Network-based analysis identifies epigenetic biomarkers of esophageal squamous cell carcinoma progression

Chun-Pei Cheng¹,²,†, I-Ying Kuo³,†, Hakan Alakus²,⁴,⁵, Kelly A. Frazer²,⁴,⁶, Olivier Harismendy²,⁴, Yi-Ching Wang³,⁷,* and Vincent S. Tseng¹,⁸,*

¹Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan 701, Taiwan, ²Moores Cancer Center, University of California San Diego, La Jolla, CA 92093, USA, ³Institute of Basic Medical Sciences, National Cheng Kung University, Tainan 701, Taiwan, ⁴Department of Pediatrics and Rady Children’s Hospital, University of California San Diego, La Jolla, CA 92093, USA, ⁵Department of General, Visceral and Cancer Surgery, University of Cologne, Köln, Germany, ⁶Institute for Genomic Medicine, University of California San Diego, La Jolla, CA 92093, USA, ⁷Department of Pharmacology and ⁸Institute of Medical Informatics, National Cheng Kung University, Tainan 701, Taiwan

Associate Editor: Ziv Bar-Joseph

ABSTRACT

Motivation: A rapid progression of esophageal squamous cell carcinoma (ESCC) causes a high mortality rate because of the propensity for metastasis driven by genetic and epigenetic alterations. The identification of prognostic biomarkers would help prevent or control metastatic progression. Expression analyses have been used to find such markers, but do not always validate in separate cohorts. Epigenetic marks, such as DNA methylation, are a potential source of more reliable and stable biomarkers. Importantly, the integration of both expression and epigenetic alterations is more likely to identify relevant biomarkers.

Results: We present a new analysis framework, using ESCC progression-associated gene regulatory network (GRNescc), to identify differentially methylated CpG sites prognostic of ESCC progression. From the CpG loci differentially methylated in 50 tumor–normal pairs, we selected 44 CpG loci most highly associated with survival and located in the promotors of genes more likely to belong to GRNescc. Using an independent ESCC cohort, we confirmed that 8/10 of CpG loci in the promoter of GRNescc genes significantly correlated with patient survival. In contrast, 0/10 CpG loci in the promoter genes outside the GRNescc were correlated with patient survival. We further characterized the GRNescc network topology and observed that the genes with methylated CpG loci associated with survival deviated from the center of mass and were less likely to be hubs in the GRNescc. We postulate that our analysis framework improves the identification of bona fide prognostic biomarkers from DNA methylation studies, especially with partial genome coverage.

Contact: tsengsm@mail.ncku.edu.tw or ycw5798@mail.ncku.edu.tw

Supplementary information: Supplementary data are available at Bioinformatics online.

Received on November 28, 2013; revised on June 6, 2014; accepted on July 3, 2014

1 INTRODUCTION

Recently, systematic biological approaches to study cancer have provided unprecedented views of molecular changes in many cancers. For example, the mutagenesis within a network of general human cancer signaling genes (Cui et al., 2007) and the protein expression within a protein–protein interaction network (Ostlund et al., 2010) have led to the discovery of subnetworks involving cancer-related genes. The combination of protein–protein networks with gene expression microarray datasets has also been used to distinguish metastatic from non-metastatic tumor samples (Chuang et al., 2007; Garcia et al., 2012) or to identify biomarkers correlated with patient survival (Li et al., 2012). More recently, Sun and Wang (2013) used a genetic network as a reference to estimate the penalty score of a conditional logistic regression model and applied it on a matched tumor–normal analysis of DNA methylation array data to identify a list of candidate CpG sites associated with hepatocellular cancer development. Kim et al. (2012) attempted to integrate more biological resources like the epigenomic, transcriptomic and protein interactome data to identify glioblastoma prognostic biomarkers using gene expression and DNA methylation-based networks. Although DNA methylation can be used as a powerful and promising prognostic indicator alone (Laird, 2003), none of the aforementioned network-based studies, integrating DNA methylation, gene expression or protein expression information performed experimental validation of the identified biomarkers. This can be because of the large number of candidate biomarkers within networks, making their validation and use in the clinic more difficult.

Affecting >450,000 patients annually, esophageal carcinoma with squamous cell carcinoma (ESCC) as the predominant histological subtype worldwide is the sixth leading cause of cancer-related mortality, with >400,000 deaths per year (Pennathur et al., 2013; van Hagen et al., 2012). Late presentation with already existing lymph node metastasis (LNM) followed by rapid progression explains the poor outcome of the disease (Bollschweiler et al., 2006). Metastasis requires certain steps like primary tumor initialization and proliferation, blood vessel/lymphatic channel intravasation, cell arrest and extravasation
Network-based analysis identifies ESCC progression biomarkers

and proliferation at secondary target sites/organs (Hunter et al., 2008). Metastasis can arise from tumor cells that have undergone phenotypic changes called epithelial-to-mesenchymal transition (EMT), gaining plasticity and circulating and seeding ability. There is no observable change in DNA methylation during the transforming growth factor beta-mediated EMT in AML12 mouse hepatocyte (McDonald et al., 2011). But such DNA methylation is able to be involved in gene regulation during the EMT of prostate cell line, EP156T (Ke et al., 2010). Identifying epigenetic alterations occurring during EMT progression is therefore not only essential for a detailed understanding of the molecular biology underlying the disease progression but also to improve clinical prognosis and develop more sophisticated treatment strategies.

Until now, only few genomic regions, such as the retrotransposon-related long interspersed element 1 (Iwagami et al., 2013) or the gene regulatory elements, showing methylation alterations have been identified as possible biomarkers for LNM and/or patient survival in ESCC. Of the gene annotation-based studies, hyper-methylation at CpG islands (CGIs) in the vicinities of PAX6 (paired box 6) and RN7SKP211 (RNA, 7SK small nuclear pseudogene 211) were significantly associated with LNM and disease-free survival in 96 patients (Gyobu et al., 2011). Hyper-methylated CGIs located within UCHL1 (ubiquitin carboxyl-terminal esterase L1) (Mandelker et al., 2005), FHIT (fragile histidine triad) (Lee et al., 2006), GRIN2B (glutamate receptor, ionotropic, N-methyl D-aspartate 2B) (Kim et al., 2007) and GADD45G (growth arrest and DNA-damage-inducible, gamma) (Guo et al., 2013) promoter regions were associated with poor survival. However, these studies only validated CpG sites in a small set of candidate genes, therefore limiting the scope of the findings. A more comprehensive analysis is likely to reveal new DNA methylation as biomarkers associated with LNM and survival.

In this study, we use a new comprehensive approach to efficiently identify and validate DNA methylation sites as putative prognostic biomarkers of ESCC progression. We propose an intuitive framework, and demonstrate its ability to identify CpG sites of prognostic value. The framework leverages an ESCC progression-associated gene regulatory network (GRNgesc) to identify methylated sites with significant prognostic value. By taking into account differentially methylated CpG sites whose corresponding gene promoters are ranked, the ranked CpG sites are purified/selected via a top-k precision test in network. We validate the results on a selection of 20 CpG loci in a separate cohort of ESCC patients and demonstrate that this framework is capable of identifying novel sites of DNA methylation with prognostic impact that had not been discovered by previous approaches.

2 METHODS

2.1 Patients and biopsy specimens

This study was verified and qualified by the institutional review board of National Cheng Kung University Hospital from May 1, 2010 to July 31, 2011 under contract number ‘HR-99-021’. The ethics committee specifically waived the need for informed consent forms because the data were publicly obtained from an observational study and analyzed anonymously. We enrolled 100 ESCC patients admitted to the Cancer Center and Pathology department, National Cheng Kung University Hospital (N = 80) and the Cancer Center, China Medical University Hospital (N = 20). Primary ESCC specimens and matched normal tissues, located >10 cm from the primary site, were collected through surgical resection. Pathologic examination of the resected surgical specimens was performed following a standardized protocol, and the specimens were classified according to the sixth edition of the UICC TNM (Union for International Cancer Control, TNM Classification of Malignant Tumours) system and the WHO classification. Although surgically resected tumor tissue and corresponding normal tissue samples were collected from two separate hospitals, the samples were processed by the same laboratory, using the same protocol, therefore limiting potential batch effects. Follow-up of enrolled patients was performed at 6 months interval, with the last follow-up performed at least 12 months and up to 104 months after diagnosis for living patients. The enrolled patients were randomly split between screening (50 patients) and validation (50 patients) cohorts. The general clinicopathological characteristics of the enrolled patients are shown in Supplementary Table S1.

2.2 Construction of an ESCC-related gene regulatory network

We built a general gene regulatory network (GRNg) using three publicly available networks: Pathway Commons (11/2011 version; Cerami et al., 2011), BioGRID 3.1.79 (Stark et al., 2006) and KEGG (Kyoto Encyclopedia of Genes and Genomes) (09/2011 version; Ogata et al., 1999). The consolidated network features 1294769 gene regulations (edges) and 12803 genes (nodes). In this study, we focused more on direct gene regulations because the DNA methylation is a major epigenetic event that blocks binding of transcription factors to promoters of target genes, or modifies chromatin structure, which in turn blocks transcription factor binding (Suzuki and Bird, 2008). Therefore, only interactions derived from transcriptional regulation were considered, excluding protein-protein and protein compound interactions. We then generated GRNgesc. We selected 186 genes by curating the literature and identified genes whose expression pattern is associated with ESCC progression (Supplementary Table S2) before January 2012. The progression refers to the cancer metastasis, proliferation, arrest, invasion and patient survival. We excluded genes showing differential expression between tumor and normal but that could not be associated with ESCC progression. We then generated the GRNgesc as a subnetwork of GRNg using the following steps: (i) Initiate anempty distance matrix with a length equal to the number of literature-curated genes, (186 x 186), (ii) calculate the shortest distance between each pair of genes projected on GRNg using the Dijkstra’s algorithm and (iii) calculate the shortest paths of each pair of genes on GRNgesc using a breadth-first search algorithm. The resulting GRNgesc contained 1013604 interactions between 4636 genes. A non-ESCC progression-associated gene regulatory network (GRNgesc) was also derived from the complement of GRNgesc in GRNg. More precisely, after generating the GRNgesc, via the above three steps, every gene (node) of GRNgesc and its connected gene–gene regulations (edges) in GRNg were removed from GRNg. We therefore called the rest part GRNgesc as a negative control used in this study.

2.3 Generation and analysis of DNA methylation microarray data

2.3.1 Microarray data generation The genomic DNA from primary ESCC and normal esophagus specimen was extracted using proteinase K digestion and phenol-chloroform extraction. One microgram of DNA was then converted using bisulfite following the directions from the EpiTect Bisulfite kit (Qiagen, Duesseldorf, Germany), converting unmethylated cytosines to uracil and then to thymidine in the subsequent PCR step. We used the Illumina’s GoldenGate Methylation Assay Cancer Panel 1 (1505 CpG dinucleotides located in the promoter of 807 genes; Illumina, San Diego, CA, USA) following the
2.3.2 Microarray data analysis The ratio of fluorescent signals was computed from the two alleles beta = (max(M, 0))/((U + |M| + 100), where U is the green fluorescent signal (Cy3) from an unmethylated allele and M is the red signal (Cy5) from a methylated allele, generated by the Illumina’s proprietary software (BeadStudio). The beta-value reflects the methylation level of each CpG site (Bibikova et al., 2006), and their distribution is shown in Supplementary Figure S1A. To allow further statistical analyses able to be applicable to these values across different samples, the beta-values were then normalized using the function of normalize.loess implemented in Bioconductor affy package with four parameters including epsilon (0.01), log.it (F), span (0.4) and maxit (5). Then we kept all normalized values positive by adding an absolute (the minimum value). Their distribution is shown in Supplementary Figure S1B. We identified significantly differentially methylated CpG sites between tumor and normal using a two-tailed Student’s t-test (P<0.05). CpG loci had a significant increase (respectively decrease) in methylation when methylation is increased by N-fold or greater (respectively ~N-fold) in the tumor compared with normal, with N corresponding to the median of absolute fold changes between tumor and normal.

2.3.3 Identification of CpG sites associated with progression We constructed a contingency table for each significantly different CpG site, counting the number of patients with or without LNM and for which the probe significantly increased or decreased methylation CpG. This table can be used to analyze the relationships between two categorical variables: methylation change in tumor (increase/decrease) and metastasis status (N0/N1). We then calculated, for each CpG site, the following six correlation metrics for each CpG site: PhiCoefficient (Cramer, 1946), OddsRatio (Edwards, 1963), PiatskyskyShapiroMeasure (Pietskysky-Shapiro, 1991), LiftMeasure (Tuffery, 2011), AddedValue (Sahar and Mansour, 1999) and KlosgenMeasure (Klosgen, 1992). The detailed equations are given as Supplementary Method S1. For each of these metrics, a positive value indicates a positive correlation between the direction of the methylation change and the LNM status, at each CpG site. This resulted in the identification of 130 progression-associated CpG sites.

2.3.4 Identification of CpG sites associated with survival The last follow-up of enrolled patients was performed at least 12 months after diagnosis for living patients. The first group (‘Good’ survival) included patients who were still alive after 12 months following tumor resection and the second group (‘Bad’ survival) consisted of patients who died within 12 months post-resection. Coincidently, the two groups have complement of this subnetwork in the global network is the most associated with progression (Fig. 1 panel III). (iv) We construct the disease-specific gene regulatory network. This is done by extracting the minimal path subnetwork containing genes important for the disease progression, as identified through manually curated references (Fig. 1 panel I). The complement of this subnetwork in the global network is extracted and used as a negative control (see also Section 2.2). (ii) We identify cancer-specific CpG methylation events (Fig. 1 panel II), then (iii) we select CpG sites where methylation change is the most associated with disease progression (Fig. 1 panel III). (iv) We then rank these candidate CpG sites by the association of the methylation change with patient’s survival (Fig. 1 panel IV), and finally (v) we use the disease-specific network to select the top candidate CpG sites associated with disease progression and patient’s survival (Fig. 1 panel V). Applying this approach to ESCC progression, we first built a GRN of using publicly available networks (Section 2). This GRN features 1294769 gene regulations (edges) and 12803 genes (nodes). We then extracted the minimal-path subnetwork featuring 186 genes.

2.5 Pyrosequencing validation and survival analysis The bisulfite-converted DNA was pyrosequenced using the PyroMark Q24 (Qiagen). We designed specific pyrosequencing primer and PCR primer using the specialized software (PyroMark Assay Design 2.0) to target the CpG sites in the promoter region of selected gene (Supplementary Table S3). Bisulfite-modified DNA was dissolved in 20μl H2O, and 1 μl of DNA template was used for PCR amplification. Hot-start PCR was performed with PyroMark PCR Kit (Qiagen), and pyrosequencing was carried out according to the manufacturer’s protocol (Qiagen). The target CpG sites were evaluated by converting the resulting pyrograms to numerical values for peak heights. The percentage of methylation was calculated as the mean of all CpG analyzed (Vaisiere et al., 2009). We finally performed a survival analysis by using these methylation percentages to validate the screening cohort-derived candidate probes containing CpG sites.

2.6 Quantitative RT-PCR For the mRNA quantifications, we performed SYBR Green and TaqMan® Gene Expression Assay (Life Technologies Corporation) qRT-PCR methods to detect mRNAs in the same validation cohort. If genes were not suitable for the primer design of SYBR Green qRT-PCR, we alternatively performed the TaqMan® method using its commercial primers (Supplementary Table S4). We analyzed the results using the cycle threshold method (Ct). Only strong signals with high expressions (Ct<35) were used for a further correlation analysis with promoter methylation.

3 RESULTS AND DISCUSSION

3.1 Overall framework and ESCC network construction Cancer metastatic progression may be associated with multiple gene regulatory changes, some of them mediated by aberrant promoter CpG methylation. To identify the CpG probes where methylation is the most associated with disease progression, we propose the following framework comprising five different steps. (i) We construct a disease-specific gene regulatory network. This is done by extracting the minimal path subnetwork containing genes important for the disease progression, as identified through manually curated references (Fig. 1 panel I). The complement of this subnetwork in the global network is extracted and used as a negative control (see also Section 2.2). (ii) We identify cancer-specific CpG methylation events (Fig. 1 panel II), then (iii) we select CpG sites where methylation change is the most associated with progression (Fig. 1 panel III). (iv) We then rank these candidate CpG sites by the association of the methylation change with patient’s survival (Fig. 1 panel IV), and finally (v) we use the disease-specific network to select the top candidate CpG sites associated with disease progression and patient’s survival (Fig. 1 panel V). Applying this approach to ESCC progression, we first built a GRN of using publicly available networks (Section 2). This GRN features 1294769 gene regulations (edges) and 12803 genes (nodes). We then extracted the minimal-path subnetwork featuring 186 genes.
associated with ESCC progression (Section 2 and Supplementary Table S2). This GRN\textsubscript{esc} contains 1,013,604 interactions between 4,636 genes. Finally, we derived a non-ESCC-associated gene regulatory network (GRNg\textsubscript{esc}) from the complement of GRN\textsubscript{esc} in GRN\textsubscript{g}. A Gene Ontology (GO) analysis (Dennis et al., 2003) revealed that the genes in GRN\textsubscript{esc} were enriched in biological processes such as chemotaxis, cell adhesion, cell migration, and angiogenesis, compared with the GRN\textsubscript{g-esc} (Fig. 2). These processes are important for metastasis and cancer progression. This observation indicates that the genes in GRN\textsubscript{esc} are related to cancer progression, extending our initial gene list and likely accounting for unsuspected regulatory patterns important for disease progression and metastasis.

### 3.2 Identification of candidate CpG sites associated with ESCC progression

Using a microarray, we measured the methylation status of 1,505 CpG sites located in the promoter of 807 cancer-related genes in 50 ESCC tumors and matched normal esophageal tissue. We identified 309 differentially methylated sites between tumor and normal (t-test nominal $P < 0.05$), of which 108 and 201 had decreased and increased methylation in the tumor, respectively. We then determined which CpG sites were associated with cancer progression, e.g., lymph node metastasis, in our cohort. Using a compendium of six correlation metrics, we looked for CpG sites with significant changes in methylation in the tumor of metastatic patients (LNM classification N1). Using this approach, we were able to identify 130 CpG sites in the promoter of 109 genes, which are associated with ESCC lymph node metastasis. To characterize these sites and extract a specific ESCC progression epigenetic signature, we further analyzed them using the global regulatory network.

### 3.3 Network-based selection of Candidate CpG sites associated with survival

To increase our confidence in the biological significance of the CpG sites identified above, we calculated their association with patient survival. Using a dynamic classification of patients with increased and decreased CpG methylation to compare groups of the same size (section 2.3.4), we ranked the 130 CpG sites by decreasing association of their methylation status and patient survival (log-rank test $P$-value). To further select the CpG sites where methylation status is the most likely to be associated with

---

**Fig. 1.** Schematic overview of data processing steps. (I) Development of literature-guided gene regulatory networks. The circles and arrows represent the regulatory genes and regulations, respectively. (II) Identification of differentially methylated CpG sites associated with ESCC. (III) Selection of differentially methylated CpG sites associated with ESCC progression. (IV) Ranking of CpG sites based on the association with patient survival. (V) Selection of the ranked CpG sites using a network-based approach. (VI) Validation of network-selected top-ranked CpG sites in a new patient cohort. The circles indicate increased methylation (upward diagonal), decreased methylation (downward diagonal) in the tumor or literature curated (filled).

**Fig. 2.** Enrichment analysis of biological processes GO terms for the three different GRNs studied. $P$-value as a function of GO terms.
To further distinguish the relative importance of increase and decrease in methylation at the CpG sites in the promoter of the genes in the networks, we split the ranked list of 130 CpG sites in equal size bins of decreasing association with survival (Fig. 3C). There was no significant bias between increased and decreased methylation at CpG in the promoters of these genes, indicating that both repression and activation of genes in the network may contribute to ESCC progression.

We finally selected 44 best candidate CpG sites (Fig. 3A – arrow) for further analysis (Supplementary Table S5). Of the 44 CpG sites, 22 were located in the promoters of genes belonging to GRNesc (referred to as In-CpG sites) and 22 where in promoters of genes outside GRNesc (referred to as Out-CpG sites). This enrichment of CpG with changing methylation is significantly different from what can be expected by chance ($\chi^2$ test $P<0.0001$). Moreover, only 5 of the 22 In-CpGs and none of the 11 Out-CpGs were located in the promoters of the 186 genes that seeded the GRNesc network, suggesting that the In-CpG methylation changes were likely to be associated with progression and close to the 186 genes in network. To confirm this possibility, we compared the average distance between the 186 seed genes and the genes whose promoters have CpG methylation changes associated with progression. Of the 22 Out-CpG sites, 11 were located in the promoters of genes belonging to GRNg-esc (referred to as Out-CpGsesc sites), and 11 were in the promoters of genes not represented in the GRNg network. The genes with In-CpG were significantly closer to a seed gene than the genes with Out-CpGsesc (average distance of 2.7 versus 3.2, $t$-test $P$-value $<1E-30$). The average distance of genes with In-CpG to a seed gene is in fact similar to the average distance of the seed genes between themselves (2.7 versus 2.6, $t$-test $P$-value $= 3.6E-09$). This suggests that network approach enables the identification of CpG methylation changes in the promoter of genes not previously associated with progression and, therefore, increases the number of potential prognostic biomarkers that can be tested. Moreover, although both inside and outside the GRNesc had the identical number (22) of genes, the 22 genes in GRNesc were originally from top-ranked genes (Fig. 3A and B and Supplementary Table S5—the median rank of In-CpG sites versus the median rank of Out-CpG sites = 18 versus 27). The 22 In-CpG sites in genes from GRNesc are therefore more likely to have a methylation status associated with ESCC metastatic progression and are good candidates to test their prognostic value.

3.4 Validation of the findings

To validate the association between these candidate CpG sites methylation and survival (Fig. 1 panel VI), we decided to measure the association in a new cohort (validation cohort) of 50 patients with ESCC and matched normal esophageal tissues. We were able to design specific primers for 10 In-CpGs as well as 10 control Out-CpGs (Section 2 and Supplementary Table S3).

We first checked the validity of the methodology by determining the methylation level of these 20 CpG sites in the screening cohort (N = 50 patients). This analysis showed that the methylation level determined by pyrosequencing was highly correlated with the one obtained from the microarray (Supplementary Fig. S2; $r = 0.78$), therefore demonstrating the technical validity of the approach. We then examined the methylation of these 20 CpGs in the validation cohort (N = 50 patients). We first noticed that, the methylation change is significant in 12/20 CpG ($P<10^{-4}$), either through increased (N = 7) or decreased methylation (N = 5) in the tumor. Additionally, the methylation changes of 8/10 In-CpGs and 0/10 Out-CpG were associated with patient survival (Table 1—log-rank test $P<0.05$). Of eight validated In-CpGs, only one is located in the promoter of a seed gene (MAPK4, Supplementary Table S2), suggesting the increased sensitivity provided by the network approach. To further confirm the association between methylation changes and survival, we performed a Cox regression analysis using SPSS v17 on the methylation changes in addition to other clinical variable (Supplementary Table S6). We calculated the hazard ratio (HR) of cancer death risk of variables including promoter methylation.
Table 1. Pyrosequencing-based validation of methylated In-CpG and Out-CpG sites

<table>
<thead>
<tr>
<th>CpGs</th>
<th>Distance from TSS</th>
<th>Gene</th>
<th>Fold change (P-value)</th>
<th>Association with survival P-value</th>
<th>Survival correlation direction</th>
</tr>
</thead>
<tbody>
<tr>
<td>In-CpG</td>
<td>+64</td>
<td>JAK3</td>
<td>1.8 (&lt;0.0001*** )</td>
<td>0.033*</td>
<td>–</td>
</tr>
<tr>
<td>In-CpG</td>
<td>–1,121</td>
<td>PAX6</td>
<td>1.3 (0.314)</td>
<td>0.041*</td>
<td>–</td>
</tr>
<tr>
<td>In-CpG</td>
<td>–115</td>
<td>CFTR</td>
<td>1.3 (0.047*)</td>
<td>0.188</td>
<td>NA</td>
</tr>
<tr>
<td>In-CpG</td>
<td>–516</td>
<td>EZF5</td>
<td>1.6 (&lt;0.0001*** )</td>
<td>0.024*</td>
<td>–</td>
</tr>
<tr>
<td>In-CpG</td>
<td>–272</td>
<td>CD81</td>
<td>1.1 (0.921)</td>
<td>0.031*</td>
<td>–</td>
</tr>
<tr>
<td>In-CpG</td>
<td>+53</td>
<td>CCL3</td>
<td>–1.4 (&lt;0.0001*** )</td>
<td>0.015*</td>
<td>+</td>
</tr>
<tr>
<td>In-CpG</td>
<td>–8</td>
<td>CSF3R</td>
<td>–1.1 (&lt;0.429*)</td>
<td>0.154</td>
<td>NA</td>
</tr>
<tr>
<td>In-CpG</td>
<td>–804</td>
<td>INS</td>
<td>–1.2 (&lt;0.0001*** )</td>
<td>0.040*</td>
<td>–</td>
</tr>
<tr>
<td>In-CpG</td>
<td>+273</td>
<td>MAPK4</td>
<td>–1.2 (&lt;0.0001*** )</td>
<td>0.001**</td>
<td>–</td>
</tr>
<tr>
<td>In-CpG</td>
<td>–456</td>
<td>PGR</td>
<td>–1.4 (&lt;0.0001*** )</td>
<td>0.023*</td>
<td>+</td>
</tr>
<tr>
<td>Out-CpG</td>
<td>–38</td>
<td>SLC5A8</td>
<td>2.0 (0.008*)</td>
<td>0.153</td>
<td>NA</td>
</tr>
<tr>
<td>Out-CpG</td>
<td>+26</td>
<td>PENK</td>
<td>2.0 (0.0001*** )</td>
<td>0.079</td>
<td>NA</td>
</tr>
<tr>
<td>Out-CpG</td>
<td>–546</td>
<td>HS3ST2</td>
<td>1.9 (&lt;0.0001*** )</td>
<td>0.323</td>
<td>NA</td>
</tr>
<tr>
<td>Out-CpG</td>
<td>+3</td>
<td>KCNK4</td>
<td>1.7 (&lt;0.0001*** )</td>
<td>0.091</td>
<td>NA</td>
</tr>
<tr>
<td>Out-CpG</td>
<td>–299</td>
<td>SEZ6L</td>
<td>1.3 (0.142)</td>
<td>0.252</td>
<td>NA</td>
</tr>
<tr>
<td>Out-CpG</td>
<td>–22</td>
<td>ZIM2</td>
<td>1.2 (&lt;0.0001*** )</td>
<td>0.206</td>
<td>NA</td>
</tr>
<tr>
<td>Out-CpG</td>
<td>–455</td>
<td>ADCYAP1</td>
<td>2.7 (&lt;0.0001*** )</td>
<td>0.536</td>
<td>NA</td>
</tr>
<tr>
<td>Out-CpG</td>
<td>–1,394</td>
<td>PIE</td>
<td>–1.3 (0.248)</td>
<td>0.245</td>
<td>NA</td>
</tr>
<tr>
<td>Out-CpG</td>
<td>+340</td>
<td>SFTPA1</td>
<td>–1.4 (&lt;0.0001*** )</td>
<td>0.400</td>
<td>NA</td>
</tr>
<tr>
<td>Out-CpG</td>
<td>–721</td>
<td>TRPM5</td>
<td>–1.1 (0.023*)</td>
<td>0.324</td>
<td>NA</td>
</tr>
</tbody>
</table>

*Note: TSS, transcription start site; NA, not applicable; Fold change, pyrosequencing values between matched ESCC and normal adjacent tissue.

**P-value of t-test.

***P-value of log-rank test.

The direction of correlation was considered as ‘+’ (respectively ‘–’) when the methylation increase in tumor led to a good (respectively poor) survival rate. *, P<0.05; **, P<0.001; ***, P<0.0001.

change, TNM stage, local LNM status, distant metastases status, age and drinking status. This analysis showed that methylation changes at five of eight validated In-CpGs and distant metastasis were associated with a significantly increased risk (HR >1) or decreased risk (HR <1) of cancer-related death, and three In-CpGs showed a borderline significance, while none of the methylation changes at Out-CpGs (negative control group) was associated with the risk of cancer-related death. A multivariate analysis further showed that 5/8 validated In-CpGs remain significant association with prognosis even after accounting for the presence of distant metastasis.

The validation results (Table 1), including the methylation changes between tumor and normal tissues and the direction of survival associations, of all of the eight validated CpGs were consistent with the results in the screening cohort (Supplementary Table S7 and Supplementary Fig. S3). These results suggest that, despite the limitation of the cohort size to identify significant methylation changes in the tumors, the network-based framework was able to enrich for CpG sites significantly associated with survival.

Promoter CpG methylation usually results in transcriptional repression. In Supplementary Figure S4, we measured the expression changes between normal esophagus and ESCC for six genes with strong reliable signals (Section 2) in the GRNesc and whose promoters contain CpG associated with survival. We can identify trends for negative correlation for four increased methylation CpGs or positive correlation for two decreased methylation CpGs at least 10 patients in the validation cohort. Referring to previous literature, a hyper-methylated CpG island located 5300 bp upstream from the transcriptional start site of PAX6 was found to date as the only biomarker to be associated with LNM (Gyobu et al., 2011). However, the authors claimed that although this CpG island was unlikely to be associated with repression of PAX6, it was quantified in four ESCC cell lines in three of which PAX6 was expressed in spite of CpG island methylation. Their results suggested that the methylation status would not always correlate with gene expression. Therefore, in agreement with this study, our results indicate that methylation changes at selected CpG sites can be good prognostic markers even in absence of a clear effect on transcriptional regulation.

A role for inflammation in tumorigenesis is emerging. Inflammatory responses play pivotal roles at tumor progression including tumor initiation, promotion, invasion and metastasis. Tumors are frequently surrounded by an inflammatory microenvironment rich in cytokines, chemokines and immune cells infiltration, which promote malignant cellular growth. These factors are produced by the tumor cells or its surrounding tissue and contribute to malignant progression (Grivennikov et al., 2010). Interestingly, we also validated several genes associated with inflammation. For example, CCL3 is a cytokine in the TNF inflammation pathway (Wang et al., 2013), and JAK3 is predominantly expressed in immune cells and transduces a signal in response to its activation via tyrosine phosphorylation by interleukin receptors (Krejsgaard et al., 2011). Different cytokines can either promote or inhibit tumor development and progression (Lin and Karin, 2007). Previous studies indicated that
interleukin 6 (IL-6), a pro-inflammatory cytokine that mediates chronic inflammation, may play an important role in inflammation-driven oral carcinogenesis. Notably, Jacqueline and associates recently found that IL-6 induces hyper-methylation and gene silencing mediated by DNMTs (mammalian DNA methyltransferases) (Gasche et al., 2011). In this study, the changes of global DNA methylation and gene-specific promoter methylation patterns by IL-6 treatment in oral cancer cells were examined. The increased promoter methylation changes were identified in several tumor suppressor genes, including CHFR, GATA4, and PAX6 (Gasche et al., 2011). Of these, we confirmed PAX6 with increased methylation in ESCC. Together, the role of inflammation in relation of promoter methylation of PAX6, CCL3 and JAK3 in ESCC tumor microenvironment is worthy of further investigation.

3.5 GRNesc network topology
In an effort to understand better the importance of the network in identifying significant associations with cancer progression and survival, we characterized further the network topology. Previous studies have shown that the network topology of certain genes might have functional implications in a cell. For example, an enrichment of genes having lethal knockout phenotypes possessed a high-degree (hub) property in a Saccharomyces cerevisiae gene co-expression network (Carter et al., 2004). Therefore, it is plausible that the CpG methylation changes of promoters of genes in GRNesc might have certain interesting distributions. We examined whether the gene promoters contain differentially methylated CpG possessed specific characteristics in the GRNesc network. We ranked these genes based on their decreasing association with patient survival. We noticed that both the barycenter score (White and Smyth, 2003) and the closeness centrality (Opsahl et al., 2010) showed a negative correlation with significance. This observation suggests that the genes associated with survival tend to deviate from the center of mass of the GRNesc and to be more located at the periphery of the network. This is similar to a recent work that age-associated epigenetic drift occurs preferentially in genes that occupy peripheral network positions (West et al., 2013). In another analysis shown in Figure 4C and D, the significance for survival was negatively correlated with the HITS Hub node topological measurements in GRNesc. (A) Barycenter scores, (B) Closeness centrality, (C) HITS hub and (D) HITS authority as a function of log-rank test P-values of ranked genes. The further right along x-axis indicates the greater propensity for ESCC progression. Red line: A linear regression. P: P-value is calculated by testing Spearman’s rank-order correlation coefficient (rho) with an F-distribution

4 CONCLUSIONS
In this study, we proposed a new framework that uses literature-guided GRN to enhance the results and interpretation of DNA methylation microarray experiments. Specifically, the framework helps prioritize differentially methylated genes for their impact on cancer progression and survival. We validated the results in an independent cohort and confirmed that the selected CpG sites were significantly associated with patient survival, even in absence of a direct correlation with the gene expression. Eight of 10 validated CpG sites significantly correlated with patient survival. These were located in the promoters of JAK3, PAX6, E2F5 and CD81 (increased methylation), and in the promoters of CCL3, INS, MAPK4 and PGR (decreased methylation). Interestingly, the position of the survival-associated genes in the GRNesc network significantly deviated from the center of mass. We postulate that the topology of progression-associated network could help identify progression-associated genes before any data collection. Our results demonstrate that the use of regulatory networks and prior expression studies can help identify bona fide DNA-methylation prognostic biomarkers. Although our focus is the identification of biomarkers for a clinical use via a methodological innovation, the functional exploration of these biomarkers is also worthy of further investigation.

Conflict of interest: none declared.

REFERENCES
Network-based analysis identifies ESCC progression biomarkers