BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 16 2014, pages 2360–2366
doi:10.1093/bioinformatics/btu316

Systems biology

Advance Access publication May 2, 2014

A systems-level integrative framework for genome-wide DNA methylation and gene expression data identifies differential gene expression modules under epigenetic control

Yinming Jiao[1], Martin Widschwendter[2] and Andrew E. Teschendorff[1,3,*]

[1]Computational Systems Genomics, CAS-MPG Partner Institute for Computational Biology, Shanghai Institute for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China, [2]Department of Women's Cancer, UCL Elizabeth Garrett Anderson Institute for Women's Health and [3]Statistical Genomics Group, Paul O'Gorman Building, UCL Cancer Institute, University College London, London WC1E 6BT, UK

Associate Editor: Jonathan Wren

ABSTRACT

Motivation: There is a growing number of studies generating matched Illumina Infinium HumanMethylation450 and gene expression data, yet there is a corresponding shortage of statistical tools aimed at their integrative analysis. Such integrative tools are important for the discovery of epigenetically regulated gene modules or molecular pathways, which play key roles in cellular differentiation and disease.

Results: Here, we present a novel functional supervised algorithm, called Functional Epigenetic Modules (FEM), for the integrative analysis of Infinium 450k DNA methylation and matched or unmatched gene expression data. The algorithm identifies gene modules of coordinated differential methylation and differential expression in the context of a human interactome. We validate the FEM algorithm on simulated and real data, demonstrating how it successfully retrieves an epigenetically deregulated gene, previously known to drive endometrial cancer development. Importantly, in the same cancer, FEM identified a novel epigenetically deregulated hotspot, directly upstream of the well-known progesterone receptor tumour suppressor pathway. In the context of cellular differentiation, FEM successfully identifies known endothelial cell subtype-specific gene expression markers, as well as a novel gene module whose overexpression in blood endothelial cells is mediated by DNA hypomethylation. The systems-level integrative framework presented here could be used to identify novel key genes or signalling pathways, which drive cellular differentiation or disease through an underlying epigenetic mechanism.

Availability and implementation: FEM is freely available as an R-package from http://sourceforge.net/projects/funepimod.

Contact: andrew@picb.ac.cn

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

Received on February 23, 2014; revised on April 11, 2014; accepted on April 29, 2014

1 INTRODUCTION

Epigenetic mechanisms are important not only in cellular differentiation (Ziller *et al.*, 2013) but also in disease (Petronis, 2010), especially cancer (Feinberg *et al.*, 2006). Among the epigenetic modifications seen in disease, DNA methylation (DNAm) is especially important for two reasons. First, unlike other epigenetic modifications such as histone marks, it is possible to measure genome-wide DNAm profiles in large numbers of samples (Sandoval *et al.*, 2011), including fresh-frozen and formalin-fixed paraffin-embedded clinical tissue specimens (Lechner *et al.*, 2013). Second, there is mounting evidence that DNAm aberrations can either predispose to or cause disease progression (Feinberg *et al.*, 2006; Issa *et al.*, 1994; Jones *et al.*, 2013; Teschendorff *et al.*, 2012; Ziller *et al.*, 2013). Such causal influences have been shown to be mediated by corresponding changes in gene expression (Jones *et al.*, 2013). Thus, DNAm has emerged as the 'epigenetic marker' of choice in epigenome-wide association studies (Rakyan *et al.*, 2011) and in The Cancer Genome Atlas (TCGA) studies that generate matched gene expression data (Kandoth *et al.*, 2013). Specifically, the Illumina Infinium 450k DNAm beadarray has emerged as a popular choice offering both scalability and coverage at a reasonable economic cost (Sandoval *et al.*, 2011). Thus, there is now an urgent need to develop statistical bioinformatic tools for the integrative analysis of Illumina Infinium 450k and gene expression data.

Here we present the Functional Epigenetic Module (FEM) algorithm for integrative analysis of Illumina Infinium 450k data with matched (or unmatched) gene expression data. The FEM algorithm performs a supervised analysis using a protein–protein interaction (PPI) network (Cerami *et al.*, 2011) as a scaffold to identify gene modules or signalling pathways which are epigenetically and functionally deregulated in a cellular phenotype. Supervised functional network analyses have been used extensively in the gene expression context, see e.g. Chuang *et al.* (2007). We have also previously demonstrated the feasibility of integrating Illumina Infinium 27k DNAm data with a PPI, identifying key signalling pathways and gene modules undergoing age-associated changes in DNA methylation, which were then validated in independent data (West *et al.*, 2013). Similarly, we also recently demonstrated the feasibility and power of integrating Illumina Infinium 27k DNAm data with gene expression and a PPI, identifying an epigenetically deregulated gene, called *HAND2*, which drives endometrial cancer (Jones *et al.*, 2013). What these latter studies have

*To whom correspondence should be addressed

demonstrated is that PPI hotspots of differential methylation (i.e. PPI subnetworks where a significant number of members exhibit statistically significant differential methylation) associated with ageing and cancer exist, and that further integration with gene expression data allows the identification of its putative target(s). Given that the Illumina 27k technology has now been superseded by the more comprehensive Illumina 450k platform, we were impelled to extend our previous algorithm to the 450k case. This extension, however, is non-trivial because for the Illumina 450k Methylation beadchip, there are typically many probes mapping to a gene, and to different regions associated with the gene, including distal transcription start site (TSS), proximal TSS, 5′UTR, first exon, gene body and 3′UTR. Thus, at present, it is still unclear how best to summarize the DNAm values at the gene level, especially in relation to its integration with matched or unmatched gene expression data. Although the reported correlations between Infinium 450k and gene expression data are not strong, they are nevertheless highly statistically significant (Lechner *et al*., 2013), indicating that valuable information can be extracted from such integrative analyses. To this end, we here develop a novel integrative approach, especially designed for Illumina 450k DNAm data, and validate this approach by demonstrating that it can successfully retrieve known genes and gene modules driving cellular differentiation or cancer. The key aspect of our approach is the identification of key genes or gene modules, which are functionally deregulated as a result of underlying DNAm changes.

2 METHODS

2.1 The FEM algorithm

The FEM algorithm is a functional supervised algorithm, which uses a network of relations between genes (in our case a PPI network) to identify subnetworks where a significant number of genes are associated with a phenotype of interest (POI). The association is measured at both the level of DNAm and gene expression. The algorithm thus consists of two main parts: (i) construction of an integrated network in which the associations with the phenotype are encapsulated as weights on the network edges, and (ii) inference of the FEMs as heavy subgraphs on this weighted network.

2.1.1 Integration of DNAm, mRNA expression and PPI network The first step in the algorithm is the integration of a DNAm data matrix with gene expression and a PPI network. We assume that the gene expression data represents normalized estimates of gene expression intensity, summarized at the gene level, so this could include RNA-Seq data or expression data generated using Illumina Beadchips or Affymetrix arrays. The PPI network is derived from the Pathway Commons resource (Cerami *et al*., 2011) and follows the procedure described by West *et al.* (2013). The PPI network consists of 8434 genes annotated to NCBI Entrez identifiers and is sparse, containing 303 600 documented interactions (edges). When integrating with the gene expression and DNAm data, we focus on the overlapping set of genes and extract the maximally connected component defined by this overlapping set. To assign DNAm values to a given gene, in the case of Illumina 27k data, we assigned the probe value closest to the TSS (Jones *et al*., 2013; West *et al*., 2013). In the case of Illumina 450k data, we assign to a gene the average value of probes mapping to within 200 bp of the TSS. If no probes map to within 200 bp of the TSS, we use the average of probes mapping to the first exon of the gene. If such probes are also not present, we use the average of probes mapping to within 1500 bp of the TSS. Justification for this procedure is provided later in Section 3. For each gene g in the maximally connected subnetwork, we then derive a statistic of association between its DNAm profile and the POI, denoted by $t_g^{(D)}$, as well as between its mRNA expression profile and the same POI, which we denote by $t_g^{(R)}$. Here, these statistics are derived using an empirical Bayesian framework and represent regularized t-statistics (Smyth, 2004; Zhuang *et al*., 2012); however, the signed statistics t_g could derive from any appropriate statistical test (e.g. Wald statistics from Cox regression or other types of regression). It is important to also point out that because the integration is done at the level of statistics, that there is no requirement for the mRNA and DNAm data to be matched (i.e. to come from the same individuals). Indeed, the algorithm works unchanged for the unmatched setting.

2.1.2 Construction of the weighted integrated network The key idea is to encapsulate the associations of the genes with the POI in terms of the edge weights, to then identify hotspots of differential methylation and differential expression as 'heavy subnetworks', i.e. subnetworks where the edge weight density is significantly higher than in the rest of the network. Before assigning weights to the network edges, the statistics of one data type (e.g. DNAm) are first scaled uniformly to ensure equal variance between data types. That is, if σ_D and σ_R denote the SD of the statistics $t_g^{(D)}$ and $t_g^{(R)}$ over all genes g, respectively, then we scale all $t_g^{(D)}$ by a factor σ_R/σ_D to ensure equal variance of the statistics from each data type. This is done to avoid one data type overly biasing the downstream inference procedure.

Because the DNAm data derive either from the TSS200, first exon or TSS1500 regions and DNAm levels for these regions are normally anti-correlated with gene expression (see Section 3), we assign an overall statistic value of zero to those genes where the DNAm and expression statistics are of equal sign. For those genes where there is the expected anti-correlation, we use the absolute value, i.e. $t_g = |t_g^{(D)} - t_g^{(R)}|$. All this can be expressed more compactly as

$$t_g = \{H(t_g^{(D)})H(-t_g^{(R)}) + H(-t_g^{(D)})H(t_g^{(R)})\}|t_g^{(D)} - t_g^{(R)}|$$

where $H(x)$ denotes the Heaviside function, defined by $H(x) = 1 \quad \forall x > 0$ and $H(x) = 0 \quad \forall x < 0$. We note that the statistics t_g, as defined, are always positive or zero. Assuming genes g and h are connected in the PPI, we then assign the edge weight as the average of the individual node (gene) statistics, i.e. $w_{gh} = \frac{1}{2}(t_g + t_h)$, which is positive semi-definite. We note that this scheme may introduce zero-weighted edges, which would alter the topology of the network. To avoid this mathematical nuisance, we reassign zero-weighted edges with the smallest positive non-zero value (typically this value is close to zero, i.e. ~0.001).

2.1.3 Inference of FEM A heavy subnetwork, or a module, is then a subgraph where the average weight density, also called *modularity*, is significantly larger than in the rest of the network. Because the modularity will be large if both the DNAm and mRNA expression statistics are large for a significant number of module members, we refer to such a heavy subgraph as an FEM. To infer them, we use a local greedy version of a spin-glass algorithm (Reichardt and Bornholdt, 2006) that we have used previously (West *et al*., 2013). This local greedy version works by specifying a number of seed nodes (genes) around which to search for such modules. By default, the algorithm searches for modules around 100 seeds, defined as the top 100 genes ranked according to the overall statistic t_g. A key parameter of the spin-glass algorithm, called γ, controls the average size of the resulting modules (West *et al*., 2013). By default, the choice is $\gamma = 0.5$, which typically leads to modules with an average size in the range 10 to 100. As demonstrated by us previously, modules in this size range are more likely to validate in independent data, and thus be of biological significance (West *et al*., 2013). Thus, assuming that seeds are uniformly distributed across the network, choosing 100 seeds amounts to a search space of ~5000 to 10 000 nodes, i.e. most of the network. However, we also note here again that not all seeds may lead to modules

(West *et al.*, 2013) because a seed could represent an isolated node of association with the POI. Therefore, in the case of many isolated nodes, it is also advisable to rerun the algorithm with a larger number of seeds.

We note that the spin-glass algorithm infers subnetworks of relatively high edge-weight density (in comparison with the average network edge-weight density). This inference step takes the network topology into account and could thus be overly biased towards specific topological features. Therefore, it is also important to assess the statistical significance of the inferred modules, purely in relation to the DNAm and mRNA expression associations. This is done using a Monte Carlo (MC) randomization procedure, which permutes (1000 permutations) the node statistics around the network and recomputes modularities for the previously inferred modules. We note again that while the inference of the modules depends on the topological features of the modules in relation to the whole network, the MC procedure provides an additional significance test, assessing significance only in relation to the weights, while keeping the network topology fixed (West *et al.*, 2013). Only modules that pass a false discovery rate MC significance threshold of 0.05 are deemed of statistical significance.

2.1.4 Identification of top targets within a FEM Once a FEM has been identified, the top targets of the FEM are defined as those genes within the module with the largest values of t_g. Clearly, for a FEM, constructed from a given seed gene, one of the top targets will be seed gene itself. However, for a module to be a FEM, there must be other genes (besides the seed gene) that contribute significantly to the observed modularity.

2.1.5 The EpiMod algorithm It is clear that the algorithm can be run in 'DNA methylation only' or 'gene expression only' modes, in which case the statistics t_g are defined simply as $|t_g^{(D)}|$ or $|t_g^{(R)}|$, respectively. In the former case, we call it the EpiMod algorithm because it infers differential methylation hotspots. The EpiMod algorithm in the Illumina 27k context was presented by us in West *et al.* (2013).

2.2 Data

To assess correlations between Infinium 450k DNAm and gene expression data, we used samples of normal physiology from TCGA. Specifically, we analyzed 10 normal colon, 3 normal cervical and 17 normal endometrial samples for which matched level-3 Infinium 450k and HiSeq RSEM gene-normalized RNA-Seq data were available. To validate the FEM algorithm on Illumina 450k data, we collected and analyzed an additional 118 endometrial cancer samples with matched RNA-Seq data from the same TCGA study profiling the 17 normal endometrial samples (Kandoth *et al.*, 2013). Integration of our PPI network with the TCGA data described above resulted in a maximally connected subnetwork of 6730 genes.

To test the FEM algorithm in the context of cellular differentiation we analysed a matched Illumina 450k and Agilent gene expression dataset, profiling blood and lymphatic endothelial cells (BECs and LECs, 16 samples) (Bronneke *et al.*, 2012). Because endothelial cells exhibit a remarkable plasticity and ease of transdifferentation, it is likely that epigenetic mechanisms control cell subtype specificity (Bronneke *et al.*, 2012).

In all cases, the DNAm 450k data were corrected for the type-2 probe bias using BMIQ (Teschendorff *et al.*, 2013).

2.3 Simulation

To assess the sensitivity (SE) and specificity (SP) of the FEM algorithm, we used a simulation model, in which we simulated statistics of differential methylation and differential expression on the PPI network. As a true module, we picked the *HAND2* module because the biological and clinical significance of the driver gene, *HAND2*, contained within this module, has been extensively validated (Jones *et al.*, 2013). We bootstrapped statistics for the member genes of this module to come from the top and lower 5% statistics quantiles, with the statistics of the rest of the network nodes bootstrapped from the middle 90% portion. For each simulation run, the SE and SP of the FEM algorithm were recorded. Here, SE was defined as the fraction of *HAND2* module members captured by the inferred FEM module, whereas SP was defined as one minus the false-positive rate.

3 RESULTS

3.1 DNAm values around TSS best predict gene expression

Because with the Illumina Infinium 450k beadarray, many probes may map to a given gene (and to different regions associated with the gene), we wanted to first assess which probes are most predictive of the gene expression state. To determine this, we collected high-quality data from the TCGA representing samples of normal physiology, for which matched Infinium 450k and RNA-Seq data were available (Section 2). For each gene in each sample, we averaged the DNAm β-values of probes mapping to the same gene region: these regions were defined as 1500 bp upstream of the TSS (TSS1500), 200 bp upstream of the TSS (TSS200), 5′UTR, first exon, gene body and 3′UTR. For a given sample, and for each genetic region, DNAm values were then binned into five levels. Boxplots of log-normalized RNA-Seq counts against binned DNAm levels for each genetic region in a given sample demonstrate that the strongest association (in terms of R^2 values) between DNAm and gene expression is for the first exon and TSS200 regions, followed by TSS1500, all exhibiting a relatively strong anti-correlation ($R^2 \sim 0.3$, i.e Pearson correlations of ~ -0.5; Fig. 1A). This pattern was replicated in each of 30 normal samples (ns) encompassing three different tissue types, with first exon, TSS200 and TSS1500 emerging as the most predictive regions (Fig. 1B and Supplementary Figs. S1–S3). Owing to the differences in the degrees of freedom, and hence to assess more objectively the three different regions, we compared their predictive power by restricting to a common pool of genes, specifically, those with probes mapping to each one of these three regions. This unbiased analysis revealed that TSS200 obtained marginally but consistently higher R^2 and t-statistic values compared with first exon, and substantially higher values compared with TSS1500 (Fig.1C and Supplementary Fig. S4). Based on these results, we thus devised the following scheme to assign a unique DNAm value to a given gene: for a gene with TSS200 probes, the average DNAm of such TSS200 probes was used. For a gene with no TSS200 probes but with first exon probes, we assigned the corresponding average over first exon probes. Finally, for a gene with no TSS200 or first exon probes, we used the average over TSS1500 probes.

3.2 Validation of the EpiMod and FEM algorithms in Illumina 450k endometrial cancer data

We previously demonstrated, using Illumina Infinium 27k data from normal endometrial and endometrial cancer samples that the interaction neighbourhood of the *HAND2* gene represents a differential methylation cancer hotspot (Jones *et al.*, 2013). By using unmatched Affymetrix gene expression data, we further demonstrated that this *HAND2* module represented a hotspot of differential methylation and differential expression, identifying

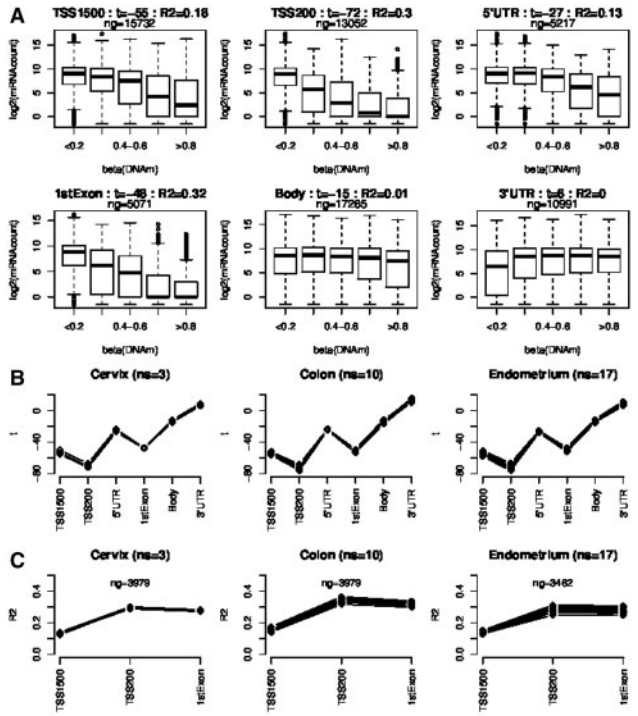

Fig. 1. (**A**) Scatterplots of log-normalized RNA-seq gene counts against corresponding average DNAm β-values for probes mapping within the given region of a gene (TSS1500, TSS200, 5′UTR, first exon, gene body, 3′UTR), for one given cervical normal sample. The *t*-statistics and R^2 of a linear regression for each region are given above plots, as well as the number of data points (genes, ng) in the regression. (**B**) Plots of the regression *t*-statistics between log-normalized RNA-Seq counts and binned DNAm levels, stratified according to genetic region. The number of curves equals the number of ns within each tissue type and is indicated above the plot. (**C**) Focusing on the regions TSS1500, TSS200 and first exon, an objective comparison of R^2 values between the three regions restricting to genes with probes mapping to all of these regions

HAND2 itself as the key target (Jones *et al*., 2013). Indeed, the biological, functional and clinical importance of the *HAND2* gene was further demonstrated by Jones *et al*. (2013), where we showed that it is causally implicated in the development of endometrial cancer.

Thus, to validate our method of summarizing Illumina 450k DNAm data at the gene level, we decided to apply the EpiMod and FEM algorithms to an independent endometrial normal/cancer Infinium 450k set generated as part of the TCGA (Section 2) (Kandoth *et al*., 2013). Specifically, our aim was to see if we could retrieve the *HAND2* module, but now using 450k data. In the first instance, we excluded the RNA-Seq data and only used the EpiMod algorithm to identify differential methylation hotspots associated with endometrial cancer (Section 2). This resulted in 23 differential methylation hotspots, including one centred around *HAND2* (Supplementary Table S1), which was also one of the top seeds. Thus, extension of the EpiMod algorithm to the 450k case has indeed identified a module highly similar to the one we inferred previously using independent Illumina 27k data. Close inspection of the *HAND2* module

Table 1. Summary of the output of the FEM algorithm listing 4 of the 17 hotspots of differential methylation and expression (FEMs) in endometrial cancer

Seed	Size	Modularity	*P*	Top targets
SFN	18	3.62	0.007	*SFN, KCNK3*
TGFB1I1	10	3.45	0.002	*TGFB1I1, LIMS2, GIT2, P2RX7*
HAND2	11	2.87	0.016	*HAND2*
LNX1	54	2.16	0.006	*LNX1, NADK, WAC, CKS2*

Note: Columns label the seed gene symbol, the size of the FEM, its modularity (defined as the average of the edge-weights), the associated *P*-value and a list of the top targets.

confirmed that many other genes in this module undergo differential methylation in cancer (Supplementary Table S2). Running the FEM algorithm (i.e. including the matched RNA-Seq data) identified 17 FEMs passing the significance threshold of 0.05, and all with sizes in the range 10 to 59 (Supplementary Table S3). A table summarizing the output of the algorithm for 4 of the top 17 FEMs shows that one of the modules is again centred around *HAND2* (Table 1). Close inspection of this module shows that *HAND2* is the main gene that is functionally deregulated, despite many other module genes being deregulated at the DNAm level (Fig. 2 and Supplementary Table S4). Specifically, *HAND2* exhibited simultaneous hypermethylation and underexpression in cancer, as observed by us previously using independent data (Jones *et al*., 2013). Thus, we can conclude that the extension of the FEM algorithm to the 450k case has successfully identified an epigenetically deregulated gene module targeting a gene that has been demonstrated to drive endometrial cancer development (Jones *et al*., 2013).

While the FEM *HAND2* module identified *HAND2* as its main target, other FEMs contained multiple putative targets, for instance, the *TGFB1I1* module (Tables 1 and 2). Thus, in the same way that a copy-number aberration in cancer can affect the gene expression levels of several genes within the altered genomic region (Chin *et al*., 2007), epigenetic changes in cancer affecting functionally related genes may also affect the gene expression of several targets. To confirm the hotspot nature and biological significance of the *TGFB1I1* module, we sought to validate it in the independent endometrial normal/cancer dataset considered by Jones *et al*. (2013), which also used different technologies (Illumina 27k for DNAm and Affymetrix for mRNA expression). We observed that this same module was a significant FEM in this independent set, with the top targets (*TGFB1I1* and *LIMS2*) showing the same pattern of coordinated hypermethylation and underexpression in cancer (Supplementary Fig. S5). Interestingly, *TGFB1I1*, also known as *HIC5*, is a known co-activator of the progesterone receptor (PGR) and has previously been implicated in endometriosis (Aghajanova *et al*., 2009). Remarkably, given that *HAND2* is a target of PGR and that it mediates the tumour suppressive effects of progesterone (Jones *et al*., 2013), it is entirely plausible that silencing of *HIC5* can have a similar effect by downregulating the PGR pathway.

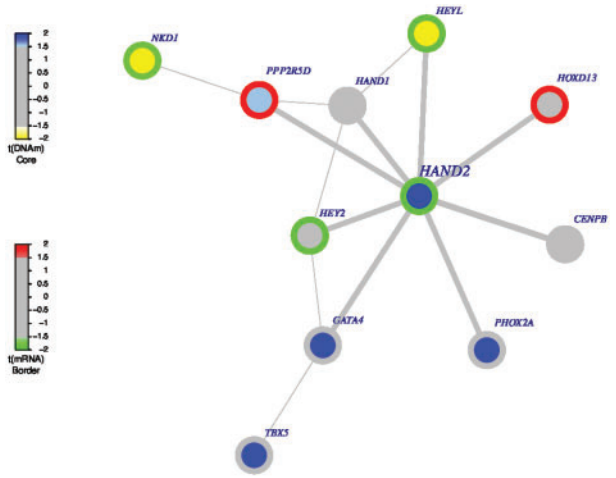

Fig. 2. Depicted is the FEM centred around seed gene *HAND2*. Edge widths are proportional to the average statistic of the genes making up the edge. Node shades denote the differential DNAm statistics as indicated. Border shades denote the differential expression statistics. Observe that despite many nodes exhibiting differential methylation and differential expression, only *HAND2* exhibits the expected anti-correlation with hypermethylation leading to underexpression

Table 2. The FEM with seed gene *TGFB1I1* inferred from the matched Illumina 450k RNA-Seq endometrial cancer dataset

Gene	*t*(DNAm)	*P*(DNAm)	*t*(mRNA)	*P*(mRNA)	*t*(Int)
TGFB1I1	11.41	1e-21	–13.03	1e-25	9.71
LIMS2	6.06	1e-8	–9.66	4e-17	6.00
P2RX7	5.62	1e-7	–7.08	7e-11	4.99
RYR3	1.44	0.15	–11.36	2e-21	4.21
GIT2	2.72	0.007	–5.37	3e-7	3.01
FKBP1A	1.93	0.06	1.33	0.18	0
SVIL	–0.1	0.92	–10.24	1e-18	0
HIPK3	–0.38	0.7	–4.49	1e-5	0
PANX1	0.45	0.65	2.21	0.03	0
LIMD1	2.56	0.01	0.89	0.38	0

Note: For each of the module members, we provide the symbol, entrez ID, the statistic and *P*-value of differential methylation, the statistic and *P*-value of differential expression and the overall statistic *t*(Int). Five putative targets are indicated in boldface.

3.3 The FEM algorithm identifies DNA methylation-regulated gene expression modules associated with endothelial cell differentiation

As a second application of the FEM algorithm, we tested it in the context of cellular differentiation. Specifically, we applied it to a matched DNAm-mRNA expression dataset of endothelial cells (Bronneke *et al*., 2012), to identify hotspots of coordinated differential methylation and differential expression between two cellular subtypes: BECs and LECs. Transdifferentiation between these two endothelial subtypes has been widely reported, with DNAm emerging as a key regulator of this phenotypic plasticity (Bronneke *et al*., 2012). Thus, we decided to test the FEM algorithm in its ability to retrieve genes or gene modules known to mark LECs/BECs, but, importantly, to also identify novel biologically plausible genes or gene modules that may determine endothelial cell subtype specificity.

Running FEM with 300 seeds, we identified 41 FEMs containing at least five genes (Supplementary Table S5). Many of these included or were centred around genes (e.g. *BATF*, *IL7*, *RTKN*, *MAF*, *NRP2*), which have been reported to be overexpressed in LECs compared with BECs (Bronneke *et al*., 2012). This list also included *PROX1*, a transcription factor required for LEC differentiation (Amatschek *et al*., 2007; Bronneke *et al*., 2012). Although many of these genes were reported to undergo DNAm changes, these changes were mainly restricted to regions farther away from the TSS (Bronneke *et al*., 2012). This explains why in our FEM analysis, which focuses mainly on the TSS200 region, many of these genes showed more modest DNAm changes (Supplementary Table S5). In spite of this, FEM was able to capture these genes, owing to their significant differential expression changes (Supplementary Table S5).

Most importantly, FEM identified a novel module mapping to major histocompatability complex (MHC) genes, of which several members (e.g *HLA-DMB*, *HLA-DRB1*, *CD74*, *HLA-DMA*, *HLA-DRB5*) showed coordinated hypomethylation and overexpression in blood endothelial relative to lymphatic cells (Fig. 3 and Table 3). We note that these MHC genes were not highlighted by Bronneke *et al*. (2012), yet the biological plausibility of this module is strongly supported by another study (Amatschek *et al*., 2007), which observed *HLA-DRB1* to be a marker of BECs. As explained by Amatschek *et al*. (2007), the overexpression of these genes in BECs is likely to be triggered by the tissue environment. Our FEM analysis further suggests that DNAm plays a key intermediary role, regulating the overexpression of this specific MHC module in BECs.

Another example of an interesting novel module is that centred around *STAT6*, which we found to be hypomethylated and overexpressed in BECs (Supplementary Table S5). Overexpression of *STAT6* in BECs relative to LECs is supported by an independent study (Nelson *et al*., 2007). Interestingly, several other genes in the same module exhibited either significant overexpression (e.g. *IFI35*, *NMI*) or differential methylation (e.g. *THY1*, *AICDA*). Most importantly, however, the FEM analysis suggests that the observed overexpression of *STAT6* may be driven by hypomethylation of its promoter region.

3.4 Assessment of FEM's operating characteristics

Although we have shown that FEM can successfully identify key hotspots of differential methylation and expression, it is nevertheless still important to assess its overall operating characteristics. To this end, we devised a realistic simulation model, using the same real PPI network as a scaffold, and using the *HAND2* module (11 genes) as an example of a realistic module. Statistics of differential methylation and differential expression were simulated, however, assigning larger values to the HAND2 module genes than for the other nodes in the network (see Section 2). This simulation thus allows the sensitivity (SE) of the inference procedure to be assessed. We performed 100 simulations, recording the SE, specificity (SP) and positive predictive value (Fig. 4). The mean SP and SE (0.79) were high, although in

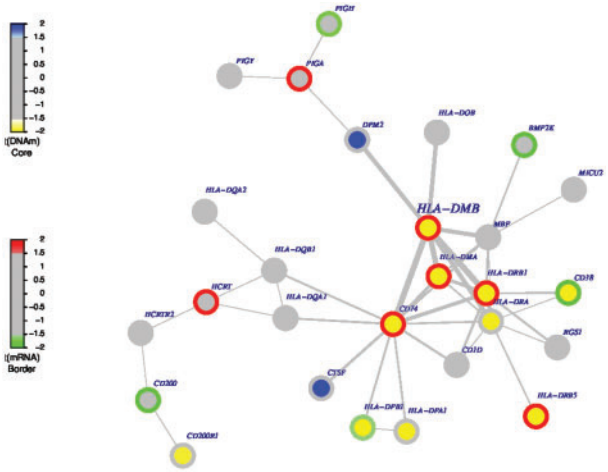

Fig. 3. Depicted is the FEM centred around seed gene *HLA-DMB*. Edge widths are proportional to the average statistic of the genes making up the edge. Node shades denote the differential DNAm statistics as indicated. Border shades denote the differential expression statistics. Observe how this module is driven by five genes exhibiting coordinated hypomethylation and overexpression in blood epithelial cells compared with lymphatic cells

Table 3. Five members of the 27-gene FEM with seed gene HLA-DMB, all overexpressed in BECs compared with LECs

Gene	*t*(DNAm)	*P*(DNAm)	*t*(mRNA)	*P*(mRNA)	*t*(Int)
HLA-DMB	–24.21	2e-13	4.6	0.0003	14.73
HLA-DRB1	–5.33	9e-5	6.49	8e-6	6.37
CD74	–7.87	1e-6	3.56	0.003	5.96
HLA-DMA	–3.79	0.002	4.16	0.0008	4.27
HLA-DRB5	–2.04	0.06	5.29	8e-5	4.04

Note: We provide the symbol, entrez ID, the statistic and *P*-value of differential methylation, the statistic and *P*-value of differential expression and the overall statistic *t*(Int). Positive statistics mean higher levels in BECs compared with LECs.

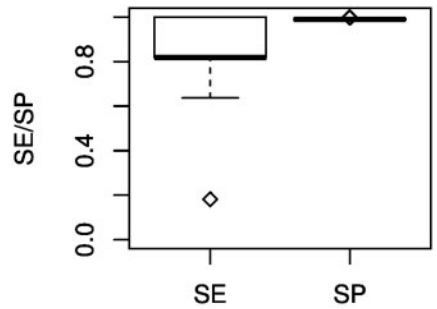

Fig. 4. Operating characteristics of FEM: SE and SP values of the FEM algorithm to identify true modules, as evaluated over a simulation ensemble of 100 runs (Section 2)

the case of SE, there was also significant variation. This indicates that in the majority of runs, the algorithm can identify most members of the true module (Fig. 4).

4 DISCUSSION

We have presented a novel algorithm for integrative functional supervised analyses of Illumina 450k DNAm and gene expression data. By applying and testing it in two different biological contexts, we have demonstrated here the feasibility of integrating Illumina 450k data with gene expression in a systems context, using a human protein interaction network as a scaffold to identify gene modules whose differential expression is regulated by differential methylation. In an application to cancer, we have seen how it successfully retrieved an epigenetically deregulated gene module centred around (*HAND2*), a gene known to mediate the tumour suppressive effects of the PGR pathway (Jones *et al.*, 2013). Specifically, silencing of *HAND2* inactivates this tumour suppressor pathway. It is therefore remarkable that FEM identified another hotspot and target gene (*TGFB1I1*) implicated in the PGR pathway. Our data suggest that hypermethylation-mediated silencing of *TGFB1I1* could also lead to downregulation of the PGR tumour suppressor pathway because *TGFB1I1* (*HIC5*) is a known co-activator of *PGR* (Aghajanova *et al.*, 2009). Importantly, the *TGFB1I1* module was validated in independent data, further supporting its biological significance. In the context of endothelial cell differentiation, we have shown how FEM retrieved known markers of endothelial cell subtypes, including an MHC gene module hypomethylated and overexpressed in BECs. Likewise, it identified DNA hypomethylation as the potential mechanism underlying *STAT6*'s overexpression in BECs. That the overexpression of the MHC module genes and *STAT6* in BECs may be regulated by DNAm is—to the best of our knowledge—an entirely novel insight. All these results clearly highlight the value of the FEM algorithm to identify novel biologically and clinically interesting gene modules of coordinated differential methylation and expression.

It is important to comment on the number and nature of the inferred FEMs. In principle, a FEM could be driven by one gene only, if this gene has exceptionally large absolute statistics of differential methylation and expression. Other FEMs could be driven by several genes, but with each one having only a marginally significant statistic. Importantly, the FEM algorithm is capable of identifying both types. For instance, in the application to endometrial cancer we have observed FEMs of the two types, with the *HAND2* module being an example of the former, and the *TGFB1I1* module an example of the latter. In the case of the *HAND2* module, many genes showed differential methylation changes, but only *HAND2* showed the expected directional change in gene expression, thus identifying it as the target of the deregulated epigenetic hotspot. In the application to endothelial cell differentiation, we observed a similar pattern, with some FEMs driven mainly by individual genes with large differential methylation and expression statistics, and others driven by a number of functionally related genes (e.g. the MHC module). The existence of 'rich' modules (i.e. modules implicating several targets, like the MHC module) should not be surprising because functionally related genes are often commonly regulated, with epigenetic mechanisms controlling this regulation. In particular, application of FEM to complex tissues such as blood may reveal

many more rich modules, which are likely to be cell subtype-specific. Indeed, it may be possible to use such rich modules as a means of correcting for cellular heterogeneity.

However, in the application to endometrial cancer and endothelial cell differentiation, we did not observe many rich FEMs. This scarcity most likely reflects the noisy nature, or the level of contextual irrelevance of the PPI network, yet it also likely reflects our conservative approach to give the TSS200 region most weight. An alternative approach, which selects the largest statistic across the different genetic regions associated with a given gene, would likely yield richer FEMs, yet the probability that the observed deregulation is because of DNAm changes would also be less clear. Our conservative approach, while identifying a smaller number of rich FEMs, is more likely to identify those which are under direct DNAm regulation. Another approach would to rerun the algorithm giving more preference to the first exon and TSS1500 regions, but this resulted in similar modules (Supplementary Tables S6 and S7). An alternative explanation for why rich FEMs are scarce could be biological. For instance, most of the genes in the inferred FEMs are characterized by differential methylation or differential expression but not both. The observed differential expression of module genes, not caused by underlying *in-cis* DNAm changes, may nevertheless still be caused by DNAm changes of neighbouring interacting genes. As shown here, FEM identified many modules exhibiting these types of alterations, and further investigation of these patterns might be of interest.

We stress again that FEM represents a functional supervised network algorithm, integrating multi-dimensional DNAm and gene expression data in the context of a human PPI network. The power of such functional supervised analyses has been previously demonstrated in the gene expression (Chuang *et al*., 2007) and DNAm (West *et al*., 2013) contexts. Specifically, these studies demonstrated that the use of a network, encoding functional relations between genes, can improve the probability of detecting a true positive. It is equally important, however, to point out that any such integrative network approach is restricted to a smaller search space because typically not all profiled genes may be present in the actual PPI network. Moreover, in the application to Illumina Infinium 450k data, averaging over probes within genetic regions, as done in the FEM algorithm, can lead to loss of probe-level information. Thus, when analysing matched DNAm and gene expression data, the FEM algorithm should be viewed as an analysis strategy, which complements the more ordinary univariate supervised and Gene Set Enrichment Analysis method.

In summary, the FEM algorithm presented here will be useful to a growing number of studies that aim to identify gene modules or molecular pathways that are epigenetically and functionally deregulated in disease. Similarly, FEM could be applied to cellular differentiation data to identify cell type-specific gene expression modules under the regulation of DNA methylation.

Funding: Y.J. and A.E.T. acknowledge support from the Chinese Academy of Sciences, the Shanghai Institute for Biological Sciences and the Max-Planck Gesellschaft.

Conflicts of Interest: none declared.

REFERENCES

Aghajanova,L. *et al.* (2009) The progesterone receptor coactivator Hic-5 is involved in the pathophysiology of endometriosis. *Endocrinology*, **150**, 3863–3870.

Amatschek,S. *et al.* (2007) Blood and lymphatic endothelial cell-specific differentiation programs are stringently controlled by the tissue environment. *Blood*, **109**, 4777–4785.

Bronneke,S. *et al.* (2012) DNA methylation regulates lineage-specifying genes in primary lymphatic and blood endothelial cells. *Angiogenesis*, **15**, 317–329.

Cerami,E.G. *et al.* (2011) Pathway Commons, a web resource for biological pathway data. *Nucleic Acids Res.*, **39**, D685–D690.

Chin,S.F. *et al.* (2007) High-resolution aCGH and expression profiling identifies a novel genomic subtype of ER negative breast cancer. *Genome Biol.*, **8**, R215.

Chuang,H.Y. *et al.* (2007) Network-based classification of breast cancer metastasis. *Mol. Syst. Biol.*, **3**, 140.

Feinberg,A.P. *et al.* (2006) The epigenetic progenitor origin of human cancer. *Nat. Rev. Genet.*, **7**, 21–33.

Issa,J.P. *et al.* (1994) Methylation of the oestrogen receptor CpG island links ageing and neoplasia in human colon. *Nat. Genet.*, **7**, 536–540.

Jones,A. *et al.* (2013) Role of DNA methylation and epigenetic silencing of HAND2 in endometrial cancer development. *PLoS Med.*, **10**, e1001551.

Kandoth,C. *et al.* (2013) Integrated genomic characterization of endometrial carcinoma. *Nature*, **497**, 67–73.

Lechner,M. *et al.* (2013) Identification and functional validation of HPV-mediated hypermethylation in head and neck squamous cell carcinoma. *Genome Med.*, **5**, 15.

Nelson,G.M. *et al.* (2007) Differential gene expression of primary cultured lymphatic and blood vascular endothelial cells. *Neoplasia*, **9**, 1038–1045.

Petronis,A. (2010) Epigenetics as a unifying principle in the aetiology of complex traits and diseases. *Nature*, **465**, 721–727.

Rakyan,V.K. *et al.* (2011) Epigenome-wide association studies for common human diseases. *Nat. Rev. Genet.*, **12**, 529–541.

Reichardt,J. and Bornholdt,S. (2006) Statistical mechanics of community detection. *Phys. Rev. E. Stat. Nonlin. Soft. Matter Phys.*, **74**, 016110.

Sandoval,J. *et al.* (2011) Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome. *Epigenetics*, **6**, 692–702.

Smyth,G.K. (2004) Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.*, **3,** Article3.

Teschendorff,A.E. *et al.* (2012) Epigenetic variability in cells of normal cytology is associated with the risk of future morphological transformation. *Genome Med.*, **4**, 24.

Teschendorff,A.E. *et al.* (2013) A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450k DNA methylation data. *Bioinformatics*, **29**, 189–196.

West,J. *et al.* (2013) An integrative network algorithm identifies age-associated differential methylation interactome hotspots targeting stem-cell differentiation pathways. *Sci. Rep.*, **3**, 1630.

Zhuang,J. *et al.* (2012) A comparison of feature selection and classification methods in DNA methylation studies using the Illumina Infinium platform. *BMC Bioinformatics*, **13**, 59.

Ziller,M.J. *et al.* (2013) Charting a dynamic DNA methylation landscape of the human genome. *Nature*, **500**, 477–481.