Bioinformatics, 36(3), 2020, 798–804
doi: 10.1093/bioinformatics/btz636
Advance Access Publication Date: 28 August 2019
Original Paper

Gene expression

NExUS: Bayesian simultaneous network estimation across unequal sample sizes

Priyam Das[1], Christine B. Peterson[1,*], Kim-Anh Do[1], Rehan Akbani[2] and Veerabhadran Baladandayuthapani[3]

[1]Department of Biostatistics, [2]Department of Bioinformatics & Computational Biology, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA and [3]Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA

*To whom correspondence should be addressed.

Associate Editor: Russell Schwartz

Received on November 8, 2018; revised on June 25, 2019; editorial decision on August 5, 2019; accepted on August 26, 2019

Abstract

Motivation: Network-based analyses of high-throughput genomics data provide a holistic, systems-level understanding of various biological mechanisms for a common population. However, when estimating multiple networks across heterogeneous sub-populations, varying sample sizes pose a challenge in the estimation and inference, as network differences may be driven by differences in power. We are particularly interested in addressing this challenge in the context of proteomic networks for related cancers, as the number of subjects available for rare cancer (sub-)types is often limited.

Results: We develop NExUS (Network Estimation across Unequal Sample sizes), a Bayesian method that enables joint learning of multiple networks while avoiding artefactual relationship between sample size and network sparsity. We demonstrate through simulations that NExUS outperforms existing network estimation methods in this context, and apply it to learn network similarity and shared pathway activity for groups of cancers with related origins represented in The Cancer Genome Atlas (TCGA) proteomic data.

Availability and implementation: The NExUS source code is freely available for download at https://github.com/priyamdas2/NExUS.

Contact: cbpeterson@mdanderson.org

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The last decade has seen a proliferation of large, complex datasets that quantify molecular variables such as gene, protein, microbiome and population-wide genetic variation. The Cancer Genome Atlas (TCGA) is a prime example of recent large-scale consortium-level efforts, which has generated multi-platform 'omics measurements from >10 K patients across 32 common and rare cancer types. This allows for systematic investigations into the molecular mechanisms behind various oncogenic processes. Multiple studies have established that cancer initiation and progression are not outcomes of a single mutation within a gene or protein, but rather the result of a perturbation to co-ordinated networks and pathways that correspond to basic oncogenic processes such as cell proliferation and apoptosis (Hanahan and Weinberg, 2000; Wang *et al.*, 2015). Therefore, it is important to understand and characterize the underlying network dependency structures within and across cancers (Cho *et al.*, 2016; Hristov and Singh, 2017). From a discovery standpoint, this is not only crucial to identify new cancer biomarkers and mechanisms, but also to distinguish the key molecular regulators of networks in different cancers (Carro *et al.*, 2010; Sonabend *et al.*, 2014).

Most existing studies focus on identifying the characteristics of molecular networks in individual cancer populations one at a time to understand tumor-specific molecular interactions (Creixell *et al.*, 2015; Gill *et al.*, 2014). On the other hand, researchers are beginning to recognize the critical importance of simultaneous analysis of similar tumor types (e.g. in terms of cell origin, organ location, biological evolution) to understand fundamental commonalities and differences (Tamborero *et al.*, 2013; Weinstein *et al.*, 2013). This has led to multiple pan-cancer studies that encompass the examination of various genomic and immunologic features in related cancers such as different squamous carcinomas (Campbell *et al.*, 2018), gynecologic and breast cancers (Berger *et al.*, 2018), gastrointestinal adenocarcinomas (Liu *et al.*, 2018) and urologic cancers (Chen *et al.*, 2017). In this article, we propose a network modeling approach for simultaneous analysis of genomic data from related tumor types.

Probabilistic graphical models (Lauritzen, 1996) are well-established statistical tools to conduct estimation and inference of network structures. In particular, Gaussian graphical models (Whittaker, 1990) have gained immense popularity because of their ability to capture global dependency structures. In high-dimensional

settings, such as genomics, sparse graphical models are widely used to identify important nodes and interactions in a network consisting of a large number of genes/proteins. Various approaches for Gaussian graphical model estimation (Baladandayuthapani *et al.*, 2014; Friedman *et al.*, 2008; Wang, 2012, 2015; Yuan and Lin, 2007) have been proposed over the years. These methods recently have been generalized to link network estimation across multiple populations using penalization-based (Danaher *et al.*, 2014) or Bayesian approaches (Peterson *et al.*, 2015). The former approach is not optimal for groups with differing levels of similarity, while the latter has scaling limitations to the prior specification. Related work on joint estimation of Gaussian graphical models in the Bayesian framework can be found in Lin *et al.* (2017) and Tan *et al.* (2017). Along with learning the sparse network for each cancer type while borrowing strength across groups, another objective of our proposed method is to avoid artefactual differences in network sparsity due to differing sample sizes. In an application to pan-cancer network analysis, Kling *et al.* (2015) proposed an algorithmic method for joint network estimation with a group-specific penalty correction based on sample size. Their method, however, lacks a formal statistical justification of the sample-size penalty term and its effect on the sparsity of the estimated networks.

In this article, we propose a Bayesian approach for simultaneous Network Estimation across Unequal Sample sizes (NExUS). The main advantages of the NExUS method over existing methods include (i) explicit incorporation of sample size correction in network estimation which not only allows control of the sparsity but also improves borrowing of strength between cancer-specific networks to enable a balancing of statistical power across groups; (ii) the ability to quantify a *network similarity index* (NSI) which can be used as a global measure of network similarity among different cancer-specific networks; (iii) automatic selection of penalty parameters within a fully Bayesian framework, avoiding the need for the cross-validation step required by most frequentist graphical modeling approaches; and (iv) an efficient strategy for updating the precision matrices, making the method more scalable both in terms of the number of variables and the number of groups than existing Bayesian methods (Kundu *et al.*, 2018; Peterson *et al.*, 2015).

Our methods are motivated by and applied to The Cancer Proteomic Atlas (TCPA, Li *et al.*, 2013) that has collected high-quality Reverse Phase Protein Array (RPPA) data across 32 cancer types. Many of these cancers can be divided into groups based on their histological origin, location, or similarity of biological and oncogenic processes. In this paper, we consider four different group of cancers, namely pan-gynecological, pan-kidney, pan-squamous and pan-gastrointestinal. The simultaneous analysis of the proteomic networks of related cancers can help in robustly identifying the common network features. In addition, it can improve power for the discovery of features for rare cancers, which are shared with other cancers. We analyze the shared network structure of these cancers and identify shared pathway activity between all pairs of protein networks within the same group of cancers. We are able to establish both global trends in network similarity (e.g. we find that uterine carcinoma, a rare cancer, is more similar to cervical squamous cell carcinoma and endocervical adenocarcinoma, than to other gynecological cancers), as well as pathway-specific activity sharing (e.g. we find that the hormone receptor pathway has common activity across the pan-gynecological group, while the RAS/MAPK and RTK pathways have shared activation across the various types of kidney cancers).

2 The NExUS method

Data structure and notation Let $\mathbf{X}_c$ represent the $n_c \times p$ matrix of observed protein or gene expression for the cth cancer type of interest, where $c = 1, 2, \ldots, C$. We assume that the same set of p proteins are observed for all subjects, but allow the sample sizes n_c for each cancer type to differ. We assume that the data for each subject i follows a multivariate normal distribution

$$\mathbf{x}_{c,i} \sim N_p(0, \boldsymbol{\Theta}_c^{-1}),\ i = 1, \ldots, n_c,$$

with mean vector $0 \in \mathrm{R}^p$ and cancer-specific precision matrix $\boldsymbol{\Theta}_c$. The multivariate normal distribution is parameterized using the precision matrix $\boldsymbol{\Theta}_c$ (rather than the covariance matrix $\boldsymbol{\Sigma}_c = \boldsymbol{\Theta}_c^{-1}$) since there is a direct correspondence between the precision matrix and the conditional independence graph among variables. Specifically, in a Gaussian graphical model, entry θ_{ij}^c in the precision matrix $\boldsymbol{\Theta}_c$ is exactly 0 if and only if the corresponding proteins i and j are conditionally independent for that cancer type i.e. they are not connected by an edge within the conditional dependence network (Dempster, 1972). Our statistical objectives are to learn a sparse network for each cancer type with an approach that both allows for borrowing of information across cancers, and avoids artefactual differences in network sparsity due to differing sample sizes. We now describe the prior formulation that allows us to achieve these goals.

2.1 Hierarchical shrinkage priors for borrowing strength

We construct a joint prior that both encourages sparsity of the precision matrices within each cancer type and similarity across the cancer types. To formulate this prior, we first divide the elements of each matrix into diagonal (D) and non-diagonal (ND) elements. Let θ_{ij}^c denote the (i, j)th element of the precision matrix $\boldsymbol{\Theta}_c$. Since $\boldsymbol{\Theta}_c$ is symmetric, the unique elements of $\{\boldsymbol{\Theta}_c\}_{c=1}^C$ can be partitioned into C p diagonal elements and $Cp(p-1)/2$ non-diagonal elements as follows:

$$\boldsymbol{\theta_D} = (\theta_{11}^1, \ldots, \theta_{pp}^1, \ldots, \theta_{11}^C, \ldots, \theta_{pp}^C)_{Cp\times 1}$$
$$\boldsymbol{\theta_{ND}} = (\theta_{12}^1, \ldots, \theta_{12}^C, \ldots, \theta_{(p-1)p}^1, \ldots, \theta_{(p-1)p}^C)_{\frac{Cp(p-1)}{2}\times 1}$$

We put independent exponential priors with mean γ^{-1} on the diagonal elements (i.e. on each element of $\boldsymbol{\theta_D}$), and a gamma hyperprior on the parameter γ, with shape parameter α_γ and rate parameter β_γ.

To enable borrowing of strength while estimating sparse versions of the precision matrices for the set of cancers of interest, we introduce the shrinkage parameters λ_1 and λ_2, where λ_1 controls the shrinkage within each cancer type and λ_2 induces similarity across the cancer types. Instead of taking the same λ_1 for each cancer type and the same λ_2 for each pair of cancers, we consider a more flexible scenario where the within-cancer and cross-cancer shrinkage parameters are dependent on cancer type. Let λ_1^c denote the individual shrinkage parameter for cancer type c, $1 \le c \le C$ and $\lambda_2^{cc'}$ denote the cross-penalty across cancer types c and c', $1 \le c < c' \le C$. Following Kyung *et al.* (2010), the conditional prior of $\boldsymbol{\theta_{ND}}$ is given by (1). Here the first term aims to achieve sparsity within each cancer and the second term to achieve similarity across cancers.

$$\pi(\boldsymbol{\theta_{ND}}|\sigma^2) \propto \exp\left(-\frac{1}{\sigma}\sum_{c=1}^{C}\lambda_1^c \sum_{i<j}|\theta_{ij}^c| - \frac{1}{\sigma}\sum_{c<c'}\lambda_2^{cc'}\sum_{i<j}|\theta_{ij}^c - \theta_{ij}^{c'}|\right) \quad (1)$$

Network similarity index: Taking a closer look into the role of the cross-group penalty parameter in the model, it is evident that higher values of $(\lambda_2^{cc'})^2$ encourage more similarity between the networks of cancer c and c' i.e. the higher the value of $(\lambda_2^{cc'})^2$ the closer the network structure (edges) will be between two cancers (see Section 2.2 for a detailed description of the prior structure on the penalty parameters, and our simulations and real data analyses for illustration of their effect). Therefore within a group of related cancers, the similarity between cancer types c and c' can be estimated by the penalty $(\lambda_2^{cc'})^2$, which we term the NSI. Within each group of related cancers, we transform the NSI values of all pairs of cancers to be within the unit interval using a linear monotonic mapping such that the pair of cancers with the lowest and the highest NSI values are mapped to 0 and 1 respectively. We term these transformed NSI values as *normalized network similarity index* (NNSI).

2.2 Incorporating sample size adjustment in the priors

When estimating networks for multiple cancer types using existing methods, network sparsity will vary depending upon sample size, with larger sample sizes generally resulting in denser graphs (more edges). From a pan-cancer context, however, sparsity should not be

dependent on the sample size—since large cancers will overwhelm under-sampled or rare cancers. To counteract this dependency, we design priors using shrinkage parameters to mitigate the sample size effect. Since larger values of λ_1^c encourage more shrinkage of the elements of the precision matrix, and hence more sparsity, our objective is to have a larger prior mean of $(\lambda_1^c)^2$ for the cancer types with larger sample sizes, and a smaller prior mean for rare cancer types with smaller sample sizes. Similarly, the shrinkage parameters $\lambda_2^{cc'}$ encourage similarity between cancer types c and c'.

To achieve these goals, we put gamma priors on the squared terms $(\lambda_1^c)^2$ with shape parameter α_1 and rate parameter β_1^c, for $1 \leq c \leq C$, and gamma priors on $(\lambda_2^{cc'})^2$ with shape parameter α_2 and rate parameter $\beta_2^{cc'}$, for $1 \leq c < c' \leq C$. Following Kling *et al.* (2015), we define β_1^c and $\beta_2^{cc'}$ as follows:

$$\beta_1^c = \frac{\beta_1}{(n_c^e)^2},\ \beta_2^{cc'} = \beta_2 \left\{ \frac{n_c^e + n_{c'}^e}{2 n_c^e n_{c'}^e} \right\}^2,$$

where $\beta_1, \beta_2 > 0$. Let $\bar{n}$ represent the average sample size across the C cancer types. We define the effective sample size of the cth cancer type as $n_c^e = \bar{n}^\delta n_c^{(1-\delta)}$, where $0 < \delta < 1$. Hence, we get the prior mean of $(\lambda_1^c)^2$ to be $\left(\frac{\alpha_1}{\beta_1}\right)(n_c^e)^2$, and the prior mean of $(\lambda_2^{cc'})^2$ to be $\left(\frac{\alpha_2}{\beta_2}\right)\left\{\frac{2n_c^e n_{c'}^e}{n_c^e + n_{c'}^e}\right\}^2$. At $\delta = 0$, $n_c^e = n_c$, so the prior mean depends only on the sample size for cancer type c, while at $\delta = 1$, $n_c^e = \bar{n}$, so the prior mean is equal across all cancer types. As δ approaches 0, relatively more sample size correction is induced by the prior. We recommend $\delta = 0.5$ as a default setting. In order to further characterize the role of δ, we perform extensive simulation studies, which are summarized in Section SK of the Supplementary Materials. We observe that as the value of δ decreases, the true positive rate (TPR), false positive rate (FPR) and estimated number of non-zero edges increases (see Supplementary Tables S12 and S13).

An illustrative example: Consider three cancer types A, B and C with sample sizes $n_A = 50$, $n_B = 100$ and $n_C = 200$. Then for $\delta < 1$, our objective is to have prior mean of $(\lambda_1^A)^2$ to be the smallest and the prior mean of $(\lambda_1^C)^2$ to be the largest, while prior mean of $(\lambda_1^B)^2$ should be in between those values. Also, for $\delta < 1$, for similarity shrinkage parameters, our objective is to have the lowest prior mean among the two cancers with the smallest sample sizes (here A and B) and the highest prior mean among the two cancers with the highest sample sizes (here B and C). The variation of the prior means of the shrinkage parameters for this scenario is plotted as a function of δ in Figure 1. Note that under the proposed way of prior construction, the desired order between the prior means of the shrinkage parameters are maintained for $\delta < 1$. Also note that as δ increases to 1, the penalty terms converge in both scenarios.

2.3 MCMC and edge selection

After introducing latent variables $\mathbf{T}^2$ and $\mathbf{\Omega}^2$, with an appropriate choice of prior, it can be shown that the conditional prior of Equation (1) can be re-expressed as the normal distribution

$$\mathbf{\Theta}_{ND} | \sigma^2, \mathbf{T}^2, \mathbf{\Omega}^2 \sim N_{Cp(p-1)/2}(0, \sigma^2 \mathbf{\Sigma}_{\mathbf{\Theta}_{ND}}),$$

where $\mathbf{\Sigma}_{\mathbf{\Theta}_{ND}}$ can be formulated in terms of $\mathbf{T}^2$ and $\mathbf{\Omega}^2$ (see Supplementary Section SA). This property is important as it allows us to construct a computationally efficient Gibbs sampler.

Since the posterior distribution is not tractable, we rely on Markov Chain Monte Carlo (MCMC) sampling to obtain a sample from the posterior distribution. Specifically, we are able to construct a Gibbs sampler using the posterior conditional distributions for each parameter or set of parameters. Here is a high-level outline of the updates that take place in each iteration of the algorithm. For additional details on the joint posterior, posterior full conditional distributions and sampling steps, see Sections SB–SE in the Supplementary Material.

- Update the ith column of $\mathbf{\Theta}_c$ following the approach in Wang (2012) by sampling a scalar from a gamma distribution and a vector from a multivariate normal, and then applying an appropriate transformation for $c = 1, \ldots, C$ and $i = 1, \ldots, p$.
- Update the latent variable $\mathbf{T}^2$ by sampling the element $(\tau_{ij}^c)^{-2}$ from an inverse-gamma distribution for $c = 1, \ldots, C$ and $1 \leq i < j \leq p$.
- Update the latent variable $\mathbf{\Omega}^2$ by sampling the element $(\omega_{ij}^{cc\prime})^{-2}$ from an inverse-gamma distribution for $1 \leq c < c' \leq C$ and $1 \leq i < j \leq p$.
- Update $(\lambda_1^c)^2$ by sampling from a gamma distribution for $c = 1, \ldots, C$.
- Update $(\lambda_2^{cc\prime})^2$ by sampling from a gamma distribution for $1 \leq c < c' \leq C$.
- Update γ by sampling from a gamma distribution.

Since the shrinkage prior results in sampled values for the precision matrices that have entries, which are very close to 0, but not exactly 0, thresholding is required to represent the sampled precision matrices as sparse. Therefore, to identify selected edges in the estimated networks, we choose a cut-off $\kappa = 0.05$ on the elements of the partial correlation matrices (derived from corresponding precision matrices). Instead of simply thresholding the elements of the posterior mean partial correlation matrices, we consider a slightly different probabilistic approach. From the posterior sample, we specifically calculate the posterior probability of each non-diagonal entry of the partial correlation matrices being greater than the cut-off κ. The

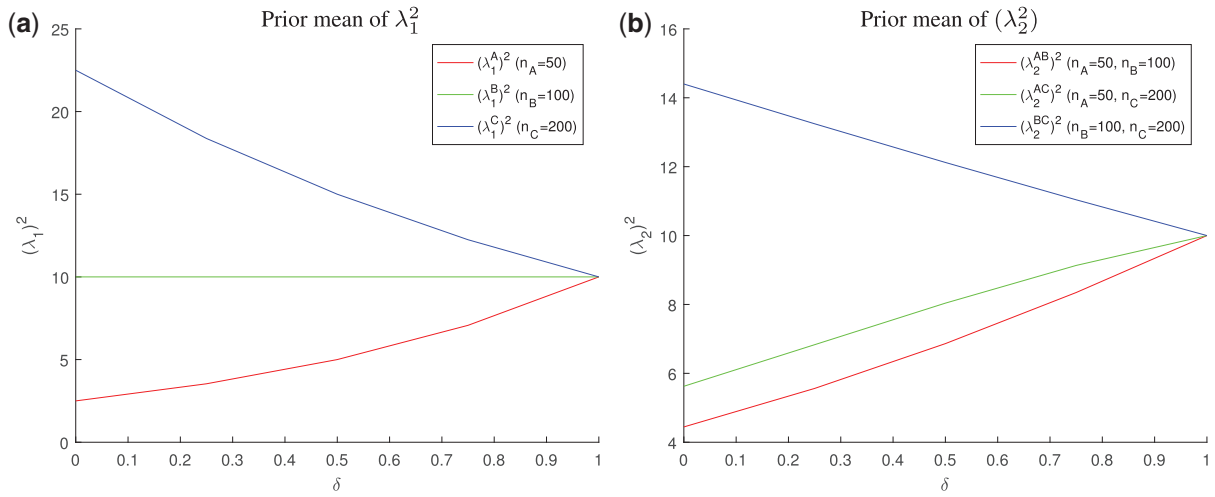

Fig. 1. Prior means of within-cancer (λ_1^2) and cross-cancer (λ_2^2) shrinkage parameters as a function of δ. The (**a**) within-cancer and (**b**) cross-cancer shrinkage parameters have been shown for a scenario with 3 cancer types A, B, C with sample sizes $n_A = 50, n_B = 100, n_C = 200$

edges corresponding to the entries with posterior probability of inclusion greater than 0.5 across all MCMC iterations are then selected (see Supplementary Material Section SF for additional details on the edge selection procedure).

3 Simulation studies

In this section, we compare the performance of the proposed NExUS method to several existing methods based on a simulated dataset. The main focus of this study is understanding the relative performance of all the methods in terms of the True Positive Rate (TPR), False Positive Rate (FPR) and the area under the ROC curve (AUC) when estimating a group of similar networks with high sample size disparity. In the TCGA data analysis (described in detail in Section 4), we consider 5 cancers in the pan-gynecological (pan-gynae) group consisting of breast invasive carcinoma (BRCA), cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC), ovarian serous cystadenocarcinoma (OV), uterine corpus endometrial carcinoma (UCEC) and uterine carcinosarcoma (UCS), with sample sizes 879, 171, 428, 404 and 48 respectively, where each group has 75 proteins. Among the pan-cancer groups considered in Section 4, the pan-gynae group has the largest variation in terms of sample sizes, which range from 48 to 879. To assess the performance of NExUS along with the performance of other existing methods, we generate a simulated dataset to mimic this real pan-gynae dataset. First we estimate the precision matrices of all 5 cancers in the pan-gynae group using NExUS. Once the precision matrices are estimated, synthetic data is generated in such a way that the sample size of the synthetic data generated from each estimated precision matrix is the same as the sample size in the real data. Note that once the precision matrices are estimated from the real data, the precision matrices obtained are considered as the true precision matrices, and further samples are generated from those precision matrices. Therefore, the dataset is not directly derived from the underlying model of NExUS. We denote the groups in the simulated dataset which are intended to mimic BRCA, CESC, OV, UCEC and UCS by C_1, C_2, C_3, C_4 and C_5, respectively.

For the methods comparison, we consider a variety of Bayesian and frequentist methods. In the Bayesian framework, in addition to the proposed NExUS method, we consider the Bayesian graphical lasso (BGL, Wang, 2012). The values of the tuning parameters for BGL are the same as considered in Wang (2012). For edge selection, however, we use the same thresholding approach as NExUS. For NExUS, we use the following hyperparameter setting: $\alpha_1 = 1, \alpha_2 = .1, \beta_1 = .1\bar{n}^2, \beta_2 = 1\bar{n}^2, \alpha_\gamma = 1,$ and $\beta_\gamma = 1$. For both Bayesian approaches, we discard the first 5000 iterations as burn-in, and use the following 15 000 iterations as the basis for inference.

In the frequentist framework, we consider both single graph and joint estimation methods. Specifically, the single graph methods we compare to are the graphical lasso (FreqGL Friedman *et al.*, 2008), adaptive graphical lasso (FreqAdapGL) and the SCAD penalized lasso (Fan *et al.*, 2009). The joint estimation methods we consider are the group graphical lasso (GGL), fused graphical lasso (FGL) (Danaher *et al.*, 2014) and Sparse Inverse Covariance Selection (SICS Kling *et al.*, 2015). We also consider the sample-size adjusted versions of FGL and GGL, denoted by FGL(ADJ) and GGL(ADJ). For GGL and FGL, to select the best value of the penalty parameters, we take an equidistant grid of 100×100 over the possible values of λ_1 and λ_2 in $[0.01, 1] \times [0.01, 1]$ and estimate the precision matrices by those methods. The best possible values of the tuning parameters are selected using AIC. For SICS, we consider the ranges of λ_1, λ_2 and δ to be 0.01–1, 0–0.02 and 0–1, respectively. We take an equidistant grid over the range of values, and select the values of tuning parameters using AIC. Once the values of the tuning parameters are selected, the rest of the implementation exactly follows the procedure described (Kling *et al.*, 2015) using their provided code (for details see Section SK.1 in the Supplementary Material).

To compare the performance of these methods, we primarily rely on the area under the ROC curve (AUC), as it provides a single summary of network learning accuracy. In Table 1, we summarize the AUC values for the various methods for each of the five simulated graph structures. The reported value is the mean of AUC values estimated from 10 simulated datasets. The standard error values are reported inside parentheses. For SICS, we note that when varying the values of the tuning parameters, the (TPR, FPR) coordinates obtained are inconsistent (possibly due to the bootstrap resampling technique which is part of the selection procedure), yielding a very small AUC value. Therefore, we did not include the AUC value for SICS in Table 1. We note that NExUS is consistently the top performing method. FGL and FGL(ADJ) perform well for groups with smaller sample sizes, e.g. C_2 and C_5. However, their performance is weaker for other cancer groups. Some of the single graph methods (BGL and FreqGL) become competitive when applied to the group with the largest sample size, but lack power in the classes with smaller sample sizes. In Table 2, we report the true and the estimated proportions of non-zero edges from the simulation study for NExUS and other joint graphical network estimation methods only, i.e. SICS, FGL, FGL(ADJ), GGL and GGL(ADJ), estimated from 10 simulated datasets. In general, the estimated proportions of non-zero elements obtained using NExUS are closest to the true proportions of non-zero elements. Unlike FGL, GGL and SICS, the FPR obtained using NExUS is always less than 0.01, along with competitive TPR values. Note that except NExUS, all other methods yield very large FPR values for the cancer type with the smallest sample size (i.e. C_5). Also, for C_5, the estimated proportion of non-zero edges is larger for FGL, GGL and SICS compared to NExUS.

We also summarize the AUC20 values for the various methods based on 10 simulated data-sets in Supplementary Table S11. In Supplementary Tables S12 and S13, we show how the estimated sparsity, TPR, FPR and the estimated number of non-zero edges vary with tuning parameter δ. Both TPR and FPR increase as δ decreases to 0. In addition, we consider another simulation study in Supplementary Section SG, along with sensitivity analysis of the tuning parameters.

4 NExUS analyses of pan-cancer proteomic data

We demonstrate the utility of NExUS to conduct pan-cancer analyses of proteomics data collected across various cancer types.

Table 1. Mean AUC values of graph structure learning across 10 simulated datasets, with standard deviations in parentheses

Methods	C_1 ($n = 879$)	C_2 ($n = 171$)	C_3 ($n = 428$)	C_4 ($n = 404$)	C_5 ($n = 48$)
NExUS	**0.99 (0.00)**	**0.98 (0.00)**	**0.99 (0.00)**	**0.99 (0.00)**	**0.98 (0.01)**
FGL	0.91 (0.01)	0.91 (0.01)	0.90 (0.01)	0.89 (0.01)	0.96 (0.01)
FGL(ADJ)	0.92 (0.00)	0.92 (0.01)	0.90 (0.01)	0.89 (0.01)	0.96 (0.00)
GGL	0.92 (0.01)	0.93 (0.01)	0.93 (0.01)	0.91 (0.01)	0.91 (0.01)
GGL(ADJ)	0.91 (0.00)	0.89 (0.01)	0.88 (0.01)	0.87 (0.01)	0.90 (0.01)
BGL	0.99 (0.00)	0.92 (0.01)	0.98 (0.01)	0.97 (0.00)	0.77 (0.02)
FreqGL	**0.99 (0.00)**	0.91 (0.02)	0.98 (0.01)	0.97 (0.01)	0.75 (0.03)
FreqAdapGL	0.77 (0.00)	0.79 (0.03)	0.88 (0.01)	0.93 (0.01)	0.50 (0.00)
SCAD	0.80 (0.01)	0.87 (0.02)	0.96 (0.01)	0.95 (0.01)	0.50 (0.00)

Note: The results for the top performing method for each cancer type are highlighted in bold.

Table 2. The true and estimated proportions of non-zero edges for each method

Methods	C_1 ($n = 879$)	C_2 ($n = 171$)	C_3 ($n = 428$)	C_4 ($n = 404$)	C_5 ($n = 48$)
TRUE	0.07	0.04	0.05	0.06	0.04
NExUS	0.07 (0.88/0.00)	0.03 (0.75/0.00)	0.04 (0.80/0.00)	0.05 (0.80/0.00)	0.03 (0.70/0.00)
SICS	0.01 (0.07/0.00)	0.12 (0.65/0.09)	0.01 (0.14/0.00)	0.01 (0.17/0.00)	0.16 (0.49/0.15)
FGL	0.07 (0.73/0.01)	0.06 (0.85/0.01)	0.06 (0.82/0.01)	0.07 (0.82/0.01)	0.14 (0.89/0.10)
FGL(ADJ)	0.09 (0.93/0.01)	0.06 (0.83/0.02)	0.06 (0.83/0.01)	0.07 (0.75/0.01)	0.06 (0.92/0.02)
GGL	0.07 (0.83/0.01)	0.08 (0.81/0.05)	0.06 (0.89/0.01)	0.07 (0.89/0.02)	0.22 (0.80/0.19)
GGL(ADJ)	0.08 (0.94/0.01)	0.03 (0.62/0.00)	0.05 (0.85/0.01)	0.06 (0.80/0.01)	0.00 (0.05/0.00)

Note: The TPR and FPR are provided in '(TPR/FPR)' format.

We focus on four groups of related cancers: pan-gynecological (pan-gynae), pan-kidney, pan-squamous and pan-gastrointestinal (pan-GI). For each group of related cancers, we perform separate NExUS analyses. Gynecologic cancers have similar embryonic origins, since female hormones influence their development (Berger *et al.*, 2018). The pan-gynae group consists of breast invasive carcinoma (BRCA), cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC), ovarian serous cystadenocarcinoma (OV), uterine corpus endometrial carcinoma (UCEC) and uterine carcinosarcoma (UCS). Since kidney cancers originate from the cells of the outer layer of the kidney (the renal cortex), we consider the following cancers in the pan-kidney group: kidney chromophobe (KICH), kidney renal clear cell carcinoma (KIRC) and kidney renal papillary cell carcinoma (KIRP). Squamous cell carcinomas (SCCs) arise from the epithelia of the aerodigestive and genitourinary tracts and share some histological characteristics that can be used for predicting the site of origin, clinical behavior, cause, prognosis, or optimal therapies (Campbell *et al.*, 2018). In the pan-squamous group, we consider esophageal squamous-carcinoma (ESCA(sq)), head and neck squamous cell carcinoma (HNSC) and lung squamous cell carcinoma (LUSC). Adenocarcinomas of the GI tract share similar endodermal developmental origins along with exposure to common insults that promote the tumor formation (Liu *et al.*, 2018); hence the pan-GI group consists of colon/rectum adenocarcinoma (CORE), esophageal adeno-carcinoma (ESCA(ad)) and stomach adenocarcinoma (STAD).

Our proteomics data arises from The Cancer Proteome Atlas (TCPA, Li *et al.*, 2013), which provides protein abundance data for TCGA samples, and consists of RPPA-based quantifications using antibodies that cover functions including proliferation, DNA damage, polarity, vesicle function, EMT, invasiveness, hormone signaling, apoptosis, metabolism, immunological and stromal function, as well as other critical cellular signaling pathways (Akbani *et al.*, 2014). The proteins profiled allow for a focused exploration of the various functional mechanisms underlying oncogenic processes across tumor types. Here we focus on the set of 75 proteins which were profiled across all of the cancers considered here.

A challenge in analyzing this data is that sample sizes for different cancers vary considerably, from 48 for UCS, to 879 for BRCA (see Supplementary Table S9 for the number of samples available for each cancer type). Instead of estimating the proteomic networks for each cancer separately, joint estimation and borrowing of strength across the networks gives us a broader picture of the similarities and dissimilarities among the cancers belonging to the same group. Specifically, the cross-group penalty terms in NExUS help in identifying some of the weak signals (edges) for the cancer networks with smaller sample sizes when those edges are shared by other cancers of the same group, and the sample size adjustment encourages networks with more similar levels of sparsity across cancers belonging to the same pan-cancer group. In addition to learning the network structures per cancer type, we also are able to estimate the global proteomic network-based similarities and the pathway-specific similarities among the cancers belonging to the same group.

For the analysis, we use the same values of the prior parameters as in the simulation studies. We perform 20 000 iterations, discarding the first 5000 iterations as burn-in for each pan-cancer group. We adopt the same posterior edge selection strategy as in the simulation studies. The posterior selected networks for each cancer type are shown in Supplementary Figures S9–S22. Here we focus on high-level take-aways of these pan-cancer analyses, in particular, on the global and pathway-specific similarity measures.

4.1 Global proteomic similarity between cancers

To understand the extent of shared network structure among the cancers belonging to the same group, we estimate the normalized network similarity indexes (NNSI) for each pair of cancers. We plot the NNSI in Figure 2 for all pan-cancer groups. In the pan-gynae group we observe 3 distinct clusters among the pairs of cancers based on NNSI: CESC-UCS are the closest; OV-UCS, CESC-OV, UCEC-UCS, CESC-UCEC and OV-UCEC belong to the second cluster with intermediate NNSI values; and the third cluster, which represents the most distant pairs of cancers, contains BRCA-OV, BRCA-UCS, BRCA-CESC and BRCA-UCEC. These results show that the protein network for the rare cancer UCS is more similar to that of CESC and OV cancers as compared to other gynae cancers. BRCA has the least network similarity with other gynae cancers, suggesting BRCA has distinctive network features. In the pan-kidney group, the KICH and KIRP networks are the closest, while KIRC and KIRP networks are least similar. For the pan-squamous and pan-GI groups, ESCA(sq)-LUSC and ESCA(ad)-STAD are most similar while HNSC-LUSC and CORE-STAD are the least similar. To provide further insight into the meaning of the network similarity index, we examined the relationship between it and the L_1 distance between pairs of precision matrices, and found the relation to be approximately linear, where pairs of cancer with relatively high values of the network similarity index have relatively lower L_1 distances between them (see Supplementary Fig. S6).

Finally, to visualize the extent of shared structure across each group of cancers, we construct heatmaps of the posterior edge inclusion probabilities (Fig. 2). The ordering of the cancer types, which was determined by hierarchical clustering of the edge probabilities, supports the NNSI results, with pairs of cancer scoring higher on the NNSI grouped more closely within the estimated hierarchy.

4.2 Pathway-specific similarity between cancers

To understand important aspects of the network similarity between cancer types, we conducted a deeper investigation of the proteomic networks using available pathway information. We focus our exploration on 12 well-established curated pathways with translational relevance (Akbani *et al.*, 2014; Ha *et al.*, 2018): apoptosis, cell cycle, DNA damage response, EMT, hormone receptor, hormone signaling breast, PI3K/AKT, RAS/MAPK, RTK, TSC/mTOR, breast reactive and core reactive. A few of the proteins profiled are shared by more than one pathway (see Supplementary Table S10). One of our main motivations for deconvolving proteomic activity into pathways is to investigate which pathways are activated in each of the cancer types, and thereby gain a better understanding of the mechanistic and regulatory sharing of information between proteomic pathways.

To quantify pathway-specific activation, we estimate the proportion of shared edges within each pathway and across each pair of

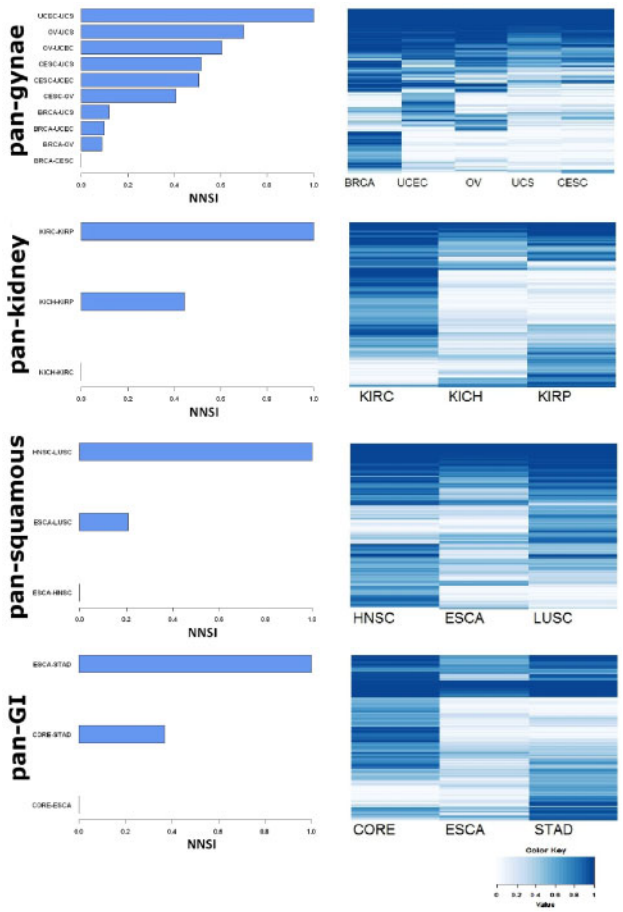

Fig. 2. (left) Bar plot of normalized network similarity index (NNSI) for each pair of cancers in the pan-gynae, pan-kidney, pan-squamous and pan-GI groups. (right) Heatmap of posterior edge probabilities for each cancer in the given group. For each cancer group, out of a possible $75 * (75 - 1)/2$ edges, we include only rows for the probabilities for edges selected in at least one graph in the group

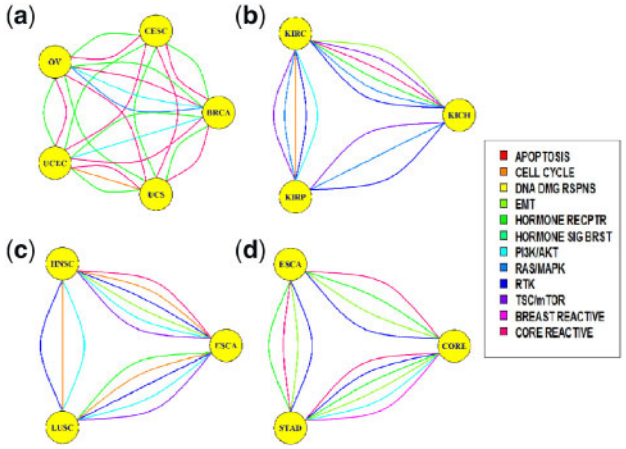

Fig. 3. Shared pathway activity: Proportion of shared edges within pathways across all pairs of cancers, with edges denoting top 20% (for gynae) and 40% (for other cancer groups) of all pathways for each pair of cancers within each group are plotted for (**a**) pan-gynae, (**b**) pan-kidney, (**c**) pan-squamous and (**d**) pan-GI cancer groups

pathways for all cancers belonging to the same group. The heatmap for the proportion of shared edges within and across pathways for each pair of cancers is given in Supplementary Figures S7 and S8. In general, the proportion of shared edges are higher within pathways than across different pathways, which is along expected lines. This makes biological sense, as core pathway activities are likely to be preserved across different cancer lineages. To identify the strongest shared pathway activity for each pair of cancers within each group, we computed the percentage of shared edges within each pathway for each pair of cancer types. In Figure 3, we include edges based on the ranking of these percentages of edges shared: specifically, we include a colored link between cancer types in the figure for the 40% highest sharing percentages across all pathways and cancer pairs for all cancer types, except pan-gynae. As the pan-gynae group has a larger number of cancer types included, we focus on the top 20% there to improve legibility.

For the pan-gynae group, the hormone receptor pathway and the core reactive pathway are the most active shared pathways across the pairs of cancers. The hormone receptor pathway is important for OV cancer, where hormonal-based systemic therapies are used to treat ovarian stromal tumors (Dacheux *et al.*, 2013). As UCS is a rare cancer, it is challenging to pinpoint its active pathways and to draw network-based inference. We are able to identify the cell cycle pathway as one of the top three actively shared pathways between UCS and UCEC. This is clinically relevant as a cell cycle pathway inhibitor has been identified as one of the therapeutic options for UCS (Cherniack *et al.*, 2017). Finally, PI3K/AKT pathway activity is shared between the BRCA and UCEC as well as BRCA and OV cancer types. TCGA Research Network (2012b) mentions many components of PI3K pathway were amplified in basal-like breast cancers.

For the pan-kidney group, our analysis identifies the TSC/mTOR, RAS/MAPK and RTK pathways as the top 3 actively shared pathways. We also observe shared pathway activity for the PI3/AKT, EMT, core reactive, cell cycle and hormone receptor pathways. Chen *et al.* (2016) and TCGA Research Network (2013) recognized the PI3K/AKT and mTOR pathways to be active for renal cell carcinoma (RCC) and clear cell renal cell carcinoma (ccRCC). In addition, Gibbons and Creighton (2018) found EMT as one of the important activated pathways in kidney cancers.

For the pan-squamous group, the cell cycle, RTK and PI3K/AKT pathways are activated between all pairs of cancers. Other activated pathways are TSC/mTOR, EMT, core reactive and hormone receptor. TCGA Research Network (2015) and TCGA Research Network (2012a) described the PI3K, RAS and AKT pathways as major pathways influencing HNSC and LUSC, respectively. TCGA Research Network (2015) and TCGA Research Network (2012a) also reported that the cell cycle and RTK pathways play an important role in HNSC and LUSC. Finally, for the pan-GI group, we find the hormone receptor, RTK, EMT and core reactive pathways as the top 3 activated shared pathways.

5 Conclusion and remarks

In this article, we propose NExUS, a fully Bayesian method for estimating a group of related networks using Gaussian graphical modeling. The incorporation of a penalty on dissimilarity across the precision matrices in the prior specification enables borrowing of strength across the networks. This aspect of the model is especially helpful for identifying edges in the networks with smaller sample sizes. In addition, the resulting NSI helps order the relative closeness of each pair of networks within the set analyzed. Another novel feature of NExUS is the inclusion of a sample size correction which encourages similar sparsity levels across networks with different sample sizes. Our simulation studies show that NExUS outperforms other existing methods for individual and joint estimation of networks. NExUS is motivated by the TCGA-based RPPA dataset, wherein we estimate the pan-cancer proteomic networks for 4 groups of related cancers lineages. The existence of rare cancers (e.g. UCS in pan-gynae, and ESCA(ad) in pan-GI) makes the proposed model particularly appropriate for this application.

We developed the NExUS model under Gaussian assumptions and applied it for datasets containing a moderate number of proteins obtained using targeted profiling. In the future, it can be modified for skewed data and discrete variables, potentially following the methods proposed in Bhadra *et al.* (2018). The method would then be applicable to non-normal datasets such as mutation or copy number variation. In the current model, the prior means of the penalty terms are only functions of the sample sizes, however, characteristics of the datasets and cancer types might be incorporated into the model by making the prior means of the penalty terms dependent on those attributes as well. NExUS could be also extended to handle the scenario where the set of proteins measured differs across the groups; how to handle this missing data is a non-trivial challenge, however. Finally, more efficient algorithms could be developed to allow NExUS to scale to a larger number of variables, enabling an application to datasets such as the Clinical Proteomic Tumor Analysis Consortium (CPTAC), which was acquired using untargeted proteomics, and therefore has a much higher dimensionality.

Funding

This work was supported by the National Institutes of Health [P30CA016672 to C.P., K.D. and R.A., P30CA046592 to V.B., SPORE CA140388 to K.D., EDRN CA086368 to K.D., CCTS TR000371 to K.D., U24CA210950 to R.A., U24 CA210949 to R.A., R01CA194391 to V.B., R21CA220299-01A1 to V.B., R01CA160736 to V.B.]; National Science Foundation [DMS1463233 to P.D. and V.B.]; Department of Defense Congressionally Directed Medical Research Programs [W81XWH-16-1-0237 to R.A.]; MD Anderson institutional Moonshot funding [to C.P. and K.D]; Cancer Prevention and Research Institute of Texas [RP150521 to C.P.]; and Rogel Cancer Center Start-up funds [to V.B.].

Conflict of Interest: none declared.

References

Akbani,R. *et al.* (2014) A pan-cancer proteomic perspective on The Cancer Genome Atlas. *Nat. Commun.*, **5**, 3887.

Baladandayuthapani,V. *et al.* (2014) Bayesian sparse graphical models for classification with application to protein expression data. *Ann. Appl. Stat.*, **8**, 1443–1468.

Berger,A. *et al.* (2018) A comprehensive pan-cancer molecular study of gynecologic and breast cancers. *Cancer Cell*, **33**, 690–705.

Bhadra,A. *et al.* (2018) Inferring network structure in non-normal and mixed discrete-continuous genomic data. *Biometrics*, **74**, 185–195.

Campbell,J. *et al.* (2018) Genomic, pathway network, and immunologic features distinguishing squamous carcinomas. *Cancer Cell*, **23**, 194–212.

Carro,M. *et al.* (2010) The transcriptional network for mesenchymal transformation of brain tumours. *Cancer Res.*, **463**, 318–325.

Chen,F. *et al.* (2016) Multilevel genomics-based taxonomy of renal cell carcinoma. *Cell Rep.*, **14**, 2476–2489.

Chen,F. *et al.* (2017) Pan-urologic cancer genomic subtypes that transcend tissue of origin. *Nat. Commun.*, **8**, 1–15.

Cherniack,A. *et al.* (2017) Integrated molecular characterization of uterine carcinosarcoma. *Cancer Cell*, **31**, 411–423.

Cho,A. *et al.* (2016) MUFFINN: cancer gene discovery via network analysis of somatic mutation data. *Genome Biol.*, **17**, 129.

Creixell,P. *et al.* (2015) Pathway and network analysis of cancer genomes. *Nat. Methods*, **12**, 615–621.

Dacheux,E. *et al.* (2013) Hormone receptors in serous ovarian carcinoma: prognosis, pathogenesis, and treatment considerations. *PLoS One*, **8**, e67313.

Danaher,P. *et al.* (2014) The joint graphical lasso for inverse covariance estimation across multiple classes. *J. R. Stat. Soc. B*, **76**, 373–397.

Dempster,A.P. (1972) Covariance selection. *Biometrics*, **28**, 157–175.

Fan,J. *et al.* (2009) Network exploration via the adaptive LASSO and SCAD penalties. *Ann. Appl. Stat.*, **3**, 521–541.

Friedman,J. *et al.* (2008) Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, **9**, 432–441.

Gibbons,D. and Creighton,C. (2018) Pan-cancer survey of epithelial-mesenchymal transition markers across The Cancer Genome Atlas. *Dev. Dyn.*, **247**, 555–564.

Gill,R. *et al.* (2014) Differential network analysis in human cancer research. *Curr. Pharm. Des.*, **20**, 4–10.

Ha,M. *et al.* (2018) Personalized integrated network modeling of the cancer proteome atlas. *Sci. Rep.*, **8**, 14924.

Hanahan,D. and Weinberg,R. (2000) The hallmarks of cancer. *Cell*, **100**, 57–70.

Hristov,B. and Singh,M. (2017) Network-based coverage of mutational profiles reveals cancer genes. *Cell Syst.*, **5**, 221–229.

Kling,T. *et al.* (2015) Efficient exploration of pan-cancer networks by generalized covariance selection and interactive web content. *Nucleic Acids Res.*, **43**, e98.

Kundu,S. *et al.* (2019) Efficient Bayesian regularization for graphical model selection. *Bayesian Anal.*, **14**, 449–476.

Kyung,M. *et al.* (2010) Penalized regression, standard errors, and Bayesian lassos. *Bayesian Anal.*, **5**, 369–412.

Lauritzen,S. (1996) *Graphical Models*. Oxford, Claredon.

Li,J. *et al.* (2013) TCPA: a resource for cancer functional proteomics data. *Nat. Methods*, **10**, 1046–1047.

Lin,Z. *et al.* (2017) On joint estimation of Gaussian graphical models for spatial and temporal data. *Biometrics*, **73**, 769–779.

Liu,Y. *et al.* (2018) Comparative molecular analysis of gastrointestinal adenocarcinomas. *Cancer Cell*, **33**, 721–735.

Peterson,C. *et al.* (2015) Bayesian inference of multiple Gaussian graphical models. *J. Am. Stat. Assoc.*, **110**, 159–174.

Sonabend,A. *et al.* (2014) The transcriptional regulatory network of proneural glioma determines the genetic alterations selected during tumor progression. *Cancer Res.*, **74**, 1440–1451.

Tamborero,D. *et al.* (2013) Comprehensive identification of mutational cancer driver genes across 12 tumor types. *Sci. Rep*, **3**, 1–9.

Tan,L. *et al.* (2017) Bayesian inference for multiple Gaussian graphical models with application to metabolic association networks. *Ann. Appl. Stat.*, **11**, 2222–2251.

TCGA Research Network (2012a) Comprehensive genomic characterization of squamous cell lung cancers. *Nature*, **489**, 519–525.

TCGA Research Network (2012b) Comprehensive molecular portraits of human breast tumours. *Nature*, **490**, 61–70.

TCGA Research Network (2013) Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature*, **499**, 43–49.

TCGA Research Network (2015) Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature*, **517**, 576–582.

Wang,E. *et al.* (2015) Predictive genomics: a cancer hallmark network framework for predicting tumor clinical phenotypes using genome sequencing data. *Semin. Cancer. Biol.*, **30**, 4–12.

Wang,H. (2012) Bayesian graphical lasso models and efficient posterior computation. *Bayesian Anal.*, 7, 867–886.

Wang,H. (2015) Scaling it up: stochastic search structure learning in graphical models. *Bayesian Anal.*, **10**, 351–377.

Weinstein,J. *et al.* (2013) The Cancer Genome Atlas pan-cancer analysis project. *Conf. Proc. IEEE Eng. Med. Biol. Soc.*, **45**, 1113–1120.

Whittaker,J. (1990) *Graphical Models in Applied Multivariate Statistics*. Wiley, Chichester.

Yuan,M. and Lin,Y. (2007) Model selection and estimation in the Gaussian graphical model. *Biometrika*, **94**, 19–35.