Published online 14 June 2016

Nucleic Acids Research, 2016, Vol. 44, No. 16 ***e134***
doi: 10.1093/nar/gkw546

NEpiC: a network-assisted algorithm for epigenetic studies using mean and variance combined signals

Peifeng Ruan[1], Jing Shen[2], Regina M. Santella[2], Shuigeng Zhou[1] and Shuang Wang[3,*]

[1]School of Computer Science and Shanghai Key Lab of Intelligent Information Processing, Fudan University, Shanghai 200433, China, [2]Department of Environmental Health Science, Mailman School of Public Health, Columbia University, New York, NY 10032, USA and [3]Department of Biostatistics, Mailman School of Public Health, Columbia University, New York, NY 10032, USA

Received December 22, 2015; Revised May 23, 2016; Accepted June 04, 2016

ABSTRACT

DNA methylation plays an important role in many biological processes. Existing epigenome-wide association studies (EWAS) have successfully identified aberrantly methylated genes in many diseases and disorders with most studies focusing on analysing methylation sites one at a time. Incorporating prior biological information such as biological networks has been proven to be powerful in identifying disease-associated genes in both gene expression studies and genome-wide association studies (GWAS) but has been under studied in EWAS. Although recent studies have noticed that there are differences in methylation variation in different groups, only a few existing methods consider variance signals in DNA methylation studies. Here, we present a network-assisted algorithm, NEpiC, that combines both mean and variance signals in searching for differentially methylated sub-networks using the protein–protein interaction (PPI) network. In simulation studies, we demonstrate the power gain from using both the prior biological information and variance signals compared to using either of the two or neither information. Applications to several DNA methylation datasets from the Cancer Genome Atlas (TCGA) project and DNA methylation data on hepatocellular carcinoma (HCC) from the Columbia University Medical Center (CUMC) suggest that the proposed NEpiC algorithm identifies more cancer-related genes and generates better replication results.

INTRODUCTION

DNA methylation plays critical roles in many biological activities, especially in the carcinogenesis process. Two common kinds of aberrant methylation in cancers are regional hypermethylation and global hypomethylation. Hypermethylation within promoter regions of genes, which may lead to the silence of associated genes, is known to be associated with various kinds of cancers, such as liver (1), renal (2), colorectal (3) and endometrial (4) cancers. Global hypomethylation mainly affects intergenic regions of the genome and may increase chromosomal instability (5).

Many epigenome-wide association studies (EWAS) have successfully identified aberrantly methylated genes in cancers (6–8) with most studies focusing on analyzing DNA methylation sites one at a time. Several methods have being developed that consider correlations among sites on a gene or correlations among genes in a pathway (9,10) using penalized regression models. In genome-wide association studies (GWAS) or gene expression studies, incorporating prior biological information has been proven to be a more effective way to identify disease-associated single nucleotide polymorphisms (SNPs) or genes that are enriched with stronger association signals and higher biological relevance (11–15). Those methods prioritize candidate SNPs or genes by incorporating prior biological information such as gene annotations, biological pathways or protein–protein interaction (PPI) networks. More specifically, network-assisted methods overlay genetic or gene expression signals onto a biological network and search for sub-networks (modules) with GWAS data or gene expression data. Jia *et al.* (13) proposed a dense module searching method for GWAS (dmGWAS) that searches for modules that are enriched with genes of higher significances (low *P*-values) within a PPI network and showed that dmGWAS is more powerful in identifying disease related genes than other methods that do not incorporate network information. There are also some EWAS studies that incorporate biological network information (16,17). Another feature of DNA methylation measures that was recently observed is the higher variation in cancer tissues than in normal tissues across human cancer types (18). A few methods for DNA methylation data that consider differences in variances between two experimental conditions have already been developed (19–21). However, there is no method that incorpo-

*To whom correspondence should be addressed. Tel: +1 212 342 4165; Fax: +1 212 305 9408; Email: sw2206@columbia.edu

rates both the prior biological information such as the network information and variance signals in DNA methylation studies. In this paper, we propose the NEpiC algorithm, a Network-assisted algorithm for Epigenetic studies that uses mean and variance Combined signals in searching for differentially methylated sub-networks in a PPI network.

In the proposed NEpiC algorithm, we first compute mean and variance signal scores for a CpG site and then summarize the two scores with weights to create a combined score for the CpG site. We then extract the gene-level score out of all CpG sites on a gene. Finally, using a PPI network, we search for dense modules that are enriched for genes with large gene-level scores with a greedy search algorithm. We conducted simulation studies to show the performance of the proposed NEpiC algorithm compared to methods that either do not use the biological network information or do not use variance signals in searching for differentially methylated genes. We applied NEpiC to the 450K DNA methylation datasets of tumor and adjacent normal tissues of breast invasive carcinoma (BRCA) and liver hepatocellular carcinoma (LIHC) from the Cancer Genome Atlas (TCGA) project as well as an independent 450K DNA methylation data of tumor and adjacent normal tissues of hepatocellular carcinoma (HCC) from the Columbia University Medical Center (CUMC) (8). The results show that the proposed NEpiC algorithm which uses the biological network information among genes and both mean and variance signals at each CpG site identifies more cancer-related genes and generates better replication results than methods that do not consider both pieces of information.

MATERIALS AND METHODS

Since matched case–control designs with tumor and adjacent normal tissues are frequently used in DNA methylation studies of cancer, we focused on studies with a matched case–control design here. However, the proposed NEpiC algorithm is readily modified and applied to other types of designs. There are three steps in the proposed NEpiC algorithm: (i) constructing site-level and gene-level signal scores using DNA methylation data; (ii) searching modules on the PPI network with the guide of gene-level scores; (iii) prioritizing modules and candidate genes; (iv) validating identified modules using permutations. Figure 1 displays the pipeline of the proposed NEpiC algorithm.

Step 1: Constructing scores

Site-level mean and variance signal scores. We use the two-sided paired t-test and the one-sided Morgan–Pitman Test (22,23) to calculate *P*-values to test if the mean methylation measures are the same between tumor and adjacent normal groups and if the variance of the methylation measures in the tumor groups is greater than that in the adjacent normal group at CpG site i, which are denoted as p_{mi} and p_{vi}, respectively. The mean and variance signal scores at CpG site i are then defined as $m_i = \Phi^{-1}(1 - p_{mi})$, and $v_i = \Phi^{-1}(1 - p_{vi})$, where Φ is the standard normal distribution function. We set those mean and variance scores that are smaller than zero (i.e., those sites with mean and variance *P*-values > 0.5) to be zero and delete the CpG sites whose mean and variance scores are both zero.

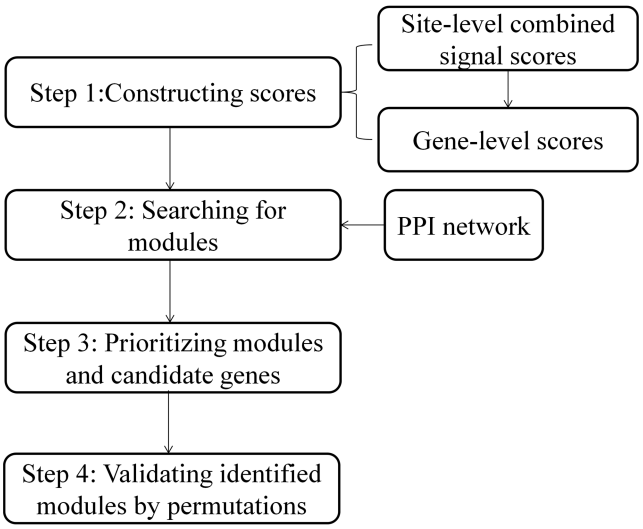

Figure 1. NEpiC algorithm pipeline.

Site-level combined signal scores. Due to potentially different scales for site-level mean and variance signal scores, we define a combined signal score c_i at CpG site i weighted by λ: $c_i = \lambda m_i + (1 - \lambda)v_i$ to balance the mean and variance scores. We calculate λ as follows: we first define a ratio $r_i = \frac{v_i}{m_i + v_i}$ at each CpG site i; we then average across all sites on a gene to obtain the gene-level ratio; lastly, we average gene-level ratios across all genes genome-wide to obtain λ.

Gene-level combined signal scores. We choose the CpG site with the largest combined signal score c_i to represent gene j, and denote this gene-level score with g_j, $j = 1, \ldots, J$, where J is the total number of genes.

Step 2: Searching for modules

To search for modules enriched with genes of high gene-level signal scores, we define module scores S as follows:

$$S = \frac{\sum_{j=1}^{m} g_j}{\sqrt{m}},$$

where m is the number of genes in a module. We use a greedy search algorithm with the following steps to search for modules. (i) Set a gene on the PPI network as the seed gene, which is considered as the starting module, and calculate the module score S. (ii) Identify the gene with the largest gene-level score g_j from all genes that are the first order neighbors of the seed gene on the PPI network and add this gene to the starting module only if the module score increases by a predefined cutoff, e.g., >10% (13). (iii) Continue with the first-order genes of the genes in the current module and keep adding genes if the module score increases by >10%, otherwise stop the algorithm and save the current module as the module corresponding to the seed gene. (iv) Choose another gene on the PPI network as the seed gene and repeat the above steps until all genes on the PPI network have been considered as a seed gene.

Step 3: Prioritizing modules and candidate genes

We now have the same number of modules as the number of genes on the PPI network. We exclude the small modules of size smaller than five genes (13). Since modules with more genes have larger module scores S in general, to make modules of different sizes comparable, we normalize module scores according to their sizes through a resampling method. More specifically, for each module obtained, we randomly generate 100 000 modules of the same size from the PPI network and calculate their module scores. We then normalize the module score S and define a normalized module score S_N as $S_N = \frac{S-\mu}{\sigma}$, where μ and σ are the mean and standard deviation of those 100 000 module scores. We then rank the modules by their normalized module scores and select the top ranked modules. Genes that appear in the selected top-ranked modules are candidate genes that are potentially associated with the tumor status. We also rank the candidate genes using the frequencies that each candidate gene is selected by the selected top-ranked modules. We define those candidate genes that are selected by more than one module as the prioritized candidate genes.

Step 4: Validating identified modules using permutations

To control for the potential bias brought by different CpG site density or different gene sizes, we perform a permutation test on identified modules where we shuffle the tumor/adjacent normal labels within tumor/adjacent normal pairs and repeat the permutation procedures 100 times. We then calculate the permutated module scores S_{perm} for the identified top-ranked modules. The permutation P-values of the identified top-ranked modules are calculated as follows:

$$P_{perm} = \frac{\#\{S_{perm} > S\}}{\#\{total\ permutation\}},$$

where S is the observed module score.

Simulation studies

We conducted simulation studies to investigate the performance of the proposed NEpiC algorithm that uses the biological network information and both mean and variance signals in DNA methylation data, and compared the performance of NEpiC with that of the following three methods: (i) NEpi method that uses only mean signals of DNA methylation data and also incorporates the biological network information in searching for dense modules, which is an extension of dmGWAS (13); (ii) EpiC method that uses combined signals in both mean and variance of DNA methylation data but not the biological network information; and (iii) Epi method that uses only mean signals of DNA methylation data and not the biological network information.

Simulation settings

PPI network. We used the PPI network from the Protein Interaction Network Analysis platform (PINA) (24) and only maintained edges established with experimental evidence. We also excluded *UBC* gene from the PPI network which has a degree of connection of 9112, far greater than the rest of the genes on the network. The final PINA PPI network has 13 932 genes and 131 778 edges. We further trimmed the PPI network to keep genes that are also in the CUMC HCC methylation data after quality control steps (see **Real Data Applications**). We ended up with 12 630 genes and 116 772 edges.

Simulating site-level signal scores. We simulated site-level mean and variance signal scores based on real data from the CUMC HCC study where there are 195 259 CpG sites from 12 630 genes for 66 matched tumor and adjacent normal pairs after quality control steps which overlap with the genes in PINA PPI network.

To assign genes out of the total 12 630 genes on the PPI network as outcome-associated genes, we first chose the 25 genes that were considered as driver genes in a recent review paper on HCC (25) as the outcome-associated genes. We then randomly selected genes that are the first-order neighbors of the 25 driver genes as the outcome-associated genes with a probability 0.02. We considered the 25 driver genes plus the selected first-order outcome-associated genes as the 'seed associated genes', and randomly selected genes that are the first-order neighbors of those 'seed associated genes' as outcome-associated genes with a probability 0.02. We chose a probability of 0.02 such that after repeating this procedure 5–6 times we will have around 500 outcome-associated genes. This is because there are currently 572 mutated genes that have been causally implicated with cancers according to the Cancer Gene Census category (CGC, as of December 2015) (26). The rest around 12 130 genes are then set as not associated with the outcome.

To assign effect sizes to the outcome-associated genes simulated above, we first defined three groups of genes with different effect sizes, genes with strong, median, and weak effects, where the effect sizes are determined based on the results from the CUMC HCC data. Within each gene category, the mean and variance signal scores were then simulated with a bivariate normal distribution with means, variances and correlations mimicking results from the CUMC HCC data. We set the 25 driver genes to have strong effects. We randomly set half of the selected outcome-associated genes excluding the 25 driver genes to have median effects and the other half to have weak effects. Within each effect size group, we divided them into three subgroups. Using the gene group with a strong effect as the example, the three subgroups are: genes whose mean and variance signal scores are correlated; genes whose mean and variance signal scores are independent when mean signal scores have strong effects and variance signal scores have no effects; and genes whose mean and variance signal scores are independent when mean signal scores have no effects and variance signal scores have strong effects. The ratio of number of genes in these three subgroups is 2:4:1. See supplementary materials for more details of the simulation settings.

RESULTS

Simulation results

We simulated 100 datasets and applied the proposed NEpiC algorithm and the three compared methods, NEpi, EpiC,

and Epi. For NEpiC and NEpi, we chose all genes in the top 1% of modules as candidates. For EpiC and Epi methods that do not use network information, we chose the same number of top-ranked genes as that in the top 1% of modules selected by NEpiC based on gene-level combined signal scores (EpiC method) or gene-level mean signal scores (Epi method). Part I in Table 1 shows the average number of candidate genes identified and the average number (percentage) of truly associated genes out of the candidate genes identified using the four methods and the enrichment *P*-value of the truly associated genes among candidate genes identified. The enrichment *P*-value was calculated using a hypergeometric distribution hyper(q, M, N, k), where q is the number of truly associated genes among identified candidate genes, M is the number of truly associated genes, N is the number of true null genes and k is the number of identified candidate genes. It shows that the proposed NEpiC algorithm identifies more truly associated genes with a higher percentage of truly associated genes out of the candidate genes identified and achieves a more significant enrichment *P*-value than the other three methods. The fact that the NEpiC algorithm outperforms the NEpi algorithm that uses the network information but not the variance signal in DNA methylation data and the EpiC algorithm that uses the variance signal in DNA methylation data but not the network information indicates the benefit of incorporating variance signals and the benefit of incorporating the network information, respectively. Part I of Table 1 also displays the average numbers of truly associated genes identified broken down by effect size categories. It is clear that the proposed NEpiC algorithm is more powerful in identifying genes with median or weak effects as expected.

We further chose the candidate genes that were selected by more than one module among the top 1% of modules as the prioritized candidate genes using the NEpiC or NEpi algorithm. For the EpiC and Epi algorithms, we then chose the number of top ranked candidate genes as the same number of prioritized candidate genes identified using the proposed NEpiC algorithm (Table 1 Part II). Part II of Table 1 also displays the average number (percentage) of truly associated genes out of the prioritized candidate genes identified. There are 31 prioritized candidate genes on average using the proposed NEpiC algorithm (there are 30 prioritized candidate genes on average using the NEpi algorithm). The enrichment *P*-values of the truly associated genes among the prioritized candidate genes are more significant than that among the candidate genes for both the NEpiC and NEpi algorithms (Table 1). Moreover, the percentage of truly associated genes out of the prioritized candidate genes is higher than that out of the candidate genes for both the NEpiC and NEpi algorithms (Table 1). These suggest that the prioritization procedure using the selection frequencies might be a useful step to further improve the performance of the proposed NEpiC algorithm to identify truly outcome-associated genes.

Real data applications

We applied the proposed NEpiC algorithm to the TCGA BRCA, TCGA LIHC and CUMC HCC DNA methylation datasets. We removed probes with missing data in more than 70% of the samples and probes on the sex chromosomes and probes that are known SNP sites. We also removed probes with no gene annotations and required CpG coverage to be at least 95% per sample. Finally, we used Bioconductor package wateRmelon to correct for the type II probe bias (27).

TCGA BRCA data

After quality control steps, there are 229 655 CpG sites from 19 270 genes for 90 matched tumor and adjacent normal pairs in the TCGA BRCA data. Of those, 12 561 genes are also in the PPI network, which contains 115 928 edges. We then applied the proposed NEpiC algorithm using the PINA PPI network to the TCGA BRCA DNA methylation data.

Figure 2 shows the module scores of modules of different sizes before and after normalization using the proposed NEpiC algorithm. The module scores before normalization increase with module sizes as expected while the module scores after normalization are comparable. With the normalized module scores, we chose genes in the top 1% of modules as the candidate genes using NEpiC and NEpi, where there are 227 and 161 genes, respectively (Table 2 Part I). All the top 1% of modules identified by NEpiC and NEpi with the TCGA BRCA data have permutation *P*-values smaller than 0.0005. We then chose the top 227 genes with the strongest combined or mean signals using EpiC and Epi. Among the candidate genes identified by NEpiC, NEpi, EpiC and Epi, there are 16, 2, 11 and 1 genes that have been reported to be differentially methylated in cancers according to a cancer methylation database developed by combining text-mining and expert annotation (Pubmeth) (28) (Table 2 Part I). According to CGC (26), there are 26, 12, 10 and 11 genes that have been causally implicated with cancers, respectively (Table 2 Part I). NEpiC has the highest percentages of reported differentially methylated genes in cancers based on Pubmeth and causally implicated cancer genes based on CGC out of the candidate genes identified. This suggests that the proposed NEpiC algorithm that uses both the biological network information and the mean and variance signals is a more powerful method in identifying potentially cancer-related genes than the methods that ignore either the biological network information or variance signals in DNA methylation data. Between the list of 16 genes reported to be differentially methylated in cancers based on Pubmeth and the list of 26 genes reported to be causally implicated in cancers based on CGC, there are four genes in common. Although the number of genes in common is not large, we found that among the 22 causally implicated cancer genes identified by the CGC database only, nine genes were also reported to be aberrantly methylated in different cancers (5 in breast cancer and four in other cancers, all were published after the Pubmeth database was generated), seven genes were reported to be aberrantly expressed or mutated in breast cancer, and the remaining 6 genes were also reported to be aberrantly expressed or mutated in other cancers (Additional Table 1 for BRCA, Part I). We then conducted pathway enrichment analyses using WebGestalt (29), where we used the enrichment *P*-value of the 'pathway in cancer' from KEGG as the benchmark

Table 1. Average numbers of candidate genes (Part I) and *prioritized* candidate genes (Part II) identified and truly associated genes among the candidate genes (Part I) and *prioritized* candidate genes (Part II) identified by NEpiC, NEpi, EpiC and Epi algorithms and the enrichment *P*-values based on 100 simulated datasets

	NEpiC	NEpi	EpiC	Epi
Part I				
k = Number of candidate genes identified	151	149	151	151
q = Number of candidate genes that are truly associated (combining strong, median or weak effects) (%[1])	21 (13.9%)	15 (10.1%)	15 (9.9%)	11 (7.3%)
Number of candidate genes that are truly associated with strong effect	11	9	14	11
Number of candidate genes that are truly associated with median effect	7	4	1	0
Number of candidate genes that are truly associated with weak effect	3	2	0	0
Enrichment *P*-value of truly associated genes among candidate genes[2]	1.37×10^{-8}	6.76×10^{-5}	7.94×10^{-5}	6.89×10^{-3}
Part II				
k = Number of *prioritized* candidate genes identified	31	30	31	31
q = Number of *prioritized* candidate genes that are truly associated (combining strong, median or weak effects) (%[3])	10 (32.3%)	8 (26.7%)	5 (16.1%)	4 (12.9)
Number of *prioritized* candidate genes that are truly associated with strong effect	9	7	5	4
Number of *prioritized* candidate genes that are truly associated with median effect	1	1	0	0
Number of *prioritized* candidate genes that are truly associated with weak effect	0	0	0	0
Enrichment *P*-value of truly associated genes among *prioritized* candidate genes[4]	3.92×10^{-9}	4.87×10^{-6}	1.22×10^{-3}	4.15×10^{-3}

[1]Percent truly associated genes out of the candidate genes identified.
[2]Enrichment *P*-value was calculated using the hypergeometric distribution: hyper(q, M, N, k), where q is the average number of truly associated genes among candidate genes identified, M is the average number of simulated truly associated genes ($M = 443$), N is the average number of simulated truly non-associated genes ($N = 12\,187$), and k is the average number of candidate genes in the top 1% of modules.
[3]Percent truly associated genes out of the *prioritized* candidate genes identified.
[4]Enrichment *P*-value was calculated using the hypergeometric distribution: hyper(q, M, N, k), where q is the number of truly associated genes among *prioritized* candidate genes identified, M is the average number of simulated truly associated genes ($M = 456$), N is the average number of simulated truly non-associated genes ($N = 12\,174$), and k is the average number of *prioritized* candidate genes in the top 1% of modules.

(Table 2 Part I). It shows that the proposed NEpiC algorithm achieves the most significant enrichment *P*-value while NEpi achieves the second most significant enrichment *P*-value.

We also compared the prioritized candidate genes that appear in more than one module selected. There are 68 prioritized candidate genes selected by more than one of the top 1% of modules identified by the proposed NEpiC algorithm, while there are 27 using the NEpi algorithm (Table 2 Part II). We then chose the top 68 genes with the strongest combined or mean signals using EpiC and Epi. According to Pubmeth, there are 8, 1, 2, 1 genes reported to be differently methylated in cancers out of the prioritized candidate genes identified using NEpiC, NEpi, EpiC and Epi, respectively. There are 14, 3, 2, 3 genes that have been causally implicated with cancers out of the prioritized candidate genes identified using NEpiC, NEpi, EpiC and Epi according to CGC, respectively (Table 2 Part II). The proposed NEpiC algorithm generates the highest percentages of reported differentially methylated genes in cancers and causally implicated cancer genes out of the prioritized candidate genes identified. Moreover, the percentages of differentially methylated genes in cancers according to Pubmeth and causally implicated cancer genes according to CGC out of the prioritized candidate genes are higher than that out of the candidate genes for both the NEpiC and NEpi algorithms, which agrees with the simulation results and suggests that the prioritization procedure using selection frequencies to further prioritize candidate genes might be a useful step for further selecting outcome-related genes. Between the list of eight genes reported to be differentially methylated in cancers according to Pubmeth and the list of 14 genes reported to be causally implicated in cancers according to CGC, there are two genes in common. Among the 12 causally implicated cancer genes identified by the CGC database only, five genes were also reported to be aberrantly methylated (two in breast cancers and three in other cancers, all were published after the Pubmeth database was generated), five genes were reported to be aberrantly expressed or mutated in breast cancer, and the remaining two genes were also reported to be aberrantly expressed or mutated in other cancers (Additional Table 1 for BRCA, Part II). We then conducted a gene set enrichment analysis of 'pathway in cancer' from KEGG among the prioritized candidate genes (Table 2 Part II). It shows that the proposed NEpiC algorithm achieves the most significant enrichment *P*-value while NEpi achieves the second most significant enrichment *P*-value. 'Pathway in cancer' was not enriched among candidate genes identified by EpiC and Epi algorithms.

We display the histogram of gene-level combined signal scores of the candidate genes in the top 1% of modules identified using NEpiC with the TCGA BRCA dataset in Supplementary Figure S3. About a quarter of the candidate genes identified have very high gene-level combined signal scores while there are also some candidate genes with rather small gene-level combined signal scores but which are connected to genes with high gene-level combined signal scores.

We further show in Figure 3 the original methylation measures of the CpG site with the largest site-level com-

Table 2. Number of candidate genes (Part I) and *prioritized* candidate genes (Part II) identified in the TCGA BRCA data and number of reported differentially methylated genes in cancers and causally implicated cancer genes out of the candidate genes (Part I) and out of the *prioritized* candidate genes (Part II) identified according to PubMeth[1] and CGC[2] and enrichment *P*-value of 'pathways in cancer' from KEGG among the candidate genes (Part I) and the *prioritized* candidate genes (Part II) identified

	NEpiC	NEpi	EpiC	Epi
Part I				
Number of candidate genes identified	227	161	227	227
Number (percentage) of reported differentially methylated genes in cancers out of the candidate genes according to Pubmeth (%[3])	16 (7.0%)	2 (1.2%)	11 (4.8%)	1 (0.4%)
Number (percentage) of causal implicated cancer out of the candidate genes according to CGC (%[4])	26 (11.5%)	12 (7.5%)	10 (4.4%)	11 (4.8%)
Enrichment *P*-value of 'pathway in cancer' among the candidate genes identified	1.96×10^{-13}	2.00×10^{-4}	NS[5]	5.00×10^{-3}
Part II				
Number of *prioritized* candidate genes identified	68	27	68	68
Number (percentage) of reported differentially methylated genes in cancers out of the *prioritized* candidate genes according to Pubmeth (%[6])	8 (11.8%)	1 (3.7%)	2 (2.9%)	1 (1.5%)
Number (percentage) of causally implicated cancer genes out of the *prioritized* candidate genes according to CGC (%[7])	14 (20.6%)	3 (11.1%)	2 (2.9%)	3 (4.4%)
Enrichment *P*-value of 'pathway in cancer' among the *prioritized* candidate genes identified	1.42×10^{-13}	4.90×10^{-5}	NS	NS

[1]There are 292 reported differentially methylated genes in cancers according to Pubmeth (28).
[2]There are 572 mutated genes that have been causally implicated with cancers according to the Cancer Gene Census category (CGC, as of December 2015) (26).
[3]Percent Pubmeth genes out of the candidate genes identified.
[4]Percent CGC genes out of the candidate genes identified.
[5]NS stands for not significant.
[6]Percent Pubmeth genes out of the *prioritized* candidate genes identified.
[7]Percent CGC genes out of the *prioritized* candidate genes identified.

bined signal score (top row) and the site-level combined scores of all sites (bottom row) in three selected genes from the top first-ranked module identified by NEpiC with the TCGA BRCA data. We note that (i) tumor samples in general have bigger variation in methylation measures and this difference in variation is not driven by a few outliers; (ii) genes with large gene-level combined signal scores usually have a high proportion of sites with large site-level combined signal scores. Similar plots for the TCGA LIHC and the CUMC HCC data are shown in Supplementary Figures S1 and S2.

Among genes uniquely identified by NEpiC but not the other three methods in the TCGA BRCA dataset, we examined *ZNF652* for illustration purpose. *ZNF652* has a gene-level combined signal score of 9.02 and ranked #6,799 by gene-level combined signal scores. Thus, *ZNF652* was not selected by EpiC. Moreover, *ZNF652* is in the identified module using NEpiC with six other genes: *CPLX2, NEUROG1, GFI1, PRDM14, ETS1 and RUNX1T1* (Figure 4), which ranked #7, #97, #251, #263, #360, and #665 by gene-level combined signal scores, respectively. *CPLX2* and *NEUROG1* were reported to be aberrantly methylated in cancers (30,31). On the other hand, using NEpi with mean signal scores, the highest ranked module with *ZNF652* does not make to the top 1% of modules, thus was not selected by NEpi. This module has six genes: *NEDD9, PIK3CA, ZBTB47, TCF3, CBFA2T3 and ZNF652*, with the mean signal scores ranked #38, #167, #313, #1465, #2848 and #3651. Thus, *ZNF652* was not selected by Epi either. Note that *ZNF652* was previously reported to be associated with breast cancer (32).

TCGA LIHC data and CUMC HCC data

After the same quality control steps, there are 229 700 CpG sites from 19,257 genes for 50 matched tumor and adjacent normal pairs in the TCGA LIHC data. Of those, 12 565 genes are also in the PPI network, which contains 115 964 edges. We then applied the proposed NEpiC algorithm using the PINA PPI network to the TCGA LIHC DNA methylation data.

Table 3 Part I displays the number of candidate genes identified in the TCGA LIHC data and the numbers (percentages) of reported differentially methylated genes in cancers according to Pubmeth (28) and causally implicated cancer genes according to CGC (26) out of the candidate genes identified. The proposed NEpiC algorithm clearly outperforms the other methods ignoring either biological network information or variance signals in DNA methylation. Between the list of 11 genes reported to be differentially methylated in cancers based on Pubmeth and the list of 18 genes reported to be causally implicated in cancers based on CGC, there are 6 genes in common. Among the 12 causally implicated cancer genes identified by the CGC database only, two genes were reported to be aberrantly methylated in cancers other than liver cancer (both were published after the Pubmed database was generated), seven genes were reported to be aberrantly expressed in liver cancer, and the remaining 3 genes were reported to be aberrantly expressed or mutated in other cancers. (Additional Table 2 for LIHC, Section I, Part I). In the pathway enrichment analysis, we selected eight core liver cancer pathways (p53 signaling pathway, cell cycle regulation pathway, TERT pathway, WNT pathway, chromatin modifying factors, growth factor signaling pathway, KEAP1–NFE2L2

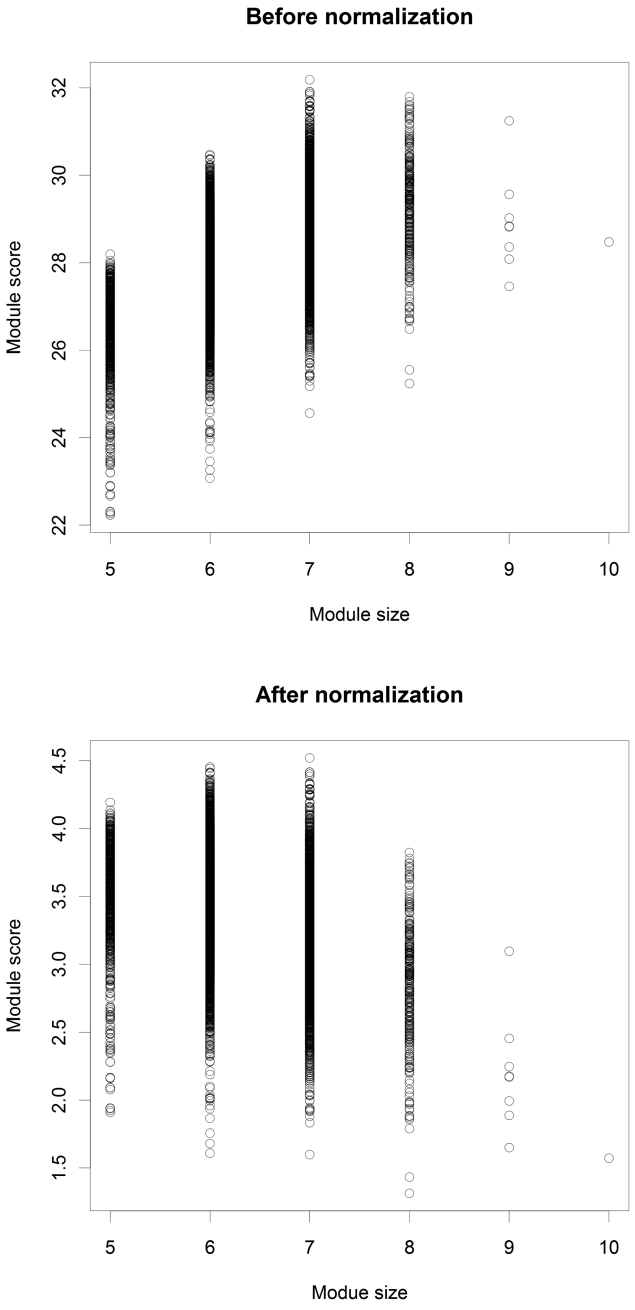

Figure 2. Module scores before and after normalization using the proposed NEpiC algorithm with the TCGA BRCA data.

pathway and NOTCH pathway) (25) and used the 'pathway in cancer' from KEGG as the benchmark to evaluate the performance of the four methods. The 'pathway in cancer', the 'p53 signaling pathway' and the 'WNT signaling pathway' are enriched among the candidate genes identified using the proposed NEpiC algorithm while only the 'pathway in cancer' is enriched among the candidate genes identified using NEpi (Table 3 Part I). All the top 1% of modules identified by NEpiC and NEpi with the TCGA LIHC data have permutation *P*-values smaller than 0.0005.

We then compared the prioritized candidate genes identified using the four methods (Table 3 Part II). There are 64 candidate genes selected by more than one of the top 1% of modules identified using NEpiC, while there are 37 using NEpi. We then chose the top 64 genes with the strongest combined or mean signals using EpiC and Epi algorithms. The percentages of reported differentially methylated genes in cancers according to Pubmeth and causally implicated cancer genes according to CGC among the prioritized candidate genes are higher than that among candidate genes before the prioritization procedure. This again suggests that the prioritization procedure may improve the performance of the proposed NEpiC algorithm. Between the list of five genes reported to be differentially methylated in cancers according to Pubmeth and the list of 10 genes reported to be causally implicated cancer genes according to CGC, there are three genes in common. Among the seven causally implicated cancer genes identified by the CGC database only, one gene was reported to be aberrantly methylated in cancer other than liver cancer (it was published after the Pubmeth database was generated), four genes were reported to be aberrantly expressed in liver cancer, and the remaining two genes were reported to be aberrantly expressed or mutated in other cancers (Additional Table 2 for LIHC, Section I, Part II). In the pathway enrichment analysis of the eight core liver cancer pathways and the 'pathway in cancer' from KEGG, three pathways ('pathway in cancer', 'p53 signaling pathway' and 'WNT signaling pathway') are enriched among the prioritized candidate genes identified by the proposed NEpiC algorithm, while only one pathway ('pathway in cancer') is enriched among the prioritized candidate genes identified by the NEpi algorithm (Table 3 Part II). No pathway is enriched among the prioritized candidate genes identified by NEpi and Epi algorithms.

We further performed a replication analysis using the CUMC HCC data and investigated the replication results using the TCGA LIHC data and the CUMC HCC data. Similar results as in Table 3 but using the CUMC HCC data are listed in Table S1 in the supplementary file. For replication analyses, we first define replicated candidate genes (replicated prioritized candidate genes) as the overlapping candidate genes (prioritized candidate genes) between the two lists of candidate genes (prioritized candidate genes) identified by the same method applied to the two HCC datasets. We then examined the reported differentially methylated genes in cancers according to Pubmeth and causally implicated cancer genes according to CGC out of the replicated candidate genes (replicated prioritized candidate genes) identified and the enrichment *P*-values of the replicated candidate genes (replicated prioritized candidate genes) in the eight liver cancer core pathways (25) and the 'pathway in cancer' from KEGG (Table 4). The proposed NEpiC algorithm generates the highest percentages of reported differentially methylated genes in cancers according to Pubmeth and causally implicated cancer genes according to CGC out of the replicated candidate genes and replicated candidate genes identified by the proposed NEpiC algorithm are enriched in two liver cancer core pathways and the 'pathway in cancer' (Table 4 Part I). Between the list of three genes reported to be differentially methylated in cancers according to Pubmeth and the list of five causally

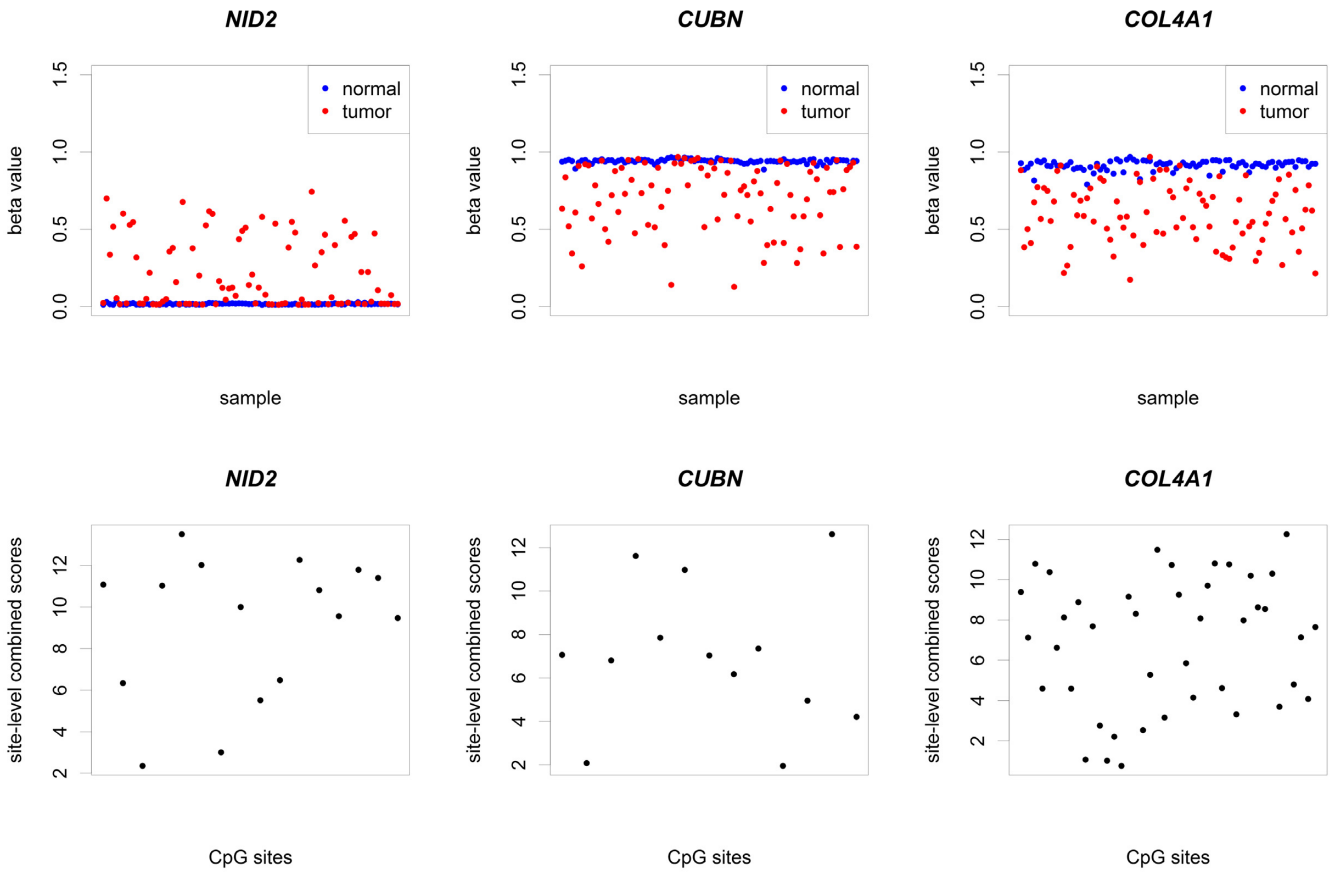

Figure 3. Original methylation measures of the CpG site with the largest site-level combined signal score (top row) and site-level combined signal scores of all sites (bottom row) in three genes in the top first-ranked module identified with the TCGA BRCA data.

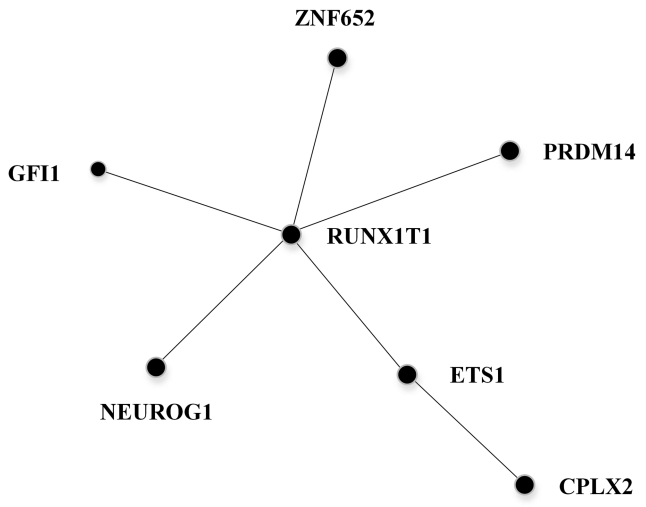

Figure 4. An example of identified module with *ZNF652* uniquely identified by the NEpiC algorithm.

implicated genes according to CGC, there are two genes in common. Among the three causally implicated cancer genes identified by the CGC database only, one gene was also reported to be differentially methylated in cancer other than liver cancer (it was published after the Pubmeth database was generated), one gene was reported to be aberrantly expressed in liver cancer, and the last gene was reported to be aberrantly expressed in other cancer (Additional Table 2 for LIHC, Section II, Part I). For the replicated prioritized candidate genes, the proposed NEpiC algorithm also outperforms the other three methods using both percentages of causally implicated cancer genes according to CGC and enrichment *P*-values of eight liver cancer core pathways (25) and the 'pathway in cancer' from KEGG as the criteria (Table 4 Part II). However, the percentage of reported differentially methylated genes in cancers according to Pubmeth for NEpiC is lower than that for EpiC. This may due to the fact that the Pubmeth database is not most up-to-date. We also repeated the replication analyses with different cutoffs to select top ranked modules, where we chose candidate genes as the genes in the top 1.5%, 2%, 2.5% and 3% of modules (Supplementary Tables S2–S5). Under all these scenarios, NEpiC outperforms the other three methods which ignore either the biological network information or variance signals in both candidate genes and prioritized candidate genes using all three comparison criteria.

A further investigation of the 42 replicated candidate genes from the top 1% of modules identified using the proposed NEpiC algorithm suggests that 33 replicated candidate genes out of the 42 replicated candidate genes have been reported to be associated with cancer. These includes genes are reported to be related with liver cancer: *CDKN2A*

Table 3. Number of candidate genes (Part I) and *prioritized* candidate genes (Part II) identified in the TCGA LIHC data and number of reported differentially methylated genes in cancers and causally implicated cancer genes out of the candidate genes (Part I) and out of the *prioritized* candidate genes (Part II) identified according to PubMeth[1] and CGC[2] and Bonferroni adjusted enrichment *P*-values of enriched core pathways in liver cancer and 'pathway in cancer' from KEGG among the candidate genes (Part I) and the *prioritized* candidate genes (Part II) identified

	NEpiC	NEpi	EpiC	Epi
Part I				
Number of candidate genes identified	201	148	201	201
Number (percentage) of reported differentially methylated genes in cancers out of the candidate genes according to Pubmeth (%[3])	11 (5.5%)	4 (2.7%)	8 (3.9%)	2 (0.1%)
Number (percentage) of causally implicated cancer out of the candidate genes according to CGC (%[4])	18 (9.0%)	4 (2.7%)	10 (4.9%)	1 (0.5%)
Enrichment *P*-values[5]				
Pathway in cancer	4.38×10^{-10}	1.78×10^{-7}	0.032	NS[6]
p53 signaling pathway	0.036	NS	NS	NS
WNT signaling pathway	6.49×10^{-5}	NS	NS	NS
Part II				
Number of *prioritized* candidate genes identified	64	37	64	64
Number (percentage) of reported differentially methylated genes in cancers out of the *prioritized* candidate genes according to Pubmeth (%[7])	5 (7.8%)	3 (8.1%)	3 (4.7%)	0 (0.0%)
Number (percentage) of causally implicated cancer out of the *prioritized* candidate genes according to CGC (%[8])	10 (15.6%)	3 (8.1%)	4 (6.3%)	0 (0.0%)
Enrichment *P*-values[9]				
Pathway in cancer	1.24×10^{-8}	1.80×10^{-3}	NS	NS
p53 signaling pathway	9.00×10^{-4}	NS	NS	NS
WNT signaling pathway	9.18×10^{-7}	NS	NS	NS

[1]There are 292 reported differentially methylated genes in cancers according to Pubmeth (28).
[2]There are 572 mutated genes that have been causally implicated with cancers according to the Cancer Gene Census category (CGC, as of December 2015) (26).
[3]Percent Pubmeth genes out of the candidate genes identified.
[4]Percent CGC genes out of the candidate genes identified.
[5]Enrichment *P*-values of significant core liver cancer pathways and 'pathway in cancer' from KEGG among the candidate genes identified were Bonferroni corrected with the number of compared pathways from KEGG.
[6]NS stands for not significant.
[7]Percent Pubmeth genes out of the *prioritized* candidate genes identified.
[8]Percent CGC genes out of the *prioritized* candidate genes identified.
[9]Enrichment *P*-values of significant core liver cancer pathways and 'pathway in cancer' from KEGG among the *prioritized* candidate genes identified were Bonferroni corrected with the number of compared pathways from KEGG.

(33), *GRASP* (34), *DLGAP1* (35), *LPAR2* (35), *STEAP4* (36), *WNT3A* (37), *TSC22D1* (38), *NKD2* (39), *TGFA* (40), *TERT* (41), *HSP90AA1* (42) and *IGF1R* (43); genes that are known to be aberrantly methylated in cancers other than liver cancer: *MYO10* (44), *BAI1* (45), *FYN* (46), *ACTA1* (47), *SPRR2A* (48) and *CARD11* (49); and genes that are associated with cancers other than liver cancer: *VIM* (50), *CTBP2* (51), *PRKCQ* (52), *PDZD2* (53), *DSCAML1* (54), *KCNQ1* (55), *KCNQ2* (56), *KCNQ3* (57), *OBSCN* (58), *FSCN1* (59), *DLGAP2* (60), *CFTR* (61), *RUNX1T1* (62), *GRID2* (63) and *SCN5A* (64). The full list of 42 replicated candidate genes is included in the supplementary materials.

DISCUSSION

In this article, we proposed the NEpiC algorithm, a network assisted method incorporating combined signals in mean and variance differences of DNA methylation data. We demonstrated that incorporating prior biological network information and utilizing the signals in variance differences of DNA methylation data could effectively improve the power of the association studies to identify aberrant methylated genes associated with the outcomes.

We demonstrated a much improved power of the proposed NEpiC algorithm that incorporates both biological network information and variance signals in DNA methylation data than the methods that do not. In simulation studies, the proposed NEpiC algorithm identifies most truly associated genes among the candidate genes identified and achieves the most significant enrichment *P*-value of truly associated genes among candidate genes identified. Using the prioritized genes that were selected in more than one of the top 1% of modules further improves the performance of the proposed NEpiC algorithm in both simulation studies and real data applications. The application to two independent liver cancer datasets, the TCGA LIHC data and the CUMC HCC data, gives us the opportunity to examine replication results. The replication results show that the proposed NEpiC algorithm identifies more genes that were already reported to be differentially methylated in cancers according to Pubmeth and causally implicated cancer genes according to CGC and identifies genes that are more enriched in known liver cancer pathways than methods that do not use both biological network information and variance signals in DNA methylation data.

Although we focused on applying the proposed NEpiC algorithm to cancer patients with tumor and adjacent normal tissues to identify genes that are related to tumor status, several publications (65,66) have shown that differential variability is most informative and meaningful in comparing precursor cancer lesions to normal cells. That is, the de-

Table 4. Number of replicated candidate genes (Part I) and replicated *prioritized* candidate genes (Part II) identified in the TCGA LIHC data and the CUMC HCC data and number of reported differentially methylated genes in cancers and causally implicated cancer genes out of the replicated candidate genes (Part I) and out of the replicated *prioritized* candidate genes (Part II) identified according to PubMeth[1] and CGC[2]and Bonferroni adjusted enrichment *P*-values of enriched core pathways in liver cancer and the 'pathway in cancer' from KEGG among the replicated candidate genes (Part I) and the replicated *prioritized* candidate genes (Part II) identified

	NEpiC	NEpi	EpiC	Epi
Part I				
Number of replicated candidate[3] genes identified	42	26	89	77
Number (percentage) of reported differentially methylated genes in cancers out of the replicated candidate genes according to Pubmeth (%[4])	3 (7.1%)	0 (0%)	3 (3.4%)	0 (0%)
Number (percentage) of causally implicated cancer out of the replicated candidate genes according to CGC (%[5])	5 (11.9%)	0 (0%)	5 (5.6%)	0 (0%)
Enrichment *P*-values[6]				
Pathway in cancer	2.56×10^{-7}	NS[7]	3.17×10^{-4}	NS
p53 signaling pathway	0.018	NS	NS	NS
WNT signaling pathway	3.60×10^{-3}	NS	0.028	NS
Part II				
Number of replicated *prioritized* candidate genes identified	20	12	24	23
Number (percentage) of reported differentially methylated genes in cancers out of the replicated *prioritized* candidate genes according to Pubmeth (%[8])	0 (0%)	0 (0%)	2 (8.3%)	0 (0%)
Number (percentage) of causally implicated cancer out of the replicated *prioritized* candidate genes according to CGC (%[9])	1 (5.0%)	0 (0%)	1 (4.2%)	0 (0.0%)
Enrichment *P*-values[10]				
p53 signaling pathway	0.031	NS	NS	NS

[1]There are 292 reported differentially methylated genes in cancers according to Pubmeth (28).
[2]There are 572 mutated genes that have been causally implicated with cancers according to the Cancer Gene Census category (CGC, as of December 2015) (26).
[3]Replicated candidate genes are defined as the candidate genes identified in both the TCGA LIHC data and the CUMC HCC data.
[4]Percent Pubmeth genes out of the replicated candidate genes identified.
[5]Percent CGC genes out of the replicated candidate genes identified.
[6]Enrichment *P*-values of significant core liver cancer pathways and the 'pathway in cancer' from KEGG among the replicated candidate genes identified were Bonferroni corrected with the number of compared pathways from KEGG.
[7]NS stands for not significant.
[8]Percent Pubmeth genes out of the replicated *prioritized* candidate genes identified.
[9]Percent CGC genes out of the replicated *prioritized* candidate genes identified.
[10]Enrichment *P*-values of significant core liver cancer pathways and 'pathway in cancer' from KEGG among the replicated *prioritized* candidate genes identified were Bonferroni corrected with the number of compared pathways from KEGG.

veloped NEpiC algorithm may be the most useful in cancer early detection.

Since bigger variation in methylation measures is usually observed in tumor tissues compared to adjacent tissues, tumor tissue purity may thus influence findings from methods that use variance signals. A method was recently developed to check the purity of tumor tissues using DNA methylation 450K arrays (67), which could be applied in the quality control steps to screen out tumor tissues with low purity.

In summary, we developed a new algorithm, NEpiC that incorporates biological network information and utilizes variance signals in DNA methylation data in detecting differentially methylated genes. Results from both simulations and real data applications demonstrate a much better performance of the NEpiC algorithm compared to several other methods that ignore either the biological network information or variance signals in DNA methylation data. The NEpiC algorithm is implemented in an R package which is freely available through CRAN.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

National Natural Science Foundation of China [61272380 to P.R. and S. Z.]; National Institutes of Health [R01ES05116 and P30ES009089 to S.J., R.S. and S.W.]. Funding for open access charge: Departmental fund from Department of Biostatistics, Columbia University.
Conflict of interest statement. None declared.

REFERENCES

1. Tischoff,I. and Tannapfel,A. (2008). DNA methylation in hepatocellular carcinoma. *World J. Gastroenterol.: WJG.*, **14**, 1741–1748.
2. Sato,Y., Yoshizato,T., Shiraishi,Y., Maekawa,S., Okuno,Y., Kamura,T., Shimamura,T., Sato-Otsubo,A., Nagae,G., Suzuki,H. *et al.* (2013). Integrated molecular analysis of clear-cell renal cell carcinoma. *Nat. Genet.*, **45**, 860–867.
3. Toyota,M., Ahuja,N., Ohe-Toyota,M., Herman,J.G., Baylin,S.B. and Issa,J.P.J. (1999). CpG island methylator phenotype in colorectal cancer. *Proc. Natl. Acad. Sci. U.S.A.*, **96**, 8681–8686.
4. Tao,M.H. and Freudenheim,J.L. (2010). DNA methylation in endometrial cancer. *Epigenetics*, **5**, 491–498
5. Eden,A., Gaudet,F., Waghmare,A. and Jaenisch,R. (2003). Chromosomal instability and tumors promoted by DNA hypomethylation. *Science*, **300**, 455–455.
6. Ruike,Y., Imanaka,Y., Sato,F., Shimizu,K. and Tsujimoto,G. (2010). Genome-wide analysis of aberrant methylation in human breast cancer cells using methyl-DNA immunoprecipitation combined with high-throughput sequencing. *BMC Genomics*, **11**, 137.

7. Ammerpohl,O., Pratschke,J., Schafmayer,C., Haake,A., Faber,W., von Kampen,O., Balschun,K., Rocken,C., Arlt,A., Schniewind,B. *et al.* (2012). Distinct DNA methylation patterns in cirrhotic liver and hepatocellular carcinoma. *Int. J. Cancer*, **130**, 1319–1328.
8. Shen,J., Wang,S., Zhang,Y.J., Wu,H.C., Kibriya,M.G., Jasmine,F., Ahsan,H., Wu,D.P., Siegel,A.B., Remotti,H. *et al.* (2013). Exploring genome-wide DNA methylation profiles altered in hepatocellular carcinoma using Infinium HumanMethylation 450 BeadChips. *Epigenetics*, **8**, 34–43.
9. Sun,H. and Wang,S. (2012). Penalized logistic regression for high-dimensional DNA methylation data with case–control studies. *Bioinformatics*, **28**, 1368–1375.
10. Sun,H. and Wang,S. (2013). Network-based regularization for matched case-control analysis of high-dimensional DNA methylation data. *Stat. Med.*, **32**, 2127–2139.
11. Wang,K., Li,M. and Bucan,M. (2007). Pathway-based approaches for analysis of genomewide association studies. *Am. J. Hum. Genet.*, **81**, 1278–1283.
12. Chen,M., Cho,J. and Zhao,H. (2011). Incorporating biological pathways via a Markov random field model in genome-wide association studies. *PLoS Genet.*, **7**, e1001353.
13. Jia,P., Zheng,S., Long,J., Zheng,W. and Zhao,Z. (2011). dmGWAS: dense module searching for genome-wide association studies in protein–protein interaction networks. *Bioinformatics*, **27**, 95–102.
14. Baranzini,S.E., Galwey,N.W., Wang,J., Khankhanian,P., Lindberg,R., Pelletier,D., Wu,W., Uitdehaag,B., Kappos,L., Polman,C.H. *et al.* (2009). Pathway and network-based analysis of genome-wide association studies in multiple sclerosis. *Hum. Mol. Genet.*, **18**, 2078–2090.
15. Rossin,E.J., Lage,K., Raychaudhuri,S., Xavier,R.J., Tatar,D., Benita,Y., Cotsapas,C. and Daly,M.J. (2011). Proteins encoded in genomic regions associated with immune-mediated disease physically interact and suggest underlying biology. *PLoS Genet.*, **7**, e1001273.
16. West,J., Beck,S., Wang,X. and Teschendorff,A.E. (2013). An integrative network algorithm identifies age-associated differential methylation interactome hotspots targeting stem-cell differentiation pathways. *Sci. Rep.*, **3**, 1630.
17. Jiao,Y., Widschwendter,M. and Teschendorff,A.E. (2014). A systems-level integrative framework for genome-wide DNA methylation and gene expression data identifies differential gene expression modules under epigenetic control. *Bioinformatics*, **30**, 2360–2366.
18. Hansen,K.D., Timp,W., Bravo,H.C., Sabunciyan,S., Langmead,B., McDonald,O.G., Wen,B., Liu,Y., Diep,D., Briem,E. *et al.* (2011). Increased methylation variation in epigenetic domains across cancer types. *Nat. Genet.*, **43**, 768–775.
19. Teschendorff,A.E. and Widschwendter,M. (2012). Differential variability improves the identification of cancer risk markers in DNA methylation studies profiling precursor cancer lesions. *Bioinformatics*, **28**, 1487–1494.
20. Teschendorff,A.E., Liu,X., Caren,H., Pollard,S.M., Beck,S., Widschwendter,M. and Chen,L. (2014). The dynamics of DNA methylation covariation patterns in carcinogenesis. *PLoS Comput. Biol.*, **10**, e1003709.
21. Chen,Y., Ning,Y., Hong,C. and Wang,S. (2014). Semiparametric Tests for Identifying Differentially Methylated Loci With Case–Control Designs Using Illumina Arrays. *Genet. Epidemiol.*, **38**, 42–50.
22. Morgan,W.A. (1939). A test for the significance of the difference between two variances in a sample from a normal bivariate population. *Biometrika*, **31**, 13–19.
23. Pitman,E.J.G. (1939). A note on normal correlation. *Biometrika*, **31**, 9–12.
24. Wu,J., Vallenius,T., Ovaska,K., Westermarck,J., Mäkelä,T.P. and Hautaniemi,S. (2009). Integrated network analysis platform for protein–protein interactions. *Nat. Methods*, **6**, 75–77.
25. Shibata,T. and Aburatani,H. (2014). Exploration of liver cancer genomes. *Nat. Rev. Gastroenterol. Hepatol.*, **11**, 340–349.
26. Futreal,P.A., Coin,L., Marshall,M., Down,T., Hubbard,T., Wooster,R., Rahman,N. and Stratton,M.R. (2004). A census of human cancer genes. *Nat. Rev. Cancer*, **4**, 177–183.
27. Pidsley,R., Wong,C.C., Volta,M., Lunnon,K., Mill,J. and Schalkwyk,L.C. (2013). A data-driven approach to preprocessing Illumina 450K methylation array data. *BMC Genomics*, **14**, 293.
28. Ongenaert,M., Van Neste,L., De Meyer,T., Menschaert,G., Bekaert,S. and Van Criekinge,W. (2008). PubMeth: a cancer methylation database combining text-mining and expert annotation. *Nucleic Acids Res.*, **36**, D842–D846.
29. Zhang,B., Kirov,S. and Snoddy,J. (2005). WebGestalt: an integrated system for exploring gene sets in various biological contexts. *Nucleic Acids Res.*, **33**, W741–W748.
30. Rodriguez,J., Muñoz,M., Vives,L., Frangou,C.G., Groudine,M. and Peinado,M.A. (2008). Bivalent domains enforce transcriptional memory of DNA methylated genes in cancer cells. *Proc. Natl. Acad. Sci. U.S.A.*, **105**, 19809–19814.
31. Fackler,M.J., Umbricht,C.B., Williams,D., Argani,P., Cruz,L.A., Merino,V.F., Teo,W.W., Zhang,Z., Huang,P., Visvananthan,K. *et al.* (2011). Genome-wide methylation analysis identifies genes specific to breast cancer hormone receptor status and risk of recurrence. *Cancer Res.*, **71**, 6195–6207.
32. Kumar,R., Manning,J., Spendlove,H.E., Kremmidiotis,G., McKirdy,R., Lee,J., Millband,D.N., Cheney,K.M., Stampfer,M.R., Dwivedi,P.P. *et al.* (2006). ZNF652, a novel zinc finger protein, interacts with the putative breast tumor suppressor CBFA2T3 to repress transcription. *Mol. Cancer Res.*, **4**, 655–665..
33. Wong,I.H., Lo,Y.D., Zhang,J., Liew,C.T., Ng,M.H., Wong,N., Lai,P.B., Lau,W.Y., Hjelm,N.M. and Johnson,P.J. (1999). Detection of aberrant p16 methylation in the plasma and serum of liver cancer patients. *Cancer Res.*, **59**, 71–73.
34. Tao,R., Li,J., Xin,J., Wu,J., Guo,J., Zhang,L., Jiang,L., Zhang,W., Yang,Z. and Li,L. (2011). Methylation profile of single hepatocytes derived from hepatitis B virus-related hepatocellular carcinoma. *PLoS One*, **6**, e19862.
35. Song,M.A., Tiirikainen,M., Kwee,S., Okimoto,G., Yu,H. and Wong,L.L. (2013). Elucidating the landscape of aberrant DNA methylation in hepatocellular carcinoma. *PLoS One*, **8**, e55761.
36. Shen,J., Wang,S., Zhang,Y.J., Kappil,M., Wu,H.C., Kibriya,M.G., Wang,Q., Jasmine,F., Ahsan,H., Lee,P.H. *et al.* (2012). Genome-wide DNA methylation profiles in hepatocellular carcinoma. *Hepatology*, **55**, 1799–1808.
37. Lee,H.C., Kim,M. and Wands,J.R. (2006). Wnt/Frizzled signaling in hepatocellular carcinoma. *Front. Biosci.*, **11**, 1901–1915.
38. Saitta,C., Tripodi,G., Barbera,A., Bertuccio,A., Smedile,A., Ciancio,A., Raffa,G., Sangiovanni,A., Navarra,G., Raimondo,G. *et al.* (2015). Hepatitis B virus (HBV) DNA integration in patients with occult HBV infection and hepatocellular carcinoma. *Liver Int.*, **35**. 2311–2317.
39. Liu,S., Cheng,J., Zhang,X., Jiang,S., Liu,X., Li,M., Zhang,J., Li,X., Xu,C., Chen,X. *et al.* (2013). Quantificational methylation analysis of APC and AXIN2 in HBV-related hepatocellular carcinoma. *Curr. Cancer Ther. Rev.*, **9**, 137–146.
40. Zender,L., Villanueva,A., Tovar,V., Sia,D., Chiang,D.Y. and Llovet,J.M. (2010). Cancer gene discovery in hepatocellular carcinoma. *J. Hepatol.*, **52**, 921–929.
41. Lee,H.S., Kim,B.H., Cho,N.Y., Yoo,E.J., Choi,M., Shin,S.H., Jang,J.J., Suh,K.S., Kim,Y.S. and Kang,G.H. (2009). Prognostic implications of and relationship between CpG island hypermethylation and repetitive DNA hypomethylation in hepatocellular carcinoma. *Clin. Cancer Res.*, **15**, 812–820.
42. Kittaka,N., Takemasa,I., Takeda,Y., Marubashi,S., Nagano,H., Umeshita,K., Dono,K., Matsubarak,K., Matsuura,N. and Monden,M. (2008). Molecular mapping of human hepatocellular carcinoma provides deeper biological insight from genomic data. *Eur. J. Cancer*, **44**, 885–897.
43. Hernandez-Vargas,H., Lambert,M.P., Le Calvez-Kelm,F., Gouysse,G., McKay-Chopin,S., Tavtigian,S.V., Scoazec,J.Y. and Herceg,Z. (2010). Hepatocellular carcinoma displays distinct DNA methylation signatures with potential as clinical predictors. *PLoS One*, **5**, e9749.
44. Dunwell,T., Hesson,L., Rauch,T.A., Wang,L., Clark,R.E., Dallol,A., Gentle,D., Catchpoole,D., Eamonn,R., Pfeifer,G.P. *et al.* (2010). A genome-wide screen identifies frequently methylated genes in haematological and epithelial cancers. *Mol. Cancer*, **9**, 44–56.
45. Vincent,A., Omura,N., Hong,S.M., Jaffe,A., Eshleman,J.R. and Goggins,M. (2011). Genome-wide analysis of promoter methylation associated with gene expression profiles of pancreatic adenocarcinomas. *Clin. Cancer Res.*, **17**, 4341–4354.

46. Yamashita,S., Tsujino,Y., Moriguchi,K., Tatematsu,M. and Ushijima,T. (2006). Chemical genomic screening for methylation-silenced genes in gastric cancer cell lines using 5-aza-2′-deoxycytidine treatment and oligonucleotide microarray. *Cancer Sci.*, **97**, 64–71.
47. Vincent,A., Omura,N., Hong,S.M., Jaffe,A., Eshleman,J.R. and Goggins,M. (2011). Genome-wide analysis of promoter methylation associated with gene expression profiles of pancreatic adenocarcinomas. *Clin. Cancer Res.*, **17**, 4341–4354.
48. Leclerc,D., Lévesque,N., Cao,Y., Deng,L., Wu,Q., Powell,J., Sapienza,C. and Rozen,R. (2013). Genes with aberrant expression in murine preneoplastic intestine show epigenetic and expression changes in normal mucosa of colon cancer patients. *Cancer Prev. Res.*, **6**, 1171–1181.
49. Mayol,G., Martín-Subero,J.I., Ríos,J., Queiros,A., Kulis,M., Suñol,M., Esteller,M., Gomez,S., Garcia,I., Torres,C.D. *et al.* (2012). DNA hypomethylation affects cancer-related biological functions and genes relevant in neuroblastoma pathogenesis. *PLoS One*, **7**, e48401.
50. Furuta,M., Kozaki,K.I., Tanaka,S., Arii,S., Imoto,I. and Inazawa,J. (2009). miR-124 and miR-203 are epigenetically silenced tumor-suppressive microRNAs in hepatocellular carcinoma. *Carcinogenesis*, **31**, 766–776.
51. Birts,C.N., Harding,R., Soosaipillai,G., Halder,T., Azim-Araghi,A., Darley,M., Cutress,R., Bateman,A.C. and Blaydes,J.P. (2011). Expression of CtBP family protein isoforms in breast cancer and their role in chemoresistance. *Biol. Cell*, **103**, 1–19.
52. Zhang,H.H., Zhang,Z.Y., Che,C.L., Mei,Y.F. and Shi,Y.Z. (2013). Array analysis for potential biomarker of gemcitabine identification in non-small cell lung cancer cell lines. *Int. J. Clin. Exp. Pathol.*, **6**, 1734.
53. Tam,C.W., Cheng,A.S., Ma,R.Y.M., Yao,K.M. and Shiu,S.Y.W. (2006). Inhibition of prostate cancer cell growth by human secreted PDZ domain-containing protein 2, a potential autocrine prostate tumor suppressor. *Endocrinology*, **147**, 5023–5033.
54. Landi,D., Gemignani,F., Pardini,B., Naccarati,A., Garritano,S., Vodicka,P., Vodickova,L., Canzian,F., Novotny,J., Barale,R. *et al.* (2012). Identification of candidate genes carrying polymorphisms associated with the risk of colorectal cancer by analyzing the colorectal mutome and microRNAome. *Cancer*, **118**, 4670–4680.
55. Than,B.L.N., Goos,J.A.C.M., Sarver,A.L., O'Sullivan,M.G., Rod,A., Starr,T.K., Fijneman,R.J., Zhao,L., Zhang,Y., Largaespade,D.A. *et al.* (2014). The role of KCNQ1 in mouse and human gastrointestinal cancers. *Oncogene*, **33**, 3861–3868.
56. Salyer,S.A., Olberding,J.R., Distler,A.A., Lederer,E.D., Clark,B.J., Delamere,N.A. and Khundmiri,S.J. (2013). Vacuolar ATPase driven potassium transport in highly metastatic breast cancer cells. *Biochim. Biophys. Acta (BBA)-Mol. Basis Dis.*, **1832**, 1734–1743.
57. Blaveri,E., Brewer,J.L., Roydasgupta,R., Fridlyand,J., DeVries,S., Koppie,T., Pajavar,S., Mehta,K., Carroll,P., Simko,J.P. *et al.* (2005). Bladder cancer stage and outcome by array-based comparative genomic hybridization. *Clin. Cancer Res.*, **11**, 7012–7022.
58. Keita,M., Wang,Z.Q., Pelletier,J.F., Bachvarova,M., Plante,M., Gregoire,J., Renaud,M.C., Mes-Masson,A.M., Paquet,E.R. and Bachvarov,D. (2013) Global methylation profiling in serous ovarian cancer is indicative for distinct aberrant DNA methylation signatures associated with tumor aggressiveness and disease progression. *Gynecol. Oncol.*, **128**, 356–363.
59. Chiyomaru,T., Enokida,H., Tatarano,S., Kawahara,K., Uchida,Y., Nishiyama,K., Fujimur,L., Kikkawa,N., Seki,N. and Nakagawa,M. (2010). miR-145 and miR-133a function as tumour suppressors and directly regulate FSCN1 expression in bladder cancer. *Br. J. Cancer*, **102**, 883–891.
60. Luedi,P.P., Dietrich,F.S., Weidman,J.R., Bosko,J.M., Jirtle,R.L. and Hartemink,A.J. (2007). Computational and experimental identification of novel human imprinted genes. *Genome Res.*, **17**, 1723–1730.
61. Xie,C., Jiang,X.H., Zhang,J.T., Sun,T.T., Dong,J.D., Sanders,A.J., Diao,R.Y., Wang,Y., Fok,K.L., Tsang,L.L. *et al.* (2013). CFTR suppresses tumor progression through miR-193b targeting urokinase plasminogen activator (uPA) in prostate cancer. *Oncogene*, **32**, 2282–2291.
62. Yeh,K.T., Chen,T.H., Yang,H.W., Chou,J.L., Chen,L.Y., Yeh,C.M., Chen,Y.H., Lin,R.I., Su,H.Y., Chen,G.C.W. *et al.* (2011). Aberrant TGFβ/SMAD4 signaling contributes to epigenetic silencing of a putative tumor suppressor, RunX1T1 in ovarian cancer. *Epigenetics*, **6**, 727–739.
63. Rozier,L., El-Achkar,E., Apiou,F. and Debatisse,M. (2004). Characterization of a conserved aphidicolin-sensitive common fragile site at human 4q22 and mouse 6C1: possible association with an inherited disease and cancer. *Oncogene*, **23**, 6872–6880.
64. House,C.D., Vaske,C.J., Schwartz,A.M., Obias,V., Frank,B., Luu,T., Sarvazyan,N., Irby,R., Strausberg,R.L., Hales,T.G. *et al.* (2010). Voltage-gated Na+ channel SCN5A is a key regulator of a gene transcriptional network that controls colon cancer invasion. *Cancer Res.*, **70**, 6957–6967.
65. Teschendorff,A.E. and Widschwendter,M. (2012). Differential variability improves the identification of cancer risk markers in DNA methylation studies profiling precursor cancer lesions. *Bioinformatics*, **28**, 1487–1494.
66. Teschendorff,A.E., Liu,X., Caren,H., Pollard,S.M., Beck,S., Widschwendter,M. and Chen,L. (2014). The dynamics of DNA methylation covariation patterns in carcinogenesis. *PLoS Comput. Biol.*, **10**, e1003709.
67. Zhang,N., Wu,H.J., Zhang,W., Wang,J., Wu,H. and Zheng,X. (2015). Predicting tumor purity from methylation microarray data. *Bioinformatics*, btv370.