SUPPLEMENTARY MATERIAL AND METHODS

Biopsy Sampling

Colonoscopy preparation was completed the day prior to endoscopy, as per a standard protocol[1] that was modified by shortening to a 1-day period. All biopsy samples were taken from the ascending colon during endoscopy after aspiration of loose fluid and debris and mucosal washing with sterile water. When macroscopically visualized, biopsies were collected from affected areas of the right colon (CoA) in IBD patients and from non-inflamed areas (CoN) of the right colon in CD and control patients. Biopsies were flash frozen on dry-ice in the endoscopy suite and immediately stored at -80°C until further processing.

Reference proteome preparation

Lysates from human HCT-116 colorectal carcinoma, HEK-293 embryonic kidney, and HuH-7 hepatocarcinoma cell lines were isotopically labeled in cell culture[2] and a combined pool used as a "heavy" reference proteome. Briefly, cells were grown in custom prepared methionine-, lysine-, and arginine-deficient DMEM media (AthenaES, Baltimore, MD, USA) that was supplemented with 30 mg/L methionine, 146 mg/L [$^{13}C_6$,$^{15}N_2$]-L-lysine, 42 mg/L (Hek293, HuH7) or 84 mg/L (HCT116) [$^{13}C_6$,$^{15}N_4$]-L-arginine (Sigma Aldrich, Oakville, ON, Can), 10% dialyzed FBS (GIBCO-Invitrogen; Burlington, ON,CAN), 1 mM sodium pyruvate (Gibco-Invitrogen), and 28 µg/mL gentamicin (Gibco-Invitrogen). Complete incorporation (>95%) of heavy amino acids was obtained by cell growth for a minimum of 10 doublings, and verified by mass spectrometry analysis of tryptic peptides, as outlined previously[2]. For proteome isolation, cells were grown to 80% confluency, washed twice with PBS prior to lysis with lysis buffer (4% SDS, 50 mM Tris, pH 8.0 in the presence of proteinase inhibitor cocktail (Roche)). Lysates were sonicated 3 times in 10 s pulses with cooling on ice between pulses. Following centrifugation at 10,000 x g for 10 minutes, protein-containing supernatant was transferred to a fresh tube. Protein concentrations were determined by *DC* protein assay (Bio-Rad). Proteins from cell lysates were diluted with additional lysis buffer to a concentration of 3 µg/µl, and stored in working aliquot volumes at -80 ^{o}C.

Biopsy processing

Biopsies held on dry ice were thawed in the presence of protease inhibitor-containing lysis buffer, as above. Following mechanical homogenization with a pellet pestle, biopsy lysates were sonicated for 10 s x3, held on ice between pulses. Samples underwent centrifugation at 10000 x g for 10 minutes and protein-containing supernatant isolated. Protein concentrations were determined by *DC* protein assay (Bio-Rad). 45 µg of sample protein was combined with 15 µg of protein from each reference cell lysate (HCT-116, HEK-293, HuH-7). The resulting 90 µg of protein underwent tryptic (TPCK-treated trypsin, Worthington) digestion by filter-aided sample preparation with minor modifications[3]. Eluted peptides were acidified prior to fractionation using SCX resin (Agilent Technologies, CA, USA), desalted with 10 µm AQUA-C_{18} (Dr Maisch, GmbH, Ammerbuch, Germany) resin and then dried by speed-vac prior to storage at -80 C.

LC MS/MS analysis

High-performance liquid chromatography/electrospray ionization tandem mass spectrometry (HPLC-ESI-MS/MS) of peptides was performed on an automated Ekspert nanoLC 400 (Eksigent, Dublin, CA, USA) coupled to an LTQ Velos Pro Orbitrap Elite MS (ThermoFisher Scientific, San Jose, CA) equipped with a nano-ESI interface operated in positive ion mode. 4 µl

of peptides, resuspended in 20 µl 0.5% formic acid, were separated on an in-house analytical column (75 µm x 10 cm) packed with C_{18} beads (1.9 µm, 100 angstrom pore size; (Dr Maisch, GmbH, Ammerbuch, Germany) using a 120 min gradient of 5-30% acetonitrile (v/v) in 0.1% formic acid 9 (v/v) at a flow rate of 300 nL/min. The spray voltage was set to 2.2 kV and the temperature of heated capillary was 300°C. A full MS scan (R = 60,000, range 350 to 1750 m/z) was followed by data-dependent MS/MS scan of the 20 most intense ions, a dynamic exclusion repeat count of 1, a repeat duration of 30 s and exclusion duration of 90 s. All data were recorded with Xcalibur software (ThermoFisher Scientific, San Jose, CA).

Bioinformatic analysis

All MS raw files were analyzed in a single run with MaxQuant version 1.5.1, against the human Uniprot database (downloaded 2014/07/11). Parameters used were: cysteine carbamidomethylation as a fixed modification; methionine oxidation and acetylation (protein N-termini) as variable modifications; enzyme specificity as trypsin with a maximum of two missed cleavages; heavy Lys-8 and Arg-10 as isotopic labels; minimum peptide length of seven amino acids; ion mass tolerance of 0.5 Da; protein and peptide false discovery rate (FDR) of 1%.

The MaxQuant output protein group file was imported into Perseus (version 1.3.0.4) for data filtering, log2 transformation, basic statistical analysis (ANOVA, T-test, Pearson correlation, hierarchical clustering) of the normalized ratios versus SILAC reference cell lysates, and for protein group annotation (KEGG, GOBP, Keywords). Only proteins that were quantified by 2 or more unique peptides were included in further analyses. The variance for each biopsy proteome was calculated, and Robust regression and Outlier (ROUT [4]) applied in Prism (Graphpad) to identify any biopsies for which the variance, and thus the MS quality of the data, are significantly different from the population. Outliers were removed prior to statistical analysis.

To ensure consistency of the MS data over time, the patterns of relative expression of proteins within all biopsy proteomes were compared. To avoid inherent issues associated with MS undersampling[5], or sample preparation variation, the top 25% most abundant proteins were identified by total intensity, and the $\log_2$ ratios of these were evaluated by heat map and hierarchical clustering. Pearson correlation between all proteomes was performed (Perseus) to ensure no gross proteomic differences were apparent between samples.

The entire proteome was segregated for subgroup analysis in Excel (Microsoft). Proteins with a relative contribution to protein ID by a given subgroup was determined, and those proteins that represent > 70% of the relative contribution and having identification in <50% of subgroup biopsies were considered to be potentially relevant and included in all further analyses ("subgroup specific"). Data sets with limited immunological proteins were created by removing all proteins that had the terms "immun" or "inflam" in the categories of GOBP, KEGG, Fasta, or Keyword. To evaluate proteins that are significantly different between non-IBD controls and IBD proteomes, sample data from CoA CD and UC biopsies were merged as the IBD group. The Q95 was calculated based upon each subgroup being analyzed, wherein proteins included were quantified in ≥95% of the biopsies for that analysis; the subgroup specific proteins were added to the Q95 for further processing. Missing value imputation and principle component analysis, using *knnimpute* and *princomp* respectively, were performed with MatLab.

To determine the minimum number of patients required for accurate diagnosis at the discovery stage, iterative analysis with Partial Least Squares Discriminant Analyses (PLSDA), Support Vector Machine (SVM), and Random Forest (RF) was applied. It was determined that a minimum of 30 IBD subjects were required to ensure the greatest consistency in the accuracy of diagnosis (supplementary figure 4). Thus, a balanced stratification approach for gender and diagnosis (Etcetera in WinPepi, BrixtonHealth.ca) was applied for the random assignment of patient biopsies to either the discovery or validation cohort. After allocation, the Q95 and subgroup specific proteins of each cohort were determined.

The proteomics data from the discovery cohort was analyzed by PLSDA, SVM and RF in the ROC Explorer module of Roccet[6] for candidate biomarker selection and testing. For each model, the performance is tested with repeated random sub-sampling cross validation with 2/3 of the samples used for training and 1/3 for testing, with 50 permutations. Ultimately, the candidate biomarkers that were selected were identified as significant in all three models, and ranked by the Area Under the Curve (AUC) value (starting with the highest).

REFERENCES

1. Jimenez-Rivera C, Haas D, Boland M, et al. Comparison of two common outpatient preparations for colonoscopy in children and youth. Gastroenterol Res Pract 2009;2009:518932.
2. Ong SE, Mann M. A practical recipe for stable isotope labeling by amino acids in cell culture (SILAC). Nat Protoc 2006;1:2650-60.
3. Wisniewski JR, Zougman A, Nagaraj N, et al. Universal sample preparation method for proteome analysis. Nat Methods 2009;6:359-62.
4. Motulsky HJ, Brown RE. Detecting outliers when fitting data with nonlinear regression - a new method based on robust nonlinear regression and the false discovery rate. BMC Bioinformatics 2006;7:123.
5. Beck M, Claassen M, Aebersold R. Comprehensive proteomics. Curr Opin Biotechnol 2011;22:3-8.
6. Xia J, Broadhurst DI, Wilson M, et al. Translational biomarker discovery in clinical metabolomics: an introductory tutorial. Metabolomics 2013;9:280-299.