Gastroenterology 2023;165:121–132

Proteomic Profiling of Colorectal Adenomas Identifies a Predictive Risk Signature for Development of Metachronous Advanced Colorectal Neoplasia

Jacob Mathias Bech,[1,2] Thilde Terkelsen,[3] Annette Snejbjerg Bartels,[1] Fabian Coscia,[4,5] Sophia Doll,[6] Siqi Zhao,[7] Zhaojun Zhang,[7] Nils Brünner,[1] Jan Lindebjerg,[8] Gunvor Iben Madsen,[9] Xiangdong Fang,[7] Matthias Mann,[5] and José Manuel Afonso Moreira[1]

[1]*Department of Drug Design and Pharmacology, University of Copenhagen, Copenhagen, Denmark;* [2]*Sino-Danish Center for Education and Research, Aarhus University, Aarhus, Denmark;* [3]*Center for Health Data Science, University of Copenhagen, Copenhagen, Denmark;* [4]*Spatial Proteomics Group, Max-Delbrück-Centrum for Molecular Medicine in the Helmholtz Association, Berlin, Germany;* [5]*Proteomics Program, Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark;* [6]*OmicEra Diagnostics GmbH, Planegg, Germany;* [7]*Beijing Institute of Genomics, China National Center for Bioinformation, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Beijing, China;* [8]*Department of Pathology, Lillebaelt Hospital, Vejle Hospital, Vejle, Denmark; and* [9]*Department of Pathology, Odense University Hospital, Odense, Denmark*

BACKGROUND & AIMS: Colonic adenomatous polyps, or adenomas, are frequent precancerous lesions and the origin of most cases of colorectal adenocarcinoma. However, we know from epidemiologic studies that although most colorectal cancers (CRCs) originate from adenomas, only a small fraction of adenomas (3%–5%) ever progress to cancer. At present, there are no molecular markers to guide follow-up surveillance programs. **METHODS:** We profiled, by mass spectrometry–based proteomics combined with machine learning analysis, a selected cohort of formalin-fixed, paraffin-embedded high-grade (HG) adenomas with long clinical follow-up, collected as part of the Danish national screening program. We grouped subjects in the cohort according to their subsequent history of findings: a nonmetachronous advanced neoplasia group (G0), with no new HG adenomas or CRCs up to 10 years after polypectomy, and a metachronous advanced neoplasia group (G1) where individuals developed a new HG adenoma or CRC within 5 years of diagnosis. **RESULTS:** We generated a proteome dataset from 98 selected HG adenoma samples, including 20 technical replicates, of which 45 samples belonged to the nonmetachronous advanced neoplasia group and 53 to the metachronous advanced neoplasia group. The clear distinction of these 2 groups seen in a uniform manifold approximation and projection plot indicated that the information contained within the abundance of the ∼5000 proteins was sufficient to predict the future occurrence of HG adenomas or development of CRC. **CONCLUSIONS:** We performed an in-depth analysis of quantitative proteomic data from 98 resected adenoma samples using various novel algorithms and statistical packages and found that their proteome can predict development of metachronous advanced lesions and progression several years in advance.

Keywords: Colorectal Cancer; Colonic Adenomatous Polyps; Biomarkers; Progression.

Colorectal cancer (CRC) is the third most prevalent type of cancer worldwide and the second most common cause of cancer death.[1] Most CRCs evolve from precursor lesions called colonic adenomatous polyps or adenomas, through accumulation of genetic alterations, in a gradual process termed the adenoma-to-carcinoma sequence (ACS) that typically spans over several years.[2] Adenomas are frequent in the adult population, occurring in about 20%–40% of individuals older than 50 years of age.[3] However, only a few of these lesions (3%–5%) ever eventually become a cancer.[4–6] Progression from adenomas to colon cancer is a protracted, multistep process, involving the accumulation of driver mutations. This gradual process provides an opportunity for intervention through the early detection of lesions; the 3 most accepted tests are the fecal occult blood test and endoscopic examination of the bowel by sigmoidoscopy or colonoscopy. National screening programs are well established in many countries in the Western world.[7] Early lesions identified during endoscopic examination are resected, and patients are placed on long-term repeated surveillance programs. The frequency and type of clinical follow-up depends on different parameters, including the number of adenomas removed as well as their size and histopathologic presentation.[8] Irrespective of the number and size of the adenomas, high-grade (HG) dysplasia qualifies for the earliest possible surveillance colonoscopy. Highly dysplastic adenomas are acknowledged as the last benign stage of CRC precursors in the ACS. This classification is important because health care providers use it to risk-stratify individuals and adjust surveillance plans accordingly. Although evidence based, this very crude

Abbreviations used in this paper: **ACS, adenoma-to-carcinoma sequence; AUC, area under the curve; CRC, colorectal cancer; ENR, elastic-net regression; FFPE, formalin-fixed, paraffin-embedded; GO, Gene Ontology; HG, high grade; MS, mass spectrometry; ROC, receiver operating characteristic curve; UMAP, uniform manifold approximation and projection.**

Most current article

0016-5085
https://doi.org/10.1053/j.gastro.2023.03.208

WHAT YOU NEED TO KNOW

BACKGROUND AND CONTEXT

Although most colorectal cancers originate from adenomas, only about 3%–5% of adenomas will ever progress to cancer. Unfortunately, reliable biomarkers that can predict adenoma progression are still lacking, and as a result, surveillance colonoscopy at yearly intervals is the current practice.

NEW FINDINGS

We developed a proteomics signature that could predict the development of metachronous advanced colorectal neoplasia based on mass spectrometry data from the analysis of formalin-fixed, paraffin-embedded tissue samples directly from histopathology glass slides. The predictive ability of this signature was dependent on multiple proteins.

LIMITATIONS

Further studies with larger cohorts are required to validate our predictive proteomic classification.

CLINICAL RESEARCH RELEVANCE

Development of new therapies that target the subset of proteins identified in the progression classifier may provide an effective preventative strategy for colorectal cancer.

BASIC RESEARCH RELEVANCE

Characterization of proteins identified in the progression classifier may further our molecular understanding of colorectal carcinogenesis.

stratification also represents a gap in our understanding of adenomas on a molecular level. At present, there are no clinically useful molecular markers to guide follow-up surveillance programs. The result is a substantial degree of overtreatment and, with the implementation of national screening programs, a significant burden on health care systems.

Accumulated data from screening and surveillance programs has shown that detection and removal of colonic polyps is effective in overall risk reduction for colon cancer.[9,10] However, this reduction in cancer-specific incidence and mortality comes with a significant risk cost of redundant diagnosis. Although the benefits of screening outweigh the potential harms, screening leads to the detection of many lesions with little or no risk of progression to invasive cancers. An increased understanding of these common precursor lesions would enable health systems to move resources from individuals at low risk to those at high risk, ultimately ameliorating the problem of redundant diagnosis associated with screening while reducing the incidence and mortality of the disease. We sought to develop a molecular risk classification system based on quantitative proteomics data obtained from HG lesions. We recently reported a streamlined and reproducible workflow for deep proteomics profiling of larger tissue cohorts.[11] This workflow addressed multiple common issues associated with quantitative proteomic profiling of patient-derived formalin-fixed, paraffin-embedded (FFPE) tissue samples, and we showed it to be applicable to the study of adenoma tissues.

We report here the in-depth analysis of quantitative data from 98 resected adenoma samples to identify risk markers for disease progression. Samples were from a non-metachronous advanced neoplasia group (G0), with no new HG adenomas or CRC at least up to 10 years later, and a metachronous advanced neoplasia group (G1) where individuals developed a new HG adenoma or CRC within 5 years. We found a proteomic signature that classified risk of metachronous disease. Implementation of such a classification could result in a reduction of the surveillance burden on those patients with adenomas with low risk of progression while enabling targeted surveillance programs and preventive interventions for those patients at high risk of metachronous advanced neoplasia.

Methods

Experimental Design

The study had a retrospective design and included unlinked anonymized samples retrieved from Patobank, the Danish pathology data bank. We carried out a search for patients who had had endoscopic removal of an HG adenoma and either had a diagnosis within 5 years after the initial diagnosis or no findings for at least 10 years. Patients were grouped into 2 groups: a nonmetachronous advanced neoplasia group (group 0, $n = 45$) and a metachronous advanced neoplasia group (group 1, $n = 53$). Patients were excluded if they had intestinal polyposis syndromes or a prior history of colon cancer. Individuals from the nonmetachronous advanced neoplasia group presented with neither CRC nor new HG adenomas for $\geq$10 years from the time of adenoma resection. Adenomas characterized in this study were primarily examined at the pathology departments in the period 2002–2012 at 2 Danish hospitals (Vejle Hospital and Odense University Hospital). All adenomas included in this study were reevaluated by a trained pathologist, and the classification as HG dysplasia was confirmed.

To rule out the possibility that samples in G1 had acquired molecular alterations associated with early carcinogenesis in the colon, whereas G0 had not, we sequenced a subset of adenoma samples at a depth of 15–30×. This subset included 7 samples from the G0 group and 25 samples from the G1 group. Because the aim of this analysis was to gauge the genetic makeup of adenomas across the 2 groups at the chromosomal level rather than to characterize the somatic mutation landscape of premalignant lesions, we did not sequence matching normal tissue or blood samples. A detailed sample collection and preparation protocol for FFPE tissues, as well as data analysis, is provided in Supplementary Methods. The study protocol was approved by the National Committee on Health Research Ethics (j.nr. 2112779) and granted exemption from obtaining informed consent (as per section 10, subsection 1, of the Committee Act).

Sample preparation and liquid chromatography–mass spectrometry analysis. A detailed sample collection and preparation protocol for FFPE tissues is provided in the Supplementary Methods. The mass spectrometry (MS) proteomics data were deposited to the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org/

cgi/GetDataset) via the PRIDE (PRoteomics IDEntifications database) partner repository with the dataset identifier PXD017269 without restriction.

Bioinformatic analysis of proteomics data from colon adenomas. A detailed description of bioinformatics tools and settings used in the analytical pipeline applied to the analysis of our proteomics data from colon adenomas can be found in Supplementary Methods. Briefly, the proteomics set encompassed 6256 protein groups. We employed the R package DEP[12] to explore the nature of missing values in the dataset. Because of the type of missing values in our dataset, we used the sample minimum method (sampMin) of substitution shown to perform better than other, more complex, methods of imputation and to work well for datasets where values are missing not at random.[13] We performed dimensionality reduction and plotting with uniform manifold approximation and projection (UMAP)[14] to visualize clustering of adenoma samples based on protein abundance patterns. Differential abundance analysis was performed using the limma package.[15] Elastic-net regression (ENR) was performed with the R packages glmnet[16] and caret.[17] To evaluate the co-abundance of proteins from adenomas, we used the R package WGCNA.[18] The input for analysis was the set of ~5000 proteins remaining after filtering, normalization, and missing value imputation. The pipeline of analysis followed the example provided by package developers Langfelder and Horvath at https://horvath.genetics.ucla.edu/html/CoexpressionNetwork/Rpackages/WGCNA/Tutorials/. Protein-protein interaction networks were done using the STRING database[19] of protein-protein interactions. Gene Ontology (GO) term and pathway enrichment analysis were performed using the R package clusterProfiler.[20]

Sensitivity and specificity of the top protein candidates was evaluated using the receiver operating characteristic (ROC).[21] The external data used for the analysis was a set of tandem mass tag (TMT)-labeled nanoscale liquid chromatography–MS/MS proteomics from 18 normal mucosa biopsies, 30 adenomas, and 30 cancers of the colon.[22] Out of the 28 top protein candidates, 23 were contained within the validation dataset, and as such, 5 candidates could not be evaluated using ROC (RAB33B, C1orf226, TTC39A, TSPAN6, GLS2). Pairwise and multiclass ROC analyses were performed with packages nnet and pROC,[23] dividing the dataset into a training set and a test set: two thirds of samples and one third of samples, respectively.

Results

Structural Variations and Copy Number Variations Do Not Distinguish Samples in the Metachronous Advanced Neoplasia Group From Those Without Metachronous Disease

Progression to CRC from adenomatous precursors is generally accepted as a sequential process, with driver mutations underlying tumor progression. Although all samples included in this study had HG dysplasia and, therefore, were histologically at the same stage of progression, it was possible that the adenomatous precursors included in the 2 respective groups, a nonmetachronous advanced neoplasia group (G0) and a metachronous advanced neoplasia group (G1), possessed a different mutation profile and complement of the driver alterations that accompany the switch from benign adenoma to malignant carcinoma. We sequenced a subset of adenoma samples, comprising 7 samples from the G0 group and 25 samples from the G1 group. Overall, we found no significant differences between the 2 groups with respect to structural variations or copy number variations (Figure 1). The mutational burden of samples was not significantly different between the 2 groups (Figure 1*A*), with G0 showing a slightly higher median mutational burden than G1 (Figure 1*A*). The 10 most frequently mutated genes in each group were identical (*ZNF717*, *MUC3A*, *MUC6*, *MUC16*, *MUC4*, *ANKRD36C*, *CDC27*, *CTBP2*, *OR4C5*, and *HLA-DRB1*) (Figure 1*B*), and although there were genes with significantly different mutational status between the 2 groups (Figure 1*C*), these were unlikely to reflect a general difference in malignant potential between the 2 sample groups.

Mass Spectrometry–Based Proteomics

We sought to identify a molecular risk classification system for CRC based on quantitative proteomics data obtained from adenomatous precursors. To be clinically useful, analytical workflows should be based on archival FFPE tissues and be robust and reproducible. We recently developed and reported a proteomics workflow suited for large-cohort proteomic analysis with small sample inputs from multiple tissues.[11] This workflow enabled us to isolate specific small (~30 mm^2) areas of adenomatous tissue from our samples from a single H&E-stained 5-μm section. Additionally, this protocol enabled sample preparation of our entire cohort of adenomas in 1 day, potentially reducing sample handling variability. We generated a proteome dataset from 98 selected HG adenoma samples (with 20 added technical replicates), of which 45 samples belonged to the nonmetachronous advanced neoplasia (G0) and 53 to the metachronous advanced neoplasia group (G1). We showed that this workflow was highly reproducible, with extraordinary proteome consistency within and across tissue sections of the same adenomas.[11] Furthermore, we demonstrated that the archival time of samples did not perturb global protein quantification because we observed high global proteome correlations between storage groups. We used a data-independent acquisition MS workflow to consistently quantify a large part of the proteome of the 98 HG adenoma samples included in the present study in 100-min single-run analyses. The MS proteomics data for our cohort were deposited to the PRIDE repository with the dataset identifier PXD017269. Technical sample replicates were filtered out along with 3 samples (identifiers: 132, 201, 202), which appeared to be outliers based on an initial hierarchical clustering (not shown), leaving 95 samples (44 in G0, 51 in G1) available for subsequent bioinformatics analyses.

Dimensionality Reduction Leads to Clear Clustering of the 2 Sample Groups

The following covariates were evaluated using a dimensionality reduction plot: sample group (G0 vs G1), patient age (30–89 years) binned, sex (male, female), year of

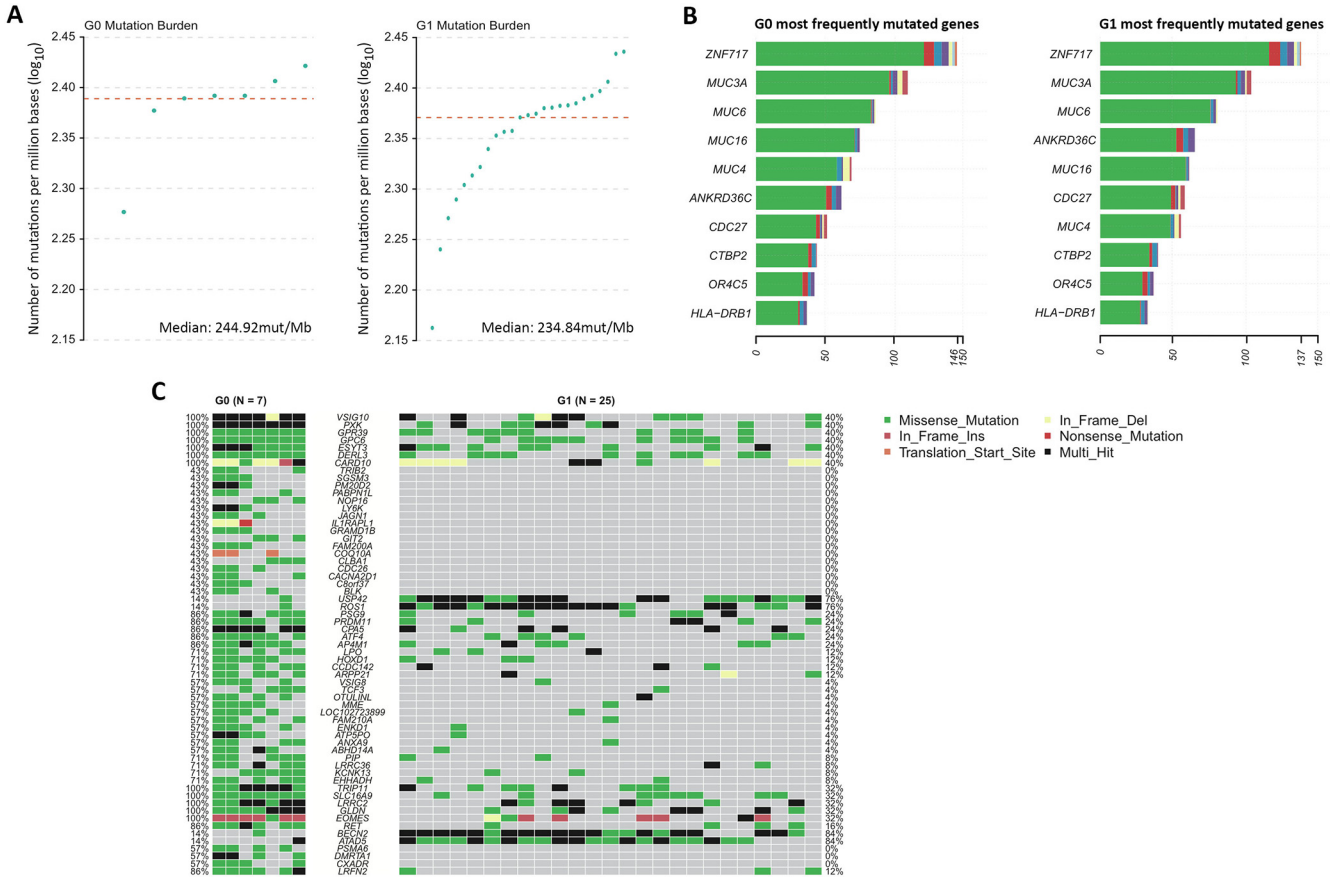

Figure 1. Mutational profiling of adenomas in the G0 and G1 groups by whole-genome sequencing analysis. (*A*) G0 and G1 mutational burden statistical scatter plots for mutational burden of each sample, defined as the number of somatic mutations per megabase of interrogated genomic sequence. (*B*) The top 10 mutated genes. (*Left*) G0 top 10 mutated genes in each sample. (*Right*) G1 top 10 mutated genes in each sample. (*C*) Waterfall map of mutated genes showing significant differences between G0 and G1 samples.

sample collection (2002–2012), localization of adenoma (colon, rectum, distal, proximal), and metachronous advanced neoplasia type (adenocarcinoma, HG adenoma) (Supplementary Table 1). Only sample group (G0, G1) displayed a clear pattern of clustering with UMAP (Figure 2*A*).

Identification of Differentially Abundant Proteins

In total, 460 proteins were identified as differentially abundant between the 2 groups in the limma analysis. Of these, 210 were significant only on the unadjusted *P* value, not after correcting for multiple testing (false discovery rate). This is most likely related to the unbalanced batch design which, when modeled in the design matrix, results in a significant loss of power. All 460 proteins were kept for further analysis. Accuracy of the elastic-net model was 0.9, with a confidence interval of 0.69–0.97, and it encompassed 101 proteins. Overlap of the 460 differentially abundant (DA) proteins with the 101 proteins from ENR resulted in 53 shared proteins. Hierarchical clustering of adenoma samples showed that the small set of 53 consensus proteins was sufficient to partition samples from the non-metachronous neoplasia vs metachronous neoplasia group (Figure 2*B*).

Modules From Weighted Coexpression Network Analysis Correlate With Patient Group

Adenoma sample group (metachronous neoplasia, non-metachronous neoplasia) was significantly correlated with multiple co-abundance modules (9 out of 17 modules). In contrast, none of the other patient covariates displayed noteworthy correlation with protein modules, supporting the initial observations from dimensionality reduction visualization with UMAP (Figure 2*C*). The 9 modules that correlate significantly with the adenoma sample group are the largest modules, and combined they encompass 60% of all proteins identified in our set (∼3000 proteins). We visualized the fraction of proteins that were either differentially abundant, retained in the elastic-net model, or returned by both analyses in a stacked bar plot (Figure 2*D*). This figure highlights that the forestGreen module, which has the most significant inverse correlation with sample group, also had the largest fraction of proteins with known cancer relevance. However, and importantly, this module was also one of the smallest modules out of the 9. Taking module size into consideration, as well as consensus between DA and EN proteins, the most interesting modules appears to be the dustyRed, lavenderPurple, and oliveGreen

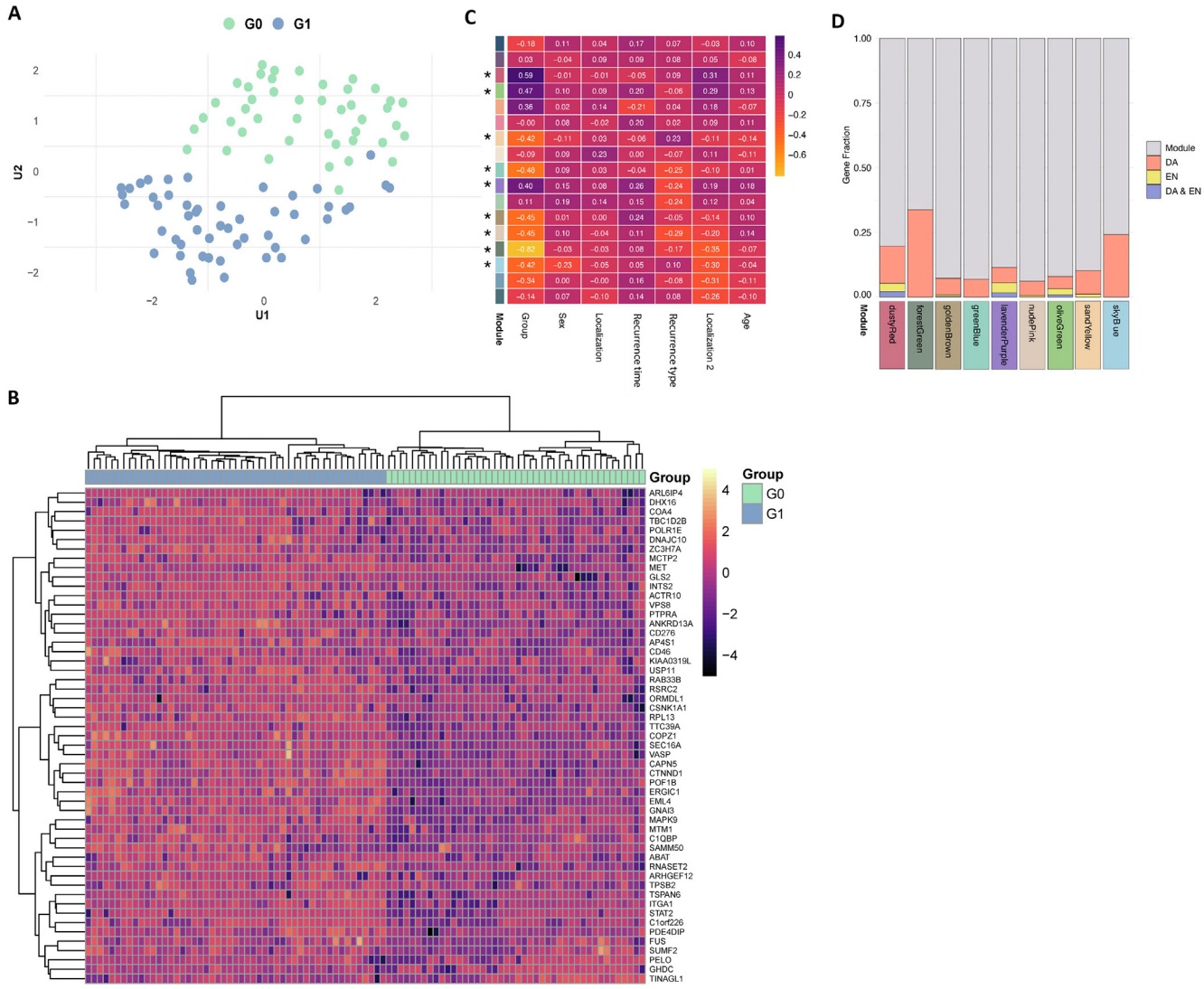

Figure 2. Dimensionality reduction and clustering. (*A*) Dimensionality reduction by UMAP. UMAP of batch-corrected samples, colored for the G0 and G1 groups, respectively. The dimensionality reduction plot shows the segregation of samples from the 2 groups, based on adenoma proteomes. (*B*) Heatmap of 53 differentially abundant proteins. Hierarchical clustering was performed using the Ward algorithm with the set of 53 consensus proteins, present in both the limma and ENR sets. (*C*, *D*) Correlation of clinical variables with eigen proteins from weighted gene coexpression network analysis (WCGNA) coexpression modules. (*C*) The 17 co-abundance modules (*left-most multicolored column*) and their correlations with 7 different clinical variables. Absolute correlation scores of $\geq$0.4 and an adjusted *P* value of $<$.05 (not shown) were considered to be significant correlations. Nine modules correlate (*stars*) with the metachronous advanced neoplasia group, whereas none of the other investigated variables correlated significantly with any of the modules. (*D*) Fraction of proteins identified in the different types of analysis for each significant module from *C*. Each bar encompasses all proteins identified in that module, with modules listed below each bar. Module: proteins only present in the co-abundance module but none of the other types of analysis. DA: proteins in the module that are also identified by limma. EN: proteins in the module that are also identified by elastic net regression. DA and EN: proteins in the module identified by both types of analysis.

modules (all positively correlated with metachronous neoplasia group).

Network Analysis of Coexpressed Modules Shows Complex Multinodal Systems That Are Not Defined by Any Single Protein

Each module displayed an intricate network reflected by the protein-protein interactions extracted from the STRING database (Figure 3). A protein-protein interaction was only included in a module network if the interaction met the minimum cutoff criteria (defined in the Methods section). The 4 largest modules are the dustyRed, oliveGreen, lavenderPurple, and goldenBrown. Three of these are also the only ones to contain proteins that are present in all 3 types of analyses (consensus proteins): 5 in oliveGreen and 8 in each of both lavenderPurple and dustyRed. The dustyRed module (Figure 3*A*, enlarged left panel) consists of 2 subclusters. The lower left cluster contains 2 consensus proteins, SEC16A and COPZ1; the latter is central to this cluster,

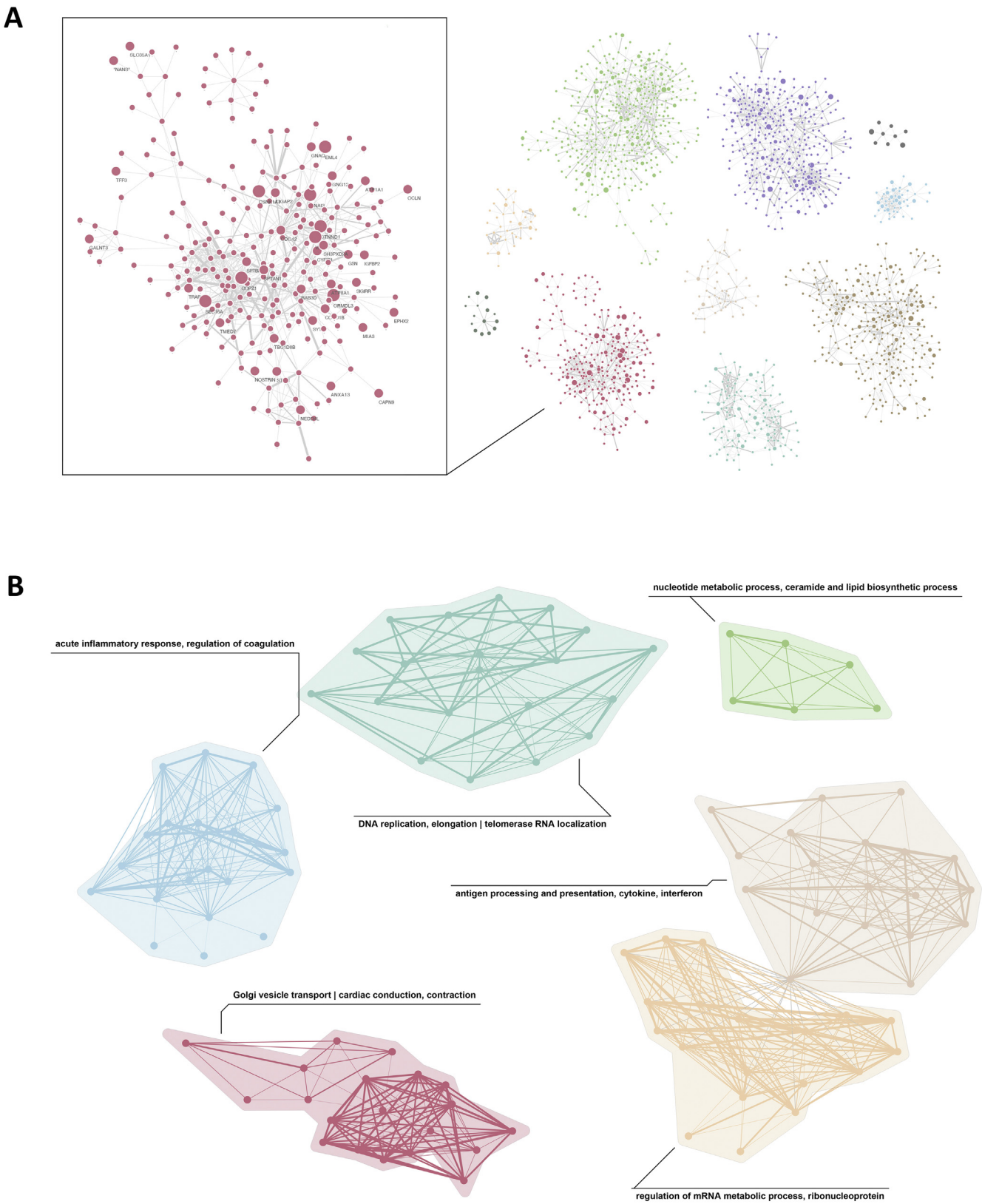

Figure 3. Network representation of coexpressed modules and differentially abundant proteins with Cytoscape. (*A*) Protein-protein interactions of the 9 coexpressed modules identified by weighted gene coexpression network analysis (WGCNA). Only interactions validated in human cells or tissue are shown, displayed by the edges. The smallest nodes represent proteins that are coexpressed, medium-sized nodes represent proteins that are also identified in either DA or ENR, and the large nodes are proteins that are present in all 3. (*B*) GO analysis and pathway enrichment using the Kyoto Encyclopedia of Genes and Genomes. Networks of different GO terms that are significantly enriched in either one of the DA, ENR, or coexpression sets. Titles of the different networks are based on the most common and significant terms. Each node represents a GO term, and edges display overlap of proteins for different GO terms.

as highlighted by the density of interactions, for example, a hub protein. The upper cluster contains the consensus proteins catenin-delta (CTNND1) and casein kinase 1 alpha 1 (CSNK1A1), as well as cell division control protein 42 (CDC42, medium-sized node). In the center of 4 consensus proteins and CDC42 is β-catenin (CTNNB1; small node, not highlighted). The lavenderPurple module displays a similar pattern of 2 large subclusters, with another consensus protein integrin-alpha 1 (ITGA1) as the most central protein in the module. Conversely, the oliveGreen module is more homogenous, with no subclusters. Common for all modules is that many of the consensus proteins are at the edges of their respective networks. Here, it is worth noting the inherent (and unavoidable) biases in databases such as STRING—that is, highly studied proteins, such as transcription factors or oncogenes, will have significantly more annotated interactions and stronger experimental support, often placing them at the center of a network.

Enrichment Analyses Show an Overrepresentation of Proteins Related to Vesicle Transport of the Golgi and Endoplasmic Reticulum as Well as to the Immune System and Inflammation

The 2 most significant GO terms (*P*-adjust = 10^{-19} and 10^{-17}) in the module dustyRed are related to vesicle transport of the Golgi or between the endoplasmic reticulum and Golgi (Figure 3*C*; red module). In the skyBlue module, we observed enrichment for proteins involved in humoral immune response, acute inflammatory response, complement activation, and regulation of blood coagulation (Figure 3*C*; blue module).

Candidate Proteins and Their Receiver Operating Characteristic Curves

We intersected our list of candidates with other proteomics sets on adenomas and excluded potentially confounding samples such as those where an adenoma was adjacent to cancer or where the adenoma had genetic characteristics assumed to influence the proteome (ie, adenomas from familial adenomatous polyposis patients with constitutional variants in *APC*).[24–27] Out of the 54 proteins tested, 28 were supported by at least 1 external study (Table 1). We continued with proteins that we also identified in at least 1 of the other proteomics sets and performed ROC analysis on the list that remained in an external dataset[22] (Figure 4). Twelve proteins fulfilled the cutoff criteria ($\geq$0.75) set for the area under the curve (AUC) analysis (Table 1). One of these, ITGA1, displays a difference in abundance between the 2 groups in our study. Furthermore, its abundance in the nonmetachronous neoplasia group samples appears to be subdivided in 2. In the validation set, 3 proteins display an adenoma-specific abundance that is different from the abundance in both normal and cancer samples (Figure 4*B*). Two of these (C1QBP and POF1B) are more abundant in adenomas than in normal and cancer samples, whereas the last one (ITGA1) is less abundant in adenomas than in normal and cancer samples. We filtered for an AUC of $\geq$0.8 in the single-protein candidates, which left us with 8 proteins for multiclass ROC. We set the confidence interval for a good model to 0.7–0.95, resulting in 37 models (Supplementary Table 2). The highest-scoring model consisted of the proteins C1QBP, ERGIC1, and ORMDL1.

Discussion

The ACS is a recognized and well-studied phenomenon that gives rise to the vast majority of CRC cases. In the present study, we characterized a cohort of colorectal adenomas with HG dysplasia. Based on long clinical follow-up of 6–16 years, we divided the cohort into a metachronous neoplasia group (new HG dysplasia adenoma or CRC within 5 years) and a nonmetachronous neoplasia group (no HG dysplasia adenoma or CRC for at least 10 years). We sequenced a subset of samples (7 from the G0 group and 25 from the G1 group) and found no structural variations or copy number variations associated with metachronous lesions. It should be noted that this analysis was designed to assess imbalance in the genetic makeup of the adenomas across the 2 groups at the chromosomal level. Given the limited number of samples included and the impossibility of comparing the mutational profile of initial lesions and following lesions, as well as adjacent unaffected tissue, we cannot support or rule out the possibility of an association between a specific mutational spectrum and the occurrence of subsequent adenoma or CRC. We analyzed the proteome of these adenomas with a focus on differences between the 2 groups. Finally, we aimed to characterize these high-dysplasia adenomas in general terms because they are the last benign stage of the ACS. The clear distinction of our 2 groups seen in the UMAP plot indicated that the information contained within the abundance of the $\sim$5000 proteins was sufficient to predict the future development of HG adenomas or CRC. One inference from our findings is that because the proteome of initial lesions could accurately predict the likelihood of developing metachronous advanced lesions, there may be a context dependence to the proteomes of polyps, which reflect the integrative outcome of multiple factors that affect polyp formation and progression. We leveraged our proteome-level analysis to predict outcome because no singular protein could predict the occurrence of subsequent adenoma or CRC. This ability may also be attributed to the newly developed type of dimensionality reduction in the form of UMAP, which appears superior to previous dimensionality reduction methods.[14]

We observed that only the group variable correlated with coexpression modules, which is in line with findings in related studies that showed little or no correlation of the same covariates. Furthermore, we speculated whether our highly specific experimental design, with one of the inclusion criteria for our samples being HG dysplasia, plays a part in this observation. In addition to this, guidelines for logging sample information were not yet fully implemented at the time of sample acquisition, resulting in a discrepancy of level of detail for each sample. An example of this is that for one part of our cohort we have detailed anatomic location of

Table 1. Candidate Proteins

Genes	Log FC	Adjusted *P* value	Weight	Support	Available for AUC	AUC		
						Adenoma vs normal	Cancer vs adenoma	Cancer vs normal
ITGA1	1.88	.77	0.07	3	Yes	0.99	0.83	
STAT2	1.65	.02	0.03	3	Yes			0.78
POF1B	0.84	.2	0.09	3	Yes	0.88	0.81	
DHX16	0.58	.02	0.26	3	Yes			0.76
CSNK1A1	0.55	.6	0.55	3	Yes		0.76	
C1QBP	0.55	.01	0.25	3	Yes	0.88	0.82	
COA4	2.07	.83	0.01	2	Yes			
MTM1	1.53	.01	0.04	2	Yes			
ABAT	1.25	.05	0.01	2	Yes			
TSPAN6	1.25	.03	0.01	2	No			
CAPN5	1.2	.14	0.11	2	Yes		0.83	0.84
TINAGL1	0.82	.04	0.02	2	Yes	0.97		0.97
ERGIC1	0.72	0	0.36	2	Yes		0.79	0.78
COPZ1	0.68	.29	0.23	2	Yes			
RSRC2	0.65	.02	0.1	2	Yes			
SEC16A	0.65	.36	0.04	2	Yes			
SEC16A	0.65	.36	0.04	2	Yes			
FUS	0.58	.01	0.3	2	Yes	0.91		0.94
SUMF2	0.57	.01	0.69	2	Yes			
VASP	0.57	.32	0.11	2	Yes			
GNAI3	0.53	.3	0.44	2	Yes	0.76		0.82
MCTP2	2.98	.53	0.05	1	Yes		0.78	
RAB33B	2.48	.83	0	1	No			
C1orf226	2.32	.77	0.06	1	No			
TTC39A	2.27	.61	0.05	1	No			
ORMDL2	1.51	.03	0.04	1	No			
GLS2	0.95	.02	0.16	0	No			
CD46	0.65	0	0.03	0	Yes			

NOTE. From external sets that support the listed candidates. Log FC and adjusted *P* value refer to results of the DA analysis, while weight pertains to the ENR model.
FC, fold change.

the adenoma, whereas for another part, the information is cruder and divided into colon or rectum. Interestingly, by generalizing the more detailed anatomic location to align with the less detailed one, we saw a marked increase in correlation level across several modules (Localization vs Localization 2, Figure 2*C*). This emphasizes one of the caveats of working with old FFPE samples where available clinical information is inadequate and highlights the importance of acquiring precise, aligned clinical information if anything is to be inferred based on this.

The observation that the type of metachronous neoplasia (AC or HG) did not display subclustering within the G1 cluster (not shown) with UMAP was intriguing. This suggests that the protein abundance differences in the metachronous neoplasia samples were similar irrespective of whether an individual later developed CRC or a new HG adenoma. At the same time, it could also mean that the HG adenoma individuals would eventually develop CRC but that the time to develop cancer is different for these individuals. Another plausible explanation is that these individuals adhered to

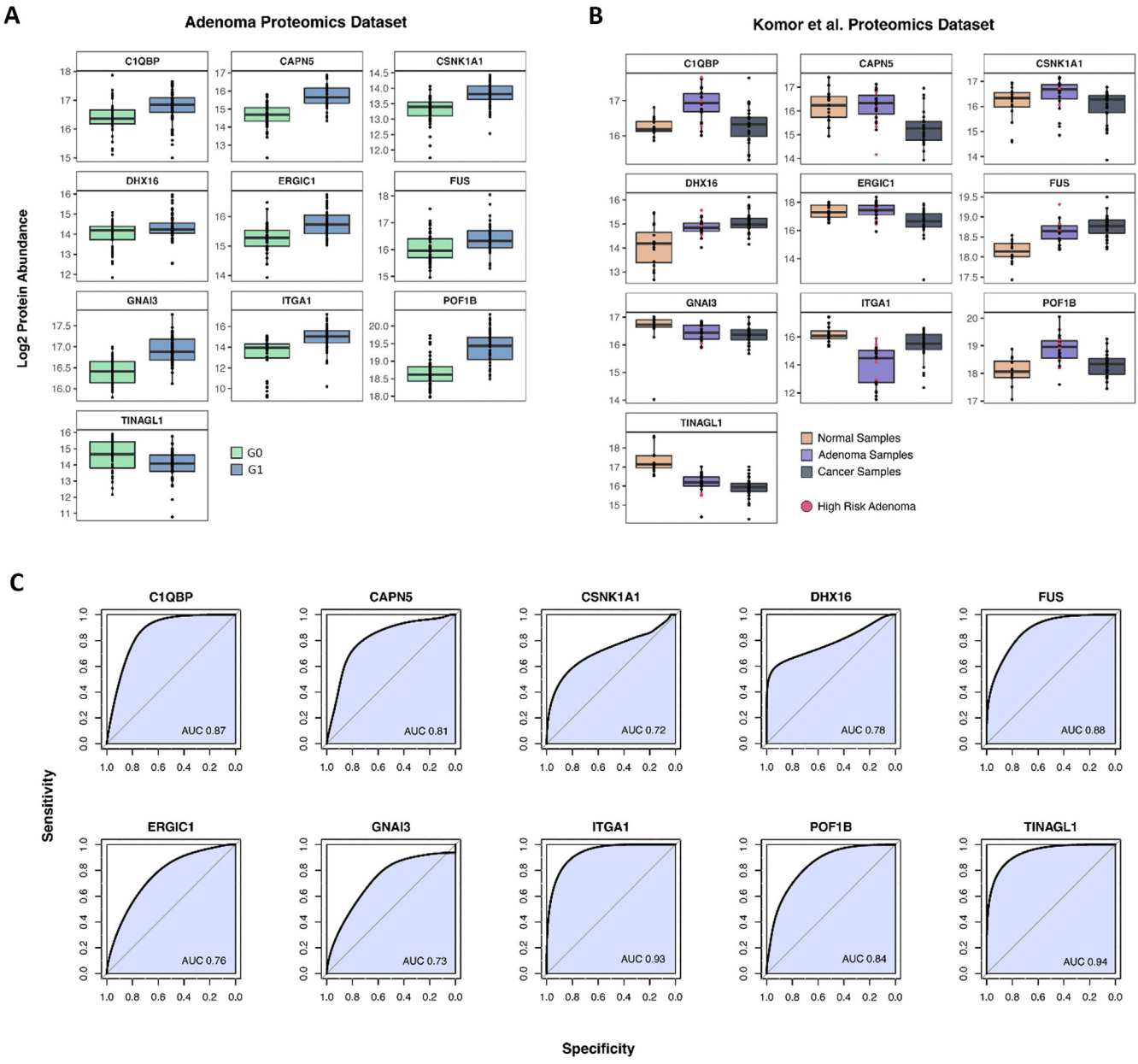

Figure 4. Boxplots of candidate proteins and their ROCs. Boxplots of 10 proteins and their abundance across metachronous advanced neoplasia and nonmetachronous advanced neoplasia groups in (*A*) our proteomics set and across (*B*) normal, adenoma, and cancer samples in a validation set. Adenomas in the validation set are divided into high risk (red dots), with presence of at least 2 CRC-linked chromosomal abnormalities, and low risk, with fewer than 2 of these as defined in Komor et al.[22] (*C*) Receiver operating characteristic curves for 10 proteins with their AUCs displayed in the bottom right corner.

checkups that would discover adenomas before turning malignant or that symptoms earlier in the ACS led to discovery of the adenomas before they could progress to CRC.

Cancer studies are abundant, and network-based analysis favors well-characterized interactions. In our case, we expected that some proteins would be linked to cancer development, representing the transitional stage of adenomas to CRC and those individuals who developed an adenocarcinoma within 5 years. Our samples are, however, not cancers, and we were just as interested in the individuals who developed a new HG adenoma. This issue might have affected our interpretation of the network analysis. Overlaying information of the consensus proteins enabled us to identify candidates who were likely important for the different modules. We observed that the consensus protein CSNK1A1 (a casein kinase) was coexpressed with, and linked to, β-catenin in the dustyRed module. The Wnt/β-catenin pathway is central in CRC development,[28] and CSNK1A1 functions as a negative regulator of Wnt signaling.[29] We wondered whether this kinase and its deregulation in our samples was another potential path to destabilizing β-catenin homeostasis already at early stages of the ACS, leading to adenoma formation and CRC. Its role at the cancer stage has been described, where it seems to correlate with poor survival and affect p53-associated prognosis.[30,31] At the outskirts of this module, we saw

mucin proteins as coexpressed, albeit with very few interactions. This highlights the aforementioned potential challenge of this type of analysis, where the uniqueness of these proteins to mucin-producing organs such as the large intestine might affect their level of annotation.

We were not able to identify studies on colorectal adenomas that were directly comparable to ours. Existing studies on colorectal adenomas where proteomics data are available compared either to cancer, to normal tissue, or between adenoma subtypes. Furthermore, none of the study designs included samples with long follow-up or looked at metachronous advanced neoplasia, which was the basis of the present study. We observed significant discordance of abundance directionality for proteins identified in our analysis compared to external sets, as well as between these sets, emphasizing the issue of comparing studies with different subsets/subtypes of adenomas. A study on single-cell genetic analysis of adenomas based on recurrence could be a reason for optimism.[32] That study also analyzed the recurrence samples, highlighting one of the shortcomings in the present study, because we did not have access to the recurring adenomas or cancers. With these limitations in mind, we analyzed the most interesting candidates by ROC analysis. In the validation set that compares normal, adenoma, and cancer, we saw an abundance of some proteins defy the expected linear progression of the ACS: 2 proteins were more abundant in adenoma than normal, but their abundance in cancer was similar to that in normal. A third protein, ITGA1, displayed the inverse relationship (down-regulation in adenoma compared to normal but abundance back to normal levels in cancer). It is plausible that some events, such as changes in protein abundance, play a role in the transitional stages of the ACS but are later lost.

The protein integrin alpha 1 (ITGA1/CD49a) also appeared differentially abundant between the 2 groups in our cohort, although not clearly. Furthermore, ITGA1 seemed down-regulated in a subgroup of our nonmetachronous advanced neoplasia group, and it would be interesting to follow up on this finding. Its possible involvement in colorectal carcinogenesis has been linked to the Ras/extracellular signal–regulated kinase pathway in CRC, and it is controlled by Myc.[33–35] Again, however, there was an issue of discordance in directionality between available studies. The referred studies found that it was increased in cancer, whereas in the validation set we investigated, we observed that the abundance is lower in cancer than in normal, or similar at best (Figure 4*B*). Finally, ITGA1 has also been linked to progression of pancreatic adenocarcinomas and their premalignant lesions, where it was even suggested as a premalignant biomarker.[36] The lack of similarity between studies was problematic, highlighted by the large discordance in protein abundance of the different studies we compared with. This made it hard to do meaningful comparisons and validate findings. We encourage continued sharing of data and code as a means to combat this issue.

In conclusion, we characterized the proteome of HG dysplasia adenomas divided into 2 groups based on history of metachronous lesions to new HG adenomas or CRC. Dimensionality reduction led to a clear separation of the 2 groups, showing that proteomics was a powerful approach to distinguish samples that were otherwise identical by pathology evaluation. We showed that the proteome of HG adenomas could potentially contain the information needed to predict development of HG adenomas or development of CRC in individuals several years in advance. It was also apparent that no single protein clearly defined metachronous advanced neoplasia from no metachronous advanced neoplasia, supporting the notion that multiple proteins, types of analysis, or even entire proteomes might be needed to predict the behavior of this complex disease. Our results raise the possibility that proteomic profiling of polyps could be used to refine our current population screening strategies for CRC. Proteome-guided risk allocation would enable better deployment of colonoscopic resources and the development of personalized screening programs, a major change from current workflows. Further studies in larger prospective cohorts will be required to validate our findings.

Supplementary Material

Note: To access the supplementary material accompanying this article, visit the online version of *Gastroenterology* at www.gastrojournal.org, and at https://doi.org/10.1053/j.gastro.2023.03.208.

References

1. Sung H, Ferlay J, Siegel RL, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2021;71:209–249.
2. Fearon ER, Vogelstein B. A genetic model for colorectal tumorigenesis. Cell 1990;61:759–767.
3. Strum WB. Colorectal adenomas. N Engl J Med 2016; 374:1065–1075.
4. Spjut H, Estrada RG. The significance of epithelial polyps of the large bowel. Pathol Annu 1977;12:147–170.
5. Vatn MH, Stalsberg H. The prevalence of polyps of the large intestine in Oslo: an autopsy study. Cancer 1982; 49:819–825.
6. Williams AR, Balasooriya BA, Day DW. Polyps and cancer of the large bowel: a necropsy study in Liverpool. Gut 1982;23:835–842.
7. Cardoso R, Guo F, Heisser T, et al. Colorectal cancer incidence, mortality, and stage distribution in European countries in the colorectal cancer screening era: an international population-based study. Lancet Oncol 2021; 22:1002–1013.
8. Hassan C, Antonelli G, Dumonceau JM, et al. Post-polypectomy colonoscopy surveillance: European Society of Gastrointestinal Endoscopy (ESGE) guideline—update 2020. Endoscopy 2020;52:687–700.
9. Bretthauer M, Loberg M, Wieszczy P, et al. Effect of colonoscopy screening on risks of colorectal cancer and related death. N Engl J Med 2022;387:1547–1556.
10. Holme Ø, Schoen RE, Senore C, et al. Effectiveness of flexible sigmoidoscopy screening in men and women and different age groups: pooled analysis of randomised trials. BMJ 2017;356:i6673.

11. Coscia F, Doll S, Bech JM, et al. A streamlined mass spectrometry-based proteomics workflow for large-scale FFPE tissue analysis. J Pathol 2020; 251:100–112.
12. Zhang X, Smits AH, van Tilburg GB, et al. Proteome-wide identification of ubiquitin interactions using UbIA-MS. Nat Protoc 2018;13:530–550.
13. Liu M, Dongre A. Proper imputation of missing values in proteomics datasets for differential expression analysis. Brief Bioinform 2021;22:bbaa112.
14. McInnes L, Healy J, Melville J. UMAP: uniform manifold approximation and projection for dimension reduction. arXiv. Preprint posted online September 18, 2020. https://doi.org/10.48550/arXiv.1802.03426.
15. Ritchie ME, Phipson B, Wu D, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res 2015; 43:e47.
16. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. J Stat Softw 2010;33:1–22.
17. Kuhn M. Building predictive models in R using the caret package. J Stat Softw 2008;28:1–26.
18. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 2008;9:559.
19. Szklarczyk D, Gable AL, Nastou KC, et al. The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. Nucleic Acids Res 2021; 49:D605–D612.
20. Wu T, Hu E, Xu S, et al. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. Innovation (Camb.) 2021;2:100141.
21. Pepe MS. The statistical evaluation of medical tests for classification and prediction. New York: Oxford University Press, 2003.
22. Komor MA, de Wit M, van den Berg J, et al. Molecular characterization of colorectal adenomas reveals POFUT1 as a candidate driver of tumor progression. Int J Cancer 2020;146:1979–1992.
23. Robin X, Turck N, Hainard A, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics 2011;12:77.
24. Uzozie A, Nanni P, Staiano T, et al. Sorbitol dehydrogenase overexpression and other aspects of dysregulated protein expression in human precancerous colorectal neoplasms: a quantitative proteomics study. Mol Cell Proteomics 2014;13:1198–1218.
25. Wisniewski JR, Dus-Szachniewicz K, Ostasiewicz P, et al. Absolute proteome analysis of colorectal mucosa, adenoma, and cancer reveals drastic changes in fatty acid metabolism and plasma membrane transporters. J Proteome Res 2015;14:4005–4018.
26. Sohier P, Sanson R, Leduc M, et al. Proteome analysis of formalin-fixed paraffin-embedded colorectal adenomas reveals the heterogeneous nature of traditional serrated adenomas compared to other colorectal adenomas. J Pathol 2020;250:251–261.
27. Tang M, Zeng L, Zeng Z, et al. Proteomics study of colorectal cancer and adenomatous polyps identifies TFR1, SAHH, and HV307 as potential biomarkers for screening. J Proteomics 2021;243: 104246.
28. Fearon ER. Molecular genetics of colorectal cancer. Annu Rev Pathol 2011;6:479–507.
29. Shen C, Nayak A, Melendez RA, et al. Casein kinase 1α as a regulator of Wnt-driven cancer. Int J Mol Sci 2020; 21:5940.
30. Fariña Sarasqueta A, Forte GI, Corver WE, et al. Integral analysis of p53 and its value as prognostic factor in sporadic colon cancer. BMC Cancer 2013; 13:277.
31. Richter J, Kretz A-L, Lemke J, et al. CK1α overexpression correlates with poor survival in colorectal cancer. BMC Cancer 2018;18:140.
32. Fiedler D, Heselmeyer-Haddad K, Hirsch D, et al. Single-cell genetic analysis of clonal dynamics in colorectal adenomas indicates *CDX2* gain as a predictor of recurrence. Int J Cancer 2019; 144:1561–1573.
33. Boudjadi S, Carrier JC, Groulx JF, et al. Integrin $\alpha1\beta1$ expression is controlled by c-MYC in colorectal cancer cells. Oncogene 2016;35:1671–1678.
34. Boudjadi S, Bernatchez G, Sénicourt B, et al. Involvement of the integrin $\alpha1\beta1$ in the progression of colorectal cancer. Cancers 2017;9:96.
35. Li H, Wang Y, Rong SK, et al. Integrin $\alpha1$ promotes tumorigenicity and progressive capacity of colorectal cancer. Int J Biol Sci 2020;16:815–826.
36. Gharibi A, La Kim S, Molnar J, et al. ITGA1 is a pre-malignant biomarker that promotes therapy resistance and metastatic potential in pancreatic cancer. Sci Rep 2017;7:10060.

Received January 23, 2023. Accepted March 14, 2023.

Correspondence
Address correspondence to: José M.A. Moreira, Department of Drug Design and Pharmacology, Faculty of Health and Medical Sciences, University of Copenhagen, Jagtvej 160, Copenhagen, Denmark. e-mail: jomo@sund.ku.dk.

Acknowledgments
The authors would like to thank Signe Lykke Nielsen for expert technical assistance.

CRediT Authorship Contributions
Jacob Mathias Bech, PhD (Formal analysis: Equal; Investigation: Equal; Software: Supporting; Visualization: Lead; Writing – original draft: Lead; Writing – review & editing: Equal).

Thilde Terkelsen, PhD (Formal analysis: Lead; Investigation: Supporting; Methodology: Lead; Software: Lead; Visualization: Lead; Writing – original draft: Lead; Writing – review & editing: Equal).

Annette Snejbjerg Bartels, BA (Formal analysis: Supporting; Investigation: Equal; Writing – review & editing: Supporting).

Fabian Coscia, PhD (Formal analysis: Equal; Investigation: Lead; Writing – review & editing: Supporting).

Sophia Doll, PhD (Formal analysis: Equal; Investigation: Lead; Writing – review & editing: Supporting).

Siqi Zhao, MSc (Formal analysis: Lead; Investigation: Lead; Software: Lead; Writing – original draft: Equal).

Zhaojun Zhang, PhD (Formal analysis: Lead; Investigation: Equal; Supervision: Lead; Visualization: Lead; Writing – original draft: Equal; Writing – review & editing: Lead).

Nils Brünner, MD PhD (Conceptualization: Lead; Resources: Equal; Writing – review & editing: Equal).

Jan Lindebjerg, MD (Investigation: Equal; Resources: Lead; Writing – review & editing: Equal).

Gunvor Iben Madsen, MD (Conceptualization: Equal; Investigation: Equal; Resources: Lead; Writing – review & editing: Equal).

Xiangdong Fang, PhD (Conceptualization: Lead; Project administration: Lead; Supervision: Lead; Writing – review & editing: Equal).

Matthias Mann, PhD (Conceptualization: Lead; Project administration: Lead; Supervision: Lead; Writing – review & editing: Equal).

Jose Manuel Afonso Moreira, PhD (Conceptualization: Lead; Funding acquisition: Lead; Project administration: Lead; Supervision: Lead; Writing – review & editing: Lead).

Conflicts of interest

The authors disclose no conflicts.

Funding

This work was supported by a grant from the Sawmill Owner Jeppe Juhl and Wife Ovita Juhl Memorial Foundation. Jacob Mathias Bech was supported by a PhD scholarship from Sino-Danish Center. Fabian Coscia acknowledges the European Union's Horizon 2020 Research and Innovation Programme under grant agreement 846795 (Marie Skłodowska-Curie grant) and support by the German Federal Ministry of Education and Research (BMBF), as part of the National Research Initiatives for Mass Spectrometry in Systems Medicine, under grant agreement 161L0222.

Data Availability

Deidentified individual participant proteome data are available indefinitely from the PRIDE repository at www.ebi.ac.uk/pride/ using the data set identifier [PXD017269]. Other individual participant data will not be shared.

Supplementary Methods

Sample Preparation and Liquid Chromatography–Mass Spectrometry Analysis

A detailed sample collection and preparation protocol for FFPE tissues is provided in Coscia et al.[1] Briefly, nanoflow liquid chromatography–MS/MS analysis of tryptic peptides was conducted on a quadrupole Orbitrap mass spectrometer (Q Exactive HF-X, Thermo Fisher Scientific) as described.[1] The mass spectrometer was operated in data-independent mode (DIA). The DIA method consisted of 1 MS1 scan (350 or 300 to 1650 m/z; resolution, 60,000 or 120,000; maximum injection time, 60 ms; automatic gain control target, 3E6) and 32 segments at varying isolation windows from 14.4 m/z to 562.8 m/z (resolution, 30,000; maximum injection time, 54 ms; automatic gain control target, 3E6). Stepped normalized collision energy was 25, 27.5, and 30. The default charge state for MS2 was set to 2.

Mass Spectrometry Data Analysis

DIA raw files were analyzed with Spectronaut Pulsar X software (Biognosys, version 12.0.20491.17) under default settings for targeted DIA analysis with "mutated" as the decoy method. We used an adenoma tissue–specific data dependent acquisition spectral library, encompassing 7725 protein groups (77,275 precursors). Data export was filtered by "No Decoy" and "Quantification Data Filtering" for peptide and protein quantifications. The human UniProtKB database (October 2017, UP000005640_9606) was used as the forward database and the automatically generated reverse database was used for the decoy search.

Isolation of Genomic DNA From Formalin-Fixed, Paraffin-Embedded Adenomas

We cut 5-μm-thick sections from FFPE samples and mounted them on SuperFrost Ultra Plus glass slides (catalog no 15.101.280, Hounisen). Before cutting the samples used for nucleic acid purification, the 3 initial sections of 3 μm each were cut, mounted on SuperFrost Ultra Plus slides, and then stored long term at 4°C for additional analyses. This removed the outermost part of the tissue block, which was the most affected by oxidation because of direct exposure. Slides were deparaffinized according to a standard protocol (twice in xylene for a total 15 minutes and then 2 × 99% ethanol, 2 × 96% ethanol, and 1 × 70% ethanol for 2 minutes in each) followed by air drying (minimum of 30 minutes). The protocol for determining and isolating tissue of interest is described in detail in Coscia et al.[1] Areas isolated for DNA analysis were at the same location as those used for proteomics, differing only by 3 or 4 sections (15–20-μm depth). Briefly, high-resolution images of H&E-stained samples were evaluated to identify regions containing adenomatous tissue, taking care to avoid vascular tissue or normal stratified colon epithelium. For each sample, we scraped 3 sections of 5 μm with a scalpel and collected in DNA LoBind Tubes (catalog no. 0030108051, Eppendorf) with 10–30 μL of ATL buffer (QIAamp DNA FFPE Tissue Kit). Nucleic acids were isolated with the QIAamp DNA FFPE Tissue Kit (catalog no. 56404, Qiagen). The integrity and purity of isolated DNA was assessed by agarose gel electrophoresis. To quantify the isolated DNA, we used a NanoDrop spectrophotometer (Thermo Fisher Scientific).

Bioinformatic Analysis of Whole-Genome Sequencing Data

DNA samples were sequenced at Novogene UK. A stringent quality control step was used, such that only the samples with the highest-quality DNA were chosen for whole-genome sequencing. We further divided samples into 2 groups depending on their age because initial analysis showed that older samples (collected before 2008) were of inferior quality, thereby affecting output. For the older samples, we sequenced up to 150 Gb of data, and for new ones, 100 Gb of data. The aim was to achieve a similar sequencing depth across samples. We sequenced samples at a depth of 15–30× to permit evaluation of structural variations at the chromosome level as well as any resulting copy number alterations. We used FastQC on the raw sequence data to assess yield and raw base qualities. Raw reads were converted to FASTQ using bwa, quality trimmed, and then mapped to the GRCh37 (hg19). We ran Picard CollectMultipleMetrics and CollectWGSMetrics on the aligned binary alignment map to collect alignment and insert size metrics. Picard CollectGcBiasMetrics was run to compute normalized coverage across multiple GC bins. Reads duplication metrics were quantified by running Picard MarkDuplicates on the binary alignment map. Single nucleotide and insertion/deletion (indel) variants were mapped using the Sentieon variant caller using 1000 Genomes as the normal indel reference. Copy number calls were pooled across individuals with bedtools and overlapped with bedIntersect to identify the regions that were recurrently gained/lost. Structural variants were pooled using bedtools and overlapped with intersectBed to identify common regions of structural variation. We identified structural variants with the Manta Structural variant caller[2] and copy number alterations using the CNVnator.[3] Annotation of the variants was performed using ANNOVAR.[4] Results were visualized with the R/Bioconductor package Maftools.[5] All variants were stored in variant call format files. We combined copy number calls with BEDTools[6] and overlaid these with bedIntersect to identify repeatedly lost and gained regions. BEDTools was used to pool structural variants and overlaid with intersectBed to find regions with frequent structural variations.

Bioinformatic Analysis of Proteomics Data From Colon Adenomas

Missing value imputation, normalization, and filtering. The proteomics set was acquired as described in Coscia et al[1] and encompassed 6256 protein groups. We used the R package DEP[7] to explore the nature of missing values in the dataset. Preliminary plots from DEP indicated that proteins with lower average abundances had more missing values, that is, not available values were not random (MNAR). Before imputation, proteins with more than 50% missing abundance counts across samples in each of the 2

groups (G0, G1) were removed from the dataset, and protein groups assigned to the sample gene symbol were averaged. Filtering reduced the proteomics set to 4958 unique proteins. Because of the type of missing values in our dataset, we used the sample minimum method (sampMin) of substitution shown to perform better than other, more complex, methods of imputation and to work well for datasets where values are missing not at random.[8]

Dimensionality reduction with uniform manifold approximation and projection. We performed dimensionality reduction and plotting with UMAP[9] to visualize clustering of adenoma samples based on protein abundance patterns. UMAP analysis was performed using the R package umap.[10] Because sample collection had been divided between 2 Danish hospitals, we checked and corrected for this technical/batch effect before plotting. Batch correction for plotting purposes only was performed using the R framework Combat/sva.[11] The following covariates were evaluated using a dimensionality reduction plot: sample group (nonrecurrence G0 vs recurrence G1), patient age (30–89 years), sex (male, female), year of sample collection (2002–2012), localization of adenoma (colon, rectum, distal, proximal), and recurrence type (adenocarcinoma, HG adenoma).

Variable Selection

Differential abundance analysis. Differential abundance analysis (DAA) was performed using the limma package.[12] Although this R package was originally developed for analysis of microarray and RNA-sequencing data, the underlying statistical framework of limma has proven itself useful for the analysis of proteomics data[13–16] and allows for inclusion of covariates and technical artifacts in the statistical design. We included hospital of origin as a batch effect.

Elastic-net regression. ENR was performed with the R packages glmnet[17] and caret,[18] for example, $\alpha = 0.5$. The dataset was split into a training set (two thirds of samples) and test set (one third of samples) to evaluate the accuracy of the model. A range of lambda values were tested using leave-one-out cross-validation to estimate the optimal value of this parameter ($\lambda = 0.001$).

Consensus set of variables. Results of DAA and ENR were combined, and proteins were scored according to whether both types of analyses identified them, regardless of which batch adjustment method was used.

Weighted gene coexpression network analysis. To evaluate the coabundance of proteins from adenomas, we used the R package WGCNA.[19] The input for analysis was the set of ∼5000 proteins remaining after filtering, normalization, and missing value imputation. The pipeline of analysis followed the example provided by package developers Langfelder and Horvath at https://horvath.genetics.ucla.edu/html/CoexpressionNetwork/Rpackages/WGCNA/Tutorials/.

In brief, we checked the data for sample outliers (3 samples were removed: S132, S201, and S202), and a soft-thresholding power was chosen from a range of tested values (softPower = 12). Correlation of protein abundances were calculated using the biweight midcorrelation (signed network), and the interconnectivity of proteins was estimated (topological overlap). Next, we performed hierarchical clustering using the Ward clustering algorithm.[20] Coabundance modules were defined based on protein dissimilarities, with a minimum module size of 30 proteins (default). Protein modules with high module eigenprotein correlations ($r > 0.8$) were merged. Finally, we correlated clinical variables with eigenproteins (recalculated from the merged modules) and tested for significance of correlation, defined as an adjusted P value of $\leq .01$ and an absolute correlation score of ≥ 0.4.

Protein-protein interaction networks. The STRING database[21] of protein-protein (P-P) interactions was downloaded (https://stringdb-static.org/download/protein.links.full.v11.0/9606.protein.links.full.v11.0.txt.gz) on March 30, 2022 (11,759,454 P-P pairs). For our analysis, we extracted experimentally validated P-P pairs (minimum of 1 experiment in human or animal tissue/cells) with a STRING score of ≥ 0.6 and removed directed edges, reducing the number of interaction pairs to 522,722.

Next, we extracted P-P pairs from strings in which both proteins were either DA, retained in the ENR model, or coabundant in the 9 modules correlated with the adenoma group variable. To get a more condensed network, we removed proteins with ≤ 5 interactions. A final set of 1565 proteins and 4095 P-P interactions (edges) were exported for visualization with Cytoscape version 3.9.1.[22]

Gene Ontology and pathway enrichment analysis. Sets of proteins from modules significantly correlated with group variable, differentially abundant proteins, and proteins retained in the elastic-net model were used for GO term and pathway enrichment analysis. The full set of 4958 proteins served as the background for enrichment testing.

Enrichment analyses was performed with the R package clusterProfiler,[23] here specifically using the Kyoto Encyclopedia of Genes and Genomes database for pathway enrichment analysis. The cutoff for a significant pathway or GO term was a Q value of <0.05 and a minimum group size of ≥ 5. Out of 10 sets of proteins used for enrichment analysis, 6 sets were enriched within either GO terms and/or Kyoto Encyclopedia of Genes and Genomes pathways (dustyRed, greenBlue, nudePink, oliveGreen, sandYellow, skyBlue). GO term similarity analysis was performed by comparing proteins annotated in all pairwise combinations of ontology terms across all sets of proteins. Pairs with a similarity score greater than 0.2 were included in the GO similarity plot.

Support of protein candidates and receiver operating characteristic curves. In total, 54 proteins meet the criteria for a well-supported candidate across analyses, that is, the protein was DAA or selected by elastic net and was retained in the coexpression/interaction analyses. The protein candidates were overlapped with lists of differentially abundant proteins from external studies on colorectal adenomas/cancers.[24–27] Out of these 54 proteins, 28 were supported by at least 1 external study.

Sensitivity and specificity of the top protein candidates was evaluated using ROC.[28] The external data used for the analysis was a set of TMT-labeled nanoscale liquid chromatography–MS/MS proteomics from 18 normal mucosa biopsy samples, 30 adenomas, and 30 cancers of the colon.[29] Before analysis, this dataset was filtered and normalized and missing values were

imputed with the R package DEP[7] using the standard pipeline for protein group TMT-labeled proteomics data (https://www.bioconductor.org/packages/release/bioc/vignettes/DEP/inst/doc/MissingValues.html). Out of the 28 top protein candidates, 23 were contained within the validation dataset and, as such, 5 candidates could not be evaluated using ROC (RAB33B, C1orf226, TTC39A, TSPAN6, GLS2). Pairwise and multiclass ROC analyses were performed with the packages nnet and pROC,[30] dividing the dataset into a training set and a test set: two thirds of samples and one third of samples, respectively.

Supplementary References

1. Coscia F, Doll S, Bech JM, et al. A streamlined mass spectrometry-based proteomics workflow for large-scale FFPE tissue analysis. J Pathol 2020;251:100–112.
2. Chen X, Schulz-Trieglaff O, Shaw R, et al. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. Bioinformatics 2016;32:1220–1222.
3. Abyzov A, Urban AE, Snyder M, et al. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome Res 2011;21:974–984.
4. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 2010;38:e164.
5. Mayakonda A, Lin D-C, Assenov Y, et al. Maftools: efficient and comprehensive analysis of somatic variants in cancer. Genome Res 2018;28:1747–1756.
6. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 2010;26:841–842.
7. Zhang X, Smits AH, van Tilburg GB, et al. Proteome-wide identification of ubiquitin interactions using UbIA-MS. Nat Protoc 2018;13:530–550.
8. Liu M, Dongre A. Proper imputation of missing values in proteomics datasets for differential expression analysis. Brief Bioinform 2021;22:bbaa112.
9. McInnes L, Healy J, Melville J. UMAP: uniform manifold approximation and projection for dimension reduction. arXiv. Preprint posted online September 18, 2020. https://doi.org/10.48550/arXiv.1802.03426.
10. Konopka T. UMAP: uniform manifold approximation and projection. R package v.0.2.4.1 (2020). Available at: https://cran.r-project.org/web/packages/umap/index.html.
11. Leek JT, Johnson WE, Parker HS, et al. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics 2012;28:882–883.
12. Ritchie ME, Phipson B, Wu D, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res 2015;43:e47.
13. D'Angelo G, Chaerkady R, Yu W, et al. Statistical models for the analysis of isobaric tags multiplexed quantitative proteomics. J Proteome Res 2017;16:3124–3136.
14. Berg P, McConnell EW, Hicks LM, et al. Evaluation of linear models and missing value imputation for the analysis of peptide-centric proteomics. BMC Bioinformatics 2019;20(Suppl 2):102.
15. Dowell JA, Wright LJ, Armstrong EA, et al. Benchmarking quantitative performance in label-free proteomics. ACS Omega 2021;6:2494–2504.
16. Theodorakis E, Antonakis AN, Baltsavia I, et al. ProteoSign v2: a faster and evolved user-friendly online tool for statistical analyses of differential proteomics. Nucleic Acids Res 2021;49:W573–W577.
17. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. J Stat Softw 2010;33:1–22.
18. Kuhn M. Building predictive models in R using the caret package. J Stat Softw 2008;28:1–26.
19. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 2008;9:559.
20. Murtagh F, Legendre P. Ward's hierarchical agglomerative clustering method: which algorithms implement ward's criterion? J Classif 2014;31:274–295.
21. Szklarczyk D, Gable AL, Nastou KC, et al. The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. Nucleic Acids Res 2021;49:D605–D612.
22. Shannon P, Markiel A, Ozier O, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 2003;13:2498–2504.
23. Wu T, Hu E, Xu S, et al. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. Innovation (Camb.) 2021;2:100141.
24. Uzozie A, Nanni P, Staiano T, et al. Sorbitol dehydrogenase overexpression and other aspects of dysregulated protein expression in human precancerous colorectal neoplasms: a quantitative proteomics study. Mol Cell Proteomics 2014;13:1198–1218.
25. Wisniewski JR, Dus-Szachniewicz K, Ostasiewicz P, et al. Absolute proteome analysis of colorectal mucosa, adenoma, and cancer reveals drastic changes in fatty acid metabolism and plasma membrane transporters. J Proteome Res 2015;14:4005–4018.
26. Sohier P, Sanson R, Leduc M, et al. Proteome analysis of formalin-fixed paraffin-embedded colorectal adenomas reveals the heterogeneous nature of traditional serrated adenomas compared to other colorectal adenomas. J Pathol 2020;250:251–261.
27. Tang M, Zeng L, Zeng Z, et al. Proteomics study of colorectal cancer and adenomatous polyps identifies TFR1, SAHH, and HV307 as potential biomarkers for screening. J Proteomics 2021;243:104246.
28. Pepe MS. The statistical evaluation of medical tests for classification and prediction. New York: Oxford University Press, 2003.
29. Komor MA, de Wit M, van den Berg J, et al. Molecular characterization of colorectal adenomas reveals POFUT1 as a candidate driver of tumor progression. Int J Cancer 2020;146:1979–1992.
30. Robin X, Turck N, Hainard A, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics 2011;12:77.

Supplementary Table 1. Sample Summary

Clinical variables	n (%)
Overall	95 (100)
Group	
G0	44 (46.3)
G1	51 (53.7)
Age, *y*	
30–50	6 (6.3)
50–60	22 (23.2)
60–70	27 (28.4)
70–80	25 (26.3)
80–90	15 (15.8)
Localization 1	
Colon	52 (54.7)
Rectum	43 (45.3)
Localization 2	
Colon	37 (38.9)
Distal	54 (56.8)
Proximal	4 (4.2)
Ki67 percentage	
High	24 (25.3)
Low	20 (21.1)
Mid	47 (49.5)
Recurrence type	
Adenocarcinoma	18 (18.9)
HG	29 (30.5)
Missing	48 (50.5)
Recurrence time, *y*	
1	15 (15.8)
2–5	24 (25.3)
6–10	8 (8.4)
Missing	48 (50.5)

Supplementary Table 2. Receiver Operating Characteristic Models

AUC	Lower CI	Upper CI	Model
0.97	0.92	1	C1QBP + ERGIC1 + ORMDL1
0.97	0.92	1	C1QBP + CSNK1A1 + ERGIC1 + ORMDL1
0.98	0.93	1	C1QBP + ERGIC1 + ITGA1 + ORMDL1
0.97	0.92	1	C1QBP + ERGIC1 + POF1B + ORMDL1
0.97	0.92	1	C1QBP + ERGIC1 + FUS + ORMDL1
0.97	0.92	1	C1QBP + ERGIC1 + ORMDL1 + TINAGL1
0.98	0.95	1	C1QBP + CAPN5 + ERGIC1 + ITGA1 + ORMDL1
0.98	0.93	1	C1QBP + CSNK1A1 + ERGIC1 + ITGA1 + ORMDL1
0.97	0.92	1	C1QBP + CSNK1A1 + ERGIC1 + POF1B + ORMDL1
0.97	0.92	1	C1QBP + CSNK1A1 + ERGIC1 + FUS + ORMDL1
0.97	0.92	1	C1QBP + CSNK1A1 + ERGIC1 + ORMDL1 + TINAGL1
0.98	0.93	1	C1QBP + ERGIC1 + ITGA1 + POF1B + ORMDL1
0.98	0.93	1	C1QBP + ERGIC1 + ITGA1 + FUS + ORMDL1
0.98	0.92	1	C1QBP + ERGIC1 + ITGA1 + ORMDL1 + TINAGL1
0.97	0.92	1	C1QBP + ERGIC1 + POF1B + FUS + ORMDL1
0.97	0.92	1	C1QBP + ERGIC1 + POF1B + ORMDL1 + TINAGL1
0.97	0.92	1	C1QBP + ERGIC1 + FUS + ORMDL1 + TINAGL1
0.98	0.95	1	C1QBP + CAPN5 + CSNK1A1 + ERGIC1 + ITGA1 + ORMDL1
0.98	0.95	1	C1QBP + CAPN5 + ERGIC1 + ITGA1 + FUS + ORMDL1
0.98	0.93	1	C1QBP + CSNK1A1 + ERGIC1 + ITGA1 + POF1B + ORMDL1
0.98	0.93	1	C1QBP + CSNK1A1 + ERGIC1 + ITGA1 + FUS + ORMDL1
0.98	0.93	1	C1QBP + CSNK1A1 + ERGIC1 + ITGA1 + ORMDL1 + TINAGL1
0.97	0.92	1	C1QBP + CSNK1A1 + ERGIC1 + POF1B + FUS + ORMDL1
0.97	0.92	1	C1QBP + CSNK1A1 + ERGIC1 + POF1B + ORMDL1 + TINAGL1
0.97	0.92	1	C1QBP + CSNK1A1 + ERGIC1 + FUS + ORMDL1 + TINAGL1
0.98	0.93	1	C1QBP + ERGIC1 + ITGA1 + POF1B + FUS + ORMDL1
0.98	0.92	1	C1QBP + ERGIC1 + ITGA1 + POF1B + ORMDL1 + TINAGL1
0.98	0.92	1	C1QBP + ERGIC1 + ITGA1 + FUS + ORMDL1 + TINAGL1
0.97	0.92	1	C1QBP + ERGIC1 + POF1B + FUS + ORMDL1 + TINAGL1
0.98	0.95	1	C1QBP + CAPN5 + CSNK1A1 + ERGIC1 + ITGA1 + FUS + ORMDL1
0.98	0.93	1	C1QBP + CSNK1A1 + ERGIC1 + ITGA1 + POF1B + FUS + ORMDL1
0.98	0.92	1	C1QBP + CSNK1A1 + ERGIC1 + ITGA1 + POF1B + ORMDL1 + TINAGL1
0.98	0.93	1	C1QBP + CSNK1A1 + ERGIC1 + ITGA1 + FUS + ORMDL1 + TINAGL1
0.97	0.92	1	C1QBP + CSNK1A1 + ERGIC1 + POF1B + FUS + ORMDL1 + TINAGL1
0.98	0.92	1	C1QBP + ERGIC1 + ITGA1 + POF1B + FUS + ORMDL1 + TINAGL1
0.97	0.92	1	CAPN5 + CSNK1A1 + ERGIC1 + ITGA1 + POF1B + FUS + ORMDL1
0.98	0.92	1	C1QBP + CSNK1A1 + ERGIC1 + ITGA1 + POF1B + FUS + ORMDL1 + TINAGL1
0.97	0.92	1	CAPN5 + CSNK1A1 + ERGIC1 + ITGA1 + POF1B + FUS + ORMDL1 + TINAGL1

CI, confidence interval.