LGALS4, *CEACAM6*, *TSPAN8*, *COL1A2*: blood markers for colorectal cancer. Validation in a cohort of subjects positive to faecal immunochemical test

Maria Teresa Rodia[1]*, Rossella Solmi[1]*, Francesco Pasini[2], Elena Nardi[3], Gabriella Mattei[1], Giampaolo Ugolini[2], Luigi Ricciardiello[2], Pierluigi Strippoli[1,4,5], Rossella Miglio[3], Mattia Lauriola[1,#]

1 Department of Experimental, Diagnostic and Specialty Medicine (DIMES), Unit of Histology, Embryology and Applied Biology, University of Bologna, Italy;

2 Department of Medical and Surgical Sciences (DIMEC), University of Bologna, Italy

3 Department of Statistical Sciences "Paolo Fortunati", University of Bologna, Italy

4 Centre of Molecular Genetics, "CARISBO Foundation", University of Bologna, Italy

5 “Giorgio Prodi” Cancer Research Center (CIRC), University of Bologna, Italy

*Equally contributed

[#] Corresponding author

mattia.lauriola2@unibo.it

+39 3402693431

0000-0002-7966-0065

Keywords

CRC, colorectal cancer; liquid biopsy, blood markers, LGALS4, CEACAM6, TSPAN8, COL1A2, colonoscopy, faecal occult blood test, faecal immunochemical test

Running title

CELTiC- blood markers for colorectal cancer.

Abbreviations: CRC = colorectal cancer; CEACAM6 = carcinoembryonic antigen-related cell-adhesion molecule 6; LGALS4 = lectin, galactoside binding soluble 4; TSPAN8 = tetraspanin 8; COL1A2 = collagen type I alpha 2 chain; qRT-PCR = quantitative reverse transcriptase-polymerase chain reaction; FOBT = faecal occult blood test; FIT = faecal immunochemical test;

Abstract

A non-invasive blood test for early detection of colorectal cancer (CRC) is highly required. In this study, we evaluated a panel of four mRNAs as putative markers of colorectal cancer.

We tested *LGALS4*, *CEACAM6*, *TSPAN8*, *COL1A2*, referred to as the CELTiC panel, by qRT-PCR, on subjects positive to the faecal immunochemical test (FIT) and undergoing colonoscopy. By using a non-parametric test and multinomial logistic model, FIT positive subjects were compared to colorectal cancer patients and healthy individuals.

All genes of the CELTiC panel displayed statistically significant differences in the comparison between healthy subjects (N) (n=67) and both low risk (LR) (n=36) and high risk/colorectal cancer subjects (HR/CRC) (n=92), as well as those in the negative colonoscopy, FIT positive group (NFIT) (n=36).

The multinomial logistic model indicated *LGALS4* as the most powerful marker discriminating the four groups. When assessing the diagnostic values by analyses of the areas (AUC) under the receiver-operating curves (ROC), the CELTiC panel reached values of 0.91 (sensitivity 79% and specificity 97%) when comparing N to LR, and 0.88 (sensitivity 75% and specificity 87%) for the comparison between N and HR/CRC; the comparison between N and NFIT resulted in an AUC of 0.93 (sensitivity of 82% and specificity 94%).

The CELTiC panel might represent a useful tool for discriminating subjects positive to FIT, as well as for early detection of precancerous adenomatous lesions and CRC.

Introduction

Over the past decade, huge efforts have been made by the scientific community to develop new and powerful tools for cancer screening and detection. This applies to colorectal cancer (CRC) in particular, since it arises as a consequence of genetic and molecular alterations of the nascent tumour, which develop over time [1]. The progression from premalignant lesion to carcinoma and metastasis is relatively slow, in some settings taking up to 15 years to progress, thus providing the timeframe for early detection [2]. Several molecular events are involved in CRC development, including oncogenic mutations and microRNAs deregulation [3]. The possibility of assessing solid tumours, by means of a simple blood test has attracted considerable interest. Indeed, the term 'liquid biopsy' indicates the approach to molecular probing blood samples for tumour gene profiles [4]. Importantly, it is now possible to test several components using human blood samples, including cell-free DNA (cfDNA) [5,6] and RNA (cfRNA)[7], as well as proteins and circulating vesicles, known as exosomes [8,9]. In addition, circular RNAs represent an emerging class of non-coding RNAs, which have gained increasing interest as markers for liquid biopsies due to their stability over time [10,11].

The purpose of screening is to identify asymptomatic populations as potential carriers of premalignant or early onset cancers, and thus suitable for complete removal, by means of a simple, non-invasive and acceptable test [12]. In recent years many countries, including Italy, have adopted the faecal immunochemical test (FIT) as the screening tool for CRC. In most Italian regions, positive patients are considered as those with an FIT >100 ng/ml.

Under these premises, about 5.5% of the screened population (aged 50-69 yrs) will be FIT positive and only 2.9‰ will receive a diagnosis of CRC and 20.1‰ a diagnosis of advanced adenomas (at first round) [13]. Thus, the rate of false positive FITs is the incentive for the investigation of new non-invasive and more specific screening tests, including blood markers.

Blood tests are widely accepted and non-invasive, making it possible to simultaneously detect multiple parameters. Furthermore, liquid biopsies are based on new technologies, thus providing high sensitivity, required for early detection and for prediction or prognosis of clinical progression. In our laboratory, over the past ten years we have employed several approaches aimed at detecting colorectal cancer patients by assaying whole blood [14,15]. Very recently, we proposed a panel of four mRNAs, namely carcinoembryonic antigen-related cell-adhesion molecule 6 (*CEACAM6)*, lectin, galactoside binding soluble 4 (*LGALS4)*, tetraspanin 8 (*TSPAN8)*, collagen type I alpha 2 chain (*COL1A2)*, hereafter referred to with the acronym of CELTiC (***CE****ACAM6*, ***L****GALS4*, ***T****SPAN8* and ***C****OL1A2*), that proved highly promising as a screening tool to detect CRC patients [16].

The CELTiC panel was identified by bioinformatic analyses [17], and subsequent data obtained by qRT-PCR assay showed a statistically significant difference between the group of CRC patients versus healthy subjects. With the current study, we validated our data on a larger sample of individuals and further assayed the possible predictive role of the CELTiC panel in the identification of patients with advanced precancerous lesions and colorectal cancer at early stages, within FIT positive subjects undergoing colonoscopy.

Our results, while confirming the previous findings, suggest that CELTiC panel might represent a promising tool for the detection of patients with high-risk lesions, as well as low risk patients from healthy subjects, a gold standard for early detection.

Materials and Methods

Patients

One hundred and one FIT-positive consecutive subjects were enrolled for this study at the Gastroenterology Unit of the Department of Medical and Surgical Sciences, University of Bologna, from May 2015 to March 2016 (Table 1). After informed consensus, 5 ml of peripheral blood was obtained from FIT positive subjects. In order to reduce contamination of samples with epithelial cells from the needle stick, the first 1 mL of blood was discarded. The family history was determined by a questionnaire that each subject filled in at the time of the enrolment. The study was conducted following approval by the ethics committee of the Sant'Orsola-Malpighi Hospital, Bologna, and complied with the Ethical Principles for Medical Research Involving Human Subjects of the Helsinki Declaration. All subjects involved were asked for informed written consent before taking part in the study.

The enrolled subjects were asymptomatic persons between the ages of 50 and 72 years, who were positive to the FIT and scheduled for colonoscopy. Participants with recurrence of colorectal neoplasia, digestive cancer or inflammatory bowel disease, were excluded from this study. Moreover, we included samples collected from our previous study [16], including 67 healthy donors with no clinical history of neoplastic disease and 63 unrelated patients with a histologically confirmed diagnosis of CRC at any stage, before elective surgery and without any chemo or radio adjuvant treatments to the surgery.

RNA Extraction

RNA was extracted from whole blood. The blood was collected in EDTA tubes and treated for lysis within one hour of collection. Briefly, one ml of whole blood was diluted with PBS (1:2 ratio) lysed with TRIzol LS reagent (Invitrogen, Carlsbad, CA, USA) and

total RNA was extracted according to the manufacturer's protocol. Total extracted RNA from 1 mL of blood was subjected to standard ethanol precipitation, and the pellet was dissolved in 15 µL RNase-free water to a final concentration of up to 0.5 µg/µL, and stored at -20°C. The concentration of all RNA samples was quantified by Nanodrop ND-2000 spectrophotomer (Thermo Fisher Scientific, Waltham, MA).

qRT-PCR

300 ng of RNA was reverse transcribed with the RevertAid First Strand cDNA Synthesis kit (Carlo Erba Reagents, Milan, Italy) and amplified using the EvaGreen system (Bio-Rad, Hercules, CA), according to the manufacturer's instructions.

Real-time PCR reactions were performed using the CFX96 instrument (Bio-Rad, Hercules, CA), in duplicate, at 95°C for 10 min, followed by 40 cycles of 95°C for 15 sec and 60°C for 1 min, with melting curve analysis. Each qPCR run always included a negative control lacking the cDNA template, and a positive control of cDNA derived from the HT-29 cell line, in which all the tested genes are expressed. Primer sequences and the calibration test are described in [16].

Statistical analysis

Mean, median, standard deviation (sd), ranges and frequencies were used as descriptive statistics. Kruskall-Wallis rank sum test was applied to compare each marker's expression among groups; two-tailed adjusted p-values less than 0.05 were considered to be significant. Correlation coefficients among markers were reported with their p-values. A multinomial logistic regression model was performed in order to study the association between outcome and a linear combination of the proposed markers; two-tailed p-values less than 0.05 were considered statistically significant; the reference group is N (healthy

subjects).

ROC (Receiving Operating Characteristic) curve analysis was applied to assess the accuracy of the model in discriminating among the 4 groups of subjects. Area Under the Curve (AUC) is reported together with the relative optimal values of sensitivity and specificity. Statistical analyses were performed by using STATA 14.0 and RStudio Version 1.0.143.

Results

Study population

In order to evaluate the efficiency of the CELTiC panel, a total of 231 subjects were analysed; 63 CRC patients undergoing surgery and 67 volunteers, healthy subjects (N), were collected from a previous study [16]. An additional 103 consecutive subjects were recruited after an FIT positive test and further evaluated by colonoscopy (Figure 1). From the 103 FIT positive subjects, two were excluded as they were not Caucasian and the remaining 101 were further stratified according to the diagnosis from the colonoscopy. Four patients with CRC plus 25 subjects with advanced adenomas (high-grade dysplasia or with ≥25% villous histologic features or measuring ≥10 mm in the greatest dimension) were grouped as high risk subjects (HR/CRC) (Table 1), along with the 63 CRC patients with a histologically confirmed diagnosis of colorectal cancer at any stage, as previously reported [16]. Next, 36 subjects with small polyps (< 10 mm) constituted the low risk (LR) group. Finally, the remaining 36 FIT positive subjects without any neoplastic or preneoplastic lesions but with diverticulitis, haemorrhoids, angiodysplasia and aphthoid lesions were grouped as the negative colonoscopy, FIT positive (NFIT) group (Figure 1).

Descriptive statistics

CELTiC panel mRNA expression values were evaluated for each group by qRT-PCR, sampling RNA extracted from the total blood. First, we analysed the ΔCT distribution of each marker upon normalization on the B2M housekeeping gene. In order to properly evaluate the following results, it is important to clarify that ΔCT values are inversely correlated to the amount of gene expression, thus high ΔCT values indicate low levels of the relative gene.

As reported in the box plot in Figure 2, ΔCT values of *CEACAM6* and *LGALS4* appeared higher in LR, HR/CRC patients and NFIT compared to the N group, with values on average of 12.3 and 12.9 in the N group and 13.6-13.3 or 15.3 or 14.7 in low risk and high risk patients (Figure 2).

Conversely, *TSPAN8* and *COL1A2* displayed lower ΔCT values in LR, HR/CRC and NFIT compared to N (Figure 2), with values on average of 9.9 to 9.6 and 9.7 to 9.6 respectively in LR and HR/CRC patients and 11.3 and 11.4 in N (Figure 2). The comparison between the groups of N and NFIT showed a different distribution of ΔCT values (Figure 2) and significant differences for each marker (p-value <0.001) (Table 2); for further evaluations and statistical analysis these groups were therefore considered separately. On the other hand, no differences were reported between the CRC and HR groups, thus we decided to combine them into the HR/CRC group.

Next, we performed an explorative analysis by applying the Kruskal-Wallis Rank Sum Test, comparing the medians of age and ΔCT among the four groups. With this test we aimed to evaluate the ability of each marker to discriminate between all possible comparisons, adjusting p-values for multiple comparisons.

Besides the statistical difference between the age of NFIT and HR/CRC subjects, no differences were found for the remaining groups, suggesting a good homogeneity on the

age distributions of our dataset. Each member of the CELTiC panel was able to statistically distinguish the N group from the HR/CRC, LR and NFIT groups (p-values <0.001), as reported in Table 2. On the other hand, when we tested the ability of the CELTiC panel to discriminate between the LR and HR/CRC groups, only *LGALS4* retained the discrimination power, with a p-value of 0.034 (Table 2). The comparison of the NFIT and HR/CRC groups showed significant differences (p-values <0.001 and 0.003 respectively) for *LGALS4* and *CEACAM6*, while no differences were seen for the comparison of NFIT with LR subjects (Table 2).

This type of analysis highlights a key role for the CELTiC panel in distinguishing the LR and HR/CCR groups from the group of healthy subjects and furthermore suggests a discriminating ability of *LGALS4* in detecting LR from HR/CRC subjects.

Multinomial Logit Model and correlation between markers

First, we investigated the relative correlations among genes of the CELTiC panel. Interestingly, *TSPAN8* and *COL1A2* displayed a strong positive correlation (r= 0.87, p-value <0.001) (Table 3), thus suggesting a shared transcriptional regulation for these two genes which requires further functional evaluations. On the other hand, only a weak correlation was identified for *CEACAM6* and *LGALS4* (r= 0.39, p-value <0.001) (Table 3). *COL1A2* displayed a very weak negative but still significant correlation with *LGALS4* and *CEACAM6* (r= -0.14 and p-value= 0.03). None of the remaining associations reached a significant level of correlation.

Next, we performed a Multinomial Logistic Regression Model in order to evaluate the global association between the CELTiC panel and the studied groups, taking into account correlations between markers. The healthy subjects N represented the reference group for all the explored comparisons. In the comparison between N and LR subjects, *LGALS4* was

the only marker with a statistically significant value, and proved to be a protective factor (OR: 4, p-value <0.001), thus LR patients displayed lower amounts of this gene in the blood (Table 4). Next, in the comparison between N and HR/CRC subjects, *TSPAN8* and *LGALS4* displayed a predictive ability. In detail, *TSPAN8* proved to be a risk factor (OR: 0.46, p-value: 0.006), while *LGALS4* confirmed its role as a protective factor (OR: 2.72 p-value <0.001) (Table 4). Age proved to be a minor albeit significant risk factor (OR: 1.04 p-value: 0.043). *LGALS4* was higher also in NFIT samples compared to the N group (OR: 5.15, p-value: <0.001) and in this comparison *COL1A2* seems to be a risk factor (OR: 0.42 and p-value: 0.041). Finally, when studying the N and NFIT comparison, *CEACAM6* proved to be a protective factor (OR: 2.08, p-value: 0.002).

To sum up, high levels of *TSPAN8* appeared to be a risk factor for subjects with high risk lesions and CRC patients, while *LGALS4* values seemed to be protective for both LR and HR/CRC groups.

Diagnostic values of CELTiC

In order to evaluate the diagnostic accuracy in terms of specificity and sensitivity of the CELTiC panel, the Receiver Operating Characteristic (ROC) curve analysis was applied. First, we performed a comparison between the healthy subjects and the low risk group (n=103), with the ROC curve reaching an impressive value of AUC of 0.91, with a sensitivity of 79% and specificity of 97% (Figure 3). When comparing N with the HR/CRC group (n=159), the AUC value was still significantly high, 0.88, with a sensitivity of 75% and specificity of 87%. Next, we performed a unified ROC evaluation, by combining together all the patient datasets, both the LR and the HR/CRC groups and comparing these to the healthy subject group N (n=195). The resulting ROC curve showed

an AUC value of 0.88, with a sensitivity of 73% and specificity of 89%. When comparing N with the NFIT group, the ROC value was 0.93, with a sensitivity of 82% and specificity of 94%.

All the remaining combinations for ROC analyses and results are reported as supplementary Figure 1.

To sum up, ROC curves confirmed the ability of the CELTiC panel to effectively select the healthy subject group from the high risk/colorectal cancer patients, as previously reported in our investigations. Interestingly, however, the significant AUC values obtained in the comparison between the healthy subject with the low risk group as well as the NFIT group suggested a discrimination power of the CELTiC panel in the detection of early lesions and FIT positive subjects.

Discussion

Transcriptional markers, based on gene expression analysis, give prompt information about the response of cells to internal and external stimulations and the microenvironment involvement [18]. We propose a simple blood assay that represents a minimally invasive technique for the detection and analysis of CELTiC panel mRNA biomarkers in blood. This approach could be periodically repeated for early detection as well as to improve the follow-up of disease progression in the perspective of personalized colorectal cancer surveillance. Indeed, the identification of promising biomarkers remains a major challenge.

This study represents an extension of our previous paper, which concluded with the identification of a panel of four genes in the detection of CRC patients [16]. Now, on comparing healthy subjects with colorectal cancer and high risk groups, we confirmed our previous findings [16] and the CELTiC panel reached values of sensitivity and specificity of

75% - 87% respectively, with a larger number of samples.

Next, *LGALS4* was able to discriminate between LR and HR/CRC groups, thus leading us to speculate for a biological role of this gene in the development and progression of colorectal cancer, from benign to aggressive form. Indeed, LGALS4 is a microvillar lipid raft stabilizer/organizer protein, involved in cell adhesion (Table 5) [19], but it can also be secreted and mediate cell responses [20]. Consistent with our findings, LGALS4 was reported strongly down-regulated in tumours [21]. Makoto and colleagues reported that the plasma or serum levels of LGALS4 were significantly reduced in subjects affected by colorectal cancer after surgical treatment [22]. As a whole, literature data about LGALS4 in colorectal cancer are incomplete and sometimes controversial, even if all studies recognize LGALS4 as a marker in colorectal cancer.

In the same vein, CEACAM6 was identified as a differentiation marker in normal colonocytes, responsible for the maintenance of tissue architecture [23,24]. Similarly to LGALS4, it is localized in specific membrane rafts, where it is involved in cell-cell and cell-matrix interactions (Table 5) [25]. CEACAM6 up regulation was reported in hyperplastic polyps and early adenomas, thus it is considered as one of the events leading to colorectal tumours [26]. In addition, high levels of CEACAM6 were detected in the sera of a large number of lung, liver, pancreatic, breast, and colorectal cancer patients [27]. TSPAN8, also named transmembrane 4 superfamily member 3 (TM4SF3), is an integral membrane protein of the tetraspanin super family, which has been reported to be involved in cell adhesion, motility [28,29] and angiogenesis [30], while high *TSPAN8* mRNAs were reported up regulated in gastric cancer tissues [31].

The data obtained by us from the expression of *TSPAN8* in blood, for N, LR and HR/CRC subjects are in line with the literature data about the protein expression on tissues and in sera and they correlate with a definition of TSPAN8 as a marker of high risk lesions. Of

interest, TSPAN8 and galectins have been described highly abundant in exosomes [32,33], thus suggesting an exosomal origin of the mRNA we collected from the whole blood, although this speculation will require a separate and further evaluation. Finally, COL1A2 is a secreted extracellular matrix protein. In our data, the level of *COL1A2* strongly correlates with *TSPAN*8, in line with the previous report from Greco et al. [28].

In conclusion, by testing a larger sample, these data confirmed our previous findings on the ability of the CELTiC panel to distinguish healthy subjects from patients with colorectal cancer or high risk lesions. Moreover, *LGALS4* emerged as a marker to distinguish subjects with low risk lesions from HR/CRC subjects, thus suggesting an involvement of this gene in colorectal cancer progression.

We are aware of the limitations of this study for the identification of a clear cut-off for the expression levels of the CELTiC panel, which will require the examination of a much larger dataset, involving additional healthy subjects as well as FIT negative subjects and an increase in the case record of FIT positive and CRC subjects.

To sum up, these preliminary data suggest that the CELTiC panel might represent a useful tool for early detection of precancerous adenomatous lesions as well as for advanced CRC patients, although additional studies need to be performed.

Compliance with Ethical Standards:

Conflict of Interest: RS has received support grants from the “Fondazione Enzo Piccinini” (Modena, Italy). ML has received support grants from Fondazione del Monte. LR has received support grants from “Italian Association for Cancer Research” [Grant number: Investigator Grant IG14281] and European Community’s Seventh Framework Program [Pathway-27, under grant agreement n. 311876 to L.R.]. MTR, GM, GU, RS and LM

declare that they have intellectual property rights on a patent pending (WO2016/185451 A1 - Method and kit for the diagnosis of colorectal cancer). All the other authors have nothing to declare.

Ethical approval: All procedures performed in this study involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Informed consent: Informed consent was obtained from all individual participants included in the study.

Acknowledgments

The authors are grateful to Susan West for her assistance in the English language review.

References

1. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz Jr. LA, Kinzler KW. Cancer Genome Landscapes. *Science (80-)*. 2013;339(6127):1546-1558. doi:10.1126/science.1235122.
2. Nash I. Intratumor heterogeneity and branched evolution. *N Engl J Med*. 2012;366(22):2132-3; discussion 2133. doi:10.1056/NEJMc1204069#SA3.
3. Li Y, Lauriola M, Kim D, et al. Adenomatous polyposis coli (APC) regulates miR17-92 cluster through β-catenin pathway in colorectal cancer. *Oncogene*. 2016;35(35):4558–4568. doi:10.1038/onc.2015.522.
4. Montagut C, Siravegna G, Bardelli A. Liquid biopsies to evaluate early therapeutic response in colorectal cancer. *Ann Oncol*. 2015;26(8):1525-1527. doi:10.1093/annonc/mdv228.
5. Diehl F, Schmidt K, Choti M a, et al. Circulating mutant DNA to assess tumor dynamics. *Nat Med*. 2008;14(9):985-990. doi:10.1038/nm.1789.
6. Aarthy R, Mani S, Velusami S, Sundarsingh S, Rajkumar T. Role of Circulating Cell-Free DNA in Cancers. *Mol Diagn Ther*. 2015;19(6):339-350. doi:10.1007/s40291-015-0167-y.
7. Tzimagiorgis G, Michailidou EZ, Kritis A, Markopoulos AK, Kouidou S. Recovering circulating extracellular or cell-free RNA from bodily fluids. *Cancer Epidemiol*. 2011;35(6):580-589. doi:10.1016/j.canep.2011.02.016.
8. Ogata-Kawata H, Izumiya M, Kurioka D, et al. Circulating exosomal microRNAs

as biomarkers of colon cancer. *PLoS One*. 2014;9(4). doi:10.1371/journal.pone.0092921.
9. Schageman J, Zeringer E, Li M, et al. The complete exosome workflow solution: From isolation to characterization of RNA cargo. *Biomed Res Int*. 2013;2013. doi:10.1155/2013/253957.
10. Enuka Y, Lauriola M, Feldman ME, Sas-Chen A, Ulitsky I, Yarden Y. Circular RNAs are long-lived and display only minimal early alterations in response to a growth factor. *Nucleic Acids Res*. 2015:gkv1367-. doi:10.1093/nar/gkv1367.
11. Li Y, Zheng Q, Bao C, et al. Circular RNA is enriched and stable in exosomes: a promising biomarker for cancer diagnosis. *Cell Res*. 2015;25(8):981-984. doi:10.1038/cr.2015.82.
12. Imperiale TF, Ransohoff DF, Itzkowitz SH, et al. Multitarget stool DNA testing for colorectal-cancer screening. *N Engl J Med*. 2014;370(14):1287-1297. doi:10.1056/NEJMoa1311194.
13. Giorgi Rossi P, Vicentini M, Sacchettini C, et al. Impact of Screening Program on Incidence of Colorectal Cancer: A Cohort Study in Italy. *Am J Gastroenterol*. 2015;110(9):1359-1366. doi:10.1038/ajg.2015.240.
14. Solmi R, Ugolini G, Rosati G, et al. Microarray-based identification and RT-PCR test screening for epithelial-specific mRNAs in peripheral blood of patients with colon cancer. *BMC Cancer*. 2006;6:250. doi:10.1186/1471-2407-6-250.
15. Lauriola M, Ugolini G, Rosati G, et al. Identification by a Digital Gene Expression Displayer (DGED) and test by RT-PCR analysis of new mRNA candidate markers for colorectal cancer in peripheral blood. *Int J Oncol*. 2010;37(2):519-525.
16. Rodia MT, Ugolini G, Mattei G, et al. Systematic large-scale meta-analysis identifies a panel of two mRNAs as blood biomarkers for colorectal cancer detection. *Oncotarget*. March 2016. doi:10.18632/oncotarget.8108.
17. Lenzi L, Facchin F, Piva F, et al. TRAM (Transcriptome Mapper): database-driven creation and analysis of transcriptome maps from multiple sources. *BMC Genomics*. 2011;12(February):121. doi:10.1186/1471-2164-12-121.
18. Byron SA, Van Keuren-Jensen KR, Engelthaler DM, Carpten JD, Craig DW. Translating RNA sequencing into clinical diagnostics: opportunities and challenges. *Nat Rev Genet*. 2016;17(5):257-271. doi:10.1038/nrg.2016.10.
19. Danielsen EM, van Deurs B. Galectin-4 and small intestinal brush border enzymes form clusters. *Mol Biol Cell*. 1997;8(11):2241-2251.
20. Ideo H, Seko A, Yamashita K. Galectin-4 binds to sulfated glycosphingolipids and carcinoembryonic antigen in patches on the cell surface of human colon adenocarcinoma cells. *J Biol Chem*. 2005;280(6):4730-4737. doi:10.1074/jbc.M410362200.
21. Rechreche H, Mallo G V, Montalto G, Dagorn JC, Iovanna JL. Cloning and expression of the mRNA of human galectin-4, an S-type lectin down-regulated in colorectal cancer. *Eur J Biochem*. 1997;248(1):225-230. doi:10.1111/j.1432-1033.1997.00225.x.
22. Makoto W, Ichiro T, Naoki K, et al. Clinical significance of circulating galectins as colorectal cancer markers. *Oncol Rep*. 2011;25(5):1217-1226. doi:10.3892/or.2011.1198.
23. C. I, S. J, LC. A, P. D, CP. S. Cell-surface levels of human carcinoembryonic antigen are inversely correlated with colonocyte differentiation in colon carcinogenesis. *Lab Invest*. 1997;76(5 PG-703-16):703-716.
24. Ilantzis C, Demarte L, Screaton R a, Stanners CP. Deregulated Expression of the Human Tumor Marker CEA and CEA Family Member CEACAM6 Disrupts Tissue

Architecture and Blocks Colonocyte Differentiation1. *Neoplasia*. 2002;4(2):151-163. doi:10.1038/sj/neo/7900201.
25. Gemei M, Mirabelli P, Di Noto R, et al. CD66c is a novel marker for colorectal cancer stem cell isolation, and its silencing halts tumor growth in vivo. *Cancer*. 2013;119(4):729-738. doi:10.1002/cncr.27794.
26. Schölzel S, Zimmermann W, Schwarzkopf G, Grunert F, Rogaczewski B, Thompson J. Carcinoembryonic antigen family members CEACAM6 and CEACAM7 are differentially expressed in normal tissues and oppositely deregulated in hyperplastic colorectal polyps and early adenomas. *Am J Pathol*. 2000;156(2):595-605. doi:10.1016/S0002-9440(10)64764-5.
27. Kuroki M, Matsushita H, Matsumoto H, Hirose Y, Senba T, Yamamoto T. Nonspecific cross-reacting antigen-50/90 (NCA-50/90) as a new tumor marker. *Anticancer Res*. 1999;19(6 C):5599-5606.
28. Greco C, Bralet MP, Ailane N, et al. E-cadherin/p120-catenin and tetraspanin Co-029 cooperate for cell motility control in human colon carcinoma. *Cancer Res*. 2010;70(19):7674-7683. doi:10.1158/0008-5472.CAN-09-4482.
29. Guo Q, Xia B, Zhang F, et al. Tetraspanin CO-029 inhibits colorectal cancer cell movement by deregulating cell-matrix and cell-cell adhesions. *PLoS One*. 2012;7(6). doi:10.1371/journal.pone.0038464.
30. Nazarenko I, Rana S, Baumann A, et al. Cell surface tetraspanin Tspan8 contributes to molecular pathways of exosome-induced endothelial cell activation. *Cancer Res*. 2010;70(4):1668-1678. doi:10.1158/0008-5472.CAN-09-2470.
31. Matsumura N, Zembutsu H, Yamaguchi K, et al. Identification of novel molecular markers for detection of gastric cancer cells in the peripheral blood circulation using genome-wide microarray analysis. *Exp Ther Med*. 2011;2(4):705-713. doi:10.3892/etm.2011.252.
32. de Menezes-Neto A, Sáez MJ, Lozano-Ramos I, et al. Size-exclusion chromatography as a stand-alone methodology identifies novel markers in mass spectrometry analyses of plasma-derived vesicles from healthy individuals. *J Extracell Vesicles*. 2015;4(1). doi:10.3402/jev.v4.27378.
33. Heiler S, Wang Z, Zöller M. Pancreatic cancer stem cell markers and exosomes - The incentive push. *World J Gastroenterol*. 2016;22(26):5971-6007. doi:10.3748/wjg.v22.i26.5971.

Legends to figures

Figure 1: Enrolment and outcomes.

Study plan that describes the admission to the study of 103 subjects positive to the FIT test and further tested by colonoscopy. The results of colonoscopy permitted the stratification of this case record into 3 groups: subjects with low risk lesions (LR) n=36, subjects with negative colonoscopy (NFIT) n=36, and subjects with high risk lesions or affected by

colorectal cancer (HR/CRC) n=29 that were combined with 63 colorectal cancer patients (CRC) n=92. 67 healthy subjects (N) from our previous study were also included in this analysis.

Figure 2. Box plots distribution of the ΔCT **values for the CELTiC panel.**

The relative ΔCT values for the four groups of healthy control subjects (N), negative colonoscopy (NFIT), low risk lesion (LR), high risk lesion or colorectal cancers (HR/CRC) are reported for each marker *CEACAM6*, *LGALS4*, *TSPAN8* and *COL1A2*.

Figure 3: ROC curves of the CELTiC panel of the indicated comparisons.

Table 1. Clinical characteristics of the FIT positive study participants

Category	**N.**	**age**	**M**	**F**	**Subcategory**	**N.**	**age**	**M**	**F**	**Subcategory**	**N.**	**age**	**M**	**F**
No lesion	36	60±6.4	10	26										
					No clinical evidence	14		2	12	No CRC risk disease	22		8	14
										haemorrhoids	7			
										diverticulitis	13			
										aphthoid lesion	1			
										angiodysplasia	1			
Polyps	61		31	30										
					Non advanced	36	62.2±6.7	17	19	Advanced	25	60.04±10.5	14	11
Number														
1					16					8				
> 1					18					17				
N.D.					2									
size														
≤ 4 mm					16					2				
> 4 mm ≤10					13					13				
>10					0					10				
N.D.					7					0				
Type														
sessile					28					12				
pedunculated					3					12				
N.S.					5					1				

Histotype						
serrated					9	6
adenomatous					29	24
hyperplastic					3	1
villous					4	16
N.S.					4	0
Position						
right					15	9
left					12	13
rectum					4	2
N.S.					5	1
Cancer	4	65±3.5	3	1		
G1	1			1		
G2	3		3			

N.D.= not determined; N.S. = not specified; non advanced = polyp < 10 mm in the greatest dimension or with < 25% villous histologic features; advanced = high-grade dysplasia or with ≥ 25% villous histologic features or measuring ≥ 10 mm in the greatest dimension; CRC = colorectal cancer; G1 = grade 1; G2 = grade 2

Table 2. Descriptive statistics

	N	**NFIT**	**LR**	**HR/CRC**	Total	Kruskal-Wallis Rank Sum Test (adjusted p-values)					
Subjects n.	67	36	36	92	231	N vs NFIT	N vs LR	N vs HR/CRC	NFIT vs LR	NFIT vs HR/CRC	LR vs HR/CRC
	ΔCt ± SD	ΔCt ± SD	ΔCt ± SD	ΔCt ± SD	ΔCt ± SD						
CEACAM6											
mean ±sd	12.3±1.9	14.2±1.1	13.6±1.2	13.3±1.2	13.2±1.6	*<0.001*	*0.004*	*0.005*	*0.116*	*<0.001*	*1.091*
min	7.6	11.5	11.4	10.6	7.6						
max	15.6	16.2	15.3	16.6	16.6						
median	12.6	14.4	13.7	13.4	13.4						
LGALS4											
mean ±sd	12.9±2.0	15.7±1.3	15.3±0.8	14.7±1.3	14.4±1.8	*<0.001*	*<0.001*	*<0.001*	*1.4*	*0.003*	*0.034*
min	6.8	13.8	14	10.3	6.8						
max	16.4	19.5	17.5	18.3	19.5						
median	13.1	15.4	15.1	14.7	14.6						
TSPAN8											
mean ±sd	11.3±1.7	10.0±1.2	9.9±1.4	9.6±1.9	10.2±1.8	*<0.001*	*<0.001*	*<0.001*	*3.82*	*1.339*	*1.505*
min	8.3	8.2	7.4	4.8	4.8						
max	17.6	12.3	13.1	13.8	17.6						
median	11	9.8	10.1	10	10.2						
COL1A2											
mean ±sd	11.4±1.9	9.7±1.3	9.7±1.4	9.6±2.0	10.2±2.0	*<0.001*	*<0.001*	*<0.001*	*3,964*	*3.747*	*3.646*
min	7.8	7.1	6.6	4.8	4.8						
max	18.2	11.8	12.8	14	18.2						
median	11.2	9.6	9.7	9.8	10						
male%	52.2	27.8	47.2	52.2							
age											
mean ± sd	64.9±14.7	60.0±6.4	62.2±6.7	67.1±11.6	65.0±11.5	*0.176*	*0.736*	*2,555*	*0.527*	*0.006*	*0.222*

N = healthy controls; NFIT = negative colonoscopy; LR = low risk lesion; HR/CRC = high risk lesion or colorectal cancer

Table 3. CELTiC marker correlations

	CEACAM		*LGALS4*		*TSPAN8*		*COL1A2*	
	r	p-value	r	p-value	r	p-value	r	p-value
CEACAM6	1							
LGALS4	0.39	<0.001	1					
TSPAN8	-0.11	0.082	-0.08	0.23	1		0.87	<0.001
COL1A2	-0.14	0.03	-0.14	0.03			1	

Table 4. Multinomial logit model

	OR	se	p-value	95% C.I.	
LR vs N					
age	1	0.03	0.856	0.95	1.04
CEACAM6	1.29	0.2	0.194	0.91	1.67
LGALS4	4	0.26	<0.001	3.49	4.52
TSPAN8	0.79	0.39	0.55	0.04	1.55
COL1A2	0.61	0.36	0.168	-0.11	1.32
intercept	1.31E-07	4.63	<0.001	-9.15	9.15
HR/CRC vs N					
age	1.04	0.02	0.043	1	1.07
CEACAM6	1.24	0.16	0.181	0.93	1.56
LGALS4	2.72	0.21	<0.001	2.3	3.14
TSPAN8	0.46	0.29	0.006	-0.11	1.02
COL1A2	0.98	0.26	0.944	0.48	1.48
intercept	2.59E-05	3.59	0.002	-7.04	7.04
NFIT vs N					
age	0.97	0.03	0.36	0.92	1.03
CEACAM6	2.08	0.23	0.002	1.62	2.53
LGALS4	5.15	0.28	<0.001	4.6	5.69
TSPAN8	1.05	0.44	0.91	0.19	1.92
COL1A2	0.42	0.42	0.041	-0.4	-1.25
intercept	2.66E-11	5.31	<0.001	-10.4	10.4

LR=low risk; N=healthy controls; HR/CRC= hight risk/colorectal cancer;
NFIT= negative colonoscopy

Table 5 . CELTiC panel characteristics

Name	Chromosomal Location	Translation specificy	Protein subcellular location	Protein Information
CEACAM6 (carcinoembryonic antigen-related cell adhesion molecule 6) **Alias names:** NCA carcinoembryonic antigen-related cell adhesion molecule 6 (non-specific cross reacting antigen) **Alias symbols:** CD66c, CEAL	19q13 2 Start: 41,750,977 bp from pter End: 41,772,211 bp from pter **gene**: 6 exons **length**: 21,235 bps	**translation length**: 344 aminoacids; **molecular mass**: 37195 Da **quaternary structure**: Homodimeric and heterodimeric (CEACAM6-CEACAM8)	Plasma membrane; lipid-anchor; GPI-anchor; extracellular	Glycosyl phosphatidyl inositol (GPI) anchored cell surface glycoprotein Roles: in cell adhesion; tumor marker in serum immunoassay determinations of carcinoma Family members: 12
LGALS4 (lectin, galactoside binding soluble 4) **Alias names:** lectin **Alias symbols**: GAL4, L36LBP	19q13 2 Start: 38,801,671 bp from pter End: 38,813,544 bp from pter **gene:** 10 exons **length**: 11,874 bps	**translation length:** 323 aminoacids**;** **molecular mass:** 35941 Da **quaternary structure:** monomer	Plasma membrane; extracellular space, cytosol	Beta-galactoside-binding protein Roles: modulation of cell-cell and cell-matrix interactions Microvillar lipid raft stabilizer/organizer; polarized membrane trafficking; involved in the assembly of adherens junctions; cancer cell invasion Expression: restricted to small intestine, colon, and rectum Family members: 15
TSPAN8 (tetraspanin 8) **Alias name:** TM4SF3 (transmembrane 4 superfamily member 3) **Alias symbol:** CO-029	12q21 1 Start: 71,125,085 bp from pter End: 71,441,898 bp from pter **gene:** 9 exons **length**: 316,814 bps	**translation length:** 237 aminoacids; **molecular mass:** 26044 Da **quaternary structure:** No Data Available	Plasma membrane	Multi-pass membrane glycoprotein that complexes with integrins Roles: mediation of signal transduction events in the regulation of cell development, activation, growth and motility; involved in the promotion of angiogenesis Acts as a "molecular facilitator" by forming a web in glycolipid-enriched membrane microdomains, called TEM (tetraspanin enriched membrane domains) Present in exosomes Family members: 33
COL1A2 (collagen type I alpha 2 chain) **Alias names**: collagen type I alpha 2, collagen, type I, alpha 2, osteogenesis imperfecta type IV **Alias symbols**: OI4	7q21 3 Start: 94,394,561 bp from pter End: 94,431,232 bp from pter **gene**: 52 exons **length:** 36,672 bps	**translation length:** 1366 aminoacids **molecular mass:**129314 Da **quaternary structure:** Trimers of one alpha 2(I) and two alpha 1(I) chains	Secreted, extracellular space, extracellular matrix	Fibril-forming collagen, the most abundant collagen in the human body Roles: structural protein that interacts with other matrix proteins (proteoglycans, fibronectin); anchors cells into the matrix by binding to the cell surface integrins The expression is necessary for angiogenesis Up- or down-regulated in certain cancers Expressed by tumor stromal fibroblasts and vascular cells infiltrating the tumor Family members: 28

Figure 1

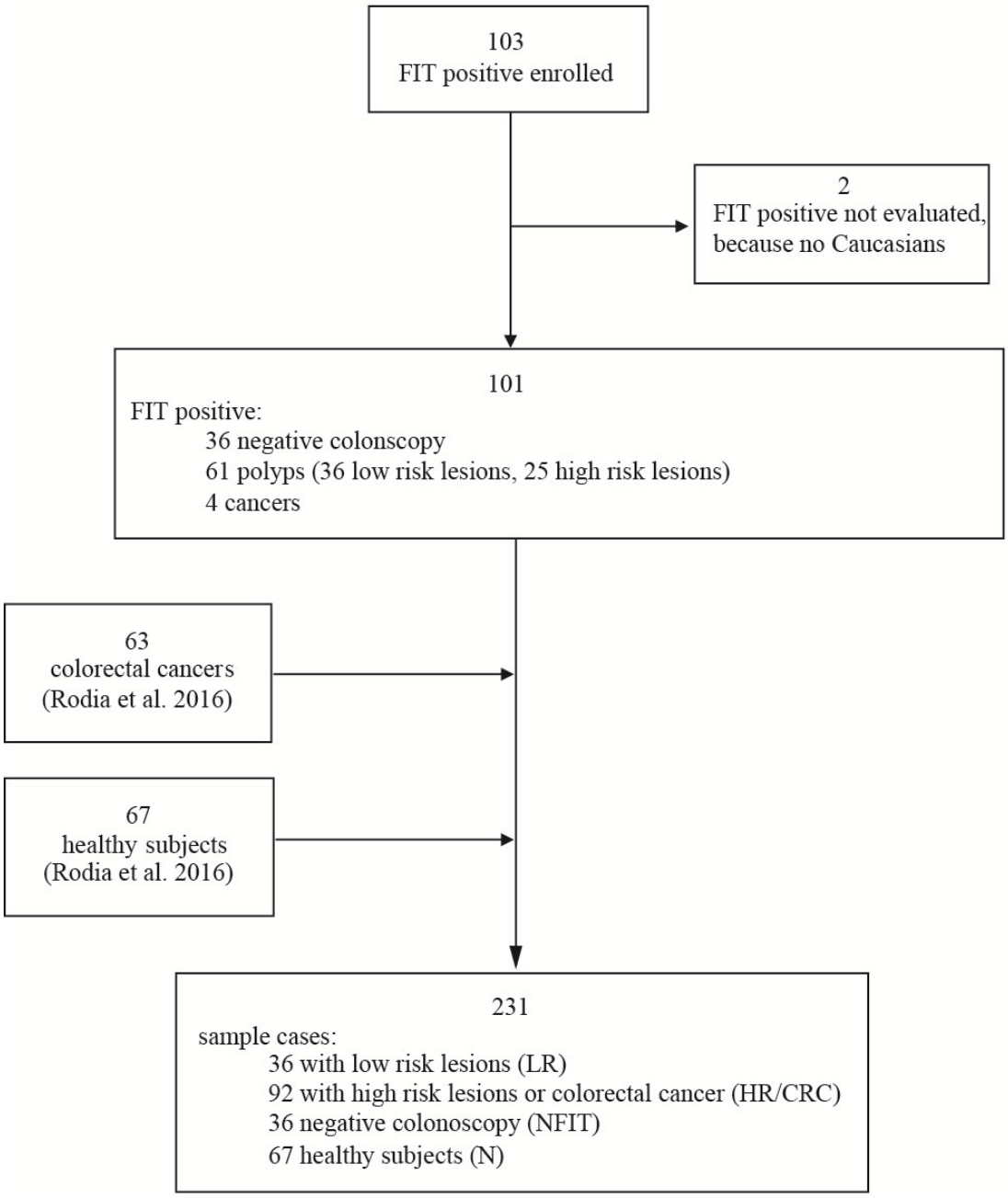

103
FIT positive enrolled
2
FIT positive not evaluated, because no Caucasians
101
FIT positive:
36 negative colonscopy
61 polyps (36 low risk lesions, 25 high risk lesions)
4 cancers
63
colorectal cancers
(Rodia et al. 2016)
67
healthy subjects
(Rodia et al. 2016)
231
sample cases:
36 with low risk lesions (LR)
92 with high risk lesions or colorectal cancer (HR/CRC)
36 negative colonoscopy (NFIT)
67 healthy subjects (N)

Figure 2

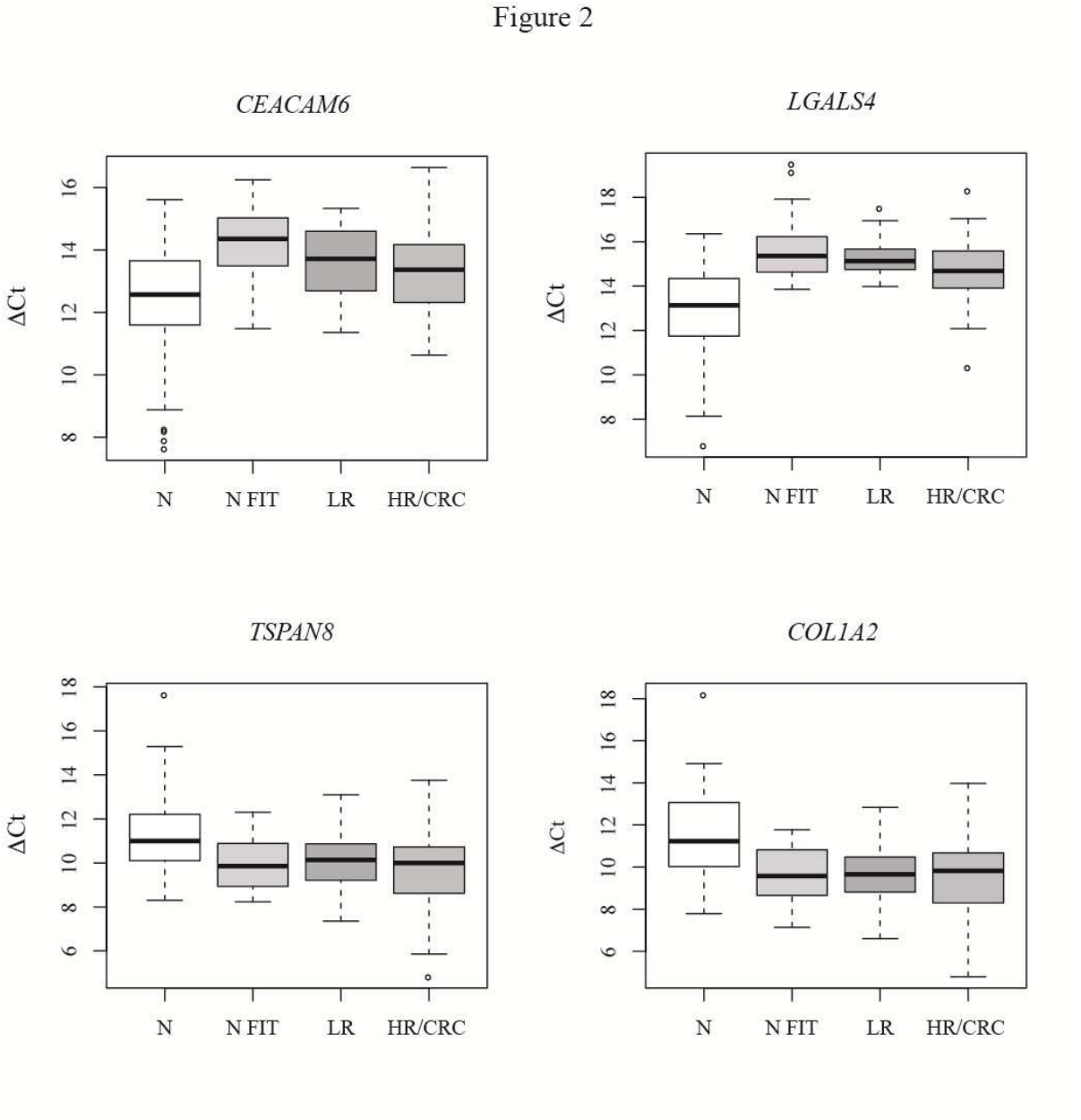
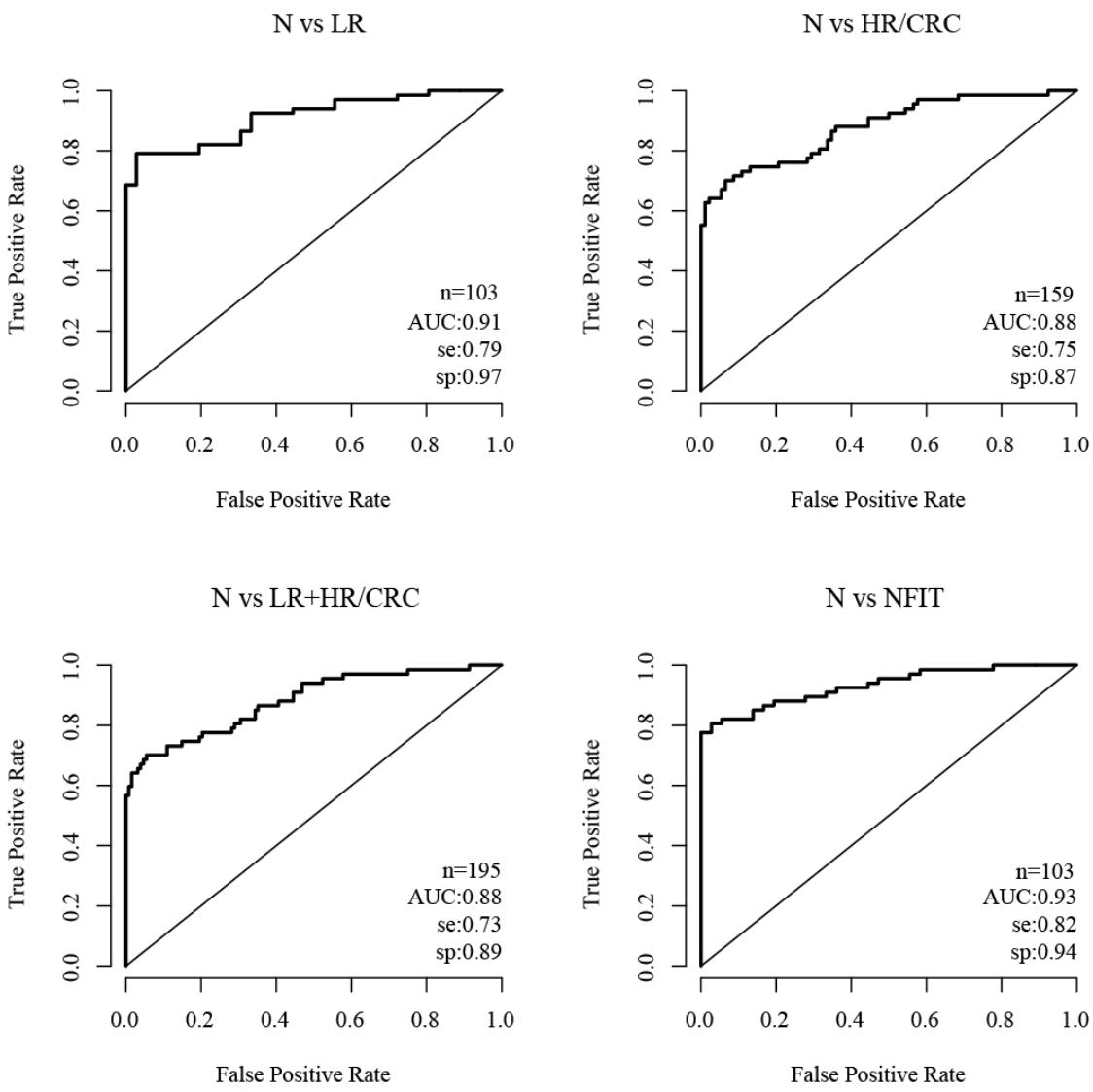


Figure 3


N vs LR
True Positive Rate
False Positive Rate
n=103
AUC:0.91
se:0.79
sp:0.97
N vs HR/CRC
True Positive Rate
False Positive Rate
n=159
AUC:0.88
se:0.75
sp:0.87
N vs LR+HR/CRC
True Positive Rate
False Positive Rate
n=195
AUC:0.88
se:0.73
sp:0.89
N vs NFIT
True Positive Rate
False Positive Rate
n=103
AUC:0.93
se:0.82
sp:0.94

Supplementary Figure 1