Gene expression model for the classification of human colorectal cancer and potential CRC biomarkers search

García Amaia[1], Freije Ana[1], Armañanzas Rubén[2], Inza Iñaki[2], Ispizua Ziortza[1], Heredia Pedro[1], Larrañaga Pedro[1], López Vivanco Guillermo[3], Calvo Begoña[3], Suárez Tatiana[1],

Betanzos Mónica[1]

[1.]Área de Biotecnología, Centro Tecnológico GAIKER, Parque Tecnológico, Edificio 202, 48170 Zamudio, Bizkaia, Spain.
[2.]Departament of Computer Science and Artificial Intelligence, University of the Basque Country, P.O Box 649, E-20080 Donostia-San Sebastián, Spain
[3.]Servicio de Oncología Médica, Plaza de Cruces s/n, 48903, Barakaldo, Bizkaia, Spain.

ABSTRACT

A genomic study of human colorectal cancer has been carried out on a total of 32 tumoral, corresponding to different stages of the disease, and 33 non-tumoral samples.
An exhaustive analysis of the quality and quantity of the RNA obtained was made with the Agilent 2100 Bioanalyzer. Only those samples fulfilling an RNA standard of quality were selected for hybridization.
Gene expression study was performed by hybridization of the tumour samples against a pool obtained from the non tumoral samples. We used the Human 1A 60-mer oligomicrorray belonging to Agilent Platform.
In the subsequent bioinformatic study, six different statistical metrics were computed and a consensus on a unique importance probe ranking was reached. This ranking can show the statistical univariate relevance of each probe in a class-prediction problem over this dataset. The statistical model achieved correctly classifies samples in non-tumoral and tumoral categories, and three tumoral stages (B, C and D).
The validation process is carried out on the seven genes with high ranking positions in the tentative model, through real time PCR and Taqman gene expression assays in 15 new colorectal cancer samples.

INTRODUCTION

Colorectal cancer (CRC) is the second cause of cancer death in western countries. The success of the therapy depends on an early diagnostic, on the knowledge of the biological behaviour in each tumor and its susceptibility to drugs. DNA microarray technology allows the measure of the mRNA expression level of thousands of genes simultaneously.

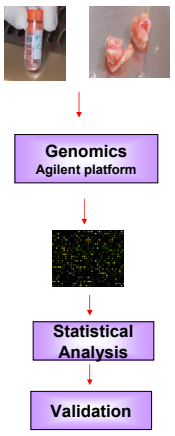

Fig. 2. Diagram of the experiment carried out.

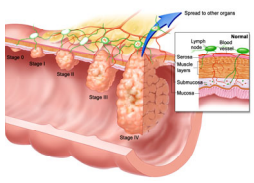

Fig. 1. Progresion of colorectal cancer

MATERIALS AND METHODS

Tissues and patients

A total of 133 tissue samples (120 form patients with CRC in different stages and 13 samples from patients with no colorectal cancer) were obtained from Cruces Hospital (BIOEF). The 120 CRC samples consisted in 60 tumour samples and 60 paired non tumoral samples.

Patient	TNM	DUKES	Age	Sex	Patient	TNM	DUKES	Age	Sex
M1	IV	D	46	F	M2	IV	D	46	F
M11	IIIA	C1	87	F	M12	IIIA	C1	87	F
M14	IIIC	C2	68	F	M13	IIIC	C2	68	F
M20	IV	D	77	F	M19	IV	D	77	F
M25	IIA	B2	71	F	M26	IIA	B2	71	F
M35	IIIB	C2	73	F	M36	IIIB	C2	73	F
M41	IIA	B2	66	F	M42	IIA	B2	66	F
M55	IIIB	C3	67	F	M56	IIIB	C3	67	F
M65	IIIB	C2	71	F	M66	IIIB	C2	71	F
M88	IIA	B2	76	F	M87	IIA	B2	76	F
M90	IIIB	C3	57	F	M89	IIIB	C3	57	F
M106	IIA	B2	75	F	M105	IIA	B2	75	F
M112	IIIA	C1	62	F	M109	IV	D	69	F
M116	IV	D	72	F	M111	IIIA	C1	62	F
M5	IIA	B2	69	M	M6	IIA	B2	69	M
M8	IIB	B3	68	M	M7	IIB	B3	68	M
M10	IV	D	63	M	M15	IV	D	81	M
M22	IA	B1	73	M	M16	IV	D	81	M
M23	IIIC	C3	47	M	M21	IA	B1	73	M
M29	IIA	B2	72	M	M24	IIIC	C3	47	M
M37	IIB	B3	73	M	M40	IIA	B2	55	M
M39	IIA	B2	55	M	M44	IA	B1	77	M
M43	IA	B1	77	M	M50	IA	B1	46	M
M49	IA	B1	46	M	M54	IV	D	50	M
M51	IIIC	C2	47	M	M59	IV	D	83	M
M53	IV	D	50	M	M63	IV	D	74	M
M60	IV	D	83	M	M69	IIIB	C2	65	M
M64	IV	D	74	M	M72	IIA	B2	57	M
M70	IIIB	C2	65	M	M81	IIA	B2	71	M
M71	IIA	B2	57	M	M95	IIA	B2	63	M
M74	IA	B1	71	M	M100			76	M
M78	IA	B1	68	M	M119	IIIB	C2	67	M
M80	IA	B1	59	M	M9	IV	D	63	M
M83	IIIC	C2	70	M					
M94	IIA	B2	63	M					
M99			76	M					
M104	IIA	B2	78	M					
M107	IIA	B2	60	M					
M120	IIIB	C2	67	M					
M98									
M114									

Table 1. Clinical and pathologic data of patient tumor (rigth) and non-tumor(left).

Genomic Analysis

Tissue samples were preserved in *RNA later Stabilization Solution* (Qiagen) and stored at -80ºC.
RNA extraction. Total RNA was extracted from all the samples using the *RNAeasy Mini Kit (Qiagen)*. RNA quality and quantity was determined with the *Agilent 2100 Bioanalyzer (Agilent Technologies)*. We used the RIN algorithm (RNA Integrity Number, *Agilent Technologies*) as a quality standard to select the samples. We synthesized and labelled the cRNA using the *Agilent Low RNA Input Fluorescent Linear Amplification Kit (Agilent Technologies)*
cRNA hybridization. The selected samples were hybridized on to the *Agilent Human 1A 60-mer oligo microarrays (Agilent Technologies*) and the microarrays were scanned using the *GenePix 4000B Scan (Axon Instruments)*. Images were analyzed with Genepix 6.0 (Axon Instruments) and data were filtered and normalized with Acuity 4.1 (Axon Instruments).
Experimental Design. Tumoral and non-tumoral samples were hybridized against a common RNA control pool. As none of the control non-cancerous samples presented an acceptable RNA quality they were discarded and we collected RNA from paired non-tumoral samples to form the NT Pool.The tumoral samples were labelled using Cy-5 dye (red) and the "pool" was labelled with Cy-3 dye (green)[1]:

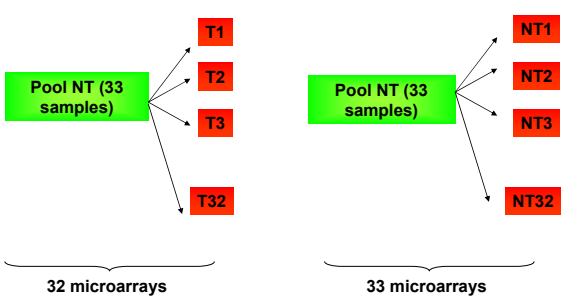

Fig. 3. Hybridization design

Data Statistical Analysis

Spot quality metrics. Reliability in the microarray probes are tackled by applying three different widely used quality metrics[2]: *fluorescent intensity measurement* quality, *background flatness* quality and *signal intensity consistency* quality. In basis of these three metrics a global quality metric with values between 0 and 1 is computed for each spot in each microarray.
Imputation of lost values. Collateral undesirable problems, such as small fibres inside the array, or an incomplete hybridization, can cause a spot value to be lost. In order to complete all these lost values we used the *KNNImpute*[3] procedure which has been proven as one of the best imputation techniques in the microarray domain.
Intraclass ratio differences. It is not expectable to find big differences between the expression ratios of a gen inbetween the same type of tissue. But, due to the heterogeneity of the cells included in the biopsias, genes with expression differences bigger than *2-fold* in the same kind of tissue are discarded.
Global machine learning approach. On the basis of the CRC state of each patient, we propose a supervised classification problem, or class prediction problem. The classification dataset is then composed of 64 instances from four different classes with cardinalities: 33 non-tumour, 13 Dukes B, 10 Dukes C and 8 Dukes D.
Discretisation policy. To apply the following statistical techniques the continuous expression values have to be discretise. Attending to the expected biological behaviour -under, baseline or over expressed-, the values are discretised using an *Equal Width* policy with three intervals.
Univariate statistical metrics. Using the supervised approach we can univariatelly measure the relevance of each gen (from now on called *variable*) in the problem. Six different statistical metrics[4] were computed: *Mutual information*, *Euclidean distance*, two versions of the *Kullback-Leibler divergence*, *Matusita* and *Battacharyya* metrics. Sorting the variables by means of their coefficients, we can construct six different importance rankings.
Consensus univariate relevance. Individually, the univariate relevance metrics may be biased owing to the low number of instances. In these scenarios and to achieve a more dependable result it is better to put all the metrics together into a consensus. The consensus among the six original rankings is made up using the average position of each variable over all the rankings. The final consensus ranking shows the statistical univariate relevance of each probe in the supervised problem.

RESULTS

The quality analysis of the total RNA isolated was made for all the 133 tissue samples collected, and based on the electropherograms and the RIN number values obtained, all those samples with a RIN below 6 were discarded. Consequently, a total of 32 tumoral and 33 non-tumoral samples were selected for the microarray gene expression analysis.
After scanning the 65 microarrays the "control spots" were removed and the data obtained were pre-treated by Lowess Normalization [5].

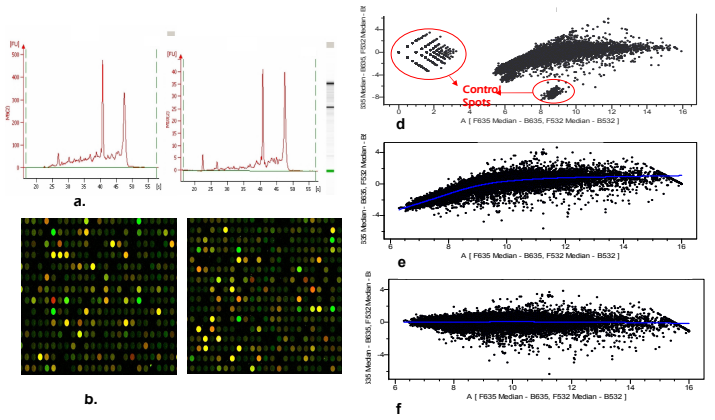

Fig 4. A) 2100 Bioanalyzer Electropherograms of total RNA isolated from a tumoral (left) and paired non tumoral samples (right). C) Microarray scanned image form a tumoral sample hybridized against the control pool. D) Data obtained form one microarray showing the "control spots". E) Unnormalized data after removing the "control spots". F) Data normalized by Lowess.

Once the control spots were removed from the data, the total number of probes descended from 22,574 to 17,986 probes. On the quality metrics filter process the acceptance threshold was set up in an average of 0.99 quality value; a total of 11,120 probes surpassed this stage. The imputation algorithm was run with a K value of 15 neighbors. From the 722,800 number of total spots, there were only 1,04% of lost values (7,534 probes) to imputate. The last filtering step removes 3,016 probes that showed differences bigger than *2-fold* in between each of the four classes of tissues. A total of 8,104 probes composed the final dataset.

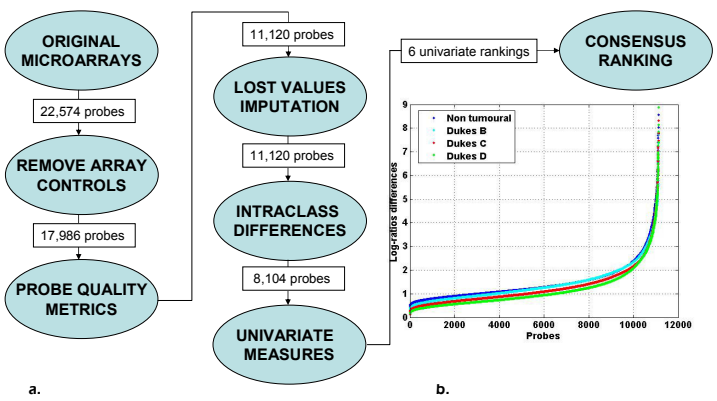

Fig 5. A) Overall process of the data analysis, for each stage the number of probes that surpass the stages are included in the boxes. B) Intraclass disperssion measure for the 11,120 filtered quality probes.

CONCLUSIONS

We have already obtained a tentative model (first ten genes are showed in table 2) for the classification of non-cancerous and cancerous samples and tumor staging, based on their gene expression profile. We are now in the validation process of this model, and it is of the outmost importance for us to check its potentiality for diagnosis/prognosis.

From the machine learning point of view we envision the building of different classification models. Furthermore, the search for statistical reliable dependencies could bring us some light regarding the complex nature of human CRC.

Rank	Gene	Description
1	ENC1	Ectodermal-neural cortex 1, overexpressed in some diseases and role in malignant transformation. Regulated by β-catenin/TCF4 pathway. High overexpression in many of primary colon cancer tissues.
2	ACAT1	Acetyl-Coenzyme A acetyltransferase 1 (mitochondrial acetoacetyl-coenzyme A thiolase); mutations in the corresponding gene are associated with 3-ketothiolase deficiency. Expressed in many types of tissues, including intestines.
3	CKLFSF7	Homo sapiens chemokine-like factor superfamily 7. Encoded protein is highly expressed in leucocytes, but unknown function.
4	HSPA5BP1	Homo sapiens heat shock 70kDa protein 5 (glucose-regulated protein 78kDa) binding protein 1.Plays a role in apoptosis and proliferation.
5	RPL23	Homo sapiens ribosomal protein L23. Inmunohistochemistry detection of ribosomal proteins in colorectal mucosa is been unknown. With RPL11 mediated activation of p53
6	FAM60A	Protein of unknown function, has high similarity to uncharacterized mouse Tera.
7	MCTS1	Multiple copies in T-cell malignancy, a putative oncogene that is involved in cell cycle regulation and participates in positive control of cellular proliferation through the regulation of CDK activity, amplified and overexpressed in T-cell lymphomas. Expressed in membrane and citoplasm of colon adenocarcinomas
8	SNRPB2	U2 small nuclear ribonucleoprotein (snRN) B, an snRNP- and snRNA-binding protein that may play a role in mRNA splicing, functions as an autoimmune antigen in patients systemic lupus erythematosus (SLE) and other rheumatic diseases.
9	MADCAM1	Homo sapiens mucosal vascular addressin cell adhesion molecule 1. Expressed in endothelial cells of the intenstine mucose, submucose and Peyer patches. In cronic intestinal inflammation it is overexpressed.
10	DDX55	Homo sapiens DEAD (Asp-Glu-Ala-Asp) box polypeptide 55. Member of the DEAD or DEAH box ATP-dependent RNA helicase family, contains a helicase conserved C-terminal domain, has moderate similarity to ATP-dependent RNA helicase, which is required for processing of 25S ribosomal RNA precursor.

Table 2. First ten genes in the consensus relevance ranking.

Predictive accuracy

From the work gene list, we select about a hundred genes highly correlated with the phenotype. By means of this gene subset, a supervised classifier can be induced to classify new unseen samples. To estimate the predictive accuracy of this supervised model a LOOCV [6] is performed using the naïve Bayes classifier as the model to induce.

The estimation achieves a 96.875% of accuracy, with only two misclassified samples in the confusion matrix. Furthermore, these two samples are swapped between the B and C stage, showing that the clinical criteria sometimes differ from the genetic profile [7]. Non cancerous samples are flawlessly distinguished from the cancerous ones.

VALIDATION

New CRC samples, not previously used in the microarray gene expression analysis, were collected for verification of gene expression levels through real time PCR. The previous RNA quality standard was applied and 15 samples were selected for the validation process(table 3).Total RNA was isolated using *RNeasy Plus Mini kit* (Qiagen) which includes a step for removing genomic DNA that could interfere in the RT-PCR reactions. 700 ng of total RNA were retrotranscripted with *Taqman RT reagents* (Applied Biosytems) at 25º for 10 minutes, followed by 1 hour at 48º and 5 minutes at 95º.

Real time PCR of seven high ranking position genes were made using *Assays on Demand* (Applied Biosystems) and the eukaryotic 18S Endogenous control (Applied Biosystems) was established as a housekeeping gene. PCR master mix was prepared with *Platinum qPCR supermix* (Invitrogen) and the reaction was carried out according to the following conditions: 1 cycle of 2 minutes at 95º and 40 cycles consisting in 15 second at 95º and 1 minute at 60º.

SAMPLES	T/NT	RIN	DUKES' STAGE	TNM STAGE
M30	T	8	B2	IIA
M46	T	6	D	IV
M52	T	9.4	C2	IIIC
M84	T	7.3	C2	IIIC
M117	T	8	C2	IIIC
V1	T	8.5	B2	IIA
V2	T	7.6	D	IV
V4	T	7	B3	IIB
V5	T	8	C2	IIIC
V6	T	6.6	D	IV
V7	T	8	B2	IIA
M104	NT	6.8	-	-
M116	NT	7.7	-	-
V10	NT	7.5	-	-
V12	NT	7.8	-	-

Table 3. Samples used for validation assays. T: tumour, NT: non-tumoral

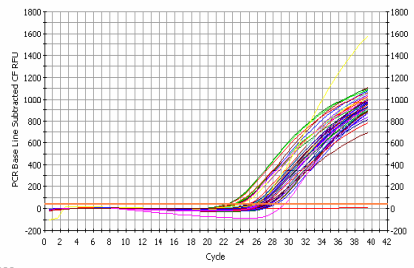

Fig. 6.. Amplification curve for ENC1 gene.

Microarray vs. qPCR profiling

The final validation step of the presented methodology involves the comparison of the genes' activities detected by the microarray platform against the activities detected by a quantitative PCR. For this purpose, seven genes with high ranking positions were monitored. For the qPCR, we compute the median of the differences between the CT values returned by the gene and the 18S control. The microarray columns (see Table 4) show the difference between the logRatios of the non cancerous and the cancerous ones. Table 4 gathers these calculations, clearly showing that both profiles agree in 20 out of the 21 possibilities.

	qPCR			Microarray		
	Dukes B	*Dukes C*	*Dukes D*	*Dukes B*	*Dukes C*	*Dukes D*
ENC1	3,104	2,515	1,946	1,66	1,405	1,512
ACAT1	-1,811	-0,900	-1,791	-0,884	-0,825	-1,287
HSPA5BP1	2,919	2,466	2,817	0,865	0,867	1,398
CKLFSF7	1,193	2,593	1,133	0,752	0,8305	1,277
FAM60A	1,150	1,044	-0,187	1,346	0,829	1,3475
MADCAM1	-1,759	-0,029	-6,826	-0,534	-0,831	-0,5605
DDX55	1,164	1,143	0,335	0,876	0,7795	0,6755

Table 4. Comparisons among the gene activities detected by the microarray and the qPCR validation technique.

Bibliografy

[1] van't Veer et al. Nature, 2002, 415, 530-535
[2] Chen Y. et al. *Bioinformatics*, 2002, 18(9), 1207-1215
[3] Troyanskaya O. et al. *Bioinformatics*, 2001, 17(6), 520-525
[4] Molloy M.P. et al. *Electrophoresis*,1998,19(5),837-44
[5] Ben-Bassat M. Handbook of Statistics, 1982, 2:773-791
[6] Stone M. *Journal of the Royal Statistical Society B*, 36, 1974, 147-211
[7] Eschrich, S. et al. Journal of Clinical Oncology, 2005, 23(15), 3526-3535