scientific reports

Check for updates

OPEN

A two-tiered targeted proteomics approach to identify pre-diagnostic biomarkers of colorectal cancer risk

Sophia Harlid[1]✉, Justin Harbs[1], Robin Myte[1], Carl Brunius[2,3], Marc J. Gunter[4], Richard Palmqvist[5], Xijia Liu[6] & Bethany Van Guelpen[1,7]

Colorectal cancer prognosis is dependent on stage, and measures to improve early detection are urgently needed. Using prospectively collected plasma samples from the population-based Northern Sweden Health and Disease Study, we evaluated protein biomarkers in relation to colorectal cancer risk. Applying a two-tiered approach, we analyzed 160 proteins in matched sequential samples from 58 incident colorectal cancer case–control pairs. Twenty-one proteins selected from both this discovery phase and the literature were then analyzed in a validation set of 450 case–control pairs. Odds ratios were estimated by conditional logistic regression. LASSO regression and ROC analysis were used for multi-marker analyses. In the main validation analysis, no proteins retained statistical significance. However, exploratory subgroup analyses showed associations between FGF-21 and colon cancer risk (multivariable OR per 1 SD: 1.23 95% CI 1.03–1.47) as well as between PPY and rectal cancer risk (multivariable OR per 1 SD: 1.47 95% CI 1.12–1.92). Adding protein markers to basic risk predictive models increased performance modestly. Our results highlight the challenge of developing biomarkers that are effective in the asymptomatic, prediagnostic window of opportunity for early detection of colorectal cancer. Distinguishing between cancer subtypes may improve prediction accuracy. However, single biomarkers or small panels may not be sufficient for effective precision screening.

Colorectal cancer is one of the most common causes of cancer-related deaths in the world, and mortality is highly dependent on stage at diagnosis[1]. Early detection and treatment of colorectal cancer could therefore lead to decreased mortality rates world-wide. Many countries have implemented, or are in the process of implementing, age-based general screening programs, typically using colonoscopy, or fecal tests followed by endoscopy[2,3]. In addition to early detection, screening has major preventive and therapeutic effects, through the removal of precancerous and early malignant lesions. Improvements to general screening programs could, therefore, translate into substantial reductions in colorectal cancer incidence and mortality.

Currently, efforts to supplement colorectal cancer screening programs with blood tests for sub-clinical disease presence are ongoing[4,5], including the FDA-approved test for Septin 9 DNA methylation[6]. Such tests use diagnostic biomarkers, i.e. biomarkers of disease, as an acceptable, minimally invasive and resource-effective means of selecting screening participants for colonoscopy. Another approach to refining general screening programs is through population-based risk stratification, in the hope of identifying higher-risk groups for earlier and/or more frequent screening. Risk algorithms using personal data such as age, family history of cancer, genetic risk variants and lifestyle-related factors show some promise for improving risk prediction[7,8], but have not achieved sufficient accuracy to majorly impact general screening programs[9].

Blood-based biomarkers for risk prediction represent an enticing avenue in the ongoing effort toward effective risk stratification and personalized colorectal cancer screening. However, given the focus on diagnostic biomarkers, the bulk of research in the field has used samples collected in a clinical setting[5], from patients with existing colorectal cancer. Markers detected may, therefore, not be applicable in the pre-carcinogenic and

[1]Department of Radiation Sciences, Oncology, Umeå University, 901 87 Umeå, Sweden. [2]Department of Biology and Biological Engineering, Chalmers University of Technology, Gothenburg, Sweden. [3]Chalmers Mass Spectrometry Infrastructure, Chalmers University of Technology, Gothenburg, Sweden. [4]Section of Nutrition and Metabolism, International Agency for Research On Cancer, World Health Organization, Lyon, France. [5]Department of Medical Biosciences, Pathology, Umeå University, Umeå, Sweden. [6]Department of Mathematics and Mathematical Statistics, Umeå University, Umeå, Sweden. [7]Wallenberg Centre for Molecular Medicine, Umeå University, Umeå, Sweden. ✉email: sophia.harlid@umu.se

early-carcinogenic phases particularly relevant for risk stratification. Some studies have used a screening setting, in which individuals with colorectal adenomas and polyps, not just carcinomas, are compared to those individuals free of neoplasms[10,11], which may be a more promising approach. One venue that has shown some success is inflammation, a hallmark of cancer and an established etiological driver in colorectal cancer. A recent case-cohort study based in Japan, using a panel of 62 inflammatory biomarkers, identified a number of chemokines putatively related to subsequent colorectal cancer risk[12]. However, these have not been replicated in an independent sample.

In the present study, we used a two-tiered approach to colorectal cancer biomarker discovery and validation. Our primary aim was to identify novel biomarkers using large panels of inflammatory and cancer-related markers in a unique set of colorectal cancer cases and controls with time-matched, repeated, prediagnostic samples[13], and to validate these in an independent sample from the same population. Our second aim was to validate findings reported in previous studies by incorporating them into a custom panel that also included our top findings.

Materials and methods

Study population. Participants were from the Västerbotten intervention programme (VIP)[14] and the northern Swedish Monitoring of Trends and Determinants in Cardiovascular Disease (MONICA)[15] cohort. The VIP is an ongoing preventive program (initiated in 1985) for cardiovascular disease and type-2 diabetes. All residents in Västerbotten County are invited for a primary health screening at 40, 50 and 60 years of age and at this time asked to fill out an extensive questionnaire covering lifestyle, diet and health as well as donate a blood sample for research purposes. The North Sweden MONICA project is part of the WHO MONICA and consists of cross-sectional questionnaire surveys and blood sample collections conducted in 1986, 1990, 1994, 1999, 2004, 2009 and 2014 (with a new collection planned for 2021). MONICA participants are randomly selected from the inhabitants of Västerbotten and Norrbotten in Northern Sweden.

Blood samples for both VIP and MONICA are collected in EDTA and Heparin tubes and aliquots of plasma, buffy coat and erythrocytes are frozen within 1 h of collection. Samples are stored at − 80 °C at the Northern Sweden Biobank (Biobanken Norr) in Umeå, Sweden. All samples are collected in the morning and participants of both cohorts are asked to fast for at least 8 h prior to sampling. If for some reason the participant has not fasted, or fasted for a shorter time-period, this information is noted in the accompanying sample file. Both VIP and MONICA are part of the Northern Sweden Health and Disease Study (NSHDS)[16].

The study was approved by the regional ethical review board at Umeå University, Umeå, Sweden (Ref number: 2015/172-32 and 2015/391-32M). All study subjects provided written informed consent at recruitment, and the study was conducted in accordance with the Declaration of Helsinki.

Study design. The discovery set included only VIP participants, and all cases were selected based on the following strict criteria: A primary colorectal cancer diagnosis within 5 years after the most recent sampling (excluding the last 3 months before diagnosis), at least two available blood samples in the biobank collected at least 10 years apart, and no other primary cancer diagnosis, except non-melanoma skin cancer, at the final date of follow-up (Dec 31st 2014). Controls were matched to cases based on age (± 12 months), sex and sampling dates (+/− 12 months for both sampling occasions). Controls had to be cancer free for at least 5 years after the colorectal cancer diagnosis of their index case or at the end of follow up, whichever came first. Only samples collected after at least 8 h of fasting were included, and no samples had been thawed prior to aliquoting for analysis. The original discovery sample set included 69 matched case–control pairs, all with time-matched repeated samples, and has been previously described[13]. Nine individuals failed in the proteomics quality control and were excluded together with their matched cases or controls. Thus, the final study population in the discovery phase consisted of 120 participants, 60 cases and 60 controls, all with time-matched repeated, pre-diagnostic samples. Later DNA methylation array analyses in another study using the same participants[17] revealed identity mismatch between repeated samples of two individuals (one case and one control). Subsequent error analysis determined that the identity mismatch had occurred at the biobank, prior to sample shipment. Although the validation phase was already underway at that point, we reran the statistical analyses on the discovery set, excluding these case sets, for comparison.

The validation sample set was selected from a larger nested colorectal cancer case–control study, comprising 1010 case–control pairs (matched on age at and year of sampling, sex, study cohort, freeze thaw cycles and fasting status) from the VIP and MONICA cohorts, described in detail elsewhere[18].

We selected 1000 samples from this nested case–control study, prioritizing sample plates to minimize the number of cases with peridiagnostic blood samples available in the related clinical colorectal cancer cohort, U-CAN[19]. This was done to reserve patients with both prediagnostic NSHDS and peridiagnostic U-CAN samples for possible future validation of potential novel diagnostic biomarkers. The final validation sample sent for analysis included 461 matched colorectal cancer case–control pairs, of which 39 had time-matched repeated samples.

Outcome variables and covariates. Outcome data in both the discovery and validation phases were obtained by linkage to Swedish national registers (the Swedish Cancer Register, The Swedish Cause of Death Register and the Swedish Register of the Total Population). Colorectal cancer cases were identified using ICD-10 codes (18.0 and 18.2–18.9 for colon cancer, 19.9 and 20.9 for rectal cancer) and verified by a gastrointestinal pathologist (Richard Palmqvist). Data on disease stage and anatomical location were retrieved from the Swedish Colorectal Cancer Register and, in cases of missing data, from patient records. Molecular tumor data (*BRAF* V600E and *KRAS* mutations) were generated in house, as previously described[18]. Additional covariates were considered based on previously established associations with colorectal cancer risk and data availability. They included: age at sampling, sex, body mass index (BMI, based on height and weight measured by a health professional), self-reported smoking status, education, alcohol intake and physical activity.

Sample collection and laboratory analysis. Blood samples were collected in EDTA tubes, centrifuged, aliquoted and frozen within 1 h of sampling. In the discovery phase of the study, plasma samples were analyzed for two panels of biomarkers using predesigned Proseek Multiplex immunoassays (Inflammation and Oncology II, Olink Proteomics, Uppsala, Sweden), as previously described[13]. We selected the Inflammation panel based on the role of inflammation in colorectal cancer etiology and progression, leading to the hypothesis that inflammatory biomarkers may have merit as potential risk predictive and/or early diagnostic biomarkers of colorectal cancer. The Oncology II panel was added to capture additional markers associated with cancer and cancer development but not included in the Inflammation panel. All proteins included in the commercially available Olink panels, including Oncology II and Inflammation, are pre-selected by Olink, and we thus had no influence on panel content. It is also worth noting that since our initial selection (performed in 2016), the number of available panels has increased substantially. The multiplex panels rely on Proximity Extension Assay (PEA) technology, which maintains specificity despite high multiplexing levels[20].

All sample processing and quality control was performed by Olink Proteomics. Data were delivered as Normalized Protein eXpression (NPX) values on a log2 scale and pre-processed as described in previous publications[13,21]. Information about limits of detection (LOD) can be found online (http://www.olink.com). The full list of markers included in both panels, together with the percentage of samples that fell below the LOD, can be found in Supplementary Table S1.

For the validation phase of the study, we selected 21 biomarkers to be analyzed on a custom-made panel, designed for us by Olink proteomics (Supplementary Table S2). For the panel, we prioritized proteins from our own discovery phase that had an FDR of less than 0.25. Fifteen proteins passed this threshold and of these; three failed quality control in the multiplex design (TNFSF13, S100A4 and CEA). TNFSF13 and S100A4 were hence excluded from the validation phase. However, as CEA is a known, and used, colorectal cancer tumor marker we deemed it to be of strong interest and therefore decided to include it, but run the analysis as a single assay. Protein biomarkers selected from the literature filled the additional eight available spots on the custom panel, however there were several restrictions limiting our selection of markers from the literature. First, the markers had to be available on one of the pre-existing Olink panels, second, they had to pass quality control for multiplexing and third their concentration in plasma had to be in the right range as to not need dilution. The final selection of literature markers included eight proteins that fit these criteria and were selected from two previous publications[22,23].

Assay performance for the custom-panel was assessed by Olink proteomics during the design stage and continuously during sample runs. Briefly, three internal controls were added to each multiplex plate and two internal controls to each singleplex plate in order to monitor the quality of assay performance and the quality of individual samples. The standard deviation from the internal controls was evaluated for each sample plate; if the deviation was above 0.2 NPX values, the plate was rerun. Individual samples were evaluated by determining the deviation from the median value of the controls, if this exceeded 0.3 NPX values the sample was excluded from the analysis. Controls in triplicate were used for calculations of Inter- and Intra-Assay Coefficients of Variability (CV). In order to reduce the impact of batch effects to a minimum, cases and their matched controls were always placed on the same plate. After quality control and pre-processing, data were delivered as NPX values. All included proteins are presented in Supplementary Table S2.

Data pre-processing. In the discovery dataset all individuals contributed two samples each. For those individuals lacking covariate data at one of the sampling occasions, information from the other occasion was used to complete the dataset. In the validation dataset, missing data were observed for BMI (N = 9), smoking status (N = 21), level of education (N = 13), alcohol consumption (N = 116) and physical activity (N = 116). As most participants in the validation phase lacked repeated measures, we used multivariate imputation by chained equation (mice R-package), to replace the missing data, under the assumption that data were missing at random[18]. For continuous variables (BMI) we used predictive mean matching, whereas for multi-categorical (N > 2) variables (smoking status and level of education) Bayesian polytomous regression was used. Predictors originally included age, sex, sampling year, smoking status, BMI, education level, alcohol consumption and cohort (MONICA or VIP). To assess the robustness of our imputation method, we repeated the imputation step multiple times with different random starting samples and compared the results to analyses in which observations with missing data were excluded.

To reduce the influence of extreme outliers, relative protein concentrations were winsorized to the 1st and 99th percentile. To simplify comparisons between protein associations, the log2 NPX values were scaled to mean 0 and SD 1 prior to data analysis. Proteins with > 50% of values below LOD were removed from the dataset (Supplementary Table S1). For the remaining proteins, values below LOD were replaced by protein specific LODs (discovery data set) or individually reported and included in the analyses (validation data set).

Statistical analysis. For each protein, conditional logistic regression was used to calculate an odds ratio (OR) of colorectal cancer risk. Multivariable models included smoking status, BMI, and education level as covariates. In the discovery phase, availability of repeated measurements for all participants allowed us to perform the analyses at two different time points. In the validation phase baseline values were used for all downstream analyses, unless otherwise specified. Pre-defined subgroup analyses were performed based on tumor location (colon/rectum), stage (I–II/III–IV), time to diagnosis (> 5 years/≤ 5 years after sampling) and molecular subtypes based on *KRAS* and *BRAF* mutations.

On the validation phase dataset, we applied logistic regression with Least absolute shrinkage and selection operator, LASSO, to identify a subset of informative features from 21 proteins. The penalty parameter was chosen using tenfold cross-validation with respect to predictive performance (area under the curve, AUC).

Cross-validation was repeated 100 times with different training/test set partitions to accommodate for randomness in the partitioning. The lasso model was adjusted for age, sex, BMI, smoking status, and level of education which were therefore excluded from penalization. In order to evaluate our extended model, including covariates and the additional proteins selected by LASSO, we compared it to a model containing only the risk factors included as covariates in the logistic regression (age, sex, BMI, smoking status, level of education) and performed a likelihood ratio test. We also compared our models by plotting receiver operator curves (ROCs) and calculating the AUCs.

For the individuals in the validation set that had repeated measures we also conducted a longitudinal analysis using linear mixed models for the top proteins associated with colorectal cancer. We included subject ID and case–control pair ID as random effects, and BMI, smoking status, level of education, and time to diagnosis as fixed effects parameters. An interaction term between case–control status and time until diagnosis was included to investigate changes in protein levels over time between cases and controls. Time until diagnosis was defined as time of sampling until time at diagnosis, for cases, and as time from sampling until time of diagnosis for their matched case, for controls. Models were fitted using the lme4 R-package and the degrees-of-freedom and p-values were estimated by Satterthwaite approximation.

All p-values in the discovery phase were adjusted for multiple comparisons using the false discovery rate (FDR), q-value framework[24]. In the discovery phase of this study, q-values below 0.25 were considered for selection to the validation panel. In the validation phase, p values < 0.05 were considered statistically significant.

All computations were conducted in R v.3.6.0 (R Foundation for Statistical Computing, Vienna, Austria).

Results

Participant characteristics. Participant characteristics are shown in Table 1. Overall, there were no clear differences, in terms of baseline characteristics, between cases and controls. In the validation set, cases tended to have a higher BMI and lower education compared with controls. Neither smoking nor alcohol intake varied between cases and controls in either of the datasets. Age at baseline was approximately 10 years higher in the validation set (59.8 years) compared to the discovery set (49.9 years), due to the requirement of repeated samples (collected at 10 year intervals) in the discovery set. Also due to the study design, the mean age at colorectal cancer diagnosis was lower in the discovery set (60.6 years compared to 66.6 years in the validation set) (Table 1).

Tumor characteristics. Among the cases, there were some important differences between the discovery set and the validation set. First, in the discovery dataset almost half (46.6%) of the cases were rectal cancers compared to 34.7% in the validation set. In addition, a higher proportion of the discovery set cases were stage IV (22.4%) compared to 14.2% in the validation set (Supplementary Table S3).

Biomarker identification and selection. After preprocessing and exclusions of proteins not passing quality control, 160 proteins remained. In our initial analyses, 15 proteins were associated with colorectal cancer at an FDR cutoff of 0.25 (Fig. 1, Table 2). Of these, we selected 13 and excluded two proteins (TNFSF13 and S100A4). TNFSF13 was excluded in favor of MIC A/B based on literature review, and S100A4 was excluded as it failed quality control in the designing of the custom multiplex panel. We also included eight additional proteins previously reported to be associated with colorectal cancer risk in the prediagnostic setting[10,22,23,25].

In the discovery analysis, PPY showed the strongest overall association with colorectal cancer with an OR of 1.79 (95% CI 1.29–2.48) per 1 SD. In total, seven proteins had higher levels in cases compared to controls and eight had lower levels (Fig. 1). Aside from PPY, top proteins included CEA (OR: 1.65, 95% CI 1.14–2.40) and 5'NT (OR: 1.62, 95% CI 1.17–2.24). ESM1 was the protein with the strongest inverse association with colorectal cancer (OR: 0.59 95% CI 0.43–0.80), closely followed by HGF (OR: 0.60 95% CI 0.42–0.85).

In analyses rerun after excluding two case–control pairs with identity mismatch 12 out of the 13 originally selected proteins remained among the top hits, and an additional six potentially significant proteins were identified (Supplementary Fig. S1, Supplementary Table S4). Although the stage of the validation analyses prevented their inclusion in the custom panel, we chose to present the results for potential future replication.

Custom panel quality control. For the custom multiplex panel the average intra-assay CV was 7% and the average inter-assay CV was 12%. Three proteins had an intra-assay CV between 5–10%, no proteins had intra CV values above 10%. For the inter-assay CV, 17 proteins had an inter-assay CV value between 10 and 20% and one protein had inter-assay CV values between 20 and 30%. The singleplex CEA assay had an intra-assay CV of 11% and inter-assay CV of 20%.

The 21 proteins included in the custom panel (Supplementary Table S2) were analyzed in a validation set consisting of 1000 samples from 461 case–control pairs (with 39 pairs including repeated samples from both cases and controls). Of these, 12 samples failed quality control and were excluded, together with their matched cases or controls, from downstream analyses (Supplementary Fig. S2). The final validation set thus included 450 complete case–control pairs, of which 38 pairs had repeated measurements.

Main analyses in the validation set. No proteins were associated with colorectal cancer risk at significance level $p < 0.05$. (Table 3). Out of the 13 proteins selected from the dicovery set, seven retained the same direction of association although with varying degrees of attenuation. The strongest positive association was for FGF-21 (OR: 1.14, 95% CI 0.99–1.30), followed by CEA (OR: 1.10, 95% CI 0.95–1.28) and PPY (OR: 1.08, 95% CI 0.93–1.26), all showing the same direction of association as previously identified. Only minimal differences in point estimates were observed when comparing models with imputed covariate data to models based only on collected data (data not shown).

Variable	Discovery set (n = 116[a])			Validation set (n = 900[a])		
	Cases (n = 58)	Controls (n = 58)	p^b	Cases (n = 450)	Controls (n = 450)	p^b
Cohort n (%)			1.0			1.0
VIP	58 (100)	58 (100)		417 (92.7)	417 (92.7)	
MONICA	0 (0)	0 (0)		33 (7.3)	33 (7.3)	
Repeated samples n (%)						
Yes	58 (100)	58 (100)	1.0	38 (50)	38 (50)	
Age median (range)						
Baseline (years)	49.9 (39.5–52.5)	49.9 (39.7–52.4)	1.0	59.8 (29.7–74.5)	59.8 (29.8–74.9)	1.0
Follow up (years)	59.9 (49.9–60.5)	59.9 (49.9–60.6)	1.0	59.9 (39.9–73.0)[c]	60.0 (40.0–73.9)[c]	1.0
Diagnosis (years)	60.6 (50.2–65.1)	N/A		66.6 (40.4–89.6)	N/A	
Sex n (%)						
Men	32 (55.2)	32 (55.2)	1.0	240 (53.3)	240 (53.3)	1.0
Anthropometrics median (range)						
Height (cm)	172.0 (157.0–195.0)	170.5 (157.0–191.0)	0.9	171.0 (150.0–201.0)	170.8 (150.0–194.0)	0.6
Weight (kg)	75.50 (51.0–123.0)	75.85 (52.0–128.0)	0.9	77.00 (48.0–143.0)	75.20 (45.0–132.0)	0.1
Body mass index median (range)						
BMI (kg/m^2)	25.32 (19.6- 37.8)	24.43 (18.8–41.3)	0.9	26.03 (17.78–43.04)	25.72 (17.15–44.62)	0.1
Body mass index groups n (%)						
< 18.5	0 (0)	0 (0)	0.3	2 (0.5)	9 (2.0)	0.2
18.5–24.9	27 (46.6)	32 (55.2)		168 (37.3)	170 (37.8)	
25–29.9	26 (44.8)	18 (31.0)		195 (43.3)	197 (43.8)	
≥ 30	5 (8.6)	8 (13.8)		80 (17.8)	70 (15.6)	
Missing[d]	0 (0)	0 (0)		5 (1.1)	4 (1.0)	
Smoking n (%)						
Current smoker	20 (34.5)	17 (29.3)	0.7	111 (24.7)	96 (21.4)	0.2
Former smoker	13 (22.4)	17 (29.3)		155 (34.4)	143 (31.8)	
Never smoker	25 (43.1)	23 (39.7)		174 (38.7)	201 (44.6)	
Missing[d]	0 (0.0)	1 (1.7)		10 (2.2)	10 (2.2)	
Education n (%)						
Elementary	26 (44.8)	26 (44.8)	0.8	332 (73.8)	307 (68.2)	0.1
Secondary	16 (27.6)	18 (31.1)		56 (12.4)	62 (13.8)	
Post-secondary	16 (27.6)	13 (22.4)		55 (12.2)	76 (16.9)	
Missing[d]	0 (0)	1 (1.7)		7 (1.6)	5 (1.1)	
Alcohol intake median (range)						
Grams/day	2.82 (0–17.9)	2.74 (0–19.2)	0.7	2.20 (0–30.6)	2.22 (0–27.9)	0.9
Missing[d] n (%)	7 (12.07)	7 (12.07)		59 (13.11)	57 (12.67)	
Physical activity n (%)						
Inactive	8 (13.79)	9 (15.52)	0.3	83 18.44	75 (16.67)	0.9
Moderately inactive	19 (32.76)	11 (18.97)		135 (30.00)	144 (32.00)	
Moderately active	18 (31.04)	18 (31.03)		111 (24.67)	109 (24.22)	
Active	8 (13.79)	14 (24.14)		63 (14.00)	64 (14.22)	
Missing[d]	5 (8.62)	6 (10.34)		58 (12.89)	58 (12.89)	

Table 1. Participant characteristics. [a]Includes complete case–control sets. [b]Paired Wilcoxon signed rank test for continuous variables, Chi-Square tests for categorical variables. [c]Includes 38 case control pairs with follow up samples. [d]Missing category not included in the statistical comparisons.

Subgroup analyses in the validation set. Stratified subgroup analyses were performed to take differences in tumor location, stage, time to diagnosis and molecular tumor subtypes into account. Results from these analyses are presented in Fig. 2 and Supplementary Table S5. In the subgroup analyses based on tumor site, FGF-21 (OR: 1.23, 95% CI 1.03–1.47) and 5'NT (OR: 0.86, 95% CI 0.79–0.99), were associated with colon but not rectal cancer, whereas PPY (OR: 1.47, 95% CI 1.12–1.92) was associated with rectal but not colon cancer. FGF-21 also retained statistical significance in stage I–II colorectal cancer and in cases with samples collected more than 5 years before diagnosis. No proteins showed statistical significance in the analyses including only stage III–IV colorectal cancer. In analyses stratified for molecular subtypes, MIC A/B was associated with a lower risk of *KRAS*-mutated colorectal cancer (OR: 0.66, 95% CI 0.47–0.93), no proteins were associated with the risk of *BRAF*-mutated or *KRAS-BRAF*-wild type colorectal cancer.

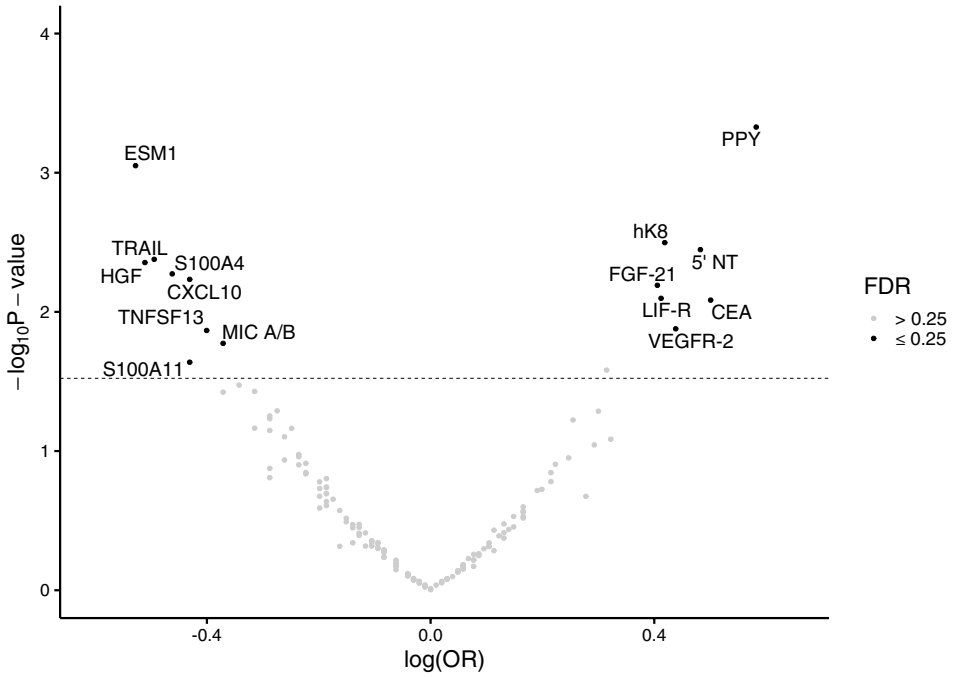

Figure 1. Volcano plot depicting results from the discovery phase of the study. Odds ratios were calculated using conditional logistic regression based on individuals' baseline- and repeated values, separated in time. The model was adjusted for smoking status, BMI, and level of education. The dotted line represents FDR = 0.25.

Protein	OR (95% CI) Crude	OR (95% CI) Adjusted*	FDR*	Included in Validation phase
PPY	1.60 (1.20-2.12)	1.79 (1.29-2.48)	0.07	Yes
CEA	1.51 (1.10-2.07)	1.65 (1.14-2.40)	0.12	Yes
5'NT	1.56 (1.15-2.12)	1.62 (1.17-2.24)	0.12	Yes
VEGFR2	1.51 (1.10-2.07)	1.55 (1.10-2.20)	0.17	Yes
hK8	1.45 (1.13-1.86)	1.52 (1.15-2.00)	0.12	Yes
LIF-R	1.30 (1.00-1.68)	1.51 (1.11-2.05)	0.12	Yes
FGF-21	1.45 (1.11-1.89)	1.50 (1.12-2.02)	0.12	Yes
MIC A/B	0.75 (0.56-1.00)	0.69 (0.50-0.93)	0.19	Yes
TNFSF13	0.70 (0.52-0.95)	0.67 (0.49-0.92)	0.17	No
S100A4	0.71 (0.53-0.94)	0.65 (0.45-0.94)	0.12	No
S100A11	0.69 (0.48-0.97)	0.65 (0.45-0.94)	0.25	Yes
CXCL10	0.71 (0.55-0.93)	0.65 (0.48-0.88)	0.12	Yes
TRAIL	0.66 (0.49-0.89)	0.61 (0.43-0.86)	0.12	Yes
HGF	0.64 (0.47-0.88)	0.60 (0.42-0.85)	0.12	Yes
ESM1	0.62 (0.46-0.83)	0.59 (0.43-0.80)	0.07	Yes

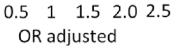

Table 2. Top 15 proteins identified in the discovery phase. *Adjusted for BMI, smoking and education and conditioned on matching criteria (age, sex and sampling date).

Lasso regression in the validation set. In order to determine if any specific combination of proteins could predict colorectal cancer risk better than individual proteins, we used a Lasso logistic regression model. Models containing only the risk factors included as covariates in the conditional logistic regression analyses (age, sex, BMI, smoking status and level of education) were compared to our Lasso-generated protein models using ROC curves, both for the main analysis (Supplementary Fig. S3) and for the subgroup analyses (Fig. 3). In the main model, FGF-21 was the only protein with sufficient predictive ability to be included, increasing the AUC

Protein	OR (95% CI) Crude	OR (95% CI) Adjusted*	FDR*
FGF-21	1.17 (1.02-1.33)	1.14 (0.99-1.30)	0.91
CEA	1.15 (1.00-1.32)	1.10 (0.95-1.28)	0.91
PPY	1.11 (0.96-1.29)	1.08 (0.93-1.26)	0.91
MSLN	1.07 (0.94-1.23)	1.08 (0.92-1.26)	0.91
S100A11	1.11 (0.96-1.28)	1.07 (0.92-1.24)	0.91
IL8	1.07 (0.94-1.22)	1.06 (0.93-1.21)	0.91
CXCL10	1.06 (0.93-1.22)	1.06 (0.92-1.22)	0.91
HGF	1.11 (0.97-1.26)	1.05 (0.90-1.21)	0.91
LIF-R	1.04 (0.90-1.20)	1.04 (0.90-1.20)	0.91
hk8	1.02 (0.89-1.17)	1.04 (0.90-1.20)	0.91
ESM1	0.97 (0.85-1.12)	1.03 (0.89-1.20)	0.91
ENG	1.00 (0.87-1.15)	1.01 (0.87-1.18)	0.94
vWF	1.02 (0.88-1.20)	1.01 (0.87-1.19)	0.94
VEGFR2	1.05 (0.91-1.20)	1.01 (0.88-1.16)	0.94
IL6	1.06 (0.92-1.22)	1.00 (0.87-1.16)	0.97
EGFR	0.97 (0.84-1.13)	0.98 (0.84-1.14)	0.94
MIC A/B	0.98 (0.86-1.12)	0.98 (0.86-1.11)	0.94
TRAIL	1.00 (0.88-1.15)	0.95 (0.83-1.10)	0.91
IL6R	0.95 (0.82-1.10)	0.95 (0.82-1.10)	0.91
DKK1	0.95 (0.83-1.08)	0.94 (0.83-1.08)	0.91
5'NT	0.93 (0.81-1.06)	0.90 (0.79-1.04)	0.91

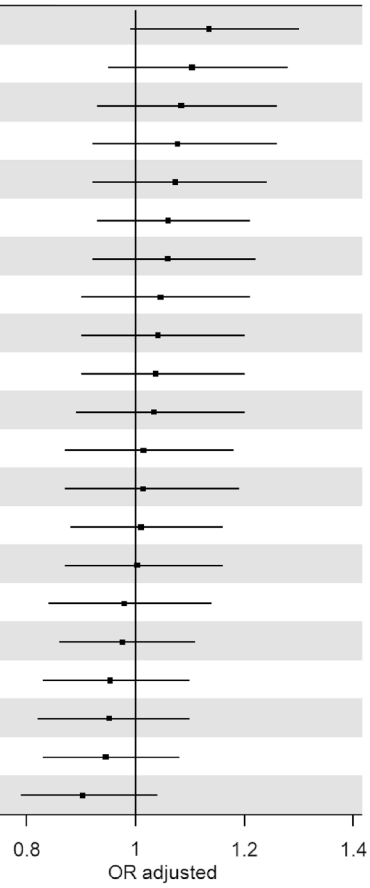

Table 3. Proteins included in the validation phase main analysis. *Adjusted for BMI, smoking and education and conditioned on matching criteria (age, sex and sampling date).

slightly from 0.55 (95% CI 0.51–0.59) in the risk-factor-only model to 0.57 (95% CI 0.50–0.53). The likelihood ratio test indicated that the increase was borderline significant ($p = 0.046$). In the subgroup analyses (Fig. 3), the highest discriminative ability was seen for colon cancer, for which the AUC increased from 0.56 (95% CI 0.52–0.61) in the risk-factor-only model to 0.63 (95% CI 0.59–0.68), $p = 0.003$ and for individuals with stage I–II colorectal cancer, for which the AUC increased from 0.58 (95% CI 0.53–0.64) to 0.62 (95% CI 0.57–0.67), $p = 0.002$.

Linear mixed models. No statistically significant differences were found for either FGF-21 or PPY on the 38 case sets with repeated samples (Supplementary Fig. S4).

Discussion

Utilizing a two-tiered approach with prospectively collected samples, we aimed to identify protein biomarkers related to colorectal cancer susceptibility or early diagnosis. In the discovery phase, we identified 15 proteins with significantly altered levels in colorectal cancer cases compared to controls, of which 13 were then selected for further analysis. None of the selected proteins retained statistical significance in the main validation analysis, but two proteins of particular interest, FGF-21 and PPY, were identified when stratifying analyses by tumor location, stage and time to diagnosis.

Multiple studies have aimed to identify protein biomarkers that could be utilized for precision screening for early detection of colorectal cancer[5,12,23]. However, the majority of these have used samples from either a clinical or a screening setting. These study designs, although successful in identifying biomarkers or biomarker combinations with relatively good predictive ability (AUC > 0.7), may not be useful for identifying prospective patients years, or even months, before diagnosis. Clearly, identifying markers that are measurable before diagnosis, to be used for refined risk stratification or even early diagnosis has proven difficult. Very few studies have utilized prospectively collected samples, and results have proven hard to replicate[12]. Because most previous studies did not include prospective samples, it is difficult to compare the results from our study to those of others. The importance of conducting additional studies in a prediagnostic, asymptomatic setting, and not only in cancer patients or screening participants therefore needs to be emphasized.

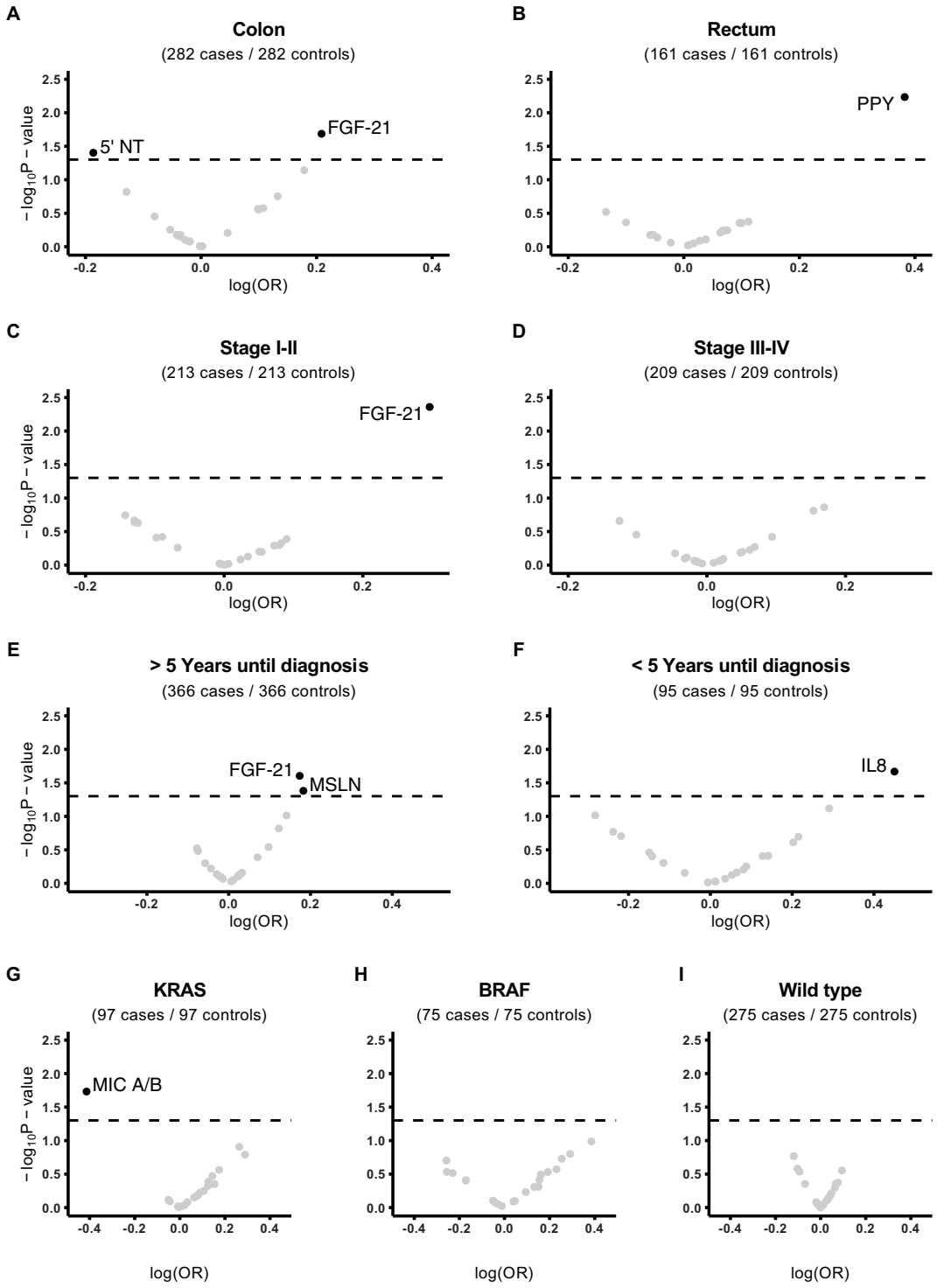

Figure 2. Volcano plots for stratified analyses comparing Colon and Rectum (**A**,**B**), Stage I–II and Stage III–IV (**C**,**D**), proximity to diagnosis (**E**,**F**) and molecular subtypes (**G**–**I**). All models were adjusted for smoking status, BMI, and level of education. The dotted line marks a *p* value cutoff of 0.05.

In our study, which focused mainly on inflammatory targets, FGF-21 was the protein that performed best throughout all stages. However, it did not reach significance in the full dataset combining both colon and rectal cancer cases. Interestingly, when stratifying by anatomical tumor sub-site, it became evident that the association between FGF-21 and colon cancer was driving this association. FGF-21 remained associated with colon cancer after adjusting for BMI, education and smoking. Notably, FGF-21 was associated with early, but not late, stages of colorectal cancer. We previously identified FGF-21 as associated with colorectal cancer risk in a study

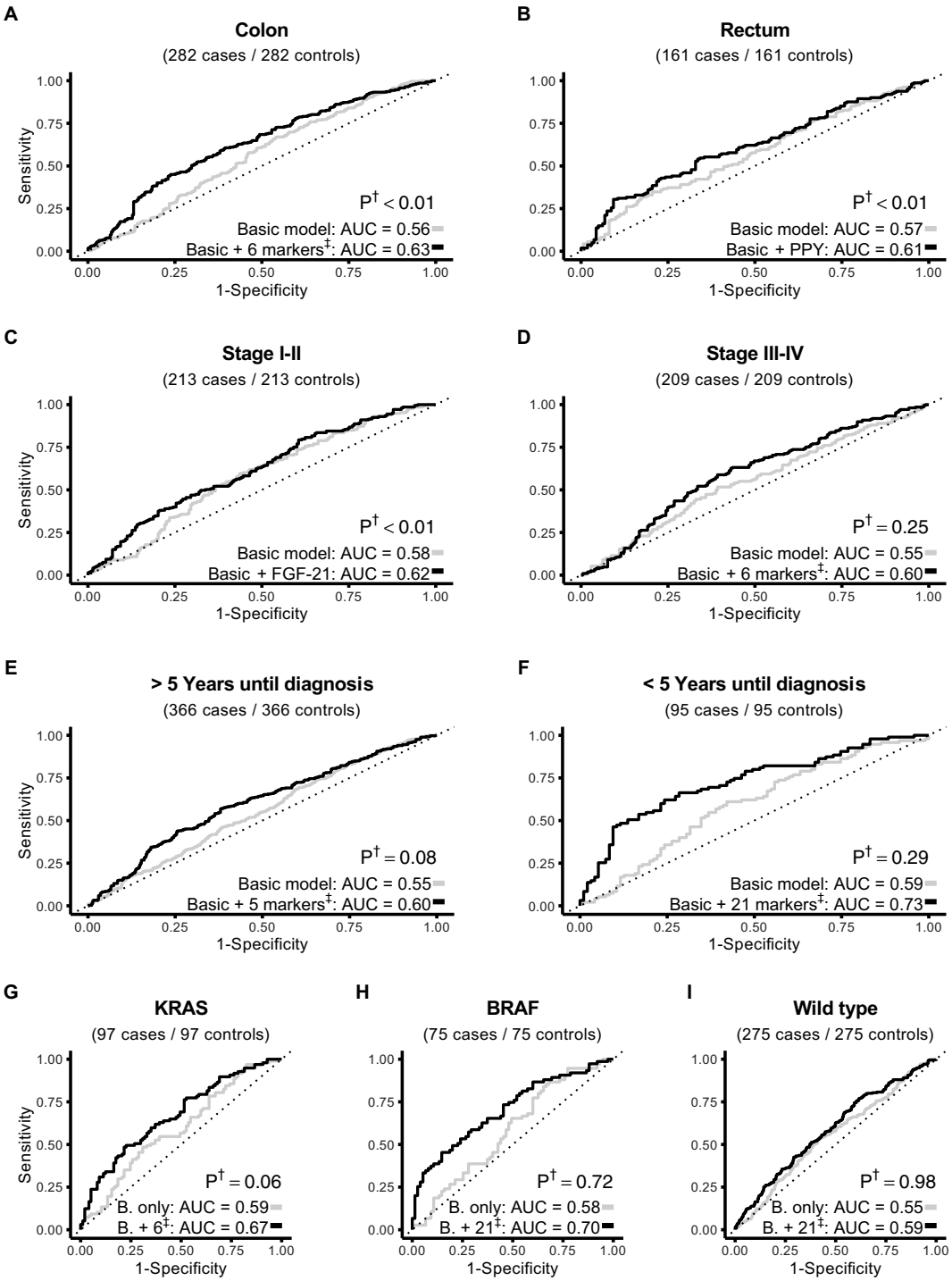

Figure 3. ROC curves for stratified analyses comparing Colon and Rectum (**A**,**B**), Stage I–II and Stage III–IV (**C**,**D**), years until diagnosis (**E**,**F**) and molecular subtypes (**G**–**I**). All models were adjusted for smoking status, BMI, level of education, age and sex. For complete protein panels see Supplementary Table S5. †Based on the loglikelihood ratio test between the two models. ‡Selected proteins are listed in Supplementary Table S6.

investigating potential protein biomarkers of metabolic syndrome[13], where we found it to be strongly associated with BMI and therefore of potential interest as a colorectal cancer screening biomarker. FGF-21 has since been shown to associate with colorectal cancer in at least one other study[26], where it was reported to be positively associated with both early and late stage colorectal cancers.

In the rectal cancer group, pancreatic prohormone (PPY), also known as pancreatic polypeptide[27], was the most prominent finding. This protein is mainly produced in the pancreas and secreted postprandially where it slows down the digestive process[28]. PPY is a marker of some pancreatic tumors including pancreatic

polypeptide-secreting tumor of the distal pancreas (PPoma) and Multiple endocrine neoplasia type 1 (MEN1) both of which are characterized by high serum levels of PPY[29]. Few studies have specifically examined levels of PPY in colorectal cancer patients compared to healthy controls. One small-scale study from Poland, including 60 colorectal cancer patients and 30 healthy controls, found elevated levels of PPY in colon cancer patients compared to rectal cancer patients and cancer free controls[30].

Despite being potentially predictive and possibly related to colorectal cancer etiology, neither FGF-21 nor PPY would be useful as standalone biomarkers of colon or rectal cancer. We therefore explored the possibility of combining different protein markers in order to identify a panel with better discriminative capabilities, an approach that has been proven successful in similar study designs[10,11,31,32]. However, the addition of the top markers selected by Lasso regression to conventional risk factors increased the predictive ability only modestly from an AUC of 0.55–0.57, still far from clinically useful. Subgroup analyses resulted in larger improvements in discriminative ability, although all AUCs remained below 0.7. It should be noted that the predictive performance of our basic model was quite low compared to previous studies[10,11,23,32]. Possible explanations may include lack of family history data in our study, and the low ages at sampling and at colorectal cancer diagnosis (mean of < 70 years in both data sets) due to the recruitment protocol with ongoing sampling at defined ages (40, 50 and 60 years).

Aside from the markers identified in our discovery phase the custom panel also included eight proteins selected from previous promising findings and available for Olink multiplexing, namely Cohen et al.[22] and Rho et al.[23]. However, none were significantly associated with colorectal cancer in our study population. In Cohen et al. the authors developed a general blood based test for cancer detection, which was tested in a set of cancer patients and controls. Although the test does not target colorectal cancer specifically, potential colorectal cancer biomarkers were evaluated. Of the additional five markers from Cohen et al. selected for our study, none were included in the final CancerSEEK panel. The only colorectal cancer marker in the CancerSEEK panel was CEA, which also reached statistical significance in our discovery dataset and was thus already selected for our custom array. CEA is used in clinical practice to follow colorectal cancer patients[33].

In contrast to Cohen et al., and Rho et al.[23] conducted a study using prediagnostic blood samples and four biomarkers of particular interest for early detection of colorectal cancer (BAG4, IL6R/ST, VWF and EGFR). One of our initial aims was to try to replicate all four of these markers, but we were limited by the proteins available on the Olink panels and therefore could not include BAG4. The lack of association with colorectal cancer for the other three markers in our study may be due to the longer time between sample donation and cancer diagnosis in our population compared to Rho et al. in which it was less than 3 years. Another reason may be that our study lacked power to detect small effect sizes, even if there were differences between prospective cases and controls more than 3 years before case diagnosis.

Our study has several strengths including the prospective approach, the two independent sample sets and the large size of the validation set. However, several limitations need to be addressed. First, the composition of the cases in the two datasets differed with respect to clinical characteristics. Rectal cancer and stage IV cancer were more common in the discovery set compared to the validation set. These discrepancies primarily reflect the small samples size in the discovery set, which was a conscious trade-off to allow the stringent selection of cases and controls with time-matched, repeated prediagnostic samples. The discovery data set is, therefore, not entirely representative of the site and stage proportions in the general population in Västerbotten, which is captured in the larger validation dataset. Since the result for FGF-21 was retained for colon cancer and stage I-II colorectal cancer in the validation dataset, the clinical differences between the two datasets do not seem to have affected the main findings. The comparatively low proportion of stage III-IV cancers in the validation set could probably explain, at least partly, why CEA did not reach significance despite being one of few commonly used biomarkers for colorectal cancer monitoring[34]. In addition, we lacked information on family history, which is known to be one of the best predictors of future colorectal cancer risk[7] and likely would have improved clinical risk prediction models. Furthermore, we chose to rely mainly on statistical cut-offs for marker selection when proceeding from the discovery to the validation stage on the study. However, the small size of our discovery dataset may have hindered the identification of true colorectal cancer biomarkers. In hindsight an approach combining statistical cut-offs with biological relevance might have resulted in more markers being validated. Finally, for sample size reasons, Lasso regression to select protein markers was performed without dividing our dataset into a training and testing set and results are therefore in need of further validation.

Our findings highlight the challenge of identifying cancer biomarkers that can be used in the pre-diagnostic window of opportunity for early detection. For risk stratification, with the vision of achieving effective precision screening, single biomarkers or small marker panels may not be sufficient. Instead, we would advocate also attempting to identify patterns, based on panels of biomarkers, which might achieve more precise risk stratification. For translation to commercially viable blood tests, the composition and size of biomarker panels will require consideration of cost effectiveness, such as numbers needed to prevent one colorectal cancer case or colorectal cancer death.

In conclusion, we identified two markers (FGF-21 and PPY) that were associated with colon and rectal cancer respectively, suggesting a potential of biomarkers to discriminate between different subtypes of colorectal cancer. Approaches for future studies of colorectal cancer risk prediction and early detection biomarkers should probably focus on large collections of prospectively collected samples and deeply phenotyped colorectal cancer cases, and perhaps use machine learning on high-dimensional biomarker platforms to identify biomarker risk patterns.

Data availability

The datasets generated and/or analyzed during the current study are considered personal data, which prohibits us from storing them in a public depository. However, all data are archived at the Biobank Research Unit at Umeå University, and access for secondary use can be granted conditional upon meeting Swedish requirements for human research.

Received: 9 November 2020; Accepted: 10 February 2021
Published online: 04 March 2021

References

1. Brenner, H., Kloor, M. & Pox, C. Colorectal cancer. *Lancet* **383**, 1490–1502. https://doi.org/10.1016/S0140-6736(13)61649-9 (2014).
2. Schreuders, E. H. *et al.* Colorectal cancer screening: a global overview of existing programmes. *Gut* **64**, 1637. https://doi.org/10.1136/gutjnl-2014-309086 (2015).
3. Pellat, A., Deyra, J., Coriat, R. & Chaussade, S. Results of the national organised colorectal cancer screening program with FIT in Paris. *Sci. Rep.* **8**, 4162. https://doi.org/10.1038/s41598-018-22481-9 (2018).
4. Issa, I. A. & Noureddine, M. Colorectal cancer screening: an updated review of the available options. *World J. Gastroenterol.* **23**, 5086–5096. https://doi.org/10.3748/wjg.v23.i28.5086 (2017).
5. Bhardwaj, M., Gies, A., Werner, S., Schrotz-King, P. & Brenner, H. Blood-based protein signatures for early detection of colorectal cancer: a systematic review. *Clin. Transl. Gastroenterol.* https://doi.org/10.1038/ctg.2017.53 (2017).
6. Lin, J. S. *et al.* Screening for colorectal cancer: updated evidence report and systematic review for the US preventive services task force. *JAMA* **315**, 2576–2594. https://doi.org/10.1001/jama.2016.3332 (2016).
7. Smith, T. *et al.* Comparison of prognostic models to predict the occurrence of colorectal cancer in asymptomatic individuals: a systematic literature review and external validation in the EPIC and UK Biobank prospective cohort studies. *Gut* https://doi.org/10.1136/gutjnl-2017-315730 (2018).
8. Schmit, S. L. *et al.* novel common genetic susceptibility loci for colorectal cancer. *J. Natl. Cancer Inst.* **111**, 146–157. https://doi.org/10.1093/jnci/djy099 (2019).
9. Robertson, D. J. & Ladabaum, U. Opportunities and challenges in moving from current guidelines to personalized colorectal cancer screening. *Gastroenterology* **156**, 904–917. https://doi.org/10.1053/j.gastro.2018.12.012 (2019).
10. Chen, H., Zucknick, M., Werner, S., Knebel, P. & Brenner, H. Head-to-head comparison and evaluation of 92 plasma protein biomarkers for early detection of colorectal cancer in a true screening setting. *Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res.* **21**, 3318–3326. https://doi.org/10.1158/1078-0432.CCR-14-3051 (2015).
11. Bhardwaj, M. *et al.* Multiplex screening of 275 plasma protein biomarkers to identify a signature for early detection of colorectal cancer. *Mol. Oncol.* **14**, 8–21. https://doi.org/10.1002/1878-0261.12591 (2020).
12. Song, M. *et al.* Circulating inflammatory markers and colorectal cancer risk: a prospective case-cohort study in Japan. *Int. J. Cancer* **143**, 2767–2776. https://doi.org/10.1002/ijc.31821 (2018).
13. Harlid, S., Myte, R. & Van Guelpen, B. The metabolic syndrome, inflammation, and colorectal cancer risk: an evaluation of large panels of plasma protein markers using repeated, prediagnostic samples. *Mediat. Inflamm* **2017**, 4803156. https://doi.org/10.1155/2017/4803156 (2017).
14. Norberg, M., Wall, S., Boman, K. & Weinehall, L. The vasterbotten intervention programme: background, design and implications. *Glob. Health Action* https://doi.org/10.3402/gha.v3i0.4643 (2010).
15. Benckert, M., Lilja, M., Soderberg, S. & Eliasson, M. Improved metabolic health among the obese in six population surveys 1986 to 2009: the Northern Sweden MONICA study. *BMC Obes.* **2**, 7. https://doi.org/10.1186/s40608-015-0040-x (2015).
16. Hallmans, G. *et al.* Cardiovascular disease and diabetes in the Northern Sweden health and disease study cohort—evaluation of risk factors and their interactions. *Scand. J. Public Health Suppl.* **61**, 18–24. https://doi.org/10.1080/14034950310001432 (2003).
17. Myte, R., Sundkvist, A., Guelpen, B. & Harlid, S. Circulating levels of inflammatory markers and DNA methylation, an analysis of repeated samples from a population based cohort. *Epigenetics* **14**, 649–659. https://doi.org/10.1080/15592294.2019.1603962 (2019).
18. Myte, R. *et al.* A longitudinal study of prediagnostic metabolic biomarkers and the risk of molecular subtypes of colorectal cancer. *Sci. Rep.* **10**, 5336. https://doi.org/10.1038/s41598-020-62129-1 (2020).
19. Glimelius, B. *et al.* U-CAN: a prospective longitudinal collection of biomaterials and clinical information from adult cancer patients in Sweden. *Acta Oncol.* **57**, 1–8. https://doi.org/10.1080/0284186X.2017.1337926 (2017).
20. Assarsson, E. *et al.* Homogenous 96-plex PEA immunoassay exhibiting high sensitivity, specificity, and excellent scalability. *PLoS ONE* **9**, e95192. https://doi.org/10.1371/journal.pone.0095192 (2014).
21. Sundkvist, A. *et al.* Targeted plasma proteomics identifies a novel, robust association between cornulin and Swedish moist snuff. *Sci. Rep.* **8**, 2320. https://doi.org/10.1038/s41598-018-20794-3 (2018).
22. Cohen, J. D. *et al.* Detection and localization of surgically resectable cancers with a multi-analyte blood test. *Science* **359**, 926–930. https://doi.org/10.1126/science.aar3247 (2018).
23. Rho, J.-H. *et al.* Protein and glycomic plasma markers for early detection of adenoma and colon cancer. *Gut* **67**, 473. https://doi.org/10.1136/gutjnl-2016-312794 (2018).
24. Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. U.S.A.* **100**, 9440–9445. https://doi.org/10.1073/pnas.1530509100 (2003).
25. Li, S. *et al.* Plasma mesothelin as a novel diagnostic and prognostic biomarker in colorectal cancer. *J. Cancer* **8**, 1355–1361. https://doi.org/10.7150/jca.18014 (2017).
26. Qian, J., Tikk, K., Weigl, K., Balavarca, Y. & Brenner, H. Fibroblast growth factor 21 as a circulating biomarker at various stages of colorectal carcinogenesis. *Br. J. Cancer* **119**, 1374–1382. https://doi.org/10.1038/s41416-018-0280-x (2018).
27. Lonovics, J., Devitt, P., Watson, L. C., Rayford, P. L. & Thompson, J. C. Pancreatic polypeptide. A review. *Arch. Surg. (Chicago, Ill.:1960)* **116**, 1256–1264. https://doi.org/10.1001/archsurg.1981.01380220010002 (1981).
28. Śliwińska-Mossoń, M., Marek, G. & Milnerowicz, H. The role of pancreatic polypeptide in pancreatic diseases. *Adv. Clin. Exp. Med.* **26**, 1447–1456. https://doi.org/10.17219/acem/65094 (2017).
29. Kamilaris, C. D. C. & Stratakis, C. A. Multiple endocrine neoplasia type 1 (MEN1): an update and the significance of early genetic and clinical diagnosis. *Front. Endocrinol. (Lausanne)* **10**, 339. https://doi.org/10.3389/fendo.2019.00339 (2019).
30. Zygulska, A. L., Furgala, A., Krzemieniecki, K., Kaszuba-ZwoiNska, J. & Thor, P. Enterohormonal disturbances in colorectal cancer patients. *Neoplasma* **64**, 421–429. https://doi.org/10.4149/neo_2017_313 (2017).
31. Chen, H. *et al.* Development and validation of a panel of five proteins as blood biomarkers for early detection of colorectal cancer. *Clin. Epidemiol.* **9**, 517–526. https://doi.org/10.2147/CLEP.S144171 (2017).
32. Werner, S. *et al.* Evaluation of a 5-marker blood test for colorectal cancer early detection in a colorectal cancer screening setting. *Clin. Cancer Res.* **22**, 1725–1733. https://doi.org/10.1158/1078-0432.ccr-15-1268 (2016).
33. Hall, C. *et al.* A review of the role of carcinoembryonic antigen in clinical practice. *Ann. Coloproctol.* **35**, 294–305. https://doi.org/10.3393/ac.2019.11.13 (2019).
34. Locker, G. Y. *et al.* ASCO 2006 update of recommendations for the use of tumor markers in gastrointestinal cancer. *J. Clin. Oncol.* **24**, 5313–5327. https://doi.org/10.1200/JCO.2006.08.2644 (2006).
35. Harlid, S. *et al.* Abstract 2353: a two-tiered targeted proteomics approach to identify biomarkers of colorectal cancer risk. *Cancer Res.* **80**, 2353–2353. https://doi.org/10.1158/1538-7445.Am2020-2353 (2020).

Acknowledgements
We thank the Biobank Research Unit at Umeå University, Västerbotten Intervention Programme, the Northern Sweden MONICA study and the County Council of Västerbotten for providing data and samples and acknowledge the contribution from Biobank Sweden, supported by the Swedish Research Council (VR 2017-00650). Special thanks to Robert Johansson, Åsa Ågren, and their colleagues at the Biobank Research Unit, Umeå University for helpful assistance. We also want to thank the staff at Biobanken Norr, Västerbotten County Council, as well as Åsa Stenberg, Anna Löfgren-Burström, Carl Zingmark and Roger Stenling at the Department of Medical Biosciences, Pathology, Umeå University, for invaluable assistance with the tumour tissue retrieval and analyses. The study abstract was submitted to, and selected for a poster presentation, at the American Association for Cancer Research (AACR) annual meeting (rescheduled to June 22–24, 2020, abstract #2353)[35].

Author contributions
S.H. and B.V.G. designed the study, R.M. analyzed the discovery dataset. S.H. and B.V.G. selected proteins for the validation phase. J.H. analyzed the validation dataset. J.H. and X.L. performed the Lasso regression analyses and constructed the ROC curves. S.H., B.V.G., C.B., M.J.G., R.P. and X.L. contributed to interpretation of the data. All authors critically reviewed and approved the final version of the manuscript.

Funding
Open access funding provided by Umea University. This study was funded by the Swedish Research Council (Grant number: 2017-01737), WCMM, Knut and Alice Wallenberg Foundation, the Swedish Cancer Society (Grant numbers: 2017/581 and 2014/780), the Cancer Research Foundation in Northern Sweden (multiple grants), the Lion's Cancer Research Foundation (multiple grants), the Faculty of Medicine at Umeå University and a regional agreement between Umeå University and Region Västerbotten (so-called ALF).

Competing interests
Where authors are identified as personnel of the International Agency for Research on Cancer/World Health Organization, the authors alone are responsible for the views expressed in this article and they do not necessarily represent the decisions, policy or views of the International Agency for Research on Cancer/World Health Organization.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-021-83968-6.

Correspondence and requests for materials should be addressed to S.H.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.