ECMarker: Interpretable machine learning model identifies gene expression biomarkers predicting clinical outcomes and reveals molecular mechanisms of human disease in early stages

Ting Jin[1], Flaminia Talos[2,3,4], Daifeng Wang[1,5, *]

[1]Department of Biostatistics and Medical Informatics, University of Wisconsin Madison, Madison, WI, 53706, USA

[2]Department of Pathology, Stony Brook Medicine, Stony Brook, NY, 11794, USA

[3]Department of Urology, Stony Brook Medicine, Stony Brook, NY, 11794, USA

[4]Stony Brook Cancer Center, Stony Brook Medicine, Stony Brook, NY, 11794, USA

[5]Waisman Center, University of Wisconsin-Madison, Madison, WI, 53705, USA

*** Correspondence:**
daifeng.wang@wisc.edu

Keywords: Interpretable machine learning, early cancer, gene expression biomarkers, gene regulatory network, semi-restricted boltzmann machine, cancer medicine

Abstract

Gene expression and regulation, a key molecular mechanism driving human disease development, remains elusive, especially at early stages. The increasing amount of genomics data gathered from patients allows for better understanding of gene regulatory mechanisms that occur during disease development at a system level. However, integrating such a large amount of population data is still a challenging task. To address this, we developed an interpretable machine learning model, ECMarker, to predict gene expression biomarkers and regulatory networks specific for disease stages such as early cancer. In particular, ECMarker was built on the semi-restricted Boltzmann machine, a machine learning model with multiple hierarchical layers. This model allows one to directly model and prioritize gene interactions for the classification of clinical phenotypes, revealing potential gene regulatory networks (GRNs) for various cancer stages. We applied ECMarker to the gene expression data of non-small cell lung cancer (NSCLC) patients. Unlike the state-of-the-art machine learning methods designed for classification only, ECMarker not only achieved an accuracy > 85% for classification but also predicted a lung cancer GRN that is highly consistent with previous findings. Additionally, ECMarker has clear clinical interpretability as it is able to prioritize biomarker genes for early lung cancer that can predict survival rates of early cancer patients (p-value < 0.0035). Finally, we identified a number of drugs currently in clinical use for late stages (e.g., Navitoclax, Cyclopamine and Erlotinib) with effects on these early cancer biomarker genes, suggesting potential novel insights on early cancer medicine. ECMarker is also general purpose for other disease types and available as an open-source tool at https://github.com/daifengwanglab/tools.

1 Background

Cancer development is a complex, dynamic process that is fundamentally driven by abnormal molecular mechanisms. However, understanding the mechanisms of cancer is still a challenging task, especially during early cancer development (Koeffler, McCormick and Denny 1991, Herbst, Morgensztern and Boshoff 2018). To this end, the tumor/node/metastasis (TNM) system has been widely used to characterize and classify cancer development into various stages (Ludwig and Weinstein 2005). The TNM stages were further associated with a number of individual molecular biomarkers and clinical outcomes such as survival rates (Ludwig and Weinstein 2005). However, with the advancement of whole genome sequencing of numerous human tumors, it became apparent that molecular profiles of various tumor stages might not necessarily be reflected by the TNM system. Therefore, using systems biology approaches to identify the biomarkers that drive cancers from early to late stages could allow for better understanding of cancer mechanisms and offer new venues for the development of new preventive and therapeutic strategies. Currently, there is a gap in academic understanding of the molecular biomarkers of early cancer and their underlying mechanisms at a system level. For example, the lung cancer, causing 27% of cancer-related deaths in the USA alone (Siegel, Miller and Jemal 2018), is localized to the lung; neither lymph nodes nor other organs are believed to be affected at the early stage. As the cancer progresses to a more advanced stage, nearby lymph nodes and other organs may be affected (Hu et al. 2008, Frost et al. 1984). This pathological difference suggests that the underlying molecular mechanisms of the early and late stages are different. Also, if cancer is diagnosed at an early stage, such as a localized stage, the five-year survival rate is approximately 50%; this is mostly due to surgical interventions involving lung removal. After the localized stages, survival rates decrease rapidly as cases involving the lymph nodes or other metastatic sites necessitate elaborate treatment strategies (Hu et al. 2008). Nearly 70% of patients with lung cancer present with locally advanced or metastatic disease at the time of diagnosis (Molina et al. 2008). Thus, although a number of studies have indicated that early localized stages are easier to treat and have better survival rates, the underlying molecular mechanisms remain elusive. Here, we hypothesized that early cancers have different molecular wiring at a system level and that understanding this wiring could reveal new biomarkers and mechanisms of early cancer development. Thus, it is essential to identify the specific biomarkers of early cancer to understand the molecular mechanisms driving cancer development; this would enhance early cancer diagnosis and therefore improve survival rates.

Detecting early cancer biomarkers, however, involves the inherent challenges of relating the complex, multi-dimensional (e.g., temporal) molecular processing that occurs in organs and tissues during early-stage cancer to observable clinical phenotypes in human patients. In particular, differential, temporal, and spatial gene expression during early cancer result from disruptions in the complex, dynamic, and multi-gene process that tightly regulates and controls the developmental integrity of organs and tissues. These temporal and spatial gene expression dynamics are fundamentally controlled by a variety of molecules called gene regulatory factors, such as transcription factors (TFs) and non-coding RNAs. These factors cooperate in a gene regulatory network (GRN) to carry out correct developmental functions on a genome scale (Iyer, Osmanbeyoglu and Leslie 2017). The nodes of a GRN are genes, and the edges of a GRN connect regulatory factors to their target genes. Disruption of the cooperation between genes and regulatory factors in a GRN can give rise to abnormal gene expression, such as that which is present in diseases such as cancer. Therefore, a fundamental challenge for uncovering early cancer mechanisms is that of understanding the gene regulatory mechanisms, especially GRNs, controlling the changes in gene expression that occur across cancer stages.

The collection of next-generation sequencing (NGS) data from large cohorts such as TCGA (Liu et al. 2018) provides measurements across multiple molecular levels, including transcriptomics and

epigenomics. This allows for studies of temporal dynamics in gene expression and regulation during cancer development and also allows for the systematic identification of stage-specific cancer biomarkers. Progress has been made in identification of some stage-specific molecular biomarkers of lung cancer, but systematic genome-wide analyses for identification of all potential early-stage biomarkers with predictive value for disease outcome are limited. For example, dysregulations in the epidermal growth factor receptor (EGFR, associated with sensitivity of lung cancers to the tyrosine kinase inhibitor gefitinib (Iressa) (Pao et al. 2004), echinoderm microtubule-associated protein-like 4 (EML4), and anaplastic lymphoma kinase (ALK) are frequently involved in oncogenic transformation (Lindeman et al. 2013). Additionally, v-raf murine sarcoma viral oncogene homologue B1 (BRAF) is a driver mutation gene in lung adenocarcinoma (Paik et al. 2011).It remains challenging to integrate the large-scale data provided by human tumors to enable the discovery of genome-wide early cancer biomarkers and underlying GRNs. Although a challenging task, finding novel ways to integrate the large-scale data provided by human tumors would enable the discovery of genome-wide early cancer biomarkers and underlying GRNs.

Traditionally, correlation-based models have been used to select biomarker genes involved in cancer development; for example, 62 genes were uncovered in this way to distinguish between the early and late stages of clear cell renal cell carcinoma (ccRCC); (Rahimi and Gönen 2018) (Jagga and Gupta 2014). However, correlation-based models only reveal linear relationships, whereas cancer development is a complex, nonlinear process. Thus, machine learning has emerged as a powerful tool to predict biomarkers for various cancer features related to clinical presentation and staging; this tool has been found to be of great help in the diagnosis and treatment of various diseases(Libbrecht and Noble 2015) . For example,(Statnikov, Wang and Aliferis 2008) applied random forests (RFs) and support vector machines (SVMs) to microarray data in order to aid in cancer diagnosis. (Xiao et al. 2018) constructed a multi-model ensemble approach to predict cancer in both normal conditions and tumor conditions. However, none of these studies revealed novel cancer mechanistic insights; these studies were limited to building an accurate classification model as a “black box” but lacked any biological or clinical interpretability from the box.

To address these challenges, we designed a novel, interpretable machine learning approach, ECMarker, that can be used to discover gene expression biomarkers in early disease stages such as early cancer. In particular, ECMarker embeds GRN structures into a hierarchical neural network model in order to classify early and late cancer patients according to their gene expression data. The prioritized genes and gene-gene interactions for early/late cancer stages revealed potential cancer stage-specific gene biomarkers and GRNs. Furthermore, we found drugs that are potentially associated with ECMarker’s early cancer biomarkers with the goal of uncovering novel genomic medicine for early cancer development.

2 Materials and Methods

2.1 ECMarker, an interpretable machine learning framework to identify gene expression biomarkers and GRNs for cancer development and outcomes

We built an integrated, supervised machine learning framework, ECMarker, to identify the gene expression biomarkers for classifying cancer development stages and predicting clinical outcomes (Fig. 1). ECMarker was built based on semi-restricted Boltzmann machine (SRBM)(Osindero and Hinton 2007b) and including the following multi-step analysis: Step 1. After feeding the gene expression data and various clinically relevant phenotypes of cancer patients (e.g., stages and survival) into the model, multiple feature selection methods were used to select a group of genes that

potentially represent cancer development. Step 2. Based on these genes, ECMarker identified the biomarker genes that classify the patients into stages of cancer, especially the early stages, and learned the connections between these genes that comprise imputed gene regulatory network. Step 3. A survival analysis stratified the early cancer patients into the groups with different survival rates using the biomarker genes.

In addition to predicting cancer stages, ECMarker is also biologically and clinically interpretable First, the gene set enrichment analysis (GSEA; (Subramanian et al. 2005) was used with the potential biomarkers to reveal the underlying mechanisms of the biological processes and pathways of cancer development. In particular, various biological functions and pathways are impacted by the cancer gene expression changes that occur across stages, which are fundamentally controlled by GRNs. Thus, we used the structures of ECMarker to extract the GRNs that control the biomarker genes along with enriched functions and pathways. These biological explanations could provide a molecular, mechanistic understanding of cancer development. Furthermore, we predicted that certain drugs may significantly affect the biomarker genes or their regulatory factors, especially for early cancer stages, thus revealing potential stage-specific, genomic medicine targets and therapies.

2.2 The ECMarker classifier was built on a semi-restricted Boltzmann machine (SRBM) for biological interpretation

The ECMarker classifier was built on SRBM, which is a supervised learning probabilistic discriminative model. (Fig 1) Specifically, SRBM is an undirected graphical model that uses hidden binary units to model the distribution and connections of variables of a visible layer. Given the expression profiles of selected genes, $V_{gene} = \{v_{gene_i}\}_{i=1}^{m}$ as visible layer nodes. In *m* input gene nodes, each gene $v_{gene_i} \in \mathbb{R}^d$, where *d* indicates the number of patients. We also involved a weights matrix $L_{gene} \in \mathbb{R}^{m\times m}$, which describes a pairwise gene-to-gene connection. The hidden layer consists of the binary variables $\mathbf{h} \in \mathbb{R}^n$ where n is number of hidden nodes; To connect the visible layer genes and the hidden layer pathways unit, $W_{hidden} \in \mathbb{R}^{m\times \mathrm{n}}$,each element $w_{h_j,gene_i}$ in this gene-to-hidden connection matrix. And the output layer consists of cancer stage $Y_{stage} = \{y_{stage_i}\}_{i=1}^{2}$. We represent y_{stage_i} as a binary vector, such that if the class is i , another element is zero except element i which takes the value 1. Further, the $W_{stage} \in \mathbb{R}^{2\times n}$, w_{h_i,s_j} is used as the weight matrix linking hidden node i and stage node j. $\theta = (L_{gene}, W_{hidden}, W_{stage}, a, b, c)$ are the model parameters. a, b, c are the biases of gene and hidden layer activation potentials.

Then the energy function of an ECMarker is given by

$$E(V_{gene}, \mathrm{h}, Y_{stage}|\theta) = -V_{gene}{}^T L_{gene} V_{gene} - V_{gene}{}^T W_{hidden}\mathbf{h} - \boldsymbol{h}^T W_{stage} Y_{stage} - a^T V_{gene} - \boldsymbol{b}^T \boldsymbol{h} - \boldsymbol{c}^T Y_{stage} \quad (1)$$

From the energy function, we assigned the probabilities to values of $V_{gene}, \mathbf{h}, Y_{stage}$ with the parameters θ as follows,

$$p(V_{gene}, \mathbf{h}, Y_{stage}|\theta) = \frac{\exp(-E(V_{gene}, \mathbf{h}, Y_{stage}|\theta))}{Z} \quad (2)$$

and

$$Z = \sum_{V_{gene}, \mathrm{h}, Y_{stage}} e^{-E(V_{gene}, \mathbf{h}, Y_{stage}|\theta)} \quad (3)$$

where Z is a partition function.

Further, we calculated the required conditional distributions.

$$p\left(V_{gene}|\mathbf{h}\right) = \prod_i p(v_{gene_i}|\mathbf{h}) \tag{4}$$

Once binary states have been set for the hidden units, a reconstruction is produced by setting v_{gene_i} to 1 with probability

$$p\left(v_{gene_i} = 1|\mathbf{h}\right) = \sigma(a_i + \sum_j w_{h_j,gene_i} h_j\ ) \tag{5}$$

where σ(x) is the logistic function $\frac{1}{1+\exp(-x)}$.

And

$$p\left(Y_{stage}|\mathbf{h}\right) = \prod_i p\left(y_{stage_i} = |\mathbf{h}\right) \tag{6}$$

Once binary states have been set for the hidden units, the possibility of y_{stage_i} belongs to group one is

$$p\left(y_{stage_i} = 1|\mathbf{h}\right) = \sigma(c_i + \sum_j w_{h_j,stage_i} h_j\ ) \tag{7}$$

$$p\left(\mathbf{h}|V_{gene}, Y_{stage}\right) = \prod_i p\left(h_i|V_{gene}, Y_{stage}\right) \tag{8}$$

Given V_{gene}, Y_{stage} , the binary state $h_{pathway}$ is set to 1 with probability

$$\prod_i p\left(h_i = 1|V_{gene}, Y_{stage}\right) = \sigma\left(b_i + \sum_j v_{gene_j} w_{h_i,gene_j} + \sum_j y_j w_{h_i,stage_j}\right) \tag{9}$$

We used the contrastive divergence(Hinton 2002, Hinton 2010) to learned the parameter θ. The update rules (Larochelle et al. 2012, Osindero and Hinton 2007a) are

$$\Delta L_{gene} = \epsilon(\langle V_{gene}{}^{\mathrm{T}} V_{gene}\rangle_{\mathrm{data}} - \langle V_{gene}{}^{\mathrm{T}} V_{gene}\rangle_{\mathrm{recon}}) \tag{10}$$

$$\Delta W_{hidden} = \epsilon(\langle V_{gene}{}^{\mathrm{T}} \mathrm{h}\rangle_{\mathrm{data}} - \langle V_{gene}{}^{\mathrm{T}} h\rangle_{\mathrm{recon}}) \tag{11}$$

$$\Delta W_{stage} = \epsilon(\langle Y_{stage}{}^{\mathrm{T}} \mathrm{h}\rangle_{\mathrm{data}} - \langle Y_{stage}{}^{\mathrm{T}} \mathrm{h}\rangle_{\mathrm{recon}}) \tag{12}$$

where ϵ is a learning rate. The angle brackets denote the expected value of the distribution calculated by V_{gene}, Y_{stage} and the hidden units distributed according to the conditional distribution $p\left(\mathbf{h}|V_{gene}, Y_{stage}\right), p\left(V_{gene}|\mathbf{h}\right)$, $p\left(Y_{stage}|\mathbf{h}\right)$. The subscript data indicates the visible gene input, stage units and the hidden units are being driven by data, and subscript recon is corresponding fraction for reconstructions. A pseudocode of the model parameter update procedure is given by the following algorithm in Table 1.

Table 1. Pseudocode for ECMarker parameter update

Algorithm 1: Update ECMarker using Contrastive Divergence

Initialization;
$W_{hidden} \in \mathbb{R}^{m \times n}, W_{stage} \in \mathbb{R}^{2 \times n}$ are initialized with values from Gaussian distribution;
$L_{gene} \in \mathbb{R}^{m \times m}$ is initialized as an identity matrix;
$\mathbf{a} \in \mathbb{R}^{m}, \mathbf{b} \in \mathbb{R}^{n}, \mathbf{c} \in \mathbb{R}^{2}$ are set to zero.

Params:
m: number of binary gene nodes in visible layer;
n: number of hidden nodes.
ϵ is learning rate.

for *i in 1*: *n.iter* **do**
  # Initial state drived by data
  $y_0 \leftarrow y_{stage_i}, v_0 \leftarrow v_{gene_i}, h_0 \leftarrow binarize(\sigma(\mathbf{b} + W_{hidden} * v_0 + W_{stage} * y_0))$

  #Reconstruction
  $\hat{h_0} \sim p(\mathbf{h}|v^0, y^0), y_1 \sim p(y|\hat{h_0}), v_1 \sim p(x|\hat{h_0})$
  $h_1 \leftarrow \sigma(\mathbf{b} + W_{hidden} * v_1 + W_{stage} * y_1)$
  $v_2 \leftarrow \sigma(\mathbf{a} + W_{hidden} * h_1 + L_{gene} * x_1)$

  # Update
  $\triangle W_{hidden} = \epsilon(v_0 * h_0 - v_1 * h_1)$
  $\triangle W_{stage} = \epsilon(y_0 * h_0 - y_1 * h_1)$
  $\triangle L_{gene} = \epsilon(v_0 * v_0 - v_2 * v_2)$
end

After training the model, the predicted matrix can be used as adjancey matrix to infer the gene regulatory network among input genes.

2.3 Ranking method based on the ECMarker classifier to prioritize cancer stage-specific biomarkers

The structure of the ECMarker classifier consists of nodes and weighted edges that link these nodes (Fig. 1). The weight values of the edges learned by ECMarker represent the strengths of the interconnections between nodes. To identify early-stage biomarkers, we ranked all the absolute values of the weighted edges, linking each early-stage node to each hidden node. We selected the ten hidden nodes with the highest absolute weight values as the early stage-associated hidden nodes. For each gene, it connected to these ten “early” hidden nodes with weighted edges. Each gene connected to these ten “early” hidden nodes with weighted edges. We selected the highest absolute weight value as each gene’s association score with the “early” hidden nodes. Then, we ranked all of the genes by their association scores with the “early” hidden nodes and picked the top-ranking genes as early-stage biomarkers. Late-stage biomarkers were generated similarly. Firstly, we chose the top ten late-stage associated hidden nodes by ranking the absolute weight value between the late stage node and all hidden nodes. Then based on strength of interactions between the ten “late” hidden nodes and each gene, we selected the highest absolute weight value as this gene’s association with late-stage hidden nodes. The genes were ranked according to highest associated with value and the top ranking genes are marked as the late-stage biomarkers.

2.4 Cancer gene expression and clinical datasets for building and testing ECMarker

We built and tested the ECMarker with two publicly accessible datasets. The Gentles2015 dataset (Gentles et al. 2015) consists of standardized gene expression data collected from 1103 non-small cell lung cancer (NSCLC) patients who had not received pre-biopsy treatment. Then, we grouped patients based on their TMN stages, with (I+IA+IB) as the early stage (N = 766) and II, III, and IV as the late stages (N = 337) and divided the dataset into training and testing datasets. Lung adenocarcinoma (LUAD) is one of the most common subtypes of NSCLC (Herbst et al. 2018) .To demonstrate the utility of ECMarker in distinct subtypes of NSCLC, we downloaded LUAD datasets from TCGA(The Cancer Genome Atlas Research et al. 2013), resulting in independent validation dataset TCGA-LUAD. After log2 transformation, each gene in the TCGA-LUAD dataset was standardized so that its mean expression was zero and its standard deviation was 1, as the same way

in Gentles2015. Furthermore, owing to the effects of some therapeutics on certain genes' expression levels(Zitvogel et al. 2008), we focused on treatment-naïve tumors by excluding patients who received pharmaceutical (chemotherapy, targeted molecular therapy, immunotherapy), radiation, or surgical treatments, according to the TCGA-LUAD treatment annotation (Campbell et al. 2016) (The Cancer Genome Atlas Research et al. 2014), to avoid drug-induced gene expression profiles. Overall, we included 436 TCGA-LUAD patients as the independent validation dataset (N = 307 and N = 129 in the early and late stages, respectively). A summary on patient numbers of various stages is in S3 Table.

2.5 Feature extraction for ECMarker

The gene expression datasets were very high-dimensional and heterogeneous, especially at the tissue level (e.g., a wide variety of genes vs. limited sample size). In order to increase the interpretability and reduce the complexity and the training time of our model, we integrated extensive feature selection as an essential part before tuning our model. For this purpose, we tested multiple feature selection methods to reduce the dimensionality of gene expression data and input features for classification, including filter feature selection methods Pearson correlation, embedding methods LASSO, ridge regression, elastic net, wrapper methods, and Boruta (Kursa and Rudnicki 2010) algorithms using R functions (Friedman, Hastie and Tibshirani 2010).

2.6 State-of-the-art classification methods

After feature selection, we compared the classification performance of ECMarker with several state-of-the-art classifiers, including an artificial neural network (ANN), random forests (RF), AdaBoost-based ensemble learning (ES), a support vector machine (SVM), naive Bayes (NB), and logistic regression (LR) toward discrimination between the early and late stages. These state-of-the-art classifiers could be implemented by the Scikit-Learn library in Python (Pedregosa et al. 2011).

2.7 Preprocessing for training the ECMarker classification model

First, we oversampled to balance the unbalanced dataset between early and late stages (69% early-stage patients vs. 31% late-stage patients). In particular, we used the oversampling method to replicate the sample from the minority class (late-stage patients) to balance the data. After oversampling, early and late-stage patients (1532 patients in total) had similar distributions in multi-clinical features such as gender, stage, age. The stages of the individual cancer patients were used as output labels in classification. We implemented oversampling using R package ROSE (Lunardon, Menardi and Torelli 2014). Second, we split the dataset into a training set (75%) and a test set (25%), while keeping some important features such as age, gender, and survival information in a similar distribution between them (S1 Table). Finally, we binarized the gene expression data. The ECMarker model, based on the SRBM model, could only handle binary inputs. Therefore, we used the k-means clustering method in the BoolNet (Müssel, Hopfensitz and Kestler 2010) R package to binarize the real-value gene expression input data. For each gene, k-means clusterings were performed to split the expression data into two groups. After the centroid values were compared, the patients belonging to the group with the smaller centroid were set to 0, and the patients belonging to the greater centroid were set to 1.

2.8 Identification of drugs targeting biomarker genes predicted by ECMarker

We applied the gene-drug association model and analysis from PharmacoGx (Smirnov et al. 2016) , in which the association between the biomarker gene and drug sensitivity was modeled using linear

regression. Because drug sensitivity is tissue-specific (Garnett et al. 2012) and batch variation exists, we adjusted the linear model of the drug-gene association for the tissue types and experimental batch.

$$D_{IC50_j} = \beta_0 + \beta_i G_i + \beta_t T_{Lung} + \beta_b B \quad (4)$$

where D_{IC50_j} is the IC50 of drug j, G_i is biomarker gene i, T_{Lung} indicates the lung cancer tumor source, and B denotes the experimental batch. β_i was used to estimate the drug-gene association. D_{IC50} and G were scaled to access the relative importance of the different biomarker genes. Therefore, the drug with the highest absolute gene-drug association value (β_i) for each biomarker would be selected as the potential biomarker's drug.

2.9 Identification of pathways and functions associated with early and late cancer stages using ECMarker biomarker genes via enrichment analysis

To identify the pathways and functions related to early cancer, we also performed the enrichment analyses using ECMarker biomarker genes. In particular, given an ECMarker biomarker $gene_j$, we defined a $gene_j$ association score for each gene to be the Pearson correlation coefficient between the expression levels of this gene and $gene_j$ across individuals. Then, we ranked all genes based on their $gene_j$ association scores and identified a set of enriched functions and pathways from the ranked list using gene set enrichment analysis (GSEA; (Subramanian et al. 2005) as $gene_j$'s associated functions, and pathways. Then, we selected the top five ECMarker genes for the early cancer stage and identified their associated functions and pathways. The shared associated pathways and functions for at least three of these genes were defined as early cancer-associated pathways and functions. Similarly, we picked the top five ECMarker late-stage cancer-related genes and identified their associated pathways and functions. The significant overlapped functions and pathways for at least three of these genes were determined as late cancer-associated pathways and functions.

3 Results

3.1 Lung cancer gene expression biomarkers identified by the ECMarker model for classification of cancer stages

We applied ECMaker to the Gentles2015 dataset (see Methods) with the aim of predicting novel biomarker genes for lung cancer development and outcomes, especially for early-stage patients. ECMarker tested various feature selection methods such as LASSO, ridge regression, elastic net, and correlation and also prior biologically known gene sets related to cancer development as biological features, including epithelial-to-mesenchymal transition (EMT) signature genes (Byers et al. 2013), lung cancer mutation genes, and transcription factor genes (Slowikowski 2016) (S2 Table). Specifically, the best subset of potentially stage-correlated biomarkers can be identified with the highest classification accuracy. Using the top 200 cancer stage-related genes, selected as input by the ridge regression, as input, we found that ECMarker achieves the highest classification accuracy. Specifically, ECMarker had a classification accuracy of 0.835 (S2 Table) for the testing dataset. Then, we identified early and late biomarkers using the trained ECMarker model (Methods and Supplementary File 1).

In order to gain biological insights into these stage-specific biomarkers at the functional level, we identified the enriched pathways and functions of the ECMarker biomarker genes associated with early and late cancer stages using GSEA MSigDB datasets (see Methods and Supplementary File 2). Figure 3 compares these enriched functions and pathways of the early and late stages. The early stage

biomarkers are enriched in some lung cancer related pathways (for example, *lung cancer poor survival* or *lung cancer KRAS)*, as well as pan-cancer functions/pathways and immunological pathways, which might reflect general early warning signs of cancer. The late stage biomarkers were enriched in more pathways and functions related with general tumour development, such as immunologic ones, or related to soft tissue tumours, myeloid cell development pathways associated with solid tumour development, and pathways related to primary immunodeficiency, differentiating T lymphocyte which might be linked to the cancer development process. These pathways and functions could be used to understand the molecular mechanisms of cancer progression at different stages. Also, the identification of these pathways and functions confirmed that ECMarker not only performs well in clinical outcome prediction but also provides mechanistic biological insights.

3.2 ECMarker biomarker genes predict clinical outcomes for early lung cancer

In addition to the biological functions associated with cancer stages, we also explored the relationships between gene biomarkers and clinical outcomes of cancer patients. In particular, using the hierarchical structure of ECMarker, we were able to identify the connections within the input gene set and reassign the cancer stages to each individual patient (see Methods). Thus, using the 25 early-stage biomarkers learned by ECMarker, we performed a survival analysis for early-stage patient groups clustered by predicted stages. Importantly, we found that the patient groups had significantly distinct survival rates, especially those at early cancer stages (i.e., ECMarker predicts early cancer survival rates). As shown in Figure 4, the patients in the testing set that were reassigned to early stages had significantly different survival rates compared their survival rates in their original pathological stages. Moreover, we also found that more patients connected to the early stages in comparison with the pathological classification. This implies that the gene expression biomarkers identified by ECMarker were not only related to cancer stages but could also be used to predict clinical outcomes such as survival rates, particularly for early cancer patients. This suggests that ECMarker is clinically interpretable.

3.3 Hierarchy in the ECMarker reveals cancer functional genomics and regulatory networks across stages

Similar to other models, ECMarker can predict phenotypes from molecular data (e.g., gene expression to cancer stages), but understanding the biological mechanisms based on prediction model is still a challenging task. To this end, we endeavored to make ECMarker biologically interpretable. In particular, we leveraged ECMarker to reveal the biological functions and pathways related to each cancer stage and provide mechanistic insights into the functional changes involved in cancer development. Then, we identified the enriched functions and pathways for stage-specific genes (see Fig. 3). Thus, we were able to show that ECMarker is biologically interpretable and able to reveal the functional modules connecting gene expression to phenotypes.

Furthermore, we pursued the elucidation of the molecular mechanisms that drive the functional connectivities, especially in terms of gene regulation. Gene expression, though complex for phenotypes, is controlled by gene regulatory mechanisms. In particular, various regulatory factors, such as transcription factors, control the expression of biomarker genes to coordinate cancer phenotypes so they do not behave randomly (i.e., forming a GRN). ECMarker has the capability to model gene interactions so that we could extract the weight value to describe the relationship of any pair of genes among the input genes. A larger absolute weight value indicated a stronger connection between genes. According to this standard, we could extract the learned GRN from any well-trained ECMarker classifier. Specifically, we identified stage-specific biomarkers and extracted the GRNs for these biomarkers to explore the changes from early to late cancer stages. Figure 2 shows the top

20 early and late-stage biomarkers. To uncover the hierarchy of transcription factors driving cancer stage development, we generated GRN representations connecting all the biomarkers, with separate representations for early and late stages of development. The highlighted TF-target connections indicated how interactions between biomarkers drive developmental progression.

Furthermore, we validated the regulatory linkages learned by ECMarker with previous predicted lung cancer regulatory networks (Mitra et al. 2014). (Mitra et al. 2014) predicted highly confident regulatory relationships among TF regulation to target genes. Therefore, we inputted 123 genes that were listed in this paper (22 TF and 101 target genes) into ECMarker and compared the GRNs learned by ECMaker with their predicted results. ECMarker was able to learn 2222 (22 TF*101 target) pairs. We extracted 173 overlapped connected pairs matched with the predicted connected pair results from their study as overlapped TF-gene pair group (Group I), and other predicted connected pairs out of these 173 pairs as not overlapped TF-gene pair group (Group II). Then, we compared the absolute weight values between these two groups. We found that the absolute weight values of Group I were significantly higher than those of Group II ($p<0.049$), which suggests the consistency between ECMarker and (Mitra et al. 2014). To further validate the connectivity results, we used the recent human TF-target database (Slowikowski 2016) to search for our predicted TF-target pairs out of those 173 connected pairs in paper (Mitra et al. 2014), and obtain a high-confident overlapped TF-gene pair group (Group III). Then, we found that the absolute weights of Group III are even more significantly larger than Group I ($p<0.0065$, S1. Fig). Therefore, the results of the present study suggested that the regulatory linkages found by ECMarker were highly consistent with previous predictions regarding GRNs involved in lung cancer. Moreover, the present results also suggested that ECMarker is biologically interpretable for revealing gene regulatory networks in the lung cancer.

3.4 Potential novel drugs for early cancer stages

The gene expression biomarkers and GRNs identified by ECMarker may provide target candidates for cancer medicine. Thus, we associated a variety of drugs with our stage-specific biomarker genes to find drugs with the potential to target specific stages of lung adenocarcinoma, which would enable different approaches to precision medicine. To produce an at-large list of potential drugs, we also included the genes co-expressed with our biomarker genes that were likely co-regulated by similar mechanisms (Methods). Thus, for each stage in NSCLC, we proceeded to identify a list of drugs that significantly correlate with the stage biomarker and co-expressed genes (e.g., IC50 value; see Supplementary File 3). For this purpose, we applied the gene-drug association model and analysis PharmacoGx(Smirnov et al. 2016) to large pharmacogenomic datasets such as the Genomics of Drug Sensitivity in Cancer database (GDSC database)(Yang et al. 2013).

The GDSC database provides gene-drug sensitivity and response data for a variety of cancer types (i.e., cell lines). Underlying the power of our approach (see Methods), several known drugs for the treatment of late lung cancer stages were also predicted to have significant activity on our late-stage biomarkers; e.g., Sorafenib(Scagliotti et al. 2010) and BI-2536, a novel Plk-1 Inhibitor which could be used to treat stage IIIB/IV NSCLC patients undergoing surgery BI-2536 (Sebastian et al. 2010). In addition, Gefitinib, which has been used in a clinical trial to treat patients with EGFR mutations (Maemondo et al. 2010) could also improve progression-free survival in certain patients at advanced lung cancer stages. Interestingly, we also found that a few drugs in clinical use for late stages significantly correlate with our early stage biomarkers; e.g., Lapatinib (Ross et al. 2010), PD.0325901(Haura et al. 2010), and Sorafenib. Therefore, these drugs could potentially have effects on early cancer stages. Figure 5 shows the different prediction p-values between stage-specific genes

for 30 prediction drugs. These results suggest that the gene expression biomarkers found by ECMarker are able to help identify potential drugs for early cancer treatment.

3.5 Comparison with state-of-the-art classification methods

A variety of state-of-the-art classification methods, such as ANN, SVM, RF, ES, LR, and NB (Methods), have been applied in cancer research for outcome detection and prediction. Even though these cancer applications have demonstrated effective and accurate performance, most of these methods lack biological and clinical interpretability, making it difficult to improve academic understanding of cancer progression using these methods. Therefore, we first compared our prediction performance with existing state-of-the-art methods; the results of this comparison are shown in S4 Table. Then, we selected some studies that made use of these methods and evaluated these works in the following aspects, computation performance, biological interpretability, and clinical interpretability demonstration in Table 2.

Table 2. Comparison of ECMarker and previous studies using state-of-the-art classification methods

Method	Biological application	Interpretability
RF, SVM	Undertook a methodologically comprehensive comparison of random forests and support vector machines for microarray-based cancer classification(Statnikov et al. 2008).	Used a large number of datasets to compare two classifiers but failed to explore the biological insights of the best performance classifier after identifying this model.
KNN, SVM, RF, ANN	Demonstrated a new strategy that applied deep learning to an ensemble approach that incorporated different machine learning models for cancer prediction with gene data selected by differential gene expression analysis(Xiao et al. 2018).	Trained the model as a black box after reaching high accuracy, but did not further explore GRNs, functions, or pathways to better understand cancer mechanisms.
ES	Performed a comparison of single supervised machine learning and ensemble methods in classifying cancerous gene expression data and presented some theoretical explanations for the performance of ensemble methods (Tan and Gilbert 2003).	Improved understanding of the correlations between gene expression profiles and cancer developmental stages but did not focus on the far more important aspects of theoretical, biological, and clinical explanations.
LR	Proposed an improved sparse logistic regression method to identify discriminatory biomarker genes based on gene expression data in cancer classification (Cawley and Talbot 2006).	Sufficiently evaluated the computational performance of various methods in multi-cancer datasets but lacked biologically or clinically interpretable analysis of the selected biomarker genes.

ANN	Trained ANNs using the gene expression signatures of small, round blue-cell tumors (SRBCTs) to classify cancers into specific diagnostic categories (Statnikov et al. 2005).	After training the high-accuracy ANN model, the authors extracted 96 genes that were highly relevant to cancer. Although the authors discussed the identification of genes specific to cancer with implications for biology and therapy, they did not conduct a GRN analysis or detection of gene-gene interactions.

4 Discussion

In this paper, we developed a machine learning approach as well as a computational framework, ECMarker, to identify gene expression biomarkers for various stages of cancer development. ECMarker is both biologically and clinically interpretable for the discovery of gene expression biomarkers. In particular, ECMarker simultaneously predicts the underlying GRNs for these biomarkers and thus possesses biological interpretability, which state-of-the-art machine learning models typically lack. We applied ECMarker to individual gene expression data for lung cancer, the leading cause of cancer-related deaths, and identified a set of novel biomarker genes classifying cancer stages. We found that these biomarkers significantly predict patients' survival rates, especially at the early stage, and can also be linked to potential novel drugs; this reveals the clinical interpretability of our approach. The predicted biomarker genes comprise novel molecular candidates for early cancer diagnosis and detection, and the GRNs could potentially guide future experimental validations for early cancer mechanisms and treatments. Although this study applied ECMarker lung cancer data specifically, ECMarker is a general-purpose method and can therefore be applied for other disease types including neurodevelopmental disorders(Li et al. 2018).

This study demonstrated that we are able to build machine learning models that are biologically interpretable. This was our first round of attempts to address the lack of interpretability and translation of machine learning applications in biology and biomedicine. Given that cancer phenotypes are driven by a variety of multi-omic mechanisms, including transcriptomics, epigenomics, metabolomics, etc., we expect to develop advanced machine learning approaches that can reveal interactions across multi-omics relating to phenotypes in the near future. In particular, by integrating genotyping data, we aim to build a deeper hierarchical model, similar to deep neural network models(Wang et al. 2018), to predict genotype-phenotype relationships and use intermediate connectivities and structures inside the model to reveal all possible molecular mechanisms from genotype to phenotype.

The present study focused on gene expression data at the individual tissue level. However, the cancer tissues consist of different cell types with various fitness and mutational profiles(Saadatpour et al. 2015). The continuous development of single-cell genomic and transcriptomic analyses for cancer research will enable us to explore how single cells contribute to cancer tissue expression and eventually affect phenotypes (e.g., single-cell deconvolution to estimate various cell fractions). Integrating single-cell data into machine learning modelling and drug association analysis might uncover novel biological mechanisms and targetable key regulators for the advancement of precision cancer medicine.

5 Author Contributions

D.W. conceived and designed the project. T.J. and D.W. performed data processing and analyses. T.J., F.T. and D.W. wrote the manuscript. All authors read and approved the final manuscript.

6 Reference

Byers, L. A., L. Diao, J. Wang, P. Saintigny, L. Girard, M. Peyton, L. Shen, Y. Fan, U. Giri, P. K. Tumula, M. B. Nilsson, J. Gudikote, H. Tran, R. J. Cardnell, D. J. Bearss, S. L. Warner, J. M. Foulks, S. B. Kanner, V. Gandhi, N. Krett, S. T. Rosen, E. S. Kim, R. S. Herbst, G. R. Blumenschein, J. J. Lee, S. M. Lippman, K. K. Ang, G. B. Mills, W. K. Hong, J. N. Weinstein, Wistuba, II, K. R. Coombes, J. D. Minna & J. V. Heymach (2013) An epithelial-mesenchymal transition gene signature predicts resistance to EGFR and PI3K inhibitors and identifies Axl as a therapeutic target for overcoming EGFR inhibitor resistance. *Clin Cancer Res,* 19**,** 279-90.

Campbell, J. D., A. Alexandrov, J. Kim, J. Wala, A. H. Berger, C. S. Pedamallu, S. A. Shukla, G. Guo, A. N. Brooks, B. A. Murray, M. Imielinski, X. Hu, S. Ling, R. Akbani, M. Rosenberg, C. Cibulskis, A. Ramachandran, E. A. Collisson, D. J. Kwiatkowski, M. S. Lawrence, J. N. Weinstein, R. G. W. Verhaak, C. J. Wu, P. S. Hammerman, A. D. Cherniack, G. Getz, N. Cancer Genome Atlas Research, J. C. Zenklusen, J. Zhang, I. Felau, J. A. Demchok, L. Yang, Z. Wang, M. L. Ferguson, R. Tarnuzzer, C. M. Hutter, H. J. Sofia, T. Pihl, Y. Wan, S. Chudamani, J. Liu, C. Sun, R. Naresh, L. Lolla, Y. Wu, C. J. Creighton, W. K. Rathmell, J. T. Auman, S. Balu, T. Bodenheimer, D. N. Hayes, K. A. Hoadley, A. P. Hoyle, C. D. Jones, S. R. Jefferys, S. Meng, P. A. Mieczkowski, L. E. Mose, C. M. Perou, J. Roach, Y. Shi, J. V. Simons, T. Skelly, M. G. Soloway, D. Tan, J. Wu, U. Veluvolu, J. S. Parker, M. D. Wilkerson, L. Boice, M. Huang, L. B. Thorne, M. S. Noble, H. Zhang, D. I. Heiman, J. Cho, N. Gehlenborg, G. Saksena, D. Voet, P. Lin, S. Frazer, L. Chin, M.-S. Tsao, F. Allison, D. Chadwick, T. Muley, M. Meister, H. Dienemann, R. Kucherlapati, P. Park, J. Bowen, J. M. Gastier-Foster, M. Gerken, K. M. Leraas, T. M. Lichtenberg, N. C. Ramirez, L. Wise, E. Zmuda, J. Stuart, E. Collisson, et al. (2016) Distinct patterns of somatic genome alterations in lung adenocarcinomas and squamous cell carcinomas. *Nature Genetics,* 48**,** 607.

Cawley, G. C. & N. L. C. Talbot (2006) Gene selection in cancer classification using sparse logistic regression with Bayesian regularization. *Bioinformatics,* 22**,** 2348-2355.

Friedman, J. H., T. Hastie & R. Tibshirani (2010) Regularization Paths for Generalized Linear Models via Coordinate Descent. *2010,* 33**,** 22.

Frost, J. K., W. C. Ball, Jr., M. L. Levin, M. S. Tockman, R. R. Baker, D. Carter, J. C. Eggleston, Y. S. Erozan, P. K. Gupta, N. F. Khouri & et al. (1984) Early lung cancer detection: results of the initial (prevalence) radiologic and cytologic screening in the Johns Hopkins study. *Am Rev Respir Dis,* 130**,** 549-54.

Garnett, M. J., E. J. Edelman, S. J. Heidorn, C. D. Greenman, A. Dastur, K. W. Lau, P. Greninger, I. R. Thompson, X. Luo, J. Soares, Q. Liu, F. Iorio, D. Surdez, L. Chen, R. J. Milano, G. R. Bignell, A. T. Tam, H. Davies, J. A. Stevenson, S. Barthorpe, S. R. Lutz, F. Kogera, K. Lawrence, A. McLaren-Douglas, X. Mitropoulos, T. Mironenko, H. Thi, L. Richardson, W. Zhou, F. Jewitt, T. Zhang, P. O'Brien, J. L. Boisvert, S. Price, W. Hur, W. Yang, X. Deng, A. Butler, H. G. Choi, J. W. Chang, J. Baselga, I. Stamenkovic, J. A. Engelman, S. V. Sharma, O. Delattre, J. Saez-Rodriguez, N. S. Gray, J. Settleman, P. A. Futreal, D. A. Haber, M. R. Stratton, S. Ramaswamy, U. McDermott & C. H. Benes (2012) Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature,* 483**,** 570-575.

Gentles, A. J., S. V. Bratman, L. J. Lee, J. P. Harris, W. Feng, R. V. Nair, D. B. Shultz, V. S. Nair, C. D. Hoang, R. B. West, S. K. Plevritis, A. A. Alizadeh & M. Diehn (2015) Integrating Tumor and Stromal Gene Expression Signatures With Clinical Indices for Survival Stratification of Early-Stage Non–Small Cell Lung Cancer. *JNCI: Journal of the National Cancer Institute,* 107.

Haura, E. B., A. D. Ricart, T. G. Larson, P. J. Stella, L. Bazhenova, V. A. Miller, R. B. Cohen, P. D. Eisenberg, P. Selaru, K. D. Wilner & S. M. Gadgeel (2010) A phase II study of PD-0325901, an oral MEK inhibitor, in previously treated patients with advanced non-small cell lung cancer. *Clin Cancer Res,* 16**,** 2450-7.

Herbst, R. S., D. Morgensztern & C. Boshoff (2018) The biology and management of non-small cell lung cancer. *Nature,* 553**,** 446.

Hinton, G. 2010. *A Practical Guide to Training Restricted Boltzmann Machines (Version 1).*

Hinton, G. E. (2002) Training products of experts by minimizing contrastive divergence. *Neural Comput,* 14**,** 1771-800.

Hu, Z., J. Chen, T. Tian, X. Zhou, H. Gu, L. Xu, Y. Zeng, R. Miao, G. Jin, H. Ma, Y. Chen & H. Shen (2008) Genetic variants of miRNA sequences and non-small cell lung cancer survival. *J Clin Invest,* 118**,** 2600-8.

Iyer, A. S., H. U. Osmanbeyoglu & C. S. Leslie (2017) Computational methods to dissect gene regulatory networks in cancer. *Current Opinion in Systems Biology,* 2**,** 115-122.

Jagga, Z. & D. Gupta (2014) Classification models for clear cell renal carcinoma stage progression, based on tumor RNAseq expression trained supervised machine learning algorithms. *BMC proceedings,* 8**,** S2-S2.

Koeffler, H. P., F. McCormick & C. Denny (1991) Molecular mechanisms of cancer. *The Western journal of medicine,* 155**,** 505-514.

Kursa, M. B. & W. R. Rudnicki (2010) Feature Selection with the Boruta Package. *2010,* 36**,** 13.

Larochelle, H., M. Mandel, R. Pascanu & Y. Bengio (2012) Learning algorithms for the classification restricted Boltzmann machine. *J. Mach. Learn. Res.,* 13**,** 643-669.

Li, M., G. Santpere, Y. Imamura Kawasawa, O. V. Evgrafov, F. O. Gulden, S. Pochareddy, S. M. Sunkin, Z. Li, Y. Shin, Y. Zhu, A. M. M. Sousa, D. M. Werling, R. R. Kitchen, H. J. Kang, M. Pletikos, J. Choi, S. Muchnik, X. Xu, D. Wang, B. Lorente-Galdos, S. Liu, P. Giusti-Rodríguez, H. Won, C. A. de Leeuw, A. F. Pardiñas, M. Hu, F. Jin, Y. Li, M. J. Owen, M. C. O'Donovan, J. T. R. Walters, D. Posthuma, M. A. Reimers, P. Levitt, D. R. Weinberger, T. M. Hyde, J. E. Kleinman, D. H. Geschwind, M. J. Hawrylycz, M. W. State, S. J. Sanders, P. F. Sullivan, M. B. Gerstein, E. S. Lein, J. A. Knowles & N. Sestan (2018) Integrative functional genomic analysis of human brain development and neuropsychiatric risks. *Science,* 362**,** eaat7615.

Libbrecht, M. W. & W. S. Noble (2015) Machine learning applications in genetics and genomics. *Nature Reviews Genetics,* 16**,** 321.

Lindeman, N. I., P. T. Cagle, M. B. Beasley, D. A. Chitale, S. Dacic, G. Giaccone, R. B. Jenkins, D. J. Kwiatkowski, J.-S. Saldivar, J. Squire, E. Thunnissen & M. Ladanyi (2013) Molecular testing guideline for selection of lung cancer patients for EGFR and ALK tyrosine kinase inhibitors: guideline from the College of American Pathologists, International Association for the Study of Lung Cancer, and Association for Molecular Pathology. *Journal of thoracic*

oncology : official publication of the International Association for the Study of Lung Cancer, 8**,** 823-859.

Liu, J., T. Lichtenberg, K. A. Hoadley, L. M. Poisson, A. J. Lazar, A. D. Cherniack, A. J. Kovatich, C. C. Benz, D. A. Levine, A. V. Lee, L. Omberg, D. M. Wolf, C. D. Shriver, V. Thorsson, S. J. Caesar-Johnson, J. A. Demchok, I. Felau, M. Kasapi, M. L. Ferguson, C. M. Hutter, H. J. Sofia, R. Tarnuzzer, Z. Wang, L. Yang, J. C. Zenklusen, J. Zhang, S. Chudamani, J. Liu, L. Lolla, R. Naresh, T. Pihl, Q. Sun, Y. Wan, Y. Wu, J. Cho, T. DeFreitas, S. Frazer, N. Gehlenborg, G. Getz, D. I. Heiman, J. Kim, M. S. Lawrence, P. Lin, S. Meier, M. S. Noble, G. Saksena, D. Voet, H. Zhang, B. Bernard, N. Chambwe, V. Dhankani, T. Knijnenburg, R. Kramer, K. Leinonen, Y. Liu, M. Miller, S. Reynolds, I. Shmulevich, W. Zhang, R. Akbani, B. M. Broom, A. M. Hegde, Z. Ju, R. S. Kanchi, A. Korkut, J. Li, H. Liang, S. Ling, W. Liu, Y. Lu, G. B. Mills, K.-S. Ng, A. Rao, M. Ryan, J. Wang, J. N. Weinstein, J. Zhang, A. Abeshouse, J. Armenia, D. Chakravarty, W. K. Chatila, I. de Bruijn, J. Gao, B. E. Gross, Z. J. Heins, R. Kundra, K. La, M. Ladanyi, A. Luna, M. G. Nissan, A. Ochoa, S. M. Phillips, E. Reznik, F. Sanchez-Vega, C. Sander, N. Schultz, R. Sheridan, S. O. Sumer, Y. Sun, B. S. Taylor, et al. (2018) An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics. *Cell,* 173**,** 400-416.e11.

Ludwig, J. A. & J. N. Weinstein (2005) Biomarkers in Cancer Staging, Prognosis and Treatment Selection. 5**,** 845-856.

Lunardon, N., G. Menardi & N. Torelli (2014) ROSE: a Package for Binary Imbalanced Learning. *R Journal,* 6**,** 79-89.

Maemondo, M., A. Inoue, K. Kobayashi, S. Sugawara, S. Oizumi, H. Isobe, A. Gemma, M. Harada, H. Yoshizawa, I. Kinoshita, Y. Fujita, S. Okinaga, H. Hirano, K. Yoshimori, T. Harada, T. Ogura, M. Ando, H. Miyazawa, T. Tanaka, Y. Saijo, K. Hagiwara, S. Morita & T. Nukiwa (2010) Gefitinib or chemotherapy for non-small-cell lung cancer with mutated EGFR. *N Engl J Med,* 362**,** 2380-8.

Mitra, R., M. D. Edmonds, J. Sun, M. Zhao, H. Yu, C. M. Eischen & Z. Zhao (2014) Reproducible combinatorial regulatory networks elucidate novel oncogenic microRNAs in non-small cell lung cancer. *Rna,* 20**,** 1356-68.

Molina, J. R., P. Yang, S. D. Cassivi, S. E. Schild & A. A. Adjei (2008) Non-Small Cell Lung Cancer: Epidemiology, Risk Factors, Treatment, and Survivorship. *Mayo Clinic Proceedings,* 83**,** 584-594.

Müssel, C., M. Hopfensitz & H. A. Kestler (2010) BoolNet—an R package for generation, reconstruction and analysis of Boolean networks. *Bioinformatics,* 26**,** 1378-1380.

Osindero, S. & G. Hinton. 2007a. *Modeling image patches with a directed hierarchy of Markov random fields.*

---. 2007b. Modeling image patches with a directed hierarchy of Markov random fields. In *Proceedings of the 20th International Conference on Neural Information Processing Systems,* 1121-1128. Vancouver, British Columbia, Canada: Curran Associates Inc.

Paik, P. K., M. E. Arcila, M. Fara, C. S. Sima, V. A. Miller, M. G. Kris, M. Ladanyi & G. J. Riely (2011) Clinical characteristics of patients with lung adenocarcinomas harboring BRAF mutations. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology,* 29**,** 2046-2051.

Pao, W., V. Miller, M. Zakowski, J. Doherty, K. Politi, I. Sarkaria, B. Singh, R. Heelan, V. Rusch, L. Fulton, E. Mardis, D. Kupfer, R. Wilson, M. Kris & H. Varmus (2004) EGF receptor gene mutations are common in lung cancers from "never smokers" and are associated with sensitivity of tumors to gefitinib and erlotinib. *Proceedings of the National Academy of Sciences of the United States of America,* 101**,** 13306-13311.

Pedregosa, F., Ga, #235, l. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, #201 & d. Duchesnay (2011) Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.,* 12**,** 2825-2830.

Rahimi, A. & M. Gönen (2018) Discriminating early- and late-stage cancers using multiple kernel learning on gene sets. *Bioinformatics,* 34**,** i412-i421.

Ross, H. J., G. R. Blumenschein, Jr., J. Aisner, N. Damjanov, A. Dowlati, J. Garst, J. R. Rigas, M. Smylie, H. Hassani, K. E. Allen, L. Leopold, T. Z. Zaks & F. A. Shepherd (2010) Randomized phase II multicenter trial of two schedules of lapatinib as first- or second-line monotherapy in patients with advanced or metastatic non-small cell lung cancer. *Clin Cancer Res,* 16**,** 1938-49.

Saadatpour, A., S. Lai, G. Guo & G. C. Yuan (2015) Single-Cell Analysis in Cancer Genomics. *Trends Genet,* 31**,** 576-586.

Scagliotti, G., S. Novello, J. von Pawel, M. Reck, J. R. Pereira, M. Thomas, J. E. Abrao Miziara, B. Balint, F. De Marinis, A. Keller, O. Aren, M. Csollak, I. Albert, C. H. Barrios, F. Grossi, M. Krzakowski, L. Cupit, F. Cihon, S. Dimatteo & N. Hanna (2010) Phase III study of carboplatin and paclitaxel alone or with sorafenib in advanced non-small-cell lung cancer. *J Clin Oncol,* 28**,** 1835-42.

Sebastian, M., M. Reck, C. F. Waller, C. Kortsik, N. Frickhofen, M. Schuler, H. Fritsch, B. Gaschler-Markefski, G. Hanft, G. Munzert & J. von Pawel (2010) The efficacy and safety of BI 2536, a novel Plk-1 inhibitor, in patients with stage IIIB/IV non-small cell lung cancer who had relapsed after, or failed, chemotherapy: results from an open-label, randomized phase II clinical trial. *J Thorac Oncol,* 5**,** 1060-7.

Siegel, R. L., K. D. Miller & A. Jemal (2018) Cancer statistics, 2018. *CA Cancer J Clin,* 68**,** 7-30.

Slowikowski, K. 2016. tftargets https://github.com/slowkow/tftargets.

Smirnov, P., Z. Safikhani, N. El-Hachem, D. Wang, A. She, C. Olsen, M. Freeman, H. Selby, D. M. Gendoo, P. Grossmann, A. H. Beck, H. J. Aerts, M. Lupien, A. Goldenberg & B. Haibe-Kains (2016) PharmacoGx: an R package for analysis of large pharmacogenomic datasets. *Bioinformatics,* 32**,** 1244-6.

Statnikov, A., C. F. Aliferis, I. Tsamardinos, D. Hardin & S. Levy (2005) A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. *Bioinformatics,* 21**,** 631-43.

Statnikov, A., L. Wang & C. F. Aliferis (2008) A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. *BMC Bioinformatics,* 9**,** 319.

Subramanian, A., P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander & J. P. Mesirov (2005) Gene set enrichment

analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences,* 102**,** 15545-15550.

Tan, A. C. & D. Gilbert (2003) Ensemble machine learning on gene expression data for cancer classification. *Appl Bioinformatics,* 2**,** S75-83.

The Cancer Genome Atlas Research, N., K. Chang, C. J. Creighton, C. Davis, L. Donehower, J. Drummond, D. Wheeler, A. Ally, M. Balasundaram, I. Birol, Y. S. N. Butterfield, A. Chu, E. Chuah, H.-J. E. Chun, N. Dhalla, R. Guin, M. Hirst, C. Hirst, R. A. Holt, S. J. M. Jones, D. Lee, H. I. Li, M. A. Marra, M. Mayo, R. A. Moore, A. J. Mungall, A. G. Robertson, J. E. Schein, P. Sipahimalani, A. Tam, N. Thiessen, R. J. Varhol, R. Beroukhim, A. S. Bhatt, A. N. Brooks, A. D. Cherniack, S. S. Freeman, S. B. Gabriel, E. Helman, J. Jung, M. Meyerson, A. I. Ojesina, C. S. Pedamallu, G. Saksena, S. E. Schumacher, B. Tabak, T. Zack, E. S. Lander, C. A. Bristow, A. Hadjipanayis, P. Haseley, R. Kucherlapati, S. Lee, E. Lee, L. J. Luquette, H. S. Mahadeshwar, A. Pantazi, M. Parfenov, P. J. Park, A. Protopopov, X. Ren, N. Santoso, J. Seidman, S. Seth, X. Song, J. Tang, R. Xi, A. W. Xu, L. Yang, D. Zeng, J. T. Auman, S. Balu, E. Buda, C. Fan, K. A. Hoadley, C. D. Jones, S. Meng, P. A. Mieczkowski, J. S. Parker, C. M. Perou, J. Roach, Y. Shi, G. O. Silva, D. Tan, U. Veluvolu, S. Waring, M. D. Wilkerson, J. Wu, W. Zhao, T. Bodenheimer, D. N. Hayes, A. P. Hoyle, S. R. Jeffreys, L. E. Mose, J. V. Simons, M. G. Soloway, S. B. Baylin, B. P. Berman, M. S. Bootwalla, L. Danilova, et al. (2013) The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genetics,* 45**,** 1113.

The Cancer Genome Atlas Research, N., E. A. Collisson, J. D. Campbell, A. N. Brooks, A. H. Berger, W. Lee, J. Chmielecki, D. G. Beer, L. Cope, C. J. Creighton, L. Danilova, L. Ding, G. Getz, P. S. Hammerman, D. Neil Hayes, B. Hernandez, J. G. Herman, J. V. Heymach, I. Jurisica, R. Kucherlapati, D. Kwiatkowski, M. Ladanyi, G. Robertson, N. Schultz, R. Shen, R. Sinha, C. Sougnez, M.-S. Tsao, W. D. Travis, J. N. Weinstein, D. A. Wigle, M. D. Wilkerson, A. Chu, A. D. Cherniack, A. Hadjipanayis, M. Rosenberg, D. J. Weisenberger, P. W. Laird, A. Radenbaugh, S. Ma, J. M. Stuart, L. Averett Byers, S. B. Baylin, R. Govindan, M. Meyerson, S. B. Gabriel, K. Cibulskis, J. Kim, C. Stewart, L. Lichtenstein, E. S. Lander, M. S. Lawrence, C. Kandoth, R. Fulton, L. L. Fulton, M. D. McLellan, R. K. Wilson, K. Ye, C. C. Fronick, C. A. Maher, C. A. Miller, M. C. Wendl, C. Cabanski, E. Mardis, D. Wheeler, M. Balasundaram, Y. S. N. Butterfield, R. Carlsen, E. Chuah, N. Dhalla, R. Guin, C. Hirst, D. Lee, H. I. Li, M. Mayo, R. A. Moore, A. J. Mungall, J. E. Schein, P. Sipahimalani, A. Tam, R. Varhol, A. Gordon Robertson, N. Wye, N. Thiessen, R. A. Holt, S. J. M. Jones, M. A. Marra, M. Imielinski, R. C. Onofrio, E. Hodis, T. Zack, E. Helman, C. Sekhar Pedamallu, J. Mesirov, G. Saksena, S. E. Schumacher, S. L. Carter, L. Garraway, R. Beroukhim, S. Lee, et al. (2014) Comprehensive molecular profiling of lung adenocarcinoma. *Nature,* 511**,** 543.

Wang, D., S. Liu, J. Warrell, H. Won, X. Shi, F. C. P. Navarro, D. Clarke, M. Gu, P. Emani, Y. T. Yang, M. Xu, M. J. Gandal, S. Lou, J. Zhang, J. J. Park, C. Yan, S. K. Rhie, K. Manakongtreecheep, H. Zhou, A. Nathan, M. Peters, E. Mattei, D. Fitzgerald, T. Brunetti, J. Moore, Y. Jiang, K. Girdhar, G. E. Hoffman, S. Kalayci, Z. H. Gümüş, G. E. Crawford, P. Roussos, S. Akbarian, A. E. Jaffe, K. P. White, Z. Weng, N. Sestan, D. H. Geschwind, J. A. Knowles & M. B. Gerstein (2018) Comprehensive functional genomic resource and integrative model for the human brain. *Science,* 362**,** eaat8464.

Xiao, Y., J. Wu, Z. Lin & X. Zhao (2018) A deep learning-based multi-model ensemble method for cancer prediction. *Comput Methods Programs Biomed,* 153**,** 1-9.

Yang, W., J. Soares, P. Greninger, E. J. Edelman, H. Lightfoot, S. Forbes, N. Bindal, D. Beare, J. A. Smith, I. R. Thompson, S. Ramaswamy, P. A. Futreal, D. A. Haber, M. R. Stratton, C. Benes,

U. McDermott & M. J. Garnett (2013) Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Res*, 41**,** D955-61.

Zitvogel, L., L. Apetoh, F. Ghiringhelli & G. Kroemer (2008) Immunological aspects of cancer chemotherapy. *Nat Rev Immunol*, 8**,** 59-73.

7 Supplementary Material

Supporting document contains Supplementary Figures 1-2 and Supplementary Tables 1-4.

Supplementary File 1 - A list of Top 50 gene expression biomarkers identified by ECMarker for early and late lung cancer stages

Supplementary File 2 - A list of Lung-cancer-stage associated functions & pathways from GSEA

Supplementary File 3 - Potential novel drugs along with drug-gene association sensitivity information that respond to the stage-specific biomarkers for lung cancer predicted by ECMarker

8 Data Availability Statement

The derived datasets are all provided by the supplemental materials.

Figures

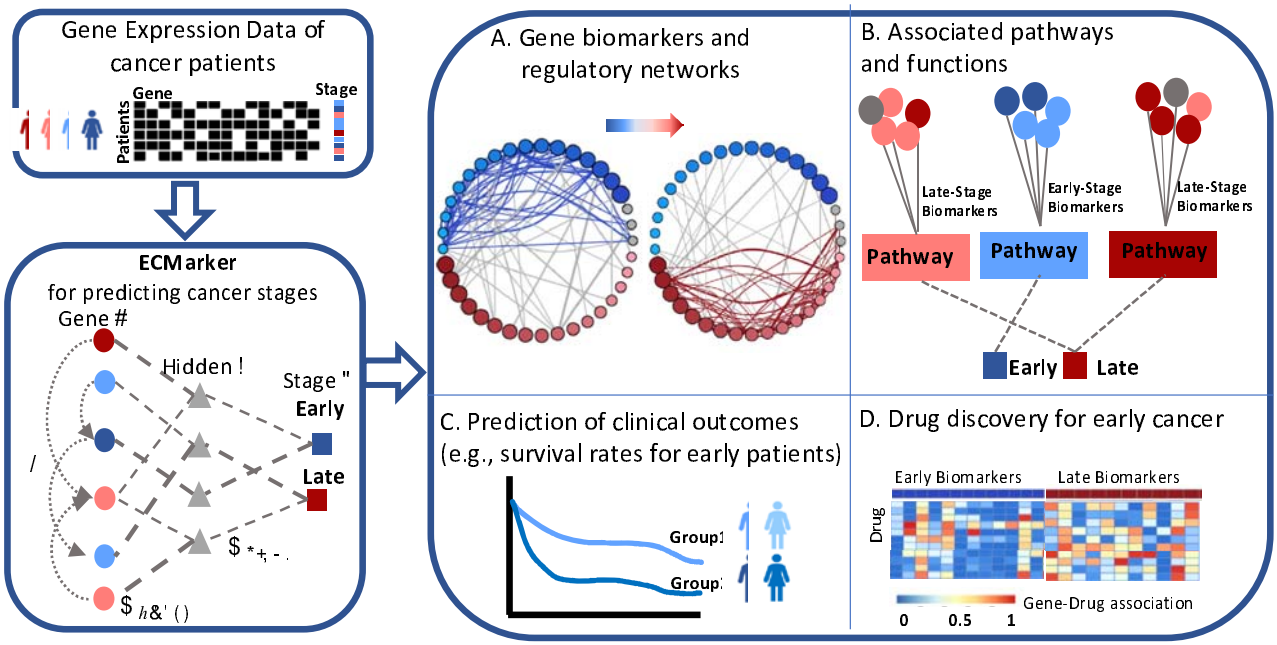

Fig. 1 ECMarker, an interpretable machine learning framework for the identification of gene expression biomarkers of cancer stages and the prediction of clinical outcomes. ECMarker was built on a hierarchical neural network model, semi-restricted Boltzmann machine (sRBM). The gene expression data of cancer patients was inputted to predict their cancer stages. ECMarker has three layers: (1) the input layer—gene, (2) the hidden layer, and (3) the output layer—cancer stage. In addition to predicting cancer stages, ECMarker is also biologically and clinically interpretable for (A) identifying the gene expression biomarkers and regulatory networks of each stage; (B) reporting the associated biological functions and pathways for each stage; (C) predicting clinical outcomes such as

survival rates, especially for early cancer patients; and (D) discovering novel drugs that could potentially affect early cancer. Red and blue represent early and late stages, respectively.

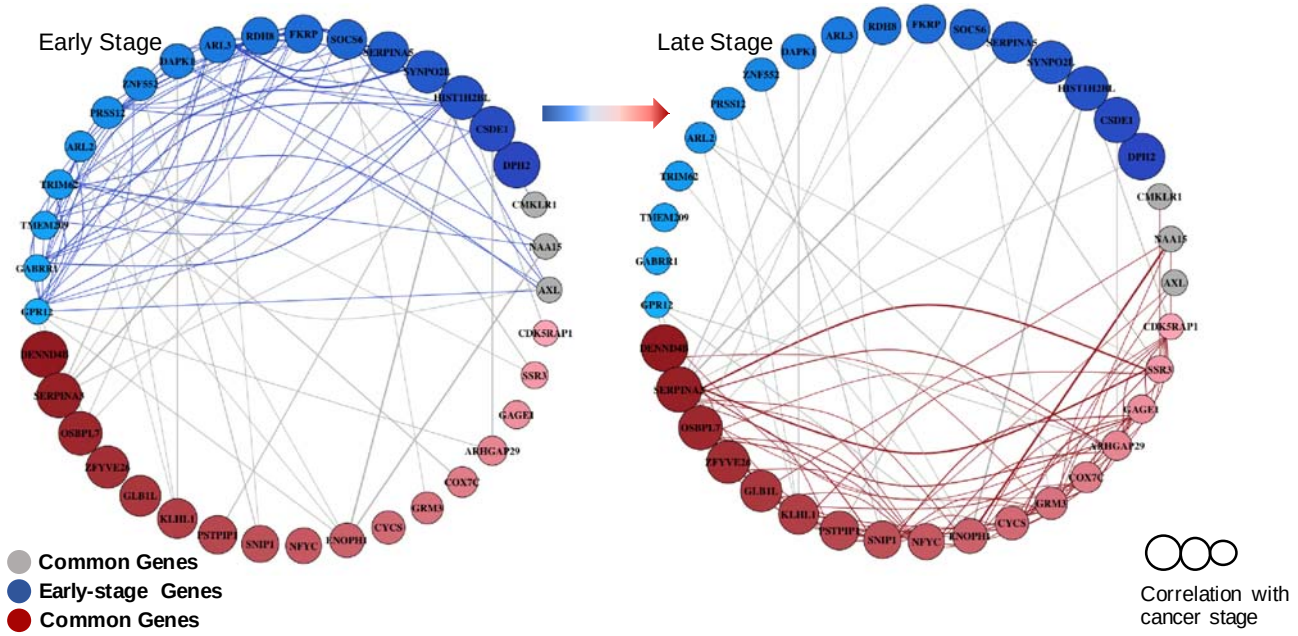

Fig. 2 ECMarker identifies cancer stage-specific GRNs for lung cancer. The GRN connects TF genes to target genes. The left network is for early-stage cancer; the right network is for late-stage cancer. Nodes are ECMarker expression biomarker genes for early-stage (blue) and late-stage (red) cancer. The grey nodes are the common biomarker genes between the early and late stages. Node size is proportional to the correlation coefficient between expression levels and stages of the gene across patients. Genes are connected with high connection weights from ECMarker. The edge width of two genes is proportional to the connection weight predicted by ECMarker.

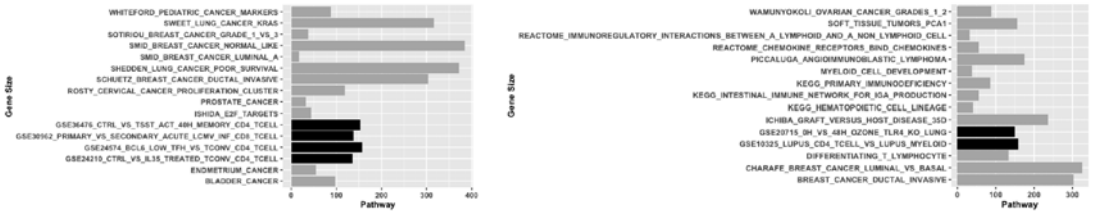

Fig. 3 ECMarker genes reveal biological functions and pathways associated with cancer stages. Select pathways and functions are found for early (left) and late (right) stages in lung cancer by gene set enrichment analysis (GSEA) for ECMarker biomarker genes. The x-axis represents the number of genes in the associated pathways/functions (y-axis). Black highlights immunological pathways. A complete list of lung cancer stage-associated functions and pathways is available in Supplement File 2.

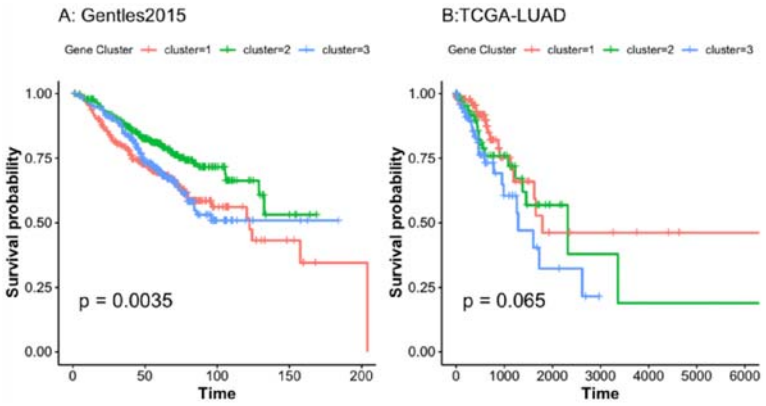

Fig. 4 ECMarker genes enable the prediction of early cancer survival rates. Early lung cancer patients were clustered into three major groups based on the top 25 ECMarker biomarker genes for early-stage cancer. These biomarker genes were found using the Gentles2015 dataset (Gentles et al. 2015). A Kaplan–Meier analysis showed that the clustering groups using ECMarker biomarkers had significantly different survival rates ($p < 0.0035$), as shown in the left panel. In addition, application of these biomarker genes to an independent cohort, TCGA-LUAD (The Cancer Genome Atlas Research et al. 2013), showed that early-stage patients also had significantly different survival rates ($p < 0.065$), as shown in the right panel.

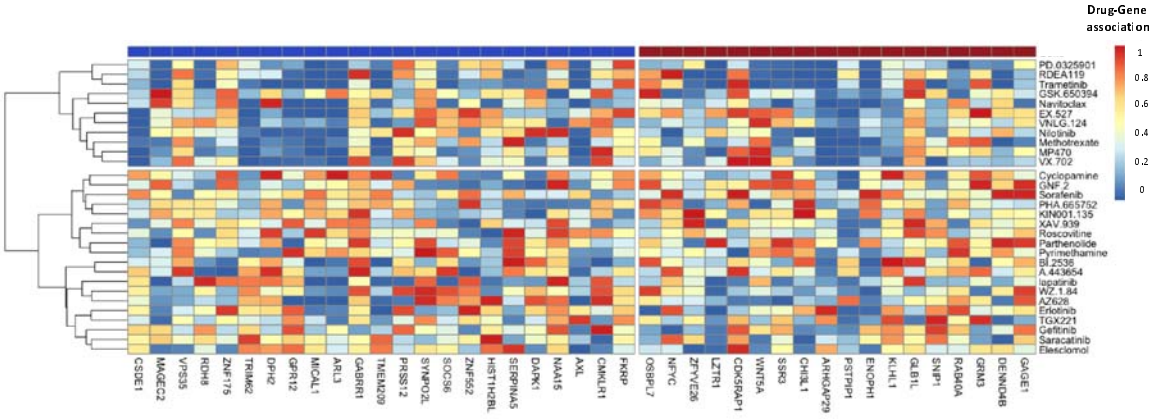

Fig. 5 ECMarker genes can be used to discover novel, potentially effective drugs for early lung cancer. Using early stage-specific and late stage-specific gene groups to predict which drugs might be effective. The heatmap shows the gene-drug association matrix that ECMarker predicted for lung cancer. ECMarker (x-axis) predicted 19 early-stage biomarker genes (blue), 18 late-stage biomarker genes (red), and 30 potentially effective drugs (y-axis). The color bar in the range of [0,1] corresponds to the gene-drug association value. A higher gene-drug association value indicates greater drug sensitivity. Out of 37 associated gene-drug pairs (estimate coefficient > 0.917), 21 pairs have been supported by previous clinical studies.