*The Annals of Applied Statistics*
2016, Vol. 10, No. 3, 1373–1396
DOI: 10.1214/16-AOAS939


# DETECTION OF EPIGENOMIC NETWORK COMMUNITY ONCOMARKERS

BY THOMAS E. BARTLETT[1] AND ALEXEY ZAIKIN[2]

*University College London*

In this paper we propose network methodology to infer prognostic cancer biomarkers based on the epigenetic pattern DNA methylation. Epigenetic processes such as DNA methylation reflect environmental risk factors, and are increasingly recognised for their fundamental role in diseases such as cancer. DNA methylation is a gene-regulatory pattern, and hence provides a means by which to assess genomic regulatory interactions. Network models are a natural way to represent and analyse groups of such interactions. The utility of network models also increases as the quantity of data and number of variables increase, making them increasingly relevant to large-scale genomic studies. We propose methodology to infer prognostic genomic networks from a DNA methylation-based measure of genomic interaction and association. We then show how to identify prognostic biomarkers from such networks, which we term "network community oncomarkers". We illustrate the power of our proposed methodology in the context of a large publicly available breast cancer dataset.

**1. Introduction.** Complex systems which can be modelled as networks are ubiquitous. Well-known examples include social/communication networks [Beguerisse-Díaz et al. (2014)] and economic networks [Saavedra et al. (2014)], as well as many others in the biological sciences such as ecological networks [Nandi, Sumana and Bhattacharya (2014)], gene networks [Li and Wang (2014), Wei and Pan (2010)], protein networks [Mardia (2013), Tran and Kwon (2013)] and metabolic networks [Reznik, Watson and Chaudhary (2013)]. Over the past few years in cell biology, much focus has shifted from investigation of individual genes, to pathways of genes, to gene networks. The interest in novel methodology for network analysis in cell biology follows from the recognition that examining the way genes work in groups often yields more accurate inference of biological processes.

The problem of finding community structure in networks has been studied for many years. Important applications of this problem include identifying groups of

Received July 2015; revised March 2016.

[1]Supported in part by EPSRC Grant no. EP/M507970/1 and previously by EPSRC and MRC via UCL CoMPLEX.

[2]Supported in part by the Deanship of Scientifc Research (DSR), King Abdulaziz University (KAU), Jeddah, under Grant No. (86-130-35-RG), and from the Russian Foundation for Basic Research (14-02-01202, 13-02-00918).

*Key words and phrases.* Computational biology, stochastic networks, community detection, epigenomics.

friends or co-workers in social networks, as well as identifying functional subnetwork modules in biological networks [Girvan and Newman (2002)]. In the biological setting, genes can be viewed as acting together as part of "subnetwork modules", which are functional units with specific biological roles [Shen-Orr et al. (2002)]. Indeed, it has been demonstrated recently that such modularity is a natural and even inevitable result of evolutionary pressures [Clune, Mouret and Lipson (2013)]. This is because modularity minimises network connectivity cost whilst maximising performance, and thus it represents the most parsimonious and efficient type of network structure for biological networks such as these. Furthermore, considering groups of genes defined together as subgraphs can lead to big increases in statistical power, aiding discovery of biological phenomena [Jacob, Neuvial and Dudoit (2012), Li and Li (2010), Peng et al. (2010)]. Therefore, it is relevant to both the biological and statistical modelling to consider the group behaviour of genes in this way. Hence, this viewpoint of modular genomic network structure is fundamental to the methodology we propose here. Epigenetic patterns are gene-regulatory patterns, meaning that they influence the activity of particular genes, among other phenomena [Jones (2012)]. Epigenetic information can be modulated during the lifetime of an organism by environmental cues [Christensen et al. (2009), Cooney (2007), Feinberg, Ohlsson and Henikoff (2006)]. As such, epigenetics can be considered to be an interface between the genome and the environment, and consequently also a conduit for environmental risk factors. Alterations in the epigenetic pattern DNA methylation are among the earliest changes in human carcinogenesis [Feinberg, Ohlsson and Henikoff (2006)], and hence DNA methylation patterns are expected to yield important prognostic information useful for biomarker development. DNA methylation patterns are thought promising for biomarker development in a wide variety of physiological systems and organs [Fleischer et al. (2014), Gao et al. (2013), Kishida et al. (2012), Van Hoesel et al. (2013), Verschuur-Maes, de Bruin and van Diest (2012), Kang et al. (2001, 2003), Bhagat et al. (2012), Luo et al. (2014), Maekawa et al. (2013), Navarro et al. (2012), Yamamoto et al. (2012)].

It is well established that DNA methylation plays an important role in gene regulation, and hence DNA methylation patterns often reflect gene regulatory behaviour [Jones (2012)]. Changes in DNA methylation are highly stochastic. The timescale over which these changes take place is much faster than DNA mutations can arise, but much slower than the transient and periodically varying activity of individual genes, and this timescale is ideal for biomarker development. DNA methylation data are extremely noisy; however, statistics which summarise DNA methylation patterns at the gene level have been shown to have much utility as analytical tools [Bartlett et al. (2013)]. It has been shown previously that DNA methylation can serve as a surrogate measure of genomic-regulatory action [Brocks et al. (2014)]. Hence, DNA methylation measurements are a natural basis from which to construct genomic regulatory and related networks. As a cancer progresses, its signalling and control networks are rearranged ("rewired"), leading

to genomic changes which are advantageous for the cancer [Barabási and Oltvai (2004)]. Previous research has found that patient survival outcome in breast cancer can be predicted well by network models of this rewiring based on gene expression data [Taylor et al. (2009)]. Hence, network models based on DNA methylation measurements are a very promising basis for the development of prognostic biomarkers.

Statistical network models are a parsimonious way to represent and analyse large numbers of variables and samples. They are efficient analytical tools appropriate for the very large datasets which are produced by the latest technologies in cell biology. When carrying out modelling of this type, it is important to balance statistical fidelity with computational efficiency. The "stochastic blockmodel" (SBM) [Bickel and Chen (2009), Holland, Laskey and Leinhardt (1983)] is an efficient network model which has been widely studied and is well understood, and hence it is a good basis for our proposed methodology. Under the SBM, there is a greater probability of observing an edge (or interaction) between a pair of nodes if they are in the same block, or community. The Newman–Girvan modularity [Newman and Girvan (2004)] quantifies the extent to which network edges are observed between community members, for a particular assignment of nodes to communities, compared to the expected number of edges between community members if there were no community structure present. It can be shown that, under certain conditions, fitting the stochastic blockmodel is equivalent to maximising the Newman–Girvan modularity over a network, and that these are both equivalent to spectral clustering [Bickel and Chen (2009), Riolo and Newman (2012)]. We use spectral clustering as an efficient computational algorithm for fitting the SBM.

It has also been shown recently that, under reasonable assumptions, the SBM can be used to represent any network as a "network histogram", whatever the generating mechanism of that network. Further, the network histogram provides a heuristic method to estimate the optimum number of blocks, or clusters, which a valid blockmodel representation of the network may contain. This is important and useful because it means that the blockmodel can be used to identify an unknown number of communities, or functional subnetwork modules, in a biological network. Genomic networks are typically scale-free, which means that they exhibit a power-law degree distribution [Wagner (2002)]. Further, they are thought to be hierarchical [Barabási and Oltvai (2004), Palla, Lovász and Vicsek (2010)], displaying multi-scale properties. This means that different functional organisation is visible at different granularities, or scales. We use the network histogram method [Olhede and Wolfe (2014)] to estimate the optimal granularity at which to identify communities, or functional subnetwork modules, in our prognostic networks by fitting the SBM.

The main contribution of this work is to propose a well-integrated, and well-validated, statistical methodology for detecting biomarkers from the biological viewpoint of modular genomic network structure using DNA-based measurements of genomic regulatory patterns. To do this, we show how to integrate our previously

proposed DNA methylation-based measure of interaction or association between pairs of genes, the "DNA methylation network interaction measure" [Bartlett, Olhede and Zaikin (2014)], into a multi-stage pipeline to construct prognostic network community-based biomarkers. This leads to our novel and generally applicable statistical methodology; we present the multiple stages of this methodology sequentially here, and thus this paper is organised as follows. In Section 2, we outline our previously proposed DNA methylation network interaction measure [Bartlett, Olhede and Zaikin (2014)], and we show how to use this measure to infer prognostic genomic networks. An edge between a pair of genes/nodes in these networks indicates that the strength of interaction or association between those genes is associated with disease progression. Also, in Section 2, we show how to identify prognostic biomarkers from such networks using community detection to identify subnetwork modules within the network. These communities are groups of nodes/genes among which there is a high density of prognostic interactive or associative behaviour, and we term them "network community oncomarkers". In Section 3, we demonstrate the utility of our proposed methodology in the context of a large, publicly available breast cancer dataset. To do so, we use each network community oncomarker to calculate a one-number prognostic score for each patient, and we use these scores to classify patients one by one into prognostic groups. Also, in Section 3, we show that among the genes of the network community oncomarkers, the DNA methylation network interaction measure is associated with co-regulatory behaviour as measured by gene expression, justifying these findings in terms of biological function.

**2. Proposed methodology.** An overview of our proposed methodology appears in Figure 1, following which component parts of this methodology are described in detail.

We note that, in principle, each of the steps illustrated in Figure 1 could be replaced with alternative choices of methodology.

2.1. *DNA methylation network interaction measure.* DNA methylation is a chemical modification to DNA which may occur at numerous locations within a gene: the pattern of these modifications within a gene forms a "DNA methylation profile". Using canonical correlation analysis (CCA) [Hotelling (1936)], we previously proposed a statistic [Bartlett, Olhede and Zaikin (2014)] which measures the strength of interaction or association between a pair of genes (network nodes) in a single sample/patient based on DNA methylation profiles (Figure 2). This statistic quantifies the extent to which the DNA methylation profiles of a pair of genes explain each other. It is based only on measurements of the DNA methylation profiles of that pair of genes, and it acts as a surrogate for a measure of the extent to which this pair of genes behave interactively or associatively. Such behaviour may include transcriptional regulation or co-regulation, or other types of biochemical interaction, influencing gene expression levels, isoforms and the presence of

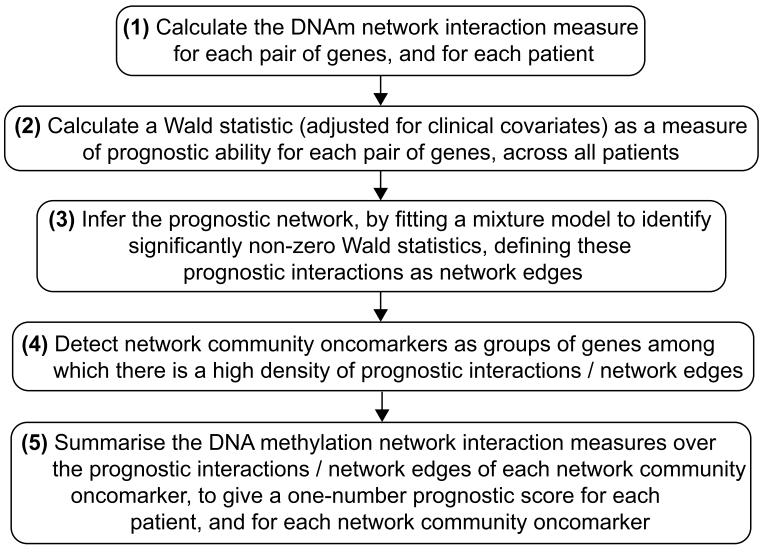


FIG. 1. *Overview of methods.*

alternatively spliced gene products, among other phenomena [Jones (2012)]. The details of this DNA methylation network interaction measure are as follows.

The DNA methylation network interaction measure is defined by analogy to CCA. CCA aims to discover linear combinations of variables of one type and linear combinations of variables of another type so that these combinations best explain each other. In this context, a particular way of combining (by scaling and adding) the deviations from the mean methylation profile at a number of locations within one gene might be particularly effective at explaining a particular combination of (again, by scaling and adding) the deviations from the mean methylation profile at a number of locations in another gene and vice versa. There will probably be fewer ways in which the methylation levels of these genes covary across the samples than there are locations at which methylation is measured along the genes; this is because the methylation level is highly correlated at many locations along a particular gene. CCA finds the most important components of this covariation across samples.

CCA seeks to find the vectors $a$ and $b$, in the $p$ and $q$ dimensional spaces of variables $\mathbf{X} = (x_1, x_2, \ldots, x_p)'$ and $\mathbf{Y} = (y_1, y_2, \ldots, y_q)'$, respectively, which maximise the correlation $\rho = \operatorname{cor}(\mathbf{a}'\mathbf{X}, \mathbf{b}'\mathbf{Y})$ defined according to equation (1):

$$\rho = \frac{\mathbf{a}'\boldsymbol{\Sigma}_{XY}\mathbf{b}}{\sqrt{\mathbf{a}'\boldsymbol{\Sigma}_{XX}\mathbf{a}}\sqrt{\mathbf{b}'\boldsymbol{\Sigma}_{YY}\mathbf{b}}}, \tag{1}$$

where

$$\boldsymbol{\Sigma}_{XX} = \mathbb{E}\big[(\mathbf{X} - \boldsymbol{\mu}_X)(\mathbf{X} - \boldsymbol{\mu}_X)'\big]$$

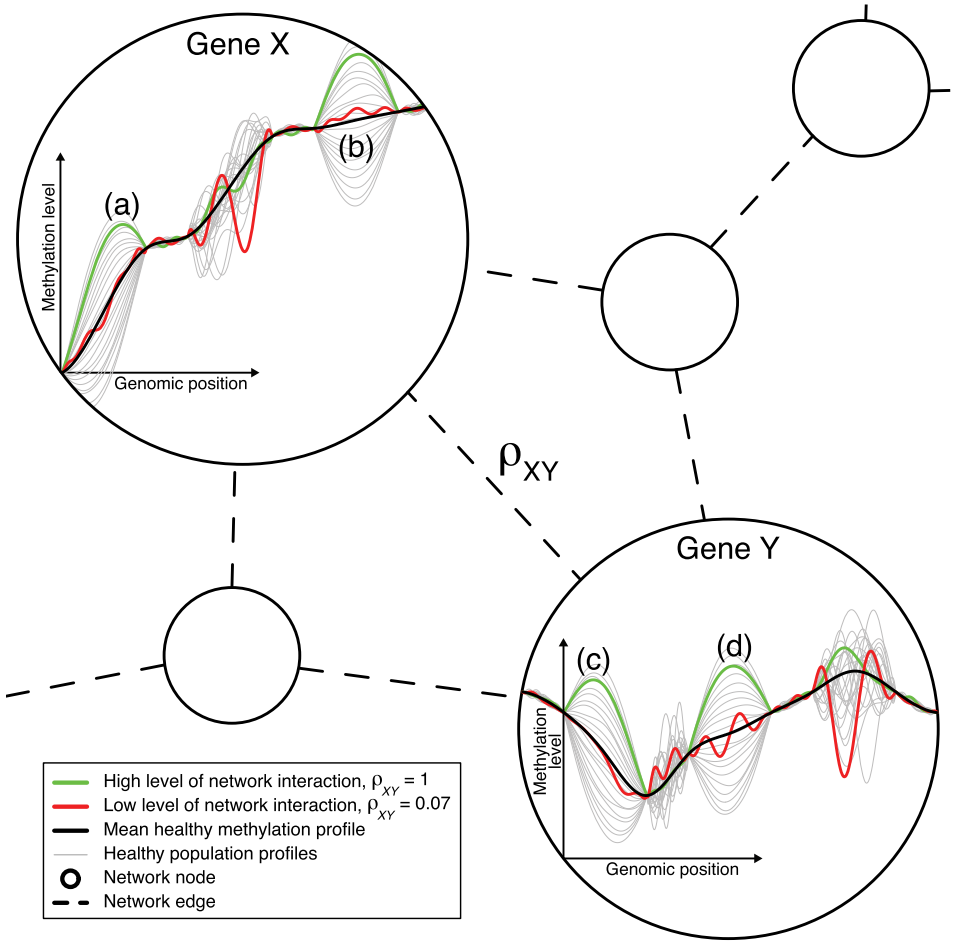


FIG. 2. *The DNA methylation network interaction measure. A combination of the variation of the healthy methylation profiles in regions* (a) *and* (b) *of gene X explains well/is well explained by a combination of the variation of the healthy methylation profiles in regions* (c) *and* (d) *of gene Y. The green cancer sample varies by a large amount about the mean methylation profile and in a typical way in these regions in both genes. Hence, the green sample corresponds to a high level of network interaction for this sample,* $\rho_{XY} = 1$. *The equivalent variations in the other regions of these genes do not explain each other well, and so the red sample, which varies by a large amount in these other regions and varies less and in an atypical way in regions* (a)–(d), *corresponds to a low level of network interaction,* $\rho_{XY} = 0.07$. *Genes X and Y are likely to have different numbers of methylation measurement locations* (*i.e., variables X and Y are of different dimension*). *The ordering of the measurement locations has no influence on the calculation of* $\rho$, *as long as the ordering is consistent across samples. This diagram was presented previously by Bartlett, Olhede and Zaikin* (2014).

and

$$\mathbf{\Sigma}_{YY} = \mathbb{E}\big[(\mathbf{Y} - \boldsymbol{\mu}_Y)(\mathbf{Y} - \boldsymbol{\mu}_Y)'\big]$$

are the covariance matrices of $\mathbf{X}$ and $\mathbf{Y}$, respectively,

$$\mathbf{\Sigma}_{XY} = \mathbb{E}\big[(\mathbf{X} - \boldsymbol{\mu}_X)(\mathbf{Y} - \boldsymbol{\mu}_Y)'\big]$$

is the cross-covariance matrix of $\mathbf{X}$ and $\mathbf{Y}$, and $\boldsymbol{\mu}_X$ and $\boldsymbol{\mu}_Y$ are the mean vectors of $\mathbf{X}$ and $\mathbf{Y}$.

Two genes $X$ and $Y$ have corresponding methylation profiles which are measured for sample/patient $k$ at $p$ and $q$ CpGs (loci), respectively, along these genes. Denoting these measurements by the variables $x_1, \ldots, x_p$ and $y_1, \ldots, y_q$ for genes $X$ and $Y$, respectively, the DNA methylation profiles for these genes, for patient $k$, can be represented by the vectors $\mathbf{x}(k)$ and $\mathbf{y}(k)$, which have $p$ and $q$ entries, respectively. A measure of DNA methylation network interaction $\rho_{XY}(k)$, of the methylation profiles of genes $X$ and $Y$ for sample $k$, can then be defined by analogy with equation (1), according to equation (2):

$$\rho_{XY}(k) = \frac{\mathbf{x}^c(k)^T \hat{\mathbf{\Sigma}}_{XY}^{(h)} \mathbf{y}^c(k)}{\sqrt{\mathbf{x}^c(k)^T \hat{\mathbf{\Sigma}}_{XX}^{(h)} \mathbf{x}^c(k)} \sqrt{\mathbf{y}^c(k)^T \hat{\mathbf{\Sigma}}_{YY}^{(h)} \mathbf{y}^c(k)}}, \tag{2}$$

where $\hat{\mathbf{\Sigma}}_{XX}^{(h)}$, $\hat{\mathbf{\Sigma}}_{YY}^{(h)}$ and $\hat{\mathbf{\Sigma}}_{XY}^{(h)}$ are estimated from healthy rather than cancer samples in the methylation dataset, according to equations (3)–(5),

$$\hat{\mathbf{\Sigma}}_{XX}^{(h)} = \frac{1}{n_h} \sum_{k \in \text{healthy}} \big(\mathbf{x}(k) - \hat{\boldsymbol{\mu}}_X^{(h)}\big)\big(\mathbf{x}(k) - \hat{\boldsymbol{\mu}}_X^{(h)}\big)^T, \tag{3}$$

$$\hat{\mathbf{\Sigma}}_{YY}^{(h)} = \frac{1}{n_h} \sum_{k \in \text{healthy}} \big(\mathbf{y}(k) - \hat{\boldsymbol{\mu}}_Y^{(h)}\big)\big(\mathbf{y}(k) - \hat{\boldsymbol{\mu}}_Y^{(h)}\big)^T, \tag{4}$$

$$\hat{\mathbf{\Sigma}}_{XY}^{(h)} = \frac{1}{n_h} \sum_{k \in \text{healthy}} \big(\mathbf{x}(k) - \hat{\boldsymbol{\mu}}_X^{(h)}\big)\big(\mathbf{y}(k) - \hat{\boldsymbol{\mu}}_Y^{(h)}\big)^T, \tag{5}$$

where

$$\hat{\boldsymbol{\mu}}_X^{(h)} = \frac{1}{n_h} \sum_{k \in \text{healthy}} \mathbf{x}(k),$$

$$\hat{\boldsymbol{\mu}}_Y^{(h)} = \frac{1}{n_h} \sum_{k \in \text{healthy}} \mathbf{y}(k),$$

$n_h$ is the number of healthy samples in the dataset, and $\mathbf{x}^c(k)$ and $\mathbf{y}^c(k)$ are the mean-centered methylation profiles $\mathbf{x}^c(k) = \mathbf{x}(k) - \hat{\boldsymbol{\mu}}_X^{(h)}$ and $\mathbf{y}^c(k) = \mathbf{y}(k) - \hat{\boldsymbol{\mu}}_Y^{(h)}$. The DNAm network interaction measure hence evaluates the extent to which, in an individual tumour sample, the combinations of the methylation-variables (i.e.,

loci) in genes $X$ and $Y$ explain each other, or covary, in the spaces determined by CCA on corresponding healthy samples; that is, the covariation in tumour sample $k$ between the methylation-variables in genes $X$ and $Y$ is assessed against typical healthy variability in these variables. When the DNA methylation network interaction measure $\rho_{XY}(k)$ is large (i.e., close to 1), the corresponding pair of genes explain each others' gene-regulatory behaviour (as reflected in their methylation profiles) well, or have otherwise well-correlated interactive or associative behaviour, for sample/patient $k$. Hence, $\rho_{XY}(k)$ measures (according to their DNA methylation profiles) the level of interaction or association between genes $X$ and $Y$ in tumour sample $k$ compared to typical interactions between these genes in healthy tissue.

2.2. *Prognostic network construction*. Our proposed methodology for inference of network oncomarkers is based on a prognostic interaction network over $m$ genes. This network is represented by the $m \times m$ adjacency matrix $\mathbf{A}$, in which an edge is defined to be present (i.e., $A_{ij} = 1$) if and only if the corresponding pair of genes (nodes) are prognostic according to the DNA methylation network interaction measure of Section 2.1. Otherwise, we set $A_{ij} = 0$. We note that $i$ and $j$ are now redefined compared to the last section so that they index genes rather than DNA methylation locations. This formulation will not be problematic because all subsequent analysis is carried out at the level of genes rather than DNA methylation locations. To identify prognostic edges, we use the Cox proportional hazards model [Cox (1972)] to calculate a Wald-statistic $z_{ij}$ for each of the $\binom{m}{2}$ pairs of genes in the network. The Wald statistic quantifies the strength of association of the DNA methylation network interaction measure $\rho_{ij}$ for the pair of genes $i$ and $j$ ($i = 1, \ldots, m$ and $j = 1, \ldots, m$) with patient survival outcome across patients $k$ ($k = 1, \ldots, n$). We use a multivariate Cox model, adjusting these Wald statistics for clinical covariates, fitting this model separately to each pair of genes $(i, j)$. We adjust in this way in order to detect novel DNA methylation biomarkers which are independent of known prognostic clinical features.

The Wald statistic is asymptotically normally distributed with unit variance [Harrell (2001)], and we can therefore model the distribution of our observed Wald statistics, $z_{ij}$, as a mixture of Gaussians. We have previously demonstrated the utility of mixture modelling to a related network inference problem [Bartlett (2015)], and a similar approach can be applied in this context. We model the $z_{ij}$ as a Gaussian mixture as follows:

$$z_{ij} \sim \begin{cases} \mathcal{N}(\mu_{ij}, \sigma^2), & \text{if } A_{ij} = 1, \\ \mathcal{N}(0, \sigma^2), & \text{if } A_{ij} = 0, \end{cases} \tag{6}$$

where $\mathcal{N}(\mu_{ij}, \sigma^2)$ is the normal distribution, and we enforce $\sigma^2 = 1$ in line with the asymptotic behaviour of the Wald statistic. We fit this mixture model to each observed statistic $z_{ij}$, and then infer whether, given $z_{ij}$, it is more likely that $\mu_{ij} =$

0, or $\mu_{ij} \neq 0$, leading to the estimates $\hat{A}_{ij} = 0$ or $\hat{A}_{ij} = 1$, respectively. We fit this model using the empirical Bayes procedure of Johnstone and Silverman (2004), defining a mixture prior distribution $f_{\text{prior}}(\mu_{ij})$ over the $\mu_{ij}$ of equation (6):

$$f_{\text{prior}}(\mu_{ij}) = (1-w)\delta(\mu_{ij}) + w\gamma(\mu_{ij}), \tag{7}$$

where $w$ is the mixing parameter between the two components, which can also be interpreted as $w = \mathbb{E}[p(A_{ij} = 1)]$, and $\gamma(\cdot|a)$ is the Laplace probability density function,

$$\gamma(\mu_{ij}|a) = \frac{a}{2}\exp\bigl(-a|\mu_{ij}|\bigr),$$

where we use the standard value of $a = 0.5$ [Johnstone and Silverman (2004)]. Taking the mixture components to have Gaussian likelihoods, $f_{\mathcal{N}}(\cdot|\mu_{ij}, \sigma^2)$, as in equation (6), it follows from equation (7) that the posterior density over the observed prognostic Wald statistic $z_{ij}$ is

$$f_{\text{posterior}}(\mu_{ij}|z_{ij}) = \frac{(1-w)\delta(\mu_{ij})f_{\mathcal{N}}(z_{ij}|0,\sigma^2) + w\gamma(\mu_{ij})f_{\mathcal{N}}(z_{ij}|\mu_{ij},\sigma^2)}{f_{\text{marginal}}(z_{ij})}, \tag{8}$$

where the marginal density is

$$f_{\text{marginal}}(z_{ij}) = (1-w)f_{\mathcal{N}}\bigl(z_{ij}|0,\sigma^2\bigr) + wg(z_{ij}), \tag{9}$$

where $g(\mu_{ij})$ is the convolution of the Laplace density with the standard normal density. If the Laplace distribution in the prior [equation (7)] were replaced with a Gaussian, then the marginal distribution [equation (9)] would be a mixture of Gaussians. However, as noted previously [Johnstone and Silverman (2004)], this empirical Bayes procedure requires a prior with tails that are exponential or heavier. Hence, we similarly use the Laplace rather than Gaussian prior, which is a slight model misspecification.

Although a separate model is fitted to each observed Wald statistic $z_{ij}$, a common weight $w_i$ is used for each gene/node $i$. We choose to do this because estimating $w_i$ separately for each gene $i$ allows adaptation to a heterogenous degree distribution in $\mathbf{A}$, as follows. For a particular gene $i$, if the $z_{ij}$ are mostly close to zero, then $\hat{w}_i$ will be set low, which means that fewer edges ($A_{ij} = 1$) will be detected; this hence corresponds to $i$ being a low-degree node. If for a different gene $i$ the $z_{ij}$ are generally further from zero, then $\hat{w}_i$ will be set high, which corresponds to more edges being detected; this hence corresponds to $i$ being a high-degree node.

The estimate $\hat{w}_i$ is found as the value which maximises the marginal likelihood [equation (10)] of the observed statistics $z_{ij}$ over all the pairwise comparisons of $i$ with $j$, $j \neq i$. This allows the model for each such pairwise comparison $(i, j)$ to "borrow strength" from all the other comparisons $(i, j')$, $j' \neq i$, $j' \neq j$:

$$\hat{w}_i = \arg\max_{w} \sum_{j \neq i} \log\bigl\{(1-w)\phi(z_{ij}) + wg(z_{ij})\bigr\}. \tag{10}$$

As in the original presentation of this methodology [Johnstone and Silverman (2004)], we use the posterior median to obtain the estimate $\hat{\mu}_{ij}$. Then we make a conservative estimate of $\mathbf{A}$ as follows:

$$\begin{aligned} &\hat{A}_{ij} = 1 \qquad \text{if } \hat{\mu}_{ij} > 0 \text{ and } \hat{\mu}_{ji} > 0 \text{ or } \hat{\mu}_{ij} < 0 \text{ and } \hat{\mu}_{ji} < 0, \\ &\hat{A}_{ij} = 0 \qquad \text{otherwise.} \end{aligned} \tag{11}$$

2.3. *Community and oncomarker detection*. Network nodes can be grouped together according to their propensity to interact with each other, for example, groups of friends in a social network or functional subnetwork modules in a biological network; this method is referred to as community detection [Girvan and Newman (2002), Newman (2004)]. We use community detection to naturally infer groups of genes in our constructed prognostic network. These groups of genes interact differently in cancer than in healthy tissue, in a way which is predictive of how advanced the disease is. We term these groups "network community oncomarkers". Within a network community oncomarker the genes may interact with each other more (relative to healthy tissue) the more serious the disease is [as in Figure 6(c)], or they may interact with each other less the more serious the disease is [as in Figure 6(a)]. We carry out the task of community detection by fitting the degree-corrected stochastic blockmodel [Bickel and Chen (2009), Holland, Laskey and Leinhardt (1983)]. We fit this model in an efficient way by regularised spectral clustering [Qin and Rohe (2013)], calculating the optimum number of communities to divide the network into by the network histogram method [Olhede and Wolfe (2014)]. Each community identified in this way represents a potential network community oncomarker.

For each network community oncomarker, we then calculate a prognostic score for each patient by summarising the DNA methylation network interaction measure over this group of genes. This prognostic score can be used as a one-number summary of disease prognosis for that patient according to that network community oncomarker. The following points are important when calculating these summaries. Some gene–gene interactions will correspond to an increasingly negative DNA methylation network interaction measure $\rho_{ij}$ for worse patient prognosis. On the other hand, some gene–gene interactions will correspond to an increasingly positive $\rho_{ij}$ for worse prognosis. This means that care must be taken when summarising the network interaction measure across the network community oncomarker. Also, for the same amount of prognostic information conveyed, the magnitude of the changes in the network interaction measure may not be the same for each prognostic pairs of genes. To address these points, we combine the $\rho_{ij}$ across the prognostic pairs of genes of the network community after first multiplying them by the corresponding fitted Cox proportional hazards model coefficients $\hat{\theta}_{ij}$, obtained as described at the start of Section 2.2. Under the Cox proportional hazards model, the fitted model coefficient $\hat{\theta}_{ij}$ for a predictor $ij$ gives the log of the

hazard ratio (HR) for that predictor in the model, that is, $\log(\mathrm{HR}_{ij}) = \hat{\theta}_{ij}$. The hazard ratio is the scale-factor increase in probability of an event (e.g., death) occurring per unit time, relative to the baseline hazard (e.g., compared to a control group). Hence, these coefficients are interpretable in the same way, without scaling issues, across fitted models. This means that, for patient $k$, we can combine the DNA methylation network interaction measures over a network community oncomarker to generate a one-number prognostic score, as follows:

$$\mathrm{Score}_k = \sum_{i \in C, j \in C, i < j} \hat{A}_{ij} \hat{\theta}_{ij} \rho_{ij}(k),$$

where $C$ is the set of nodes in the network community oncomarker, $\hat{\mathbf{A}}$ is the inferred adjacency matrix, $\rho_{ij}(k)$ is the DNA methylation network interaction measure for genes/nodes $i$ and $j$ and patient $k$, and $\hat{\theta}_{ij}$ is the corresponding fitted Cox multivariate proportional-hazards model coefficient. Network edges/DNA methylation network interaction measures $\rho_{ij}$ which increase with poor prognosis (i.e., pairs of genes which interact more as the disease progresses, coloured green in Figure 6) will correspond to $\hat{\theta}_{ij} > 0$. Hence, an increase in such a $\rho_{ij}$ will increase the prognostic score. Equivalently, network edges/DNA methylation network interaction measures $\rho_{ij}$ which decrease with poor prognosis (i.e., pairs of genes which interact less as the disease progresses, coloured red in Figure 6) will correspond to $\hat{\theta}_{ij} < 0$. Hence, a decrease in such a $\rho_{ij}$ will also increase the prognostic score.

2.4. *An equivalent gene-expression interaction measure.* To examine further the hypothesis that the DNA methylation network interaction measure is a reflection of co-regulatory or co-regulated gene-expression patterns (among other genomic effects), we need an equivalent measure of gene–gene interaction or association in terms of gene expression. We can calculate such a measure, $\rho_{XY}^{\mathrm{expr}}(k)$, for gene expression measurements $x^{\mathrm{expr}}(k)$ and $y^{\mathrm{expr}}(k)$ for the genes $X$ and $Y$ and patient $k$, as follows [equation (12)]:

$$\rho_{XY}^{\mathrm{expr}}(k) = \frac{(x^{\mathrm{expr}}(k) - \hat{\mu}_{x^{\mathrm{expr}}}^{(h)})}{\hat{\sigma}_{x^{\mathrm{expr}}}^{(h)}} \cdot \frac{(y^{\mathrm{expr}}(k) - \hat{\mu}_{y^{\mathrm{expr}}}^{(h)})}{\hat{\sigma}_{y^{\mathrm{expr}}}^{(h)}}, \tag{12}$$

where

$$\hat{\mu}_{x^{\mathrm{expr}}}^{(h)} = \frac{1}{n_h} \sum_{k \in \mathrm{healthy}} x^{\mathrm{expr}}(k) \quad \text{and} \quad \hat{\mu}_{y^{\mathrm{expr}}}^{(h)} = \frac{1}{n_h} \sum_{k \in \mathrm{healthy}} y^{\mathrm{expr}}(k),$$

$$\bigl(\hat{\sigma}_{x^{\mathrm{expr}}}^{(h)}\bigr)^2 = \frac{1}{n_h} \sum_{k \in \mathrm{healthy}} \bigl(x^{\mathrm{expr}}(k) - \hat{\mu}_{x^{\mathrm{expr}}}^{(h)}\bigr)^2$$

and

$$\bigl(\hat{\sigma}_{y^{\mathrm{expr}}}^{(h)}\bigr)^2 = \frac{1}{n_h} \sum_{k \in \mathrm{healthy}} \bigl(y^{\mathrm{expr}}(k) - \hat{\mu}_{y^{\mathrm{expr}}}^{(h)}\bigr)^2.$$

The intuition of equation (12) is that when the gene expression measurements $x^{\text{expr}}(k)$ and $y^{\text{expr}}(k)$ deviate *in the same sample* from the corresponding healthy mean expression levels, this measure will be nonzero. When this occurs in the same samples as the DNA methylation network interaction measure $\rho_{XY}(k)$ is also nonzero, we will see a correlation between $\rho_{XY}(k)$ and $\rho_{XY}^{\text{expr}}$. These interaction measures for methylation and expression, $\rho_{XY}(k)$ and $\rho_{XY}^{\text{expr}}$, are equivalent because they both measure deviation from typical interactive behaviour in healthy/control samples.

**3. Examples.** We present an example application of the methodology proposed in Section 2 to a large publicly available breast cancer invasive carcinoma (BRCA) dataset downloaded from the Cancer Genome Atlas (TCGA). We downloaded an initial batch of DNA methylation data for tumour samples taken from 175 individuals (the training set), together with clinical data for these samples relating to patient survival outcome, and the covariates age, disease stage and residual disease. These training data were used to detect potential network community oncomarkers. We then downloaded DNA methylation data for a further 528 tumour samples (the test set), together with data for the same clinical features: these independent samples were used to validate the potential network community oncomarkers. We also downloaded corresponding DNA methylation data for healthy breast tissue samples from 98 individuals to form a reference population of DNA methylation profiles for this analysis, and we downloaded gene expression data for 216 of the tumours for which DNA methylation data were also available. To proceed, we estimated from the training set the healthy population means, covariances and cross-covariances required to calculate the $\rho_{ij}$ ($i = 1, \ldots, m$ and $j = 1, \ldots, m$), as well as the corresponding log hazard ratios $\hat{\theta}_{ij}$ and adjacency matrix $\hat{\mathbf{A}}$. Additionally, from the training data we estimated the communities in the adjacency matrix (including the number of communities) and the prognostic score thresholds used to assign patients to better and worse prognostic groups. We then used these estimates to verify the prognostic ability of the methodology in the test set.

We first inferred the binary prognostic adjacency matrix $\hat{\mathbf{A}}$ for the 175 samples of the BRCA training dataset according to the methods set out in Sections 2.1–2.2. DNA methylation data were available for 14,829 genes, and hence the number of nodes/genes $m$ in the inferred adjacency matrix $\hat{\mathbf{A}}$ is $m = 14{,}829$. The presence of an edge in $\hat{\mathbf{A}}$, that is, $\hat{A}_{ij} = 1$, indicates that the interaction between genes $i$ and $j$ is associated with disease progression. The edge density of $\hat{\mathbf{A}}$ is 0.0035, that is, $p(\hat{A}_{ij} = 1) = 0.0035$. We then extracted the connected component from this inferred network and carried out community detection on this connected component as described in Section 2.3. This resulted in 33 communities ranging from 116 to 285 nodes in size. The reduced adjacency matrix relating to these communities [with $m = 5668$ and $p(\hat{A}_{ij} = 1) = 0.023$] is shown in Figure 3. We note that the

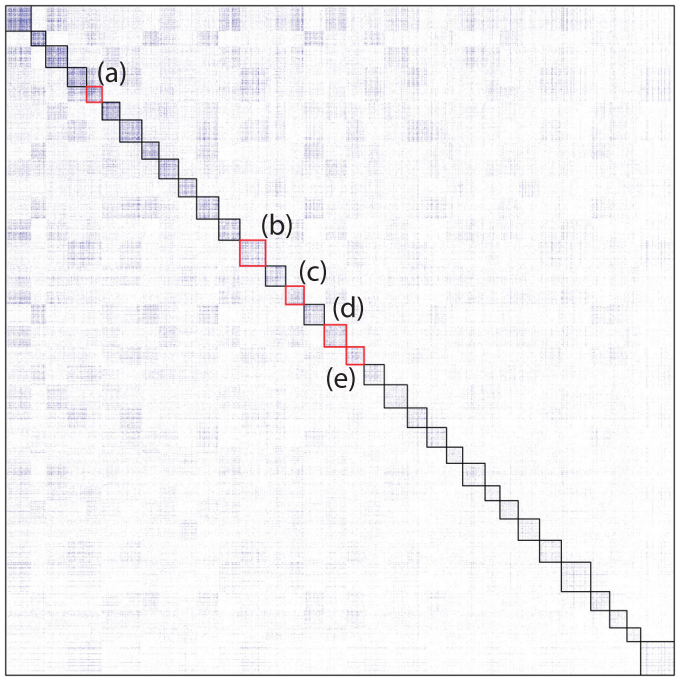


FIG. 3. *The inferred prognostic adjacency matrix after community detection. Entries in the adjacency matrix equal to* 1 (*representing a network edge*) *are coloured blue. Detected communities are outlined in black. The potential network community oncomarkers which are analysed further in Figures* 4–7, *Tables* 1–2 *and Tables* S1–S5 *in the supplement are outlined in red and labelled* (a)–(e).

stochastic blockmodel, fitted in this way via spectral clustering, does not provide any uncertainty as to the inferred community assignments: if this is desired, then mixed-membership stochastic blockmodels are available as an alternative [Airoldi et al. (2008)]. In the analysis we present here, uncertainties arising from these inferred community assignments are considered in the subsequent analyses (Figures 4 and 5, and Tables 1 and 2).

We validated each of the 33 potential network community oncomarkers in the 528 independent tumour samples of the test/validation set. We note that these 528 samples were not used in any way to identify the 33 potential network community oncomarkers shown in Figure 3. Hence, in this validation, each of these 528 patients were classified individually according to prognosis without reference to the other validation samples. This means that comparing these prognostic classifications assigned to the validation samples is a true test of prognostic ability of the network community oncomarkers. To carry out the validation, we calculated the prognostic score for the 528 independent/unseen samples of the test set based on the inferred prognostic adjacency matrix $\hat{\mathbf{A}}$ and the fitted Cox multivariate proportional hazards model coefficients $\hat{\boldsymbol{\theta}}$ obtained from the initial 175 samples of

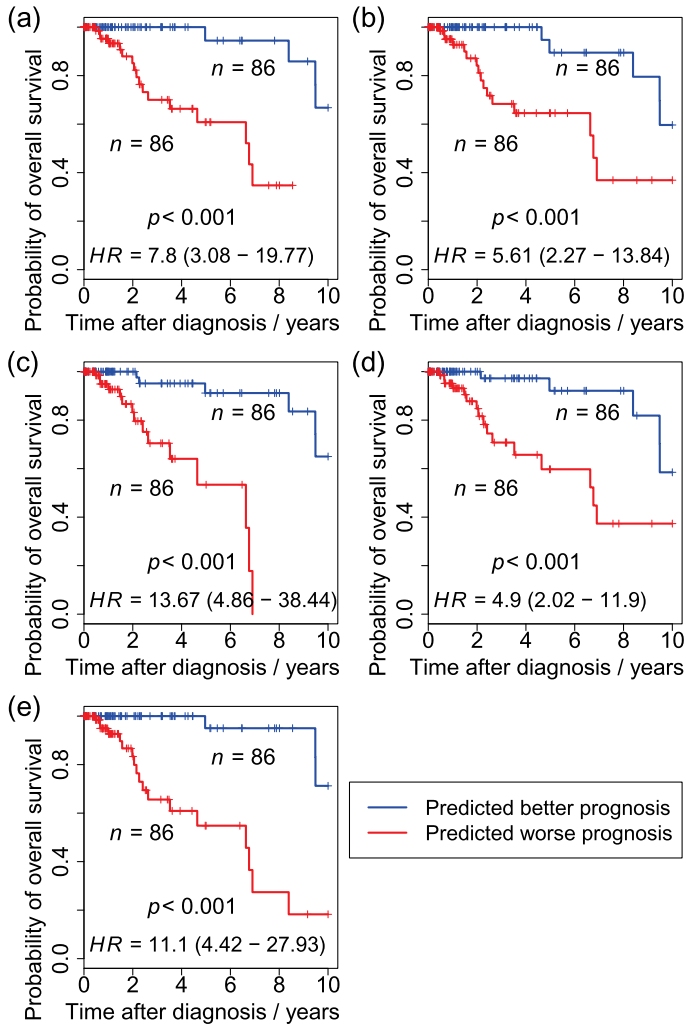


FIG. 4. *Network community oncomarkers*: *Kaplan–Meier plots for the training set*. *Comparison of survival curves for the patient groups defined by the prognostic score for each network community oncomarker. The groups are divided by the median prognostic score in the* 175 *samples of the training dataset*. *The hazard ratio* (HR) *is displayed with* 95% *C.I. in brackets*, *with the corresponding p-value calculated by univariate Cox regression*. (a)–(e) *indicate network community oncomarkers* 1–5, *as shown in Figure* 3.

the training set. Using this trained model, we calculated one prognostic score for each potential network community oncomarker for each of the 528 unseen test-set samples. We then tested the prognostic score, for each potential network community oncomarker, for significant prediction of patient survival outcome in these 528 unseen test-set samples. The five potential network community oncomarkers

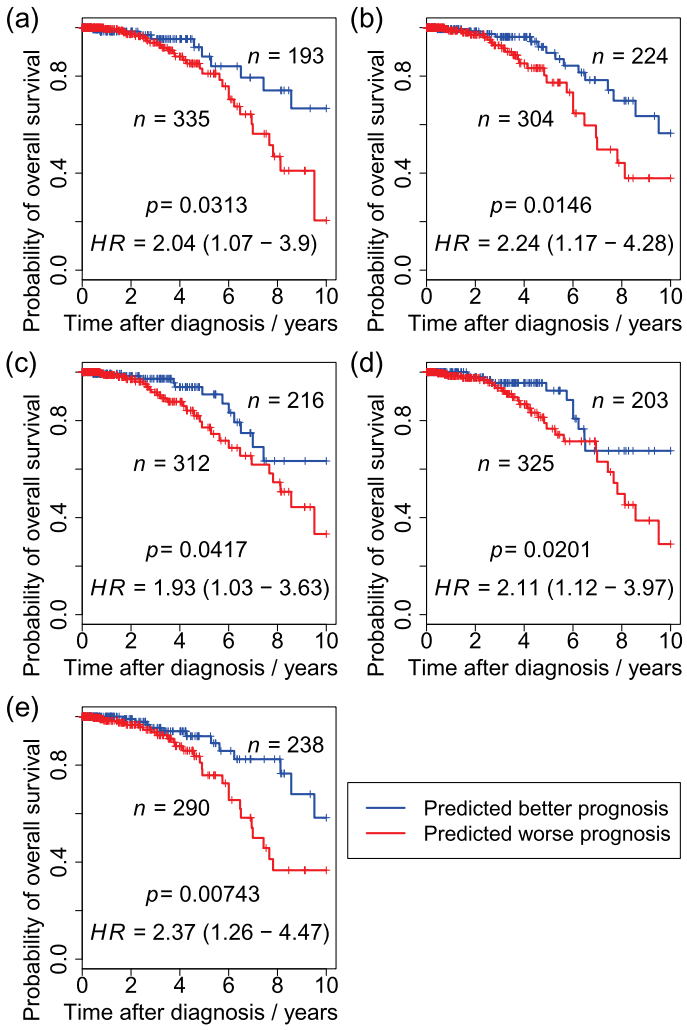


FIG. 5. *Network community oncomarkers*: *Kaplan–Meier plots for the test set. Comparison of survival curves for the patient groups defined by the prognostic score for each network community oncomarker. The groups are divided by the median prognostic score in the* 175 *samples of the training dataset. The hazard ratio* (HR) *is displayed with* 95% *C.I. in brackets, with the corresponding p-value calculated by univariate Cox regression.* (a)–(e) *indicate network community oncomarkers* 1–4, *as shown in Figure* 3.

which validated in this way with the highest level of significance are outlined in red in Figure 3. The results of univariate and multivariate Cox regression for these five best network community oncomarkers are shown in Figures 4 and 5, and in Tables 1 and 2, for the training and test sets, respectively. Plots equivalent to Figures 4 and 5 for all 33 detected network communities appear in Supplementary

TABLE 1

*Network community oncomarkers—training set prognosis. Multivariate Cox regression was used to test significance of the prognostic scores obtained from the network community oncomarkers.* (a)–(e) *indicate network community oncomarkers* 1–5, *as shown in Figure* 3

| | **HR (95%CI)** | ***p*** | ***n*** |
|---|---|---|---|
| | (a) Network community oncomarker 1 | | |
| Prognostic score | 77.1 (10.5–567) | $<0.001$ | 172 |
| Age | 1.79 (0.66–4.84) | 0.249 | 172 |
| Residual disease | 15.4 (4.68–50.9) | $<0.001$ | 172 |
| Stage | 2.85 (0.96–8.46) | 0.060 | 172 |
| | (b) Network community oncomarker 2 | | |
| Prognostic score | 51.3 (8.35–315) | $<0.001$ | 172 |
| Age | 1.42 (0.48–4.23) | 0.53 | 172 |
| Residual disease | 30.4 (5.82–158) | $<0.001$ | 172 |
| Stage | 1.95 (0.68–5.54) | 0.212 | 172 |
| | (c) Network community oncomarker 3 | | |
| Prognostic score | 50.1 (9.77–256) | $<0.001$ | 172 |
| Age | 2.16 (0.81–5.8) | 0.125 | 172 |
| Residual disease | 13.3 (4.54–39.1) | $<0.001$ | 172 |
| Stage | 2.41 (0.81–7.18) | 0.114 | 172 |
| | (d) Network community oncomarker 4 | | |
| Prognostic score | 22.7 (5.52–93.1) | $<0.001$ | 172 |
| Age | 3.49 (1.3–9.42) | 0.0135 | 172 |
| Residual disease | 16.3 (5.24–50.7) | $<0.001$ | 172 |
| Stage | 1.05 (0.38–2.91) | 0.928 | 172 |
| | (e) Network community oncomarker 5 | | |
| Prognostic score | 46.0 (8.17–259) | $<0.001$ | 172 |
| Age | 2.91 (1–8.44) | 0.0493 | 172 |
| Residual disease | 7.04 (2.68–18.5) | $<0.001$ | 172 |
| Stage | 3.74 (1.23–11.4) | 0.02 | 172 |

Figures S1–S2 [Bartlett and Zaikin (2016)]. For the multivariate analysis, samples with missing data for any of the clinical covariates were removed, leaving 172 and 396 samples for the training and test sets, respectively. We note that, as would be expected, the level of significance in the training set (to which the model was fitted, Figure 4 and Table 1) is much higher than in the test set (Figure 5 and Table 2).

Figure 6 shows the five network community oncomarkers which validated most significantly. Green edges indicate gene–gene interactions which become stronger with disease progression. Red edges indicate interactions which become weaker with disease progression. Hence, the network community oncomarkers of Figure 6(a) and (b) can be considered to be functional subnetwork modules which become less active as the cancer progresses (comprised of 99% and 96% red edges, respectively). On the other hand, Figure 6(c) and (d) can be considered to be func-

TABLE 2

*Network community oncomarkers—test/validation set prognosis. Multivariate Cox regression was used to test significance of the prognostic scores obtained from the network community oncomarkers.* (a)–(e) *indicate network community oncomarkers* 1–5, *as shown in Figure* 3

| | **HR (95%CI)** | ***p*** | ***n*** |
|---|---|---|---|
| | (a) Network community oncomarker 1 | | |
| Prognostic score | 4.89 (1.65–14.5) | 0.00429 | 396 |
| Age | 3.52 (1.46–8.49) | 0.00513 | 396 |
| Residual disease | 12.5 (5.32–29.3) | <!0.001 | 396 |
| Stage | 1.62 (0.66–4) | 0.294 | 396 |
| | (b) Network community oncomarker 2 | | |
| Prognostic score | 5.07 (1.81–14.1) | 0.00195 | 396 |
| Age | 3.67 (1.49–9.03) | 0.00458 | 396 |
| Residual disease | 8.72 (3.78–20.1) | <0.001 | 396 |
| Stage | 1.47 (0.6–3.61) | 0.406 | 396 |
| | (c) Network community oncomarker 3 | | |
| Prognostic score | 2.63 (1.01–6.89) | 0.0484 | 396 |
| Age | 2.07 (0.86–5) | 0.106 | 396 |
| Residual disease | 11.3 (4.97–25.5) | <0.001 | 396 |
| Stage | 2.04 (0.76–5.45) | 0.157 | 396 |
| | (d) Network community oncomarker 4 | | |
| Prognostic score | 4.92 (1.8–13.5) | 0.00189 | 396 |
| Age | 1.91 (0.78–4.69) | 0.159 | 396 |
| Residual disease | 17.2 (6.76–43.9) | <0.001 | 396 |
| Stage | 0.92 (0.34–2.48) | 0.871 | 396 |
| | (e) Network community oncomarker 5 | | |
| Prognostic score | 2.5 (0.94–6.65) | 0.0668 | 396 |
| Age | 2.23 (0.94–5.27) | 0.0677 | 396 |
| Residual disease | 8.17 (3.47–19.3) | <0.001 | 396 |
| Stage | 1.59 (0.64–3.95) | 0.321 | 396 |

tional subnetwork modules which become more active as the cancer progresses (both comprised of 99% green edges). Then the network community oncomarker of Figure 6(e) contains a mixture of these effects (comprised of 87% red and 13% green edges). However, each of these network community oncomarkers represents a functional subnetwork module which is rewired in a way which is advantageous for the cancer, in favour of proliferation, and against cell death and immune function. The genes/nodes of these network community oncomarkers are shown in Tables S1–S5 in the supplement [Bartlett and Zaikin (2016)]; they list many genes related to cell proliferation (e.g., *CDKL1*, *NKAPL*, *MAPK6*), developmental processes (e.g., *HOXD10*, *HOXB9*, *HOXC10*, *HOXA13*, *HOXC12*, *HOXD13*) and immune function (e.g., *VSIG2*, *IL36B*, *RBPJ*).

We hypothesise that the DNA methylation network interaction measure is a reflection of co-regulatory or co-regulated gene-expression patterns, among other

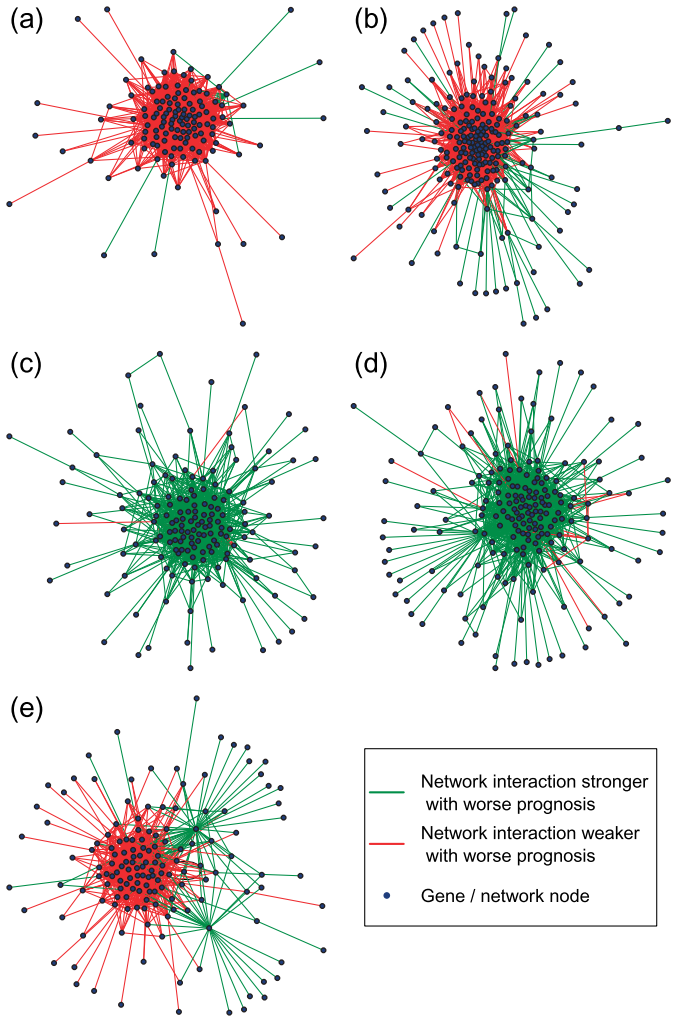


FIG. 6. *Detected network community oncomarkers*. (a)–(e) *indicate network community oncomarkers* 1–5, *as shown in Figure* 3.

genomic effects. We tested this hypothesis by comparing the DNA methylation network interaction measure $\rho_{XY}$ for a pair of genes $XY$ [equation (2)] with an equivalent measure of interactive behaviour of these genes in terms of their expression levels, $\rho_{XY}^{\text{expr}}$ [equation (12)]. Correlation test $p$-values for the comparison between $\rho_{XY}$ and $\rho_{XY}^{\text{expr}}$ appear in Figure 7. It is clear that, in these histograms, there is a concentration of significant $p$-values close to zero, indicating a departure from the null hypothesis uniform distribution, and demonstrating an association between $\rho_{XY}$ and $\rho_{XY}^{\text{expr}}$ for many of the edges/interactions of each network community oncomarker. However, there are also many nonsignificant $p$-values

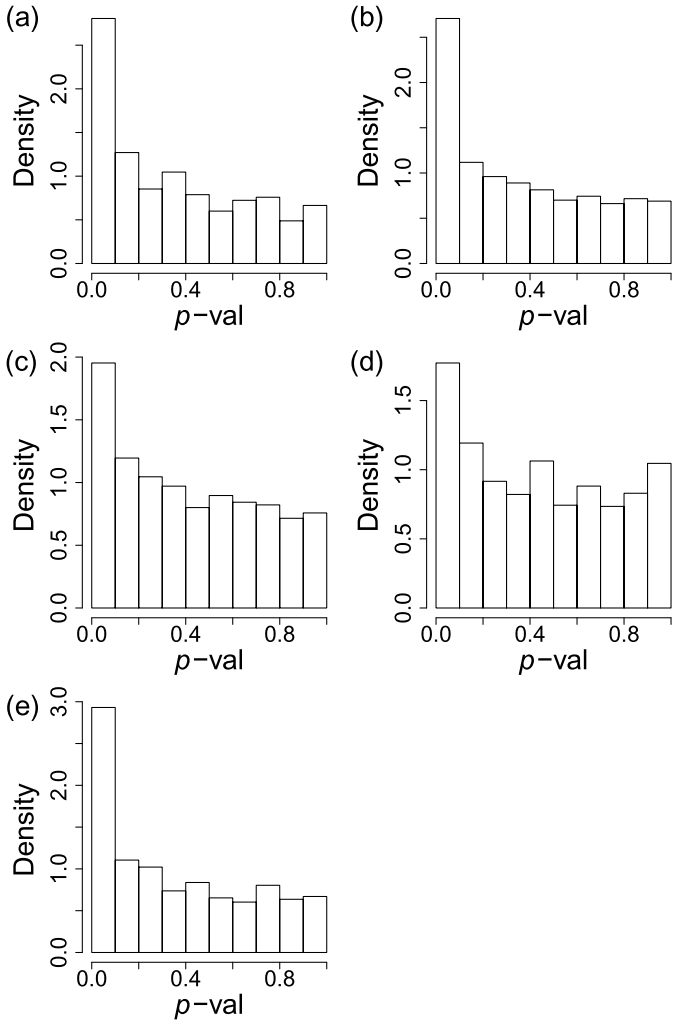


FIG. 7. *Correlation of DNA methylation with gene expression for the network community oncomarkers*. (a)–(e) *indicate network community oncomarkers* 1–5, *as shown in Figure* 3.

visible in these histograms, indicating that there are other genomic interactive effects present which cannot be explained in terms of gene expression (as assessed by mRNA levels) alone. Such effects are expected to include the influence of alternatively spliced products or isoforms [Jones (2012)] and the interaction between noncoding transcripts and the epigenome [Lai and Shiekhattar (2014)].

**4. Discussion.** In this paper, we have proposed methodology to detect cancer biomarkers based on the epigenomic pattern DNA methylation. This methodology builds on a previously proposed measure of pairwise interaction between genes

based on the epigenomic gene-regulatory pattern DNA methylation [Bartlett, Olhede and Zaikin (2014)]. Based on this DNA methylation network interaction measure, the methodology we describe in this paper allows inference of prognostic genomic networks and identification of prognostic biomarkers from such networks using community detection methodology. Community detection has previously proved powerful as well as realistic in a range of fields, including social as well as biological networks [Girvan and Newman (2002)]. In the context of genomic networks, such modular groups of genes are known to correspond to specific physiological functions [Shen-Orr et al. (2002)]. The modular prognostic biomarkers which we detect are termed "network community oncomarkers"; they are groups of nodes/genes among which there is a high density of prognostic genomic interactive or associative behaviour. We have demonstrated that within these communities, the DNA methylation network interaction measure is highly associated with co-regulatory behaviour linked to gene expression (at the mRNA level), giving functional relevance to the findings. However, there are also likely to be a range of genomic interactive effects present which are measured by the DNA methylation network interaction measure but which are not reflected in mRNA levels. Our proposed methodology also allows a one-number prognostic score for a network community oncomarker to be calculated for each patient/sample: this prognostic score is a measure of disease progression in that patient.

Our proposed methodology uses mixture modelling to infer network structure from prognostic association between genes, and draws on practical approaches to community detection to obtain oncomarkers from this prognostic network. Mixture modelling has previously been shown to be an effective approach to the related problem of clustering in networks [Vu, Hunter and Schweinberger (2013)]. This suggests that more general methodology could be developed here, in which network and community inference are both carried out simultaneously by model fitting. Network inference has also been carried out previously using multiple node attributes in cell biological data [Katenka and Kolaczyk (2012)], and those findings could be used as a basis upon which data from other genomic sources could be integrated into the methodology proposed here. Genes also frequently carry out multiple roles in different biological contexts, and hence may be involved in more than one functional subnetwork module within a genomic network. Work has been carried out on overlapping stochastic blockmodels [Latouche, Birmelé and Ambroise (2011)], and hence this would be a natural context in which to develop an application for such methodology.

The field of epigenomics is progressing fast and promises many new insights in the near future into unexplained or undiscovered genomic phenomena, for example, relating to the so-called "dark matter" of the genome [Venters and Pugh (2013)]. Epigenomics is also expected to provide new understanding of the mechanisms of disease progression. The discovery that some genomic loci gain or lose methylation in ways which may be unique to cancer suggests that understanding

changes in DNA methylation machinery may be essential to understanding oncogenesis [Xie et al. (2013)]. The field of network science is also advancing rapidly. Networks are an efficient way to represent and analyse large numbers of variables, which is particularly relevant in modern, large-scale genomic studies. Networks of interactions are a natural way to represent and analyse genomic interactions, associations and processes. Therefore, the study of genomic and epigenomic networks promises to be productive over the coming years for the fields of biology, medicine and statistics.

**5. Datasets.** DNA methylation (DNAm) data from breast cancer invasive carcinoma (BRCA) tumour samples, collected via the Illumina Infinium HumanMethylation450 platform, were downloaded from The Cancer Genome Atlas (TCGA) project [Bonetta (2006), Collins and Barker (2007), Hampton (2006)] at level 3. These data were preprocessed by first removing probes with nonunique mappings and which map to SNPs (as identified in the TCGA level 3 data); probes mapping to sex chromosomes were also removed; in total, 98,384 probes were removed in this way from all datasets. After removal of these probes, 270,985 probes with known gene annotations remained. Probes were then removed if they had less than 95% coverage across samples; probe values were also replaced if they had corresponding detection $p$-value greater than 5% by KNN ($k$ nearest neighbour) imputation ($k = 5$). The loci of analysed CpGs were mapped to genes based on annotation information for the Illumina Infinium platform obtained from the *R*/*Bioconductor* package "IlluminaHumanMethylation450k". The data were also checked for batch effects by hierarchical clustering and correlation of the significant principle components with phenotype and batch: no significant batch effects (which would warrant further correction) were found. We downloaded DNA methylation data for tumour samples from 175 samples/individuals, from TCGA in July 2013, with clinical data available for patient survival outcome, and the clinical covariates age, disease stage and residual disease. At the same time, we also downloaded corresponding DNA methylation data for healthy tissue for 98 individuals. These data were used to detect potential network community oncomarkers. We then downloaded DNA methylation data for a further 528 tumour samples from TCGA in September 2014, with data for the same clinical features available. These independent samples were used to validate the potential network community oncomarkers. At this time we also downloaded gene expression data from TCGA at level 3, for 216 of the tumours for which we also obtained DNA methylation data.

**Acknowledgment.** We are very grateful to Professor S. C. Olhede for literature suggestions, fruitful discussions and helpful comments.

## SUPPLEMENTARY MATERIAL

**Supplementary tables and figures** (DOI: 10.1214/16-AOAS939SUPP; .pdf). Supplementary Tables S1–S5 and Supplementary Figures S1–S2.

## REFERENCES


AIROLDI, E. M., BLEI, D. M., FIENBERG, S. E. and XING, E. P. (2008). Mixed membership stochastic blockmodels. *J. Mach. Learn. Res.* **9** 1981–2014.

BARABÁSI, A.-L. and OLTVAI, Z. N. (2004). Network biology: Understanding the cell's functional organization. *Nat. Rev. Genet.* **5** 101–113.

BARTLETT, T. E. (2015). Network inference and community detection, based on covariance matrices, correlations and test statistics from arbitrary distributions. Preprint. Available at arXiv:1506.04928.

BARTLETT, T. E., OLHEDE, S. C. and ZAIKIN, A. (2014). A DNA methylation network interaction measure, and detection of network oncomarkers. *PLoS ONE* **9** e84573.

BARTLETT, T. E. and ZAIKIN, A. (2016). Supplement to "Detection of epigenomic network community oncomarkers." DOI:10.1214/16-AOAS939SUPP.

BARTLETT, T. E., ZAIKIN, A., OLHEDE, S. C., WEST, J., TESCHENDORFF, A. E. and WIDSCHWENDTER, M. (2013). Corruption of the intra-gene DNA methylation architecture is a hallmark of cancer. *PLoS ONE* **8** e68285.

BEGUERISSE-DÍAZ, M., GARDUÑO-HERNÁNDEZ, G., VANGELOV, B., YALIRAKI, S. N. and BARAHONA, M. (2014). Interest communities and flow roles in directed networks: The Twitter network of the UK riots. *J. R. Soc. Interface* **11** 20140940.

BHAGAT, R., CHADAGA, S., PREMALATA, C. S., RAMESH, G., RAMESH, C., PALLAVI, V. R. and KRISHNAMOORTHY, L. (2012). Aberrant promoter methylation of the RASSF1A and APC genes in epithelial ovarian carcinoma development. *Cellular Oncology* **35** 473–479.

BICKEL, P. J. and CHEN, A. (2009). A nonparametric view of network models and Newman–Girvan and other modularities. *Proc. Natl. Acad. Sci. USA* **106** 21068–21073.

BONETTA, L. (2006). Genome sequencing in the fast Lane. *Nature Methods* **3** 141.

BROCKS, D., ASSENOV, Y., MINNER, S., BOGATYROVA, O., SIMON, R., KOOP, C., OAKES, C., ZUCKNICK, M., LIPKA, D. B., WEISCHENFELDT, J. et al. (2014). Intratumor DNA methylation heterogeneity reflects clonal evolution in aggressive prostate cancer. *Cell Reports* **8** 798–806.

CHRISTENSEN, B. C., HOUSEMAN, E. A., MARSIT, C. J., ZHENG, S., WRENSCH, M. R., WIEMELS, J. L., NELSON, H. H., KARAGAS, M. R., PADBURY, J. F., BUENO, R. et al. (2009). Aging and environmental exposures alter tissue-specific DNA methylation dependent upon CpG island context. *PLoS Genetics* **5** e1000602.

CLUNE, J., MOURET, J.-B. and LIPSON, H. (2013). The evolutionary origins of modularity. *Proc. R. Soc. Lond., B Biol. Sci.* **280** 20122863.

COLLINS, F. and BARKER, A. (2007). Mapping the cancer genome. *Scientific American Magazine* **296** 50–57.

COONEY, C. A. (2007). Epigenetics–DNA-based mirror of our environment? *Dis. Markers* **23** 121–137.

COX, D. R. (1972). Regression models and life-tables. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **34** 187–220. MR0341758

FEINBERG, A. P., OHLSSON, R. and HENIKOFF, S. (2006). The epigenetic progenitor origin of human cancer. *Nat. Rev. Genet.* **7** 21–33.

FLEISCHER, T., FRIGESSI, A., JOHNSON, K. C., EDVARDSEN, H., TOULEIMAT, N., KLAJIC, J., RIIS, M. L., HAAKENSEN, V., WÄRNBERG, F., NAUME, B. et al. (2014). Genome-wide DNA methylation profiles in progression to in situ and invasive carcinoma of the breast with impact on gene transcription and prognosis. *Genome Biol* **15** 435.

GAO, F., SHI, L., RUSSIN, J., ZENG, L., CHANG, X., HE, S., CHEN, T. C., GIANNOTTA, S. L., WEISENBERGER, D. J., ZADA, G. et al. (2013). DNA methylation in the malignant transformation of meningiomas. *PloS One* **8** e54114.

GIRVAN, M. and NEWMAN, M. E. J. (2002). Community structure in social and biological networks. *Proc. Natl. Acad. Sci. USA* **99** 7821–7826 (electronic). MR1908073

HAMPTON, T. (2006). Cancer genome atlas. *JAMA*: *The Journal of the American Medical Association* **296** 1958–1958.

HARRELL, F. E. (2001). *Regression Modeling Strategies*: *With Applications to Linear Models*, *Logistic Regression*, *and Survival Analysis*. Springer, Berlin.

HOLLAND, P. W., LASKEY, K. B. and LEINHARDT, S. (1983). Stochastic blockmodels: First steps. *Social Networks* **5** 109–137. MR0718088

HOTELLING, H. (1936). Relations between two sets of variates. *Biometrika* **28** 321–377.

JACOB, L., NEUVIAL, P. and DUDOIT, S. (2012). More power via graph-structured tests for differential expression of gene networks. *Ann. Appl. Stat.* **6** 561–600. MR2976483

JOHNSTONE, I. M. and SILVERMAN, B. W. (2004). Needles and straw in haystacks: Empirical Bayes estimates of possibly sparse sequences. *Ann. Statist.* **32** 1594–1649. MR2089135

JONES, P. A. (2012). Functions of DNA methylation: Islands, start sites, gene bodies and beyond. *Nat. Rev. Genet.* **13** 484–492.

KANG, G. H., SHIM, Y.-H., JUNG, H.-Y., KIM, W. H., RO, J. Y. and RHYU, M.-G. (2001). CpG island methylation in premalignant stages of gastric carcinoma. *Cancer Research* **61** 2847–2851.

KANG, G. H., LEE, S., KIM, J.-S. and JUNG, H.-Y. (2003). Profile of aberrant CpG island methylation along multistep gastric carcinogenesis. *Laboratory Investigation* **83** 519–526.

KATENKA, N. and KOLACZYK, E. D. (2012). Inference and characterization of multi-attribute networks with application to computational biology. *Ann. Appl. Stat.* **6** 1068–1094. MR3012521

KISHIDA, Y., NATSUME, A., KONDO, Y., TAKEUCHI, I., AN, B., OKAMOTO, Y., SHINJO, K., SAITO, K., ANDO, H., OHKA, F. et al. (2012). Epigenetic subclassification of meningiomas based on genome-wide DNA methylation analyses. *Carcinogenesis* **33** 436–441.

LAI, F. and SHIEKHATTAR, R. (2014). Where long noncoding RNAs meet DNA methylation. *Cell Res.* **24** 263–264.

LATOUCHE, P., BIRMELÉ, E. and AMBROISE, C. (2011). Overlapping stochastic block models with application to the French political blogosphere. *Ann. Appl. Stat.* **5** 309–336. MR2810399

LI, C. and LI, H. (2010). Variable selection and regression analysis for graph-structured covariates with an application to genomics. *Ann. Appl. Stat.* **4** 1498–1516. MR2758338

LI, C. and WANG, J. (2014). Quantifying the underlying landscape and paths of cancer. *J. R. Soc. Interface* **11** 20140774.

LUO, Y., WONG, C.-J., KAZ, A. M., DZIECIATKOWSKI, S., CARTER, K. T., MORRIS, S. M., WANG, J., WILLIS, J. E., MAKAR, K. W., ULRICH, C. M. et al. (2014). Differences in DNA methylation signatures reveal multiple pathways of progression from adenoma to colorectal cancer. *Gastroenterology* **147** 418–429.

MAEKAWA, R., SATO, S., YAMAGATA, Y., ASADA, H., TAMURA, I., LEE, L., OKADA, M., TAMURA, H., TAKAKI, E., NAKAI, A. et al. (2013). Genome-wide DNA methylation analysis reveals a potential mechanism for the pathogenesis and development of uterine leiomyomas. *PloS One* **8** e66632.

MARDIA, K. V. (2013). Statistical approaches to three key challenges in protein structural bioinformatics. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **62** 487–514. MR3060628

NANDI, A. K., SUMANA, A. and BHATTACHARYA, K. (2014). Social insect colony as a biological regulatory system: Modelling information flow in dominance networks. *J. R. Soc. Interface* **11** 20140951.

NAVARRO, A., YIN, P., MONSIVAIS, D., LIN, S. M., DU, P., WEI, J.-J. and BULUN, S. E. (2012). Genome-wide DNA methylation indicates silencing of tumor suppressor genes in uterine leiomyoma. *PLoS ONE* **7** e33284.

NEWMAN, M. E. (2004). Detecting community structure in networks. *The European Physical Journal B-Condensed Matter and Complex Systems* **38** 321–330.

NEWMAN, M. E. and GIRVAN, M. (2004). Finding and evaluating community structure in networks. *Phys. Rev. E* (3) **69** 026113.

OLHEDE, S. C. and WOLFE, P. J. (2014). Network histograms and universality of blockmodel approximation. *Proc. Natl. Acad. Sci. USA* **111** 14722–14727.

PALLA, G., LOVÁSZ, L. and VICSEK, T. (2010). Multifractal network generator. *Proc. Natl. Acad. Sci. USA* **107** 7640–7645.

PENG, J., ZHU, J., BERGAMASCHI, A., HAN, W., NOH, D.-Y., POLLACK, J. R. and WANG, P. (2010). Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer. *Ann. Appl. Stat.* **4** 53–77. MR2758084

QIN, T. and ROHE, K. (2013). Regularized spectral clustering under the degree-corrected stochastic blockmodel. In *Advances in Neural Information Processing Systems* 3120–3128. Lake Tahoe, Nevada.

REZNIK, E., WATSON, A. and CHAUDHARY, O. (2013). The stubborn roots of metabolic cycles. *J. R. Soc. Interface* **10** 20130087.

RIOLO, M. A. and NEWMAN, M. E. J. (2012). First-principles multiway spectral partitioning of graphs. Preprint. Available at arXiv:1209.5969.

SAAVEDRA, S., ROHR, R. P., GILARRANZ, L. J. and BASCOMPTE, J. (2014). How structurally stable are global socioeconomic systems? *J. R. Soc. Interface* **11** 20140693.

SHEN-ORR, S. S., MILO, R., MANGAN, S. and ALON, U. (2002). Network motifs in the transcriptional regulation network of Escherichia coli. *Nat. Genet.* **31** 64–68.

TAYLOR, I. W., LINDING, R., WARDE-FARLEY, D., LIU, Y., PESQUITA, C., FARIA, D., BULL, S., PAWSON, T., MORRIS, Q. and WRANA, J. L. (2009). Dynamic modularity in protein interaction networks predicts breast cancer outcome. *Nat. Biotechnol.* **27** 199–204.

TRAN, T.-D. and KWON, Y.-K. (2013). The relationship between modularity and robustness in signalling networks. *J. R. Soc. Interface* **10** 20130771.

VAN HOESEL, A. Q., SATO, Y., ELASHOFF, D. A., TURNER, R. R., GIULIANO, A. E., SHAMONKI, J. M., KUPPEN, P. J. K., VAN DE VELDE, C. J. H. and HOON, D. S. B. (2013). Assessment of DNA methylation status in early stages of breast cancer development. *British Journal of Cancer* **108** 2033–2038.

VENTERS, B. J. and PUGH, B. F. (2013). Genomic organization of human transcription initiation complexes. *Nature* **502** 53–58.

VERSCHUUR-MAES, A. H., DE BRUIN, P. C. and VAN DIEST, P. J. (2012). Epigenetic progression of columnar cell lesions of the breast to invasive breast cancer. *Breast Cancer Res. Treat.* **136** 705–715.

VU, D. Q., HUNTER, D. R. and SCHWEINBERGER, M. (2013). Model-based clustering of large networks. *Ann. Appl. Stat.* **7** 1010–1039. MR3113499

WAGNER, A. (2002). Estimating coarse gene network structure from large-scale gene perturbation data. *Genome Res.* **12** 309–315.

WEI, P. and PAN, W. (2010). Network-based genomic discovery: Application and comparison of Markov random-field models. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **59** 105–125. MR2750134

XIE, W., SCHULTZ, M. D., LISTER, R., HOU, Z., RAJAGOPAL, N., RAY, P., WHITAKER, J. W., TIAN, S., HAWKINS, R. D., LEUNG, D. et al. (2013). Epigenomic analysis of multilineage differentiation of human embryonic stem cells. *Cell* **153** 1134–1148.

YAMAMOTO, E., SUZUKI, H., YAMANO, H., MARUYAMA, R., NOJIMA, M., KAMIMAE, S., SAWADA, T., ASHIDA, M., YOSHIKAWA, K., KIMURA, T. et al. (2012). Molecular dissection of premalignant colorectal lesions reveals early onset of the CpG island methylator phenotype. *Am. J. Pathol.* **181** 1847–1861.

DEPARTMENT OF STATISTICAL SCIENCE
UNIVERSITY COLLEGE LONDON
1-19 TORRINGTON PLACE
LONDON WC1E 7HB
UNITED KINGDOM
E-MAIL: thomas.bartlett.10@ucl.ac.uk