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, Andrew E Teschendorff 1,3*  

1Computational Systems Genomics, CAS-MPG Partner Institute for Computational Biology, Shanghai Institute for Biological Sciences, Chinese Academy of Sciences, 320 Yue Yang Road, Shanghai 200031, China.  
2Department of Women’s Cancer, UCL Elizabeth Garrett Anderson Institute for Women’s Health, University College London, London WC1E 6BT, United Kingdom.  
3Statistical Genomics Group, Paul O’Gorman Building, UCL Cancer Institute, University College London, 72 Huntley Street, London WC1E 6BT, United Kingdom.

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

1 INTRODUCTION

Epigenetic mechanisms are important not only in cellular differentiation (Ziller et al., 2013), but also in disease (Petronis, 2010), specially cancer (Feinberg et al., 2006). Among the epigenetic modifications seen in disease, DNA methylation (DNAm) is specially 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 (Issa et al., 1994; Feinberg et al., 2006; Teschendorff et al., 2012; Jones et al., 2013; Ziller et al., 2013). Such causal influences have been shown to be mediated by corresponding changes in gene expression (Jones et al., 2013). Thus, DNA methylation has emerged as the “epigenetic marker” of choice in Epigenome Wide Association Studies (EWAS) (Rakyan et al., 2011) and in 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 demonstrated is that PPI hotspots of differential methylation (i.e. PPI subnetworks were 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 since 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, 1st Exon, Gene Body and 3'UTR. Thus, at present, it is still unclear how best to summarize the DNA methylation values at the gene level, specially 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, specially 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

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 DNA methylation and gene expression. The algorithm thus consists of two main parts: (1) Construction of an integrated network in which the associations of the genes with the POI in terms of the edge weights, in order 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 $\sigma_D$ and $\sigma_E$ denote the standard deviations of the statistics $D^{(g)}$ and $E^{(g)}$ over all genes $g$, respectively, then we scale all $D^{(g)}$ by a factor $\sigma_R/\sigma_D$ to ensure equal variance of the statistics from each data type.

This is done in order to avoid one data type overly biasing the downstream inference procedure.

Since the DNAm data derives either from the TSS200, 1st exon or TSS1500 regions and DNAm levels for these regions are normally anticorrelated with gene expression (see Results), 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 anticorrelation, we use the absolute value, i.e. $t_g = |D^{(g)}| - E^{(g)}|$. All this can be expressed more compactly as

$$t_g = \begin{cases} D^{(g)} & \text{if } H(x) = 1, x > 0, H(x) = 0, x < 0, \text{the statistics } t_g \text{, as defined, are always positive or zero.} \\
-D^{(g)} & \text{otherwise} \end{cases}$$

where $H(x)$ denotes the Heaviside function, defined by $H(x) = 1$ if $x > 0$ and $H(x) = 0$ if $x < 0$. 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 very close to zero, i.e. $\approx 0.001$).

Inference of Functional Epigenetic Modules (FEM) A heavy subgraph, 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. Since 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 a Functional Epigenetic Module (FEM). To infer them, we use a local greedy version of a spin-glass algorithm (Reichardt and Bornholdt, 2006) which 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 $\gamma$, 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 approximately 5000 to 10000 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), since a seed could represent an isolated node of association with the phenotype of interest. Therefore, in the case of many isolated nodes, it is also advisable to return 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 to 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. It is therefore also important to assess the statistical significance of the inferred modules, purely in relation to the whole network, the MC procedure provides an additional significance test, assessing significance only in relation to the whole network, 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.

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 the 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.

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, since it infers differential methylation hotspots. The EpiMod algorithm in the Illumina 27k context was presented by us in West et al. (2013).

Data To assess correlations between Illumina 450k DNA methylation and gene expression data we used samples of normal physiology from The Cancer Genome Atlas (TCGA). Specifically, we analyzed 10 normal colon, 3 normal cervical, and 17 normal endometrial samples for which matched TSS1500, 200bp upstream of the TSS (TSS200), 5'UTR, 1st 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 1st Exon and TSS200 regions, followed by TSS1500, all exhibiting a relatively strong anti-correlation ($R^2 \sim 0.3$ i.e Pearson correlations of $\sim -0.5$).

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 neighborhood 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 methylation and differential expression on the PPI network. As a true module we picked the HAND2 module, since 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 sensitivity and specificity of the FEM algorithm was recorded. Here, sensitivity was defined as the fraction of HAND2 module members captured by the inferred FEM module, whereas specificity was defined as one minus the false positive rate.

3 RESULTS DNA methylation values around transcription start site (TSS) best predict gene expression Since 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 was available (Methods). For each gene in each sample, we averaged the DNA methylation $\beta$-values of probes mapping to the same gene region: these regions were defined as 1500bp upstream of the TSS (TSS1500), 200bp upstream of the TSS (TSS200), 5'UTR, 1st 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 1st Exon and TSS200 regions, followed by TSS1500, all exhibiting a relatively strong anti-correlation ($R^2 \sim 0.3$ i.e Pearson correlations of $\sim -0.5$).
differential methylation and differential expression, identifying HAND2 itself as the key target (Jones et al., 2013). Indeed, the biological, functional and clinical importance of the HAND2 gene was further demonstrated in Jones et al. (2013), where we showed that it is causally implicated in the development of endometrial cancer.

Thus, in order 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 (Methods) (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 (Methods). This resulted in 23 differential methylation hotspots, including one centred around HAND2 (Supp.Table.1), 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 confirmed that many other genes in this module undergo differential DNA methylation in cancer (Supp.Table.2). 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 (Supp.Table.3). 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 which is functionally deregulated, despite many other module genes being deregulated at the DNAm level (Fig. 2 & Supp.Table.4). 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 TGFBI1 module (Table 1 & Table 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
Table 1. Summary of the output of the FEM algorithm listing 4 of the 17 modules that this same module was a significant FEM in this independent endometrial normal/cancer data set considered in Jones et al. (2013), which also used different technologies (Illumina 27k Illumina 450k RNA-Seq endometrial cancer data set). 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.

<table>
<thead>
<tr>
<th>Seed</th>
<th>Size</th>
<th>Modularity</th>
<th>P</th>
<th>Top Targets</th>
</tr>
</thead>
<tbody>
<tr>
<td>SFN</td>
<td>18</td>
<td>3.62</td>
<td>0.007</td>
<td>SFN, KCNK3</td>
</tr>
<tr>
<td>TGFβ1I</td>
<td>10</td>
<td>3.45</td>
<td>0.002</td>
<td>TGFβ1I, LIMS2, GIT2, P2RX7</td>
</tr>
<tr>
<td>HAND2</td>
<td>11</td>
<td>2.87</td>
<td>0.016</td>
<td>HAND2</td>
</tr>
<tr>
<td>LNX1</td>
<td>54</td>
<td>2.16</td>
<td>0.006</td>
<td>LNX1, NADK, WAC, CKS2</td>
</tr>
</tbody>
</table>

Table 2. The FEM with seed gene TGFβ1I inferred from the matched Illumina 450k RNA-Seq endometrial cancer data set. For each of the module members, we provide the symbol, entrez ID, the statistic and P-value of differential methylation and expression (FEMs) in endometrial cancer. Columns label the seed gene symbol, the size of the FEM, its modularity (defined as the average of the edge-weights), the associated significance of the DNA methylation, the statistic and P-value of differential methylation and expression and the overall statistic t(Int). Five putative targets are indicated in boldface.

<table>
<thead>
<tr>
<th>Gene</th>
<th>t(DNAm)</th>
<th>P(DNAm)</th>
<th>t(mRNA)</th>
<th>P(mRNA)</th>
<th>t(Int)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TGFβ1I</td>
<td>11.41</td>
<td>1e-21</td>
<td>-13.03</td>
<td>1e-25</td>
<td>9.71</td>
</tr>
<tr>
<td>LIMS2</td>
<td>6.06</td>
<td>1e-8</td>
<td>-9.66</td>
<td>4e-17</td>
<td>6.00</td>
</tr>
<tr>
<td>P2RX7</td>
<td>5.62</td>
<td>1e-7</td>
<td>-7.08</td>
<td>7e-11</td>
<td>4.99</td>
</tr>
<tr>
<td>RVR3</td>
<td>1.44</td>
<td>0.15</td>
<td>-11.36</td>
<td>2e-21</td>
<td>4.21</td>
</tr>
<tr>
<td>GIT2</td>
<td>2.72</td>
<td>0.007</td>
<td>-5.37</td>
<td>3e-7</td>
<td>3.01</td>
</tr>
<tr>
<td>FKBP1A</td>
<td>1.93</td>
<td>0.06</td>
<td>1.33</td>
<td>0.18</td>
<td>0</td>
</tr>
<tr>
<td>SVIL</td>
<td>-0.1</td>
<td>0.92</td>
<td>-10.24</td>
<td>1e-18</td>
<td>0</td>
</tr>
<tr>
<td>HIPK3</td>
<td>-0.38</td>
<td>0.7</td>
<td>-4.49</td>
<td>1e-5</td>
<td>0</td>
</tr>
<tr>
<td>PANX1</td>
<td>0.45</td>
<td>0.65</td>
<td>2.21</td>
<td>0.03</td>
<td>0</td>
</tr>
<tr>
<td>LIMD1</td>
<td>2.56</td>
<td>0.01</td>
<td>0.89</td>
<td>0.38</td>
<td>0</td>
</tr>
</tbody>
</table>

functionally related genes may also affect the gene expression of several targets. To confirm the hotspot nature and biological significance of the TGFβ1I module, we sought to validate it in the independent endometrial normal/cancer data set considered in Jones et al. (2013), which used also used different technologies (Illumina 27k for DNA methylation and Affymetrix for mRNA expression). We observed that this same module was a significant FEM in this independent set, with the top targets (TGFβ1I and LIMS2) showing the same pattern of coordinated hypermethylation and underexpression in cancer (Supp. Table 5). Interestingly, TGFβ1I, 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.

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 DNA methylation and expression data set of endothelial cells (Bronneke et al., 2012), in order to identify hotspots of coordinated differential methylation and differential expression between two cellular subtypes: blood and lymphatic endothelial cells (BECs and LECs). Transdifferentiation between these two endothelial subotypes has been widely reported, with DNA methylation 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 also to identify novel biologically plausible genes or gene modules which may determine endothelial cell subtype specificity.

Running FEM with 300 seeds, we identified a total of 41 FEMs containing at least 5 genes (Supp. Table 5). Many of these included co-ordinated or were centred around genes (e.g. BATF, IL7, RTKN, MAF, NRP2), which have been reported to be overexpressed in LECs compared to BECs (Bronneke et al., 2012). This list also included PROX1, a transcription factor required for LEC differentiation (Bronneke et al., 2012; Amatschek et al., 2007). Although many of these genes were reported to undergo DNA methylation changes, these changes were mainly restricted to regions further 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 DNA methylation changes (Supp. Table 5). In spite of this, FEM was able to capture these genes, owing to their significant differential expression changes (Supp. Table 5).

Most importantly, FEM identified a novel module mapping to major-histocompatibility (MHC) genes, of which several members
The MHC module in BECs.

Intermediary role, regulating the overexpression of this specific FEM-analysis further suggests that DNA methylation plays a key role. Our study observed that Amatschek et al. observed endothelial relative to lymphatic cells (Fig. 3 & Table 3). We showed coordinated hypomethylation and overexpression in blood endothelial relative to lymphatic cells (Fig. 3 & Table 3). We note that these MHC genes were not highlighted in 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 in 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 DNA methylation 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 (Supp. Table.5). 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.

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 however simulated, assigning larger values to the HAND2 module genes than for the other nodes in the network (Methods). This simulation thus allows the sensitivity of the inference procedure to be assessed. We performed a total of 100 simulations, recording the sensitivity, specificity and positive predictive value (PPV) (Fig. 4). The mean specificity and sensitivity (0.79) were high, although in the case of sensitivity 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).

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

<table>
<thead>
<tr>
<th>Gene</th>
<th>t(DNAm)</th>
<th>P(DNAm)</th>
<th>t(mRNA)</th>
<th>P(mRNA)</th>
<th>t(Int)</th>
</tr>
</thead>
<tbody>
<tr>
<td>HLA-DMB</td>
<td>-24.21</td>
<td>2e-13</td>
<td>4.6</td>
<td>0.0003</td>
<td>14.73</td>
</tr>
<tr>
<td>HLA-DRB1</td>
<td>-5.33</td>
<td>9e-5</td>
<td>6.49</td>
<td>8e-6</td>
<td>6.37</td>
</tr>
<tr>
<td>CD74</td>
<td>-7.87</td>
<td>1e-6</td>
<td>3.56</td>
<td>0.003</td>
<td>5.96</td>
</tr>
<tr>
<td>HLA-DMA</td>
<td>-3.79</td>
<td>0.002</td>
<td>4.16</td>
<td>0.0008</td>
<td>4.27</td>
</tr>
<tr>
<td>HLA-DRB5</td>
<td>-2.04</td>
<td>0.06</td>
<td>5.29</td>
<td>8e-5</td>
<td>4.04</td>
</tr>
</tbody>
</table>

(see HLA-DMB, HLA-DRB1, CD74, HLA-DMA, HLA-DRB5) showed coordinated hypomethylation and overexpression in blood endothelial relative to lymphatic cells (Fig. 3 & Table 3). We have presented a novel algorithm for integrative functional supervised analyses of Illumina 450k DNA methylation and gene expression data. By applying and testing it in two different biological contexts, we have here demonstrated 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 (TGFBI1) implicated in the PGR pathway. Our data suggests that hypermethylation mediated silencing of TGFBI1 could also lead to downregulation of the PGR tumour suppressor pathway, since TGFBI1 (HIC5) is a known co-activator of PGR (Aghajanova et al., 2009). Importantly, the TGFBI1 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 blood endothelial cells. 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 DNA methylation 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 TGBRII 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 since 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 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 very 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 due to 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 1st Exon and TSS1500 regions, but this resulted in very similar modules (Supp. Table S6 & S7). An alternative explanation for why rich FEMs are scarce could be biological. For instance, most of the genes in the inferred FEMs are characterised 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 neighboring interacting genes. As shown here, FEM identified many modules exhibiting these type 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 DNA methylation 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 DNA methylation (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, since 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 (GSEA) method.

In summary, the FEM algorithm presented here will be useful to a growing number of studies which aim to identify gene modules or molecular pathways which 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.

GLOSSARY

TGFB1I1

TSS: transcription start site; DNAm: DNA methylation; UTR: untranslated region; FEM: functional epigenetic module; TCGA: The Cancer Genome Atlas; POI: Phenotype of interest.

ACKNOWLEDGEMENT

Funding YJ and AET acknowledge support from the Chinese Academy of Sciences, the Shanghai Institute for Biological Sciences and the Max-Planck Gesellschaft.

REFERENCES


