Published in final edited form as:

Epigenomics. 2014 ; 6(5): 463–467. doi:10.2217/epi.14.44.

The era of integrative genomics: more data or better methods?

Devin C Koestler[‡],

Department of Biostatistics, University of Kansas Medical Center, 3901 Rainbow Blvd, Kansas City, KS 66160, USA, Tel.: +1 913 588 4788

Meaghan Jones[‡], and

Centre for Molecular Medicine & Therapeutics, Department of Medical Genetics, Child & Family Research Institute, The University of British Columbia, Vancouver, BC, Canada, V5Z 4H4

Michael Kobor

Centre for Molecular Medicine & Therapeutics, Department of Medical Genetics, Child & Family Research Institute, The University of British Columbia, Vancouver, BC, Canada, V5Z 4H4

Keywords

DNA methylation; epigenetic silencing; integrative genomics; protein–protein interaction (PPI) networks

In this commentary we highlight a recently proposed statistical methodology called Functional Epigenetic Modules for the integrative analysis of DNA methylation and gene expression data. We begin by providing a high-level overview of this technique, along with a discussion of the perceived strengths and limitations of this methodology. This serves as the jumping off point for dialogue on some of the frequently encountered challenges in the analysis of data from integrative ‘omic studies; particularly those involving the collection of both DNA methylation and gene expression data. We shed light on these challenges and offer our perspective on potential solutions.

Integrating ‘omics data

According to PubMed, published research studies involving integrated analysis of multiple ‘omic data types have witnessed a nearly 10-fold increase over the past decade. The reasons for the surge in integrative genomics studies are both biological – fueled by an increasing understanding that the development and progression of complex diseases are due to the

Correspondence to: Devin C Koestler.

[‡]Authors contributed equally

For reprint orders, please contact: reprints@futuremedicine.com

Financial & competing interests disclosure

This work was supported by a CTSA grant from NCATS awarded to the University of Kansas Medical Center for Frontiers: The Heartland Institute for Clinical and Translational Research # KL2TR000119 (DCK) and a Mining for Miracles post doctoral fellowship from the Child and Family Research Institute, University of British Columbia (MJ). MS Kobor is a Senior Fellow of the Canadian Institute for Advanced Research and a Canada Research Chair in Social Epigenetics. The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in this manuscript.

No writing assistance was utilized in the production of this manuscript.

confluence of alterations in the genome, epigenome, transcriptome, proteome, etc. (and their complex interactions) – and logistical – while large-scale integrative ‘omic studies were once economically infeasible, the declining cost of high-throughput technologies for interrogating the ‘ome have opened the door for such studies to become a reality. Leveraging multiple ‘omic data types for improving our understanding complex disease cannot be better exemplified than through the efforts of The Cancer Genome Atlas (TCGA), which since 2005 has sought to comprehensively catalog the entire spectrum of genomic and proteomic changes involved in human cancers [1]. However, procurement of multiple ‘omic data types is only the beginning of a continuum that involves the transformation of such data into information that improves our capacity to better diagnose, treat and prevent complex human diseases. While the assimilation of various types of ‘omic data provides the raw material for improved understanding, the ‘more data’ philosophy that is at the core of integrative genomics studies has the tendency to detract from the significance of statistical/bioinformatics tools that facilitate the transformation of this raw material into new knowledge. Hence, with the rapid increase in studies involving the collection of multiple ‘omic data types comes the companion need for novel and innovative statistical/bioinformatics tools for integrated analyses. Thus, left with the question of ‘more data’ or ‘better methods’, the truth is that both are necessary elements in our ability to better understand the molecular basis of complex disease and, ultimately, improve human health. Here however, our objective is to emphasize the importance and offer our perspective on what is needed moving forward to develop ‘better methods’ for integrative genomics studies.

In this commentary we focus our attention on a recently published article by Jiao *et al.*, which describes a systems-level framework for the integrative analysis of DNA methylation and gene expression data [2]. Our choice to focus on this specific technique as the basis of this commentary is twofold. First, we aim to highlight what we believe is a novel and innovative methodology for the integrated analysis of DNA methylation and gene expression data that exemplifies the harmonization of the fields of bioinformatics and statistics. Second, as the aforementioned article was published within the last year, it stands as a representative example of current practices in the joint analysis of gene expression and DNA methylation data and represents an ideal case on which to initiate discussion on considerations and challenges that arise in integrative genomic studies. Our contention is that the path to developing better statistical/bioinformatics tools for integrative genomics studies is best served by first understanding what has been done; assessing of the assumptions of these approaches and the extent to which they align with our current understanding of the underlying biology; and finally based on this information, what the necessary steps are for moving forward.

Functional Epigenetic Module algorithm

The study builds off previous work from their group [3,4], in which they demonstrated the feasibility of integrating array-based DNA methylation data with protein–protein interaction (PPI) networks for identifying signaling pathways undergoing age-associated changes in DNA methylation. Here, Jiao *et al.* proposed a method called Functional Epigenetic Modules (FEM) for the integrated analysis of DNA methylation data assayed using the

Illumina Infinium Human Methylation450 BeadChip and gene expression data generated using one of several possible platforms such as RNA-seq, Illumina BeadChips, Affymetrix arrays, for example [2]. The FEM methodology uses PPI networks [5] as the backbone for identifying sub-networks of genes that are epigenetically and functionally dis-regulated based on some phenotype or exposure of interest. Broadly speaking, FEM can be distilled into two main parts: computation of edge weights for connected genes in the PPI network, where the weights are a composite measure of each gene's strength of association between both gene expression and DNA methylation and the phenotype of interest; and identification of sub-networks of genes where the average weight density is significantly larger than the rest of the network. Algorithmically, FEM involves the following five steps:

- Subset the data to consist of the set of G genes that overlap between the gene expression data, DNA methylation data and genes represented in the PPI network.
- Summarize DNA methylation information at the gene level by computing the average methylation of CpG sites mapping to within 200 bp of the transcription start site (TSS200); if there are no probes mapping to within 200 bp of the transcription start site, compute the average methylation of CpGs mapping to within the 1st exon of the gene; if there are no probes mapping to within the 1st exon of the gene, compute the average methylation of CpGs mapping to within 1500 bp of the TSS (TSS1500).
- Record the test statistics, $t_g^{(R)}$ and $t_g^{(D)}$, obtained from testing the association between both gene expression and DNA methylation, respectively, and the phenotype of interest for each of the G genes.
- Create a composite test statistic for each gene t_g , $g = 1, 2, \dots, G$ that is a function of both the gene expression and DNA-methylation-based test statistics generated in step 3. For genes exhibiting anticorrelation between gene expression and DNA methylation (i.e., $\text{sign}(t_g^{(R)}) \neq \text{sign}(t_g^{(D)})$), composite test statistics are taken to be the absolute difference of the DNA methylation- and gene expression-based test statistics (i.e., $t_g = |t_g^{(D)} - t_g^{(R)}|$); otherwise $t_g = 0$ if $\text{sign}(t_g^{(R)}) = \text{sign}(t_g^{(D)})$. Weights between connected genes, gene g and gene h , in the PPI are then taken to be the average of the composite test statistics for those two genes (i.e., $w_{gh} = 1/2(t_g + t_h)$).
- Scan the PPI network using a version of the spin-glass algorithm [6] to identify subnetworks where the average weight density of connected genes is significantly larger than the rest of the network.

Since this approach enables the identification of epi-genetically disregulated gene sub-networks within a PPI network, FEM is likely to yield more biologically relevant and reproducible findings in comparison to simple locus-by-locus DNA methylation analyses, which have been cornerstone of most Epigenome Wide Association Analyses (EWAS) [7,8]. Following the description of the FEM methodology, the authors proceeded by testing and validating this technique in two different biological contexts using previously published data [2]. In an endometrial cancer application involving DNA methylation and RNA-seq on 118 endometrial cancer samples and 17 matched normal endometrial samples, FEM successfully

identified a previously identified epigenetically dysregulated sub-network centered around HAND2, a gene known to mediate the tumor suppressor effects of the progesterone receptor pathway [3]. It also identified 16 previously unidentified novel dysregulated modules. In an endothelial cell differentiation application involving DNA methylation and Agilent gene expression data on 16 blood and lymphatic endothelial cells, FEM was capable of identifying well-established markers of endothelial cell types, including a major histocompatibility complex gene sub-network.

Beyond the potential of FEM to become a valuable addition to the EWAS analysis toolbox, there are several other aspects of this work that deserve to be highlighted. A major strength of FEM is that it does not require matched gene expression and DNA methylation data (i.e., collected from the same biological specimens). Hence, FEM is applicable in scenarios in which DNA methylation and gene expression measurements are from entirely different studies, so long as study populations and biospecimens used for DNA methylation/gene expression profiling are similar between studies; in other words, same tissue type used for molecular profiling, patient characteristics are comparable between the two studies, etc. This is particularly noteworthy as barriers may preclude collection of matched gene expression and DNA methylation data within a single study (i.e., cost constraints, insufficient biological material), allowing researchers to turn to the wealth of publicly available genomic data deposited in repositories such as: the Gene Expression Omnibus [9], ArrayExpress [10], European Genome-phenome archive [11] and The Cancer Genome Atlas [1]. The authors should also be applauded for ensuring that the data, models and code used in their research is publicly available; FEM methodology is freely available as an R-package from [21]. As a research community, it is our obligation to encourage this level of transparency to authenticate research findings and conclusions and to promote the streamlined application of new methods and tools for future research studies. Also, while the current FEM algorithm was designed for DNA methylation data assayed using the Illumina Infinium Human Methylation450 BeadChip, which has emerged as the technology of choice for EWAS, theoretically any platform or technology for DNA methylation assessment could be used; for example, whole genome bisulfite sequencing (WGBS) or CHARM assay [12], provided that DNA methylation measurements can be reliably summarized to the gene level. This is important point as the increased coverage and resolution of techniques like WGBS compared with array-based techniques, coupled with their anticipated decrease in cost over time, are likely to result in an increasing number studies in the future that involve the application of WGBS in larger cohorts.

'Omics challenges

While the authors should be commended for their efforts in developing an innovative systems-based framework for the simultaneous analysis of gene expression and DNA methylation data, this study and the authors' solutions illustrate several inherent challenges in integration of different types of 'omic data. For a multitude of reasons, DNA methylation data is particularly difficult to integrate with gene expression. One reason is that CpGs are not organized neatly into genes and are scattered throughout the genome, mostly concentrated in CpG islands that are often, but not always, associated with a gene promoter [13]. In addition, different genomic features have different patterns, relationships and

functions of DNA methylation, such that the level of DNA methylation at a particular region might have a completely different association with the expression of a nearby gene compared with DNA methylation of some other region [14]. For example, while it is well established that in most cases promoter methylation of CpG-Islands is associated with decreased transcription of downstream genes and may have a direct effect on transcription, DNA methylation in the gene body has been observed to show the opposite trend, and is thought to be a consequence of gene expression rather than a cause [15]. In effect, DNA methylation data have characteristics similar to gene expression data in its continuous nature, but its uneven distribution and different functional roles based on context are shared with genetic analyses. For these reasons, it is particularly hard to come up with a ‘one-size-fits-all’ approach based on established gene expression or genetic analysis methods alone for summarizing DNA methylation to the gene level and determining its relationship to the nearby gene; challenges that were explicitly acknowledged by the authors of this study.

To solve this conundrum, the authors proposed the binning of CpGs into three categories based on their genomic relationship to the TSS, as described in step 2 above. This has the advantage of assigning a single DNA methylation value to each gene, simplifying downstream analyses that involve assessing the association between DNA methylation and phenotype on a gene-by-gene basis. The divisions used are common and biologically informed; many studies use 200 and 1500 bp from the TSS as functional units, although few collapse DNA methylation values in those regions down to a single value. The challenge here is that the physical distances between CpG islands and nearby TSSs are not consistent. CpG islands themselves tend to be approximately 1 kb long, but their position relative to the TSS can result in a large and highly variable gap between the island and the TSS [16]. In disease-free somatic cells, CpG islands are generally unmethylated with low variability, whereas DNA methylation levels in island shores (regions flanking CpG islands) increase rapidly with distance from the island, and exhibit the most variable patterns of DNA methylation across both tissues and individuals [17]. Because of this, the relationship of each island to the bins is vitally important, and could have a major impact on the DNA methylation level in the gene-associated bin. In this study, the hierarchical nature of the binning accounted for some of the differences in island location – since the authors used the sites within 200 bp, if present, and only if none were present did they move outward, presumably they have captured at least some of each promoter-associated island or island shore. However, this could vary widely from gene to gene. For example, if the CpGs within 200 bp of the TSS of gene A are all in an island, those of gene B are not and those of gene C are in an island shore, the levels of DNA methylation, their variability across tissues and individuals and their relationships to gene expression will differ. With this in mind, it would have been interesting if this study had presented a comparison of the genomic composition of each bin to determine whether the differences in association between DNA methylation and gene expression between the bins was due in part to differences in CpG density. One possible alternative method of collapsing DNA methylation values to a gene level would be to use a method like BumpHunting [18] to determine regions of the genome that show similar and consistent patterns of DNA methylation. This would help ensure the similarity of methylation levels within a region and may lead to more stable results when computing a summary measure of DNA methylation within those regions.

After computing the gene-specific associations between DNA methylation and gene expression and the phenotype of interest, another point of consideration involves the nature of the relationships between DNA methylation and gene expression. The strategy used by the authors for computing composite test statistics for each gene is given in step 4 above.

Provided that $t_g^{(R)}$ and $t_g^{(D)}$ have opposite signs (i.e., indicative of an inverse correlation between DNA methylation and gene expression), the composite test statistic for a given gene g , t_g is proportional to the strength of association between gene expression and DNA methylation and the phenotype, as reflected by $t_g^{(R)}$ and $t_g^{(D)}$, respectively. That is, t_g is large when either or both $|t_g^{(R)}|$ and $|t_g^{(D)}|$ are large, indicating strong associations with the phenotype. On the other hand, in instances where $t_g^{(R)}$ and $t_g^{(D)}$ are of the same sign (i.e., indicative of a positive correlation between DNA methylation and gene expression) the composite test statistic is set to zero or some very small value to avoid edges in the connected network with zero weight. Although the motivation for the later stems from observations that DNA methylation in the TSS200, 1st exon and TSS1500 is normally anticorrelated with gene expression, this has the effect of downweighting connections that involve genes exhibiting a positive correlation between DNA methylation and gene expression, and in doing so reduces the likelihood of identifying subnetworks that contain those genes. Across all genes within an individual, the relationship between gene expression and DNA methylation does tend to be negative. When examining a single gene across individuals, however, the relationship can be negative, positive or nonexistent [14,19,20]. Thus, while most genes show the expected – increased DNA methylation results in decreased gene expression – some genes show the opposite pattern, and some show no pattern at all. Therefore, in current FEM formulation, potentially interesting subnetworks may have been missed because some of the genes do not exhibit the common negative relationship between DNA methylation and gene expression. Taking these differing relationships into account could then increase the number of potentially important and interesting subnetworks identified via FEM. How to do this effectively however, remains an open research question.

Network analysis of DNA methylation data

Although PPI networks formed the scaffold on which the FEM algorithm was based, it can easily be extended to other types of networks, for example: transcription factor, co-expression, miRNA, genetic interaction, functional interaction networks or even disease-, tissue- or developmental stage-specific PPI networks. It will be an important decision, then, to choose the particular network based on the unique aims and objectives of a given study. Different types of networks could reveal very different patterns in the data, which is essentially a snapshot in a single time point. PPI networks like the one examined in this study show downstream effects of the current state – which pathways and processes are most affected by the disease or exposure, and thus what the outcomes are likely to be. Transcription factor networks, on the other hand, could give insight into the upstream effects that resulted in the current gene expression and DNA methylation patterns. For cancer analyses like the one described, PPI networks are a logical choice, since discovery of commonly dis-regulated pathways could lead to innovations in drug targeting or treatment.

On the other hand, for disease or environmental exposures, transcription factor networks might represent a more appropriate choice. If a particular transcription factor is mis-regulated by the disease or exposure, it could affect expression and DNA methylation of a number of its targets.

Conclusion & future perspective

In order to properly put the wealth of data arising from epigenomic studies into context, it is of utmost importance that we continue to emphasize the need for ‘more data’, integrating epigenomic and other types of ‘omics data. This will require the development of new tools and ‘better methods’ that reflect our evolving understanding of these biological mechanisms and their interrelationships. Here we have focused on one such new tool, FEM, an innovative and novel application of network analysis for the integration of gene expression and DNA methylation data. The promise of FEM for aiding in the discovery of epigenetically dysregulated hotspots and the freely available software implementation of this methodology underscore the potential of FEM to become a key resource for integrative ‘omics studies. This approach serves to highlight some of the challenges in integrating DNA methylation and gene expression data in particular, and common assumptions that are made in integrative analyses involving these two data types. These assumptions, however, come with their own limitations that will likely be re-visited and revised as the field continues to grow.

References

1. The Cancer Genome Atlas. www.cancergenome.nih.gov
2. Jiao Y, Widschwendter M, Teschendorff AE. A systems-level integrative framework for genome-wide DNA methylation and gene expression data identifies differential gene expression modules under epigenetic control. *Bioinformatics*. 2014 pii:btu316. (Epub ahead of print).
3. Jones A, Teschendorff AE, Li Q, et al. Role of DNA methylation and epigenetic silencing of HAND2 in endometrial cancer development. *PLoS Med*. 2013; 10(11):e1001551. [PubMed: 24265601]
4. West J, Beck S, Wang X, Teschendorff AE. An integrative network algorithm identifies age-associated differential methylation interactome hotspots targeting stem-cell differentiation pathways. *Scientific Rep*. 2013; 3:1630.
5. Cerami EG, Gross BE, Demir E, et al. Pathway commons, a web resource for biological pathway data. *Nucleic Acids Res*. 2011; 39:D685–D690. [PubMed: 21071392]
6. Reichardt J, Bornholdt S. Statistical mechanics of community detection. *Phys Rev E Stat Nonlin Soft Matter Phys*. 2006; 74(1 Pt 2):016110. [PubMed: 16907154]
7. Michels KB, Binder AM, Dedeurwaerder S, et al. Recommendations for the design and analysis of epigenome-wide association studies. *Nat Methods*. 2013; 10(10):949–955. [PubMed: 24076989]
8. Wilhelm-Benartzi CS, Koestler DC, Karagas MR, et al. Review of processing and analysis methods for DNA methylation array data. *Br J Cancer*. 2013; 109(6):1394–1402. [PubMed: 23982603]
9. Barrett T, Troup DB, Wilhite SE, et al. NCBI GEO: archive for high-throughput functional genomic data. *Nucleic Acids Res*. 2009; 37:D885–D890. [PubMed: 18940857]
10. Parkinson H, Kapushesky M, Kolesnikov N, et al. ArrayExpress update--from an archive of functional genomics experiments to the atlas of gene expression. *Nucleic Acids Res*. 2009; 37:D868–D872. [PubMed: 19015125]
11. Leinonen R, Akhtar R, Birney E, et al. The European Nucleotide Archive. *Nucleic Acids Res*. 2011; 39:D28–D31. [PubMed: 20972220]

12. Irizarry RA, Ladd-Acosta C, Carvalho B, et al. Comprehensive high-throughput arrays for relative methylation (CHARM). *Genome Res.* 2008; 18(5):780–790. [PubMed: 18316654]
13. Illingworth RS, Bird AP. CpG islands – 'a rough guide'. *FEBS Lett.* 2009; 583(11):1713–1720. [PubMed: 19376112]
14. Gutierrez-Arcelus M, Lappalainen T, Montgomery SB, et al. Passive and active DNA methylation and the interplay with genetic variation in gene regulation. *Elife (Cambridge).* 2013; 2:e00523.
15. Jones PA. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat Rev Genet.* 2012; 13(7):484–492. [PubMed: 22641018]
16. Weber M, Hellmann I, Stadler MB, et al. Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. *Nat Genet.* 2007; 39(4):457–466. [PubMed: 17334365]
17. Irizarry RA, Ladd-Acosta C, Wen B, et al. The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. *Nat Genet.* 2009; 41(2):178–186. [PubMed: 19151715]
18. Jaffe AE, Murakami P, Lee H, et al. Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies. *Int J Epidemiol.* 2012; 41(1):200–209. [PubMed: 22422453]
19. Lam LL, Emberly E, Fraser HB, et al. Factors underlying variable DNA methylation in a human community cohort. *Proc Natl Acad Sci USA.* 2012; 109 (Supl 2):17253–17260. [PubMed: 23045638]
20. Jones MJ, Fejes AP, Kobor MS. DNA methylation, genotype and gene expression: who is driving and who is along for the ride? *Genome Biol.* 2013; 14(7):126. [PubMed: 23899167]
21. Sourceforege.net. <http://sourceforege.net/projects/funepimod>