A heuristic algorithm to select genes potentially regulated by methylation

Alex Sanchez-Pla [*], Berta Miro

1 Departament of Genetics Microbiology and Statistics, Avda Diagonal 645, Barcelona, 08028
2 Department, Street, City, State, Zip

* Corresponding author: asanchez@ub.edu

Abstract

Methylation is a key process in cancer. Usually it acts by inhibiting the expression of the gene but if methylation is low then any values of expression, high or low, can be found. This suggests that to select genes regulated by methylation one may look for patterns in the relation between gene expression and methylation showing either an L-shape or negative correlation between expression and methylation. We have developed a heuristic algorithm that mimics the process of visually selecting an "L-shape", that is genes that can show a wide range of expression values (low to high) when methylation is low, but only low expressions for intermediate or high methylation. We have compared the method with naïve correlation and, despite not being able to quantify its accuracy -because no dataset with "TRUE" L-shaped genes is available- its performance seems to be very good especially due to its flexibility. The method has been implemented in an R package, "Lheuristic" and a Shiny application, both available from GitHub (http://github.com/alexsanchezpla). Given two matrices -expression and methylation values - with the same row and column names the program offers the possibility to select genes based on either negative correlation, the heuristic algorithm or both methods at once. Once genes have been selected, results can be interactively reviewed, plotted or downloaded.

Introduction and Background

Introduction to methylation

Epigenetic marks modulate gene expression without affecting the DNA nucleotide sequence. These potentially heritable changes are, for example, DNA methylation or histone acetylation ([1]). DNA methylation is the most studied epigenetic process in humans. The process is based on the addition of a methyl group, mostly in CpG dinucleotides. The CpG dinucleotides tend to group in areas of less than 500kb and with higher than 55% C and G content , these regions are named islands; further from the island the region is called shore and further from the shore it is called shelf. More than 60% of promoter regions are associated with CpG islands ([2]) and the methylation of these is linked to gene silencing and gene expression inhibition. DNA methylation has been linked to the regulation of numerous cellular processes, including embryonic development, or X-chromosome inactivation and preservation of chromosome stability among others. DNA methylation has also been observed in autoimmune diseases, metabolic disorders, neurological disorders, and other processes that despite being

natural they are debilitating, like ageing for example; and it can also be correlated with drug or treatment response ([3]; [4]; [5]; [6]). Most research on this area has been, however, focused on tumor repressor genes, which are often silenced in cancer cells due to hypermethylation. This is an important mechanism of gene silencing during tumor progression ([7]). On the contrary, a general level of hypomethylation has been observed in human tumors ([8]); therefore, hypomethylation is a useful mechanism to distinguish genes of some human cancers from their normal counterparts.

In the human genome, about 80% of cytosines in the 56 million CpG sites are methylated to 5-methylcytosines. The methylation pattern of DNA is highly variable among cells types and developmental stages and influenced by disease processes and genetic factors. The relationship between gene expression and methylation has been associated with cancer and extensively studied, therefore it has produced fruitful results ([9]).

Analysis of genes regulatated by methylation

With the abundance of emerging evidence indicating the important role of DNA methylation in common diseases, researchers have attempted to use DNA methylation as a biomarker to identify epigenetic changes that are associated with disease status. While the genetic events that drive the tumorigenic process are relatively well characterized for colorectal cancer, the epigenetic events and their impact on the transcriptional reprogramming observed in colorectal tumors have not been extensively characterized. Although recent genome-wide studies have analyzed the genomic distribution of hypermethylated CpGs in a small number of colorectal tumors (ref), a detailed analysis of the subset of these events that are important for gene expression regulation is currently lacking. Just as gene expression microarrays accelerated and revolutionized the study of transcriptional regulation, rapidly improving technologies are increasingly enabling researchers to assess locus-specific DNA methylation on a genome-wide scale. Recently various high-throughput approaches based on bisulfite conversion combined with next generation sequencing have been developed and applied for the genome wide analysis of DNA methylation. These methods provide single base pair resolution, quantitative DNA methylation data with genome wide coverage. There are various experimental types of methylation assays, but overall, methylation levels can be represented in one of three types: discrete, continuous or categorical. Therefore, methylation can be quantified by directly using read count information , ratio data (which may lose biological variability) or both. Once the DNA samples are processed, an important issue to be considered is the influence of the statistical analysis on the accuracy of the genomic methylation level estimation from bisulfite sequencing data. The accuracy of the statistical approach to methylation quantification increases with the sequencing depth of the particular cytosine residue ([10]). However, there are regression and neighboring analysis techniques that can counteract the lack of sequence depth in a particular CpG ([11]).

Existing methods and analyses

The association between gene expression and DNA methylation in the CpG islands in particular has been long studied; and as a result, mostly negative correlations have been found to relate to cancer driven mechanisms (**???**), but this inverse relationship between DNA methylation of the first intron in particular and gene expression is a broad mechanism to down-regulate gene expression and it is found in numerous processes, organisms and tissues ([12]). There have been various studies analysing this correlation using various approaches. For example, Massie et al., (2017) looked at the relationship between gene expression and DNA methylation at the probe level rather

than at the gene level. They narrowed a list of genes regulated by methylation that were identified in more than 3 out of 17 studies. Another study analysed the TCGA database to identify patterns in DNA CpG methylation and gene expression and tumor status. They found that the association involved a reduced number of genes linked to cancer than originally anticipated (around the hundreds) and that not all correlations were negative ([13]). Another recent paper reported two different models for analysis of DNA methylation and regulation of gene expression, one for negatively correlated genes and one for positively correlated genes (Klett et al., 2018). They used expression (GSE106582) and methylation datasets (GSE101764) containing 194 samples, 77 tumors and 117 of the mucose. By random forest analysis they were able to classify genes into cancer related and not related. Still methodologies to find tune classification into cancer/disease related and not cancer/disease related are still needed. A previously developed method was the selection of genes with an L-shape association between the expression and the methylation datasets (Sanchez-Pla et al., 2015). In this research, they focused on the CMI and on a method based on spline regression. They observed that the first method would detect L-shaped genes more accurately in big datasets. On the other hand, the splines clustering was not size dependent, but it would yield a smaller number of samples. Other research exists that aimed to identify genes regulated by methylation according to the expression methylation patterns; however, they only use a particular methodology like the CMI (**???**) with positive results. A paper focused on the identification of genes regulated by methylation through unsupervised clustering techniques to identify CRC subtypes was able to confirm existing subtypes ([14]).There has been other work that focused on the development of platforms for the identification of genes regulated by methylation. One of these packages is MEXPRESS (Koch et al., 2017). This package has a web interface that allows the user to visualize expression and methylation data from genes in the TCGA data. The visualization collocates for each selected gene, CpG islands, with transcripts expression together with other clinical values such as gender and age. The tool also generates p-values in relation to the variables specified. Another one of these packages is Methylmix (ref). The algorithm is based on a beta mixture model that identifies methylation states and compares them with what they call normal conditions to find hypo- and hyper-methylated genes. They developed a new statistic coeficient, the Differential Methylation value or DM-value which is defined as the difference of a methylation state with the normal methylation state. Then, they correlate that coefficient with gene expression data to characterize the association between methylation level and gene expression. For expression and methylation correlation analyses of both RNA and DNA molecules there is also a web based tool that analyses methylated genes based on TCGA data, called MethHC (http://methhc.mbc.nctu.edu.tw/php/search.php?opt=gene, Huang et al., 2015). This database has an analysis tool that provides gene-specific analysis for various diseases, and the information is displayed as a comparison between diseased and normal (non-diseased) conditions; list of highest and lowest methylated (hyper and hypo) genes; as well as correlations between expression and methylation. In this, methylation is a binary value (0,1) Other methodologies to identify methylated genes associated with cancer is through text mining analysis, as in the PubMeth database (www.pubmeth.org, [15]). In this, they identified 5000 genes of 1000 publications. However, high-throughput methodologies that offer an impartial approach to the identification of genes regulated by methylation still need further development and fine-tuning. Here we present such a methodology that will select, out of a gene expression and DNA methylation subsets, those genes that present a negative correlation, and are therefore regulated by methylation. The L-shaped heuristic method to identify genes regulated by methylation was tested and tuned for experimental expression and methylation paired datasets after normalization using other standard methods.

Material and Methods

Synthetic dataset generation for the simulation studies

The R package simstudy was used to create 4 artificial datasets by using the splines method (https://cran.r-project.org/web/packages/simstudy/simstudy.pdf). The package allows for designing data points on a pre-defined spline, in which knots, limits and dispersion can be tuned. The splines are generated based on a fixed X variable representing the methylation values. The artificial datasets contained a total of 1000 genes, and the data points were developed based on 2 parameters with 2 levels each. The first parameter was the number of samples and the second the % of true regulated by methylation genes that a sample would contain (with an expression by methylation scatterplot or spline following an L-shape). The number of samples considered was of 50 and 1000, and the % of true methylated genes in each dataset was 1% and 10%. Additionally, the shape of the negative genes (not regulated by methylation) was also pre-defined and classified into 3 different scatterplot shapes (Figure XXX) and the percentage of genes in each category equaled to 1/3 prior subtraction of the true GRM genes.

Simulation of heuristic classification with the synthetic datasets

The heuristic method was tested with the 4 artificial datasets: 50 samples, 1% of GRM genes; 50 samples, 10% GRM genes; 1000 samples, 1% GRM genes; and 1000 samples, 10% GRM genes. After running the model, sensitivity, specificity, and accuracy were measured and compared between datsets.

Results

Measures of performance for the heuristic method with synthetic datasets

Sensitivity, specificity and accuracy for the heuristic model were measured for the 4 synthetic datsets (Figure [ssa]) with predefined parameters described in the above section. Specificity was the parameter that scored highest in all datsets, with values between 0.99 (for the datasets with 50 samples) and 1 (for the datasets with 1000 samples). The second highest parameter was accuracy. In this, both datasets containing 1% of GRM genes scored 0.99, whereas the datasets with 10% of GRM genes scored 0.93 (for the one with 50 samples) and 0.92 (for the one with 1000 samples). Finally, the sensitivity values were the lowest in the combination of 1% GRM and 1000 samples (0.1) and highest for 1% GRM and 50 samples (0.5). These results indicate that the classification scored better non-L shaped scatterplots (true negatives) than L-shaped scatterplots (true positives).

Discussion

Supporting information

References

1. Berger SL, Kouzarides T, Shiekhattar R, Shilatifard A. An operational definition of epigenetics. Genes and Development. 2009;23: 781–783. doi:10.1101/gad.1787609

2. Saxonov S, Berg P, Brutlag DL. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. Proceedings of the National Academy of Sciences of the United States of America. 2006;103: 1412–1417. doi:10.1073/pnas.0510310103

3. Reik W. Stability and flexibility of epigenetic gene regulation in mammalian development. Nature Publishing Group; 2007. pp. 425–432. doi:10.1038/nature05918

4. Portela A, Esteller M. Epigenetic modifications and human disease. 2010. pp. 1057–1068. doi:10.1038/nbt.1685

5. Feil R, Fraga MF. Epigenetics and the environment: Emerging patterns and implications. 2012. pp. 97–109. doi:10.1038/nrg3142

6. Benayoun BA, Pollina EA, Brunet A. Epigenetic regulation of ageing: Linking environmental inputs to genomic stability. Nature Publishing Group; 2015. pp. 593–610. doi:10.1038/nrm4048

7. Jones PA, Baylin SB. The fundamental role of epigenetic events in cancer. 2002. pp. 415–428. doi:10.1038/nrg816

8. Feinberg AP, Vogelstein B. Hypomethylation distinguishes genes of some human cancers from their normal counterparts. Nature. 1983;301: 89–92. doi:10.1038/301089a0

9. Yang X, Han H, DeCarvalho DD, Lay FD, Jones PA, Liang G. Gene body methylation can alter gene expression and is a therapeutic target in cancer. Cancer Cell. Cell Press; 2014;26: 577–590. doi:10.1016/j.ccr.2014.07.028

10. Zhang Y, Jeltsch A. The application of next generation sequencing in DNA methylation analysis. 2010. pp. 85–101. doi:10.3390/genes1010085

11. Wreczycka K, Gosdschan A, Yusuf D, Grüning B, Assenov Y, Akalin A. Strategies for analyzing bisulfite sequencing data. Journal of Biotechnology. 2017;261: 105–115. doi:10.1016/j.jbiotec.2017.08.007

12. Anastasiadi D, Esteve-Codina A, Piferrer F. Consistent inverse correlation between DNA methylation of the first intron and gene expression across tissues and species. Epigenetics and Chromatin. BioMed Central Ltd. 2018;11. doi:10.1186/s13072-018-0205-1

13. Long MD, Smiraglia DJ, Campbell MJ. The Genomic Impact of DNA CpG Methylation on Gene Expression; Relationships in Prostate Cancer. Biomolecules. Multidisciplinary Digital Publishing Institute (MDPI); 2017;7. doi:10.3390/biom7010015

14. Barat A, Ruskin H, Byrne A, Prehn J. Integrating Colon Cancer Microarray Data: Associating Locus-Specific Methylation Groups to Gene Expression-Based Classifications. Microarrays. MDPI AG; 2015;4: 630–646. doi:10.3390/microarrays4040630

15. Ongenaert M, Van Neste L, De Meyer T, Menschaert G, Bekaert S, Van Criekinge W. PubMeth: a cancer methylation database combining text-mining and expert annotation. Nucleic Acids Research. 2007;36: D842–D846. doi:10.1093/nar/gkm788