Journal of Biometrics & Biostatistics
ISSN: 2155-6180

Wang and Liu, J Biomet Biostat 2013, 4:1
DOI: 10.4172/2155-6180.1000e123

Research Article Open Access

Integrative Analysis of Genome-wide Expression and Methylation Data

Ke-Sheng Wang*and Xuefeng Liu

Department of Biostatistics and Epidemiology, College of Public Health, East Tennessee State University, Johnson City, TN, USA

Abstract

Microarray technology has provided a tool for investigating expression levels and methylation signatures, of thousands of genes simultaneously, in a biological sample. Recent advances in next-generation sequencing and microarray technology make it possible to study genome-wide mRNA expression and DNA methylation profiles at a high resolution and in a large number of samples. However, integration of expression and methylation data turns to be a challenge. First, we briefly introduce the gene expression and methylation technology, data structure, and basic statistical methods. Furthermore, we review recent advances in integrative analysis of expression and methylation data. In addition, we stress some future directions.

Keywords: Genome; Expression; Epigenetics; Methylation; Sequencing; Statistics

Expression Data

Microarray technology has provided a tool for monitoring gene expression for tens of thousands of genes. A microarray is a glass slide with tens of thousands of spots on an array. Single-stranded DNA molecules are attached at fixed spots and each spot is related to a single gene. One of the most popular experimental platforms is used for comparing mRNA abundance in two different samples (or a sample and a control). The raw microarray data are images, which have to be transformed into gene expression matrices-tables where rows represent genes and columns represent various samples. For most microarray technology platforms, one measure of gene expression is the ratio of background-subtracted signals of the given sample and the control. For example, it is assumed that abundance ratios of 1.5-2 are indicative of a change in gene expression [1].

A simple microarray experiment may be conducted to compare the expression differences between two conditions. The simplest method is to evaluate the log ratio between two conditions and consider all genes that differ by more than an arbitrary cut-off value to be differentially expressed. For example, a two-fold difference has been used as a cut-off. Genes are considered as being different in expression if the expression under one condition is over two-fold greater or less than that under the other condition. *t*-test is commonly used to detect differential expression in comparison of two conditions. With more than two conditions, analysis of variance (ANOVA) has been used. The mixed ANOVA model is a general and powerful approach for microarray experiments with multiple factors and/or several sources of variation [2].

RNA sequencing (RNA-Seq) makes it possible to survey an entire transcriptome at single-base resolution and construct genome-wide gene expression profiles [3]. Global ChIP-chip and ChIP-seq analyses have accelerated the pace of discovery, allowing insights unattainable with other methods, and have played an important role in identifying the elongation phase of transcription as a critical point of biological regulation [4]. RNA-Seq has provided a powerful tool for detecting differential gene expression with both high-throughput and high resolution capabilities. Next-generation sequencing (NGS) provides a better approach to gene expression profiling [5].

DNA methylation Data

DNA methylation is one of the most important mechanisms of epigenetic regulation in eukaryotes. It occurs most frequently at cytosines that are followed by guanines (CpG). It has been shown that high levels of DNA methylation in promoter regions are typically associated with robust gene silencing while DNA methylation has normal function in embryonic development, X-chromosome inactivation, and genomic imprinting [6]. It is firmly established that hypermethylation of CpG islands located in the promoter regions of tumor suppressor genes is one of the most common mechanisms for gene regulation in cancer [7]. Interestingly, aberrant DNA methylation has been seen in a variety of human diseases ranging from neurological and autoimmune disorders to cancer [8]. It has been predicted that DNA methylation may provide a lifetime record of environmental exposures and a useful source of biomarkers for risk stratification and disease diagnostics [9,10].

The data structure for DNA methylation is generally dichotomous (methylated or unmethylated) or reported as a ratio between a methylation reaction and a neutral control reaction [11]. To date, two methods have been proposed to measure the methylation level. The first one is called Beta-value, ranging from 0 to 1, which has been widely used to measure the percentage of methylation. The Beta-value is the ratio of the methylated probe intensity and the overall intensity (sum of methylated and unmethylated probe intensities). Under ideal conditions, a value of zero indicates that all copies of the CpG site in the sample are completely unmethylated (no methylated molecules were measured) while a value of one indicates that every copy of the site is methylated. If we assume the probe intensities are Gamma distributed, then the Beta-value follows a Beta distribution. For this reason, it has been named the Beta-value. The second method is the log2 ratio (M-value) of the intensities of methylated probe versus unmethylated probe. The M-value has been widely used in expression microarray analysis, especially two-color microarray analysis. Therefore, many existing microarray statistical frameworks using an M-value method

***Corresponding author:** Ke-Sheng Wang, Department of Biostatistics and Epidemiology, College of Public Health, East Tennessee State University, PO Box 70259, Lamb Hall, Johnson City, TN 37614-1700, USA, Tel: +1-423 439 4481; Fax: +1-423 439 4606; E-mail: wangk@etsu.edu

Received December 19, 2012; **Accepted** December 21, 2012; **Published** December 29, 2012

Citation: Wang KS, Liu X (2013) Integrative Analysis of Genome-wide Expression and Methylation Data. J Biomet Biostat 4:e123. doi:10.4172/2155-6180.1000e123

can also be applied to methylation data analysis. The two methods are related by a Logit transformation [12].

For discrete data, chi-square test can be used for independent samples while McNemar's test or Cochran's test can be used for data correlated samples analysis. For continuous traits, t-test or ANOVA can be used for normal distributed samples, whereas for non normal traits, rank-based tests can be used [11].

Recent advances in NGS and microarray technology make it possible to map DNA methylation genome-wide, at a high resolution and in a large number of samples [13]. Genome-wide DNA methylation has been mapped with one of the three most commonly used assays, resulting in methylation-specific DNA sequencing or microarray data [10]: 1) Bisulphite sequencing: DNA treatment with bisulphite specifically introduces mutations at unmethylated Cs and these mutations are mapped by NGS, 2) Bisulphite microarrays: DNA-methylation-specific mutations are introduced by bisulphite treatment and these mutations are mapped using a genotyping microarray that covers a selection of Cs, 3) Enrichment-based methods: Methylated (alternatively, unmethylated) DNA fragments are enriched in a DNA library and the library composition is quantified by NGS.

Integrative Analysis of Genome-wide Methylation and Expression Data

The advent of global DNA methylation arrays and next-generation RNA sequencing transcriptome studies have made it possible to explore the global relationship between gene methylation and expression during cell development and tissue differentiation [3].

Generally, methylation of regulatory CpG islands is thought to down regulate transcription by promoting the formation of heterochromatin and preventing the binding of transcription factors [14]. For example, spearman rank correlations were used to assess the relationship between methylation data using the Illumina Human Methylation 27 DNA Analysis Bead Chip assay (Beta-value and the log of the ratio) and gene expression data using the RNA-sequencing data, which is presented as the number of GC-corrected reads mapping to a gene in an individual, divided by the length of the gene. Significant negative correlations were found between promoter methylation and gene expression levels [15].

However, one large-scale study has failed to demonstrate a significant relationship between genome-wide methylation and gene expression [16]. Furthermore, there exists a weak negative correlation between DNA methylation at promoter regions and gene expression, which implies that the relation between alterations in DNA methylation at promoter region and gene expression is gene-specific [17].

Interestingly, it has been observed that the relationship between DNA methylation and transcription is bidirectional based on single loci cancer studies, where transcription apparently causes intragenic CGI methylation in addition to CGI promoter methylation inhibiting transcription [18]. A recently integrative analysis of methylation from normal human heart, lung, and kidney using the Illumina Infinium 27 K methylation arrays and gene expression by RNA sequencing shows that gene methylation and its transcriptional levels are comprehensively correlated; however, DNA methylation has been found either positively or negatively to be correlated with gene expression [19].

Another study investigates the relationship between genetic variation, DNA methylation using Illumina Human Methylation 27 bead chips and gene expression generated on Illumina H12 bead chip in a sample of 148 healthy subjects. To determine whether a significant association exists between expression and methylation levels they use a multivariate linear regression model for regressing the gene expression level (dependent variable) on the methylation level (independent variable) with age and gender as covariates. The authors have found both negative (DNA methylation levels and gene expression levels in opposite direction) and positive (DNA methylation levels and gene expression levels in same direction) associations between *cis*-acting DNA methylation probes and corresponding gene expression levels, confirming previous reports that DNA methylation and gene expression located within a *cis*-region can be both positively and negatively associated. In addition, a structural equation model based analysis has strong support in particular for a traditional causal model in which gene expression is regulated by genetic variation via DNA methylation instead of gene expression affecting DNA methylation levels [20].

A more recently study conducted genome-wide expression analyses using the Affymetrix Human Genome U133 2.0 plus array (which contains 29,098 gene specific oligonucleotide probes) and compared two genome-wide methylation assays: Nimble Gen 385K Ref Seq Whole Genome Promoter Array and the Illumina Human Methylation 450 Bead Chip. Pearson's correlation coefficient and Spearman's rank-order correlation coefficients were used to compare the values from each methylation platform to genome-wide mRNA expression. To further elucidate the sources of variance, the authors performed an ordinary least squares (OLS) regression, using both the Illumina and the NimbleGen array to simultaneously predict gene expression. They found weak negative correlations between gene expression and DNA methylation within individuals across all queried loci. However, the current findings have certain limitations including the small number of samples examined [21].

Future Direction

To understand and identify the molecular mechanisms that underpin complex disorders, it has been suggested that NGS-based approaches that integrate genomewide, high-resolution, quantitative, and allelic epigenetic, genetic and transcriptomic information will be essential [22,23]. Furthermore, to better understand human development and health, novel methods to integrate data of different types (genetic, epigenetic, RNA and protein expression will be required [24]. Future research will focus on lookup tables for associations between methylation, gene expression, and genotype, as well as methylome and transcriptome modules [20]. In short, integration of gene expression, genetic and epigenetic studies may identify novel therapeutic targets and strategies for the treatment of complex diseases. Future studies will focus on integrative analysis of genome-wide expression and epigeneitcs data from high-solution NGS and microarray technology, and genome-wide genetic data (SNP data and copy number variations) using large samples and advanced statistical methods.

References

1. Brazma A, Vilo J (2000) Gene expression data analysis. FEBS Lett 480: 17-24.
2. Cui X, Churchill GA (2003) Statistical tests for differential expression in cDNA microarray experiments. Genome Biol 4: 210.
3. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 5: 621-628.
4. Gilchrist DA, Fargo DC, Adelman K (2009) Using ChIP-chip and ChIP-seq to study the regulation of gene expression: genome-wide localization studies reveal widespread regulation of transcription elongation. Methods 48: 398-408.
5. Robles JA, Qureshi SE, Stephen SJ, Wilson SR, Burden CJ, et al. (2012) Efficient experimental design and analysis strategies for the detection of differential expression using RNA-Sequencing. BMC Genomics 13: 484.

6. Bird A (2002) DNA methylation patterns and epigenetic memory. Genes Dev 16: 6-21.

7. Herman JG, Baylin SB (2003) Gene silencing in cancer in association with promoter hypermethylation. N Engl J Med 349: 2042-2054.

8. Portela A, Esteller M (2010) Epigenetic modifications and human disease. Nat Biotechnol 28: 1057–1068.

9. Walker CL, Ho SM (2012) Developmental reprogramming of cancer susceptibility. Nat Rev Cancer 12: 479–486.

10. Bock C (2012) Analysing and interpreting DNA methylation data. Nat Rev Genet 13: 705-719.

11. Siegmund KD, Laird PW (2002) Analysis of complex methylation data. Methods 27: 170-178.

12. Du P, Zhang X, Huang CC, Jafari N, Kibbe WA, et al. (2010) Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. BMC Bioinformatics 11: 587.

13. Laird PW (2010) Principles and challenges of genome-wide DNA methylation analysis. Nat Rev Genet 11: 191-203.

14. Suzuki MM, Bird A (2008) DNA methylation landscapes: provocative insights from epigenomics. Nat Rev Genet 9: 465-476.

15. Bell JT, Pai AA, Pickrell JK, Gaffney DJ, Pique-Regi R, et al. (2011) DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. Genome Biol 12: R10.

16. Fan S, Zhang X (2009) CpG island methylation pattern in different human tissues and its correlation with gene expression. Biochem Biophys Res Commun 383: 421-425.

17. Jeong J, Li L, Liu Y, Nephew KP, Huang TH, et al. (2010) An empirical Bayes model for gene expression and methylation profiles in antiestrogen resistant breast cancer. BMC Med Genomics 3: 55.

18. Jones P (1999) The DNA methylation paradox. Trends Genet 15: 34–37.

19. Xie L, Weichel B, Ohm JE, Zhang K (2011) An integrative analysis of DNA methylation and RNA-Seq data for human heart, kidney and liver. BMC Syst Biol 5: S4.

20. van Eijk KR, de Jong S, Boks MP, Langeveld T, Colas F, et al. (2012) Genetic analysis of DNA methylation and gene expression levels in whole blood of healthy human subjects. BMC Genomics 13: 636.

21. Plume JM, Beach SR, Brody GH, Philibert RA (2012) A cross-platform genome-wide comparison of the relationship of promoter DNA methylation to gene expression. Front Genet 3: 12.

22. Tycko B (2010) Allele-specific DNA methylation: beyond imprinting. Hum Mol Genet 19: R210–R220.

23. Meaburn EL, Schalkwyk LC, Mill J (2010) Allele-specific methylation in the human genome: implications for genetic studies of complex disease. Epigenetics 5: 578–582.

24. Siegmund KD (2011) Statistical approaches for the analysis of DNA methylation microarray data. Hum Genet 129: 585-595.