of the jth gene containing at least one mutation in the nth samples under the null model can be estimated as:

$$\widetilde{P}_j(y_n) = 1 - \left(1 - a_j \Big/ \sum_{i=1}^{J} a_i \right)^{m_n}, 1 \le j \le J, 1 \le n \le N. \quad (5)$$

Once we estimated $\{\widetilde{P}_j(y_n)\}$, a null mutation data matrix $\widetilde{\mathbf{M}}$ was generated via random sampling, where $\widetilde{M}(n,j)$ took a value of 0 or 1 following a Bernoulli distribution specified by $\widetilde{P}_j(y_n)$. By using the same procedure described above, a null statistic can be computed, and the P-value of the jth gene can be computed as the occurrence frequency of the null statistics being larger than or equal to the observed test statistic. Finally, after we computed the P-values of all genes, we controlled the false discover rate (FDR) using the Benjamini-Hochberg procedure [48].

Point Set Registration for Microarray Data Alignment

Our progression modeling analysis performed on the METABRIC and TCGA RNA-seq data revealed a bifurcating progression process. To investigate whether a similar progression pattern could be derived from independent datasets, we performed a large-scale validation analysis on additional 25 breast cancer datasets. Since the 25 datasets were generated by different studies using a variety of gene expression profiling techniques, it is necessary perform data alignment in order to conduct biologically meaningful comparisons.

We used the iterative closest point algorithm for point set registration [49] to align two datasets. Let $\mathcal{A} = \{\mathbf{x}_1, \cdots, \mathbf{x}_N\}$ be the METABRIC dataset that was used as a discovery (or reference) dataset and was kept fixed, and $\mathcal{B} = \{\mathbf{y}_1, \cdots, \mathbf{y}_M\}$ be a validation (or source) dataset to be aligned against

the reference. Our goal is to find a spatial transformation to be applied to the source dataset so that a certain cost function is minimized. For the purpose of this study, we defined the cost function as the sum of the least square difference between each data point in $\mathcal{B}$ after alignment and its closest point found in $\mathcal{A}$. In order to maintain the geometric structure of the source dataset, we only considered rigid transformation that consists of only translation and rotation. In this study, the molecular subtype information of each tumor sample was also available. It is reasonable to assume that the samples of the same subtype in the two datasets should be aligned as close as possible. We thus limited the searching space of a transformation operation by searching for the closest point of $\mathbf{y}_m$ only in the samples in $\mathcal{A}$ with the same subtype as $\mathbf{y}_m$. This significantly reduced computational complexity and effectively alleviated the local minimum issue. The modified iterative closest point algorithm was formulated as a least-square optimization problem:

$$\min_{\{\mathbf{R},\mathbf{t}\}} \sum_{m=1}^{M} \left\| T\left(\mathbf{y}_m|\mathbf{R},\mathbf{t}\right) - f_{\mathcal{A}}\left(T(\mathbf{y}_m)\right)\right\|^2, \tag{6}$$

where $T(\mathbf{y}_m|\mathbf{R},\mathbf{t}) = \mathbf{R}\mathbf{y}_m + \mathbf{t}$ is a transformation operator specified by rotation matrix $\mathbf{R}$ and translation vector $\mathbf{t}$, and $f_{\mathcal{A}}(T(\mathbf{y}_m))$ is a function that finds the closest point of $T(\mathbf{y}_m)$ in the reference dataset $\mathcal{A}$ that has the same subtype label as $\mathbf{y}_m$. Note that the closest points of $T(\mathbf{y}_m)$ in $\mathcal{A}$ is unknown before learning. To address the issue, an iterative process was carried out. Specifically, for each sample $\mathbf{y}_m$ in $\mathcal{B}$, we first found the closest point in $\mathcal{A}$ with the same subtype as $\mathbf{y}_m$, and then estimated the combination of rotation and translation by minimizing the squared error cost function of (6) through singular value decomposition. Once we obtained a transformation operator, we transformed the source points and re-estimated their closest points. The above two steps were iterated until convergence. A more detailed discussion of the algorithm is given in **Section S7.2.1**.

Statistical Methods and Code Availability

The statistical methods used and developed in this study, including survival data analysis, enrichment analysis, Spearman's test for non-uniformly sampled data, are presented in **Section S5**. An open-source software package including all the methods used for cancer progression modeling and mutation pattern analysis was developed and is available at `http://www.acsu.buffalo.edu/~yijunsun/lab/cancer_progression_modeling.html`.

Results

Breast Cancer Progression Modeling

We applied the developed bioinformatics pipeline to the METABRIC data [17] to construct a progression model of breast cancer. The dataset contains the expression profiles of $25,160$ genes obtained from $1,989$ surgically excised primary breast tumor samples. Since only a small fraction of genes are likely to be involved in the biological processes of cancer development, we first performed feature selection by using the method described in **Methods** that identified 359 disease related genes (see **Section S3.2** for detailed experimental procedures and parameter estimation and **Table S2** for the identified genes). To obtain a general overview of data distribution supported by the selected features, we then performed a data visualization analysis using principal component analysis [45]. The dataset also contains 144 normal breast tissue samples, which we used as the baseline to determine the origin of cancer progression. By projecting each sample onto a three-dimensional space spanned by the top

three leading principal components, some small structures (i.e., clusters and small branches) may not be visible, but we can clearly see that the tumor samples form a linear, bifurcating progression path, starting from the normal tissue samples and diverging initially to either luminal A or basal subtypes. The linear trajectory through luminal A continues to luminal B and gradually transits to the HER2+ subtype (**Figures 3a** and **S10**). The two trajectory termini (i.e., HER2+ and basal) represent the most aggressive breast tumor subtypes [23, 24].

To formally describe the disease progression process, we applied the proposed progression modeling approach to the selected genes. First, we applied the K-means method [42] to the expression measures of the selected genes to detect genetically homogeneous groups. By using gap statistic [43], the number of clusters was estimated to be ten (**Figure 3b**). To promote a robust clustering assignment, a resampling based consensus clustering analysis [44] was performed. From the generated consensus matrix that measures the probability of a pair of samples being grouped into the same cluster (**Figure 3c**), we can clearly identify ten blocks along the anti-diagonal line. The robustness of clustering assignment was further confirmed by a silhouette width analysis that classified $1,652$ out of $1,989$ (83%) samples with a positive silhouette width (**Figure 3d**). Next, we used a new principal tree method described in **Methods** to formally describe the cancer progression process. The optimal regularization parameter and kernel width of the method were estimated by using the elbow method [41] (**Figure S9**). Finally, by using the constructed principal tree as a backbone, we combined the clustering and principal tree results to build a progression model of breast cancer presented in **Figure 3e**. Each node in the figure represents an identified cluster and the node size is proportional to the number of samples in the corresponding cluster. Two connected nodes indicate a possible inter-relationship, and the length of an edge connecting two nodes is proportional to the distance of the curve connecting the centers of the two nodes. The pie chart of each node depicts the percentage of the samples in the node belonging to one of the five PAM50 subtypes [50]. The overall structure of the constructed model is consistent with the data visualization result (**Figures 3a** and **S10**), suggesting that the model faithfully reflects the data distribution.

To help visualization and put the result into the context by referral to previous classification systems, we added the PAM50 subtype labels to the model. However, as shown by the continued subdivision of PAM50 subtypes [8, 51, 52], the PAM50 classification system does not represent the full complexity of breast tumor molecular profiles. Indeed, the consensus matrix clearly showed that the luminal subtypes can be further refined (**Figure 3c**). In the constructed progression model, significant side-branches are evident for both luminal A and luminal B subtypes, and further analysis of these luminal nodes showed that they had distinct copy number profiles, significantly different genome instability levels, and distinct clinical outcomes (**Figure S12**). Notably, starting from node 2, through nodes 7 and 3 and diverging to either node 1 or 9, the proportion of luminal A samples gradually decreased as luminal B samples increased (**Figure 3e**), and the genome instability increased monotonically along with a worsening prognosis (**Figure S12**). This result suggests that the luminal subtype is not a genetically homogenous group and can be further refined beyond the current luminal A/B classification. However, the identified luminal nodes do not form clear-cut clusters and have significant overlaps, particularly between adjacent nodes (e.g., nodes 2 and 7, nodes 3 and 9. See **Figure 3c, e**). Significant overlap was also evident between luminal B and HER2+ (nodes 1 and 5), suggesting that they share a progression relationship. This explains why several recent large-scale

benchmark studies found that existing subtyping methods could only achieve moderate concordance, particularly when classifying the luminal and HER2+ subtypes [25–27]. It may also explain why, in the TCGA breast cancer study [12], half of clinically defined HER2+ tumors were in the HER2+ mRNA group, and another half were predominantly in the luminal mRNA subtype.

A number of conceptual progression models have recently been proposed regarding the origins of breast cancer subtypes and associated biological mechanisms [8,9]. One model proposes a distinct-path scenario where each discrete subtype follows a path of initiation and progression independently of the others. The alternative is a linear evolution model, which proposes that tumors gradually evolve from normal cells to malignant states through the accumulation of genetic alterations [8]. The third model describes two distinct pathways to breast cancer malignancy, either directly to basal-like subtype, or a stepwise path to luminal and HER2+ subtypes [9]. While all three models embrace the notion of cancer evolution, the first model implies that the subtypes are different diseases, while the alternative models suggest that subtypes are different stages of the same disease. Clarifying this issue could have a profound impact, as patient management and research strategies in the two scenarios could be entirely different. The bifurcation structure revealed in our model supports the third model as a representation of the breast cancer progression process. We should emphasize that our method is a generic, unbiased approach that makes no model assumption *a priori*. If the four major subtypes evolve directly from normal cells, we should be able to detect four independent paths connecting normal samples with the four subtypes, but this was clearly not the case. Our result suggests that basal and luminal subtypes are differentially derived from a normal cell origin, an idea consistent with the notion that ER+ (estrogen-receptor-positive) and ER- tumors are two fundamentally different biological entities. The idea that HER2+ phenotypes are derived from luminal tumors may also make biological sense. Through association of CNA data and putative driver gene expression (data not shown), we found that the copy numbers of the genes involved in the HER2 signaling pathway are significantly amplified in HER2+ samples relative to luminal samples, suggesting that the HER2+ phenotype develops from luminal through gene copy number alterations, and this event is distinct from progression to basal phenotypes. Our findings support recent studies that suggested that while cancer is a genetically and clinically heterogeneous disease, molecular subtypes are not hardwired, and genotypes and phenotypes can shift over time [2], as commonly seen across multiple organisms.

Progression Model Validation

We performed a series of interrogations that provided substantial support for the constructed model, and showed the utility of such a model for testing and generating hypotheses and providing novel insights into previous observations from the cancer-evolution perspective. Our modeling analysis revealed four major progression paths, referred to as N-B (normal to basal), N-H (normal through luminal A/B to HER2+), N-LB (normal through luminal A to the luminal B terminus), and N-LA (normal to the luminal A terminus) in the downstream analysis.

Similar progression patterns repeatedly observed in 28 independent datasets

To investigate whether a progression model with a similar topological structure could be derived from an independent dataset, we performed a computational analysis on the TCGA RNA-Seq breast cancer dataset [12]. Since the data sources are not entirely compatible (the TCGA study employed a different gene expression-profiling platform from that used in the METABRIC study), we first mapped the 359 genes selected from the METABRIC data analysis back to the TCGA data, and then performed the

described clustering and principal curve analyses and progression model construction. A total of 354 genes were also present in the TCGA data. By applying the same analytical protocol, eight robust clusters were identified (**Figure S23**). The different number of clusters may be attributed to various factors, including the different sample sizes or microarray platforms used, but despite these differences, the overall structure of the progression model constructed using the TCGA data (i.e., the bifurcation structure and the order of cluster connections) was almost identical to that constructed using the METABRIC data (**Figures 3f** and **S22**). The learned structure also supported a linear bifurcating progression path, distinctly diverging from a normal tissue origin to either luminal A or basal subtypes. The linear trajectory through luminal A transitioned to luminal B, and on to the HER2+ subtype. Side-branch termini were also evident for clusters comprised of predominantly luminal A/B subtypes.

We went on to demonstrate that the same progression pattern was repeatedly observed in additional 25 breast cancer datasets (**Table S1**). The majority of the validation datasets had small numbers of samples (ranging from 50 to 300), which precluded the construction of progression models directly from individual datasets, since the prerequisite of progression modeling by using static data is that the number of samples are large enough so that progression paths are well populated. Another major difficulty of model validation using gene expression data generated by different studies is that they are not always directly comparable, due to various factors including RNA quantity and quality, different gene expression profiling techniques, and different technical protocols used in data preparation. Thus, it was necessary to perform data alignment in order to conduct biologically meaningful comparisons. To this end, we developed a new validation method (see **Methods** and **Section S7.2** for details). Briefly, we first performed point set registration to align individual datasets against the METABRIC data that was used as a reference dataset, and then mapped the aligned validation samples onto the METABRIC progression path to examine whether the sample distribution of the validation data is consistent with that observed in the METABRIC model. We found that the distribution pattern was markedly similar in all 27 datasets analyzed (**Figures S46-S71**). Given the total number and diversity of samples, and the range of analytical platforms included in the combined datasets, the above analysis provides strong evidence suggesting that the data pattern presented in **Figure 3e, f** is unlikely to be an artifact but universally present in breast cancer.

Mapping of clinical and genetic data back onto progression models

To further validate the constructed models, we mapped clinical and genetic variables onto the constructed model to investigate how they correlate with identified progression paths. Examples with implications for cancer progression that we can test include increasingly poor survival functions, deviation of morphological traits of tumor cells from normal cells, and the accumulation of genetic alterations.

We first performed a survival data analysis to examine the relationship between the clinical outcomes of the identified subgroups and the modeled disease progression paths to malignancy. Due to the lack of follow-up data for the TCGA data (the median overall follow-up was 17 months vs 98 months for the METABRIC data, and there were only 93 overall survival events), we only performed the analysis on the METABRIC model. Kaplan-Meier plots of disease-specific survival for the ten groups identified from the METABRIC data revealed a clear trend of worsening survival function along the major trajectories to malignancy through normal to basal (node 10 to node 6), or to luminal A dead-end (node 2 to node 8), or to luminal B dead-end (node 2 through nodes 7, 3 to node 9), or

through luminal types to HER2+ (node 2 through nodes 7, 3, 1, 5 to node 4) (**Figure 4a**). As would be expected, each cluster, or node, on a linear path generally had a worse survival function than the preceding cluster.

We next mapped histological tumor grades onto the constructed model. Tumor grade is a measure of the extent to which tumor cells morphologically deviate from normal cells [53]. If our constructed model is valid, we would expect low-grade and high-grade tumors to be distributed at early and late steps of a progression path, respectively. This was investigated using data available from the METABRIC dataset (the TCGA data did not contain the grade information). Each sample was projected onto the specific progression path, and then a running sum score of grades was calculated. Enrichment analysis indeed identified a strong association between increasing grades and progression paths (**Figure S25**). Since the evaluation of histological grades particularly the intermediate grade is rather subjective, a method to derive molecular grades has been developed [54]. Mapping of the data onto the METABRIC model revealed that molecular grades were also highly correlated with the four major progression paths (**Figure 4b**). Statistical significance was determined by Spearman's test. Since the tumor samples mapped onto the model are non-uniformly distributed along the progression paths, an improved Spearman's test was developed (see **Section S5.3**). Strong correlation was observed (N-B: $\rho = 0.91$, P-value $= 5.5 \times 10^{-108}$; N-H: $\rho = 0.82$, P-value $= 8.7 \times 10^{-282}$; N-LB: $\rho = 0.89$, P-value $= 0$; N-LA: $\rho = 0.65$, P-value $= 3.9 \times 10^{-110}$). Consistent with the results obtained on the METABRIC model, the molecular grade index was also found to be highly correlated with the progression trajectories modeled using the TCGA data (N-B: $\rho = 0.74$, P-value $= 5.9 \times 10^{-36}$; N-H: $\rho = 0.85$, P-value $= 5.7 \times 10^{-150}$; N-LB: $\rho = 0.92$, P-value $= 7.0 \times 10^{-215}$; N-LA: $\rho = 0.74$, P-value $= 4.1 \times 10^{-68}$. **Figure S27**). These findings support the validity of the progression model in that statistically significant correlations were identified, but also because it aligns with established grade associations. The majority of luminal A tumors are reported as low grade, luminal B are typically graded higher than luminal A, and HER2+ tumors are primarily high-grade [55]. It would be difficult to interpret this pattern in a discrete disease model. It has been proposed that this pattern can be explained by a more complex inventory of luminal B phenotypes [8], but seen now in the context of cancer progression, it could also be explained by a progressive transition from luminal A through luminal B to the aggressive HER2+.

Finally, we mapped two genetic variables, namely overall mutation rate and genome instability index (GII), onto the constructed progression models. Here, GII of a sample is defined as the sum of the magnitude of copy number alterations including amplification and deletion in all genes in the sample. Cancer evolution theory suggests that cancer development is accompanied by the accumulation of genetic alterations in somatic cells [2–5]. Among them, mutations and copy number alterations play a central role in tumorigenesis [56, 57], and genome instability is generally considered an enabling characteristic of cancer progression [6]. Thus, if the model is valid, we might expect both somatic mutation rates and GII to be positively correlated with the modeled progression trajectories. Mapping data from each sample on to the TCGA progression tree revealed that this is indeed the case for both overall mutation rate (N-B: $\rho = 0.42$, P-value $= 1.2 \times 10^{-8}$; N-H: $\rho = 0.59$, P-value $= 9.8 \times 10^{-47}$; N-LB: $\rho = 0.54$, P-value $= 4.6 \times 10^{-36}$; N-LA: $\rho = 0.4$, P-value $= 3.5 \times 10^{-13}$. **Figure 4d**), and GII index (N-B: $\rho = 0.61$, P-value $= 3.5 \times 10^{-18}$; N-H: $\rho = 0.62$, P-value $= 7.4 \times 10^{-50}$; N-LB: $\rho = 0.89$, P-value $= 7.3 \times 10^{-170}$; N-LA: $\rho = 0.7$, P-value $= 6.2 \times 10^{-51}$. **Figure S27**). Despite the fact that individual

patients have significantly different mutation incidents, a clear monotonically increasing trend was also observed for both silent and non-silent mutation rates along all four progression paths (**Figures S28** and **S29**). Consistent with the results obtained on the TCGA model, GII was also significantly correlated with progression in the model constructed using METABRIC data (N-B: $\rho = 0.64$, P-value $= 5.9 \times 10^{-36}$; N-H: $\rho = 0.48$, P-value $= 6.7 \times 10^{-70}$; N-LB: $\rho = 0.78$, P-value $= 6.7 \times 10^{-196}$; N-LA: $\rho = 0.53$, P-value$= 3.7 \times 10^{-66}$. **Figure 4c**). The mutation data analysis was not performed on the METABRIC data since only 170 genes have mutation information. The significant correlations of both somatic mutation rate and genome instability with progression models built from two independent datasets provide strong evidence supporting the validity of the proposed model.

Identifying Gene Mutations Associated with Cancer Progression

Discerning driver gene mutations from copious passenger mutations is a central task of large-scale cancer studies [12,17,58–60]. By definition [61], driver gene mutations are those that confer a selective growth advantage to the cells where they reside and cause clonal expansion, while passenger mutations do not have a direct or indirect impact on the cell survival-to-death ratio and are simply passed on through disease progression. The mainstay methods used today (e.g., MutSig [47] and MuSic [46]) are prevalence-based methods that work by searching a large number of samples for genes that are mutated more frequently than random chance [59]. While existing methods somehow embrace the notion of cancer evolution as they aim to identify driver genes, by lumping tumor samples together, they can only catalog frequently mutated genes and do not provide information regarding how a gene mutation promotes cancer progression. Different gene mutations may play specific roles. While some tumors carrying certain mutations can become dormant, other mutations may be responsible for the splitting of progression paths. With the development of a cancer progression model, it is now possible to delineate the dynamic patterns of individual gene mutations and place their possible roles into a disease progression context.

We applied the developed MutationPattern method to the TCGA mutation data to detect putative driver genetic mutations. The dataset contains $54,013$ non-silent mutations in $13,870$ genes in 958 breast tumor samples. Each sample harbored an average of 54 mutations, however, the distribution of the numbers of mutations was highly heterogeneous (**Figure S33**). Sixteen hyper-mutated samples that harbored a significantly large number of mutations compared to other samples were removed from the analysis. Since it is not reliable to estimate the mutation rate of a gene if there are only a few samples containing mutations in that gene, we restricted analysis to genes that were mutated in $> 1\%$ samples in a given progression path. A total of 51 genes were identified that had a significant change in their mutation incidence in at least one progression path (FDR < 0.05. **Table S3**, **Figure 5f, g**). Candidates included previously reported cancer driver genes (*TP53*, *CDH1*, *PIK3CA*, *GATA3*, *MAP3K1*, and *MAP2K4*) and some yet to be characterized (*DOCK11*, *QSER1*, and *ITSN2*). Based on mutation dynamic patterns, we found that genes could be classified into four categories, each with potential biological and clinical implications.

Passenger mutation pattern: The mutation pattern of a gene across the progression model is similar to those generated from its null models. Most genes belong to this category. As an example, **Figure 5b** shows the estimated mutation rate of *TTN* and those derived from its null model along the N-H progression path. Due to the prevalence of *TTN* mutations in breast cancer ($\sim 20\%$), it was nominated as a cancer gene in a number of studies [62–64]. However, our analysis showed that while

the mutation rate of *TTN* exhibits an upward trend along a progression path, the observed data had no difference from those derived from its null model (P-value = 0.57). This suggests that a mutation in the gene provides no competitive advantage with respect to breast cancer malignancy, and that the observed high mutation rate of *TTN* is simply due to the extreme length of the gene [12, 47]. The mutation patterns of other putative passenger genes [47] that code large proteins (*MUC16*, *RYR1*, *DNAH11*, *USH2A*) were also examined, but none of these genes exhibited a distinct mutation pattern associated with breast cancer progression (**Figure S35**).

Monotonically increasing mutation pattern: The mutation rate of a gene significantly increases along one or more progression paths compared with its null model. Most of the 51 progression-associated genes are in this category (**Table S3**). Notably, the detected genes on the four paths were mutually exclusive (P-value $< 10^{-5}$, the exact test), implying distinct differences in the major biological processes involved in specific cancer progression paths. *TP53* is the only mutated gene identified as being significantly associated with progression along all four paths (**Figure S37**, **Table S3**), a finding consistent with the pivotal role played by this gene product in DNA repair and genome stability [65,66], but distinct differences between the two major pathways to malignancy were described by the model. At the onset of the N-B path, about 35% tumors already contained a mutation in *TP53*, and the percentage quickly reaches 80%. In contrast, $< 1\%$ of tumors in the N-H path had *TP53* mutations at the onset, and the percentage gradually increases to 90%. Interestingly, there is an inflection point at the progression distance of 0.6, which corresponds to the bifurcation that leads to the N-H and the N-LB branches (**Figure S37**). Associated with the elevation of the mutation rate in *TP53* along the N-H path is a markedly worsening survival function (**Figure 4a**).

A total of 13 mutated genes were found to be significantly enriched at the luminal A side-branch, including *PIK3CA*, *GATA3*, *MAP2K4*, *CBFB*, and *CTCF* (**Figure 5f**). The genes that are not associated with progression beyond the N-LA path may play a role in tumorigenesis or early tumor establishment, but may not drive tumors to the most malignant phenotypes. The 14 genes detected to have an upward mutation trend along the N-LB path include *TP53*, *GATA3*, *RP1*, and *PTPRD*. *GATA3* is an example of a mutated gene that is associated with a luminal phenotype but does not extend into either basal or HER2+ phenotypes. The 21 genes identified in the N-H path with an upward mutation trend include *TP53*, *ERBB3*, *RB1*, *DOCK11*, and *QSER1*. Notably, except for *TP53*, these genes have no mutations (or very few) present prior to the inflection to the HER2+ branch (**Figure 5f**). This suggests that these genes play a late role in the development of HER2+ tumors, and although mutation rates are low when viewed across all breast cancers, in the context of a progression model these mutations are strongly associated with a shift to malignancy. Considering that HER2+ tumors have extremely unfavorable clinical outcomes, our result provides a way to prioritize experimental evaluation of the functions of the identified cancer driver genes.

Bell-shaped mutation pattern: Cancer evolution theory states that during disease development while some tumor cell clones carrying certain genetic mutations thrive due to selective genetic advantages, others become dormant or extinct [2]. If this is the case, we might expect to observe some bell-shaped mutation patterns where the mutation rate of a gene first increases and then decreases significantly along a progression path. We did observe such a patterns in our study for a number of genes. A typical example is *CDH1* (**Figure 5e** and **S43**). It can be seen that the majority of *CDH1* mutations occurred before the intersection between the HER2+ and luminal B branches (the second

broken line in **Figure 5e**). If such mutations ceased to be driving events at some point and became passenger mutations, then the observed rate would at least level off or even increase slightly (as seen with *TTN* in **Figure 5b**), but this is not the case. A bell-shaped pattern indicates that tumors with *CDH1* mutations become dormant or extinct, and do not progress further into either of the lethal N-LB or N-H paths. The specific mutations may individually, or coordinately, functionally inhibit progression and thus favor other clones, or are part of a rate-limiting environment that drives further tumor cell evolution. A similar mutation pattern was also observed on *RUNX1*, *PIK3CA*, *KIF21B*, *MED23*, *SF3B1*, and *HLA-DRB1* (**Figure S43**).

We found that bell-shaped mutation patterns in other genes suggest a driving role in luminal A (*CBFB*, *MYB*, *CTCF*, *MAP2K4*) or luminal B tumors (*GATA3*), but these are lost along the N-H path (**Figure S44**). An example of the latter is *GATA3*, where mutations were monotonically increasing in early luminal samples and highly enriched in the luminal B side-branch, but only a few *GATA3* mutations were located at the beginning of the HER2+ branch (**Figure S44**). This suggests that the mutated *GATA3* gene is a driver of luminal A/B tumors but may also influence the selective dominance event that occurs at the inflection of N-LB and N-H progression trajectories.

It was observed that while the overall mutation rate is lowest in luminal A and highest in HER2+ and basal subtypes, the significantly mutated genes are considerably more diverse within luminal A tumors [12]. Our analysis suggests that luminal A may be an early, intermediate stage in cancer progression that provides a mutated gene repertoire for subsequent natural selection. After several rounds of selection, tumors with specific gene mutations may become dormant or extinct. Indeed, as shown in **Figure 5g**, the peaks of the bell-shaped mutation patterns mostly occur before entering the HER2+ branch, explaining why HER2+ has less significantly mutated genes than luminal A, even though its overall mutation rate is much higher.

High-level mutation pattern: Mutations in a number of well-known putative cancer genes, including *FOXA1*, *ERBB2*, *MLL3*, *NCOR1* and *PTEN*, follow another pattern. Interestingly, although they were mutated more frequently than expected by random chance and thus generally regarded as cancer driver genes by prevalence-based methods [46, 47], relative to their respective null models, the mutation rates of these genes did not change significantly along any progression path (**Figure 5d** and **S45**). This suggests that mutations in these genes do not offer any malignant growth advantage. The observed high mutation levels in these genes indicate that their function may primarily be in cancer initialization or in core tumor cell maintenance, but do not drive cancer progression.

We should emphasize that the above described analysis can only be performed after a progression model is constructed. Similar analyses can be performed on other genetic alterations (e.g., copy number, microRNA, and methylation), and integration of information on gene interactions and patient genome wide information would reveal more genetic and epigenetic insights into cancer development at both gene and pathway levels (**Section S9.4**, **Figure S42**).

Comparison with Existing Approaches

We performed an extensive experiment comparing the developed CancerMapp pipeline with four existing approaches, namely SPD [20], PAD [21], DPT [67] and SCMC [36]. While DPT and SCMC were not designed for cancer progression modeling, SPD and PAD have been applied to small cancer datasets to construct preliminary models. However, unlike the presented study, the constructed models were not validated and no further analyses were performed to demonstrate model utilities. Our anal-

ysis showed that the existing methods are not sufficient to construct a cancer progression model with branching structures. CancerMapp overcomes many technical limitations of the existing approaches. Specifically, our method includes feature selection, relies on automatic parameter optimization, is robust again noise, and works well for high-dimensional data without making any assumptions about biological processes. Methodological comparison and experimental results are presented in **Section S4**.

Discussion

Advancing sequencing and molecular profiling techniques are enabling the cataloging of cancer associated genetic events in unprecedented detail, but to date, it has been difficult to put the observed changes into the context of the dynamic disease process. In order to understand how cancer progresses to a malignant, life-threatening disease, we require models of disease progression. This is difficult because we typically can only obtain genetic data from excised tissues. In this study, we developed a systematic approach that can overcome the static sampling limitation and enable researchers to leverage the vast tissue archive for the study of disease dynamics. The application of the proposed method to large-scale breast cancer genomic data identified a bifurcating progression model describing two distinct pathways to breast cancer malignancy, either directly to the aggressive basal-like subtype with little deviation, or a stepwise, more indolent path through the luminal subtypes to the HER2+ phenotype. The replication of the detailed data structure in the TCGA dataset, the observation of the bifurcating structure in additional independent datasets, and the post-construction association analysis of survival data and other genetic and clinical variables support the validity of the model. To demonstrate the utility of the constructed model, we performed a mutation data analysis to identify putative cancer driver genetic mutations within the cancer-progression framework.

As with any biology process, cancer development is inherently dynamic. We should emphasize that a progression model constructed through a computational study has to be verified experimentally, but such a model could provide investigators with testable hypotheses and inform a range of research fields, as demonstrated above and discussed in **Section S10**. The utility of a progression model will increase as its resolution is further refined. In this study, we used cancer subtypes as a template to select cancer related genes, and the proposed strategy outperformed existing methods (**Section S4**). However, there is no guarantee that the selected genes would enable us to identify small branches within each subtype. We are developing a method for selecting relevant features that will enable us to uncover subtle structures while maintaining a sample distribution that is compliant with existing subtyping systems. Analyses performed in this study also provided evidence that the incorporation of the complete range of quantitative molecular data could further increase the model resolution (**Section S9.4.1**), and this work provides a technical foundation for performing such analysis. Although here we focus on breast cancer, the analytical strategy is equally applicable to model other cancers and other human progressive diseases where the lack of time-series data to study system dynamics is an unavoidable problem.

References

[1] Nowell, P. C. (1976) The clonal evolution of tumor cell populations. *Science,* **194**(4260), 23–28.

[2] Greaves, M. and Maley, C. C. (2012) Clonal evolution in cancer. *Nature,* **481**(7381), 306–313.

[3] Yates, L. R. and Campbell, P. J. (2012) Evolution of the cancer genome. *Nature Reviews Genetics,* **13**(11), 795–806.

[4] Podlaha, O., Riester, M., De, S., and Michor, F. (2012) Evolution of the cancer genome. *Trends in Genetics,* **28**(4), 155–163.

[5] Stratton, M. R., Campbell, P. J., and Futreal, P. A. (2009) The cancer genome. *Nature,* **458**(7239), 719–724.

[6] Hanahan, D. and Weinberg, R. A. (2011) Hallmarks of cancer: the next generation. *Cell,* **144**(5), 646–674.

[7] Beerenwinkel, N., Schwarz, R. F., Gerstung, M., and Markowetz, F. (2015) Cancer evolution: mathematical models and computational inference. *Systematic Biology,* **64**(1), e1–e25.

[8] Creighton, C. J. (2012) The molecular profile of luminal B breast cancer. *Biologics: Targets & Therapy,* **6**, 289.

[9] Anderson, W. F., Rosenberg, P. S., Prat, A., Perou, C. M., and Sherman, M. E. (2014) How many etiological subtypes of breast cancer: two, three, four, or more?. *Journal of the National Cancer Institute,* **106**(8), 165.

[10] Navin, N., Krasnitz, A., Rodgers, L., Cook, K., Meth, J., Kendall, J., Riggs, M., Eberling, Y., Troge, J., Grubor, V., et al. (2010) Inferring tumor progression from genomic heterogeneity. *Genome Research,* **20**(1), 68–80.

[11] Navin, N., Kendall, J., Troge, J., Andrews, P., Rodgers, L., McIndoo, J., Cook, K., Stepansky, A., Levy, D., Esposito, D., et al. (2011) Tumour evolution inferred by single-cell sequencing. *Nature,* **472**(7341), 90–94.

[12] The Cancer Genome Atlas Network (2012) Comprehensive molecular portraits of human breast tumours. *Nature,* **490**, 61–70.

[13] Desper, R., Jiang, F., Kallioniemi, O.-P., Moch, H., Papadimitriou, C. H., and Schäffer, A. A. (1999) Inferring tree models for oncogenesis from comparative genome hybridization data. *Journal of Computational Biology,* **6**(1), 37–51.

[14] Caravagna, G., Graudenzi, A., Ramazzotti, D., Sanz-Pamplona, R., De Sano, L., Mauri, G., Moreno, V., Antoniotti, M., and Mishra, B. (2016) Algorithmic methods to infer the evolutionary trajectories in cancer progression. *Proceedings of the National Academy of Sciences,* **113**(28), E4025–34.

[15] Fearon, E. R., Vogelstein, B., et al. (1990) A genetic model for colorectal tumorigenesis. *Cell,* **61**(5), 759–767.

[16] The International Cancer Genome Consortium (2010) International network of cancer genome projects. *Nature,* **464**(7291), 993–998.

[17] Curtis, C., Shah, S. P., Chin, S.-F., Turashvili, G., Rueda, O. M., Dunning, M. J., Speed, D., Lynch, A. G., Samarajiwa, S., Yuan, Y., et al. (2012) The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature,* **486**(7403), 346–352.

[18] Magwene, P. M., Lizardi, P., and Kim, J. (2003) Reconstructing the temporal ordering of biological samples using microarray data. *Bioinformatics,* **19**(7), 842–850.

[19] Gupta, A. and Bar-Joseph, Z. (2008) Extracting dynamics from static cancer expression data. *IEEE/ACM Trans. Computational Biology and Bioinformatics,* **5**(2), 172–182.

[20] Qiu, P., Gentles, A. J., and Plevritis, S. K. (2011) Discovering biological progression underlying microarray samples. *PLoS Computational Biology,* **7**(4), e1001123.

[21] Nicolau, M., Levine, A. J., and Carlsson, G. (2011) Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival. *Proceedings of the National Academy of Sciences,* **108**(17), 7265–7270.

[22] Sun, Y., Yao, J., Nowak, N., and Goodison, S. (2014) Cancer progression modeling using static sample data. *Genome Biology,* **15**(8), 440.

[23] Sørlie, T., Perou, C. M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., et al. (2001) Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences,* **98**(19), 10869–10874.

[24] Sørlie, T., Tibshirani, R., Parker, J., Hastie, T., Marron, J., Nobel, A., Deng, S., Johnsen, H., Pesich, R., Geisler, S., et al. (2003) Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proceedings of the National Academy of Sciences,* **100**(14), 8418–8423.

[25] Mackay, A., Weigelt, B., Grigoriadis, A., Kreike, B., Natrajan, R., A'Hern, R., Tan, D. S., Dowsett, M., Ashworth, A., and Reis-Filho, J. (2011) Microarray-based class discovery for molecular classification of breast cancer: analysis of interobserver agreement. *Journal of National Cancer Institute,* **103**(8), 662–673.

[26] Haibe-Kains, B., Desmedt, C., Loi, S., Culhane, A. C., Bontempi, G., Quackenbush, J., and Sotiriou, C. (2012) A three-gene model to robustly identify breast cancer molecular subtypes. *Journal of National Cancer Institute,* **104**(4), 311–325.

[27] Weigelt, B., Mackay, A., A'hern, R., Natrajan, R., Tan, D. S., Dowsett, M., Ashworth, A., and Reis-Filho, J. S. (2010) Breast cancer molecular profiling with single sample predictors: a retrospective analysis. *The Lancet Oncology,* **11**(4), 339–349.

[28] Dy, J. G. and Brodley, C. E. (2004) Feature selection for unsupervised learning. *Journal of Machine Learning Research,* **5**, 845–889.

[29] Yao, J., Mao, Q., Goodison, S., Mai, V., and Sun, Y. (2015) Feature selection for unsupervised learning through local learning. *Pattern Recognition Letters,* **53**, 100–107.

[30] Yamada, M., Jitkrittum, W., Sigal, L., Xing, E. P., and Sugiyama, M. (2014) High-dimensional feature selection by feature-wise kernelized lasso. *Neural Computation,* **26**(1), 185–207.

[31] Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004) Least angle regression. *The Annals of Statistics,* **32**(2), 407–499.

[32] Bottou, L. (2004) Stochastic learning. In *Advanced Lectures on Machine Learning,* pp. 146–168, Springer, New York.

[33] Sun, Y., Todorovic, S., and Goodison, S. (2010) Local-learning-based feature selection for high-dimensional data analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* **32**(9), 1610–1626.

[34] Murata, N. (1998) A statistical study of on-line learning. In *Online Learning and Neural Networks,* pp. 63 – 92, Cambridge University Press, Cambridge, UK.

[35] Hastie, T. and Stuetzle, W. (1989) Principal curves. *Journal of the American Statistical Association,* **84**, 502–516.

[36] Ozertem, U. and Erdogmus, D. (2011) Locally defined principal curves and surfaces. *Journal of Machine Learning Research,* **12**, 1249–1286.

[37] Wu, B. Y. and Chao, K.-M. (2004) Spanning Trees and Optimization Problems, CRC Press, New York.

[38] Friedman, J. H. and Meulman, J. J. (2004) Clustering objects on subsets of attributes (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology),* **66**(4), 815–849.

[39] Gorski, J., Pfeuffer, F., and Klamroth, K. (2007) Biconvex sets and optimization with biconvex functions - a survey and extensions. *Mathematical Methods of Operations Research,* **66**, 373–407.

[40] Kruskal, J. B. (1956) On the shortest spanning subtree of a graph and the traveling salesman problem. *Proceedings of the American Mathematical Society,* **7**(1), 48–50.

[41] Sugar, C. A. Techniques for clustering and classification with applications to medical problems, PhD thesis Stanford University (1998).

[42] Lloyd, S. (1982) Least squares quantization in PCM. *IEEE Transactions on Information Theory,* **28**(2), 129–137.

[43] Tibshirani, R., Walther, G., and Hastie, T. (2001) Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B,* **63**(2), 411–423.

[44] Monti, S., Tamayo, P., Mesirov, J., and Golub, T. (2003) Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Machine Learning,* **52**(1-2), 91–118.

[45] Hastie, T., Tibshirani, R., and Friedman, J. (2009) The Elements of Statistical Learning, Springer, New York.

[46] Dees, N. D., Zhang, Q., Kandoth, C., Wendl, M. C., Schierding, W., Koboldt, D. C., Mooney, T. B., Callaway, M. B., Dooling, D., Mardis, E. R., et al. (2012) MuSiC: identifying mutational significance in cancer genomes. *Genome Research,* **22**(8), 1589–1598.

[47] Lawrence, M. S., Stojanov, P., Polak, P., Kryukov, G. V., Cibulskis, K., Sivachenko, A., Carter, S. L., Stewart, C., Mermel, C. H., Roberts, S. A., et al. (2013) Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature,* **499**(7457), 214–218.

[48] Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological),* **57**(1), 289–300.

[49] Besl, P. J. and McKay, N. D. (1992) Method for registration of 3-D shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* **14**(2), 239–256.

[50] Parker, J. S., Mullins, M., Cheang, M. C., Leung, S., Voduc, D., Vickery, T., Davies, S., Fauron, C., He, X., Hu, Z., et al. (2009) Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of Clinical Oncology,* **27**(8), 1160–1167.

[51] Guiu, S., Michiels, S., Andre, F., Cortes, J., Denkert, C., Di Leo, A., Hennessy, B., Sorlie, T., Sotiriou, C., Turner, N., et al. (2012) Molecular subclasses of breast cancer: how do we define them? The IMPAKT 2012 Working Group Statement. *Annals of Oncology,* **23**(12), 2997–3006.

[52] Ciriello, G., Sinha, R., Hoadley, K. A., Jacobsen, A. S., Reva, B., Perou, C. M., Sander, C., and Schultz, N. (2013) The molecular diversity of Luminal A breast tumors. *Breast Cancer Research and Treatment,* **141**(3), 409–420.

[53] Rakha, E. A., El-Sayed, M. E., Lee, A. H., Elston, C. W., Grainge, M. J., Hodi, Z., Blamey, R. W., and Ellis, I. O. (2008) Prognostic significance of Nottingham histologic grade in invasive breast carcinoma. *Journal of Clinical Oncology,* **26**(19), 3153–3158.

[54] Sotiriou, C., Wirapati, P., Loi, S., Harris, A., Fox, S., Smeds, J., Nordgren, H., Farmer, P., Praz, V., Haibe-Kains, B., et al. (2006) Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. *Journal of the National Cancer Institute,* **98**(4), 262–272.

[55] Calza, S., Hall, P., Auer, G., Bjöhle, J., Klaar, S., Kronenwett, U., Liu, E. T., Miller, L., Ploner, A., Smeds, J., et al. (2006) Intrinsic molecular signature of breast cancer in a population-based cohort of 412 patients. *Breast Cancer Research,* **8**(4), R34.

[56] Santarius, T., Shipley, J., Brewer, D., Stratton, M. R., and Cooper, C. S. (2010) A census of amplified and overexpressed human cancer genes. *Nature Reviews Cancer,* **10**(1), 59–64.

[57] Watson, I. R., Takahashi, K., Futreal, P. A., and Chin, L. (2013) Emerging patterns of somatic mutations in cancer. *Nature Reviews Genetics,* **14**(10), 703–718.

[58] Beroukhim, R., Mermel, C. H., Porter, D., Wei, G., Raychaudhuri, S., Donovan, J., Barretina, J., Boehm, J., Dobson, J., Urashima, M., et al. (2010) The landscape of somatic copy-number alteration across human cancers. *Nature,* **463**(7283), 899–905.

[59] Stratton, M. R. (2011) Exploring the genomes of cancer cells: progress and promise. *Science,* **331**(6024), 1553–1558.

[60] Wood, L. D., Parsons, D. W., Jones, S., Lin, J., Sjöblom, T., Leary, R. J., Shen, D., Boca, S. M., Barber, T., Ptak, J., et al. (2007) The genomic landscapes of human breast and colorectal cancers. *Science,* **318**(5853), 1108–1113.

[61] Vogelstein, B., Papadopoulos, N., Velculescu, V. E., Zhou, S., Diaz, L. A., and Kinzler, K. W. (2013) Cancer genome landscapes. *Science,* **339**(6127), 1546–1558.

[62] Greenman, C., Stephens, P., Smith, R., Dalgliesh, G. L., Hunter, C., Bignell, G., Davies, H., Teague, J., Butler, A., Stevens, C., et al. (2007) Patterns of somatic mutation in human cancer genomes. *Nature,* **446**(7132), 153–158.

[63] Hofree, M., Shen, J. P., Carter, H., Gross, A., and Ideker, T. (2013) Network-based stratification of tumor mutations. *Nature Methods,* **10**(11), 1108–1115.

[64] Stephens, P. J., Tarpey, P. S., Davies, H., Van Loo, P., Greenman, C., Wedge, D. C., Nik-Zainal, S., Martin, S., Varela, I., Bignell, G. R., et al. (2012) The landscape of cancer genes and mutational processes in breast cancer. *Nature,* **486**(7403), 400–404.

[65] Hollstein, M., Sidransky, D., Vogelstein, B., and Harris, C. C. (1991) p53 mutations in human cancers. *Science,* **253**(5015), 49–53.

[66] Olivier, M., Hollstein, M., and Hainaut, P. (2010) TP53 mutations in human cancers: origins, consequences, and clinical use. *Cold Spring Harbor Perspectives in Biology,* **2**(1), a001008.

[67] Haghverdi, L., Buettner, M., Wolf, F. A., Buettner, F., and Theis, F. J. (2016) Diffusion pseudotime robustly reconstructs lineage branching. *Nature Methods,* **13**, 845–848.

Figure 1: Overview of the presented stepwise study that consists of three major parts.

Figure 2: Overview of the bioinformatics pipeline for cancer progression modeling.

Figure 3: Progression modeling analysis performed on the METABRIC breast cancer data. (**a**) Principal component (PC) analysis provided a general view of sample distribution supported by the selected genes. To aid in visualization, each sample was annotated by its PAM50 subtype label, and mapped onto a principal tree (black line) in a three-dimensional space. **Figure S10** provides a more clear picture of data distribution. (**b-d**) Clustering analysis performed to detect genetically homogenous tumor groups. (**b**) The optimal number of clusters was estimated to be ten by gap statistic. (**c**) Re-sampling based-consensus clustering analysis to identify robust and stable clusters. The samples in the red box are predominantly luminal A/B tumors. (**d**) Silhouette width analysis to assess the robustness of clustering assignment. (**e, f**) Progression models of breast cancer built from the METABRIC and TCGA RNA-seq data, respectively. The overall structure of the progression models constructed using the two independent datasets is almost identical.

Figure 4: Model validation analysis provided support for the validity of the constructed progression models. (**a**) Disease-specific survival of ten breast cancer subgroups detected in the METABRIC data. A clear trend of worsening survival function was identified that was associated with progression along the four major malignant trajectories. (**b-d**) Spearman's rank correlation analysis of molecular grade, genome instability index, and overall mutation rate along the progression paths. Since only 170 genes in the METABRIC data have mutation information, mutation data analysis was performed using the TCGA data (see **Figure 3f** for the TCGA model).

Figure 5: Pseudo-time series analysis performed on the TCGA mutation data to identify gene mutations associated with cancer progression. Fifty one genes were found to have significant changes in their mutation incidences along progression paths (FDR < 0.05). (**a**) Overview of the proposed MutationPattern method used to delineate the dynamic patterns of individual gene mutations along a progression path. (**b-e**) Four distinct mutation patterns were observed. Examples of each are depicted: (**b**) *TTN*, (**c**) *TP53*, (**d**) *MLL3*, and (**e**) *CDH1*. The red line depicts the estimated mutation rate, and blue lines were generated from null models built by assuming that the corresponding gene plays no role in cancer development. Each red or blue line in the bar above a figure represents the presence or absence of a mutation in a sample, respectively. The first and second broken lines in (**e**) indicate the locations where the N-H path intersects with the LA terminal and LB terminal, respectively. (**f**) Genes showing an upward mutation trend along the N-LA, N-LB, N-H and N-B progression paths. (**g**) Mapping of identified progression-associated genes onto the TCGA model. Genes reported at the end of a path are those with an upward trend along the entire path. Genes with a bell-shaped pattern are marked at the bell-peak locations. Genes associated with normal samples are those mutated more frequently than random chance, but do not have significant changes along any progression path.

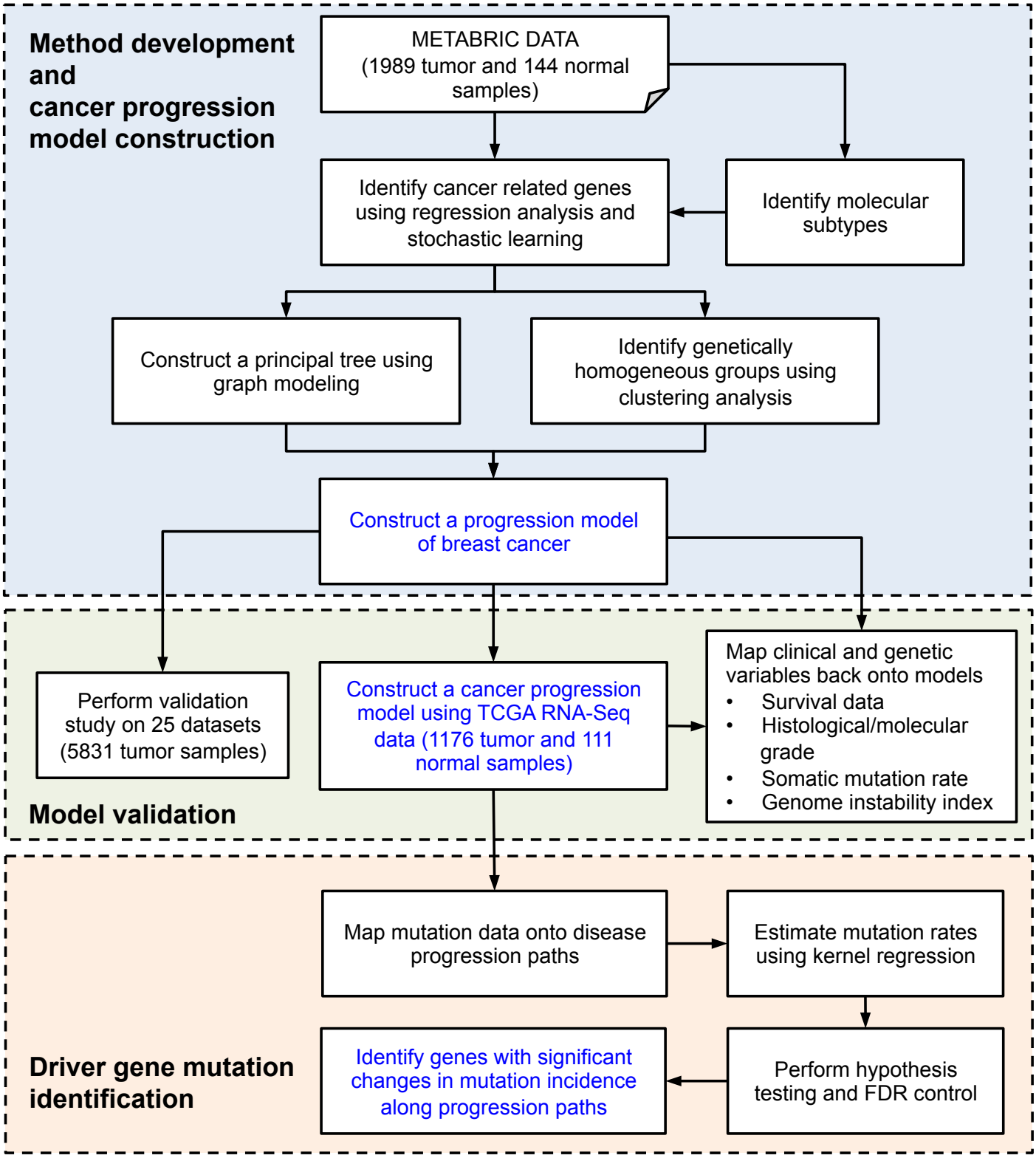

Figure 1: Overview of the presented stepwise study consisting of three major components.

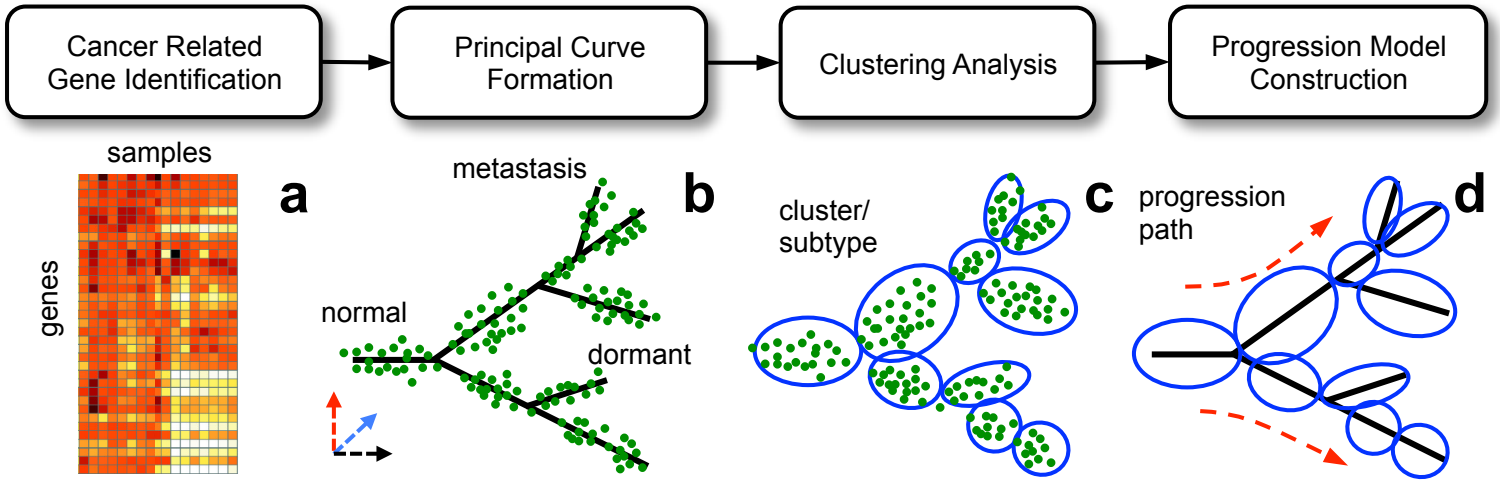

Figure 2: Overview of the bioinformatics pipeline for cancer progression modeling.

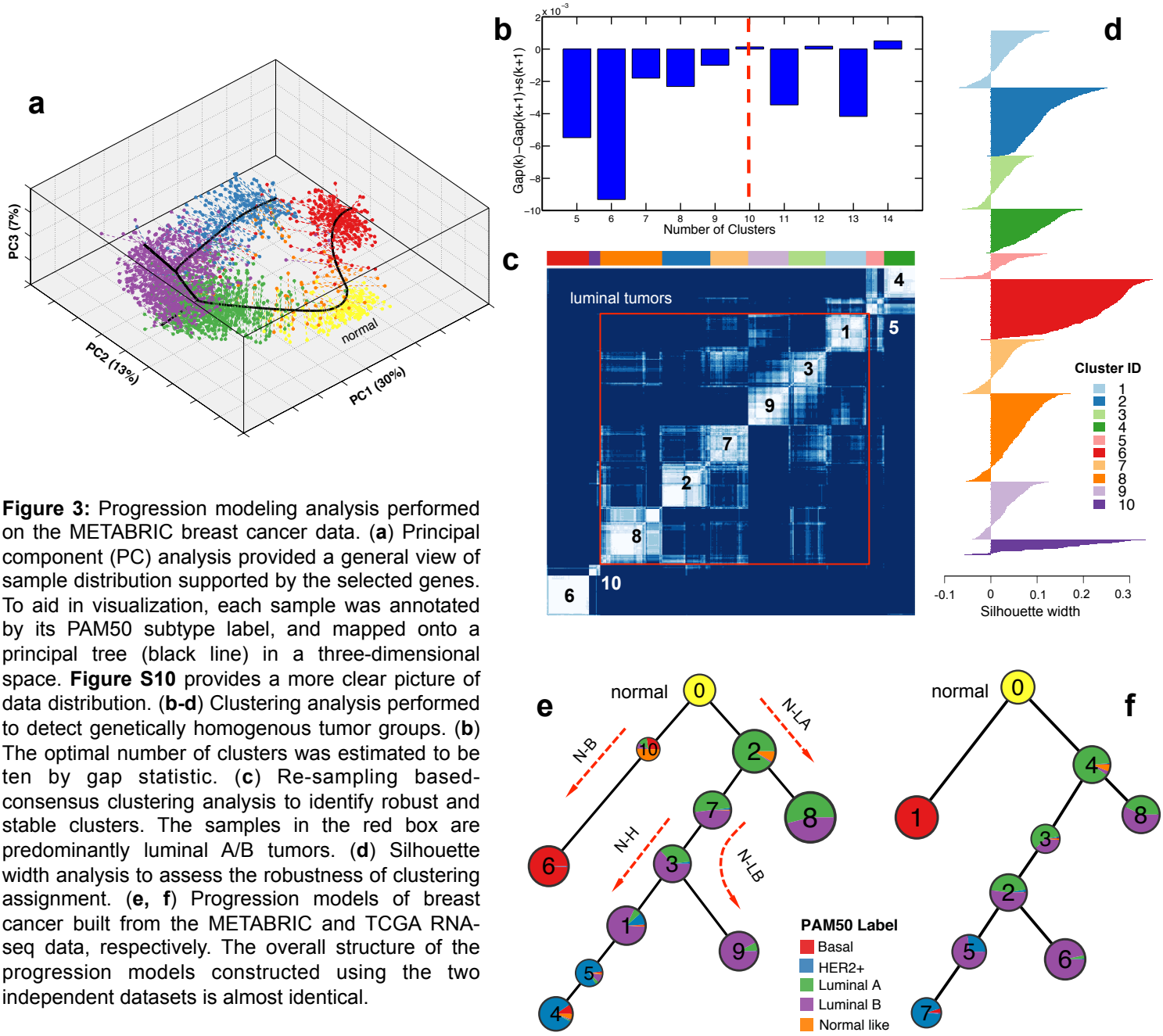

Figure 3: Progression modeling analysis performed on the METABRIC breast cancer data. (**a**) Principal component (PC) analysis provided a general view of sample distribution supported by the selected genes. To aid in visualization, each sample was annotated by its PAM50 subtype label, and mapped onto a principal tree (black line) in a three-dimensional space. **Figure S10** provides a more clear picture of data distribution. (**b-d**) Clustering analysis performed to detect genetically homogenous tumor groups. (**b**) The optimal number of clusters was estimated to be ten by gap statistic. (**c**) Re-sampling based-consensus clustering analysis to identify robust and stable clusters. The samples in the red box are predominantly luminal A/B tumors. (**d**) Silhouette width analysis to assess the robustness of clustering assignment. (**e, f**) Progression models of breast cancer built from the METABRIC and TCGA RNA-seq data, respectively. The overall structure of the progression models constructed using the two independent datasets is almost identical.

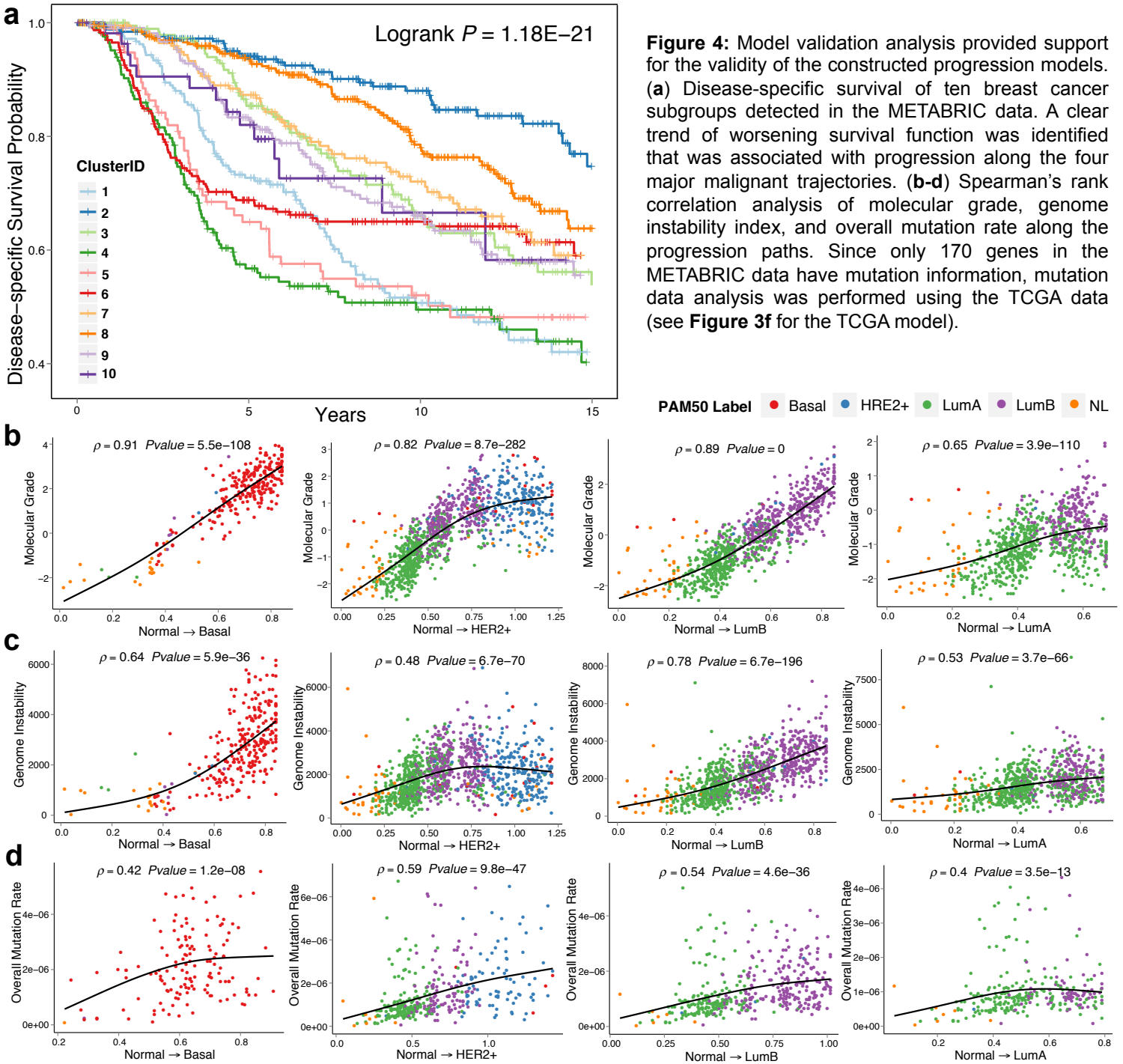

Figure 4: Model validation analysis provided support for the validity of the constructed progression models. (**a**) Disease-specific survival of ten breast cancer subgroups detected in the METABRIC data. A clear trend of worsening survival function was identified that was associated with progression along the four major malignant trajectories. (**b-d**) Spearman's rank correlation analysis of molecular grade, genome instability index, and overall mutation rate along the progression paths. Since only 170 genes in the METABRIC data have mutation information, mutation data analysis was performed using the TCGA data (see **Figure 3f** for the TCGA model).

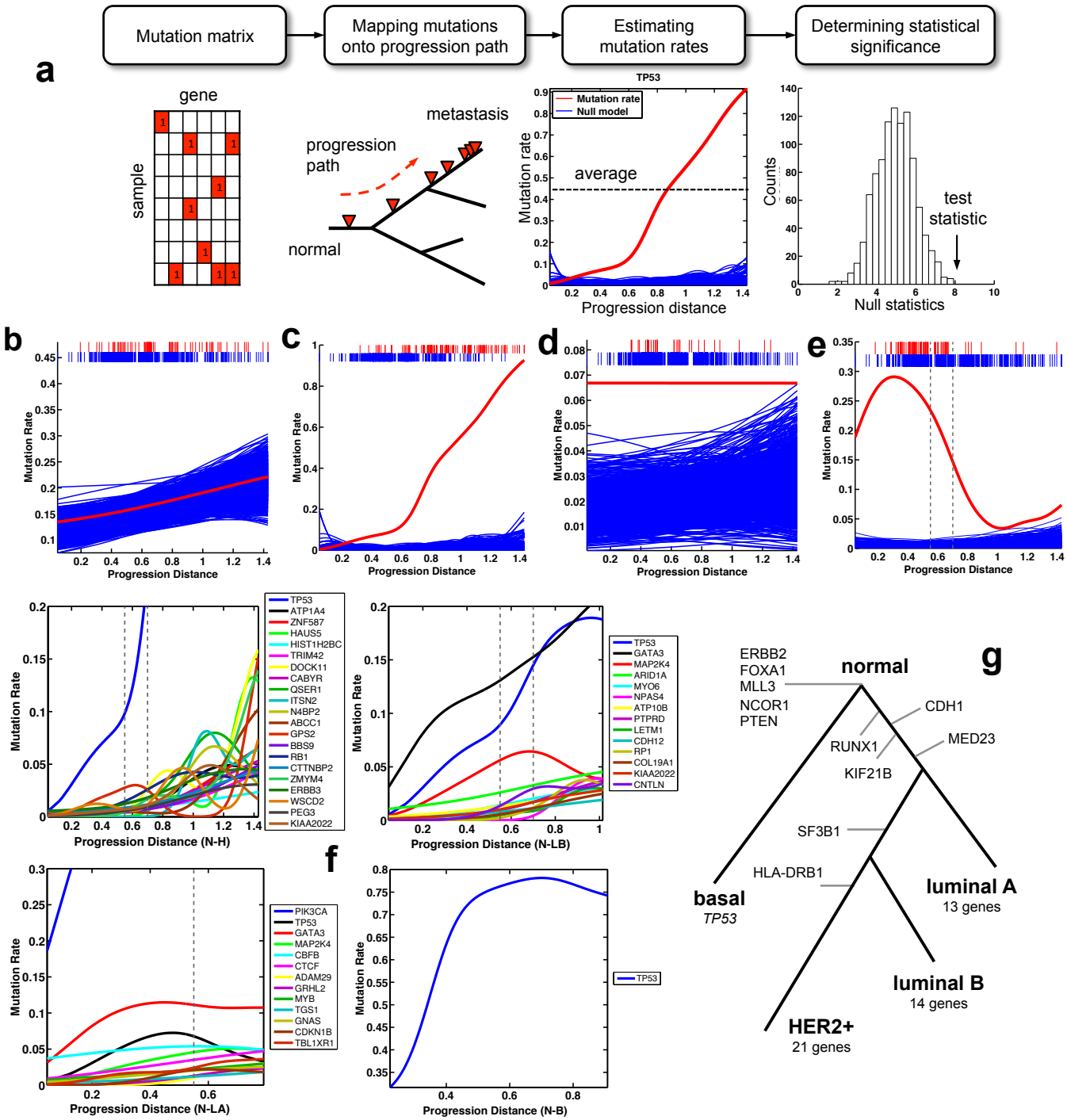

Figure 5: Pseudo-time series analysis performed on the TCGA mutation data to identify gene mutations associated with cancer progression. Fifty one genes were found to have significant changes in their mutation incidences along progression paths (FDR<0.05). (**a**) Overview of the proposed MutationPattern method used to delineate the dynamic patterns of individual gene mutations along a progression path. (**b-e**) Four distinct mutation patterns were observed. Examples of each are depicted: (**b**) *TTN*, (**c**) *TP53*, (**d**) *MLL3*, and (**e**) *CDH1*. The red line depicts the estimated mutation rate, and blue lines were generated from null models built by assuming that the corresponding gene plays no role in cancer development. Each red or blue line in the bar above a figure represents the presence or absence of a mutation in a sample, respectively. The first and second broken lines in (**e**) indicate the locations where the N-H path intersects with the LA terminal and LB terminal, respectively. (**f**) Genes showing an upward mutation trend along the N-LA, N-LB, N-H and N-B progression paths. (**g**) Mapping of identified progression-associated genes onto the TCGA model. Genes reported at the end of a path are those with an upward trend along the entire path. Genes with a bell-shaped pattern are marked at the bell-peak locations. Genes associated with normal samples are those mutated more frequently than random chance, but do not have significant changes along any progression path.