AP&T Alimentary Pharmacology and Therapeutics

A novel gene expression signature in peripheral blood mononuclear cells for early detection of colorectal cancer

C. Nichita*[,1], L. Ciarloni[†,‡,1], S. Monnier-Benoit[†], S. Hosseinian[†], G. Dorta* & C. Rüegg[‡,§]

*Gastroenterology and Hepatology Department, Centre Hospitalier Universitaire Vaudois (CHUV) and University of Lausanne (UNIL), Lausanne, Switzerland.
[†]Diagnoplex SA, Lausanne, Switzerland.
[‡]National Center for Competence in Research (NCCR), Molecular Oncology, Swiss Institute for Experimental Cancer Research (ISREC)-Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland.
[§]Department of Medicine, Faculty of Science, University of Fribourg, Fribourg, Switzerland.

Correspondence to:
Prof. C. Rüegg, Department of Medicine, Faculty of Science, University of Fribourg, 1, Rue Albert Gockel, CH-1700 Fribourg, Switzerland.
E-mail: curzio.ruegg@unifr.ch

Publication data
Submitted 4 November 2013
First decision 6 December 2013
Resubmitted 21 December 2013
Accepted 22 December 2013

[1]These authors are regarded as joint first authors, having contributed equally to the manuscript.

SUMMARY

Background
Early detection and treatment of colorectal adenomatous polyps (AP) and colorectal cancer (CRC) is associated with decreased mortality for CRC. However, accurate, non-invasive and compliant tests to screen for AP and early stages of CRC are not yet available. A blood-based screening test is highly attractive due to limited invasiveness and high acceptance rate among patients.

Aim
To demonstrate whether gene expression signatures in the peripheral blood mononuclear cells (PBMC) were able to detect the presence of AP and early stages CRC.

Methods
A total of 85 PBMC samples derived from colonoscopy-verified subjects without lesion (controls) (n = 41), with AP (n = 21) or with CRC (n = 23) were used as training sets. A 42-gene panel for CRC and AP discrimination, including genes identified by Digital Gene Expression-tag profiling of PBMC, and genes previously characterised and reported in the literature, was validated on the training set by qPCR. Logistic regression analysis followed by bootstrap validation determined CRC- and AP-specific classifiers, which discriminate patients with CRC and AP from controls.

Results
The CRC and AP classifiers were able to detect CRC with a sensitivity of 78% and AP with a sensitivity of 46% respectively. Both classifiers had a specificity of 92% with very low false-positive detection when applied on subjects with inflammatory bowel disease (n = 23) or tumours other than CRC (n = 14).

Conclusion
This pilot study demonstrates the potential of developing a minimally invasive, accurate test to screen patients at average risk for colorectal cancer, based on gene expression analysis of peripheral blood mononuclear cells obtained from a simple blood sample.

doi:10.1111/apt.12618

INTRODUCTION

Colorectal cancer (CRC) is the third most common cancer and second-leading cause of cancer-related death among men and women in Europe[1] and it fulfils the World Health Organisation (WHO) criteria for mass screening.[2] Importantly, CRC is often curable, when diagnosed at early stages. Moreover, adenomatous polyps (AP) detection and removal prevents CRC formation and decreases mortality due to CRC. Screening modalities for CRC have already been adopted by several countries, and clinical practice guidelines recommend that average-risk individuals begin regular screening at the age of 50.[3]

Colonoscopy is the 'gold standard' for CRC diagnosis; however, it is not the preferred method for screening because of its cost, invasiveness, low compliance and limited accessibility. Currently recommended non-invasive methods for mass screening include immunochemical and guaiac faecal occult blood testing (iFOBT, gFOBT). Yet, compliance with faecal tests is still suboptimal in countries with an FOBT screening program.[4–6] Therefore, there is still a large unmet screening need calling for a non- or minimally invasive, compliant, cost-effective and accurate test to detect AP and CRC at early stages.

A blood-based screening test is highly attractive due to its minimal invasiveness and high acceptance among patients. Several attempts have been made in the past to search for tumour markers in the blood to develop screening tests for CRC.[7–9] However, validation data obtained from a large screening-eligible population are still missing or showed mitigated test performances.[10]

Searching for blood-borne tumour markers could leverage different concepts such as the release of tumour-derived molecules (proteins, nucleic acids) into the blood stream, the presence of circulating tumour cells or the generation of a host response to tumour-derived signals. The latter is supported by the evidence that solid malignancies, to progress to clinically relevant tumours, require support from the tumour microenvironment, in particular from tumour-mobilised bone marrow-derived cells (BMDC).[11, 12] Tumour-recruited BMDC are likely to initiate differentiation and effector programs during their mobilisation from the bone marrow, which might be detectable during their transition in the blood.[13, 14] The report of signatures derived from peripheral blood mononuclear cells (PBMC) gene expression profiles and associated with breast,[15] renal,[16, 17] pulmonary,[18] bladder[19] and digestive cancers[8, 9, 20] further corroborates these observations. The aim of this study was to demonstrate whether the feasibility of identifying gene expression signatures in PBMC was able to discriminate patients with AP and CRC from subjects without these lesions. Moreover, we wanted to define predictive classifiers that could correctly classify Control, AP and CRC samples.

METHODS

Subjects

Participants in this monocentric case–control study were recruited from the Endoscopy Unit and from Urology, Gynecology, Pulmonary and Surgery units of the Centre Hospitalier Universitaire Vaudois (CHUV), Lausanne, Switzerland between March 2007 and March 2009. They included subjects without any colon lesion (control group) (n = 41), patients with adenoma (AP group) (n = 21) or colorectal carcinoma (CRC group) (n = 23). In addition, patients with inflammatory bowel disease (IBD) (n = 23) were also recruited. All these subjects underwent colonoscopy examination. In addition, patients with diagnosed tumours other than CRC (n = 14) were included to test the sensitivity of our classifiers to non-CRC tumours. These patients were not assessed by colonoscopy. Blood from all subjects was drawn before or immediately after colonoscopy, but before polypectomy or biopsy. The exclusion criteria included: age <18 years, alcohol or drug abuse, severe cardio-respiratory, liver, renal or gastrointestinal diseases, Hereditary Non Polyposis Colon Cancer (HNPCC) or Familial Adenomatous Polyposis (FAP). For the control, CRC and AP group, the concomitant presence of IBD or a malignant tumour other than CRC was also an exclusion criterion. Subjects signed a written informed consent before entering the study. The study was performed in accordance with the ethical principles of the Declaration of Helsinki and Good Clinical Practice guideline and was approved by the ethical committee of the Canton Vaud.

PBMC separation and RNA extraction

Blood was collected in 4 × 5 mL heparin tubes (Vacutainer, Becton Dickinson, Basel, Switzerland). PBMC separation was performed within 6 h using Histopaque®-1077 (Sigma-Aldrich, Buochs, Switzerland). RNA was extracted with the RNeasy Mini Kit (Qiagen, Basel, Switzerland); DNase was treated according to the manufacturer's instructions and stored at −80 °C.

Digital gene expression (DGE)-tag profiling and data analysis

Tag profiling libraries were prepared and sequenced at Fasteris SA (Plan-les-Ouates, Switzerland) using the

Illumina NlaIII DGE protocol and the Illumina Genome analyzer II, starting from 1 μg of total RNA. Tag alignment to a reference genome and tag counting were performed with GeneSifter® Analysis edition software (Geospiza-Perkin Elmer, Oftringen, Switzerland). After mapping and counting, a gene list with digital gene expression values was available for each sample. Normalised gene expression was calculated by multiplying each value with a linear scaling factor that was defined as the total number of tag reads obtained for a certain sample divided by the average number of tag reads obtained in all samples.[21]

Normalised data were log2 transformed, and Wilcoxon rank test, Negative binomial distribution[22, 23] and nonparametric t statistics[24] were used to identify differentially expressed genes. All statistical analyses were performed with R software (R-CRAN free software environment for statistical computing and graphics).

Three-dimensional principal component analysis (PCA), performed with Partek® Discovery Suite, was used to easily visualise the multi-dimensional PCR-derived data and reveal the internal structure of the data in a way that best explains the variance in the data. It was not used as a tool for gene selection.

Reverse transcription and single-channel quantitative multiplex PCR (scqmPCR)

200 ng of total RNA was reverse transcribed into cDNA in a final volume of 20 μL containing 4 U of Omniscript reverse transcriptase (Qiagen) in the manufacturer's buffer, 0.5 mmol/L of each dNTP, 10 U RNase inhibitor (Promega, Dübendorf, Switzerland) and 1 μmol/L NVd (T)'s (5′TTTTTTTTTTTTTTTTTTTTVN3′).

A modified version of the single-channel quantitative multiplex PCR (scqmPCR) described by Therianos *et al.*[25] was used. The first step of the scqmPCR consisted of the pre-amplification of the target sequences. Each 100 μL of the PCR reaction contained 2 μL of cDNA or plasmid, 1 U of HotStarTaq Plus DNA polymerase (Qiagen) in the manufacturer's buffer, 0.2 mmol/L of each dNTP and 15 μL of primer mixture. The primer mixture consisted of forward and reverse primers for a defined subset of the genes of interest, at a final concentration of 2 μmol/L each. The 106 different primer pairs (103 target genes and 3 reference genes) were equally split into three different primer mixtures. The PCR program consisted of 15 min at 95 °C to activate the polymerase, followed by 10 cycles of 30 s of denaturation at 95 °C, 30 s of annealing at 60 °C and 30 s of extension at 72 °C, and a final step of extension of 3 min at 72 °C.

The second, quantitative step of the scqmPCR was performed with 96-well plates loaded on the StepOnePlus real-time PCR instrument (Applied Biosystem, Zug, Switzerland). Each 20 μL of PCR reaction contained 1 μL of the first round scqmPCR reaction, 0.2 μmol/L of forward and reverse primer for one target gene, and the KAPA SYBR Green Fast qPCR Master Mix (Kapa Biosystems, Labgene Scientific, Châtel-St-Denis, Switzerland). Real-time PCR program consisted of 2 min at 95 °C for the Taq DNA Polymerase activation and 40 cycles of 3 s at 95 °C and 20 s at 60 °C. A melting curve analysis was performed to verify the specificity of each amplification product. All reactions were performed in duplicates. The StepOnePlus 2.1 software was used for the Ct determination, using a manually set threshold of 0.1 for all the PCR runs. Copy number values for each transcript were calculated by reference to standard curves. 10-fold serial dilutions of target-specific plasmids, ranging from 1 000 000 copies to 10 copies, were used to determine the linear relationship between copy number and Ct.

The normalised copy numbers were obtained by dividing the gene copy number by the median copy number of three housekeeping genes, RPLP0, NACA, B2M multiplied by a 10^5 factor.

PCR primers were designed using Primer3 software (available at http://www-genome.wi.mit.edu/genomesoftware/other/primer3.html) to specifically amplify between 180 and 200 base pairs for the target genes (Table S1).

Statistical analysis

In the gene selection phase, several univariate (*t*-test, Wilcoxon rank test[26] and univariate logistic regression[27, 28]) and multivariate statistical methods (classification and regression tree,[29] logistic regression[27, 28] and top scoring pair[30]) were applied on normalised PCR-derived gene expression values. All test results for each gene were summarised into a score and used for selecting the genes with the highest power in group discrimination (data not shown). This multi-test approach was chosen to maximise the capture of the information carried by the genes.

In the modelling phase, penalised logistic regression models[31, 32] were fitted on two subsets of the training set, one including all CRC patients and controls and one including all adenoma patients and controls, and two different classifiers were established, the CRC and AP classifier respectively.

Fitted models were validated by non-overlapped bootstrap method[33]: 500 random data sets were drawn with replacement from training set; each bootstrap had the same size as the training set. The model was re-fitted at

each bootstrap and validated with the out-of-bag samples. The specificity and sensitivity average values over 500 bootstraps were calculated and Receiver Operating Characteristics (ROC) curves were generated. All analyses were performed with R software unless otherwise indicated.

RESULTS

Characteristics of the study population

Clinical and demographic characteristics of the study population are summarised in Table 1. The study was not designed to be age-matched; therefore, differences were present between the age of control and case groups. The study population included 41 colonoscopy-verified controls, 21 patients with adenomas and 23 patients with colorectal cancer. The great majority of patients in the AP group had advanced adenomas (17/21), being larger than 1 cm or having a villous component, and the greater part of CRC were at an early stage (Table 1). Adenomas and CRCs were mainly located in the left colon (62% and 77%, respectively, Table S2A), thereby reflecting the typical enrichment of these pathologies in the lower third of the colon. Moreover, 14 patients with malignant tumours other than CRC and 23 IBD patients were recruited and included in the study as separate groups (Table 1, Table S2B, S2C).

Table 1 | Clinical and demographic characteristic of the study population

	Patients (*n*)	Age (years) mean /median (min-max)	Male
Controls	41	52/56 (21–85)	39%
Adenomas	21	68/67 (48–84)	52%
Advanced	17		
<5 mm	4		
CRC	23	65/62 (30–88)	52%
I	6		
II	6		
III	5		
IV	2		
Unknown Stage	4		
IBD	23	42.5/40 (25–70)	52%
Crohn's disease	16		
Ulcerative colitis	6		
Behcet	1		
Other cancers	14	65/66 (54–77)	71%
Prostate	8		
Lung	4		
Breast	1		
Pancreas	1		

CRC, colorectal cancer; IBD, inflammatory bowel disease.

Biomarker identification

Biomarker candidates were identified through two different approaches (Figure 1). In a first candidate marker approach, we reviewed literature for potential biomarkers associated with CRC and we mined unpublished microarray-based gene expression data, previously generated in our laboratory (C. Rüegg, unpublished results). This approach led to the selection of a panel of 49 candidate biomarkers related to angiogenesis and colorectal cancer to be further evaluated by PCR. In a second, complementary approach, we conducted a whole-transcriptome analysis of a subset of the study population, i.e. 33 PBMC samples from 16 control subjects, 13 patients with adenoma and from 4 patients with CRC, by DGE-tag profiling. This method entails the capture of a 17-nucleotide (nt) sequence immediately downstream of the 3′-most NlaIII site in each polyadenylated RNA. These 17 nt 'tags' are sequenced in a high-throughput manner and the number of occurrences of each unique tag is counted, resulting in digital gene expression

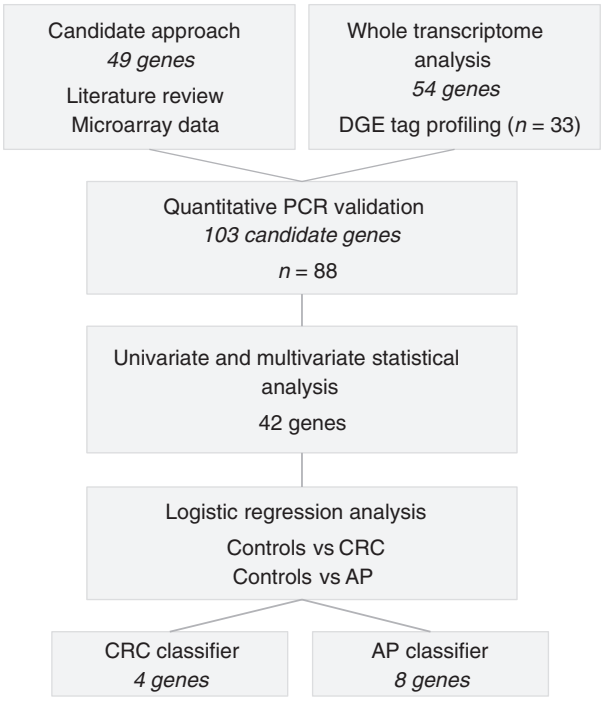

Figure 1 | Classifiers definition workflow. After the identification of a pool of 103 potential biomarkers by two complementary approaches, the marker pool underwent validation by quantitative PCR. Statistical analysis retained 42 genes that were used to fit logistic regression models and generate a CRC and an AP classifier. DGE, digital gene expression; CRC, colorectal cancer; AP, adenomatous polyp.

profiles where tag counts reflect expression levels of the corresponding transcript.[34]

A total of 20 288 unique gene transcripts with a tag count greater than 1 were identified. The transcripts showing a median count equal to zero across all samples were filtered out. The remaining 8843 genes underwent differential gene expression analysis based on significant p value (<0.05) and a gene expression fold change greater than 2 between control and case groups. Eighty-eight genes were identified as significantly overrepresented or underrepresented in the adenoma group compared with the control group, and 54 were retained for PCR validation based on gene function and overall expression levels (data not shown).

PCA was applied to the 54-gene data set to visualise in a three-dimensional space the overall variability in the data (Figure 2). Controls, adenomas and CRC samples showed distinct spacial distribution with few overlaps, suggesting that they possess specific gene expression patterns.

Biomarker validation by quantitative PCR

A total of 103 potential biomarkers identified by candidate gene approach and by whole-transcriptome analysis (Table S1) were subjected to scqmPCR validation on PBMC samples derived from control (n = 41), adenoma (n = 24) and CRC (n = 23) patients (Figure 1). Univariate analysis identified 24 genes with significant test (P < 0.05) and gene expression fold change greater than 1.5 between the control and adenoma or carcinoma groups (Table 2). Most of the genes showed up-regulation, whereas only two genes were down-regulated (BANK1, VPS18), confirming the findings from the tag profiling and the literature. In general, genes were able to discriminate both adenoma and CRC groups from the controls, although with lesser power for the adenoma group. CTSL1 was the gene most significantly up-regulated in the CRC group compared with the control group. ITIH4 and TUG1 were the genes most significantly up-regulated in the adenoma group compared with the control group. Only two genes (LST1 and TRIM24) were specifically significant for the discrimination of the adenoma group. In the CRC group, the two genes were also up-regulated, but statistical significance did not reached the cut-off value, although it was very close to. This suggests that LST1 and TRIM24 expression in cancer patients is more heterogeneous than in adenoma ones, possibly depending on disease stage, rather than a full down-regulation of the genes as the disease progress from benign to malignant.

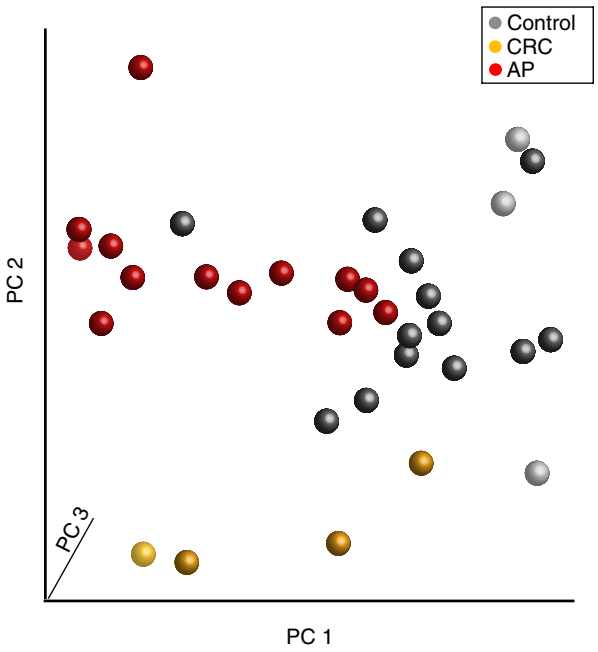

Figure 2 | Three-dimensional PCA of the 54 differentially expressed genes identified from the DGE-tag profiling data. PBMC samples from adenoma patients (n = 13) are represented by red dots, from CRC patients (n = 4) by orange dots and from control subjects (n = 16) by grey ones.

In order not to miss genes that might have a discriminatory power only when used in combinations, we conducted a multivariate analysis of the 103 genes. Eighteen genes that were not initially retained after the univariate analysis were additionally identified as significant by this analysis (data not shown, Table S1). In conclusion, a total of 42 biomarkers were validated by scqmPCR and retained for the subsequent modelling phase (Figure 1).

Interestingly, functional analysis performed with Ingenuity Pathway Analysis software (Table 3) revealed that some of these genes were associated with inflammatory conditions. Moreover, the 42-gene panel is enriched in genes involved in leucocytes trafficking, suggesting that up-regulation of these genes in patients may reflect an increased or specific capacity of colorectal tumours to attract monocytes/macrophages. Taken together, the results of this analysis support the notion that CRC is able to elicit an inflammatory-like reaction in circulating PBMC.

Discriminatory power of predictive classifiers

The 42-gene data set was utilised to generate predictive models. Two sample subsets, controls and adenomas or controls and CRCs, were used to define an AP classifier and a CRC classifier. Penalised logistic regression models

Table 2 | Differentially expressed genes among control, AP and CRC groups determined by univariate analysis of the quantitative PCR data set. *t*-test *P* values were representative of the univariate analysis and therefore reported in the table

Gene Name	Description	Discovery	*t*-test *P* value (AP vs. Con)	*t*-test *P* value (CRC vs. Con)	Fold Change (AP/Con)	Fold Change (CRC/Con)
CTSL1	Cathepsin L1	DGE	0.043	9.40E-05	4.42	9.10
GK	Glycerol kinase	DGE	0.006	0.0002	3.83	3.96
CDA	Cytidine deaminase	Literature[8]	0.018	0.00047	1.49	1.77
SET	SET translocation (myeloid leukaemia-associated)	DGE	0.070	0.001	4.06	2.37
PFDN5	Prefoldin subunit 5	DGE	0.082	0.002	3.88	2.43
PECAM1	Platelet/endothelial cell adhesion molecule 1 (CD31 antigen)	DGE	0.013	0.0017	3.46	3.09
APOBEC3A	Apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like 3A	DGE	0.196	0.002	2.37	2.50
UBXD5	UBX domain containing 5	DGE	0.0494	0.002	4.47	3.48
MSL1	Male-specific lethal-1 homolog	DGE	0.078	0.002	3.04	3.98
MMP9	Matrix metallopeptidase 9 (gelatinase B, 92 kDa gelatinase, 92 kDa type IV collagenase)	Literature[43]	0.007	0.0030	1.60	3.97
BANK1	B-cell scaffold protein with ankyrin repeats 1	Literature[8]	0.405	0.003	−2.14	−4.80
C9orf78	Chromosome 9 open reading frame 78	DGE	0.082	0.003	3.42	2.80
VPS18	Vacuolar protein sorting 18 homolog (S. cerevisiae)	DGE	0.876	0.004	−2.52	5.51
ISCU	Iron-sulphur cluster scaffold homolog (E. coli)	DGE	0.191	0.006	2.91	2.67
DYNC1LI2	Dynein, cytoplasmic 1, light intermediate chain 2	DGE	0.031	0.006	4.15	2.14
DYM	Dymeclin	DGE	0.013	0.0070	3.50	3.16
PIP4K2B	Phosphatidylinositol-5-phosphate 4-kinase, type II, beta	DGE	0.016	0.009	4.17	2.34
TUG1	Taurine up-regulated gene 1	DGE	0.006	0.023	5.35	1.87
EPHX2	Epoxide hydrolase 2, cytoplasmic	DGE	0.056	0.026	3.19	3.66
ITIH4	Inter-alpha (globulin) inhibitor H4 (plasma Kallikrein-sensitive glycoprotein)	DGE	0.026	0.028	5.48	1.71
CDCA4	Cell division cycle associated 4	DGE	0.085	0.035	2.58	1.80
S100A8	S100 calcium-binding protein A8	Literature[48]	0.159	0.047	1.32	2.11
LST1	Leucocyte-specific transcript 1	DGE	0.008	0.069	3.52	1.59
TRIM24	Tripartite motif-containing 24	DGE	0.007	0.077	2.19	3.20

DGE, digital gene expression; AP, adenoma; CON, control; CRC, colorectal cancer.

were fitted on these two subsets and validated by bootstrap (Figure 1). The two best classifiers were identified as follow:

AP classifier:

$$\log\left(\frac{\Pr(y_i=1)}{1-\Pr(y_i=1)}\right) = -3.959 + 0.061 \times \mathrm{GK} + 0.10 \times \mathrm{MMP9} + 0.053 \times \mathrm{TRIM24} + 0.081 \times \mathrm{TUG1}$$

CRC classifier:

$$\log\left(\frac{\Pr(y_i=1)}{1-\Pr(y_i=1)}\right) = -4.42 - 0.06 \times \mathrm{BANK1} + 0.162 \times \mathrm{CDA} + 0.29 \times \mathrm{CTSL1} + 0.121 \times \mathrm{GK} + 0.094 \times \mathrm{MMP9} + 0.034 \times \mathrm{PECAM1} + 0.005 \times \mathrm{PIP4K2B} + 0.046 \times \mathrm{TUG1}$$

Table 3 | Functional analysis of the 42 biomarkers selected for the modelling phase. The table reports the most significantly represented biological functions and diseases within the 42-biomarker panel. *P* values measure the likelihood that the association between a set of biomarkers and a given IPA functional category is random. The number of genes associated with a specific function is reported in the last column. The analysis was performed with the Ingenuity Pathway Analysis (IPA) software

Category	Diseases or functions	*P* value	Number of genes
Inflammatory and immunological disease	Rheumatic disease, arthritis, systemic autoimmune syndrome, lupus erythematosus	7.99E-07	17
Immune cell trafficking and haematological system development/function	Phagocyte migration and transmigration	3.13E-04	4
Cellular growth and proliferation	Cell proliferation	6.01E-04	20
Dermatological diseases	Psoriasis	6.02E-04	10
Inflammatory response	Organ inflammation	9.43E-04	10
Cell death and survival	Cell death	1.04E-03	18

with y_i being either 0 or 1 according to the pre-defined classification groups.

The bootstrap validation showed for the CRC classifier an average sensitivity and specificity of 78% and 92%, and for the adenoma classifier an average sensitivity and specificity of 46% and 92% respectively. ROC analysis determined an average area under the curve (AUC) of 0.91 (0.83–0.98, 95% CI) for the CRC classifier and an average AUC of 0.76 (0.59–0.88, 95% CI) for the AP one (Figure 3a and b).

Classifiers performance in other diseases

The specificity of the two predictive classifiers was independently evaluated in subjects with IBD. The CRC and AP classifiers showed a specificity of 91% (2/23) and 96% (1/23) towards IBD respectively. Subjects with malignant tumours other than CRC were also tested to verify the sensitivity of the two classifiers towards tumours other than CRC. None of these subjects were classified as CRC or AP by our classifiers, suggesting that both classifiers are highly specific to colorectal tumours. Unfortunately, these subjects could not be assessed by colonoscopy and therefore the concomitant presence of colorectal lesions could not be formally excluded. For this reason, they were not used as an independent set to assess the classifiers specificity.

DISCUSSION

The aim of the study was to identify PBMC-derived biomarkers and to develop predictive classifiers, which are able to discriminate patients with CRC and AP from healthy controls. The lack of non-invasive detection tools for the adenoma prompted us to focus our study not only on the identification of specific biomarkers/

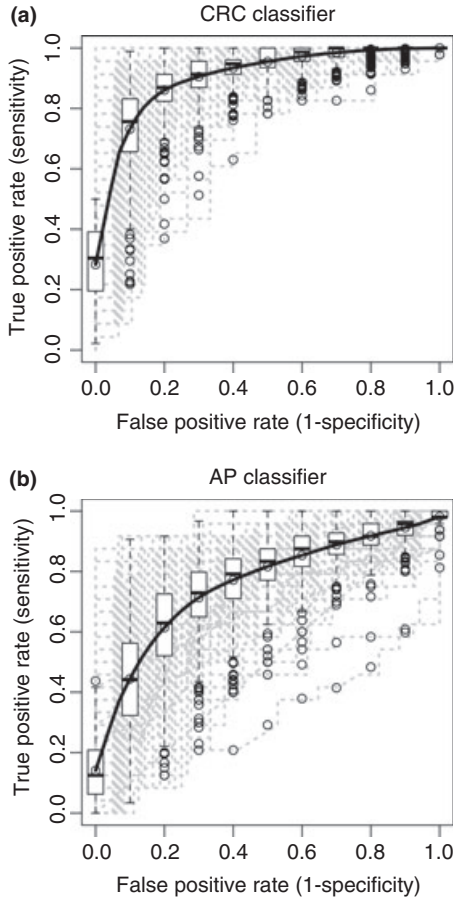

Figure 3 | Receiver Operating Characteristics (ROC) analysis of the CRC (a) and the AP classifiers (b) using 500 bootstraps validation. The boxplots represent the distribution of the 500 bootstraps. The black line represents the average values over 500 bootstraps for clinical specificity and sensitivity.

classifiers for CRC but also for those specific for adenomatous polyps, the main precursor lesion of CRC.

In particular, the biomarker discovery approach based on whole-transcriptome analysis was conceived with this goal in mind and this was reflected by the inclusion of samples issued almost exclusively from adenoma patients. Using combined candidate and whole-transcriptome approaches, we identified and validated by quantitative PCR a set of 42 genes that are able to discriminate between the control group and the CRC or AP group. Specific biomarker combinations for CRC and AP discrimination were determined by logistic regression analysis, resulting in an eight- and four-gene predictive classifier respectively. With a specificity of 92%, the CRC classifier showed a sensitivity of 78% and the AP classifier a sensitivity of 46%. The early stage of development of our test does not yet allow a full comparison with existing screening methods. Nevertheless, it is of interest to note that the adenoma detection rate (46%) was superior to the ones reported for faecal immunochemical testing (FIT, 20%-29%),[35, 36] or for new tests such as the Septin 9 test (11%).[10]

Our biomarker discovery approach identified new candidate genes, as well as it confirmed already known CRC biomarkers. In particular, TUG1, TRIM24, PIP4K2B, GK, among the novel ones and BANK1, CDA, PECAM1/CD31, CTSL1, MMP9 among the known ones were identified by our eight- and four-gene classifiers as the most discriminant markers.

TUG1 (Taurine Up-regulated Gene 1) is a long intergenic noncoding RNAs, a class of noncoding RNAs that regulate gene expression via chromatin reprogramming. Besides its role in retina development,[37] little is known about its function, including in leucocytes. Recently, it was shown that it was up-regulated in urothelial carcinoma of the bladder and that silencing of TUG1 inhibited cell proliferation and induced apoptosis in urothelial carcinoma cells.[38]

TRIM24, also known as HTIF1 alpha, functions as a co-regulator that positively or negatively modulates the transcriptional activities of several nuclear receptors, including estrogen receptor, retinoic acid receptor, and of p53. TRIM24 is overexpressed and associated with poor prognosis in breast cancer[39] and head and neck squamous cell carcinoma,[40] and in nonsmall cell lung carcinoma, TRIM24 overexpression correlates with tumour progression.[41] Interestingly, TRIM24 is expressed in blood cells and it was suggested to play a role in myeloid differentiation. In particular, its down-regulation in immature myeloid cells was required for monocyte-macrophage maturation.[42]

MMP9 is a proteolytic enzyme produced by inflammatory cells, promoting tumour progression by remodelling the extracellular matrix and basal membranes and favouring tumour angiogenesis. MMP-9 is strongly expressed in colorectal cancer with a significantly higher expression in high-grade adenoma.[43] Tumour-associated macrophages (TAM), whose presence is strongly associated with CRC progression, are the main source of MMP9 in the tumour microenvironment.[44] Of interest, we previously reported elevated levels of circulating MMP-9 in the blood of CRC patients.[45] MMP-9 up-regulation alone, however, is not specific to CRC, as it has also been reported in other malignant tumours and in autoimmune or inflammatory diseases, including ulcerative colitis or lymphocytic colitis.[46]

About 40% of the 42-gene panel was also shown to be involved in other inflammatory processes (Table 3). This observation raised the possibility that inflammatory conditions, in particular intestinal ones, could also be detected by our classifiers, resulting in a decreased specificity for CRC detection. A potential risk was largely ruled out by the observation that 91–96% of the IBD subjects were correctly identified as controls and not as CRC patients. The reasons for this high level of specificity for CRC of our signature rich in inflammatory genes are not clear at this point. It is tempting to speculate that inflammatory and related genes expressed in CRC are distinct from, or largely non-overlapping with, genes expressed in IBD or other inflammatory conditions. Importantly, none of the subjects with other type of malignancies was misclassified, thereby confirming the specificity of the classifier for CRC. The combination of multiple biomarkers is indeed a way to palliate for powerful but nonspecific information carried by single biomarkers.

Cancer detection using a peripheral blood test is an attractive screening method because of its simplicity in clinical practice, which is expected to translate into improved compliance compared with colonoscopy and FOBT, with potential major impact in public health. Several recent reports confirmed the blood cells as a precious source of candidate gene expression markers for cancer detection. Using a 25-gene model, Honda *et al.* found a distinct gene expression profile in the blood of patients with digestive cancers compared with healthy individuals.[20] Similar results have been reported for other type of tumours, such as breast, bladder, small cell lung cancer and renal carcinoma.[15–19] In particular, different panels of peripheral blood biomarkers based on gene expression have been described in recent studies to

differentiate CRC from controls. Han *et al.* identified and validated a five-gene combination that could discriminate CRC from non-CRC samples with sensitivity and specificity of 94% and 77% respectively.[8] Two genes of the combination, CDA and BANK1, were also confirmed to be valuable biomarkers for CRC detection in our study. The same group developed a seven-gene, blood-based biomarker signature that could stratify subjects according to their current relative risk to develop a CRC across a broad range in an average-risk population.[9]

In spite of the promising results, this study has two intrinsic limitations. The first one is the small sample size. This did not allow us to validate the signatures with an independent set of control, cancer and adenoma samples. Internal validation methods such as non-overlapped bootstrap have been demonstrated to be effective in reducing the risks of model over-fitting.[41] However, this risk of over-fitting still exists, as this validation method is not prospective and utilises the same sample population used for modelling; thus, the model's performances might be overoptimistic. A validation study with a fully independent data set is therefore necessary for an accurate estimation of the predictor performance.[47] The small sample size also prevented us to stratify the classifier sensitivity by cancer stage. The CRC classifier has been developed starting with a slight majority of CRC stage I-II; thus, sensitivity for the detection of early stage CRCs is expected to be similar to the overall sensitivity.

The second limitation is that the study was not designed to be age-matched between the control and the case subjects. Also, patients were not recruited according to the age criteria fixed by colorectal cancer screening guidelines and therefore patients younger than 50 year were also included.

In conclusion, in this study, we identified in PBMC new gene expression signatures and predictive classifiers specific for colorectal cancer and adenomas. This study provided the proof of concept that developing a minimally invasive, first intention test to screen patient with average risk of colorectal cancer from a simple blood draw is feasible. However, the road towards a marketable test is still long. A large prospective multi-centric study, which was recently concluded and included an independent validation set, will be used to develop the test prototype and to provide its clinical validation. For instance, the two identified classifiers could be combined by a decisional algorithm that release a positive or negative binary answer to be used for the triage of average-risk asymptomatic subjects before colonoscopy. Ultimately, a screening study performed on a screening-eligible population and a study comparing the final test to a currently used CRC screening test such as FIT is necessary to determine the clinical utility of the test.

AUTHORSHIP

Guarantor of the article: Curzio Rüegg.

Author contributions: C. Rüegg performed the conception and design. L. Ciarloni, S.Hosseinian, S. Monnier-Benoit performed the development of methodology. C. Nichita, G. Dorta, L. Ciarloni, S. Monnier-Benoit performed the acquisition of data. C. Nichita, L. Ciarloni, S. Monnier-Benoit, S. Hosseinian performed the analysis and interpretation of data. C. Nichita, L. Ciarloni, C. Rüegg performed the writing, review and/or revision of the manuscript. C. Nichita, G. Dorta, L. Ciarloni, S. Monnier-Benoit, S. Hosseinian performed the administrative, technical or material support. C. Rüegg, L. Ciarloni, S. Monnier-Benoit performed the study supervision. All authors read and approved the final manuscript.

ACKNOWLEDGMENTS

Declaration of personal interests: L. Ciarloni, S. Hosseinian and S. Monnier-Benoit are employees of Diagnoplex SA. Also, they own stock options in Diagnoplex SA. C. Rüegg is a co-founder, advisory board member and owns stocks and shares of Diagnoplex SA. C. Nichita has served as speaker for Diagnoplex SA. G. Dorta has served as speaker, consultant and advisory board member for Diagnoplex SA, and has received research funding from Diagnoplex SA.

We would like to thank S. Therianos for initiating and promoting the project, N. Rochat, M. Corboz, Y. Risse, J. Wyneger and N. Levi for their technical support in data acquisition, Dr. N. Vietti-Violi for helping in recruiting the patients and N. Imaizumi for her advises in drafting the manuscript.

Declaration of funding interests: This study was supported in part by Diagnoplex SA. Moreover, it was supported by the Gebert Rüef Stiftung, grant number 063/05, the Commission for Technology and Innovation (CTI/KTI), grant number 8463.1 LSPP-LS, the Swiss Cancer League, grant number CCRP OCS-01812-12-2005 and by the National Center of Competence in Research (NCCR) Molecular Oncology, a research instrument of the Swiss National Science Foundation.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article:

Table S1. List of the 103 biomarkers identified in the study and analysed by qPCR. The primer sequences used

for qPCR amplification are reported. The statistics column indicates whether the gene was significant after univariate or multivariate analysis. The discovery column reports the method by which the gene was identified.

Table S2. (A) Adenoma and CRC localisation in the colon. (B) Type of tumours other than CRC and TNM staging. (C) IBD patients' clinical characteristics.

REFERENCES

1. Jemal A, Siegel R, Xu J, Ward E. Cancer statistics, 2010. *CA Cancer J Clin* 2010; **60**: 277–300.
2. Wilson JMG, Jungner G. *Principles and Practice of Screening for Disease*. Geneva: WHO, 1968.
3. Levin B, Liberman DA, McFarland B, *et al.* Screening and surveillance for the early detection of colorectal cancer and adenomatous polyps, 2008: a joint guideline for the American Cancer Society, the US Multi-Society Task Force on Colorectal Cancer, and the American College of Radiology. *Gastroenerology* 2008; **134**: 1570–95.
4. Parente F, Boemo C, Ardizzoia A, *et al.* Outcomes and cost evaluation of the first two rounds of a colorectal cancer screening program based on immunochemical fecal occult blood test in northern Italy. *Endoscopy* 2013; **45**: 27–34.
5. Jean-Jacques M, Kaleba EO, Gatta JL, Gracia G, Ryan ER, Choucair BN. Program to improve colorectal cancer screening in a low-income, racially diverse population: a randomized controlled trial. *Ann Fam Med* 2012; **10**: 412–7.
6. Von Wagner C, Baio G, Raine R, *et al.* Inequalities in participation in an organized national colorectal cancer screening programme: results from the first 2.6 million invitations in England. *Int J Epidemiol* 2011; **40**: 712–8.
7. Grützmann R, Molnar B, Pilarsky C, *et al.* Sensitive detection of colorectal cancer in peripheral blood by septin 9 DNA methylation assay. *PLoS ONE* 2008; **3**: e3759.
8. Han M, Liew CT, Zhang HW, *et al.* Novel blood-based, five-gene biomarker set for the detection of colorectal cancer. *Clin Cancer Res* 2008; **14**: 455–60.
9. Marshall KW, Mohr S, Khettabi F, *et al.* A blood-based biomarker panel for stratifying current risk for colorectal cancer. *Int J Cancer* 2010; **126**: 1177–86.
10. Church TR, Wandell M, Lofton-Day C, *et al.* Prospective evaluation of methylated SEPT9 in plasma for detection of asymptomatic colorectal cancer. *Gut* 2014; **3**: 317–25.
11. Lorusso G, Rüegg C. The tumor microenvironment and its contribution to tumor evolution toward metastasis. *Histochem Cell Biol* 2008; **130**: 1091–103.
12. Coussens LM, Werb Z. Inflammation and cancer. *Nature* 2002; **420**: 860–7.
13. Laurent J, Hull EF, Touvrey C, *et al.* Proangiogenic factor PIGF programs CD11b(+) myelomonocytes in breats cancer during differentiation of their hematopoietic progenitors. *Cancer Res* 2011; **71**: 3781–91.
14. Cofflet SB, Tal AO, Scholz A, *et al.* Angiopoietin-2 regulates gene expression in TIE2-expressing monocytes and augments their inherent proangiogenic functions. *Cancer Res* 2010; **70**: 5270–80.
15. Sharma P, Sahni NS, Tibshirani R, *et al.* Early detection of breast cancer based on gene-expression patterns in peripheral blood cells. *Breast Cancer Res* 2005; **7**: R634–44.
16. Twine NC, Stover JA, Marshall B, *et al.* Disease-associated expression profiles in peripheral blood mononuclear cells from patients with advanced renal cell carcinoma. *Cancer Res* 2003; **63**: 6069–75.
17. Burczynski ME, Twine NC, Dukart G, *et al.* Transcriptional profiles in peripheral blood mononuclear cells prognostic of clinical outcomes in patients with advanced renal carcinoma. *Clin Cancer Res* 2005; **11**: 1181–9.
18. Showe MK, Vachani A, Kossenkov AV, *et al.* Gene expression profiles in peripheral blood mononuclear cells can distinguish patients with non-small cell lung cancer from patients with nonmalignant lung disease. *Cancer Res* 2009; **69**: 9202–10.
19. Osman I, Bajorin DF, Sun TT, *et al.* Novel blood biomarkers of human urinary bladder cancer. *Clin Cancer Res* 2006; **12**(11 Pt 1): 3374–80.
20. Honda M, Sakai Y, Yamashita T, *et al.* Differential gene expression profiling in blood from patients with digestive system cancers. *Biochemic Biophys Res Commun* 2010; **400**: 7–15.
21. 't Hoen PA, Ariyurek Y, Thygesen HH, *et al.* Deep sequencing-based expression analysis shows major advances in robustness, resolution and inter-lab portability over five microarray platforms. *Nucleic Acids Res* 2008; **36**: e141.
22. Robinson MD, Smyth GK. Small-sample estimation of negative binomial dispersion, with applications to SAGE data. *Biostatistics* 2008; **9**: 321–32.
23. Robinson MD, Smyth GK. Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics* 2007; **23**: 2881–7.
24. White JR, Nagarajan N, Pop M. Statistical methods for detecting differentially abundant features in clinical metagenomic samples. *PLoS Comput Biol* 2009; **5**: e1000352.
25. Therianos S, Zhu M, Pyun E, Coleman PD. Single-channel quantitative multiplex reverse transcriptase-polymerase chain reaction for large numbers of gene products differentiates nondemented from neuropathological Alzheimer's disease. *Am J Pathol* 2004; **164**: 795–806.
26. Wilcoxon Frank. Individual comparisons by ranking methods. *Biometrics Bulletin*. 1945; **1**: 80–3.
27. Dobson AJ. *An Introduction to Generalized Linear Models*. 2nd ed. Boca Raton: Chapman & Hall/CRC, 2002.
28. McCullagh P, Nelder JA. *Generalized Linear Models*. 2nd ed. Boca Raton: Chapman & Hall/CRC, 1989.
29. Breiman L, Friedman JH, Olshen RA, Stone CJ. *Classification and Regression Trees (CART)*. Belmont: Wadsworth, 1984.
30. Geman D, d'Avignon C, Naiman DQ, Winslow RL. Classifying gene expression profiles from pairwise mRNA comparisons. *Stat Appl Genet Mol Biol*. 2004; **3**: Article 19.
31. Park MY, Hastie T. L1 regularization path algorithm for generalized linear models. *J R Stat Soc B* 2007; **69**: 659–77.
32. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J Stat Softw* 2010; **33**: 1–22.
33. Steyerberg EW, Harrell FE Jr, Borsboom GJ, Eijkemans MJ, Vergouwe Y, Habbema JD. Internal validation of predictive models: efficiency of some

procedures for logistic regression analysis. *J Clin Epidemiol* 2001; **54**: 774–81.
34. Morrissy AS, Morin RD, Delaney A, *et al.* Next-generation tag sequencing for cancer gene expression profiling. *Genome Res* 2009; **19**: 1825–35.
35. Morikawa T, Kato J, Yamaji Y, Wada R, Mitsushima T, Shiratori Y. A comparison of the immunochemical fecal occult blood test and total colonoscopy in the asymptomatic population. *Gastroenterology* 2005; **129**: 422–8.
36. de Wijkerslooth TR, Stoop EM, Bossuyt PM, *et al.* Immunochemical fecal occult blood testing is equally sensitive for proximal and distal advanced neoplasia. *Am J Gastroenterol* 2012; **107**: 1570–8.
37. Young TL, Matsuda T, Cepko CL. The noncoding RNA taurine upregulated gene 1 is required for differentiation of the murine retina. *Curr Biol* 2005; **15**: 501–12.
38. Han Y, Liu Y, Gui Y, Cai Z. Long intergenic non-coding RNA TUG1 is overexpressed in urothelial carcinoma of the bladder. *J Surg Oncol* 2013; **107**: 555–9.
39. Chambon M, Orsetti B, Berthe ML, *et al.* Prognostic significance of TRIM24/TIF-1α gene expression in breast cancer. *Am J Pathol* 2011; **178**: 1461–9.
40. Cui Z, Cao W, Li J, Song X, Mao L, Chen W. TRIM24 overexpression is common in locally advanced head and neck squamous cell carcinoma and correlates with aggressive malignant phenotypes. *PLoS ONE* 2013; **8**: e63887.
41. Li H, Sun L, Tang Z, *et al.* Overexpression of TRIM24 correlates with tumor progression in non-small cell lung cancer. *PLoS ONE* 2012; **7**: e37657.
42. Gandini D, De Angeli C, Aguiari G, *et al.* Preferential expression of the transcription coactivator HTIF1alpha gene in acute myeloid leukemia and MDS-related AML. *Leukemia* 2002; **16**: 886–93.
43. Herszényi L, Sipos F, Galamb O, *et al.* Matrix metalloproteinase-9 expression in normal mucosa-adenoma-dysplasia-adenocarcinoma sequence of the colon. *Pathol Oncol Res* 2008; **14**: 31–7.
44. Kang JC, Chen JS, Lee CH, Chang JJ, Shieh YS. Intratumoral macrophage counts correlate with tumor progression in colorectal cancer. *J Surg Oncol* 2010; **102**: 242–8.
45. Zaman K, Driscoll R, Werffeli P, *et al.* Monitoring multiple angiogenesis-related molecules in the blood of cancer patients shows a correlation between VEGF-A and MMP-9 levels before treatment and divergent changes after surgical vs. conservative therapy nt. *J Cancer* 2006; **118**: 755–64.
46. Lakatos G, Sipos F, Miheller P, *et al.* The behavior of matrix metalloproteinase-9 in lymphocytic colitis, in collagenous colitis and ulcerative colitis. *Pathol Oncol Res* 2012; **18**: 85–91.
47. Bleeker SE, Moll HA, Steyerberg EW, *et al.* External validation is necessary in prediction research: a clinical example. *J Clin Epidemiol* 2003; **56**: 826–32.
48. Gebhardt C, Németh J, Angel P, Hess J. S100A8 and S100A9 in inflammation and cancer. *Biochem Pharmacol* 2006; **72**: 1622–31.