Identification of a biomarker panel for colorectal cancer diagnosis

Garcia-Bilbao Amaia1, Armananzas Rubén2, Ispiuza Zlotza1, Calvo Begoña1, Alonso-Varona Ana1, Inza Iñaki1, Larrañaga Pedro2, López Vivanco Guillerumo2, Suárez-Merino Blanca1, Betananzas Mónica1

1 GIACER-BRAK Technology Centre, Parque Tecnológico, Edificio 202, 41170 Zamudio (Bikiala), Spain
2 Medical Oncology Unit, Hospital de Cruces, Plaza de Cruces s/n, 48903 Barakaldo (Bikiala), Spain
3 Department of Cell Biology and Histology, Facultad de Medicina y Odontología, Universidad del País Vasco UPV/EHU, 48940 Leioa (Bikiala), Spain
4 Department of Computer Science and Artificial Intelligence, Facultad de Ciencias de la Comunicación, Universidad del País Vasco UPV/EHU, 20018, Donostia San Sebastián (Gipuzkoa), Spain

Introduction

Colorectal cancer (CRC) is the third most common form of cancer and the second leading cause of death among cancers worldwide. There is an urgent necessity of new diagnostic approaches in order to improve the outcome of CRC screening programs. In particular, there is a clinical need for identifying specific biomarkers for early detection of CRC. Moreover, there is at the present time a widespread awareness, that the development of biomarkers will offer the major advances in CRC detection.

The recent advances in genomics and proteomics have contributed to our understanding of CRC. Genomic techniques allow high-throughput analysis of genes, rendering big volumes of data, increasing the possibilities for uncovering potential biomarkers. Our aim in this study was to develop a model or biomarker for the objective diagnosis of CRC based on gene expression patterns. For this purpose we used the microarray technology in combination with advanced statistics analysis techniques.

Materials and Methods

TISSUE SAMPLES

Collection. A total of 64 tissue samples-33 non-tumoural tissues (NT) and 31 tumoral tissues (T) were obtained from patients with CRC diagnosed at different stages in Cruces Hospital (BIODEF) after informed consent. After surgery, an anatopathologic analysis was carried out to confirm diagnosis as well as tumour staging.

Experimental design. To search for genetic markers, the experimental design comprised the hybridisation of each sample against a reference pool consisting of the non-tumoral samples.

Genomics. Total RNA was extracted from 64 samples using the RNEasy Mini kit (Qiagen). Quantity and integrity were determined by the 2100 Bioanalyzer (Agilent Technologies). RNA algorithm was used as a quality standard for the selection/inclusion of samples in the study. The selected samples were labelled and hybridised onto the Agilent Human 1A 60-mer oligo microarray and scanned using the GenePix 4000B Axon Scanner.

Biostatistics. All the microarrays were pre-processed to ensure probe quality: lost values imputation, probes quality, data smoothing and intraclass variability filters were applied to microarray data. The final dataset (8,104 probes) was statistically analysed (supervised classification and data analyses) and the most relevant genes were obtained. Finally, a supervised classifier was induced to classify new unseen samples.

Quantitative real-time PCR. Expression levels of the differentially expressed genes were measured using Taqman reverse transcription reagents (Applied Biosystem) and Platinum Quantitative PCR Supremix-UDG with ROX (Invitrogen).

Machine learning validation. A total of 36 new tissue samples were used. These samples were hybridised against the same reference pool (E-MTAB-770). Three different classification paradigms were used to validate the panel: naïve Bayes, support vector machine and k-nearest neighbours. We used the leaving-one-out cross validation scheme.

BLOOD SAMPLES

Collection. A total of 57 blood samples-47 tumoural and 10 healthy samples were collected in Cruces Hospital after informed consent.

RNA isolation. Total RNA was extracted with Qiap DNA Blood Mini kit (Qiagen) and quantified using Nanodrop ND-1000.

Quantitative real-time PCR. Expression levels of the differentially genes obtained from tissue were measured following the same protocol used in tissue validation.

Data analysis. Parametric and non-parametric tests as well as different classification models were applied to the data.

Results and Discussion

From the final dataset of 8,104 probes, using Feature Subset Selection (FSS) we selected a subgroup of relevant genes that were able to distinguish between tumour and non-tumour samples. In order to ensure stable approaches we performed a random resampling of the dataset, a multivariate feature subset selection and the induction of a k-dependence Bayesian network classifier.

The graphical dependency structure gathered a total of 14 genes. From all of them, 10 genes showed dependences with other genes, being ENC1 the core gene. As an exploratory approach, we subjected these 10 genes to Ingenuity Pathway Analysis (IPA) software to assess the involvement of the relevant genes in known molecular pathways and networks. Therefore, 6 of the genes (ENC1, ACAT1, Mستان, MADCAM1, RPL23 and SNPBD2) were found to be associated with cancer or genetic disorders. The rest of the genes were involved in different networks.

To examine the reliability of our microarray results, we did a qPCR experiment with the 10 genes identified. The results confirmed that the expected gene expression profiling was validated for 7 of the genes analysed. Although no direct relationships were found between our set of validated genes, all of them (except CMTM7) were involved in a network were TNF and RELA played a central role. Both genes are related to inflammatory processes and take part in CRC metastasis canonical pathway.

Nevertheless, the whole gene panel (7 genes) is required as a group to identify new unseen samples as tumoral or non-tumoral with a 96.92% of accuracy. In order to validate the classification power of the biomarker panel, we tackled the classification of a new cohort of samples (14 non-tumoral and 22 tumoral), but using only the expression of the genes previously identified within the panel. A leave-one-out cross validation (LOOCV) scheme was used. All the non-tumoral samples were always correctly classified as non-tumours, whereas just two of the tumoral samples were classified and non-tumoral (estimated accuracy 94.45%).

So as to ensure that this tissue-discovered (and validated) biomarker panel was differentially expressed in the blood of CRC patients, qPCR assays were performed for each of the 7 genes. Only TME1L32 was removed from the analysis, as it was not expressed in blood. The 6 genes left were included in a multivariate diagnostic model, obtaining an estimated accuracy of 94.64% with the following associated confusion matrix:

<table>
<thead>
<tr>
<th></th>
<th>Tumoral</th>
<th>Non-Tumoral</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tumoral</td>
<td>46</td>
<td>1</td>
</tr>
<tr>
<td>Non-tumoral</td>
<td>8</td>
<td>8</td>
</tr>
</tbody>
</table>

Table 1. Confusion matrix associated to our data. Rows represent the real observed phenotype value and columns represent the prediction of the model

The sensitivity value of the model is 0.978, while specificity is 0.8. The strongest correlation relationship was found between ENC1 and ACAT1, as well as between ACAT1 and CMTM7, and both of them strongly condition the values of FAM60A.

Conclusions

We have achieved a tentative biomarker panel of 7 genes capable of correctly classifying cancerous and non-cancerous colon tissue samples. As a result of this study, this panel has been tested on the use of this panel as a diagnostic tool in colorectal cancer (EP2169078A1).

Nevertheless, translating tissue findings into clinically useful endpoints is crucial. To achieve this, we have validated the biomarker panel in 47 blood samples, with 94.64% of accuracy. In a second step, more blood samples will be re-collected and the panel will be tested by means of qPCR technique.

Acknowledgments

The authors are grateful to all the patients for their participation. This work was supported by the Eustres and Sanzor 2000-2007 programs (Bilbao County Government). AGB was supported by Fundación Itako Goenaga Noldeko for graduate studies. RA was supported by the Spanish Ministry of Science and Innovation through a Juan de la Cierva Fellowship, TIN2010-20000-C03-04 and the Cajal Blue Brain Project. IC was supported by the Spanish Ministry of Health (BILGC11/0026). It was supported by the Spanish Ministry of Science and Innovation (TIN2010-14932) and COMBINE/BD network of the Carlos III Institute.