InFAMOS - an integrated method and system to identify functional patient-specific somatic aberrations using multi-omics cancer profiles

A. Background of the intervention

The pathobiology of cancer is associated with significant aberrations within the natural complex biological processes governing the growth and differentiation of normal cells. However, there exists significant heterogeneity within cancers originating even within the same tissue type, possibly reflecting the multiple ways in which the normal signaling networks can be pathologically altered. This heterogeneity underlies the significant challenges posed in the development of diagnostic and theranostic biomarkers as well as potential therapeutic interventions in oncology, and points to the need for a systems-level understanding of cancer etiology and progression.

For instance, the ERBB2 gene which encodes a member of the epidermal growth factor (EGF) receptor family of receptor tyrosine kinases and plays a significant role in cell proliferation is highly overexpressed in multiple cancers, especially breast, GI and ovarian cancers. This gene is mutated in approximately 15% of breast cancer and in most cases its overexpression is associated with copy number amplifications, and has resulted in the definition of a specific subtype of breast cancer named after this gene, HER2-positve breast cancer. Despite the availability of a targeted therapeutic intervention for this particular subtype of breast cancer, namely Herceptin, the response rate of breast cancer patients to this therapy remains in the 50-55% range. This heterogeneity in response points to the existence of other genetic modulators of tumor progression. Indeed, it has been shown that aberrations in the AKT/PI3K pathway, such as deletions of the tumor suppressor gene PTEN, and mutations in the PIK3CA gene result in resistance to Herceptin. However, no systems-level pathway model currently exists that can integrate all of these factors into a single integrative biomarker for therapy resistance.

Since pathway level aberrations can result from multiple sources, such as somatic mutations, copy-number alterations, epigenetic variations and the regulatory gene expression changes, jointly modeling these sources of variability is essential to developing comprehensive pathway-based predictive models of use in oncology. Furthermore, with the recent advances in low-cost genome-wide data acquisition techniques in molecular biology, measurements of the different sources of variability are becoming increasingly available. However, modeling frameworks that can fully utilize the information present in these multi-omics profiles are lacking in both the research and diagnostic communities. Development of computational frameworks to integrate various data sources including RNA expression level, copy number variations, DNA methylation patterns, and somatic mutations with the objective of finding clinically useful biomarkers is therefore an essential need in the oncology community.

Recently, several integrative approaches are proposed to incorporate various sources of information into a unified framework to facilitate early cancer diagnosis, clinical outcome prediction and more relevant therapeutic interventions. Majority of these approaches take one of the two extreme perspectives of either i) totally ignoring the conceptual biological information and relying solely on data-driven techniques or ii) fully trusting the conceptual biological information that is presented in a network of interacting molecular entities.

Ignoring biological interactions among the cell molecular entities (e.g. genes and proteins) is highly inefficient in finding biologically relevant subset of entities with significant collective predictive powers, due to the potential of data over-fitting. Indeed, this problem is particularly accentuated in cancer research since the number of cancer samples in any given study tends to be an order of magnitude lower than the number of molecular features measured. On the other hand, a full reliance on descriptive biological networks ignores their limitations: the pathway networks are typically constructed based on experimental evidence in a specific cellular context, which may not always be translatable to other tissue and pathological contexts.

In this work, we take a hybrid approach and incorporate both measurement-based omics data and partially trusted pathway information into a unified Bayesian framework to build a gene-gene influence network which is capable of predicting a particular gene expression level given the regulatory network status. The Bayesian framework not only refines and extends our knowledge of tissue-specific protein-protein interactions but also provides patient-specific predictions and conditional distributions of network entities (e.g. genes). These patient-specific gene expression predictions are then leveraged to find significant deviations and inconsistencies in gene expression levels from expected levels in individual patient samples, thus allowing for the discovery of potential associations with phenotypes such as therapy response and prognosis.

B. Key elements of the invention

Our invention overcomes several significant limitations in integrating biological information and various molecular measurement data sources into a unified network-based computational framework. This paves the road to reveal more relevant patient-specific malfunctioning genes and perturbed biological processes.

(a) The pure data-driven methods are based on extracting cancer-associated genes from RNA expression data, where genes presenting with significant differential expression levels between the normal and cancer samples are reported as cancer associated biomarkers. This method, although successful in finding relevant genes (good sensitivity), presents with poor specificity, resulting in substantial false discovery rates.

Our method incorporates the biological information and reports only genes that show significant inconsistency with the underlying network-based predictions and patient-specific measurements. This approach, therefore, results in higher specificity as well as sensitivity in identifying the most functionally-relevant genes associated with the phenotype in consideration.

(b) The current set-based methods take biological information into account by first annotating sets of genes that are jointly associated with a particular phenotype or cellular/biological process based on a prior biological knowledge. For instance GSEA [2], provides a powerful tool to assess the relevance of a candidate gene set based on ranking all genes according to their differential expression levels and then assigning an enrichment score for the test set as a measure of deviation from the

random appearance of the test set in the ranked list. However, this method and similar set-based methods are not capable of adaptive integration and the software user should include the biological information manually via forming potentially more relevant gene sets.

Our method, in contrast, identifies genes connected by specific regulatory relationships as defined in the pathway network annotations. The resulting pathway subnetworks associated with a phenotype provide functional insights along with robust biomarkers and is therefore widely applicable across cancers.

(c) The currently available network-based methods such as Paradigm, Pathologist and SPIA [3] aim at integrating pathway information with measurement data, in order to identify perturbed pathways and genes exhibiting significant deviations from predictions obtained from the network. However, these approaches have two important drawbacks. Firstly, these approaches fully trust the biological pathway network relationships without allowing for the potential of tissue-specific variations in pathway network connectivity. The second and even more important issue is that these techniques overlook the potential for functional heterogeneity amongst the interaction links in the network. They presume an equal influence of all the direct parent nodes, while in reality the influence of some regulatory parent genes may be significantly higher than the other parent genes.

Our algorithm does not rely fully on the pathway network but rather refines the influence network by assigning different coefficients to the network edges that are learned from the multi-omics data. Therefore, our method highlights and discovers the heterogeneous relations among network nodes (e.g. Genes, RNAs, proteins).

(d) Recently, various algorithms have been proposed to take into account this variability and develop a more relevant influence network. For instance, HotNet algorithm intends to discover this heterogeneity in the pairwise influence through defining a measure of pairwise influence among gene pairs based on their distances and the number of connecting paths [4]. A similar approach is taken in [5] to identify highly mutated subnetworks using the idea of smoothed networks by propagating somatic mutations over the network with a predefined attenuation at network nodes. However, these methods suffer from ignoring potential variations in the strength of the influence between nodes, which may be inferred using transcriptional measurements. In other words, these techniques only consider network topology and do not leverage information derived from multi-omics measurements that may be available.

Our method, in contrast, uses both biological pathways and multi-omics measurement data to capture not only the topology but also the strength of the influence between nodes in the network as mentioned above. Therefore, it provides a more accurate and realistic influence among network nodes. Secondly, our method is not limited to finding paths that are frequently hit by somatic mutations, but also finds the malfunctioning nodes.

(e) In a recent work [6], the authors provide an elegant gene interaction model based on Boolean networks to predict a specific gene expression level. Although insightful in

describing the underlying biological processes, this approach is not a natural representation of genes expression levels that take on continuous values, as opposed to the overly simplistic ON/OFF states as presumed by the Boolean network approach, and hence may not be capable of capturing exact relationships among gene expression levels in general.

Our method takes a more natural perspective of treating gene expression levels as continuous variables that is more coherent with the actual biological process. Therefore, the data driven-learned influence network is highly capable of capturing the real regulatory relationships among genes.

C. Detailed description of how to build and use the invention

The proposed algorithm consists of four main modules. In Module 1, the whole pathway network breaks down into multiple downstream trees, where each tree includes a particular gene as its root and a network of **upstream regulators of gene's transcription**. We capture the relation between the gene and its upstream regulators as gene-gene relations, hence the leaves are usually genes. We use the terms *"ancestor genes"* or simply *"ancestors"* to refer these genes. In Module 2, each tree subnetwork is used to learn a non-linear function to predict the corresponding gene expression level from its own epigenetic information (e.g. DNA Methylation and Copy Number) and its regulatory ancestor gene expression levels. This provides a bank of functions each corresponding to a specific gene in the context of specific tissue type. This function database is learned once and can be used for patient-specific analysis in the two subsequent steps performed by Modules 3 and 4. Module 3 receives information for a given patient and performs prediction of gene expression levels for all genes with regulatory network using the function bank. This module further calculates the consistency scores for each gene by comparing the actual measurement and the predicted value. Module 4, identifies the genes whose expression levels are significantly inconsistent with the prediction values obtained from the regulatory network. These genes are likely malfunctioning due to somatic mutations in this gene or its ancestors. Module 4 further provides statistics to evaluate the significance of ancestor gene mutations that potentially are associated with the inconsistencies in the child gene expression level. The following flowchart presents the overall block-diagram of the proposed algorithm.

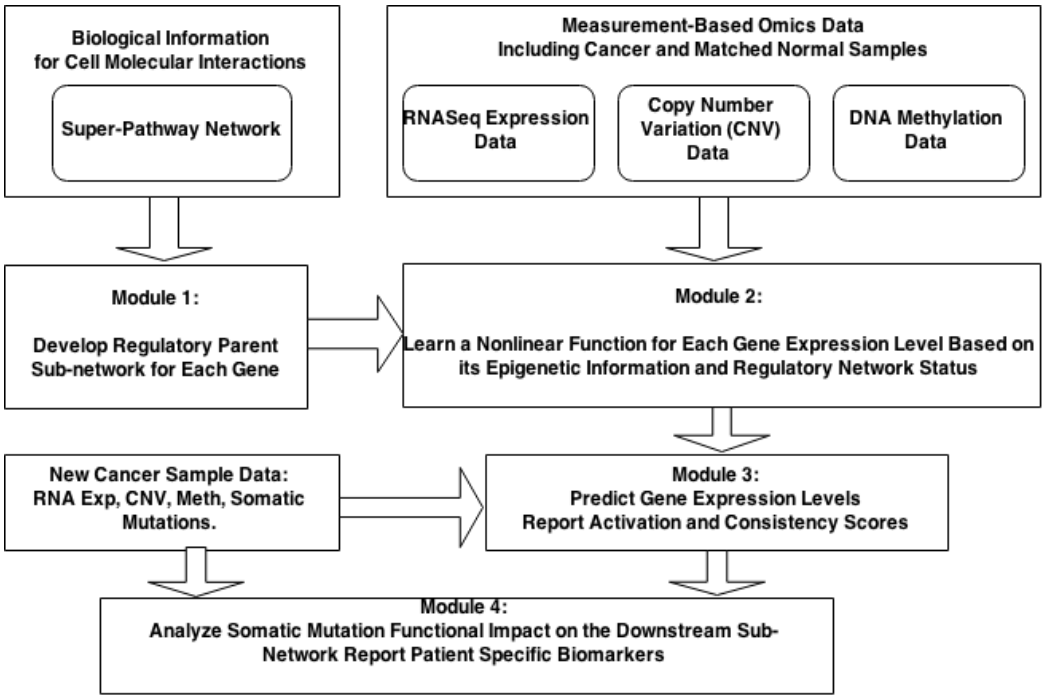


Figure 1: The block-diagram of the proposed algorithm presenting the four modules

Module 1: Incorporating Pathway Networks

Pathway networks are widely used to present the intra-cellular interactions and gene regulatory networks in a network format. The network builds a directed graph of nodes and edges. The nodes may consist of a diverse range of entities such as genes, proteins, RNAs, miRNA, protein complexes, signal receptors, and even abstract processes such as apoptosis, meiosis, mitosis, and cell proliferation. The network edges determine the pairs of interacting nodes and specify the type of each interaction. Several publicly available pathway networks are developed to model intra cellular activities various species and issue types, some of which under specific contexts (e.g. metabolism, various cancer types , ...). Currently, several well-curated pathway collections are available on-line including NCI-PID, wiki-pathways, Biocarta, KEGG, and Reactome.

In this work, we used a comprehensive network that brings together pathways from various pathway sources including NCI-PID, Biocarta and Reactome for homo sapiens and is used in several studies [7]. This network consists of six node types including Proteins or the corresponding genes, RNAs, Protein complexes, Gene Families, miRNAs, and abstracts. These nodes interact with one another in six different ways of i) positive transcription , ii) negative transcription, iii) positive activation, iv) negative activation, v) member, and vi) component. Usually, transcription is terminated only to genes represented by the corresponding proteins, while activation is applicable to all node types. Likewise, component and member relations are proprietary input commands for protein complexes and Gene families, respectively.

In order to learn a function relating a gene's mRNA expression level to its epigenetic parameters (DNA methylation and copy number variation) as well as its regulatory network, we build an upstream tree for each gene. We start from each gene and traverse the upstream network in the opposite direction of links to collect all upstream nodes. We use the depth first traversal algorithm with some modifications.

Firstly, we terminate traversing a branch once we reach a predefined maximum depth level, where the depth is defined as the number of links from the visiting node to the root node. We then eliminate all the branches do not terminate to a gene node; therefore, the leaves of tree are always genes. We pass through all nodes, except abstract nodes that represent the conceptual abstract processes in order to avoid unnecessary network complications and inclusion of irrelevant interactions.
While reaching a gene node, we only pass through the links that are not of type "transcription". The reason is that the part of the upstream regulatory network which terminates to a gene node via a "transcription" link, is already accounted for by considering the expression level of this particular gene. The only exception for this rule is the root node, where we do the exact reverse.
Passing from the root node to the direct neighbors in the first ring of root neighborhood is only allowed if the connecting edge is of *"transcription"* type to limit the parents to those who impact the expression level of the gene residing in the tree root. We also keep track of the lengths from the leaves to the root node that is further used in the function learning process. Finally, if we meet a node via two disjoint paths, the shortest path is considered. The pseudo-code for the module 1 algorithm is summarized below and a sample upstream tree extracted for gene *PPP3CA* from the network is depicted in Fig. 2.

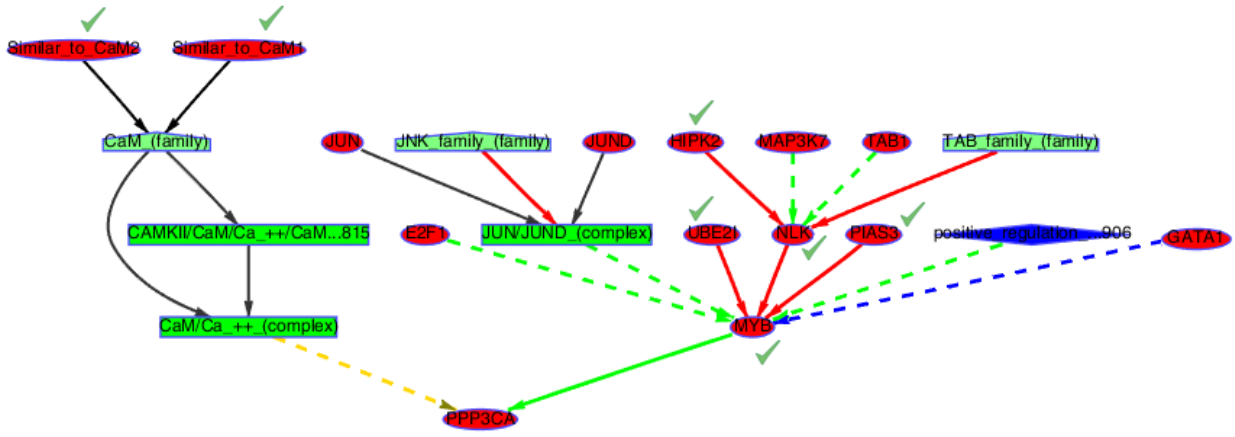


Figure 2: Sample tree showing the regulatory network of gene *PPP3CA*. The subnetwork includes ancestor genes with depth 1 up to 3rd level. Shapes define the node types with Genes (oval), Protein complexes (rectangle), gene families (house-like shapes), and abstracts (diamonds). The edges are colored according to their regulatory relationships with positive activation (yellow), negative activation (red), positive transcription (green), negative transcription (blue), component (gray) and member (black). The nodes connecting through dashed links are not included in the ancestor list.

Module 1: Building Regulatory Network for Each Gene using Modified Depth First Traverse Algorithm
Inputs: Pathway network, gene id: (g), maxDepth
Output: Regulatory Ancestor Gene Set, Depth Information

1- Set root node to input gene (r = g)
2- Set visit node to input gene (v = g), Set depth = 1;
3- Initialize visit_list, depth_list, connType_list and ancestor_list to empty sets.
4- Visit node (v, depth)
    a. If node v is in the visit_list and depth < depth_list(v)
        i. Update depth_list(v)=depth
        ii. return success
    b. If node v is in the visit_list and depth >= depth_list(v)
        i. Avoid this node
        ii. return fail (already visited via a shorter path)
    c. If node v is not in the visit_list
        i. Add v to visit_list
        ii. set depth_list(v) = depth
        iii. return success
5- If Visit node return fails [node is already visited], exit
6- If node v is a gene
    a. Add v to the ancestor list (g, depth, ancestor list)
        i. If node v is in the ancestor_list and depth < depth_list(v)
            1. Update depth_list(v)=depth
        ii. If node v is in the ancestor_list and depth >= depth_list(v)
            1. Avoid this node (already visited via shorter path)
        iii. If node v is not in the ancestor_list
            1. Add v to ancestor_list
            2. set depth_list(v) = depth

7- If depth < maxDepth [pass through this node to the next level]
    a. depth = depth+1;
    b. If node v is the root node (v==r)
        i. Remove the direct parents not connected via edges of type "transcription"
    c. If node v is of type gene and is not the root node (v<>r)
        i. Remove the direct parents connected via edges of type "transcription"
    d. for all nodes u in the parent list of the node v
        i. Goto step 4, call Visit node(u)
    e. depth = depth – 1; [return to previous level]
8- End

We see that the first level ancestors (direct parents) of the root node *PPP3CA* are connected via "transcription" edges that regulate the gene expression level. For instance, the complex *CAM/Ca++* is connected to the root node via activation link, hence does not regulate gene expression level. Therefore, all the genes connecting via complex *CAM/Ca++* in the left side of Fig. 2 are excluded from the final ancestor list. While passing through other genes, only non-transcriptional links are allowed. For instance, the upstream subnetwork of MYB is limited to the non-transcriptional nodes such as *PIAS3* and *MAP3K7*

genes, whose impact is not already captured via the MYB expression level. The impact of genes *GATA3* and *E2F1* is implicitly accounted for by the expression level of gene *MYB*.

In Fig. 3, the empirical distribution of the number of ancestors is presented in logarithmic scale. A large number of genes are upstream isolated orphan genes. Only 839 genes have ancestors ranging from only one ancestor for 23 genes up to 1152 ancestors for gene CDKN1A, when traversing up to 7 links upstream of the root node.

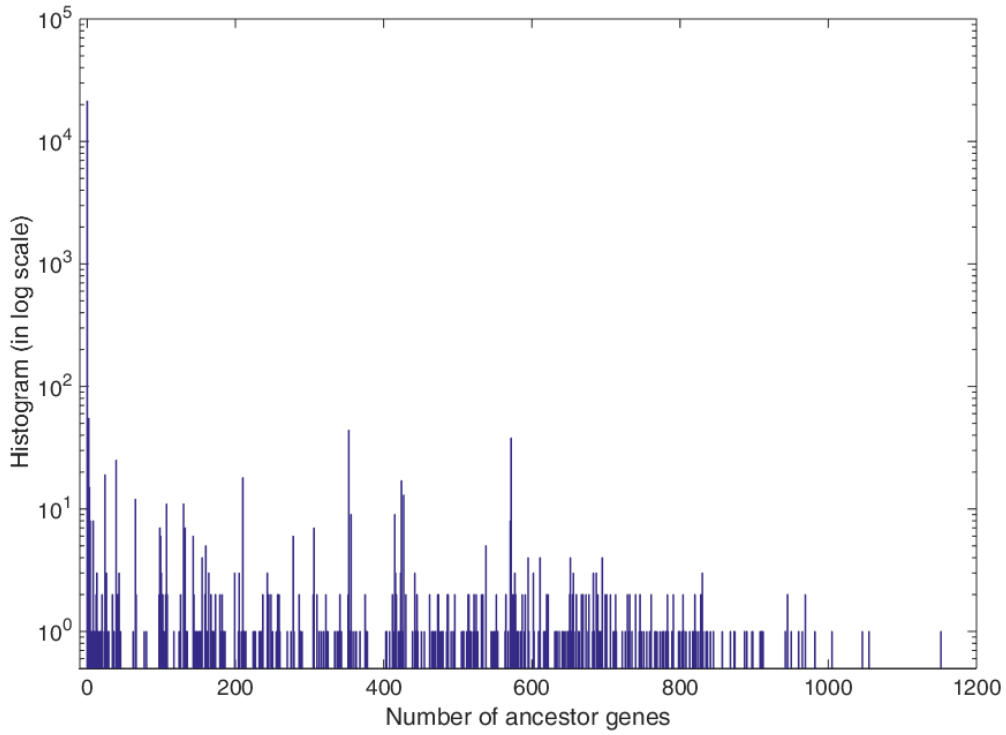


Figure 3: Histogram of ancestors count for genes. Number of ancestors vary widely, ranging from 1 to a maximum of 1150. Genes with zero ancestors were not represented in the pathway network.

Module 2: Learning a nonlinear function for each gene

In the previous section, we have built an upstream tree for each network node of type gene which is not upwards isolated. In this section, we learn a function relating the expression level of the gene residing at root node to its regulatory network and its own epigenetic information (e.g. methylation and CNV). Since, multiple DNA methylation probes may overlap with a gene's coding or regulatory regions, we leverage methylation measurements to the model through several representative statistics including minimum, maximum, and weighted mean value, where in calculating the weighted mean we exclude the regions with less than 10 probes for more accuracy. If gene g, overlaps with n_g^m regions, each having probe numbers $p_1, p_2, \dots, p_{n_g^m}$; and corresponding methylation measurement of $m_1, m_2, \dots, m_{n_g^m}$, then the weighted mean is calculated as

$$\bar{m}_g = \frac{\sum_{i=1}^{n_g^m} p_i m_i I(p_i \geq 10)}{\sum_{i=1}^{n_g^m} p_i I(p_i \geq 10)},$$

where I(.) is identity function.
To include copy number variations, we use the segment mean value provided for the region that harbors the particular gene. Most genes fall into a single CNV segment. Otherwise, if a gene falls in the border of two segments, we simply take the mean value of both segment measurements.

In order to learn a function for each gene, we use mRNA expression, somatic copy-number alteration and DNA methylation measurements for n_g samples to form the following classical regression model.

$$y_g = \mu_g \mathbf{1}_{n_g} + X_g \beta_g + \epsilon, \epsilon \sim \mathcal{N}(0, \sigma_g^2 \mathbb{I}_{n_g})$$

Where y_g is a $n \times 1$ vector of expression levels for gene g for all n_g samples. $X_g = [X_g^S, X_g^P]$ is a $n \times p$ data matrix composed of two parts including X_g^S (self-methylation and CNV data) and X_g^P (the expression levels of the ancestor genes). The term 1_{n_g} is all one column vector of length n_g and ϵ is the model noise with i.i.d zero-mean unit-variance Gaussian elements. μ_g is the expected value of gene g expression level.

The objective here is to find the optimal model parameters $\beta_i, \mathrm{i} = 1,2, \dots, \mathrm{p}$ that provides the best prediction power via minimizing the Mean Squared Error (MSE).

One may use normal samples in learning phase to avoid model crash due to severely perturbed interactions in the highly contaminated/disordered cancer cells. However, this may led to a poor predictive power when the number of predictors are large or comparable with respect to the number of samples ($n < O(p)$). In most studies, the number of cancer samples profiled tend to be significantly higher than the number of normal samples. For instance, in the case of TCGA data for breast cancer, the number of cancer samples exceed the normal samples by a factor of 10. Consequently, excluding all cancer samples is highly inefficient. On the other hand, including cancer samples in the training set, may

deteriorate the model performance for specific genes that significantly deviate from the true underlying biological function in some samples due to genomic events as stated above; Therefore, we choose to include all the normal samples and part of the cancer samples that have not impacted by somatic mutations in this particular gene and its ancestors in order to learn the predictive function. This approach leads to a different training set size for each gene, but provides a considerable improvement in model prediction power as demonstrated in section D.

One natural solution for this problem is the Least Squared Error (LSE) Solution that minimizes the squared error for the training set, when no prior information is available about the model parameters β_i.

$$\beta = \underset{\beta}{\operatorname{argmin}}(y - X\beta)^T(y - X\beta) \Rightarrow \beta_{LSE} = (X^T X)^{-1} X^T y$$

The LSE solution is not optimal when we have prior information about the model parameters. Here, we have prior knowledge about the model that can be used to enhance the model accuracy. Firstly, it is likely that not all of the ancestor genes may have a substantial impact of a given gene's expression levels. Therefore, we expect that a significant number of the model parameters β_i could be shrunk towards zero. Therefore, imposing sparsity enhances the model generalization property by avoiding noise over-fitting. Although part of sparsity is already accounted for by using the pathway network and including only ancestor genes instead of using all genes as the input data; but still a higher level of sparsity is expected, when the number of ancestor genes grow higher (in order of tens and hundreds).

One of the common optimization-based solutions to impose sparsity is regularizing the norm of model parameters. The penalization can be applied to the $l_p, p \geq 0$ norm of coefficient vector $\beta = [\beta_1, \beta_2, \dots, \beta_p]^T$, which is called bridge regression. Important special cases of this approach are Lasso, ridge, and subset selection for l_1, l_2, l_0 norm penalization, respectively. In elastic net, the penalty term is the linear combination of l_1 and l_2 penalty [8].

$$\beta_L = \underset{\beta}{\operatorname{argmin}}(y - X\beta)^T(y - X\beta) + \lambda_1|\beta|_1 + \lambda_2|\beta|_2,$$

where λ_1 and λ_2 are shrinkage parameters to impose sparsity and generalizability. Efficient algorithms based on convex optimizations, Basis pursuit, LARS, coordinate descent, Dantzig Selector, Orthogonal matching pursuit, and approximate message passing are proposed to solve this problem. However, the most limiting drawback of these methods is that it only provides point estimates for the regression coefficients.

We choose to employ a Bayesian framework that provides more detailed information about the model parameters through posterior distribution to be used in the consistency check analysis. It also enables us to incorporate other prior knowledge in addition to sparsity as explained below.

An important fact which is highly overlooked in analyzing gene expression studies is potentially non-linear relations among the biological measurements. In order to capture

such nonlinear relationships, we use a centered sigmoid function $f_1(x;\ c) = \frac{1-e^{-\frac{x}{c}}}{1+e^{-\frac{x}{c}}}$ to capture the sensitivity around the mean value and the soft thresholding function $f_2(x;\ c) = sign(x)\left(\sqrt{x^2+c^2}-c\right)$ to account for the cases in which only extremely high or low values contribute to the model. $f_2(x;\ c)$ can be considered as a softer version of the commonly used peace-wise linear soft-thresholding function, $f(x;\ c) = sign(x)(|x|-c)_+$.
These functions are depicted in Fig. 4 compared to the linear function. We have applied the element-wise non-linear extension $X_g \to \Psi(X_g) = [X_g^s, f_{1(X_g^s)}, f_{2(X_g^s)}, X_g^p]$ only to the self data (e.g. Methylation and CNV data), hence the number of predictors increases slightly compared to the number of ancestors for each gene. It is notable, that if the actual underlying function is linear, the coefficients of the nonlinear terms tend to zero in the proposed model, hence no performance degradation is observed while learning a nonlinear function for true linear relations. The results in section D confirm the nonlinear relation between the gene expression level and its epigenetic information.

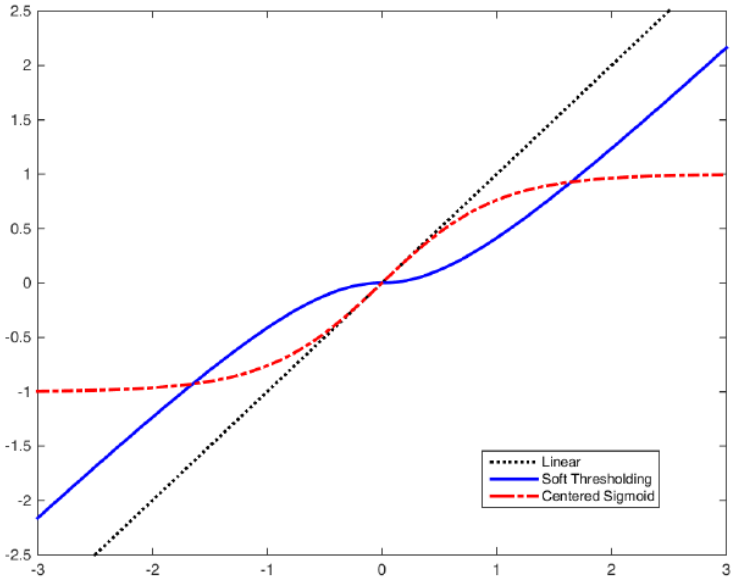


Figure 4: Nonlinear function including centered sigmoid and soft thresholding to capture near mean-sensitivity and near-mean ignorance properties.

Another important biological consideration in developing the ancestor-set for each gene when traversing the pathway network upwards is the variation in the distance of leaf nodes to the root node. One may expect the closer ancestors contribute more to the descendant downstream gene expression level than farther nodes that are connected via a long chain of intermediate nodes. Hence, the closer nodes tend to pose higher coefficient in the regression model. We leverage this fact into our model through depth penalization mechanism in our Bayesian framework.

In this section, we propose the Bayesian framework to predict the gene expression level via a nonlinear transformation/projection of its self-epigenetic data as well as the expression levels of the its regulatory ancestor genes. The proposed Bayesian framework provides the

desired statistics (e.g. median, mean, moments and ...) via full posterior distributions of the model parameters. Moreover, we can incorporate a prior knowledge about the model parameters using hierarchical Bayesian models. The Bayesian approach results in higher computational costs since the full posterior is not usually in a closed-form format and requires sampling from the posterior distributions to find the desired point estimates. However, the resulting posterior distributions provide significant insights into the functional effects of aberrations in the pathway and therefore justify the relatively higher computational costs.

We use the idea of global and local shrinkages with penalization based on the distance of the ancestor gene (i.e. the number of links from leaves to root in the regulatory network) from the gene whose expression is being predicted. We propose constructing the following model, where the subscript g is omitted for notation convenience:

$$
\begin{aligned}
& X \to \Psi(X) \\
& y|X,K,\beta,\sigma^2 \sim N(\Psi(X)\beta, \sigma^2 I_n), && K = \mathrm{diag}([k_1^2,\ldots,k_p^2]) \\
& && (4) \\
& \beta|\sigma^2,\tau_1^2,...,\tau_p^2 \sim N_p(0_p, \sigma^2 D_\tau K^{-1}) && D_\tau = \mathrm{diag}(\tau_1^2,...,\tau_p^2) \\
& k_1^2,k_2^2,\ldots,k_p^2 \sim \prod_{j=1}^{p} \frac{(a_k/d_j^2)^a}{\Gamma(a_k)} k_j^{2(a_k-1)} e^{-a_k k_j^2/d_j^2} && k_j^2 \sim \mathrm{Gamma}(a_k, a_k/d_j^2) \quad j=1,2,3,\ldots \\
& \tau_1^2,...,\tau_p^2 \sim \prod_{j=1}^{p} \frac{\lambda^2}{2} e^{-\lambda^2\tau_j^2/2} && (\tau_j^2 \sim \mathrm{Expo}(\lambda'=\lambda^2/2, \mu'=2/\lambda^2) \\
& \pi(\sigma^2) = \frac{b^a}{\Gamma(a)} (\sigma^2)^{-a-1} e^{-b/\sigma^2} && = \mathrm{InvGamma}(a_{\sigma^2}>0, b_{\sigma^2}>0) d\sigma^2
\end{aligned}
$$

The above formulation extends the normal gamma prior construction in order to incorporate the link depth information to the prior. We choose to leverage this information via coefficients k included in the variance of the model parameters. Thus, the variance of β_i is chosen to be inversely proportional to the link depth of the corresponding ancestor via setting $var(\beta_i) = \sigma^2\tau_i^2/k_i^2$, where σ^2 controls the global shrinkage, τ_i^2 accounts for the local shrinkage and k_i^2 enforces the link depth impact. To provide more flexibility, we use a Gamma prior distribution for k_i^2 . Using gamma prior has the advantage of yielding closed-form posterior distribution for k_i and hence facilitates employing computationally efficient Gibbs sampler. Therefore, we use $k_i^2 \sim Gamma(a_k, b_k)$ such that the mean of variance is inversely proportional to depth parameter, i.e. $E[k_i^2] = \frac{a_{k_i}}{b_{k_i}} = d_i^2 c$. The constant c is a normalizing term to ensure $\frac{1}{p}|K|_1 = 1$, which is obtained by setting $c = \frac{p}{\sum_{i=1}^{p} d_i^2}$. Therefore, we only have one free hyperparameter a_{k_i} for k_i prior distribution and the second parameter $b_{k_i}$is automatically obtained from $b_{k_i} = a_{k_i}/cd_i^2$. We note that $var(k_i) = a_{k_i/b_{k_i}^2} = c^2 d_i^4/a_{k_i}$. Setting a_{k_i} to small values provides higher variance for k_i and hence is less formative, while large values of a_{k_i} provides low variance reflecting a high certainty about the network topology and the fact that node pairs with shorter paths are associated with higher influences to one another. In this case, the gamma distribution approaches a Gaussian

distribution concentrated around d_i. We choose the relatively large value of $a = a_{k_i} = cd_i^2 b_{k_i} = 10$ to highlight on the significance of the underlying biological network.
The above hierarchical model yields the following full joint distribution:

$$p(y, X, K, \beta, \sigma^2) = p(y|X, K, \beta, \sigma^2)\pi(\beta|\sigma^2, \tau_1^2, ..., \tau_p^2)\pi(k_1, \ldots, k_p)\pi(\tau_1^2, \ldots, \tau_p^2)\pi(\sigma^2)$$
$$= \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{1}{2\sigma^2}(y-X\beta)^T(y-X\beta)} \prod_{j=1}^{p} \frac{(a_k/d_j^2)^a}{\Gamma(a_k)} k_j^{2(a_k-1)} e^{-a_k k_j^2/d_j^2}$$
$$\prod_{j=1}^{p} \frac{1}{\sqrt{2\pi\sigma^2\tau_j^2/k_j^2}} e^{-\frac{\beta_j^2 k_j^2}{2\sigma^2\tau_j^2}} \ldots \prod_{j=1}^{p} \frac{\lambda^2}{2} e^{-\lambda^2\tau_j^2/2} \frac{b^a}{\Gamma(a)} (\sigma^2)^{-a-1} e^{-b/\sigma^2},$$

Which immediately provides the following posterior distributions using the fact that the full conditional posterior distribution for each parameter is simply the product of the terms including that variable with other terms serving as a normalization constant to ensure the resulting probability integrates to one. This method is called completion of terms:

$$\beta|\mu, \Psi(X), y, \sigma^2, \tau_1^2, ..., \tau_p^2 \sim N_p(A^{-1}\Psi(X)^T\tilde{y}, \sigma^2 A^{-1}), \quad A = \Psi(X)^T\Psi(X) + KD_\tau^{-1}$$
$$\gamma_j = 1/\tau_j^2|\mu, \Psi(X), y, \beta, \sigma^2 \sim \text{inverse Gaussian}(\frac{\lambda\sigma}{|\beta_j|}, \lambda^2)I(\gamma_j > 0),$$
$$\sigma^2|\mu, \beta, \Psi(X), y, \tau_1^2, ..., \tau_p^2 \sim \text{inverse Gamma}(a + \frac{n+p}{2}, b + \frac{1}{2}(\tilde{y} - \Psi(X)\beta)^T(\tilde{y} - \Psi(X)\beta) + \frac{1}{2}\beta^T D_\tau^{-1} K\beta)$$
$$k_j^2|\mu, \beta, \Psi(X), y, \sigma^2, \tau_1^2, ..., \tau_p^2 \sim \text{Gamma}(a_k + \frac{1}{2}, \frac{a_k}{d_j^2} + \frac{\beta_j^2}{2\sigma^2\tau_j^2})$$

We use Woodbury Matrix inversion formula to calculate A^{-1} when $n < p$ to obtain more stable results and save in computations by converting a $p \times p$ square matrix inversion to a $n \times n$ one [9]. We apply a Gibbs sampler with burn-in iterations 1000 and computation iterations 5000 to obtain the approximate posterior distributions for the model parameters β_i, σ.

Module 3: Predict gene level expression for a new sample and report activation and consistency level for all genes

Running Module 2 using training samples from both normal and cancer cohorts provides a function bank, where each function corresponds to a specific gene. This function bank can be used to analyze test samples to identify potential inconsistencies. This module basically performs gene expression level prediction for all genes. For each gene, we extract the expression levels of the ancestor genes as well as the self-epigenetic information for all samples. Then, we predict expression level of this specific gene for all samples using the corresponding function learned for this gene. The prediction process provides the conditional posterior distribution for the expression level of this gene. We use the maximum a-posteriori (MAP) method to obtain the expected gene expression levels.

In order to calculate consistency scores for unisolated genes for which a function is learned, we take the following approach. We note that the predictive distribution for the RNA expression of any gene for each new test sample y^{new} is obtained by marginalizing out the model parameters from the conditional posterior distribution for given input x^{new} (self epigenetic info and ancestors expression levels):

$$f(y^{new}|x^{new}) = \int f(y^{new}|x^{new}, \beta, \sigma^2) f(\beta, \sigma^2|y, X)\, d\beta d\sigma^2$$

While the first term which is the conditional distribution is available in-closed form, the second term which is the posterior distribution of model parameters is not. This distribution can be approximated with the following expression, where the realizations of model parameters ($\beta^{(i)}, \sigma^{2^{(i)}}$) are obtained using Gibbs sampling method.

$$f(y^{new}|x^{new}) \cong \frac{1}{N}\sum_{i=1}^{N} f\left(y^{new}\middle|x^{new}, \beta^{(i)}, \sigma^{2^{(i)}}\right)$$

It is noticeable that the above distribution is in fact a Gaussian mixture model (GMM) with large number of equi-probable components of mean ($\Psi(x^{new})^T\beta^{(i)}$) and variance ($\sigma^{2^{(i)}}$). We note that if the Gibbs sampler converges, $\beta^{(i)}$ is concentrated around β_{MAP} with covariance matrix $\Sigma_\beta = diag\left(\left[\sigma^2_{\beta_1}, \sigma^2_{\beta\_2}, \ldots \sigma^2_{\beta_p}\right]\right)$, where the entities σ^2_β are small compared to $\sigma^{2^{(i)}}$. Therefore, $\Psi(x^{new})\beta^{(i)}$ approaches a normal distribution for large number of predictors regardless of β_i distribution according to central limit theorem. In order to save in computations and storage, we use the following Normal distribution as a surrogate for the predictive distribution:

$$f(y^{new}|x^{new}) = N\left(y^{new};\ \mu_{y^{new}|x^{new}}, \sigma^2_{y^{new}|x^{new}}\right)$$
$$\mu_{y^{new}|x^{new}} = \Psi(x^{new})^T\beta_{MAP}, \quad \sigma^2_{y^{new}|x^{new}} = \|\ \Psi(x^{new})^T\Sigma_\beta\ \|_2^2 + \ \sigma^2_{MAP})$$

where $\|.\|_2$ is the matrix induced norm. Based on this distribution, we calculate the z-score or the equivalent likelihood for the observed value as follows:

$$z_c^{new} = \frac{y^{new} - \ \Phi(x^{new})^T\beta_{MAP}}{\sigma_{y^{new}|x^{new}}},$$
$$L_c^{new} = \log f(y^{new}|x^{new}) = \ const - \log\left(\sigma_{y^{new}|x^{new}}\right) - \ .5\ (z_c^{new})^2$$

Moreover, due to huge complexity of the underlying biological process for each gene and different level of inherit randomness, natural regularity and impact of unknown factors, the predictive power of the learned functions maybe significantly different for each gene.

As an intuitive approach, we consider the average empirical predictability of each gene for normal samples as a ground level for the consistency check. Hence, only cancer samples that their consistency level is far below the average inconsistency of normal samples are reported as inconsistent samples. Therefore, we use the following normalized likelihood

$$LN_c^{new} = \log\left(\frac{f\left(y^{new}|x^{new}\right)}{\left(\sqrt[n_0]{\prod_{i=1}^{n_0} f(y_i|x_i)}\right)^{1-\alpha}\left(\sqrt[n_1]{\prod_{i=1}^{n_1} f(y_i|x_i)}\right)^{\alpha}}\right) \approx \text{cnst} + \ \mathrm{L_c^{new}} - \frac{1-\alpha}{n_0}\sum_{i=1}^{n_0} L_c^i \quad - \frac{\alpha}{n_1}\sum_{i=1}^{n_1} L_c^i$$

Where n_0 and n_1 are the number of normal and cancer samples and α is a tuning parameter between 0 and 1 in order to push different emphasis on normal and cancer cohorts, We usually choose lower values for α in order to emphasis more on the normal cancers and compensate for the lower number of normal samples. In this work, we arbitrarily set $\alpha = \frac{1}{10}$, which is almost equal to ratio of normal samples to cancer samples in the training set for TCGA Breast Cancer dataset. The inequality becomes equality if the variances of the predictive distribution are equal for all samples. The above process is repeated for all genes in parallel. In addition to consistency score, the activation score of each gene is simply obtained using the gene expression level distribution modeled as a normal distribution.

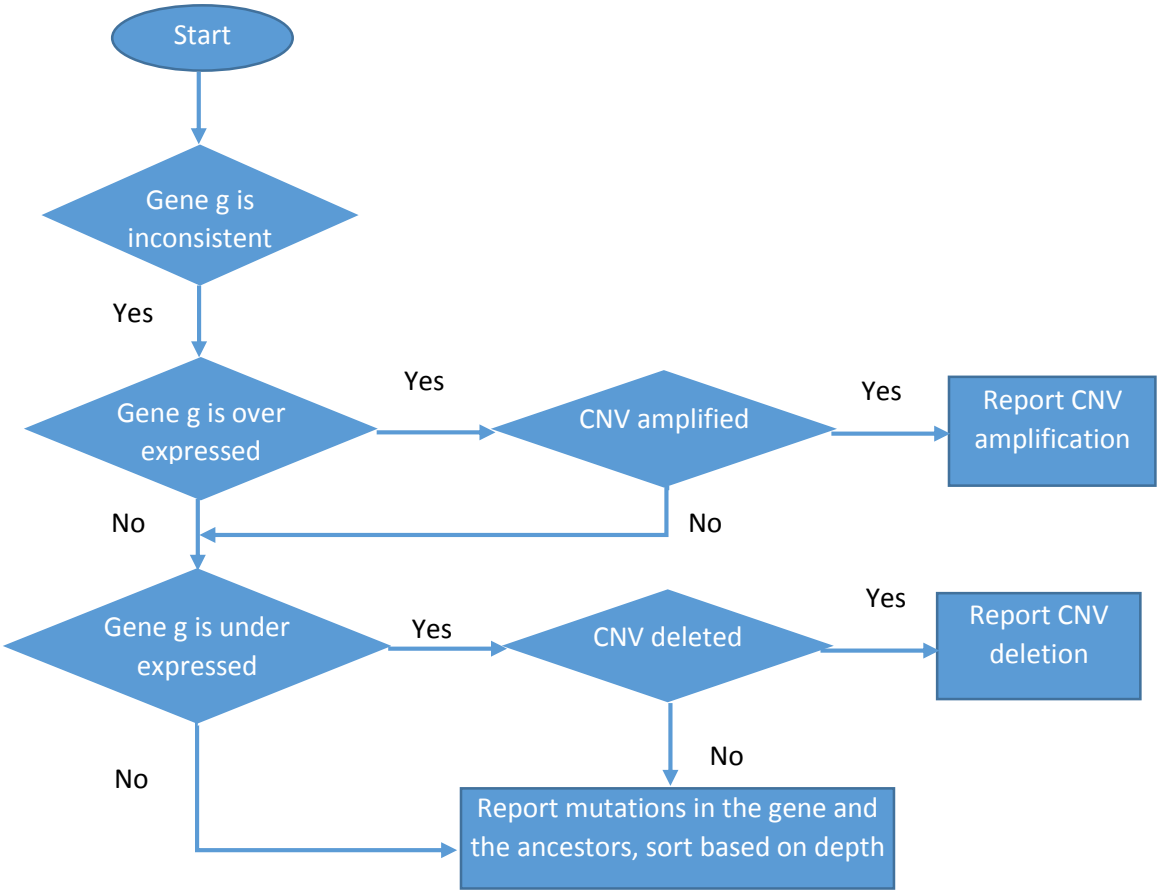
$$y \sim N(\mu, \sigma^2) \Rightarrow \quad z_A^{new} = \frac{y^{new} - \mu}{\sigma}, \qquad L_A^{new} = \log f(y^{new}; \mu, \sigma^2) = \ const - \log(\sigma) - \ .5\ (z_A^{new})^2$$

Where μ and σ is the mean and standard deviation of the normal distribution learned for each gene expression level after iteratively excluding the outliers. The postscript g is omitted for notation convenience and similar normalization is used for activation scores.

Module 4: association between somatic mutations and inconsistencies

This module takes the activation and consistency scores provided by Module 3 for further analysis. For each new test sample, Module 4, identifies the genes that are significantly inconsistent and examines if they are potentially driven by CNV aberrations or somatic mutations in the current gene or in its regulatory subnetwork.

First, we filter out those inconsistencies driven by CNV aberration events. If the inconsistency is due to overexpression of the gene and the gene experiences copy number amplifications (CNV > 0.5) , then CNV amplification is reported as the main cause of the inconsistency. Likewise, if copy number deletion (CNV<-0.5) is associated with the down expression of the gene, we report CNV deletion as the inconsistency driver.
For the genes, which do not experience significant copy number aberrations, we traverse the upstream regulatory network of the gene and report mutated ancestors as potential inconsistency drivers. We assign a global depth penalization parameter $0 < \rho \leq 1$, such that the mutated genes with k hops to the root node are assigned with value ρ^k. This is due to the fact that the closer genes are most likely to drive the inconsistency. As $\rho \to 1$, the impact of depth becomes less significant. This provides a patient-specific candidate drivers for significantly malfunctioning genes. Below is the simple decision-tree/flowchart used in this Module.



Repeating this procedure for all samples, and sorting the genes based on their assigned values filters out the passenger events and determines the most influential parent genes whose mutations functionally impact the downstream transcription factor gene.

D. Performance of Method

In order to validate the accuracy of the proposed model, the gene state success rate of the proposed Bayesian method is compared to several near-optimal state-of-the art point-estimators including LASSO, RIDGE AND Elastic-Net Regressions. To calculate state success rate, we first learn a Gaussian distribution for each gene expression level via

maximum likelihood method after iteratively excluding significant outliers. Briefly, we learn a Gaussian distribution for the samples at each iteration and then we remove the samples which are not in the second standard deviation neighborhood of the mean value. In the subsequent iteration we repeat the process for the remaining samples until the algorithm converges. The empirical distribution for a sample gene *PTEN* and the learned Normal distribution is presented in Fig. 5. We also learn a Student –t distribution for comparison purpose. Student –t distribution has the advantage of robustness to outliers and is very close to the normal distribution after outlier exclusion as shown in Fig. 5.

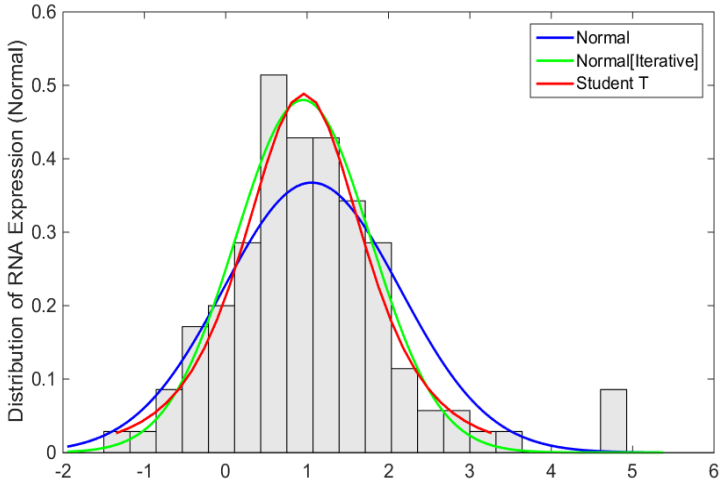


Figure 5. RNA Expression for Gene PTEN. The empirical histogram and learned distributions. The iterative normal distribution excluding outliers and t-distributions are robust to outliers.

Then, we divide gene expression levels into three states (neutral, over-expressed and under-expressed) based on predefined thresholds. We arbitrarily set thresholds such that the probability of down-expression, neutral and over-expression states become 10%, 80% and 10%, respectively .Module 3 provides patient-specific gene expression predictions for all [839] un-isolated genes. The state change rate is calculated via averaging state change events over all genes and patients. The results are calculated for each cohort separately. If the observed and predicted expression state for sample i and gene g are $s^o_{g,i}$ and $s^p_{g,i}$, respectively. The state change rate is calculated as:

$$SER = \frac{1}{|G||S|}\sum_{g=1}^{|G|}\sum_{i=1}^{|S|} I(s^o_{g,i} <> s^p_{g,i})$$

In Table 1, the prediction error is calculated for some important genes that are highly associated with cancer and have a valid set of upstream regulatory genes in the global pathway network. It is seen that the proposed method outperforms the state of the art sparsity-imposing regression models with the additional advantage of providing full posterior distribution for the gene expression level.

Gene	Number of Parents	Normal Cohort				Cancer Samples			
		Lasso	Ridge	Elastic Net	Proposed	Lasso	Ridge	Elastic Net	Proposed
CBFB	57	0.1532	0.1622	0.1441	0.1622	0.1854	0.1909	0.1937	0.1881
CCND1	836	0.2703	0.2883	0.2703	0.2432	0.2919	0.2947	0.2873	0.2614
CDH1	159	0.1802	0.1261	0.1712	0.1622	0.2484	0.2391	0.2456	0.2428
CDKN1B	456	0.2162	0.1892	0.1982	0.1982	0.2994	0.2799	0.2780	0.2530
CTCF	417	0.1261	0.0901	0.1261	0.1171	0.1409	0.1353	0.1474	0.1325
ERBB2	99	0.2252	0.2973	0.0811	0.2523	0.3883	0.4013	0.3818	0.3272
FOXA1	206	0.1712	0.1441	0.2072	0.1261	0.1196	0.1242	0.1242	0.1177
GATA3	223	0.3423	0.2703	0.3333	0.3423	0.1613	0.1529	0.1603	0.1585
MYB	219	0.0901	0.0811	0.0901	0.0901	0.1891	0.1752	0.1900	0.1891
PTEN	226	0.0541	0.0631	0.0721	0.0631	0.1631	0.1585	0.1696	0.1603
RB1	342	0.0721	0.0811	0.0811	0.0901	0.1835	0.1854	0.1724	0.1696
RUNX1	57	0.2162	0.2613	0.2432	0.1982	0.1965	0.1946	0.1993	0.1891
TP53	325	0.2162	0.1441	0.2162	0.1171	0.3309	0.3216	0.3309	0.2956
		0.1795	0.1691	0.1719	**0.1663**	0.2229	0.2195	0.2216	**0.2065**

Table 1- Gene state prediction error rate for the proposed method in comparison with the benchmark optimization-based sparse regression models

Another important observation is that despite the fact of higher contribution of cancer samples to the model due to the larger number of cancer samples with respect to normal samples, the normal cohort presents a better predictability. This observation holds for all models and reveals that the functional states of the gene expression in normal tissues are more consistent with the upstream regulatory network. This observation is also observed in Fig.6, where the observation and prediction values for two genes *PTEN* and *TP53* are presented for normal and cancer samples.

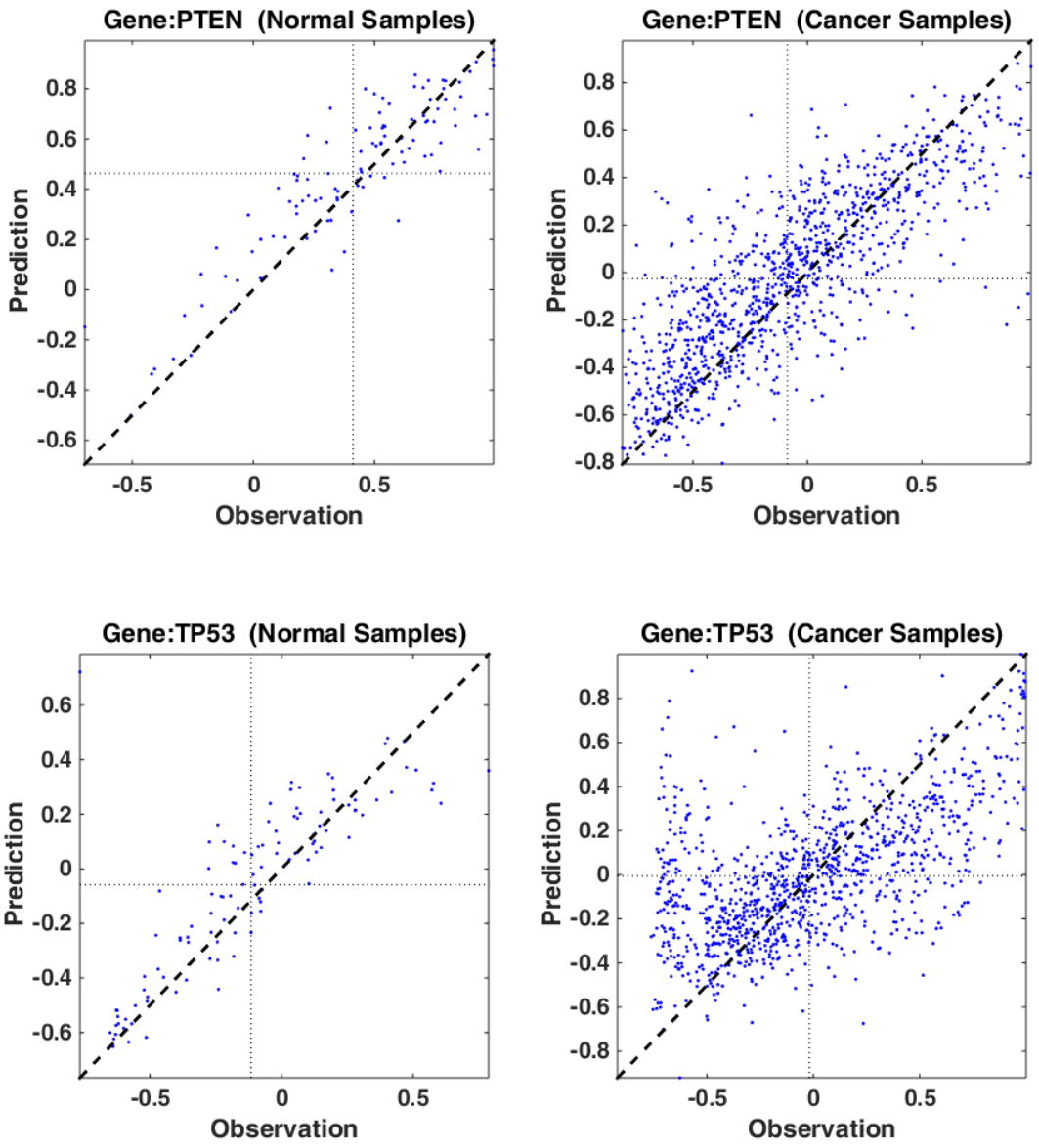


Figure 6. Prediction versus observation for two genes *PTEN* and *TP53*. The gene expression level in normal samples are more consistent with the predictions obtained from the gene self-epigenetic data and its upstream transcription regulation network.

In order to gain more insight about the model coefficients, we present the model parameters obtained for two genes *ERBB2* and *GATA3* in Table 2. Each row presents the corresponding coefficient value obtained by different learning methods and for the proposed non-linear Bayesian method the standard deviation for the posterior distribution is also presented in brackets in the last column. It is shown that the expression level of *ERBB2* is highly dependent on copy number aberration events affecting its locus as seen in the model parameter of the

proposed nonlinear soft-thresholding function. This nonlinearity reflects the ignorance of the model to small turbulences around zero which is likely measurement noise, therefore the copy-number associated logRatio values derived from SNP-arrays can be used directly in the model without the need to discretize the logRatios into amplified/neutral/deleted states. The relevance of the nonlinear function is interestingly picked up by all learning methods. Fig. 7 verifies this relevance, where the relation between the observed RNA as well as the predicted RNA versus CNV is depicted for gene *ERBB2*. This Fig. demonstrates that the nonlinear CNV terms with coefficients obtained from the learning process well define the RNA expression level for ERBB2 with some minor variability due to other terms such as DNA methylation and ancestor gene expression levels. In fact, the coefficients of DNA methylation and majority of ancestors are explicitly removed from the predictor list by LASSO and Elastic Net Methods and also notable is that the proposed model assigns negligible coefficients for DNA methylation.

On the other hand, there RNA expression level for GATA3 is more influenced by DNA methylations as well as upstream regulatory network. The expected negative sign for DNA methylation coefficients are suggestive of an inverse relationship between the gene expression level and DNA methylation for both genes. Finally, for GATA3, the upstream regulatory network plays a crucial role in regulating the expression of this gene, suggesting that most of the variation of this gene's expression in breast cancer arises primarily due to the activity of transcription factors.

The two gene model coefficients reveals that the model coefficients can be significantly different for genes due to high heterogeneity of the gene functionalities.

Model Parameters for (ERBB2)	LASSO	RIDGE	ELASTIC NET	PROPOSED [Std Deviation]
$Methylation_{min}$	0.0000	-0.0057	0.0000	-0.0015: [0.0120]
$Methylation_{max}$	0.0000	-0.0185	0.0000	-0.0027: [0.0134]
$Methylation_{mean}$	0.0000	-0.0215	0.0000	-0.0152: [0.0355]
$f_1(Methylation_{mean})$	0.0000	0.0530	0.0054	0.0220: [0.0247]
$f_2(Methylation_{mean})$	-0.0504	-0.0734	-0.0626	-0.0588: [0.0319]
CNV_{mean}	0.0000	0.3685	0.0181	0.0453: [0.0638]
$f_1(CNV_{mean})$	**-0.0986**	**-0.1579**	**-0.1268**	**-0.1457: [0.0376]**
$f_2(CNV_{mean})$	**0.9958**	**0.6272**	**0.9951**	**0.9858: [0.0524]**
$\sqrt{\sum \lvert\beta_i\rvert^2}$ for ancestors	**0.1232**	**0.2149**	**0.1528**	**0.1663**

(a)

Model Parameters for (GATA3)	LASSO	RIDGE	ELASTIC NET	PROPOSED
$Methylation_{min}$	0.0000	0.0120	0.0000	0.0201[0.0183]
$Methylation_{max}$	-0.0842	-0.0510	-0.0838	-0.0864[0.0201]
$Methylation_{mean}$	0.0000	-0.0450	-0.0142	-0.0327[0.0422]
$f_1(Methylation_{mean})$	**-0.1888**	**-0.0544**	**-0.1667**	**-0.1221[0.0426]**
$f_2(Methylation_{mean})$	0.0000	-0.0342	0.0000	-0.0099[0.0273]
CNV_{mean}	0.0000	-0.0009	0.0000	0.0068[0.0270]
$f_1(CNV_{mean})$	0.0000	0.0015	0.0000	0.0199[0.0247]
$f_2(CNV_{mean})$	0.0000	-0.0033	0.0000	0.0003[0.0232]
$\sqrt{\sum \lvert\beta_i\rvert^2}$ for ancestors	**0.3157**	**0.3329**	**0.3085**	**0.4903**

(b)

Table 2. Model parameters for two genes: *ERBB2* and *GATA3*

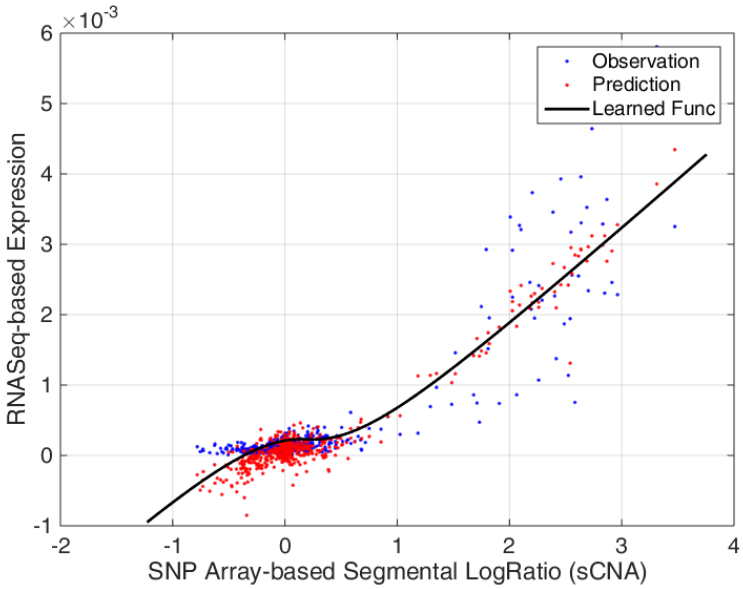


Figure 7. RNA expression level versus copy number variation CNV for gene *ERBB2*. The blue and red points correspond to the observations and the predictions obtained from the model. The black curve is the nonlinear RNA CNV relation obtained by the model parameters in Table 2.

E. Intended use of system and method

The broader area for this invention is oncology diagnostic solutions and services in Clinical Informatics and in Global Services. More specifically this invention is applicable in genomic medicine. This invention can be used in any genome informatics or clinical bioinformatics software or genomics service that is used for clinical research or clinical practice in oncology.

Specifically, our invention has at least two major applications in both the clinical and research settings:

a) The first application is *in-silico* evaluation of the functional effects of novel somatic mutations identified in cancer samples:

   As the cost of whole exome and/or whole genome sequencing continues to fall, the number of sequencing profiles of cancers is expected to continue to rise. With each new study, we find new somatic events affecting genes. Firstly, hyperactivating somatic mutations within oncogenes tend to fall within hotspots along the length of specific genes, it is quite difficult to assess the functional significance of missense mutations within these genes. Secondly, while it is relatively easy to assess the functional significance of loss-of-function mutations such as frame-shift and stop-gain mutations in tumor suppressor genes, it is again difficult to evaluate the functional significance of missense mutations in these genes. Finally, missense mutations in genes whose function haven't been previously characterized are notoriously difficult to ascertain, with standard algorithms such as SIFT and PolyPhen2 only providing estimates of potential deleterious effects of these genes.

   Our invention can more comprehensively characterize the functional effects of these mutations by deciphering their effects on downstream pathway components, thus enabling the identification of more specific functional targets within cancers.

b) The second application is *in-silico* evaluation of the functional effects of novel somatic mutations identified in clinical samples:

In this application, somatic mutations identified in a clinical tumor sample derived from a single patient that are characterized as variants of unknown significance in clinical databases (ClinVar, COSMIC etc) can be prioritized for additional follow-up based on their likely effect on the downstream elements in the pathway. Thus, in a system such as PAPAyA, being developed by Philips, the inclusion of our invention as a module can enable prioritization of somatic events for clinical follow-up. This is particularly important and useful when no clear clinical recommendations exist for that patient based on mutation databases. Such events happen quite often in genomics tumor boards, with over 50% of patients not presenting with previously well-characterized somatic mutation events.

F. References

[1] Kamalakaran S, Varadan V, Giercksky Russnes HE, Levy D, Kendall J, Janevski A et al. *"DNA methylation patterns in luminal breast cancers differ from non-luminal subtypes and can identify relapse risk independent of other clinical variables,"* Molecular Oncology, 5(1):77-92, Feb 2011.

[2] Aravind Subramaniana, Pablo Tamayoa, *"Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles," PNAS 2005,* 102 (43): 15545-15550;

[3] Varadan V, Mittal P, Vaske CJ, Benz SC. *"The integration of biological pathway knowledge in cancer genomics: a review of existing computational approaches,"* IEEE Signal Processing, 29 (1):35-50, 2012.

[4] Fabio V , Eli Upfal , Benjamin J. Raphael , *"Algorithms for Detecting Significantly Mutated Pathways in Cancer,"* Journal of Computational Biology. March 2011, 18(3): 507-522.

[5] Matan Hofree, John P Shen, et al, *"Network-based stratification of tumor mutations,"* Nature Methods, 10 (11): 640 - 646, 2013.

[6] Mohammad Shahrokh Esfahani and Edward R. Dougherty, *"Incorporation of Biological Pathway Knowledge in the Construction of Priors for Optimal Bayesian Classification,"* IEEE/ACM Transactions on Computational Biology and Bioinformatics, 11 (1), Jan 2014.

[7] Charles J. Vaske, Stephen C. Benz, J. Zachary Sanborn, Dent Earl, Christopher Szeto, Jingchun Zhu, David Haussler and Joshua M. Stuart, *"Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM,"* Bioinformatics Journal, 26(12): 237-245, 2010.

[8] W. J. Fu, *"Penalized regressions: the bridge versus the lasso,"* Journal of Computational and Graphical Statistics, 7(3): 397-416, 1998.

[9] Kaare B. Petersen, Michael S. Pedersen, "The Matrix Cookbook", web:http://matrixcookbook.com, November 15, 2012

List of Commercial Entities that may be interested in this invention

1) Agilent technologies
2) GenomOncology
3) Foundation Medicine
4) Philips Healthcare
5) GE Healthcare
6) University Hospitals Case Medical Center
7) Hospital based entities offering CLIA-certified tests for personalized medicine