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

David Narganes-Carlon[1], Graham Ball[1]

[1] School of Science and Technology, Nottingham Trent University, Clifton Campus, Nottingham NG11 8NS, UK

Correspondence to: David Narganes-Carlon, **email**: n0712780@my.ntu.ac.uk

Keywords: DNA methylation profiling, integrative analysis, pan-cancer, HNSCC, bladder cancer, ovarian cancer, breast cancer, Artificial Neural Networks, CpG sites

Received: April 20, 2017 **Accepted**: - **Published**: -

ABSTRACT

Cancer is a complex, heterogeneous, evolving and dynamic disease that has hundreds of different forms and subtypes depending on the location, morphology, affected molecular pathways, mutation status, and cell of origin and spectrum of both genomic and epigenetic alterations. Heterogeneity within tumour cells has been widely reported whereas an integrative analysis of epigenetics changes in CpG sites from gene promoters from peripheral blood cell types as a result of different stages, types, and subtypes of cancer presence has not. Furthermore, there is a total lack of universal cancer biomarkers. Four genome-wide methylation profiling array datasets from peripheral blood samples from healthy patients and individuals with head and neck squamous carcinoma (HNSCC), ovarian, bladder and breast cancers were analyzed. A linear regression analysis, student's t-test and a stepwise ANN approach were performed. The aim was to identify significant, common, homogeneous, and recurrent changes in the methylation pattern of CpG sites from promoters of consensus coding sequences in peripheral blood cell types as a result of the presence of cancer, regardless of the subtype and stage of the disease. Three gene promoters of AIM2, NOD2, and GSMDC were significantly and recurrently under-methylated in peripheral blood cell lines as a result of four cancer types. Furthermore, these three genes participated in the NOD Receptor Signaling Pathway and could be applied to compliment, support and increase the robustness for the early diagnostic of cancer in combination with currently available biomarkers.

INTRODUCTION

Cancer heterogeneity

Cancer is a complex, heterogeneous, evolving and dynamic disease that has hundreds of different forms and subtypes depending on the location, morphology, affected molecular pathways, mutations, cell of origin and spectrum of both genomic and epigenetic alterations (Hanahan and Weinberg, 2011; Galon *et al.*, 2014). Substitutions, insertions, deletions, translocations, breakages, rearrangements of chromosomes at multiple loci (e.g. chromothripsis and chromoplexy), hypermutational processes (e.g. kataegis) as well as non-mutational changes such as histone and DNA-methylation changes confer this apparent uniqueness to each form of cancer (Network *et al.*, 2013). The currently accepted clonal evolution hypothesis that follows the Darwinian model for natural selection allows the existence of both genetically and phenotypically diverse sub-populations of transformed cells with a biological fitness to adapt to a fluctuating microenvironment within the primary neoplasm (Greaves and Maley, 2012; Fisher, Pusztai and Swanton, 2013). As a result this heterogeneity, sampled sub-clonal lineages may be unrepresentative of the whole tumour cell population (Galon *et al.*, 2014). This sampling error can bias the discovery and validation of predictive biomarkers. Consequently, the current biomarker repertoire is not sensible, specific and clinically useful enough to detect treatable early-stages of cancer, histopathological diagnostics are not accurate enough, common benign conditions are misclassified (Liotta, Ferrari and Petricoin, 2003), and, the most important issue, there is a total lack of universal biomarkers to detect any form, subtype or stage of this diverse set of diseases: cancer.

Microenvironment homogeneity?

Heterogeneity within tumour cells has been widely reported but does it applies to the tumour microenvironment as well? Chronic proliferation of the neoplasm causes nutrient, chemokines, oxygen, carbon dioxide, pro-angiogenic, pro-invasive matrix-degrading enzymes, and pro-inflammatory molecules imbalance as well as release of cell contents into the bloodstream upon programmed cell death, necrosis, or cell senescence (Liotta, Ferrari and Petricoin, 2003; Hanahan and Weinberg, 2011). Transformed cells interact not only within the transformed cells but also with adjacent cells in the microenvironment and disrupt both ligand and cell-cell adhesion interactions, aiming to evade contact inhibition and immunosurveillance (Hanahan and Weinberg, 2011; Galon *et al.*, 2014). All in all, every single cell in the organism leaves a record of its physiological state in the molecules it sheds to the microenvironment, either as waste or as signals to neighboring cells, altering the physiological homeostasis and generating a unique signature in the blood microenvironment that may transcend the apparent singularity of each type of cancer (Liotta, Ferrari and Petricoin, 2003; Network *et al.*, 2013).

Methylation patterns in peripheral blood cell lines

Blood is a circulating ‘connective’ tissue composed of plasma (55%) and formed elements (45%): enucleated cells such as erythrocytes (96%), and platelets (3%) and leukocytes (1%). Blood is available in relatively high abundance, amendable to various biopsies and histologically less invasive tests, useful when the target neoplastic tissue is not readily available (Mohr and Liew, 2007). Nevertheless, biomarker dilution into the bloodstream and the immunoediting role of cancer to mask its stress in the tumor microenvironment will impede the discovery of high sensible biomarkers (Schreiber, Old and Smyth, 2011). On the other hand, the quest for a single cancer-specific biomarker has the illusion of analytical simplicity but makes little sense from a biological perspective (Liotta, Ferrari and Petricoin, 2003). Why not to look for a combination of multiple biomarkers in the blood that could enhance the predictive power and may

reach a clinically useful levels of specificity, sensitivity and reproducibility with biological relevance rather than expecting a neoplasm to recurrently generate a new protein? (Petricoin *et al.*, 2006; Galon *et al.*, 2014) More concretely, why not look for changes in 5-cytosine methylation of CpG sites from gene promoters? Changes in methylation of CpG sites are (i) preserved or regenerated during cell division, (ii) a stable and homogeneous signature in genomic DNA that may surpass cancer heterogeneity, (iii) potentially identifiable by bisulphite-based methylation profiling, (iv) a potential 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) or pan-cancer would be a highly valuable asset to the current diagnostic and screening processes, as well as contributing to the current knowledge of systemic changes associated with this disease.

Artificial Neural Networks

Artificial Neural networks (ANN) attempt to mimic the capacity to learn of biological neural systems by (i) modeling the low-level structure of the brain, (ii) performing minor mathematical adjustments on the connections of a neuron-based structure, and (iii) handling non-linear features within data to generalize or predict future cases (Lancashire, Lemetre and Ball, 2009). ANN are fault tolerant, not limited by linear functionality (unlike other statistical approaches e.g. hierarchical clustering or principal components analysis), able to handle noisy information, and able to endure incomplete data (Lancashire, Lemetre and Ball, 2009). Multilayer perceptron (MLP) was selected for this project due to its useful application for biomarker identification in previous studies (Albarakati *et al.*, 2015; Abdel-Fatah *et al.*, 2017).

Biological questions and data

Four datasets of genome-wide methylation profiling from peripheral blood samples available in ArrayExpress were used for this experiment. These datasets comprised total genomic DNA of peripheral blood samples from healthy patients and individuals with head and neck squamous carcinoma (HNSCC), ovarian, bladder and breast cancers. A linear regression analysis, student's t-test and a stepwise ANN approach were performed. The biological question was: Are there any significant, common, homogeneous, and recurrent changes in the methylation pattern of CpG sites from gene promoters in peripheral blood cell types as a result of the presence of cancer (regardless of classification or subtype of cancer)?

MATERIAL AND METHODS

Different cohorts: ArrayExpress data

Methylation ratio of 5-cytosine methylation of a total of 27579 CpG sites from gene promoters from peripheral blood cell lines was obtained from ArrayExpress (https://www.ebi.ac.uk/arrayexpress/). ArrayExpress is a public database of microarray data at the European Bioinformatics Institute (EBI) that aims to act as a data repository by supporting publication, facilitating access to high quality data and allowing the sharing of hybridization, sample treatments, microarray designs and image analysis protocols (Parkinson *et al.*, 2007). DNA converted by bisulphite from the whole blood genomic samples were hybridized to the Illumina Infinium 27k Human Methylation Beadchip v1.2 in the following cohorts:

GSE19711: Consisted on a total of 540 whole blood genomic samples from postmenopausal women with primary epithelial ovarian cancer that acted as case (n=266) and healthy postmenopausal women that acted as a control (n=274) from the UK Ovarian Cancer Population Study (Teschendorff *et al.*, 2009) (Gene Expression Omnibus (GEO): GSE19711). Detailed clinical features are described elsewhere ('*E-GEOD-19711.sdrf.txt*' file available in http://www.ebi.ac.uk/arrayexpress/experiments/E-GEOD-19711/).

GSE30229: Consisted on a total of 184 whole genomic samples of peripheral blood samples from patients with HNSCC that acted as cases (n=92) and healthy subjects (n=92) with no prior history of cancer (GEO: GSE30229) (Langevin *et al.*, 2012). Detailed clinical features are described elsewhere ('*E-GEOD-30229.sdrf.txt*' file available in https://www.ebi.ac.uk/arrayexpress/experiments/E-GEOD-30229/).

GSE50409: Consisted on a total of 428 genomic samples from peripheral lymphocytes from patients with bladder cancer that acted as case (n=223) from National Health State Cancer Registry (July 1, 1994-June 30, 1998) and healthy patients (n=205) (GEO: GSE50409) (Langevin *et al.*, 2014). Detailed clinical features are described elsewhere ('*E-GEOD-50409.sdrf.txt*' file available in https://www.ebi.ac.uk/arrayexpress/experiments/E-GEOD-50409/).

GSE57285: consisted on a total of 84 whole blood genomic samples of healthy women either with BRCA1 wild-type (n=42) or mutant BRCA1 allele (n=7) acted as control whereas women with breast cancer and mutant BRCA1 gene (n=35) acted as case (GEO: GSE57285) (Anjum *et al.*, 2014). Detailed clinical features are described elsewhere ('*E-GEOD-57285.sdrf.txt*' file available in https://www.ebi.ac.uk/arrayexpress/experiments/E-GEOD-57285/).

The different stages and subtypes of each concrete cancer were not discriminated for their subsequent analysis.

Linear regression analysis and Student's t-test

A MatLab 2015 Version 8.6 R2015b script was written to perform a linear regression analysis by comparing the CpG methylation ratios for all control samples (i) against all case samples (j) in a $i \times j$ loop. Standard residual (SR) values were calculated based on the difference between the abscissa and the ordinate values (Equation 1) for all Illumina probes all samples in all datasets (GSE19711, GSE30229, GSE50409 and GSE57285). Residual values were calculated based on its distance to the identity function (*Supplementary Figure 1*). SR means (SRM) and residual standard deviations (RSD) were calculated for all probes in all datasets. A factor value was defined (Equation 2).

Equation 1 $$SR = Case\ MRV - Control\ MRV$$

Equation 2 $$factor = ABS\ (SMR) \times RSD^{-1}$$

An unpaired Student's t method for 2 tails and heteroscedastic samples and unequal variances was performed with the previously normalized CpG methylation ratios for all probes in all four datasets.

Table 1 - Gene Expression Omnibus [GEO] accession number, total number of samples, gender of the total population, information related to cell lines used, number of control samples, cancer type affecting the case samples, number of case samples and Illumina technology used for the hybridization.

GEO	Total	Gender (M/F)	Sample description	#Control	Cancer	#Case	Illumina array
GSE19711	n=540	F	Whole blood genomic samples	n=274	Ovarian	n=266	Infinium Methylation 27 v1.2
GSE30229	n=184	Both	Genomic sample from peripheral blood cells	n=92	HNSCC	n=92	Infinium Methylation 27 v1.2
GSE50409	n=428	Both	Genomic sample from peripheral lymphocytes	n=205	Bladder	n=223	Infinium Methylation 27 v1.2
GSE57285	n=84	F	Whole blood genomic sample	n=49	Breast	n=35	Infinium Methylation 27 v1.2

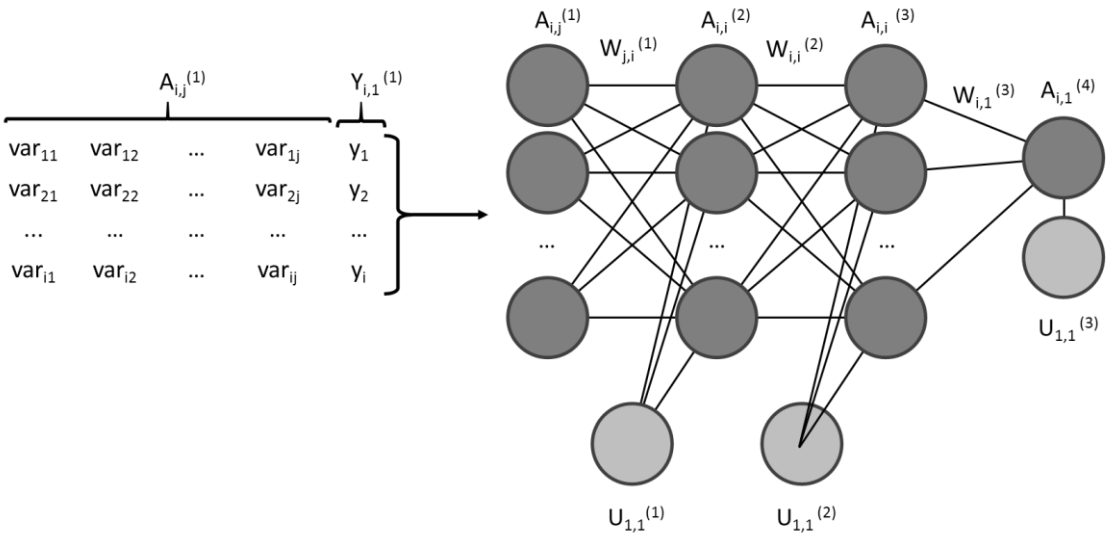

Figure 1 – All normalized CpG methylation ratios for an each discrete dataset were comprised in a $A_{i,\,j}{}^{(1)}$ matrix where indexes *i* and *j* corresponded to the number of Illumina Infinium 27k Human Methylation Beadchip v1.2 probes and the number of samples, respectively. Control (0) and cancer (1) values were included in the $Y_{1,\,i}{}^{(1)}$ matrix. $W^{(1)}$, $W^{(2)}$ and $W^{(3)}$ along with $U^{(1)}$, $U^{(2)}$ and $U^{(3)}$ matrices of random values were modified in each step of the algorithm to eventually minimize the error function (Equation 3). The basic structure MLP is represented with $A^{(2)}$ and $A^{(3)}$ as two hidden layers. $A^{(4)}$ contain the ANN predicted values.

ANN MatLab script: Multilayer Perceptron

A multilayer perceptron (MLP) algorithm script was written in a MatLab 2015 Version 8.6 R2015b. An MLP is composed by multiple layers of nodes fully connected to the next layer (Scheme of an MLP in Figure 1). Except for the input nodes, each node is a processing element with a non-linear activation function. MLP utilizes a supervised learning technique and a backpropagation algorithm (see below) for training the network. Normalized CpG methylation ratios from ArrayExpress (available at the links provided in Different Cohorts section) were disposed in $A_{i,\,j}{}^{(1)}$ matrix whereas the relationship to cancer ('0' for control samples and '1' for cancer case samples, regardless of classification or subtype of cancer) for each of the samples were comprised in the $Y_{1,\,i}{}^{(1)}$ matrix. Indexes *i* and *j* represented the total number of probes and samples in a discrete dataset, respectively. The algorithm repeated a two-phase cycle: propagation (forward algorithm) and weight update (backpropagation algorithm) in each epoch.

Forward algorithm. The input $A_{i,\,j}{}^{(1)}$ matrix is propagated forward through the network, layer by layer, until an output is generated: $A_{i,\,1}{}^{(4)}$ matrix. The activation values of all layers were calculated based on Equation 4. A non-linear transfer function was applied (Sigmoid function in Equation 5). $W^{(1)}$, $W^{(2)}$ and $W^{(3)}$ along with $U^{(1)}$, $U^{(2)}$ and $U^{(3)}$ matices of random values were redefined in each step of the algorithm to eventually minimize the ANN error function (Equation 3).

Backpropagation algorithm. As loss function (Equation 3) describes, the output values of $A_{i,\,1}{}^{(4)}$ matrix were compared to the experimental condition ('0' for control samples and '1' for cancer case samples, the different stages and subtypes of the disease were not discriminated) in a supervised method. The obtained error matrix $J(A^{(4)})$ was then propagated backwards, starting from the output, until each neuron had an associated error value which represented its contribution to the original output. The aim of this cycle was an attempt to minimize the loss function (Equation 3). Mathematically, the contribution of each neuron to the total error was calculated by partially deriving Equation 3 by all its variables: the previously mentioned matrices of random values ($W^{(1)}$, $W^{(2)}$ and $W^{(3)}$ as well as $U^{(1)}$, $U^{(2)}$ and $U^{(3)}$). A generalization formula to calculate all derived equations was written to optimize the ANN (Equations 6 and 7). The random values of mentioned matrices were updated in each cycle (or epoch) based on the gradient descent and the defined learning rate (η) (Equation 8).

$$\text{Equation 3} \qquad J(A^{(4)}) = 0.5\times(Y^{(1)} - A^{(4)})^2$$

$$\text{Equation 4} \qquad A^{(n+1)} = f(A^{(n)} \cdot W^{(n)} + U^{(n)})$$

$$\text{Equation 5} \qquad f(x) = (1 + e^{-x})^{-1}$$

The developed script was not successfully applied to complete datasets as a consequence of (i) the high number of dimensional Euclidean spaces (equal to 27579, number of Illumina probes) and (ii) due to the lack of an optimization to overcome the high-dimensionality of the data of this experiment.

$$\text{Equation 6} \qquad {\partial J}/{\partial V^{(n)}} = \begin{cases} A'^{(n)} \times \left(A^{(4)} - Y^{(1)}\right) \times f'\left(A^{(3)} \times W^{(3)} + U^{(3)}\right) \times m, & W = V \\ \left(A^{(4)} - Y^{(1)}\right) \times f'\left(A^{(3)} \times W^{(3)} + U^{(3)}\right) \times m, & U = V \end{cases}$$

$$\text{Equation 7} \qquad m = 1; \quad m = m \times W'^{(n+1)} \times f'\left(A^{(n)} \times W^{(n)} + U^{(n)}\right); \quad for\ n = 3, 2, 1$$

$$\text{Equation 8} \qquad V^{(n)} = V^{(n)} - \eta \times {\partial J}/{\partial V^{(n)}} \qquad V = (W, U)$$

Stepwise ANN

A stepwise ANN approach was performed with all datasets. The aim of this method was (i) to identify the best predictive probe that could explain the cancer phenotype in each one of the 10 performed loops and, mainly, (ii) to obtain a value that reflected the potentiality for each one of the 27,579 probes to explain the cancer phenotype. The stepwise ANN has previously been shown to identify high sensitivity and specificity biomarkers with excellent clinical validity (Lancashire *et al.*, 2010). This approach used a multi-layered supervised learning approach with a non-linear (sigmoidal) transfer function. The backward propagation of errors updated the weights in the same way as described above. The algorithm parameters were set relatively to previous effective similar datasets (Lancashire, Rees and Ball, 2008): 2 hidden layers, 50 Monte–Carlo cross-validations (a random and blind sample process for an independent validation), 1 step, 0.5 momentum, 0.1 learning rate, 100 size of window step, 10 loops; the algorithm would have automatically stopped if the performance had failed to improve on the test data split in 1% for 300 epochs; and the percentages of training, validation (to asses model performance during the training process) and test (to independently validate the model on previously unseen data) sets were 60, 20, and 20 % (Lancashire, Rees and Ball, 2008; Lancashire, Lemetre and Ball, 2009), respectively. Both a panel of the top predictive 10 probes in each loop and a 10 summaries for all probes were obtained. Values for the average performance and the average error for the training, validation and test sets were also obtained. Median accuracy (degree to which it can be used to identify diseased patients) and median squared error for the 10 loops were calculated in accordance to the percentages of training, test and validation sets. Average Test Error (ATE) was obtained in the mentioned random and blind approach for an independent validation of the predictions. ATE was the considerate value for determining the individual contribution of each probe to predict cancer in the samples.

Selection of data for a Multiple Correlation Analysis, Venn diagram, probe ranking and gene identifications

The 27,579 probes from each dataset were divided into two groups with regards to their significance value: non-significant ($p>0.05$) probes were included *Significant Groups* whereas significant ($p<0.05$) probes were comprised *Non-significant Groups* for each dataset. The *Significant Groups* were subsequently divided into two subgroups based on the value that resulted from Equation 1: probes with positive SRM formed the *Positive Subgroup* whereas probes with a negative SRM comprised the *Negative Subgroup* (Explanatory scheme of the overall division process in Figure 2). The idea was to solely compare the probes that reflected over-methylation (negative SRM) or under-methylation (positive SRM) in gene promoters of samples obtained from patients with cancer. A multiple correlation analysis (MCA) was performed with MatLab 2015 Version 8.6 R2015b for the three considered parameters: *ATE*, negative value of natural logarithm of p-Value (*LPV*) and *factor* obtained from the stepwise ANN, the student's t-test and the linear regression analysis, respectively. This MCA was independently performed for both the *Significant Groups* and the *Non-Significant Groups* of each dataset. All *Positive Subgroups* on one hand and all *Negative Subgroups* on the other hand were analyzed at the level of probe identifier within themselves to calculate the intersections into a Venn diagram (http://bioinformatics.psb.ugent.be/webtools/Venn/). Furthermore, all probes from each dataset were sorted with Microsoft Excel 2016 ® with regards to Ranking Value (Assigned to each probe with regards to Ecuation 9). The file containing the Array Design information on ArrayExpress (link mentioned in Methods, Different Cohorts) was used to identify RefSeq gene identifications. Furthermore, the Pathway sub-database (http://www.genome.jp/kegg/pathway.html) included in the KEGG database (Kanehisa and Goto, 2000; Kanehisa *et al.*, 2016, 2017) was used to systematically analyze the gene pathways.

Equation 9 $$Ranking\ Value = factor \times \mathrm{LPV} \times \mathrm{ATE}^{-1}$$

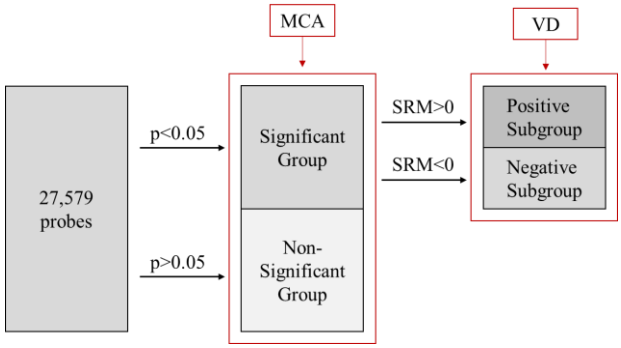

Figure 2 – Scheme of the division of the probes for the subsequent analysis: multiple correlation analysis (MCA) and Venn diagram (VD). The 27579 probes from each methylation profiling dataset were divided into two groups (a total of 8 groups, 2 from each dataset) with regards to the significance value: non-significant ($p>0.05$) and significant ($p<0.05$) probes. The four *Significant Groups* were subsequently divided into two subgroups (a total of 8 subgroups, 4 *Positive Subgroups* and 4 *Negative Subgroups*) based on the value that resulted from Equation 1: probes with positive standard residuals (SRM > 0) formed a *Positive Subgroup* whereas probes associated to a negative standard residuals (SRM < 0) comprised a *Negative Subgroup*. MCA was independently applied to the four *Significant Groups* and four *Non-Significant Groups* from all datasets whereas VD comparison was performed with both the all four *Positive Subgroups* on one hand and all four *Negative Subgroups* on the other hand.

RESULTS

Stepwise ANN modelling-panel genes

The stepwise ANN approach analyzed a total of 110,316 probes (27,579 for each one of the four discrete datasets). The algorithm was conducted in 10 loops to determine (i) the top predictive probe that could better explain the cancer in each loop and (ii) the *ATE* of each probe to explain cancer phenotype. Nonetheless, none of the panel of 10 probes (1 top probe per loop) reached high levels of accuracy (superior to 90%) (Figure 3). This low performance was consistent among the datasets: cg00645579 with 67.5% in loop 6, cg23547429 with 75.04% in loop 3, cg25307081 with 62.84% in loop 7, and cg22892110 with 78.60% in loops 3 and 8 were the highest median accuracies for the 10 loops of the stepwise ANN in GSE19711, GSE30229, GSE50409 and GSE57285 datasets, respectively. The lowest median squared errors were: cg25634666 with 10.72% in loop 4, cg23547429 with 9.28% for loop 9, cg25307081 with 11.46% in loop 7, and cg22892110 with 8.32% in loop 6, for GSE19711, GSE30229, GSE50409 and GSE57285 datasets, respectively (Numerical values for Median Accuracies and Median Squared Errors as well as the probe identifiers are available in *Supplementary Table 1*). Furthermore, no common genes were found among the four datasets. Consequently, no further analysis was performed with this approach.

Correlation of the three approaches

A strong correlation (Superior in all cases to 91.8% in *a* condition in Table 2) was reported among *ATE*, *LPV* and *factor* for the *Significant Groups* of all datasets but not for the *Non-Significant Groups* (*b* condition in Table 2). Concretely, in the MCA applied to the *Non-Significant Group*, the coefficients of determination for *factor* and *LPV* were consistently higher than 95.8% whereas coefficients of determination inferior to 8% were reported for correlations with *ATE* in the *Non-Significant Group* (*b* condition in Table 2).

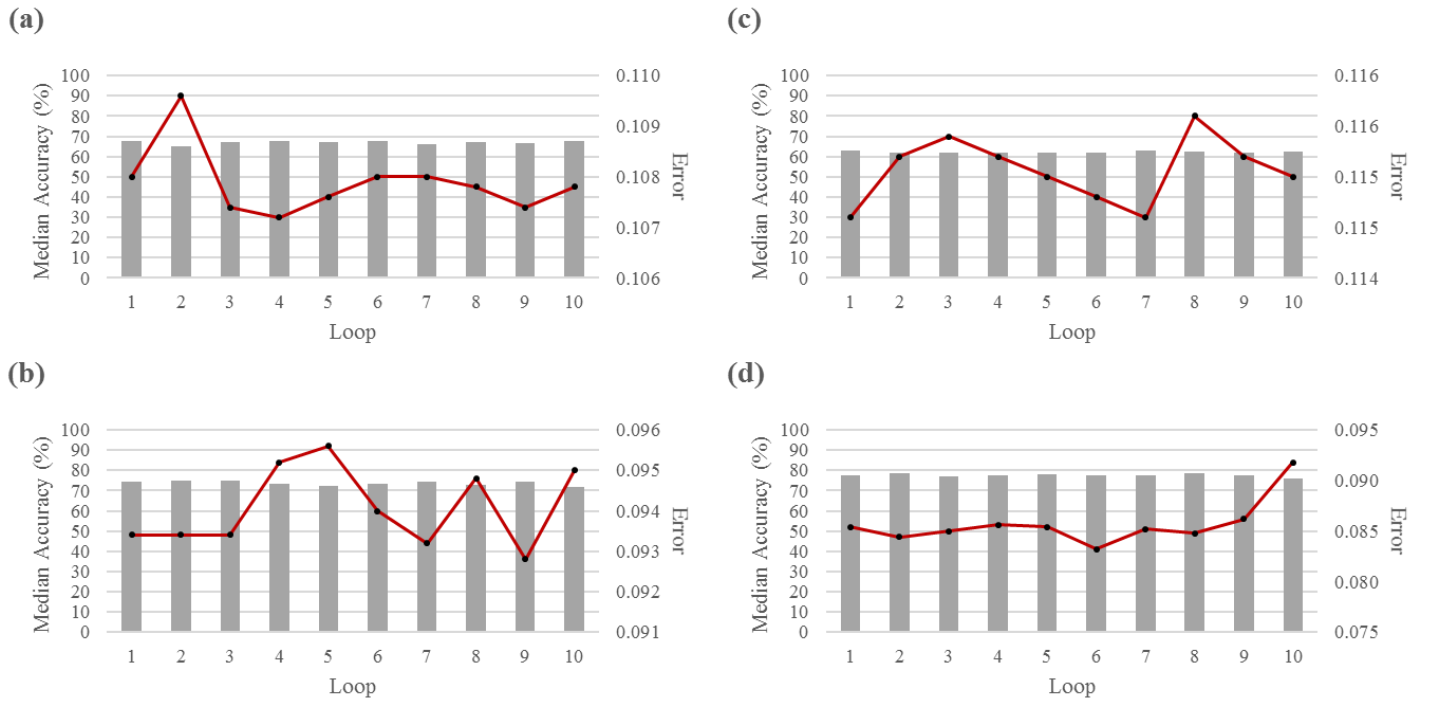

Figure 3 - Performance of the top predictive probe that could better explain the cancer condition at each of the 10 total loops of the stepwise ANN for (a) ovarian cancer in GSE19711, (b) HNSCC in GSE30229, (c) bladder cancer in GSE50409, and (d) breast cancer in GSE57285 datasets. In none of the 10 loops of the algorithm, the accuracy of the top predictive probe reached high values (superior to 90% of median accuracy). Columns represent median model accuracy; lines represent mean squared error for the predictions at each loop. Numerical values can be found in *Supplementary Table 1*.

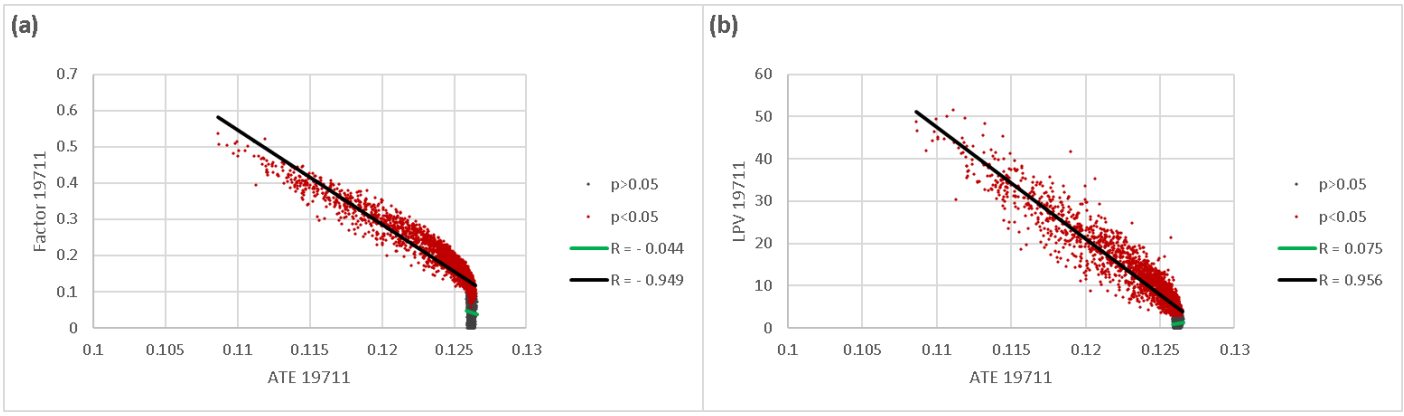

Figure 4 – Linear regression analysis for *ATE* and *Factor* (a) and *ATE* and the *LPV* (b) values for the discrete GSE19711 dataset. Grey and red markers correspond to the representation of the mentioned values in each graph for the probes comprised in *Non-Significant Group* (grey) and *Significant Group* (red), respectively. Regression lines were calculated: the regression line for the correlation of ATE-Factor (a) and ATE-LPV (b) for the *Non-Significant Subgroup*s is represented in green whereas the regression line for the correlation of ATE-Factor (a) and ATE-LPV (b) for the *Significant Subgroups* is represented in black. The coefficients of determination (R) were calculated: -0.044 for *Factor-ATE* in green for the *Non-Significant Group* in grey (a), -0.959 for *Factor-ATE* in black for the *Significant Group* in red (a); 0.075 for *LPV-ATE* in green for the *Non-Significant Group* in red (b); and 0.956 for *LPV-ATE* in black for the *Significant Group* in red (b). Complete results for all datasets are in Table 2.

Table 2 – Results for a multiple correlation analysis for ATE, factor and LPV within each dataset (GSE19711, GSE30229, GSE50409 and GSE57285). Explanatory graph in Figure 4. The coefficients of determination (R) obtained from the linear regression analysis were disposed in the following table.

	GSE19711				**GSE30229**				**GSE50409**				**GSE57285**		
	ATE	Factor	LPV		ATE	Factor	LPV		ATE	Factor	LPV		ATE	Factor	LPV
ATE	1.000	-.044[b]	.075[b]	ATE	1.000	-.012[b]	-.054[b]	ATE	1.000	-.010[b]	-.040[b]	ATE	1.000[b]	.056[b]	.080[b]
Factor	-.949[a]	1.000	.958[b]	Factor	-.981[a]	1.000	.980[b]	Factor	-.946[a]	1.000	.989[b]	Factor	-.918[a]	1.000	.977[b]
LPV	-.956[a]	.981[a]	1.000	LPV	-.977[a]	.985[a]	1.000	LPV	-.965[a]	.991[a]	1.000	LPV	-.933[a]	.966[a]	1.000

a, correlation of the three variables within each of the 4 *Significant Groups*; *b*, correlation of the three variables within each of the four *Non-Significant Groups*.

Venn diagram

Despite that some significant probes did not overlap within the four *Negative Subgroups* (A Diagram in Figure 5 with 1479, 602, 1964, and 2843 unique, significant probes with negative SRM for GSE19711, GSE30229, GSE50409, and GSE57258 datasets, respectively) and within the four *Positive Subgroups* (B Diagram in Figure 5 with 1120, 169, 2017, and 1700 unique, significant probes with positive SRM for GSE19711, GSE30229, GSE50409, and GSE57258 datasets, respectively), three probes (cg05316065, cg10636246, and cg26954174) overlapped within the four *Negative Subgroups* (A Diagram in Figure 5) whereas no common probes were found within the four *Positive Subgroups* (B Diagram in Figure 5). RefSeq Identifiers, Gene symbol and Entrez NCBI number as well as MSR, RSD, *factor*, p-Value, *LPV*, *ATE* numerical values for the three common probes (cg05316065, cg10636246, and cg26954174) can be found in Table 3. Ranking Value (Equation 9) was also included in Table 3.

Based on Table 3, the significance levels of cg10636246 were: $p<0.05$ for GSE57285, $p<0.01$ for GSE50409, and $p<0.001$ for GSE19711 and GSE30229; ATE values were inferior to a threshold of 0.12 just in GSE30229; and, based on Ranking Value (Equation 9), this gene was located in the top 500 for GSE19711, top 1000 for GSE30229 and top 5000 for the remaining datasets. Similarly, the significance levels of cg26954174 were: $p<0.01$ for GSE50409 and GSE57285, and $p<0.001$ for GSE19711 and GSE30229; ATE values were inferior to a threshold of 0.12 in GSE30229; and, based on Ranking Value, this gene was located in the top 2000 for all datasets. The significance levels of cg05316065 were: $p<0.05$ for GSE19711 and GSE50409, $p<0.01$ for GSE30229, and $p<0.001$ for GSE57285; the ATE values for this probe were superior in all cases to 0.124; and, based on Ranking Value, this gene was located in the top 500 for GSE57285, top 1000 for GSE30229 and top 5000 for the remaining datasets. Furthermore, KEGG database (Kanehisa and Goto, 2000; Kanehisa *et al.*, 2016, 2017) suggested that GSDMC, AIM2, and NOD2 participated in the NOD Receptor Signaling Pathway (Figure 6).

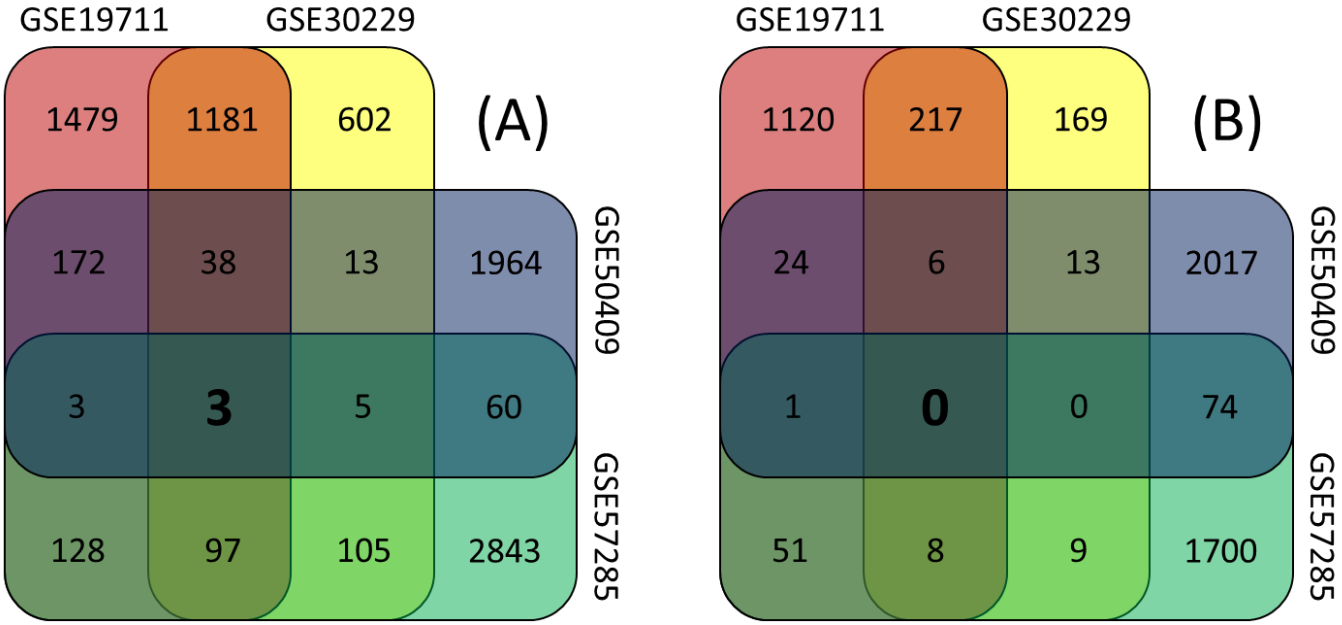

Figure 5 – Venn diagram for the comparison among all four *Negative Subgroups* (A) and comparison among all four *Positive Subgroups* (B) from all datasets (Red GSE1971 in red, yellow GSE30229 on the top, blue GSE50409, and green GSE57285 on the right) for ovarian, HNSCC, bladder, and breast cancer, respectively. 3 significant probes with negative standard residual mean (cg05316065, cg10636246, and cg26954174) were shared in all datasets (A). No commonalities were found among the *Negative Subgroups* (B) from all four datasets.

Table 3 – GEO of the Illumina Infinium methylation array datasets, name for the Illumina Infinium probe, RefSeq mRNA ID, gene symbol, entrez gene ID, mean of the standard residual (MSR), residual standard deviation (RSD), factor value obtained from Equation 2, p-Value for heteroscedastic samples and two tails, negative value of natural logarithm of the p-Value, average test error (ATE) from the stepwise ANN and number of better predictors based on Equation 9 for the three shared Illumina Infinium probes that resulted under-methylated in the blood samples from patients with cancer phenotype in comparison to samples from healthy patients.

Dataset	Illumina	RefSeq ID	Gene symbol	Entrez	MSR	RSD	factor	p-Value	LPV	ATE	ranking
GSE19711	cg10636246	NM_004833	AIM2	9447	-0.010	0.055	0.178	9.42×10^{-05} ***	9.270	0.125	416
GSE30229	cg10636246	NM_004833	AIM2	9447	-0.026	0.063	0.412	1.26×10^{-04} ***	8.980	0.113	614
GSE50409	cg10636246	NM_004833	AIM2	9447	-0.016	0.087	0.184	7.44×10^{-03} **	4.901	0.125	1660
GSE57285	cg10636246	NM_004833	AIM2	9447	-0.014	0.044	0.330	3.58×10^{-02} *	3.330	0.125	4890
GSE19711	cg26954174	NM_022162	NOD2	64127	-0.015	0.053	0.278	5.45×10-^{10} ***	21.330	0.121	1226
GSE30229	cg26954174	NM_022162	NOD2	64127	-0.029	0.067	0.433	8.96×10^{-05} ***	9.320	0.111	1496
GSE50409	cg26954174	NM_022162	NOD2	64127	-0.013	0.073	0.179	9.30×10^{-03} **	4.678	0.125	1841
GSE57285	cg26954174	NM_022162	NOD2	64127	-0.024	0.052	0.462	3.46×10^{-03} **	5.668	0.122	1362
GSE19711	cg05316065	NM_031415	GSDMC	56169	-0.009	0.102	0.090	4.05×10^{-02} *	3.208	0.126	4553
GSE30229	cg05316065	NM_031415	GSDMC	56169	-0.018	0.067	0.273	9.64×10^{-03} **	4.642	0.125	566
GSE50409	cg05316065	NM_031415	GSDMC	56169	-0.011	0.080	0.139	3.65×10^{-02} *	3.310	0.126	3895
GSE57285	cg05316065	NM_031415	GSDMC	56169	-0.061	0.099	0.619	4.57×10^{-04} ***	7.690	0.124	334

*** Represented $p<0.0001$, ** represented $p<0.001$, and * represented $p<0.05$ in the subscripts associated to the p-Value of the probes above described.

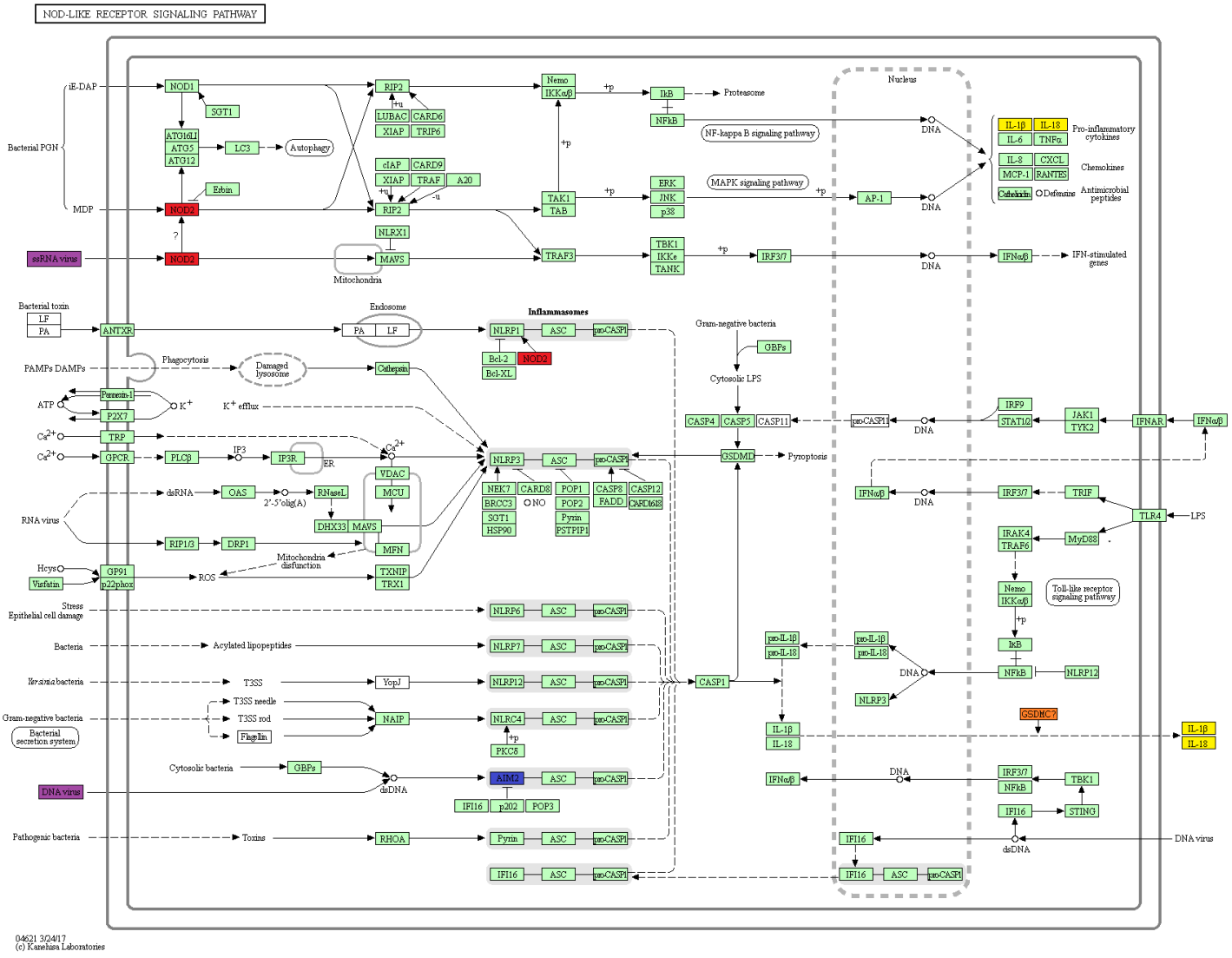

Figure 6 – Modified version of the image of NOD Receptor Signaling Pathway obtained from KEGG online platform (Kanehisa and Goto, 2000; Kanehisa *et al.*, 2016, 2017). NOD2 in red, AIM2 in blue, and GSMCD in orange as well as IL-1 beta and IL-18 in yellow as the byproducts of the pathway and the molecular triggers of the pathway in purple (ssRNA and dsDNA).

DISCUSSION

Restrictions of the research: number of assessed cancers and limitations of bisulphite-based approach

This research assessed four types of cancer, a minimal set of the total spectrum of cancers (ovarian, HNSCC, bladder, and breast cancers). Further experiments will be required to corroborate these results in other cancer types. Furthermore, despite the advantages of a methylation analysis mentioned in the introduction, there are two main disadvantages: (i) events of parallelization derived from the low complexity of the conversion of the un-methylated base and (ii) uncertainty about an exhaustive coverage of all CpG sites (Adusumalli *et al.*, 2014). Furthermore, this research was solely focused in CpG methylation with the Illumina Infinium Methylation 27 v1.2 array but the scope of the epigenetic marks assessed in this experiment is also low. Epigenetic modifications also include non-CpG sites such as CHG or CHH (where H correspond to A, T or C, especially relevant in hematopoietic cell lines) as well as acetylations (Meissner, 2010). Other methylation arrays could have been analyzed either from Illumina (Infinium HumanMethylation450 Beadchip Kit or Infinium MethylationEPIC Kit) or different suppliers (Agilent Technologies or Affymetrix) to increase the robustness of the experiment.

Validity of linear regression analysis

A linear regression analysis is a non-parametric approach that predicts the relationship of one or several independent variables and a dependent variable by generating a linear model. Requirements for this analysis are both linearity within the variables and normality within the residuals (CpG methylation ratios from ArrayExpress were already normalized). Furthermore, the number and magnitude of outliers (in our experiment, aberrant changes in the methylation of CpG sites as a consequence of the cancer) can bias the results by deviating the regression line in a particular direction (http://www.statsoft.com/Textbook/Multiple-Regression). The number and magnitude of the outliers was not consistent among the samples as a consequence of the singular influence of cancer in each particular patient (*Supplementary Figure 1*). The heterogeneity of a tumor applies to tumor cells but to the microenvironment as well (Galon *et al.*, 2014). Nevertheless, this problem was solved by performing a linear regression analysis that used the standardized criteria of using the identity function for all comparison in the $i \times j$ loop (see Methods, Linear regression analysis). This model hypothesized that the overall relationship of CpG methylation ratio had to remain identical except for several outliers (of our interest) despite of cancer presence. With this method, a null residual would have been obtained (Equation 1) when the CpG methylation

ratios were equal in cancer and non-cancer samples. A typical regression model would have assigned a non-null residual based on the deviated slope and intercept (Graphical demonstration in *Supplementary Figure 1*). Furthermore, this approach minimized the computing time in a dimensionally less complex algorithm. The coeficients of determination (Table 2) demonstrated that this approach was as useful as the unpaired Student's t-test and the stepwise ANN.

Validity of unpaired Student's t-test

An unpaired Student's t method for unequal variances is a parametric approach to test a null hypothesis: 'two means from independent, random samples from an approximately normal distribution are equal'. This is, the approach compared two independent averages (in our experiment, the average of the CpG methylation ratios for one probe in both cancer and control cases) with unequal variances and established a parameter (p-Value) to quantify how significantly different were these means with regards to a random event. In blood cell lines constitutive heterochromatin (genes with CpG methylation ratio close to 1) or constitutively active genes (CpG methylation ratio close to 0) were supposed to remain with its condition as a result of the presence of cancer (Saksouk, Simboeck and Déjardin, 2015). Nevertheless, conditionally methylated (intermediate CpG methylation ratio) genes were supposed to vary their CpG methylation ratio more often. As a consequence, the size of the residuals were assumed to vary, that is, the 'noise' or random disturbance in the relationship between the two variables. Hence, heteroscedasticity was also assumed for the samples. Additionally, when multiple hypothesis are tested, the probability of finding rare values increases in proportion to the number of tested associations (27,579 probes) and therefore the risk of incorrectly reject a null hypothesis. A Bonferroni correction could have compensated that risk by defining the significant threshold as the ratio between the desirable level of significance and the number of hypothesis (alpha/number of probes). Nevertheless, the stringency of this method resulted in no commonalities among all four datasets (note that no p-Value was minor than 0.05/27,579 except for cg26954174 in GSE19711 dataset in Table 3). Thus, the significance level was set to 0.05. Despite this low stringency to find commonalities among the datasets, three probes were identified (Figure 5). Considering the number of best predictor probes (ranking values in Table 3), the probability of finding a common probe for each was: (4,553/27,579) × (566/27,579) × (3895/27,579) × (334/27,579) for cg05316065; (416/27,579) × (614/27,579) × (1,660/27,579) × (4,890/27,579) for cg10636246; and (1,226/27,579) × (1,496/27,579) × (1,841 /27,579) × (1,362/27,579) for cg26954174. Based on p-Values, the probability of finding a common probe was: 6.51×10^{-09} for cg05316065, 3.16×10^{-12} for cg10636246, and 1.57×10^{-18} for cg26954174. The probability that the probes were shared in the four cancers as consequence of random chance was the independent product of both sets of probabilities: 1.7×10^{-16} based on ranking values and 3.23×10^{-38} based on p-Values. Furthermore, the three identified genes participated in the same pathway (Figure 6), increasing the robustness and the biological meaning of our findings.

Validity of stepwise ANN modelling algorithm

ANN are fault tolerant, capable to handle noisy, non-normal, non-linear and incomplete data, and capable of generalize hypothesis (Lancashire, Lemetre and Ball, 2009). Nevertheless, a problem that ANN face within biological datasets is 'the curse of dimensionality'. This is, the high complexity of '-omic' datasets as a consequence of the elevated number of Euclidean spaces (defined by the number of probes analyzed in our experiment). The stepwise ANN model overcomes this problem by using a single input at a time, building upwards, and subdividing the samples for testing the performance of the method (Lancashire, Lemetre and Ball, 2009). Another problem is the risk of overfitting. ANNs have a limited power in selecting important variables in small datasets (Lancashire, Lemetre and Ball, 2009). Concretely, GSE30229 (n=184) and GSE57285 (n=84) datasets had both a relative low number of samples. There is a risk that the ANN memorizes training cases, causing poor prediction performance in future cases. Thus, this results in an inefficient capability of generalization. Nonetheless, the stepwise algorithm applied an independent test to demonstrate its capability of generalization to new data. Hence, overfitting was not evident. Additionally, the number of steps was defined to 1, in other words, just one probe at each loop of the algorithm was tested to model the data. Considering just one probe, median accuracies were in all cases inferior to 80% (Figure 3). The intrinsic feature of the algorithm was not used: its capability to perform a stepwise analysis and generate a panel of probes that, combined, could predict at a 100% accuracy cancer as demonstrated in previous studies (Lowery *et al.*, 2009). By using the best input at the first step and sequentially adding the next best input in the next step to determine the combinatorial performance, higher accuracies and lesser median squared errors could have been reached. Further experiments with our data will be required to demonstrate that a panel of probes could potentially model the data and accurately predict cancer. Furthermore, no common probes were reported among the four described datasets and therefore no further analysis was conducted with this approach.

Threshold for the probe selection

As a consequence of the failure of the ANN to identify highly predictive probes for cancer (the accuracy was <80% in all cases), two further intra-comparisons (Venn diagram in Figure 5) were conducted with the probes comprised in the four *Positive Subgroups* on one hand or in the four *Negative Subgroups* on the other hand. The significance level (alpha) was set to 0.05 because it is a parametric value conventionally defined as 5% (less than 1 in 20 chance of being wrong) that could be easily applied to all discrete datasets. No internationally conventional definition exist for the thresholds of factor and ATE values. The fact of arbitrarily select the top 1000 o 5000 probes could lead to bias: this is, the 1000 probe of one dataset may be associated to a significance superior to 5%. Furthermore, a strong correlation was found among factor, p-Value and ATE within the probes comprised in the *Significant Group* (Table 2). Hence, no matter if probes were sorted the probes by either one or another value, the top probes had low *p-Values*, low *ATE*, and high *factor* values.

Identified common genes

AIM2. Absent in melanoma 2 transcript variant 1 (NM_004833), is a 38954 Da cytosolic protein member of the PYHIN or HIN200 family (Pruitt *et al.*, 2014). It is coded for a gene of the same name located on chromosome 1q23.2 (Pruitt *et al.*, 2014). Additionally, it is involved in innate immune response and mainly expressed in peripheral macrophages and leukocytes. AIM2 has two domains: a C-terminal HIN domain can bind dsDNA, and a PYD domain (Burckstummer *et al.*, 2009; Pruitt *et al.*, 2014; The UniProt Consortium, 2015). Upon non-sequence-specific recognition of exogenous dsDNA (either bacterial, viral, or host) in the cytosol, AIM2 recruits ASC via PYD-PYD (Pruitt *et al.*, 2014; The UniProt Consortium, 2015). ASC also contains a CARD domain that recruits procaspase-1 to the complex (Burckstummer *et al.*, 2009). This leads to the auto-activation of caspase-1, an enzyme that processes pro-inflammatory cytokines (e.g. IL-1B and IL-18) (Pruitt *et al.*, 2014; The UniProt Consortium, 2015).

NOD2. Nucleotide-binding oligomerization domain-containing protein 2 transcript variant 1 (NM_022162) is also known as CARD15 or inflammatory bowel disease protein 1 (IBD1) (Pruitt *et al.*, 2014). NOD2 is an 115283 Da protein coded for a gene of the same name located on chromosome 16q12.1 and mainly expressed in monocytes, macrophages, dendritic cells, and intestinal Paneth cells (Pruitt *et al.*, 2014). NOD2 has two CARD domains and eleven N-terminal LRR (Hsu *et al.*, 2008; The UniProt Consortium, 2015). Upon recognition of muramyl dipeptide or both synthetic and viral ssRNA, NOD2 is self-oligomerized. NOD1 and NOD2 signaling involve an interaction via CARD-CARD with RIPK2 (RIP2/RICK0) (Hsu *et al.*, 2008). This interaction recruits ubiquitin ligases such as XIAP, BIRC2, BIRC3 and the LUBAC complex, triggering activation of MAP kinases and activation of NF-kappa-B signaling

(Sabbah *et al.*, 2009). Molecular events that are positively regulated by NOD2 include: cell antigen processing and presenting activity of dendritic cells, prostaglandin-endoperoxide synthase activity, and IL-1B, IL-6, IL-8, IL-10, and IL-17 synthesis (Hsu *et al.*, 2008; Pruitt *et al.*, 2014; The UniProt Consortium, 2015). NOD2 is required for CASP1 activation and IL-1B release in macrophages (Pruitt *et al.*, 2014; The UniProt Consortium, 2015).

GSDMC. Gasdermin C with sequence (NM_031415) is also denominated MLZE. GSDMC has an N-terminal region that promotes pyroptosis (Ding *et al.*, 2016). Apart from the lack of information among curated databases and bibliography, the N-terminal moiety of GSDMC has been proposed to form homooligomers liked by disulfide bonds. This information was inferred due to the similarity of this mentioned N-terminal domain to the same domain of its paralog, GSDMD (Ding *et al.*, 2016). Additionally, GSDMC may form a 16-mer complex that creates pores of 10-15 nanometers in the cell plasma membrane (Ding *et al.*, 2016). This pore may facilitate the release of mature IL-1B (The UniProt Consortium, 2015).

Despite the low stringency of the threshold applied to obtain commonalities ($p<0.05$, top 5000 probes, ATE superior to 0.12 in Table 3), the probability of discovering not one but three common probes was 1.7×10^{-16} based on the Ranking Value (Equation 9) and 3.23×10^{-38} based on p-Values (Table 3). Furthermore, the three probes were related to the NOD Receptor Signaling Pathway, a dsDNA- and ssRNA-sensing pathway (Figure 6) annotated in KEGG online platform (Kanehisa and Goto, 2000; Kanehisa *et al.*, 2016, 2017). Discarding the idea that transformed cells recurrently express a tumour antigen similar to the potentially recognizable MDP 'epitope' by the NOD pathway (Kanehisa and Goto, 2000; Kanehisa *et al.*, 2016, 2017), the most plausible idea is that genomic and transcriptomic content shed by the neoplasm may trigger the demethylation of the AIM2, NOD2, and GSDMC.

Circulating-tumour nucleic acids

In 1948, Mandel and Métis described the presence of cell-free nucleic acid (cfNA) in human blood for the first time (Mandel and Metais, 1948). In 1977, increased cfDNA was reported in serum of cancer patients (Leon *et al.*, 1977). In 1989, raised concentrations of cfDNA were reported to be mainly originated from cancer cells and gave rise to the term circulating-tumour DNA (ctDNA) (Wan *et al.*, 2017). Since then, multiple investigations reported not only genomic cfDNA but also raised mitochondrial DNA, mRNA, and microRNA concentrations originated from tumour cell lines in cancer patients serum (Schwarzenbach, Hoon and Pantel, 2011; Wan *et al.*, 2017). These transcriptomic and genomic released content from transformed cells will be denominated circulating-tumor nucleic acids (ctNA) in this report. ctNA levels were demonstrated to be directly related to tumour size and stage (Wan *et al.*, 2017). The origins of all forms of cfNA are thought to be through a combination of apoptosis, tissue damage or necrosis and exosome secretion (Schwarzenbach, Hoon and Pantel, 2011; Wan *et al.*, 2017). Concretely, ctNA is produced by tumour cells and two main sources have been proposed: the primary tumor and viable circulating tumor cells within the bloodstream (Schwarzenbach, Hoon and Pantel, 2011; Alix-Panabières, Schwarzenbach and Pantel, 2012). This information supports our findings: genes involved in a dsDNA- and ssRNA-sensing pathway were significantly un-methylated in all four datasets, presumably in response to increased levels of cfNA in bloodstream. In healthy patients, homeostatic concentrations of cfNA in bloodstream can be justified by basal activities of nucleases and the combination of infiltrating macrophages, hepatic, and renal clearances (Qin *et al.*, 2016; Wan *et al.*, 2017). Tumour cells are characterized by its uncontrolled growth, and subsequently loss of viability as a consequence of their high metabolic demands that could eventually lead to the nucleic acid release into the bloodstream. Macrophages as literature suggest (Schwarzenbach, Hoon and Pantel, 2011; Wan *et al.*, 2017) and lymphocytes as our experiment suggest for GSE50229 (Langevin *et al.*, 2014) seem capable to detect high levels of both genomic and transcriptomic content presumably shed by transformed cells into the bloodstream. Upon ctNA release, internalization of ctNA in response to raised levels of cfNA may occur as *in vitro* experiments suggest (Gartler, 1959; Dvorakova *et al.*, 2013; Thierry *et al.*, 2016). Afterwards, the NOD Receptor Signaling Pathway may be triggered and AIM2, NOD2 and GSMDC genes de-methylated in a positive feedback for an efficient clearance of cytosolic ctNA. Possible consequences of the demethylation of these genes may be IL-1B and IL-18 production and secretion (Figure 6) (Kanehisa and Goto, 2000; Kanehisa *et al.*, 2016, 2017).

Further applicability of the findings

A study of 640 patients with various types and stages of cancer found an increase in the median ctDNA concentration in patients with stage IV disease compared with those who had stage I disease (Bettegowda *et al.*, 2014). Basal levels of ctNA may be maintained as a consequence of the increase of both macrophage and lymphocyte clearance activity in the early stages of the disease (Wan *et al.*, 2017). This activity may be a direct consequence of (i) the NOD pathway activation in peripheral blood cells and (ii) demethylation of its constituent genes (e.g. AIM2, GSDMC, and NOD2). Epigenetic marks are generally preserved or regenerated during cell division (Laird, 2010) and therefore may be a conserved and detectable signature in the genome for more advanced stages of the disease. On the other hand, raised levels of ctNA in advanced stages could be justified by immunoediting, when cancer cell lines are capable of 'escaping' the immune system (Wan *et al.*, 2017). In a Darwinian selection process, tumour cells not just acquire a way to circumvent the immune system and become 'invisible' but also they produce immunosuppressive environments (Schreiber, Old and Smyth, 2011). Hypothetically, ctNA clearance might be inefficient in immunocompromised patients. Nevertheless, in our experiment, we considered samples from patients with cancer as case study. Different stages of the disease were not discriminated. Further experiments could test a direct correlation between the progressive increase of the CpG methylation ratio in our discovered genes and the progression of the disease

There are promising advantages of ctNA detection (known as liquid biopsy) in clinic: early detection of cancer, real-time monitoring of disease burden because of the rapid turnover (16min to 2.5h) of cfNA, amendable to multiregional and reiterated sampling, low-invasive approach, potential monitoring of clonal cancer evolution, and response to therapy (Schwarzenbach, Hoon and Pantel, 2011; Pishvaian *et al.*, 2016; Wan *et al.*, 2017). Nevertheless, the disadvantages exceed the advantages: expensive technology in current development, background noise from wild-type sequences, low sensitivities due to sampling error (especially in early stages of cancer), unknown driver mutations of ctNA, and high ctNA levels in non-cancer-specific pathologies that may lead to false positives (e.g. acute traumas, infarction, exercise, transplantation, and mainly viral or bacterial infections) (Schwarzenbach, Hoon and Pantel, 2011). Hence, neither ctNA nor our reported genes (due to the low stringency applied to find commonalities) are likely to be provide a complete landscape of tumor heterogeneity. Nevertheless, here we have described five different proteins with their respective coding mRNAs that could compliment currently used biomarkers in the early detection of cancer: three under-methylated genes that will presumably be overexpressed in peripheral blood cell lines (AIM2, NOD2, and GSDMC) and two protein as a byproducts of the activation of the NOD Receptor Signaling Pathway (IL-1B and IL-18). An integrative analysis of methylation and transcription profiling combined with immunocytochemistry, mass spectrometry and Western blots may generate valuable data for the described stepwise ANN modelling approach to reach high accuracies in the early prediction of cancer. Possible drawbacks will arise as a consequence of low specificities derived from false positives from non-tumor pathologies that also trigger the NOD pathway. Nevertheless, this approach may become a complementary diagnosis in the early detection of cancer in combination with alternative methods.

CONCLUSION

The stepwise ANN, the linear regression analysis approach that also considered the standard deviation of the residuals, and the unpaired Student's t method for 2 tails and heteroscedastic samples and unequal variances demonstrated their potential to support, complement and increase the robustness in further identifications of CpG methylation changes to predict possible biomarkers with high levels of reliability, accuracy, sensitivity, specificity and clinical application.

Our hypothesis was tested positive: three gene promoters were significantly and recurrently under-methylated in peripheral blood cell lines as a result of cancer. Even though the function of GSMDC remains unclear, the all three reported genes seem to participate in the NOD Receptor Signaling Pathway, presumably triggered by high levels of genomic and transcriptomic content that can be shed due to an increase in apoptosis, necrosis and DNA exocytosis events. Thus, these genes can be considered in further experiments to test their potentiality for early detection of cancer. An integrative analysis of methylation and transcription profiles coupled with immunocytochemistry assays, mass spectrometry and western blot analysis of with all described genes (AIM2, NOD2, GSDMC, IL-1B, and IL-18) may generate valuable data for a further analysis in the stepwise ANN. Nevertheless, specificities, sensitivities and accuracies of these possible biomarkers should be rigorously tested in: (i) the same cancers to test our results, (ii) other types of cancer, and (ii) other non-tumour pathologies to corroborate our results.

ACKNOWLEDGEMENTS

We would like to thank L. J. Lancashire for the development of the stepwise ANN model. This study was supported by the John Van Geest Cancer Research Centre, Nottingham Trent University.

REFERENCES

Abdel-Fatah, T. M. A., Agarwal, D., Liu, D.-X., Russell, R., Rueda, O. M., Liu, K., Xu, B., Moseley, P. M., Green, A. R., Pockley, A. G., Rees, R. C., Caldas, C., Ellis, I. O., Ball, G. R. and Chan, S. Y. T. (2017) 'SPAG5 as a prognostic biomarker and chemotherapy sensitivity predictor in breast cancer: a retrospective, integrated genomic, transcriptomic, and protein analysis', *The Lancet Oncology*. Elsevier, 17(7), pp. 1004–1018. doi: 10.1016/S1470-2045(16)00174-1.

Adusumalli, S., Omar, M. F. M., Soong, R. and Benoukraf, T. (2014) 'Methodological aspects of whole-genome bisulfite sequencing analysis', *Briefings in Bioinformatics*, 16(3), pp. 369–379. doi: 10.1093/bib/bbu016.

Albarakati, N., Abdel-Fatah, T. M. A., Doherty, R., Russell, R., Agarwal, D., Moseley, P., Perry, C., Arora, A., Alsubhi, N., Seedhouse, C., Rakha, E. A., Green, A., Ball, G., Chan, S., Caldas, C., Ellis, I. O. and Madhusudan, S. (2015) 'Targeting BRCA1-BER deficient breast cancer by ATM or DNA-PKcs blockade either alone or in combination with cisplatin for personalized therapy', *Molecular Oncology*, 9(1), pp. 204–217. doi: 10.1016/j.molonc.2014.08.001.

Alix-Panabières, C., Schwarzenbach, H. and Pantel, K. (2012) 'Circulating Tumor Cells and Circulating Tumor DNA', *Annual Review of Medicine*, 63(1), pp. 199–215. doi: 10.1146/annurev-med-062310-094219.

Anjum, S., Fourkala, E.-O., Zikan, M., Wong, A., Gentry-Maharaj, A., Jones, A., Hardy, R., Cibula, D., Kuh, D., Jacobs, I. J., Teschendorff, A. E., Menon, U. and Widschwendter, M. (2014) 'A BRCA1-mutation associated DNA methylation signature in blood cells predicts sporadic breast cancer incidence and survival', *Genome Medicine*. BioMed Central, 6(6), p. 47. doi: 10.1186/gm567.

Bettegowda, C., Sausen, M., Leary, R. J., Kinde, I., Wang, Y., Agrawal, N., Bartlett, B. R., Wang, H., Luber, B., Alani, R. M., Antonarakis, E. S., Azad, N. S., Bardelli, A., Brem, H., Cameron, J. L., Lee, C. C., Fecher, L. A., Gallia, G. L., Gibbs, P., Le, D., Giuntoli, R. L., Goggins, M., Hogarty, M. D., Holdhoff, M., Hong, S.-M., Jiao, Y., Juhl, H. H., Kim, J. J., Siravegna, G., Laheru, D. A., Lauricella, C., Lim, M., Lipson, E. J., Marie, S. K. N., Netto, G. J., Oliner, K. S., Olivi, A., Olsson, L., Riggins, G. J., Sartore-Bianchi, A., Schmidt, K., Shih, le-M., Oba-Shinjo, S. M., Siena, S., Theodorescu, D., Tie, J., Harkins, T. T., Veronese, S., Wang, T.-L., Weingart, J. D., Wolfgang, C. L., Wood, L. D., Xing, D., Hruban, R. H., Wu, J., Allen, P. J., Schmidt, C. M., Choti, M. A., Velculescu, V. E., Kinzler, K. W., Vogelstein, B., Papadopoulos, N. and Diaz, L. A. (2014) 'Detection of Circulating Tumor DNA in Early- and Late-Stage Human Malignancies', *Science translational medicine*, 6(224), p. 224ra24-224ra24. doi: 10.1126/scitranslmed.3007094.

Burckstummer, T., Baumann, C., Bluml, S., Dixit, E., Durnberger, G., Jahn, H., Planyavsky, M., Bilban, M., Colinge, J., Bennett, K. L. and Superti-Furga, G. (2009) 'An orthogonal proteomic-genomic screen identifies AIM2 as a cytoplasmic DNA sensor for the inflammasome.', *Nature immunology*, 10(3), pp. 266–272. doi: 10.1038/ni.1702.

Ding, J., Wang, K., Liu, W., She, Y., Sun, Q., Shi, J., Sun, H., Wang, D. and Shao, F. (2016) 'Pore-forming activity and structural autoinhibition of the gasdermin family', *Nature*, 535(7610), pp. 111–116. doi: 10.1038/nature18590.

Dvorakova, M., Karafiat, V., Pajer, P., Kluzakova, E., Jarkovska, K., Pekova, S., Krutilkova, L. and Dvorak, M. (2013) 'DNA released by leukemic cells contributes to the disruption of the bone marrow microenvironment', *Oncogene*. Macmillan Publishers Limited, 32(44), pp. 5201–5209. Available at: http://dx.doi.org/10.1038/onc.2012.553.

Ebi.ac.uk. (2017). *ArrayExpress < EMBL-EBI*. [online] Available at: https://www.ebi.ac.uk/arrayexpress/ [Accessed 20 Apr. 2017].

Ebi.ac.uk. (2017). *E-GEOD-19711 < Browse < ArrayExpress < EMBL-EBI*. [online] Available at: http://www.ebi.ac.uk/arrayexpress/experiments/E-GEOD-19711/ [Accessed 20 Apr. 2017].

Ebi.ac.uk. (2017). *E-GEOD-30229 < Browse < ArrayExpress < EMBL-EBI*. [online] Available at: https://www.ebi.ac.uk/arrayexpress/experiments/E-GEOD-30229/ [Accessed 20 Apr. 2017].

Ebi.ac.uk. (2017). *E-GEOD-50409 < Browse < ArrayExpress < EMBL-EBI*. [online] Available at: https://www.ebi.ac.uk/arrayexpress/experiments/E-GEOD-50409/ [Accessed 20 Apr. 2017].

Ebi.ac.uk. (2017). *E-GEOD-57285 < Browse < ArrayExpress < EMBL-EBI*. [online] Available at: https://www.ebi.ac.uk/arrayexpress/experiments/E-GEOD-57285/ [Accessed 20 Apr. 2017].

Fisher, R., Pusztai, L. and Swanton, C. (2013) 'Cancer heterogeneity: implications for targeted therapeutics', *British Journal of Cancer*, 108(3), pp. 479–485. doi: 10.1038/bjc.2012.581.

Galon, J., Mlecnik, B., Bindea, G., Angell, H. K., Berger, A., Lagorce, C., Lugli, A., Zlobec, I., Hartmann, A., Bifulco, C., Nagtegaal, I. D., Palmqvist, R., Masucci, G. V, Botti, G., Tatangelo, F., Delrio, P., Maio, M., Laghi, L., Grizzi, F., Asslaber, M., D'Arrigo, C., Vidal-Vanaclocha, F., Zavadova, E., Chouchane, L., Ohashi, P. S., Hafezi-Bakhtiari, S., Wouters, B. G., Roehrl, M., Nguyen, L., Kawakami, Y., Hazama, S., Okuno, K., Ogino, S., Gibbs, P., Waring, P., Sato, N., Torigoe, T., Itoh, K., Patel, P. S., Shukla, S. N., Wang, Y., Kopetz, S., Sinicrope, F. A., Scripcariu, V., Ascierto, P. A., Marincola, F. M., Fox, B. A. and Pagès, F. (2014) 'Towards the introduction of the "Immunoscore" in the classification of malignant tumours', *The Journal of Pathology*. Chichester, UK: John Wiley & Sons, Ltd, 232(2), pp. 199–209. doi: 10.1002/path.4287.

Gartler, S. M. (1959) 'Cellular Uptake of Deoxyribonucleic Acid by Human Tissue Culture Cells', *Nature*, 184(4697), pp. 1505–1506. Available at: http://dx.doi.org/10.1038/1841505a0.

Genome.jp. (2017). *KEGG PATHWAY Database*. [online] Available at: http://www.genome.jp/kegg/pathway.html [Accessed 20 Apr. 2017].

Greaves, M. and Maley, C. C. (2012) 'Clonal evolution in cancer', *Nature*, 481(7381), pp. 306–313. doi: 10.1038/nature10762.

Hanahan, D. and Weinberg, R. A. (2011) 'Hallmarks of Cancer: The Next Generation', *Cell*, 144(5), pp. 646–674. doi: http://dx.doi.org/10.1016/j.cell.2011.02.013.

Hsu, L.-C., Ali, S. R., McGillivray, S., Tseng, P.-H., Mariathasan, S., Humke, E. W., Eckmann, L., Powell, J. J., Nizet, V., Dixit, V. M. and Karin, M. (2008) 'A NOD2-NALP1 complex mediates caspase-1-dependent IL-1beta secretion in response to Bacillus anthracis infection and muramyl dipeptide.', *Proceedings of the National Academy of Sciences of the United States of America*, 105(22), pp. 7803–8. doi: 10.1073/pnas.0802726105.

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. doi: 10.1093/nar/gkw1092.

Kanehisa, M. and Goto, S. (2000) 'KEGG: Kyoto Encyclopedia of Genes and Genomes', *Nucleic Acids Research*. Oxford, UK: Oxford University Press, 28(1), pp. 27–30. Available at: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC102409/.

Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. and Tanabe, M. (2016) 'KEGG as a reference resource for gene and protein annotation', *Nucleic Acids Research*. Oxford University Press, 44(Database issue), pp. D457–D462. doi: 10.1093/nar/gkv1070.

Laird, P. W. (2010) 'Principles and challenges of genome-wide DNA methylation analysis', *Nat Rev Genet*. Nature Publishing Group, 11(3), pp. 191–203. Available at: http://dx.doi.org/10.1038/nrg2732.

Lancashire, L. J., Lemetre, C. and Ball, G. R. (2009) 'An introduction to artificial neural networks in bioinformatics—application to complex microarray and mass spectrometry datasets in cancer studies', *Briefings in Bioinformatics*, 10(3), pp. 315–329. Available at: http://dx.doi.org/10.1093/bib/bbp012.

Lancashire, L. J., Powe, D. G., Reis-Filho, J. S., Rakha, E., Lemetre, C., Weigelt, B., Abdel-Fatah, T. M., Green, A. R., Mukta, R., Blamey, R., Paish, E. C., Rees, R. C., Ellis, I. O. and Ball, G. R. (2010) 'A validated gene expression profile for detecting clinical outcome in breast cancer using artificial neural networks', *Breast Cancer Research and Treatment*, 120(1), pp. 83–93. doi: 10.1007/s10549-009-0378-1.

Lancashire, L. J., Rees, R. C. and Ball, G. R. (2008) 'Identification of gene transcript signatures predictive for estrogen receptor and lymph node status using a stepwise forward selection artificial neural network modelling approach', *Artificial Intelligence in Medicine*. Elsevier, 43(2), pp. 99–111. doi: 10.1016/j.artmed.2008.03.001.

Langevin, S. M., Houseman, E. A., Accomando, W. P., Koestler, D. C., Christensen, B. C., Nelson, H. H., Karagas, M. R., Marsit, C. J., Wiencke, J. K. and Kelsey, K. T. (2014) 'Leukocyte-adjusted epigenome-wide association studies of blood from solid tumor patients', *Epigenetics*. Landes Bioscience, 9(6), pp. 884–895. doi: 10.4161/epi.28575.

Langevin, S. M., Koestler, D. C., Christensen, B. C., Butler, R. A., Wiencke, J. K., Nelson, H. H., Houseman, E. A., Marsit, C. J. and Kelsey, K. T. (2012) 'Peripheral blood DNA methylation profiles are indicative of head and neck squamous cell carcinoma: An epigenome-wide association study', *Epigenetics*. Landes Bioscience, 7(3), pp. 291–299. doi: 10.4161/epi.7.3.19134.

Leon, S., Shapiro, B., Sklaroff, D. and Yaros, M. (1977) 'Free DNA in the serum of cancer patients and the effect of therapy', *Cancer research*, 37, pp. 646–650. Available at: http://cancerres.aacrjournals.org/content/37/3/646.short.

Liotta, L. A., Ferrari, M. and Petricoin, E. (2003) 'Clinical proteomics: Written in blood', *Nature*, 425(6961), p. 905. Available at: http://dx.doi.org/10.1038/425905a.

Lowery, A. J., Miller, N., Devaney, A., McNeill, R. E., Davoren, P. A., Lemetre, C., Benes, V., Schmidt, S., Blake, J., Ball, G. and Kerin, M. J. (2009) 'MicroRNA signatures predict oestrogen receptor, progesterone receptor and HER2/neu receptor status in breast cancer.', *Breast cancer research : BCR*, 11(3), p. R27. doi: 10.1186/bcr2257.

Mandel, P. and Metais, P. (1948) 'Les acides nucléiques du plasma sanguin chez l'homme', *C. R. Seances Soc. Biol. Ses Fil.*, 142(3–4), pp. 241–243. doi: 10.1007/BF00832140.

Meissner, A. (2010) 'Epigenetic modifications in pluripotent and differentiated cells.', *Nature biotechnology*, 28(10), pp. 1079–88. doi: 10.1038/nbt.1684.

Mohr, S. and Liew, C.-C. (2007) 'The peripheral-blood transcriptome: new insights into disease and risk assessment', *Trends in Molecular Medicine*, 13(10), pp. 422–432. doi: http://dx.doi.org/10.1016/j.molmed.2007.08.003.

Network, T. C. G. A. R., Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., Ellrott, K., Shmulevich, I., Sander, C. and Stuart, J. M. (2013) 'The Cancer Genome Atlas Pan-Cancer analysis project', *Nat Genet*. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved., 45(10), pp. 1113–1120. Available at: http://dx.doi.org/10.1038/ng.2764.

Parkinson, H., Kapushesky, M., Shojatalab, M., Abeygunawardena, N., Coulson, R., Farne, A., Holloway, E., Kolesnykov, N., Lilja, P., Lukk, M., Mani, R., Rayner, T., Sharma, A., William, E., Sarkans, U. and Brazma, A. (2007) 'ArrayExpress—a public database of microarray experiments and gene expression profiles', *Nucleic Acids Research*, 35(suppl_1), pp. D747–D750. Available at: http://dx.doi.org/10.1093/nar/gkl995.

Petricoin, E. F., Belluco, C., Araujo, R. P. and Liotta, L. A. (2006) 'The blood peptidome: a higher dimension of information content for cancer biomarker discovery', *Nat Rev Cancer*. Nature Publishing Group, 6(12), pp. 961–967. Available at: http://dx.doi.org/10.1038/nrc2011.

Pishvaian, M. J., Bender, R. J., Matrisian, L. M., Rahib, L., Hendifar, A., Hoos, W. A., Mikhail, S., Chung, V., Picozzi, V., Heartwell, C., Mason, K., Varieur, K., Aberra, M., Madhavan, S., Petricoin, E. 3rd and Brody, J. R. (2016) 'A pilot study evaluating concordance between blood-based and patient-matched tumor molecular testing within pancreatic cancer patients participating in the Know Your Tumor (KYT) initiative.', *Oncotarget*. United States. doi: 10.18632/oncotarget.13225.

Pruitt, K. D., Brown, G. R., Hiatt, S. M., Thibaud-Nissen, F., Astashyn, A., Ermolaeva, O., Farrell, C. M., Hart, J., Landrum, M. J., McGarvey, K. M., Murphy, M. R., O'Leary, N. A., Pujar, S., Rajput, B., Rangwala, S. H., Riddick, L. D., Shkeda, A., Sun, H., Tamez, P., Tully, R. E., Wallin, C., Webb, D., Weber, J., Wu, W., Dicuccio, M., Kitts, P., Maglott, D. R., Murphy, T. D. and Ostell, J. M. (2014) 'RefSeq: An update on mammalian reference sequences', *Nucleic Acids Research*, 42(D1). doi: 10.1093/nar/gkt1114.

Qin, Z., Ljubimov, V. A., Zhou, C., Tong, Y. and Liang, J. (2016) 'Cell-free circulating tumor DNA in cancer', *Chinese Journal of Cancer*. doi: 10.1186/s40880-016-0092-4.

Sabbah, A., Chang, T. H., Harnack, R., Frohlich, V., Tominaga, K., Dube, P. H., Xiang, Y. and Bose, S. (2009) 'Activation of innate immune antiviral responses by Nod2.', *Nature immunology*, 10(10), pp. 1073–80. doi: 10.1038/ni.1782.

Saksouk, N., Simboeck, E. and Déjardin, J. (2015) 'Constitutive heterochromatin formation and transcription in mammals.', *Epigenetics & chromatin*, 8, p. 3. doi: 10.1186/1756-8935-8-3.

Schreiber, R. D., Old, L. J. and Smyth, M. J. (2011) 'Cancer Immunoediting: Integrating Immunity's Roles in Cancer Suppression and Promotion', *Science*, 331(6024), pp. 1565–1570. doi: 10.1126/science.1203486.

Schwarzenbach, H., Hoon, D. S. B. and Pantel, K. (2011) 'Cell-free nucleic acids as biomarkers in cancer patients', *Nat Rev Cancer*, 11(6), pp. 426–437. doi: 10.1038/nrc3066.

Statsoft.com. (2017). *Multiple Regression*. [online] Available at:

http://www.statsoft.com/Textbook/Multiple-Regression [Accessed 20 Apr. 2017].

Sterck, L. (2017). *Draw Venn Diagram*. [online] Bioinformatics.psb.ugent.be. Available at: http://bioinformatics.psb.ugent.be/webtools/Venn/ [Accessed 20 Apr. 2017].

Teschendorff, A. E., Menon, U., Gentry-Maharaj, A., Ramus, S. J., Gayther, S. A., Apostolidou, S., Jones, A., Lechner, M., Beck, S., Jacobs, I. J. and Widschwendter, M. (2009) 'An Epigenetic Signature in Peripheral Blood Predicts Active Ovarian Cancer', *PLOS ONE*. Public Library of Science, 4(12), p. e8274. Available at: http://dx.doi.org/10.1371%2Fjournal.pone.0008274.

The UniProt Consortium (2015) 'UniProt: a hub for protein information.', *Nucleic acids research*, 43(Database issue), pp. D204-12. doi: 10.1093/nar/gku989.

Thierry, A. R., El Messaoudi, S., Gahan, P. B., Anker, P. and Stroun, M. (2016) 'Origins, structures, and functions of circulating DNA in oncology', *Cancer and Metastasis Reviews*, 35(3), pp. 347–376. doi: 10.1007/s10555-016-9629-x.

Wan, J. C. M., Massie, C., Garcia-Corbacho, J., Mouliere, F., Brenton, J. D., Caldas, C., Pacey, S., Baird, R. and Rosenfeld, N. (2017) 'Liquid biopsies come of age: towards implementation of circulating tumour DNA', *Nat Rev Cancer*. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved., 17(4), pp. 223–238. Available at: http://dx.doi.org/10.1038/nrc.2017.7.

Supplementary material

Table 1 – GEO of the Illumina Infinium methylation array, number of loop for 1 step of the stepwise ANN, Illumina Infinium ID probe, average train performance, average test performance, average train error, average test error (ATE), and average valid error obtained from the stepwise ANN.

				GSE19711			
Loop	Input ID	Average Train Perf	Average Test Perf	Average Valid Perf	Average Train Error	Average Test Error	Average Valid Error
1	cg25634666	0.674	0.685	0.667	0.108	0.107	0.109
2	cg11822932	0.647	0.657	0.644	0.110	0.107	0.111
3	cg25634666	0.671	0.676	0.662	0.107	0.106	0.110
4	cg25634666	0.676	0.671	0.676	0.107	0.107	0.108
5	cg00645579	0.670	0.685	0.667	0.108	0.106	0.108
6	cg00645579	0.673	0.685	0.671	0.108	0.106	0.110
7	cg24777950	0.660	0.667	0.657	0.108	0.107	0.109
8	cg25634666	0.670	0.676	0.676	0.108	0.108	0.107
9	cg25634666	0.670	0.671	0.657	0.107	0.107	0.109
10	cg00645579	0.673	0.676	0.676	0.108	0.107	0.108
				GSE30229			
Loop	Input ID	Average Train Perf	Average Test Perf	Average Valid Perf	Average Train Error	Average Test Error	Average Valid Error
1	cg23547429	0.736	0.757	0.757	0.095	0.091	0.091
2	cg23547429	0.741	0.757	0.757	0.093	0.093	0.095
3	cg23547429	0.755	0.757	0.730	0.092	0.091	0.100
4	cg07285167	0.723	0.757	0.730	0.096	0.091	0.097
5	cg12089698	0.723	0.730	0.703	0.095	0.093	0.100
6	cg22242539	0.727	0.757	0.730	0.094	0.091	0.097
7	cg02679745	0.736	0.770	0.730	0.093	0.089	0.098
8	cg25623459	0.727	0.757	0.703	0.095	0.090	0.099
9	cg23547429	0.745	0.757	0.730	0.092	0.090	0.098
10	cg17709873	0.718	0.730	0.703	0.095	0.091	0.099
				GSE50409			
Loop	Input ID	Average Train Perf	Average Test Perf	Average Valid Perf	Average Train Error	Average Test Error	Average Valid Error
1	cg25307081	0.634	0.622	0.612	0.114	0.114	0.117
2	cg25307081	0.619	0.622	0.612	0.115	0.115	0.116
3	cg25307081	0.623	0.628	0.600	0.115	0.115	0.117
4	cg25307081	0.621	0.616	0.624	0.115	0.115	0.116
5	cg25307081	0.623	0.616	0.612	0.115	0.115	0.115
6	cg25307081	0.621	0.622	0.618	0.114	0.115	0.117
7	cg25307081	0.630	0.628	0.624	0.114	0.115	0.116
8	cg25307081	0.626	0.628	0.612	0.115	0.116	0.117
9	cg25307081	0.619	0.628	0.612	0.115	0.115	0.116
10	cg25307081	0.626	0.622	0.624	0.115	0.114	0.116
				GSE57285			
Loop	Input ID	Average Train Perf	Average Test Perf	Average Valid Perf	Average Train Error	Average Test Error	Average Valid Error
1	cg22892110	0.780	0.765	0.765	0.085	0.086	0.086
2	cg22892110	0.800	0.765	0.765	0.082	0.087	0.089
3	cg22892110	0.770	0.765	0.765	0.087	0.078	0.086
4	cg22892110	0.780	0.765	0.765	0.085	0.087	0.086
5	cg22892110	0.780	0.794	0.765	0.087	0.079	0.087
6	cg22892110	0.780	0.765	0.765	0.082	0.081	0.089
7	cg22892110	0.780	0.765	0.765	0.085	0.082	0.089
8	cg22892110	0.800	0.765	0.765	0.083	0.086	0.089
9	cg22892110	0.780	0.765	0.765	0.086	0.085	0.088
10	cg21505886	0.760	0.765	0.765	0.093	0.088	0.092

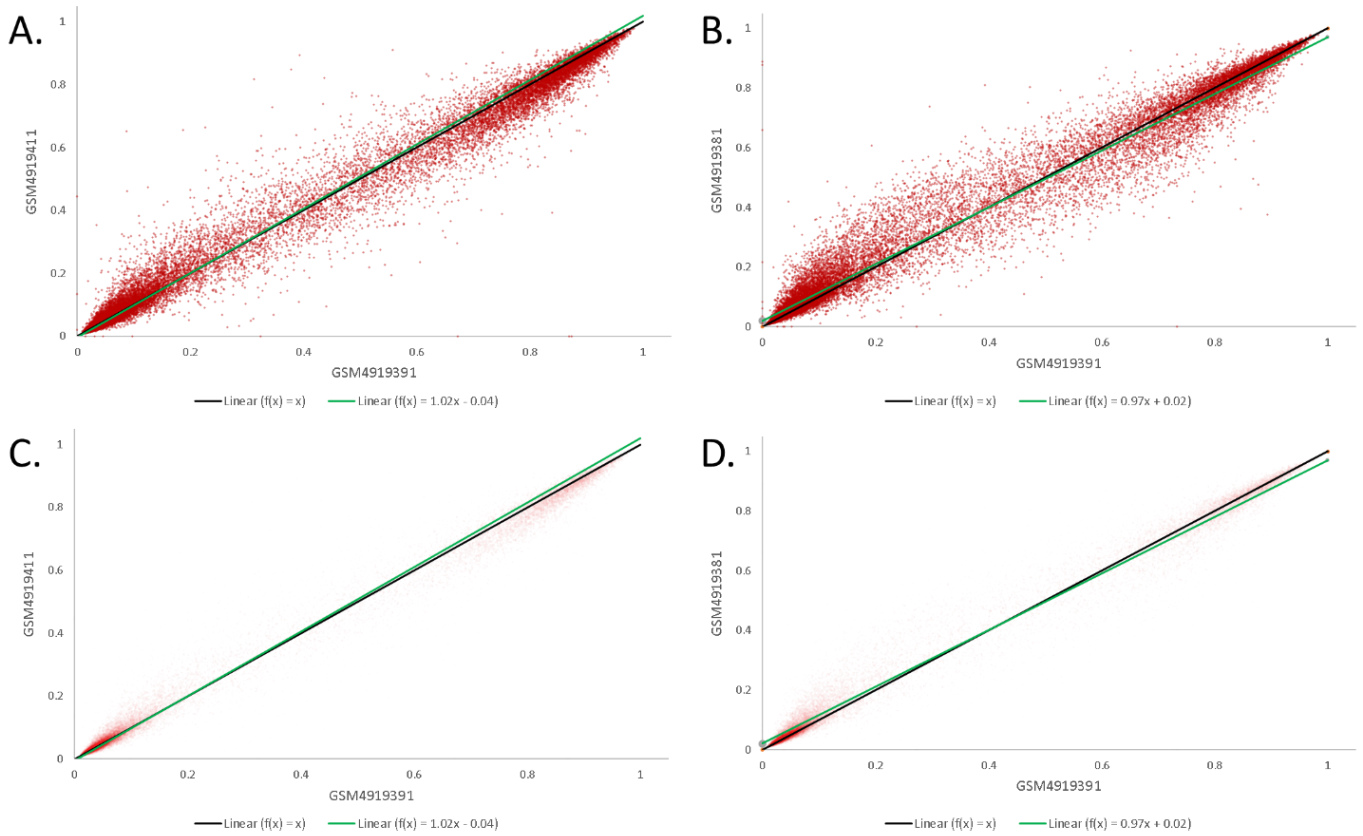

Figure 1 – Representation of the normalized CpG methylation ratios for all Illumina Infinium 27k Human Methylation Beadchip v1.2 probes of three samples comprised in GSE19711 dataset: CpG methylation ratios for all 27,579 probes from control sample GSM491939 in the X-axis (A, B, C, D) versus CpG methylation ratios for all probes from cancer case samples GSM491941 1 (A, C) and GSM491938 1 (B, D) in the Y-axis. The transparency of the point markers was increased from 50% (A, B) to 98% (C, D) to distinguish both linear functions: calculated regression line in green and identity function in black.