A graphical model of smoking-induced global instability in lung cancer‡

Yanbo Wang [1], Weikang Qian [2], Bo Yuan [1,†]

[1] Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, 200240, P.R.China
[2] UM-SJTU Joint Institute, Shanghai Jiao Tong University, Shanghai, 200240, P.R.China

Smoking is the major cause of lung cancer and the leading cause of cancer-related death in the world. The most current view about lung cancer is no longer limited to individual genes being mutated by any carcinogenic insults from smoking. Instead, tumorigenesis is a phenotype conferred by many systematic and global alterations, leading to extensive heterogeneity and variation for both the genotypes and phenotypes of individual cancer cells. Thus, strategically it is foremost important to develop a methodology to capture any consistent and global alterations presumably shared by most of the cancerous cells for a given population. This is particularly true that almost all of the data collected from solid cancers (including lung cancers) are usually distant apart over a large span of temporal or even spatial contexts. Here we report a multiple non-Gaussian graphical model to reconstruct the gene interaction network using two previously published gene expression datasets. Our graphical model aims to selectively detect gross structural changes at the level of gene interaction networks. Our methodology is extensively validated, demonstrating good robustness, as well as the selectivity and specificity expected based on our biological insights. In summary, gene regulatory networks are still relatively stable during presumably the early stage of neoplastic transformation. But drastic structural differences can be found between lung cancer and its normal control, including the gain of functional modules for cellular proliferations such as EGFR and PDGFRA, as well as the lost of the important IL6 module, supporting their roles as potential drug targets. Interestingly, our method can also detect early modular changes, with the ALDH3A1 and its associated interactions being strongly implicated as a potential early marker, whose activations appear to alter LCN2 module as well as its interactions with the important TP53-MDM2 circuitry. Our strategy using the graphical model to reconstruct gene interaction work with biologically-inspired constraints exemplifies the importance and beauty of biology in developing any bio-computational approach.

***Index Terms*—Lung cancer, graphical model, alternating direction method of multipliers.**

I. Introduction

Lung cancer is the leading cause of death among all malignancies in the world, most of which ($> 85\%$) are resulted from smoking. Over the years, we become more aware of the fact that cancer including lung cancer is not limited to individual genetic changes but a phenotype possibly conferred by many systematic and global alterations [35]. However, still a number of key biological questions remain: Firstly, do individual lung cancers resulted from smoking share any non-random molecular changes, implicating a smoking-induced cascade? Secondly, even though only about 10-20% of smokers eventually develop lung cancer, it is still important to find out what would be the early changes, mainly at the systems level, that might potentially contribute to the risks of smoking (early markers). Finally, it is unclear why individuals still remain at high risk of developing lung cancer long after their cession of smoking. The overall goal of this paper is therefore to develop a constraint-based searching methodology to capture any significant and systematic changes occurring during the neoplastic transformation of lung. It is our belief that gene interactions instead of individual genes at the global level would be better (and more robust) markers for lung cancers, potentially facilitating our understanding of smoking-induced lung cancer.

Previous attempts have been largely focusing on individual genes by large-scale comparisons [40],[41]. In both human and mouse, a number of important oncogenes and tumor suppressor genes are highly associated with the neoplastic transformation of lung, including ki-Ras, c-Myc, TP53, IL6, IL10, CASP family [11],[27],[42],[45] as well as loss of heterozygosity and change of epigenetics, etc. However, collection of individual events has failed to provide a global and systematic view of lung cancer, both structurally and temporally at the level of gene interaction and regulation.

Using high-throughput technologies, a much larger group of genes are compared for their differential expressions, leading to the notion that tumorigenesis is a systematic problem, affecting many more genes and modules than what we would previously expect. Clinical problems previously unclassified are now known to exhibit substantial molecular subtypes, contributing to new diagnosis and treatment regiments, as well as better differential prognosis. However, much of the computational methodologies previously developed are centered on clustering and classification based problems [14],[16],[17], without taking into account of any simultaneous interactions involving multiple genes. This is important because of the possibility that despite multiple genes and their expressions might be just slightly altered (insignificant in pairwise comparisons) under the assumption of conditional dependence, a large number of genes (a module and/or highly interactive genes involving hubs) might become apparent as a group using a graphical model.

For instance, using genechips, smokers with and without lung cancers were compared for their expression profiles. A set of biomarkers for lung cancer were obtained from the can-

† Corresponding author, boyuan@sjtu.edu.cn
‡ Supported by NNSFC Grants 14Z103010221

cerous samples by unsupervised learning methods. However, these biomarkers obtained were largely based on differential comparisons of individual genes, which inevitably will not include their related genes at the network level. Multiple interacting genes might undergo similar changes but missed due to the lack of statistical power if considered individually. In this paper, we aim to extend the current graphical model approach with a novel structure-based constraint to focus on systematic alternations.

Graphical models is a probabilistic model for which a network structure is used to express the conditional dependencies (i.e., possible biological interactions or regulations) among individual genes [15],[53]. Since the learning of a graphical model with partial observations is ill-posed (large dimensions and small number of samples), regularizations with certain constraints must be used in order to limit the searching space and result. Mathematically, a number of structural priors can be included, such as sparsity, group, and gradual variation [4]. In practice, both individual snapshots and joint models of multiple different stages of cancers were used to model (compare) the progression of tumorigenesis. However, the central issue of applying any graphical model for biological network is how to design appropriate regularization schemes according to biology.

The most commonly used constraint so far assumes network evolution is gradual and local, capturing mainly local and individual structural changes (loss or gain of certain edges), with most of the systems remaining intact. For instances, Guo developed a joint Gaussian graphical model to learn multiple snapshots, assuming their biological networks being only partially changed [19]. Peterson interpreted the joint Gaussian graphical model in the framework of Bayesian [36]. Xing and his colleagues [1] first applied the fused Lasso constraint to model the development of the fruit fly (Drosophila Melanogaster). Recently, Danaher [13] used a similar fused-lasso scheme to integrate multiple stage of tumorigenesis. All of these efforts assumed that most of the network remain largely unchanged while allowing only minor and local variations involving individual edges.

The assumption of gradual and local changes might be valid for very similar cellular stages. However, since almost all of the data collected from cancers (including lung cancers) are usually distant apart over a large span of temporal and even spatial contexts, it is our strong belief that gross structural changes must have already occurred thus any models must first focus on global signatures instead. Our motivation is also supported by the fact that gene expression data are error-prone and noisy, depending heavily on the source of samples [30], as well as experimental conditions.

In addition, our strategy searching for consistent and global changes lies on the recent fact that cancer cells are extremely heterogeneous, diverging very quickly over the progression of tumorigenesis, resulting in many and even random molecular changes that might not be statistically significant when considered as "cancerous markers" [26]. The previous notion of genomic instability being the tumor phenotype is thus well received.

Since we are only going to focus on global changes at the network level, many of the minor changes involving individual genes and their interactions are considered local and insignificant while only global changes are discovered instead (i.e., hubs involving many genes). Though using similar methodology (fused lasso), our assumption is different from the previous works where only minor and local changes were to be selected. Thus, the primary goal of this work is to capture any consistent global (group) changes at the structural level which can be used as the key biological markers (instead of individual genes).

Finally, we want to note the possible disadvantage of assuming a Gaussian distribution for a given gene expression data. It is well known that gene regulations are extremely complex, consisting of many individual subsystems which themselves are interacting among each other, both temporally and spatially, resulting in expression profiles characteristically mixed and complex. Such a complexity has to be approximated with an appropriate distribution. Thus, here we report a non-Gaussian graphical model instead, so that hopefully our model will be more flexible and inclusive for gene expression profiles. Specifically, we provide a fused plus group lasso to constrain our search for only global changes with the multiple D-Loss (1) as our objective function, which does not require the Gaussian assumption.

All of data used in this report are obtained from two previous publications [40],[41]. Our strategy is to develop a novel regularization to detect global changes in lung cancer using the best established datasets available in the field. Specifically, we ascertain if smoking-resulted cancers might exhibit any consistent and presumably cancerous global alternations compared to their normal controls. We chose the dataset [41], consisting of 92 patients with lung cancer and 90 normal controls of smokers. The second dataset [40] was obtained from the same investigators because of our interests to detect any smoking-induced changes (34 cases against 23 controls), potentially as the risk factors. In addition, using the same dataset, we intent to go further to possibly detect any structural changes already mimicking cancerous state for former smokers (18 cases), accounting at least in part for why individuals can still remain at high risk long after their cession of smoking. We also want to note that these two different datasets were from the same investigators, which hopefully would minimize any systematic and other variations.

II. Methodology

A. Notation

We defined our notations as follows. For n dimensional vector $x \in \mathcal{R}^n$, for $q > 0$ we define l_q norm $||x||_q = (\sum_{i=1}^{n} |x_i|^q)^{\frac{1}{q}}$. Here we note l_q norm is a quasi-norm for $0 < q < 1$. $I_p \in \mathcal{R}^{p\times p}$ represents the identity matrix. For rectangular matrix M in $\mathcal{R}^{p\times q}$, the spectral norm $||M||$ is the largest singular value, $||M|| = \sup_{x\in\mathcal{R}^q} \frac{||Mx||}{||x||}$. The Frobenius norm $||M||_F$ is the l_2 norm of singular values, $||M||_F = (tr(M^{'}M))^{\frac{1}{2}}$. The l_∞ norm is defined by $||M||_\infty = \max_{i,j} M_{ij}$. $\langle X, Y\rangle$ means $Tr(XY^T)$. For square matrix $A, B \in \mathcal{R}^{p\times p}$, the dot division is defined as $(A./B)_{ij} = (a_{ij}/b_{ij})$. Let $C \in \mathcal{R}^{n\times s_1}$ and

$F \in \mathcal{R}^{m\times s_2}$, the Kronecker product of $C \otimes F \in \mathcal{R}^{nm\times s_1 s_2}$ is defined as $(C \otimes F)_{ij} = (C_{ij}F)_{ij}$. The $vec(\cdot)$ operator creates a column vector from the matrix A by stacking the column vectors of $A = [\mathbf{a}_1, \mathbf{a}_2 \cdots, \mathbf{a}_n]$ below one another: $vec(A) = [\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_n]^T$.

B. Multiple D-Loss Graphical Model

To model the structural progression of multiple K stage tumorigenesis, we formulate a partial correlation-based multiple Graphical model as $G(V, E^{(i)}), i = 1, 2, \cdots, K$. The node set V represents n individual genes and the edge set $E^{(i)}$ represents possible interactions among all of the V at a snapshot k. The gene expression matrix $X^{(i)} \in \mathcal{R}^{p\times n_i} = (x_1^{(i)}, x_2^{(i)}, \cdots x_p^{(i)})$ consists of n_i observations for p individual genes ($p \gg n$) at a snapshot i. The goal is to model any significant conditional independencies among all of the variables of V. In another word, the searching process is equivalent to obtaining the edges so that $e_{kj}^{(i)} \in E^{(i)}$, if and only if $x_k^{(i)}$ and $x_j^{(i)}$ are conditionally dependent on the remaining variables $x_{V/\{k,j\}}^{(i)}$. Note that these independencies are obtained with all of the remaining components being taken into account. Furthermore, the conditional independence of two nodes $x_k^{(i)}$ and $x_j^{(i)}$ is equivalent to their partial correlation $\rho_{kj}^{(i)}$ being zero, so that

$$e_{kj}^{(i)} \notin E^{(i)} \Leftrightarrow \rho_{kj}^{(i)} = \frac{(\Sigma_i^{-1})_{kj}}{\sqrt{(\Sigma_i^{-1})_{kk}(\Sigma_i^{-1})_{jj}}} = 0.$$

where Σ_i denotes the sample covariance matrix of $X^{(i)}$. It is established that such an equivalence requires that the random variables $X^{(i)}$ be sampled from the families of distributions characterized with a semi-group property [3]. This semi-group condition is rather inclusive, with the typical multivariate Gaussian being one of its examples. In addition, this condition also holds for elliptical, multivariate hyper-geometric, multivariate negative hyper-geometric, multinomial and Dirichlet distributions [3]. Gene expression profiles are typically non-Gaussian with heavy-tailed distributions on both the complete-experiment and the individual-gene level [31]. Since the elliptical family contains numerous multivariate distributions, many of which also exhibit heavy-tails (including the multivariate t-distribution, Gaussian copula distribution, transelliptical distributions [7],[28]), thus our graphical model is inclusive to model gene expressions with heavy tail effects.

We formulate a joint estimator for the precision matrix Ω_i, where $\Omega_i = \Sigma_i^{-1}$ using a multiple D-Loss

$$Loss(\Omega_1, \Omega_2, \cdots, \Omega_K) = \sum_{i=1}^{K}\{\frac{1}{2}\langle\Omega_i^2, \Sigma_i\rangle - Tr(\Omega_i)\}.$$

Here the D-Loss is a novel loss function presumably more suitable to estimate a sparse precision matrix. The D-Loss was originally developed by Zhou and his colleagues, which also possesses the nice irrepresentable condition hence requiring fewer number of samples while achieving similar statistical power [54]. Note that our multiple loss $Loss(\Omega_1, \Omega_2, \cdots, \Omega_K)$ is a smooth and convex function for Ω_i with a global minimum if and only if $\Omega_i = \Sigma_i$. Our proof is straightforward and similar to the techniques used in [54].

C. Global Variation Lasso (GV-Lasso)

To join multiple graphical models together, we design a novel regularization scheme named global variation Lasso(GV-Lasso). The basic idea of our GV-Lasso is to decompose the network structure into two parts (Figure 1): the known and stable interactions (possible background signals); and the other interactions in modular structures (possible modular structures to be focused). The motivation is to control the possible background effects from our observations of interests.

We consider the known (prior) and the conserved interactions as our possible sources of background signals. Firstly, we denote the known interactions (our structural prior) as D_i, an $\mathcal{R}^{p\times p}$ matrix projected according to the adjacency matrix of given prior $\mathcal{GI}$ (with other entries to be zeros). In this paper, we use no prior information that D_i is a diagonal matrix. In the Appendix C, we show an example incorporating a prior extracted from literatures. This option is potentially very useful as it is possible in the future to include any priors, which could reduce search feature space and facilitate its convergence.

Secondly, we denote the conserved interactions as the Z_i, which is extracted by a fused penalty between two different networks to be compared. We apply a fused lasso to constrain the structural similarities between Z_i and Z_j, with only minor and local variations being allowed. Biologically, certain gene interactions are conserved over the course of evolution (housekeeping genes and/or interlogs), which are supposed to remain unchanged over tumorigenesis. In addition, we can control the stringency of similarity of Z_i for $i = 1, 2\cdots K$ over the course of a progression.Since the progression of the Z_i is largely considered as the background noise or at most some minor and local alternations, our method will likely detect consistent and especially gross changes (see below). This strategy is particularly relevant with data collected from human tissues, to account for the heterogeneous nature of neoplastic transformation.

Central to our methodology is to focus on global variation using a group-lasso to capture only modular changes, such as one-to-many gene or one-to-key module interactions, implicated as the possible changes occurring on major regulatory apparatuses. To do so, we use $U_i + U_i^T$ to assure the symmetry for all of the interactions in the U_i. We restrict the U_i as a group sparse matrix, i.e., a row sparse matrix.

Biologically, the s-th row of U_i denotes the gene x_s might interact with a group of genes at a snapshot i. Since the group-lasso will either select a row as a whole or dismiss a row completely, the non-zero selections involving many genes might be the good markers for global variation (formation of new hubs). In contrast, the zero selection would suggest a possible loss of global structures (loss of original hubs). Mathematically, we formulate our multiple D-loss graphical

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2016.2599867, IEEE/ACM Transactions on Computational Biology and Bioinformatics

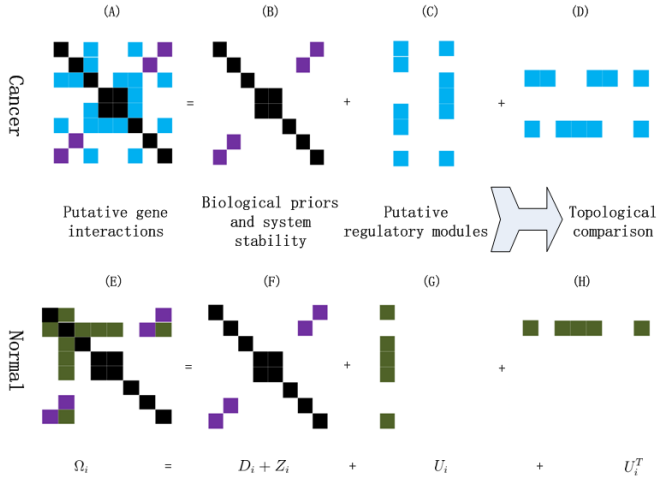

Fig. 1. A schematic view of our regularization. A putative gene interaction network Ω_i can be learned and decomposed through our GV-Lasso ((A) and (E)), constrained by a combinatory regularization consisting of a fused penalty ((B) and (F)) plus a group-based penalty to enhance a global sparsity involving individual groups ((D) and (H)). Note the (C) and (G) are the transposes of (D) and (H), respectively, a transformation to guarantee the symmetry of (A) and (E). The black color in (B) and (F) denotes the biological prior, which do not have to be identical in cancers and normal controls. Note also the prior can be readily included as a projection of any adjacency matrix, according to biology. The purple color denotes the portion of a putative gene interaction network, which is supposed to undergo only local or other minor changes (systems stability). The blue and green colors represent the groups (arrays) of one-to-many interactions (putative regulatory modules).

model regularized with GV-Lasso as follows:

$$\begin{aligned}\hat{\Omega}_i &= \arg\min_{\Omega_i \succ 0} \sum_{i=1}^{K} (\tfrac{1}{2}\langle \Omega_i^2, \Sigma_i\rangle - Tr(\Omega_i)) + \sum_{i=1}^{K} \alpha_1 |Z_i|_1 \\ &+ \sum_{i=1}^{K} \beta_1 |U_i|_1 + \sum_{i=1}^{K} \beta_2 \sum_{k=1}^{p} \sqrt{\sum_{j=1}^{p} ((U_i)_{kj})^2} \\ &+ \sum_{i=2}^{K} \alpha_2 |Z_i - Z_{i-1}|_1, \\ \text{subject to } &\begin{matrix} \Omega_i & = Z_i + D_i + V_i + U_i, \\ V_i & = U_i^T. \end{matrix}\end{aligned} \tag{1}$$

Here $\alpha_1, \alpha_2, \beta_1, \beta_2$ are all tuning parameters. $(U_i)_{kj}$ represents the kj-th entries of matrix U_i. Our regularization scheme (1) is an extension to the node based regularization multiple Gaussian graphical model originally developed by Mohan [33] as follows:

$$\begin{aligned}\hat{\Omega}_i &= \arg\min_{\Omega_i \succ 0} \sum_{i=1}^{K} (-\log\det(\Omega_i) + Tr(\Sigma_i \Omega_i)) \\ &+ \sum_{i=1}^{p} \alpha_1 |\Omega_i|_1 + \sum_{i=1}^{K} \beta_1 \sum_{k=1}^{p} \sqrt{\sum_{j=1}^{p} ((U_i)_{kj})^2}, \\ \text{subject to } &\ \Omega_i - diag(\Omega_i) = U_i + U_i^T.\end{aligned} \tag{2}$$

The differences between our GV-Lasso (1) and the classical node-based Gaussian graphical model (2) are:

(1) Known Biological prior D_i can be readily included in (1) compared to the restriction where only diagonal adjustments are allowed in (2);

(2) The relative invariance Z_i can be adjusted according to the stringency desired using the fused penalty (your assumption about how similar you would expect for the stages over the progression to be modeled). The strength of the background effect (the similarity among the Z_i) can be controlled by our tuning parameter α_2 (1), depending on if and how much the background signal is to be considered (α_2 being small for strong background effect and vice and versa). Alternatively, if we want to focus only on global and consistent variations, this background effects can be considered as heterogeneous noises (genomic instability), thus purposely minimized (to increase α_2). (2) does not have this important option;

(3) Once again, the use of multiple D-loss allow us to use less and more heterogeneous data, particularly suitable for gene expression data derived from human tissues. In contract, (2) assumes that the observations are sampled from a multivariate Gaussian distribution, thus not appropriate for heterogeneous gene expressions;

(4) Computationally, (1) is easier to implement with the alternating direction method of multipliers (ADMM). The derivatives of U_i and V_i in (1) can be explicitly obtained.

D. Choice of parameters

We select the tuning parameters for our GV-Lasso model by an approximation of the commonly-used Bayesian Information Criterion (BIC),

$$BIC = \sum_{i=1}^{K} \{\frac{1}{2}\langle \Omega_i^2, \Sigma_i\rangle - Tr(\Omega_i) + E_i \log n_i\}.$$

where E_i is the numbers of nonzero entries in Ω_i. Since the BIC doesn't have a closed form with respect to the $\alpha_1, \alpha_2, \beta_1, \beta_2$, a grid-based screen for the entire space of $\mathcal{R}^4$. We justify this assumption as we treat each of our snaps equally in terms of their structural configurations (sparsity, modularity, background effect, etc). Thus any variations observed would be something possibly intrinsic (based on the same controls). This is in fact the most compute-intensive part of our approach. We use BIC because it is intrinsically more likely to identify the 'true' model while being asymptotically consistent. Other measures such as the Akaike information criterion (AIC) were not used because of their higher risks of overfitting [39].

E. Algorithm

We develop an alternating direction method of multipliers (ADMM)[6] to implement our GV-Lasso: We write the augmented Lagrange for (1), consisting both the objective function and penalties as:

$$\begin{aligned}\mathcal{L} &= \sum_{i=1}^{K} (\tfrac{1}{2}\langle \Omega_i^2, \Sigma_i\rangle - Tr(\Omega_i)) + \sum_{i=1}^{K} t_i^T (\Omega_i - Z_i - D_i - V_i \\ &- U_i) + \sum_{i=1}^{K} r_i^T (V_i - U_i^T) + \sum_{i=1}^{K} \tfrac{\rho_1}{2} ||\Omega_i - Z_i - D_i - V_i - \\ &U_i||_F^2 + \sum_{i=1}^{K} \tfrac{\rho_2}{2} ||V_i - U_i^T||_F^2 + \sum_{i=2}^{K} \alpha_2 |Z_i - Z_{i-1}|_1 + \sum_{i=1}^{K} \alpha_1 |Z_i|_1 \\ &+ \sum_{i=1}^{K} \beta_2 \sum_{k=1}^{p} \sqrt{\sum_{j=1}^{p} ((U_i)_{kj})^2} + \sum_{i=1}^{K} \beta_1 |U_i|_1.\end{aligned}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2016.2599867, IEEE/ACM Transactions on Computational Biology and Bioinformatics

Algorithm 1: ADMM

initialize $\{\Omega_i, Z_i, D_i, U_i, V_i, T_i, R_i\}_{i=1:K}$;
parameters $\rho_1, \rho_2, \alpha_1, \alpha_2, \beta_1, \beta_2$;
repeat *for* $s = 1, 2 \cdots$

$$Z_{i=1:K}^{(s+1)} = \arg\min_{Z_{i=1:K}} \sum_{i=1}^{K} \frac{\rho_1}{2} ||X_i^{(s)} - Z_i - D_i^{(s)} - V_i^{(s)} - U_i^{(s)} + T_i^{(s)}||_F^2 + \sum_{i=1}^{K-1} \alpha_2 |Z_i - Z_{i-1}|_1 + \sum_{i=1}^{K} \alpha_1 |Z_i|_1; \quad (3)$$

for $i = 1, 2 \cdots K$

$$\Omega_i^{(s+1)} = \arg\min_{\Omega_i \succeq 0} \frac{1}{2}\langle \Omega_i^2, \Sigma_i \rangle - Tr(\Omega_i) + \frac{\rho_1}{2} ||\Omega_i - Z_i^{(s+1)} - D_i^{(s)} - V_i^{(s)} - U_i^{(s)} + T_i^{(s)}||_F^2; \quad (4)$$

$$D_i^{(s+1)} = \arg\min_{D_i} ||\Omega_i^{(s+1)} - Z_i^{(s+1)} - D_i - V_i^{(s)} - U_i^{(s)} + T_i^{(s)}||_F^2; \quad (5)$$

$$U_i^{(s+1)} = \arg\min_{U_i} \frac{\rho_1}{2} ||\Omega_i^{(s+1)} - Z_i^{(s+1)} - D_i^{(s+1)} - U_i - V_i^{(s)} + T_i^{(s)}||_F^2 + \frac{\rho_2}{2} ||V_i^{(s)} - U_i{}^T + R_i^{(s)}||_F^2 + \sum_{i=1}^{K} \beta_2 \sum_{k=1}^{p} \sqrt{\sum_{j=1}^{p} ((U_i)_{kj})^2} + \sum_{i=1}^{K} \beta_1 |U_i|_1; \quad (6)$$

$$V_i^{(s+1)} = \arg\min_{V_i} \frac{\rho_1}{2} ||\Omega_i^{(s+1)} - Z_i^{(s+1)} - D_i^{(s+1)} - V_i - U_i^{(s+1)} + T_i^{(s)}||_F^2 + \frac{\rho_2}{2} ||V_i - U_i^{(s+1)^T} + R_i^{(s)}||_F^2; \quad (7)$$

$$T_i^{(s+1)} = T_i^{(s)} + (\Omega_i^{(s+1)} - Z_i^{(s+1)} - D_i^{(s+1)} - V_i^{(s+1)} - U_i^{(s+1)});$$
$$R_i^{(s+1)} = R_i^{(s)} + (V_i^{(s+1)} - U_i^{(s+1)^T});$$

end;
until *convergence*;
Return $\Omega_i = Z_i + D_i + U_i^T + U_i$

corresponding to the dual variables t_i and r_i. We rescale the two dual variables T_i and R_i, as $T_i = \frac{t_i}{\rho_1}$ and $R_i = \frac{r_i}{\rho_2}$. We apply the ADMM discretization method to optimize the Lagrange as below (Algorithm 1):

We divide the algorithm into 5 consecutive loops. First of all, we use the fused penalty for (3) to purposely select for any invariant structures between our compared networks. The computational complexity is $\mathcal{O}(K \log K)$ for (3), very much acceptable for the number of stages to be observed [23].

Specifically, we compute the proximity operator as the solution of the following fused penalized problem

$$FLasso(z_1, z_2, \cdots, z_N) = \arg\min_{z_1, z_2, \cdots, z_N} \frac{1}{2} \sum_{k=1}^{N} ||z_k - a_k||^2 + \sum_{k=2}^{N} \tau |z_k - z_{k-1}|_1,$$

with $z_k, a_k \in \mathcal{R}$. Here the τ is a tuning parameter.

We introduce the dual problem of (8)

$$(u_1, u_2, \cdots, u_{N-1}) = \arg\min_{u_1, u_2, \cdots, u_{N-1} \in \mathcal{R}} \frac{1}{2} \sum_{k=1}^{N} ||a_k - u_k + u_{k-1}||^2 \quad \text{subject to } |u_k|_1 \leq \tau, \forall k = 1, 2, \cdots, N-1, \text{ and } u_0 = u_N = 0. \quad (9)$$

An equivalence between the primal and the dual solutions is $z_k = a_k - u_k + u_{k-1}$ with $k = 1, 2 \cdots, N$. The deviation from (8) to (9) will be given in Appendix A. The Karush-Kuhn-Tucker condition is

$$u_0 = u_N = 0, \text{ and } \forall k = 1, 2 \cdots N-1, \begin{cases} u_k \in [-\tau, \tau], \text{ if } x_k = x_{k+1}, \\ u_k = -\tau, \text{ if } x_k < x_{k+1}, \\ u_k = \tau, \text{ if } x_k > x_{k+1}, \end{cases} \quad (10)$$

We note a special case with $K = 2$ (Appendix B). We obtain an analytical solution leading to $Z_i^{(s+1)} = Lasso(A_i, \alpha_1)$ for $i = 1, 2$, with $Lasso(x, c) = sign(|x| - c)_+$[4], where

$$(A_1, A_2)_{hj} = FLasso((A_1^*)_{hj}, (A_2^*)_{hj}, \frac{\alpha_2}{\rho_1}) \begin{cases} ((A_1^*)_{hj} - \frac{\alpha_2}{\rho_1}, (A_2^*)_{hj} + \frac{\alpha_2}{\rho_1}), \text{ if } (A_1^*)_{hj} - (A_2^*)_{hj} > 2\frac{\alpha_2}{\rho_1}, \\ (\frac{(A_1^*)_{hj} + (A_2^*)_{hj}}{2}, \frac{(A_1^*)_{hj} + (A_2^*)_{hj}}{2}), \text{ if } |(A_1^*)_{hj} - (A_2^*)_{hj}| < 2\frac{\alpha_2}{\rho_1}, \\ ((A_1^*)_{hj} + \frac{\alpha_2}{\rho_1}, (A_2^*)_{hj} - \frac{\alpha_2}{\rho_1}), \text{ if } (A_1^*)_{hj} - (A_2^*)_{hj} < -2\frac{\alpha_2}{\rho_1}, \end{cases}$$

and $A_i^* = \Omega_i^{(s)} - D_i^{(s)} - V_i^{(s)} - U_i^{(s)} + T_i^{(s)}$, for $i = 1, 2$. For general case ($K > 2$), we borrow a linear time screening method to solve equation (9) which was proposed by [10].

And then, we solve (4) based on the following theorem:

Theorem 1: Supposing Y is a symmetry matrix and Φ is a given semi-definite matrix with an eigen-decomposition written as $\Phi = Q\Lambda Q^T$. We assume the parameter κ be a constant and M a given symmetry matrix. Accordingly, we claim that the following matrix equation

$$Y\Lambda + \Lambda Y + \kappa Y - M = 0. \quad (11)$$

has an unique solution $Y = QSQ^T$, where the symmetry matrix S satisfies

$$(S)_{hj} = \frac{(Q^T M Q)_{hj}}{\Lambda_{jj} + \Lambda_{hh} + \kappa}. \quad (12)$$

Proof: We let $S = Q^T Y Q$, then (11) will be equivalent to

$$S\Lambda + \Lambda S + \kappa S - Q^T M Q = 0.$$

Since Λ is a diagonal matrix, we have

$$S_{hj}\Lambda_{jj} + \Lambda_{hh} S_{hj} + \kappa S_{hj} - (Q^T M Q)_{hj} = 0, \quad (13)$$

We can thus directly obtain the equation (12) as a tensor form

$$Y = Q(Q^T M Q./(vec(\Lambda)^T \otimes \vec{1} + \vec{1}^T \otimes vec(\Lambda) + 2\rho_1 \vec{1}^T \otimes \vec{1}))Q^T.$$

To optimize (4), we first take a derivative on (4) and obtain

$$\frac{1}{2}(\Omega_i \Sigma_i + \Sigma_i \Omega_i) - I + \rho_1(\Omega_i - Z_i^{(s+1)} - U_i^{(s)} - V_i^{(s)} - D_i^{(s)} + T_i^{(s)})) = 0. \quad (8)$$

Let $\kappa = 2\rho_1$, $M = 2(I + \rho_1(Z_i^{(s+1)} + U_i^{(s)} + V_i^{(s)} + D_i^{(s)} - T_i^{(s)}))$, directly leading a closed form solution.

Next, we solve (5) by projecting the pattern matrix D_i into an adjacency of $\mathcal{GI}$, which encodes the biological prior. Hence

$$(D_i)_{hj} = \begin{cases} (\Omega_i^{(s+1)} - Z_i^{(s+1)} - D_i - V_i^{(s)} - U_i^{(s)} + T_i^{(s)})_{hj} \\ \quad \text{if } \mathcal{GI}_{hj} \neq 0 \\ 0, \text{ else.} \end{cases}$$

We use the group lasso [52] to enhance our search for any gross topological changes between our compared networks (6). We define

$$U_i^* = GLasso(\frac{\rho_1(\Omega_i - Z_i - D_i - V_i + T_i) + \rho_2(V_i^T + R_i^T)}{\rho_1 + \rho_2}, \beta_2),$$

with $GLasso(x, c) = (1 - \frac{c}{||x||_2})_+ x$. Thus, we have $U_i = Lasso(U_i^*, \beta_1)$.

Finally, we solve(7) by taking a derivative, resulting in

$$V_i = \frac{\rho_1(\Omega_i - Z_i - D_i - U_i + T_i) + \rho_2(U_i^T - R_i)}{\rho_1 + \rho_2}.$$

We note here, the convergence property of Algorithm 1 can be derived from the convergence theory of ADMM [20].

F. Network topology analysis

All of the global and local parameters describing network topology according to [5] were calculated using the network analysis tool box in matlab.

We obtain our "hubs" directly from our decomposition of $\Omega_i = Z_i + D_i + U_i + U_i^T$, thus the "hubs" is U_i, or more specifically the rows of U_i, each representing "one to many interaction" for a given hub.

III. Result and Discussion

A. Validation of Our Methodology

Synthetic Data

Our method makes three extensions of existing efforts:

1) Gaussian distribution not to be assumed for gene expression (to model any even mixture of distributions);
2) Adjustable instead of remaining identical for the invariant part to be regularized (to account to any noisy, minor, and/or local changes);
3) Biological prior to be included (to facilitate convergence).

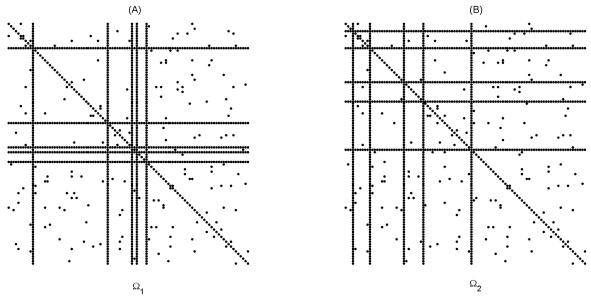

Fig. 2. The sparsity patterns of Ω_1 and Ω_2 , synthetic data as the standard for the accuracy of recovery. Note the random selection for group sparsity (rows and columns).

To validate our methodology, we compare ours with two other classical examples:

1) Two individual node-based Gaussian graphical models (NBGGM) as two independent snapshots (non-fused) [44];
2) Node-based joint Gaussian graphical model (NBJGGM) assuming absolute invariance between two individual graphs [33];

We generated two Gaussian distributions as $\mathcal{N}_1(0, \Sigma_1)$ and $\mathcal{N}_2(0, \Sigma_2)$, in which the $\Sigma_i = \Omega_i^{-1}$,for $i = 1, 2$. The precision matrices Ω_i are obtained by the following three steps (Figure 2): (a) The invariant components Ω_0 in Ω_1 and Ω_2 are constructed as $\Omega_0 = \hat{\Omega}_0 + \hat{\Omega}_0^T$, where $\hat{\Omega}_0$ contains 10% nonzero entries which are uniformly distributed in $\mathcal{R}^{100\times100}$. Each of the non-zeros entries in $\hat{\Omega}_0$ is generated according to a Gaussian distribution $\mathcal{N}(0, \sigma^2)$ (we take $\sigma = \sqrt{2}$); (b) The variable components V_i (for $i = 1, 2$) is each generated using a 100×100 zero matrix, in which 5 randomly selected rows are each replaced with a vector sampled from an uniform distribution of $\mathcal{U}(0, 1)$; (c) We let $\Omega_1 = \Omega_0 + V_1 + V_1^T$ and $\Omega_2 = \Omega_0 + V_2 + V_2^T$. Here, in order to guarantee the positive definiteness of Ω_i (for $i = 1, 2$), we iterate the matrix $\Omega_i = \Omega_i + 1.1I_{100}$ until the Ω_i's smallest eigen-value is greater than zero.

To assess the consistence of model selection for our methods, we use the Matthews correlation coefficient($MCC \in [-1, 1]$) to quantify recovery merits, where

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}},$$

with TP, TN, FP, FN denote the numbers of true positives, true negatives, false positives, and false negatives, respectively. Figure 3 depicts the MCC over the typical range for two of the most important parameters used in our model, the local sparsity (method's selectivity), and the group sparsity (method's specificity). The assessments were performed using synthetic data, purposely reflecting the degree of ill-posed nature of gene expression data (large p small n). Note the use of $\frac{1}{2}$ (sample vs. node numbers) fits roughly appropriate to the actual data we are going to use in this study, in which about a thousand nodes (genes) are independently observed for about few hundred individual experiments. We attribute the irrepresent condition [55] as the possible major factor influencing our method, using the multiple D-loss as the objective function.

Cross-Validation Given the ill-posed nature of using graphical models for gene expression data (large p small n), one of the foremost important criteria is to demonstrate that our methodology be robust, insensitive to issues such as insufficient observations, noise, or even missing data. This is particularly relevant to this investigation because cancers including lung cancers are composed of heterogeneous cells, each of which could have some very different gene expression profiles, inevitably resulting in multiple individual distributions.

One of the specific aims of our effort is to ascertain if neoplastic transformation can still be largely described as a certain cascade, diverging but still sharing some key events. Mathematically, it is our strategy to hence treat any local

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2016.2599867, IEEE/ACM Transactions on Computational Biology and Bioinformatics

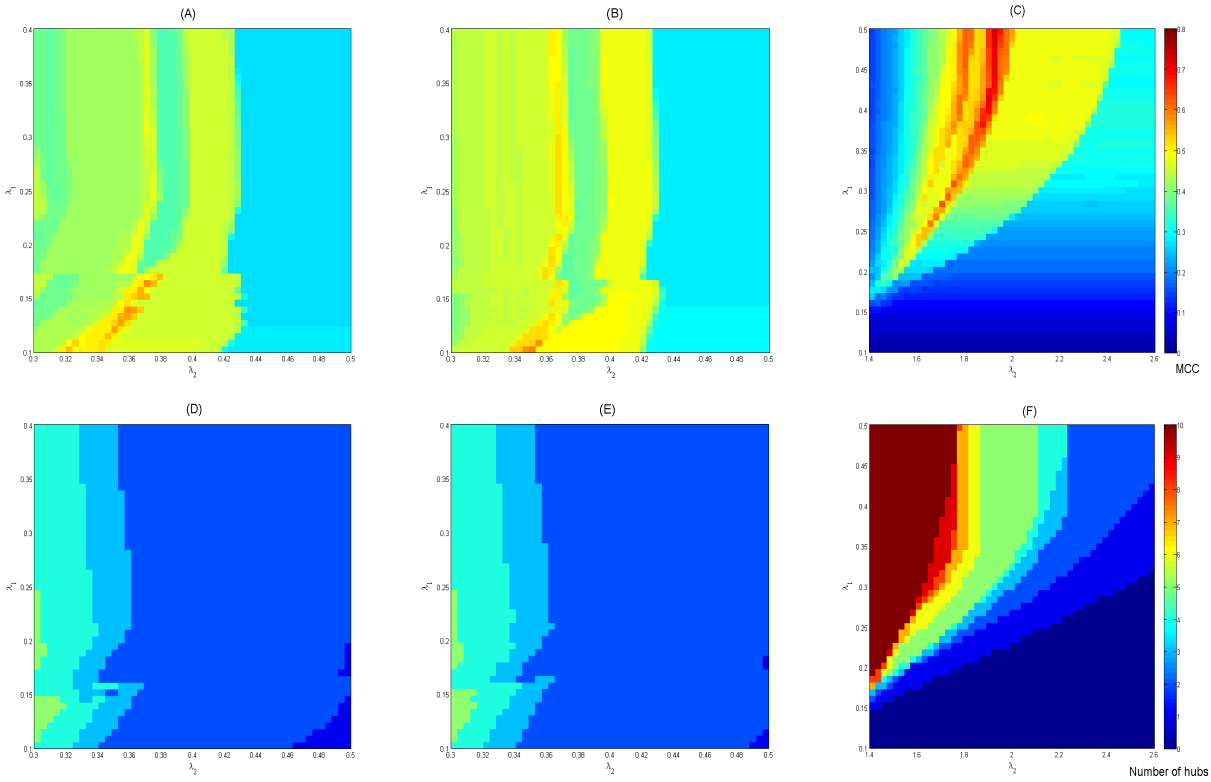

Fig. 3. The assessment of recovery accuracy based on MCC rate. Panels A-C are the MCC rates obtained using the identical synthetic data, with A: NBGGM; B: NBJGGM; and C: our GV-Lasso; Panels D-F are the projection to the identical ranges for the two original parameters assessed in Panels A-C. The purpose of the projection is to validate three methods under the identical conditions (here the number of the hubs is actually used as a latent variable). The X axis represents the group sparsity, Y the local sparsity. The intensity of the color for A-C indicates the MCC accuracy; D-F the number of recovered hubs. Demonstrated here is the supremacy of our GV-Lasso, particularly efficient for large p small n problems (such as gene expression data). .

and diverging events as noises while hopefully capture some common events shared by most (if not all) of the cells for a given tumor tissue sample. Since our method is purposely designed to focus on gross structural variation (aimed to detect consistent biological changes) while ignoring minor and local alternations, its robustness must first be absolutely validated (not detecting any structures which are random or data-dependent).

As demonstrated in the Figure 4, our method does possess the good reproducibility regardless of the data used. Note that we use the likelihood over the number of edges as a global criterion for structural similarity. Certainly it is still possible that some of the local variations might not be specifically detected, which will not significantly affect our strategy of treating any sporadic variations as noises. In contrast, we argue that the classical Gaussian graphical model (such as the cases in Figure 3, A and B and Figure 3, D and E) can become error-prone especially when a small portion of the data is selected (hard to fit any good Gaussian distribution, owning to insufficient samples to satisfy the central limit theorem).

Global-Variation Since we are going to compare and model various structural variations occurring during and even before tumorigenesis, we need to further validate the specificity and selectivity of our methodology. Ideally, such a methodology would be able to reveal smoking-induced neoplastic transformation as a multistage process, during which only minor and benign changes be initially observed from early exposure while more gross and major alterations be detected

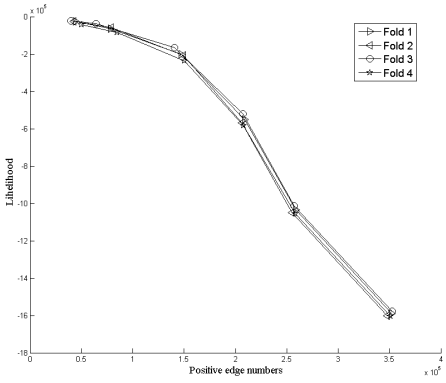

Fig. 4. Cross-Validation using data divided randomly into 4 folds (shown as 4 different labels: circle, up triangle, down triangle, and star). We use the random fractions of the original gene expression data consisting of 97 cancers plus 90 normal controls. The reproducibility (robustness) of our GV-Lasso methodology is revealed by the consistent number of edges over their likelihood regardless of the fraction of the data employed.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2016.2599867, IEEE/ACM Transactions on Computational Biology and Bioinformatics

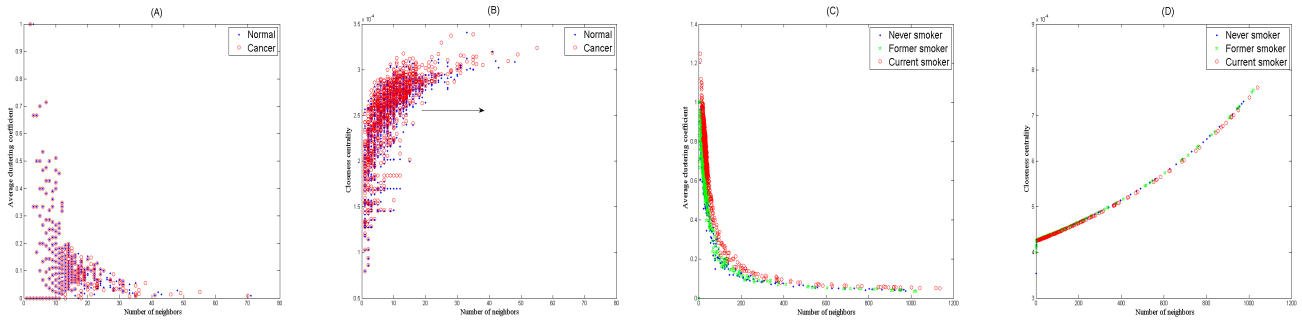

Fig. 5. Global structural variation is cancer-specific. To globally assess the degree of gross structural changes occurring during or before lung tumorigenesis, we use two statistical criteria, the average clustering coefficient (pairwise affinity) and the closeness centrality (distance from the center) to describe the local topology with respect to each of the nodes (genes). Panels A and B demonstrate the structural differences between the lung cancers and normal lung controls (both smokers, thus the control for cancer-specific changes). The arrowheads in (A) and (B) indicate the hubs (larger than which) are more likely to undergo gross structural variations (note the phenomena being more apparent in (B)). In contrast, Panels C and D reveal no significant gross alterations from the benign epithelial cells exposed to smoking, including the cells isolated from those who have long ceased smoking before this investigation. The cells from the individuals who never smoke are used as the control for smoking.

in cancerous cells. Gross topological changes between cancer and normal become apparent, especially among the hubs with connections roughly larger than 10-20 (Figure 5B). Using a different statistical criterion the average clustering coefficient, however, similar observations appear not as distinguishable. Interestingly, similar changes are not detected among any of the normal individuals, exposed to smoking, currently or formally (Figure 5, C and D). These preliminary observations establish the foundation of our GV-Lasso, which can be used to specifically scan for cancer-related gross changes while selectively ignoring minor or local alternations, presumably accumulating during the early stage of smoking-induced tumorigenesis.

We recommend our GV-Lasso to be generally applicable to scan for gross structural changes in any network evolution problems. Mathematically, we use a convex optimization (GV-Lasso) to make a threshoulding of signals in multi-scales. Specifically, our strategy captures gross variations by detecting significant changes within groups using a scale different from the criterion used when evaluating local similarities. Such a use of different scales enhances the global signatures with local variations being treated simply as noises.

B. Smoking-induced cancer cascade (common and shared events)

Having validated our method using both synthetic and real data, we next want to pursue our scan using the gene expression data (Affymetrix Human U133A Array) publically available from 97 lung cancer patients with 90 normal controls (GDS2771). Here both of the cancers and controls are collected from individuals who have smoked, with smoking being implicated as the predominant etiological factor leading to their cancers. We select 1185 genes whose expressions show significantly larger variance across both the cancer samples and controls. We apply our GV-Lasso, with the parameters setting and convergence as showed in Figure 8. The overall goal of this comparison is to capture any significant gross structural changes associated with tumorigenesis.

We purposely focus on global changes because tumorigenesis is a very heterogeneous and complex process, diverging into individual cells harboring possible different mutations over time. It is thus desirable to pay our first attentions to the more consistent and global events, thus our strategy to focus on modular changes, including the gain or loss of hubs, or hub transient as first proposed by Gerstein's group [29]. Also, it is conceivable that highly-connected hubs are generally more influential than leaves in a gene interaction network.

Lung cancers appear to undergo significant gross topological changes, involving many important genes, previously implicated in the neoplastic transformation of lung (Figures 6 and 7). We detect a total of 11.2% of the nodes (133 out of 1185) as hubs (non-zeros rows in U_i) between the lung cancers and their controls. We consider these changes consistent (or common) because a mixture of multiple individual cancers is used as the subject. Of the 133 hubs selected, 13 happened to be the ones consistent with the 36 candidate cancerous genes in lung cancers (based on literatures for their associations with lung cancers [21], [48], see the details in the Appendix C). Under the null hypothesis that these priors would be included randomly, the p-value is 4.6860e-05, strongly suggesting that our method can discover the biology consistent with literature.

Cancer network appears to gain more hubs compared to their controls. A total of 117 hubs are either newly formed or significantly expanded while losing about 12. In contrast, normal controls only have about 16 hubs (ALDH1A1, CALM1, CDK4, EEF1A1, EIF1, EIF4H, FGFR1, FOXA2, HINT1, HIPK1, IL6, PERP, PPP2CB, PTGES3, TM9SF2, TMCO1). One of the possible mechanisms by which such extensive topological changes can occur would be the "hub-transient" model proposed by Gerstein and his colleagues [29], that regulatory genes can switch their interaction partners during the progression of tumorigenesis, involving many new hub formations or alternations (Figure 7, C and D). Besides, nearly all of the hubs in the normal controls are drastically changed during

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2016.2599867, IEEE/ACM Transactions on Computational Biology and Bioinformatics

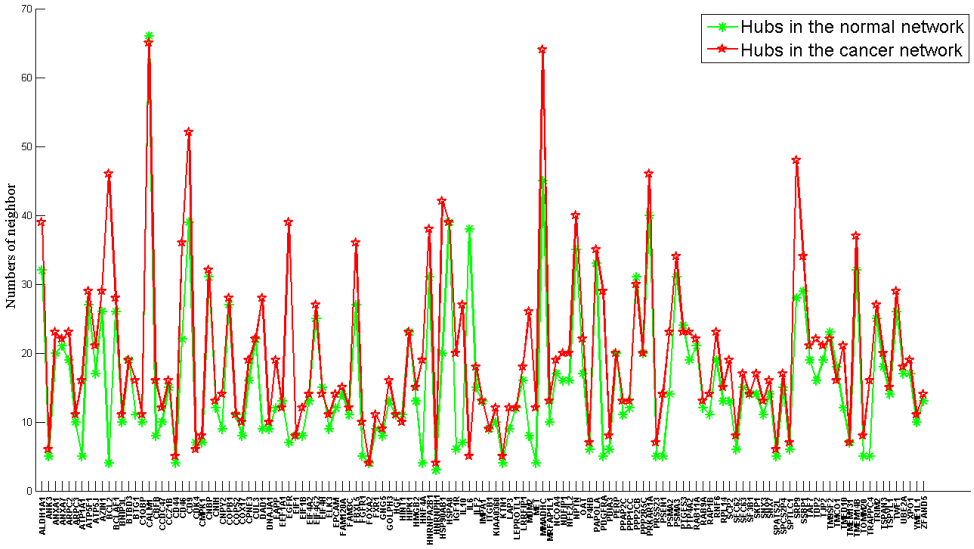

Fig. 6. Gross topological changes detected in lung epithelial cancer cells. The genes listed on the bottom are those which have undergone significant global variations. The dashed green line indicates the connectivity (the number of edges) for all of the 133 genes in the normal controls. The dashed red line indicates the connectivity for the same 133 genes in the lung cancers. These 133 genes are the hubs (non-zeros rows in U_i) selected in our scan either in the normal state or cancer.

tumorigenesis except 4 (ALDH1A1,CALM1,FGFR1,PERP). Interestingly, these stable hubs including ALDH1A1, FGFR1 and PERP are either oncogene or associated with smoking (see below).

One of the four stable hubs is ALDH1A1, a tumor stem cell-associated marker which is known to be activated by smoking, presumably due to the carcinogenic aldehydes in cigarette smoke [24]. PERP, an apoptosis-associated target of TP53, is a member of the PMP-22/gas3 family [2]. FGFR1 amplification is found to be a prognostic marker in early-stage non-small cell lung cancer [9]. CALM1 is an important mediator of signal transductions by Ca++ for cellular proliferation and cell cycle progression [8].

We note the loss of the IL6 motif in tumorigenesis, an important cytokine mediating cellular immunity against neoplastic transformation through the famous JAK1/STAT3 pathway [38], supporting the notion that gain of function-s promoting cellular proliferation (oncogenes) and loss of functions inhibiting tumor growth (tumor suppressor)(Figure 7C). Interestingly, another cytokine IL10 undergoes opposite changes, gaining a significant number of new partners during tumorigenesis. Unlike IL6, IL10 promotes tumorigenesis by stimulating cell proliferation and inhibiting cell apoptosis [37]. High systemic levels of IL-10 correlate with poor survival of some cancer patients. Our results thus support the principle that cytokines play the important roles in the immunity of lung tumorigenesis.

We demonstrate the emergences of new hub structures in the cancers associated with BCL2, EGFR, MDM2, IL10, DAD1, MMADHC, SRP9, HSP90AB, and IGF1R). Interestingly, all of these changes are previously implicated in tumorigenesis. Most strikingly to note is the important EGFR, BCL2, MDM2, DAD, HSP90, and IL10 genes and their known associations with lung cancers [48]. EGFR is only recently discovered to be an crucial factor in lung neoplastic transformation and now a new drug target of this dreadful malignancy (Figure 7D) [34]. In our result, EGFR undergoes the most obvious changes, gaining the most number of interactions during tumorigenesis. Note that our results also justify why EGFR has been an ideal target for drug development against lung cancers, being the largest hub thus more effective to be perturbed. In addition, a similar growth factor receptor PDGFRA appears to be also associated with lung tumorigenesis as well. As a member of the platelet-derived growth factor family, PDGFRA is a cell surface tyrosine kinase receptor, promoting mitosis for cells of mesenchymal origin, which are widely dispersed in lung and highly associated with smoking-induced non-small cell lung cancers [46]. MDM2 is involved in the important TP53-MDM2 circuitry. Also it is known that IGF-1R promotes cellular proliferation in several cancers, including lung, liver, and breast [47].

Correlations underlying gene expression data are largely considered indirect and at the best suggestive of possible gene-gene interactions (or regulation). In addition, the levels of gene expressions are also not necessarily associated with any direct effects of biological functions. However, a graphical model aimed to infer a possible interaction (or regulation) under the global conditional dependence for all of the other genes involved is a more sensible approach. Such a global conditional dependence is scientifically more appropriate because here each of the interaction is not assessed alone but is inferred by taking rest of the system into account as a whole. Figure 7 demonstrates the importance of using a graphical model compared to the original pairwise-based method (Pearson

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2016.2599867, IEEE/ACM Transactions on Computational Biology and Bioinformatics

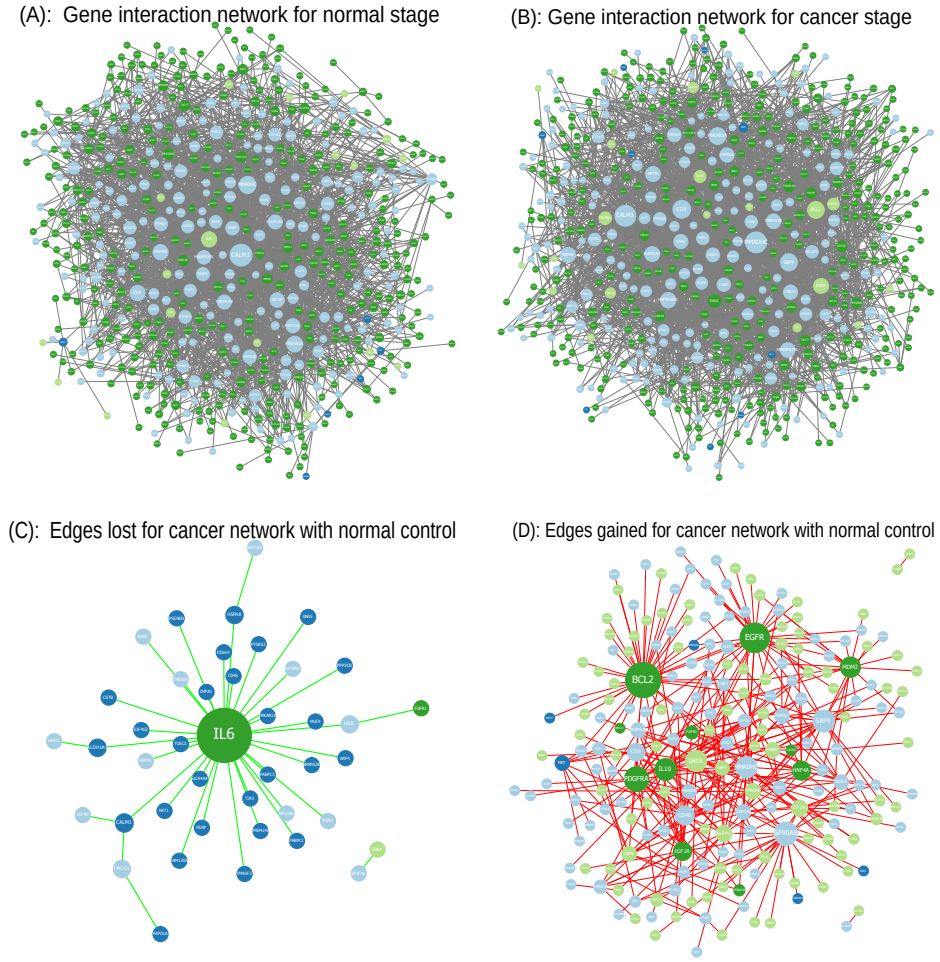

Fig. 7. Topological characteristics around the hubs detected by our GV-Lasso with lung tumorigenesis. Shown here in the Panels A, B, C and D are the sub-networks around the aforementioned 133 genes. The sizes of nodes correspond to the degrees of their interactions. The colors represent similar classes according the pairwise Pearson correlations of gene expressions. Demonstrated here is the absolute essentiality of using conditional independence to interpret the apparent correlations between gene expressions. Panel (A) represents the normal controls; (B) the cancers; (C) the interactions lost in cancer; and (D) the interactions gained in cancer.

correlation). Without the graphical model, nearly all of the genes selected particularly in the cancers (Figure 7B: the nodes in pink) would have been clustered together, forming a highly connected giant component, computationally inseparatable nor biologically interpretable. Over the years, many efforts have been devoted to work on this problem, using either betweeness or other measures. Wille has first established the importance of using a Gaussian graphical approach to model biological networks [50]. Here, we demonstrate again that the previously observed correlations among individual genes can be further divided into multiple individual modules, anchored by some important hubs (which would have been indistinguishable using the previous approach). Therefore, graphical model provides a biologically relevant framework, allowing us to revisit many of the gene expression datasets previously published for more meaningful and biologically-inspired investigations.

C. Direction of Neoplastic Transformation

Having established that lung cancers undergo consistent and gross topological alterations, implicating lung cancers might occur as a consequence (or activation) of some intrinsic cascade, induced as a result of exposure to smoking. We seek to address an important biological question as to the direction of neoplastic transformation in terms of the topological structure and organization. We are inspired by Alon and his colleagues for their work using random graphs to assess the level of organization (or orderness) for any given networks [32]. Accordingly, we calculate the average shortest path of our cancer-based network compared to its normal control. Additional topological features, such as the number of cliques and other local orderness are also assessed for their relative significances over their random permutations (a null hypothesis suggested by Alon and his colleagues). Our result does indicate the cancer network appears to become slightly more disorganized, with the average shortest distance

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2016.2599867, IEEE/ACM Transactions on Computational Biology and Bioinformatics

TABLE I
TOPOLOGICAL CHARACTERISTICS ASSOCIATED WITH SMOKING

	Never-smokers	Former-smokers	Current-smokers
Vertices	1180	1180	1180
Edges	8806	10064	9042
Degree	14.93	17.06	15.33
Vertices in the giant component	1107	1135	1116
Edges in the giant component	8733	10019	8978
Average shortest path in the giant component	2.16	2.11	2.14
The longest shortest path in the giant component	5	5	5
Independent loops	7700	8930	7927
Star motifs s_4(one hub and three spokes)	260555538	422713879	301211468

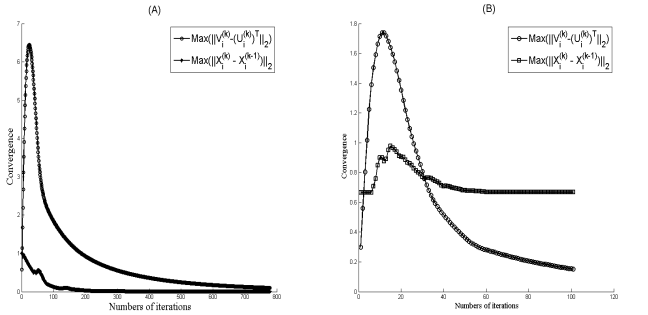

Fig. 8. The convergence of our GV-Lasso. Panels A for the cancer dataset with normal control. Here, the parameters are: $\alpha_1 = 0.8$, $\alpha_2 = 0.5$, $\beta_1 = 0.38$, $\beta_2 = 7.5$, $\rho_1 = 1$, $\rho_2 = 1.5$. Panels B for the early changes caused by smoking. Here, the parameters are: $\alpha_1 = 1.5$, $\alpha_2 = 5$, $\beta_1 = 1.5$, $\beta_2 = 15$, $\rho_1 = 1.5$, $\rho_2 = 1.5$.

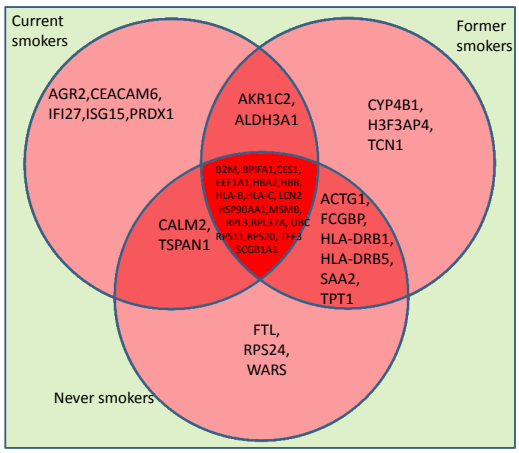

Fig. 9. Detailed topological comparisons for the normal epithelial cells exposed to smoking. Human airway epithelial cells are isolated from 34 current smokers, 18 former smokers, and 23 never smokers (using the identical genes aforementioned), respectively. Identical parameters and genes as the Figures 8 are used to model any possible early changes by comparisons. Demonstrated here are the relative proportions for each of the unique and the joint. The name of genes in each of the fractions are all of hubs discovered, using our GV-Lasso.

being 4.39 in cancer compared to 4.51 in normal (both are significant compared to their random ensembles see the Table 1 in the Appendix). However, because of the nature of our approach, intentionally missing many of the local and minor events occurring during tumorigenesis, we are unable to reach a strong conclusion based purely on our empirical extrapolations. As a future direction of this effort, one of our goals is to perform a genome-wide permutation to assess the structural and organizational significances for both the cancer and normal cells.

D. Detection of early changes as potential early markers

Our preliminary results (see Figure 5, C and D) have demonstrated the relative stability for the topological structures among the samples presumably still at their early stages of neoplastic transformation (histologically normal). This result nicely validates our methodology, linking biological structures to their phenotypes.

We then seek to explore in detail, both globally and locally, for any structural variations among the samples employed in our study. The purpose is to capture any early signs which would be likely attributable to future tumorigenesis, thus used as potential early markers for smoking related malignancies. The original work using the same dataset (GDS534) has reported a number of candidates with differential expressions between smoking and its control, including genes responsible for xenobiotic metabolism and redox-regulations, all being consistent with the consequence of smoking. Their results also reveal the importance of inflammation, immunity, and secretion being potential early markers because of their presumptive roles protective against smoking. Interestingly, our method discover similar events but with a few extra insights.

The parameters used for our GV-Lasso and their convergences are given in Figure 8. First, similar to the previous reports, we also detect the same two modules involving AGR2 and AKR1C2, both of which are well-known markers for smoking-related risks (tumorigenesis, miscarriage, birth defect, etc) [25],[49]. Interesting to note is the additional advantage of our methodology that the markers are discovered along with their interacting genes as the hubs, allowing modular structures instead of individual genes as the potential markers while filtering out all other minor and local variations. Obviously, the changes of these modular structures would be statistically more significant.

We next seek to find what would be the potential benefits to quit smoking. It turns out that at least five functional modules

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2016.2599867, IEEE/ACM Transactions on Computational Biology and Bioinformatics

can remain intact if the exposure to smoking is only temporary, including FCGBP, HLA-DRB1/5, SAA2 and TPT1. FCGBP is a known cancer marker for a number of different malignancies, including lung as well as prostate, gallbladder, and thyroid cancers [51]. The expression of FCGBP has been shown to be inversely related to the progression of tumorigenesis, presumably as a protective and/or favorable marker. TPT1 belongs to another protective system, thought to promote remodeling and repair of pulmonary vascular cells. On the other hand, all members of HLA family are perhaps the most important mediators of cellular-immunity, without which cells would lost much of the protection by the T-cell mediated self-recognition and the consequential cytotoxicity. Interestingly, the loss of SAA2 has recently been reported to be a potential marker for lung cancer [43].

It is important to also find out what the possible protective markers might be lost as a result of smoking. Here three modules are found to specifically retain only in cells without any exposure to smoking, anchored by FTL, RSP24, and WARS. Interestingly, all three modules appear relevant to smoking: FTL (ferritin) is responsible for the storage of iron in a soluble and nontoxic state, which has been shown to be affected following smoking [22]; RPS24, a ribosomal protein, is known to regulate the Mdm2-TP53-MdmX circuitry, the most important tumor suppressor [12]; and WARS, a tRNA synthetase, whose inactivation has been used as an informative cancer marker [18]. However, all these markers have not been previously reported to be directly related to smoking.

It is unclear why individuals can still remain at higher risk developing lung cancer long after their cession of smoking compared to the general public. ALDH3A1 and its associated module are captured from the epithelial cells isolated from both of the short and long term smokers, which is also well-documented in previous literatures as a potential cancer marker, including the original work using the same data as this study. We are not completely sure how this epigenetic effect would be directly or indirectly associated with previous smoking. But we want to speculate that it is possible that smoking has been such an insult, which has left cells with some permanent imprints, presumably affecting gene expressions. Overall, cession of smoking could retain, or possibly reverse at least some of the important and beneficial biological functions, many of which would perhaps prevent tumorigenesis from further development.

We suggest that a topological structure (a functional module consisting of many genes) instead of individual genes be a more robust biological marker. Besides ALDH3A1 and CYP4B1, our results do not overlap with the original report using exactly the same dataset. For instance, even though CYP1B1 over-expresses significantly (30 folds) while RPS24 does not (often used as an internal control for gene expressions) based on the differential analysis, our method in fact considers RPS24 and its associated network a potential marker but not CYP1B1. Interestingly, RPS24 mediates many interactions both in normal controls and smokers, while undergoing drastic changes as well. Though we are not sure about the biological significance of this event, it is conceivable that a ribosomal protein like RPS24 might still be an important regulator, synchronizing protein synthesis.

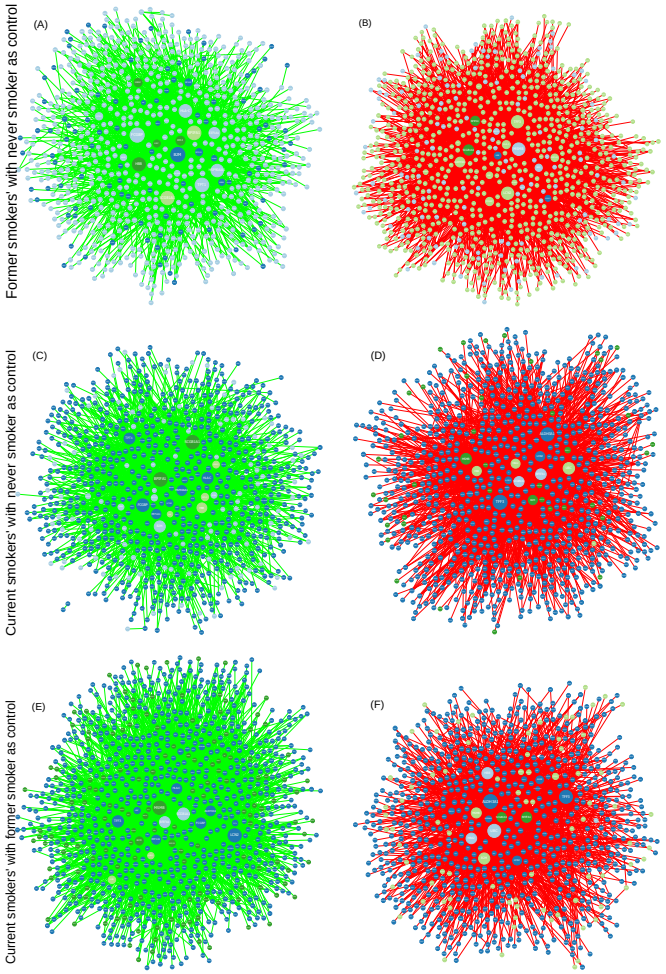

Fig. 10. Topological characteristics around the hubs detected by our GV-Lasso with the comparison among the never, former and current smokers. The sizes of nodes correspond to the degrees of their interactions. The colors represent similar classes according the pairwise Pearson correlations of gene expressions (see Figure 7). Panel (A) represents the interactions lost in former smokers with never smokers as the control; (B) the interactions gained in the former smokers with the never smokers as control; (C) the interactions lost in the current smokers with the never smokers as control; (D) the interactions gained in current smokers with the never smokers as control; (E) the interactions lost in the current smokers with the former smokers as control; and (F) the gained interactions in current smokers with former smokers as control

Finally, we want to explore the possible mechanism by which ALDH3A1 might mediate the neoplastic transformation of lung epithelial cells. Since LCN2 is presumably responsible as a protective agent against smoking (suppression of invasiveness and metastasis), the continued exposure of smoking will consistently activate ALDH3A1, while eventually inactivating LCN2 (Figure 11, B and C). Consequentially, we speculate that the ALDH3A1 will serve at least as one of the key events, influencing the important TP53-MDM2 system by directly interacting with TP53 (Figure 11, B and C). Thus, it is our hypothesis that LCN2 might be used as early marker for potential tumorigenesis. In summary, smoking will not necessarily lead to tumorigenesis; only about 10-20% of smokers would

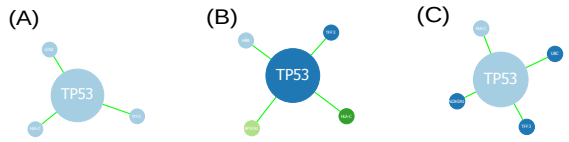

Fig. 11. A putative model by which ALDH3A1 could promote lung neoplastic transformation. Shown here are the TP53-MDM2 circuitry captured using our GV-Lasso. The colors indicate classes of co-expression profiles based on the Pearson correlations. Panel A: Normal controls (never smoking); Panel B: Former smokers; and Panel C: Current smokers. Note the loss of LCN2 after smoking; also note the gain of ALDH3A1 in current smokers, all associated with the important TP53 gene.

eventually develop lung cancer. Cession of smoking is good but can still possess an eventual risk, perhaps depending at least in part on the luck as to whether the exposure of smoking has been long or strong enough to have already turned on the ALDH3A1, in a way irreversible to its original state.

IV. Perspective

The consistency of the global variations among at least a majority of lung cancers studied implicates a possible common cascade induced by smoking, leading to the eventual genomic instability as a tumor phenotype. The notion is well taken that cancers of solid tissues are extremely heterogeneous thus our methodology developed here is purposely aimed to focus only on the common and global variations. Our results support the traditional position that neoplastic transformation is still a clonal and evolutionary process, with at least some of the progenitor events being largely shared. As a further direction, we plan to expand the temporal application of our GV-Lasso, with a hidden Markov model for the truly dynamic process of a cancer.

References

[1] A. Ahmed and E. P. Xing, "Recovering time-varying networks of dependencies in social and biological studies," *Proceedings of the National Academy of Sciences*, vol. 106, no. 29, pp. 11 878–11 883, 2009.

[2] L. D. Attardi, E. E. Reczek, C. Cosmas, E. G. Demicco, M. E. McCurrach, S. W. Lowe, and T. Jacks, "Perp, an apoptosis-associated target of p53, is a novel member of the pmp-22/gas3 family," *Genes & development*, vol. 14, no. 6, pp. 704–718, 2000.

[3] K. Baba, R. Shibata, and M. Sibuya, "Partial correlation and conditional correlation as measures of conditional independence," *Australian & New Zealand Journal of Statistics*, vol. 46, no. 4, pp. 657–664, 2004.

[4] F. Bach, "Optimization with sparsity-inducing penalties," *Foundations and Trends® in Machine Learning*, vol. 4, no. 1, pp. 1–106, 2011.

[5] G. Bounova and O. de Weck, "Overview of metrics and their correlation patterns for multiple-metric topology analysis on heterogeneous graph ensembles," *Physical Review E*, vol. 85, no. 1, p. 016117, 2012.

[6] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends® in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2011.

[7] S. Cambanis, S. Huang, and G. Simons, "On the theory of elliptically contoured distributions," *Journal of Multivariate Analysis*, vol. 11, no. 3, pp. 368–385, 1981.

[8] J. G. Chafouleas, W. E. Bolton, H. Hidaka, A. E. Boyd, and A. R. Means, "Calmodulin and the cell cycle: involvement in regulation of cell-cycle progression," *Cell*, vol. 28, no. 1, pp. 41–50, 1982.

[9] N. Cihoric, S. Savic, S. Schneider, I. Ackermann, M. Bichsel-Naef, R. Schmid, D. Lardinois, M. Gugger, L. Bubendorf, I. Zlobec *et al.*, "Prognostic role of fgfr1 amplification in early-stage non-small cell lung cancer," *British journal of cancer*, vol. 110, no. 12, pp. 2914–2922, 2014.

[10] L. Condat, "A direct algorithm for 1d total variation denoising," *IEEE Signal Processing Letters*, vol. 20, no. 11, pp. 1054–1057, 2013.

[11] C. M. Croce, "Oncogenes and cancer," *New England Journal of Medicine*, vol. 358, no. 5, pp. 502–511, 2008.

[12] L. Daftuar, Y. Zhu, X. Jacq, and C. Prives, "Ribosomal proteins rpl37, rps15 and rps20 regulate the mdm2-p53-mdmx network," *PLoS One*, vol. 8, no. 7, p. e68667, 2013.

[13] P. Danaher, P. Wang, and D. M. Witten, "The joint graphical lasso for inverse covariance estimation across multiple classes," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 76, no. 2, pp. 373–397, 2014.

[14] M. C. de Souto, I. G. Costa, D. S. de Araujo, T. B. Ludermir, and A. Schliep, "Clustering cancer gene expression data: a comparative study," *BMC bioinformatics*, vol. 9, no. 1, p. 497, 2008.

[15] D. Edwards, *Introduction to graphical modelling*. Springer Verlag, 2000.

[16] T. S. Furey, N. Cristianini, N. Duffy, D. W. Bednarski, M. Schummer, and D. Haussler, "Support vector machine classification and validation of cancer tissue samples using microarray expression data," *Bioinformatics*, vol. 16, no. 10, pp. 906–914, 2000.

[17] Y. Gao and G. Church, "Improving molecular cancer class discovery through sparse non-negative matrix factorization," *Bioinformatics*, vol. 21, no. 21, pp. 3970–3975, 2005.

[18] A. Ghanipour, K. Jirström, F. Pontén, B. Glimelius, L. Påhlman, and H. Birgisson, "The prognostic significance of tryptophanyl-trna synthetase in colorectal cancer," *Cancer Epidemiology Biomarkers & Prevention*, vol. 18, no. 11, pp. 2949–2956, 2009.

[19] J. Guo, E. Levina, G. Michailidis, and J. Zhu, "Joint estimation of multiple graphical models," *Biometrika*, vol. 98, no. 1, pp. 1–15, 2011.

[20] B. He and X. Yuan, "On the o(1/n) convergence rate of the douglas-rachford alternating direction method," *SIAM Journal on Numerical Analysis*, vol. 50, no. 2, pp. 700–709, 2012.

[21] R. S. Heist, M. Mino-Kenudson, L. V. Sequist, S. Tammireddy, L. Morrissey, D. C. Christiani, J. A. Engelman, and A. J. Iafrate, "Fgfr1 amplification in squamous cell carcinoma of the lung," *Journal of thoracic oncology: official publication of the International Association for the Study of Lung Cancer*, vol. 7, no. 12, p. 1775, 2012.

[22] M. W. Hentze, M. U. Muckenthaler, B. Galy, and C. Camaschella, "Two to tango: regulation of mammalian iron metabolism," *Cell*, vol. 142, no. 1, pp. 24–38, 2010.

[23] H. Hoefling, "A path algorithm for the fused lasso signal approximator," *Journal of Computational and Graphical Statistics*, vol. 19, no. 4, pp. 984–1006, 2010.

[24] F. Jiang, Q. Qiu, A. Khanna, N. W. Todd, J. Deepak, L. Xing, H. Wang, Z. Liu, Y. Su, S. A. Stass *et al.*, "Aldehyde dehydrogenase 1 is a tumor stem cell-associated marker in lung cancer," *Molecular Cancer Research*, vol. 7, no. 3, pp. 330–338, 2009.

[25] Q. Lan, J. L. Mumford, M. Shen, D. M. DeMarini, M. R. Bonner, X. He, M. Yeager, R. Welch, S. Chanock, L. Tian *et al.*, "Oxidative damage-related genes akr1c3 and ogg1 modulate risks for lung cancer due to exposure to pah-rich coal combustion emissions," *Carcinogenesis*, vol. 25, no. 11, pp. 2177–2181, 2004.

[26] C. Lengauer, K. W. Kinzler, and B. Vogelstein, "Genetic instabilities in human cancers," *Nature*, vol. 396, no. 6712, pp. 643–649, 1998.

[27] C. D. Little, M. M. Nau, D. N. Carney, A. F. Gazdar, and J. D. Minna, "Amplification and expression of the c-myc oncogene in human lung cancer cell lines," 1983.

[28] H. Liu, F. Han, and C.-h. Zhang, "Transelliptical graphical models," in *Advances in Neural Information Processing Systems*, 2012, pp. 809–817.

[29] N. M. Luscombe, M. M. Babu, H. Yu, M. Snyder, S. A. Teichmann, and M. Gerstein, "Genomic analysis of regulatory network dynamics reveals large topological changes," *Nature*, vol. 431, no. 7006, pp. 308–312, 2004.

[30] D. Marbach, R. J. Prill, T. Schaffter, C. Mattiussi, D. Floreano, and G. Stolovitzky, "Revealing strengths and weaknesses of methods for gene network inference," *Proceedings of the National Academy of Sciences*, vol. 107, no. 14, pp. 6286–6291, 2010.

[31] N. F. Marko and R. J. Weil, "Non-gaussian distributions affect identification of expression patterns, functional annotation, and prospective classification in human cancer genomes," *Plos One*, vol. 7, no. 10, pp. 1310–1315, 2012.

[32] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, "Network motifs: simple building blocks of complex networks," *Science*, vol. 298, no. 5594, pp. 824–827, 2002.
[33] K. Mohan, P. London, M. Fazel, D. Witten, and S.-I. Lee, "Node-based learning of multiple gaussian graphical models," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 445–488, 2014.
[34] J. G. Paez, P. A. Jänne, J. C. Lee, S. Tracy, H. Greulich, S. Gabriel, P. Herman, F. J. Kaye, N. Lindeman, T. J. Boggon *et al.*, "Egfr mutations in lung cancer: correlation with clinical response to gefitinib therapy," *Science*, vol. 304, no. 5676, pp. 1497–1500, 2004.
[35] D. Pe'er and N. Hacohen, "Principles and strategies for developing network models in cancer," *Cell*, vol. 144, no. 6, pp. 864–873, 2011.
[36] C. Peterson, F. Stingo, and M. Vannucci, "Bayesian inference of multiple gaussian graphical models," *Journal of the American Statistical Association*, no. just-accepted, pp. 00–00, 2014.
[37] N. Shivapurkar, J. Reddy, P. M. Chaudhary, and A. F. Gazdar, "Apoptosis and lung cancer: a review," *Journal of cellular biochemistry*, vol. 88, no. 5, pp. 885–898, 2003.
[38] L. Song, B. Rawal, J. A. Nemeth, and E. B. Haura, "Jak1 activates stat3 activity in non-small–cell lung cancer cells and il-6 neutralizing antibodies can suppress jak1-stat3 signaling," *Molecular cancer therapeutics*, vol. 10, no. 3, pp. 481–494, 2011.
[39] D. J. Spiegelhalter, N. G. Best, B. P. Carlin, and A. Linde, "The deviance information criterion: 12 years on," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 76, no. 3, pp. 485–493, 2014.
[40] A. Spira, J. Beane, V. Shah, G. Liu, F. Schembri, X. Yang, J. Palma, and J. S. Brody, "Effects of cigarette smoke on the human airway epithelial cell transcriptome," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 27, pp. 10 143–10 148, 2004.
[41] A. Spira, J. E. Beane, V. Shah, K. Steiling, G. Liu, F. Schembri, S. Gilman, Y.-M. Dumas, P. Calner, P. Sebastiani *et al.*, "Airway epithelial gene expression in the diagnostic evaluation of smokers with suspect lung cancer," *Nature medicine*, vol. 13, no. 3, pp. 361–366, 2007.
[42] B. W. Stewart, P. Kleihues, I. A. for Research on Cancer *et al.*, *World cancer report*. IARC press Lyon, 2003, vol. 57.
[43] H.-J. Sung, S.-A. Jeon, J.-M. Ahn, K.-J. Seul, J. Y. Kim, J. Y. Lee, J. S. Yoo, S.-Y. Lee, H. Kim, and J.-Y. Cho, "Large-scale isotype-specific quantification of serum amyloid a 1/2 by multiple reaction monitoring in crude sera," *Journal of proteomics*, vol. 75, no. 7, pp. 2170–2180, 2012.
[44] K. M. Tan, P. London, K. Mohan, S.-I. Lee, M. Fazel, and D. Witten, "Learning graphical models with hubs," *The Journal of Machine Learning Research*, vol. 15, pp. 3297–3331, 2014.
[45] K. Vahakangas, R. Metcalf, J. Welsh, W. Bennett, C. Harris, J. Samet, and D. Lane, "Mutations of p53 and ras genes in radon-associated lung cancer from uranium miners," *The Lancet*, vol. 339, no. 8793, pp. 576–580, 1992.
[46] R. G. Verhaak, K. A. Hoadley, E. Purdom, V. Wang, Y. Qi, M. D. Wilkerson, C. R. Miller, L. Ding, T. Golub, J. P. Mesirov *et al.*, "Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in pdgfra, idh1, egfr, and nf1," *Cancer cell*, vol. 17, no. 1, pp. 98–110, 2010.
[47] J. Villanueva, A. Vultur, J. T. Lee, R. Somasundaram, M. Fukunaga-Kalabis, A. K. Cipolla, B. Wubbenhorst, X. Xu, P. A. Gimotty, D. Kee *et al.*, "Acquired resistance to braf inhibitors mediated by a raf kinase switch in melanoma can be overcome by cotargeting mek and igf-1r/pi3k," *Cancer cell*, vol. 18, no. 6, pp. 683–695, 2010.
[48] G. Wang, X. Zhu, L. Hood, and P. Ao, "From phage lambda to human cancer: endogenous molecular-cellular network hypothesis," *Quantitative Biology*, vol. 1, no. 1, pp. 32–49, 2013.
[49] Z. Wang, Y. Hao, and A. W. Lowe, "The adenocarcinoma-associated antigen, agr2, promotes tumor growth, cell migration, and cellular transformation," *Cancer research*, vol. 68, no. 2, pp. 492–497, 2008.
[50] A. Wille, P. Zimmermann, E. Vranová, A. Fürholz, O. Laule, S. Bleuler, L. Hennig, A. Prelic, P. von Rohr, L. Thiele *et al.*, "Sparse graphical gaussian modeling of the isoprenoid gene network in arabidopsis thaliana," *Genome Biol*, vol. 5, no. 11, p. R92, 2004.
[51] L. Xiong, Y. Wen, X. Miao, and Z. Yang, "Nt5e and fcgbp as key regulators of tgf-1-induced epithelial–mesenchymal transition (emt) are associated with tumor progression and survival of patients with gallbladder cancer," *Cell and tissue research*, vol. 355, no. 2, pp. 365–374, 2014.
[52] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 68, no. 1, pp. 49–67, 2006.
[53] ——, "Model selection and estimation in the gaussian graphical model," *Biometrika*, vol. 94, no. 1, pp. 19–35, 2007.
[54] T. Zhang and H. Zou, "Sparse precision matrix estimation via lasso penalized d-trace loss," *Biometrika*, p. ast059, 2014.
[55] P. Zhao and B. Yu, "On model selection consistency of lasso," *The Journal of Machine Learning Research*, vol. 7, pp. 2541–2563, 2006.

Yanbo Wang received the Bachelors degree from Nanjing University Of Finance & Economics(NJUE)in Applied Mathematics, the M.S.degree from Nanjing Agriculture University(NJAU)in Computational Mathematics.

He is now a Ph.D. candidate in the Department of Computer Science and Engineering, Shanghai Jiao Tong University(SJTU). His research interests include statistical machine learning and stochastic process.

Weikang Qian is an assistant professor in the University of Michigan-Shanghai Jiao Tong University Joint Institute at Shanghai Jiao Tong University. He received his Ph.D. degree in Electrical Engineering at the University of Minnesota in 2011 and his B.Eng. degree in Automation at Tsinghua University, China in 2006. His main research interests include electronic design automation and digital design for emerging technologies. In recognition of his doctoral research, he received the Doctoral Dissertation Fellowship at the University of Minnesota. One of his papers was nominated for the William J. McCalla Best Paper Award at the 2009 International Conference on Computer-Aided Design (ICCAD).

Bo Yuan received the Bachelors degree from Peking University Medical School, Beijing, China, in 1983, the M.S. degree in biochemistry, and the Ph.D. degree in molecular genetics from the University of Louisville, Louisville, KY, in 1990 and 1995, respectively.

He is currently a Professor in the Department of Computer Science and Engineering, Shanghai Jiao Tong University (SJTU), Shanghai, China. Before joining SJTU, he was a Tenure-Track Assistant Professor with Ohio State University (OSU), Columbus, OH, in 2006, and served as a Co-Director for the OSU Program in Pharmacogenomics. At OSU, he was the founding director of OSUs genome initiative during the early 2000s, leading one of the only three independent efforts in the world (besides the Human Genome Project and the Celera Company), having assembled and deciphered the entire human and mouse genomes.

His current research interests include biological networks, network evolution, stochastic process, biologically inspired computing, and bioinformatics, particularly on how these frameworks might impact the development of intelligent algorithms and systems.