Genes Nutr (2008) 3:167–171
DOI 10.1007/s12263-008-0095-0

RESEARCH PAPER

Bioinformatic interrogation of expression array data to identify nutritionally regulated genes potentially modulated by DNA methylation

J. A. McKay · M. E. Adriaens · D. Ford · C. L. Relton · C. T. A. Evelo · J. C. Mathers

Received: 8 October 2008 / Accepted: 12 November 2008 / Published online: 26 November 2008

Abstract DNA methylation occurs at CpG dinucleotide sites within the genome and is recognised as one of the mechanisms involved in regulation of gene expression. CpG sites are relatively underrepresented in the mammalian genome, but occur densely in regions called CpG islands (CGIs). CGIs located in the promoters of genes inhibit transcription when methylated by impeding transcription factor binding. Due to the malleable nature of DNA methylation, environmental factors are able to influence promoter CGI methylation patterns and thus influence gene expression. Recent studies have provided evidence that nutrition (and other environmental exposures) can cause altered CGI methylation but, with a few exceptions, the genes influenced by these exposures remain largely unknown. Here we describe a novel bioinformatics approach for the analysis of gene expression microarray data designed to identify regulatory sites within promoters of differentially expressed genes that may be influenced by changes in DNA methylation.

Keywords Bioinformatics · CpG islands · DNA methylation · Gene expression · In silico promoter analysis · Transcription factor binding sites

J. A. McKay (✉) · D. Ford · C. L. Relton · J. C. Mathers
Human Nutrition Research Centre, Medical School, Newcastle University, Framlington Place, Newcastle NE2 4HH, UK
e-mail: jill.mckay@ncl.ac.uk

M. E. Adriaens · C. T. A. Evelo
Department of Bioinformatics, BiGCaT, Maastricht University, Universiteitssingel 50, 6229 ER Maastricht, The Netherlands

Introduction

DNA is methylated by the covalent addition of methyl groups to the 5′ position on cytosine residues, usually when the cytosine is followed by a guanine residue—i.e. in a CpG dinucleotide. Although CpG dinucleotides are underrepresented in the genome, there are dense accumulations of CpGs (CpG islands; CGI) in the promoter regions of many genes. When CpG dinucleotides in CGI located near the transcription start site of a gene are methylated, this is usually associated with gene repression. Other epigenetic marks, especially post-translational modifications of histone tails, also contribute to regulation of gene expression. Therefore the pattern of CpG methylation impacts upon phenotype by altering gene expression. For example, in the agouti mouse methylation of specific CpG sites within the intra-cisternal A particle (IAP) region of the Agouti gene can influence coat colour, body weight and longevity [20]. Aberrant DNA methylation is associated with several diseases, e.g. cancer, occurs during ageing and has been implicated as one possible mechanism involved in the developmental origins of adult health and disease.

The main patterns of DNA methylation are established during early embryonic and fetal life, but methylation marks are plastic and can be influenced by environmental factors especially when these factors are applied during development [4, 10, 11, 16, 20–22]. Although most research to date has been carried out using animal models, it is likely that environmental factors also influence DNA methylation, and therefore phenotype, in humans. In patients with hyper-homocysteinaemia, genomic DNA methylation is lower than that in controls [7]. Hyperhomocysteinaemia is characterised by increased cellular concentrations of *S*-adenosylhomocysteine (SAH), which is an inhibitor of DNA methyltransferase 1 (DNMT1)—the

enzyme responsible for maintaining DNA methylation. Supplementing the diet of such patients with folate, which provides methyl groups for the synthesis of *S*-adenosylmethionine (SAM; the universal methyl donor) and alters the SAM:SAH ratio, restored genomic DNA levels to normal and also 'corrected' CpG methylation within the *IGF2-H19* locus [7].

A large proportion (70–80%) of global DNA methylation occurs in non-coding regions, exons and repetitive DNA sites within the genome [8] and little is known about the functional consequences of changes in genomic DNA methylation. In contrast, methylation at specific CpG sites in or around the promoter regions of genes can influence transcription so it is imperative to understand which loci within the genome are susceptible to environmentally-determined modification of DNA methylation patterns. A candidate gene approach has been used successfully in studies of cancer aetiology and pathophysiology, where it is reasonable to predict that targets would include tumour suppressor genes since their silencing is of obvious aetiological significance. Such an approach is likely to be less successful for other complex diseases where potential candidate genes are less readily identified. A candidate gene approach is also less appropriate in the context of (relatively mild) nutritional exposures, where the effects on gene expression may be small and widespread across the genome. In addition, since dietary factors can influence gene expression by several other mechanisms [12], the relationship between changes in gene expression and corresponding changes in DNA methylation patterns are difficult to decipher. To help address this problem, we have developed a novel strategy to identify target loci that have shown differential gene expression in response to nutritional exposure, which potentially could be due to altered DNA methylation. To do so, we used a mouse model in which dams were fed a folate deplete diet prior to and during pregnancy, and investigated the effects of folate depletion in utero on fetal liver gene expression at 17.5 days' gestation. In this paper we outline our strategy and describe how we have used bioinformatic tools to analyse these gene expression array data to identify regulatory sites within promoters of differentially expressed genes that could potentially be influenced by differential DNA methylation.

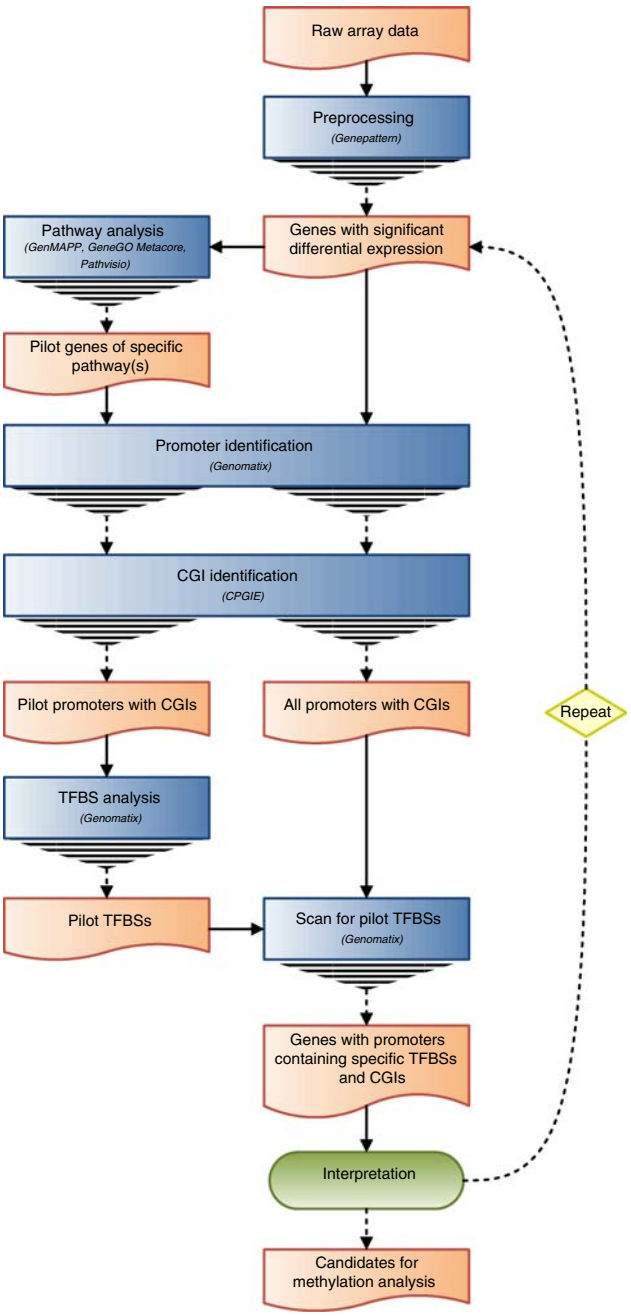

Fig. 1 Overview of strategy developed to identify target genes whose expression may have been altered by DNA methylation aberrations. Genepattern can be accessed through any NuGO Blackbox (NBX). Web address for bioinformatics tools are as follows: Genomatix; http://www.genomatix.de, CpG Island Explorer; http://bioinfo.hku.hk/cpgieintro.html

Strategy

Figure 1 provides an overview of the strategy employed to find gene targets with possible aberrant methylation. Firstly, to identify target genes, microarray data were analysed to identify a list of differentially expressed genes. We used the automatic transcriptomics analysis pipeline that was developed by NuGO (the European Nutrigenomics Organisation) and that is accessible from the NuGO Blackboxes through Genepattern. The pipeline employs several packages from the Bioconductor project (http://www.bioconductor.org) and includes procedures for quality control, normalisation and statistical analysis of microarray data. The resulting annotated gene list was

filtered on fold change and P values to identify genes with altered expression. As (relatively mild) nutritional exposures are likely to lead to subtle changes in gene expression when compared with pharmaceutical interventions, we routinely filter for genes with a 20% change in expression, i.e. either above 1.2 or below −1.2. As we are interested only in genes for which the changes in expression are statistically significant, we filter using a threshold of $P = 0.05$. P values were corrected for multiple testing using the Benjamini and Hochberg [1] approach. From this analysis we obtained a list of genes showing significant differential expression. By mapping these genes to pathways, it was possible to filter the list further by selecting only those genes occurring in pathways of interest, i.e. those pathways plausibly linked biologically to the processes being considered. Examples of useful pathway tools are Pathvisio [18], GenMAPP [15] and GeneGO's Metacore [5]. Pathway analysis can only be as good as the pathway information used. To allow evaluation of gene expression changes in processes directly related to DNA methylation we created a one-carbon metabolism pathway on WikiPathways [13] and used that in PathVisio and GenMAPP.

The next stage was to identify the promoter sequences of the selected genes. The promoter is the site responsible for regulation of gene transcription and houses several regulatory DNA motifs including transcription factor binding sites (TFBS) and, in many cases, CGI. Characterising the promoter is therefore essential to understand the regulatory networks responsible for differential gene expression. Promoter identification within genomic sequences can be difficult, although there are bioinformatic tools that can be used to predict mammalian promoters (reviewed in [24]). To be sure that a predicted promoter is indeed a promoter requires wet-lab verification. We use the validated promoter database of the commercially available Genomatix software (http://www.genomatix.de). Promoter sequences in this database are scored as gold (experimentally verified 5′ complete transcript), silver (transcript with 5′ end confirmed by PromoterInspector prediction) or bronze (annotated transcript, no confirmation for 5′ completeness) and this scoring system can be used to eliminate less likely candidates. At this stage, one may choose to leave out promoters with less relevant transcripts, e.g. those where the functional role of the gene product is unknown.

CpG sites are relatively under-represented in the mammalian genome but occur in unusually dense groupings in CGI. These are areas of the genome between 0.5 and 4 kb in length, with >50% GC content and an observed/expected CpG ratio of over 0.6. In the human genome, roughly half of all genes contain CpG islands. CpG sites within CpG islands tend to be largely unmethylated in normal tissues. Furthermore, hypermethylation of CpG islands within promoters of tumour suppressor genes is observed commonly in tumour cells [6]. Therefore, we used presence of a CGI as one criterion in our strategy to identify potential target genes that may have been regulated by DNA methylation.

There are several freely available web based tools, such as MethPrimer (http://www.urogene.org/methprimer/index1.html) [9] that can predict the presence of a CGI within a given sequence. However, most of these tools can analyse only one sequence at a time. Given that the list of candidate genes arising from array data can be quite large, even after filtering at the pathway level, this approach is unnecessarily laborious and time consuming. This problem can be addressed by using CpG Island Explorer (http://bioinfo.hku.hk/cpgieintro.html) [19], which is freely available for download and is able to predict the presence of CGI in multiple sequences simultaneously. It is based on the algorithm and Perl script created by Takai and Jones [17], which is still considered as the gold standard in CpG island searching.

Identification of methylation-sensitive TFBS

Hypermethylation of CGI is associated with gene silencing and this is generally associated with a 'closed' chromatin structure that reduces access of transcription factors and other transcriptional machinery to the DNA. However, it is also known that methylation of single specific CpG sites can be associated with decreased expression of some genes [2, 14]. In some cases, methylation of these specific sites prevents binding of transcription factors essential for gene expression. Therefore part of our strategy to identify genes whose expression may be altered due to DNA methylation changes was to search for the presence of CpG sites within TFBS of promoters.

Promoter function is governed by the binding of transcription factors to the DNA sequence. An isolated TFBS is often not functional. Werner [23] argued that TFBS act in a modular fashion allowing the support of protein complexes for transcriptional activation. His definition of a transcription factor (TF) module is 'two or more transcription factor binding sites in a defined order and orientation that comprise promoter modules'. If one TFBS within a module is functionally impaired, by a point mutation in the DNA sequence or due to methylation of a CpG site within the binding site, this causes the complete TF module to be inactivated.

We employed Genomatix software, specifically MatInspector, to investigate TFBS within promoters since this software is optimized for searching for matrices of TFBS within sequences [3]. Further analysis of the resultant TF module data list was used to identify modules with TFBS

that contained conserved CpG sites. From the list of modules we identified common TF modules that occurred regularly in our data set. Promoters containing common modules in which TFBS have CpG sites are potential targets for altered DNA methylation in response to a dietary (or other environmental) exposure since it is likely that these genes are regulated by a common mechanism. However, it is important to be aware that it is also likely that the TFs common to these modules could themselves be the cause of altered expression of these genes. One potential way to resolve this issue is to check the original array expression data for changes in expression of the TF in question. If the TF do not display differential expression, then it is more likely that the observed changes in gene expression are due to altered CpG methylation within TFBS.

Summary

We have developed a novel in silico strategy to identify target genes that could potentially be regulated by DNA methylation in response to a dietary (or other) exposure. This strategy utilises the expression data from whole genome transcriptomics arrays together with a comprehensive bioinformatics workflow to narrow the list of gene targets for DNA methylation analysis. This strategy is powerful and attractive in identifying genes that are susceptible to DNA methylation changes, and therefore may be of particular utility in pinpointing common genes susceptible to aberrant DNA methylation using different array datasets from a range of studies. With this approach, it may be possible to identify specific DNA motifs that are susceptible to aberrant DNA methylation in response to environmental (nutritional) factors.

Acknowledgments This paper was produced by the Nutritional Epigenomics Focus Team which is funded by NuGO (the European Nutrigenomics Organisation).

Conflict of interest statement The authors report no conflicts of interest, financial or otherwise.

References

1. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B Stat Methodol 57:289–300
2. Cao YX, Jean JC, Williams MC (2000) Cytosine methylation of an Sp1 site contributes to organ-specific and cell-specific regulation of expression of the lung epithelial gene t1alpha. Biochem J 350(3):883–890
3. Cartharius K, Frech K, Grote K, Klocke B, Haltmeier M, Klingenhoff A, Frisch M, Bayerlein M, Werner T (2005) MatInspector and beyond: promoter analysis based on transcription factor binding sites. Bioinformatics 21(13):2933–2942
4. Dolinoy DC, Weidman JR, Waterland RA, Jirtle RL (2006) Maternal genistein alters coat color and protects Avy mouse offspring from obesity by modifying the fetal epigenome. Environ Health Perspect 114(4):567–572
5. Ekins S, Nikolsky Y, Bugrim A, Kirillov E, Nikolskaya T (2007) Pathway mapping tools for analysis of high content data. Methods Mol Biol 356:319–350
6. Esteller M (2007) Cancer epigenomics: DNA methylomes and histone-modification maps. Nat Rev Genet 8(4):286–298
7. Ingrosso D, Cimmino A, Perna AF, Masella L, De Santo NG, De Bonis ML, Vacca M, D'Esposito M, D'Urso M, Galletti P, Zappia V (2003) Folate treatment and unbalanced methylation and changes of allelic expression induced by hyperhomocysteinaemia in patients with uraemia. Lancet 361(9370):1693–1699
8. Kim Y-I (2005) Nutritional epigenetics: impact of folate deficiency on DNA methylation and colon cancer susceptibility. J Nutr 135:2703–2709
9. Li LC, Dahiya R (2002) MethPrimer: designing primers for methylation PCRs. Bioinformatics 18(11):1427–1431
10. Lillycrop KA, Phillips ES, Jackson AA, Hanson MA, Burdge GC (2005) Dietary protein restriction of pregnant rats induces and folic acid supplementation prevents epigenetic modification of hepatic gene expression in the offspring. J Nutr 135(6):1382–1386
11. Lillycrop KA, Phillips ES, Torrens C, Hanson MA, Jackson AA, Burdge GC (2008) Feeding pregnant rats a protein-restricted diet persistently alters the methylation of specific cytosines in the hepatic PPAR alpha promoter of the offspring. Br J Nutr 100(2):278–282
12. Mathers JC (2006) Candidate mechanisms for interactions between nutrients and genes. In: Choi Sang-Woon, Friso Simonetta (eds) Nutrient–gene interactions in cancer. CRC Press, Taylor & Francis Group, Boca Raton, pp 19–36
13. Pico AR, Kelder T, van Iersel MP, Hanspers K, Conklin BR, Evelo C (2008) WikiPathways: pathway editing for the people. PLoS Biol 6:7
14. Pogribny IP, Pogribna M, Christman JK, James SJ (2000) Single-site methylation within the *p53* promoter region reduces gene expression in a reporter gene construct: possible *in vivo* relevance during tumorigenesis. Cancer Res 60:588–594
15. Salomonis N, Hanspers K, Zambon AC, Vranizan K, Lawlor SC, Dahlquist KD, Doniger SW, Stuart J, Conklin BR, Pico AR (2007) GenMAPP 2: new features and resources for pathway analysis. BMC Bioinform 8(1):217
16. Sinclair KD, Allegrucci C, Singh R, Gardner DS, Sebastian S, Bispham J, Thurston A, Huntley JF, Rees WD, Maloney CA, Lea RG, Craigon J, McEvoy TG, Young LE (2007) DNA methylation, insulin resistance, and blood pressure in offspring determined by maternal periconceptional B vitamin and methionine status. Proc Natl Acad Sci USA 104(49):19351–19356
17. Takai D, Jones PA (2002) Comprehensive analysis of CpG islands in human chromosomes 21 and 22. Proc Natl Acad Sci USA 99(6):3740–3745
18. van Iersel MP, Kelder T, Pico AR, Hanspers K, Coort S, Conklin BR, Evelo C (2008) Presenting and exploring biological pathways with PathVisio. BMC Bioinform 9(1):399
19. Wang Y, Leung FCC (2004) An evaluation of new criteria for CpG islands in the human genome as gene markers. Bioinformatics 20(7):1170–1177
20. Waterland RA, Jirtle RL (2003) Transposable elements: targets for early nutritional effects on epigenetic gene regulation. Mol Cell Biol 23(15):5293–5300

21. Waterland RA, Dolinoy DC, Lin JR, Smith CA, Shi X, Tahiliani KG (2006) Maternal methyl supplements increase offspring DNA methylation at Axin Fused. Genesis 44(9):401–406
22. Waterland RA, Lin JR, Smith CA, Jirtle RL (2006) Post-weaning diet affects genomic imprinting at the insulin-like growth factor 2 (Igf2) locus. Hum Mol Genet 15(5):705–716
23. Werner T (2001) Target gene identification from expression array data by promoter analysis. Biomol Eng 17:87–94
24. Werner T (2003) The state of the art of mammalian promoter recognition. Brief Bioinform 4(1):22–30