Integrative Analysis of Methylation Changes in CpG Sites from Peripheral Blood Genomic Samples as a Result of the Presence of Ovarian, HNSCC, Bladder, and Breast Cancers using a Linear Regression Analysis and Artificial Neural Networks

Narganes-Carlón, D., Ball, G.

The John van Geest Cancer Research Centre, NTU, School of Science and Technology, Clifton Campus, NG11 8NS.

INTRODUCTION

Cancer is a complex, heterogeneous, evolving disease with different subtypes depending on its location, morphology, cell of origin, and genomic and epigenetic alterations (Hanahan and Weinberg, 2011). Every single cell alters the physiological homeostasis and generates a unique signature in the bloodstream that may transcend the apparent singularity of each cancer (Liotta, Ferrari and Petricoin, 2003).

CpG methylation is (i) preserved and regenerated during cell division, (ii) a stable signature in the genome that may surpass cancer heterogeneity, (iii) identifiable by bisulphite-based arrays, (iv) a link between environmental conditions with genetic penetrance and expressivity (Laird, 2010). Thus, the identification of various blood-based DNA methylation biomarkers capable of predicting cancer regardless of classification or subtype would be a highly valuable asset to the current diagnostic and screening processes.

BIOLOGICAL QUESTION

Are there any significant, common, recurrent changes in the CpG methylation ratio in PBCs as a result of cancer?

MATHERIAL AND METHODS

Two scripts were written in MatLab 2015 Version 8.6 R2015b:

1. A **linear regression analysis (LRA)** in a case × control loop to obtain the mean of all standard residuals (SRM) and its standard deviation (RSD) (Equation 1). A factor value was defined (Equation 2).

Equation 1. $$\mathrm{SR} = \text{Case MR} - \text{Control MR}$$

Equation 2. $$\mathrm{Factor} = \mathrm{ABS}\,(\mathrm{SMR}) \times \mathrm{RSD}^{-1}$$

2. An **unpaired Student's t test (UST)** for 2 tails, heteroscedastic samples and unequal variances. P-values were obtained.

A **stepwise ANN** was performed to test the potentiality of each probe to predict cancer. Average Test Error (ATE) was obtained in a random, blind, independent validation of the predictions.

Table 1 - Gene Expression Omnibus (GEO) accession number, sample description, number of control samples, cancer affecting the case samples, number of case samples and Illumina technology.

GEO	Sample	#Control	Cancer	#Case	Illumina array
GSE19711	DNA from PBCs	n=274	Ovarian	n=266	Infinium Methylation 27
GSE30229	DNA from PBCs	n=92	HNSCC	n=92	Infinium Methylation 27
GSE50409	DNA from PL	n=205	Bladder	n=223	Infinium Methylation 27
GSE57285	DNA from PBCs	n=49	Breast	n=35	Infinium Methylation 27

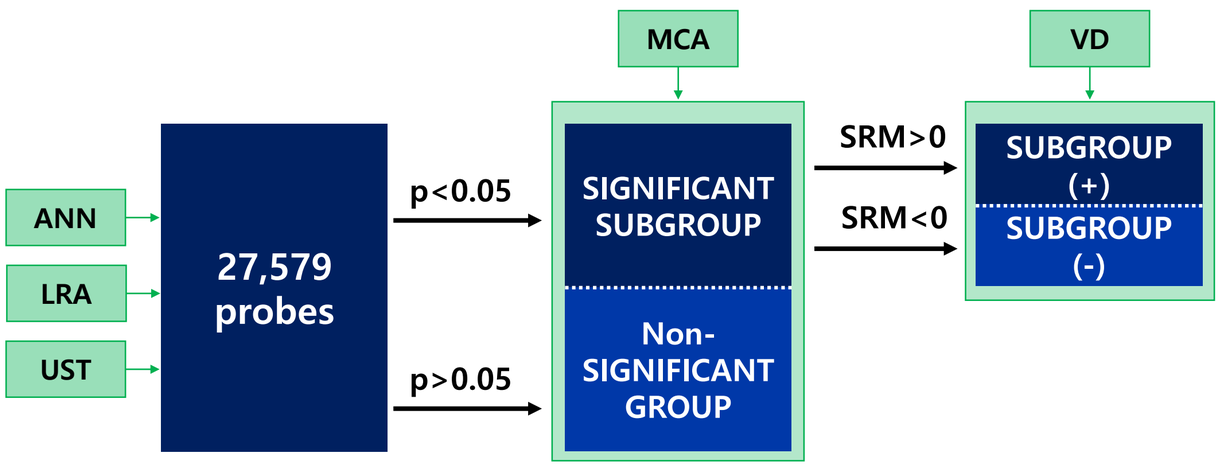

Figure 1 – Division of the four set of probes for LRA, UST, and stepwise ANN as well as the subsequent multiple correlation analysis (MCA) and Venn diagram (VD). The 27,579 probes of each dataset were divided into two groups with regards to the significance level (p=0.05). The four *Significant Groups* were subdivided based on the SRM in Equation 1. MCA was independently applied to the four *Significant Groups* and four *Non-Significant Groups* whereas a VD comprised (i) all four *Positive Subgroups* and (ii) all four *Negative Subgroups*.

RESULTS

Table 2 – Coefficients of determination (%R) from the MCA for *ATE*, *Factor* and negative value of natural logarithm of p-Value (*LPV*) within each dataset.

	GSE19711			GSE30229			GSE50409			GSE57285		
	ATE	Factor	LPV	ATE	Factor	LPV	ATE	Factor	LPV	ATE	Factor	LPV
ATE	100	4^b	7^b	100	1^b	6^b	100	1^b	4^b	100	6^b	8^b
Factor	95^a	100	96^b	98^a	100	98^b	95^a	100	99^b	92^a	100	98^b
LPV	96^a	98^a	100	98^a	99^a	100	97^a	99^a	100	93^a	97^a	100

a, %R of *ATE*, *Factor*, and *LPV* within each of the 4 *Significant Groups*;
b, %R of *ATE*, *Factor*, and *LPV* within each of the 4 *Non-Significant Groups*.

Gene	TS	RK
AIM2	3	416
AIM2	2	614
AIM2	1	1660
AIM2	1	4890
NOD2	7	1226
NOD2	3	1496
NOD2	1	1841
NOD2	1	1362
GSDMC	1	4553
GSDMC	1	566
GSDMC	1	3895
GSDMC	2	334

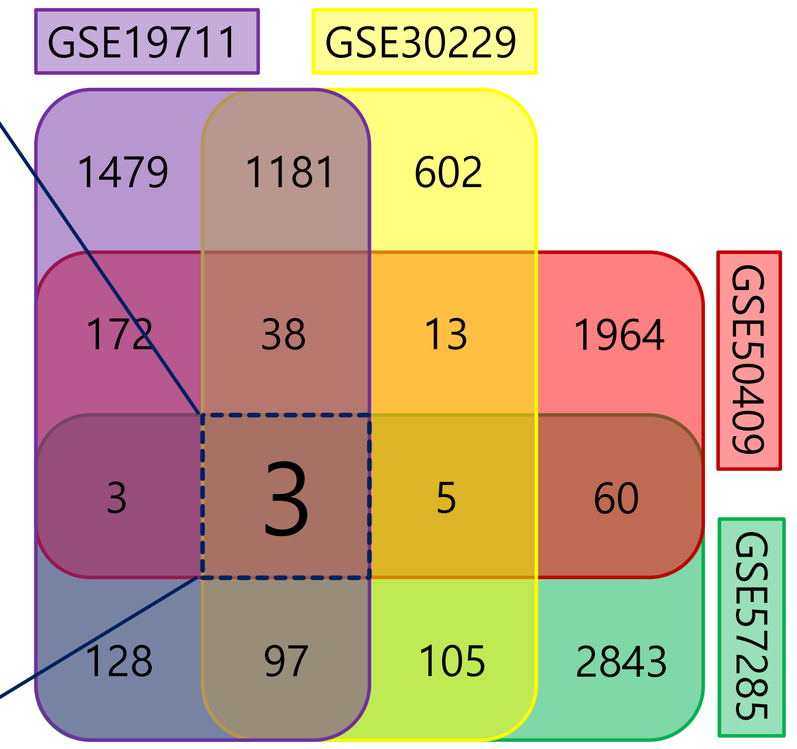

Figure 2 – Venn diagram for all four *Negative Subgroups* of probes. 3 CpG sites from gene promoters of AIM2, NOD2, and GSDMC were significantly ($p<0.05$) undermethylated ($SRM<0$) as a result of the four cancers. Times of significance ($p<0.05^{TS}$) and probe index for each dataset based on a sorting Largest to Smallest with regards to the product (Factor x LPV x ATE^{-1}) are represented. The combined probability of obtaining 3 genes was 1.7×10^{-16} based on the product (Index x $27{,}579^{-1}$) and 3.23×10^{-38} considering the products of all p-Values.

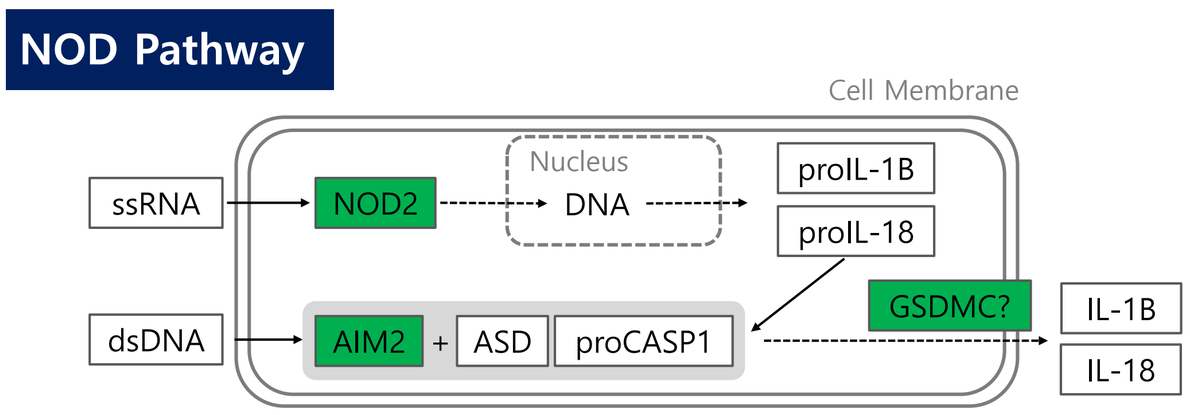

Figure 3 – Simplified NOD Pathway from KEGG platform (Kanehisa et al. 2017).

CONCLUSIONS

1. (i) The stepwise ANN, (ii) the linear regression analysis, and (iii) the unpaired Student's t-method could complement the robustness of further research of CpG methylation changes to predict highly specific, sensible and clinically useful biomarkers.

2. Three CpG sites from gene promoters of AIM2, NOD2, GSDMC were significantly and recurrently undermethylated in PBCs as a result of cancer.

FURTHER RESEARCH will focus on an integrative analysis of (i) methylation and (ii) transcription profiles, (iii) immunocytochemistry, (iv) mass spectrometry and (v) western blot assays for AIM2, NOD2, GSDMC, IL-1B, and IL-18 not only in the same but other cancers.

REFERENCES

Hanahan, D. and Weinberg, R. A. (2011) 'Hallmarks of Cancer: The Next Generation', **Cell**, 144(5), pp. 646–674.
Liotta, L. A., Ferrari, M. and Petricoin, E. (2003) 'Clinical proteomics: Written in blood', **Nature**, 425(6961), p. 905.
Laird, P. W. (2010) 'Principles and challenges of genome-wide DNA methylation analysis', **Nat Rev Genet**. Nature Publishing Group, 11(3), pp. 191–203.
Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. and Morishima, K. (2017) 'KEGG: new perspectives on genomes, pathways, diseases and drugs', **Nucleic Acids Research**. Oxford University Press, 45(Database issue), pp. D353–D361.